Pre-hoc and Post-hoc Diagnosis and Interpretation of Breast Magnetic Resonance Volumes

THE UNIVERSITY of ADELAIDE

Author: Gabriel Maicas Suso

Supervisors: Assoc. Prof. Gustavo Carneiro

Prof. Ian Reid

Australian Institute for Machine Learning

School of Computer Science

This dissertation is submitted for the degree of

Doctor of Philosophy

The University of Adelaide 2nd December 2018

Table of contents

Publications

- **Gabriel Maicas**, Gustavo Carneiro, Andrew P. Bradley. Globally optimal breast mass segmentation from DCE-MRI using deep semantic segmentation as shape prior. IEEE International Symposium on Biomedical Imaging (ISBI), 2017. **Oral Presentation.**
- **Gabriel Maicas**, Gustavo Carneiro, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid. Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.
- **Gabriel Maicas**, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Training Medical Image Analysis Systems like Radiologists. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018. **Oral Presentation.**
- **Gabriel Maicas**, Gerard Snaauw, Andrew P. Bradley, Ian Reid, Gustavo Carneiro. Model Agnostic Saliency for Weakly Supervised Lesion Detection from Breast DCE-MRI. Under Review at IEEE International Symposium on Biomedical Imaging (ISBI).
- **Gabriel Maicas**, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Pre and Post-hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI. Under Review at Medical Image Analysis.

Abstract

Breast cancer is among the leading causes of death in women. Aiming at reducing the number of casualties, breast screening programs have been implemented to diagnose asymptomatic cancers due to the correlation of higher survival rates with earlier tumour detection. Although these programs are normally based on mammography, magnetic resonance imaging (MRI) is recommended for patients at high-risk. The interpretation of such MRI volumes is time-consuming and prone to inter-observer variability, leading to missed cancers and a relatively high number of false positives provoking unnecessary biopsies. Consequently, computer-aided diagnosis systems are being designed to help improve the efficiency and the diagnosis outcomes of radiologists in breast screening programs.

Traditional automated breast screening systems are based on a two-stage pipeline consisting of the localization of suspicious regions of interest (ROIs) and their classification to perform the diagnosis (i.e. decide about their malignancy). This process is typically ineffective due to the usual expensive inference involved in the exhaustive search for ROIs and the employment of non-optimal hand-crafted features in both stages. These issues have been partially addressed with the introduction of deep learning methods that unfortunately need large strongly annotated training datasets (voxel-wise labelling of each lesion), which tend to be expensive to acquire. Alternatively, the use of weakly labelled datasets (i.e volume-level labels) allows diagnosis to become a supervised classification problem, where a malignancy probability is estimated after examining the entire volume. However, large weakly labelled training sets are still required. Additionally, to facilitate the adoption of such weakly trained systems in clinical practice, it is desirable that they are capable of providing the localization of lesions that justifies the automatically produced diagnosis for the whole volume. Nonetheless, current methods lack the precision required for the problem of weakly supervised lesion detection.

Motivated by these limitations, we propose a number of methods that address these deficiencies. First, we propose two strongly supervised deep learning approaches that not only can be trained with relatively small datasets, but are efficient in the localization of suspicious tissue. In particular, we propose: 1) the global minimization of an energy functional containing information from the semantic segmentation produced by a deep

learning model for lesion segmentation, and 2) a reinforcement learning model for suspicious region detection. Diagnosis is performed by classifying suspicious regions yielded by the reinforcement learning model.

Second, aiming to reduce the burden associated to strongly annotating datasets, we propose a novel training methodology to improve the diagnosis performance on systems trained with weakly labelled datasets that contain a relatively small number of training samples. We further propose a novel 1-class saliency detector to automatically localize lesions associated with the diagnosis outcome of this model. Finally, we present a comparison between both of our proposed approaches for diagnosis and lesion detection.

Experiments show that whole volume analysis with weakly labelled datasets achieves better performance for malignancy diagnosis than the strongly supervised methods. However, strongly supervised methods show better accuracy for lesion detection.

List of Figures

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Author: Gabriel Maicas Suso

1st October 2018

Acknowledgements

Pursuing my PhD. has been an amazing, interesting, exciting and challenging time of my life. It has definitely seen me grow, enjoy beautiful moments and fulfil an almost lifetime dream. I would like to take this opportunity to express my appreciation to the people that have made this journey much more than just a research project, but an incredibly valuable life experience.

My panel Gustavo and Ian, and close collaborators Andrew and Jacinto, have been outstanding. I am thankful to have had four different perspectives towards the same goal.

Gustavo, I owe you my deepest gratitude for the extensive influence and enormous contribution to my career, thank you for giving me the opportunity to come to Adelaide. Thank you for your patience, tolerance, understanding and support in the difficult moments, as well as always being ready to help and answer my questions. Thank you for your research guidance (including the organization of BIA 18), feedback, and for letting me explore until we came up with a good idea. Thank you for your work and support to finish our work on time, and especially thank you for teaching me how to make an impact.

Ian, thank you for your simple but thorough questions that always pointed me in the right direction. I can humbly say that I have improved my thinking by analysing how you try to understand our work and how you combine our ideas with your previous knowledge, which is one of the most valuable features I could learn about research. My deepest appreciation because I will never forget to specify and clarify my contributions. Of course, thank you also for organizing those gatherings at your place every year.

Andrew, thank you for adding a needed clinical perspective and higher level view that drives our research to have a more expansive impact. I appreciate that our chats remind me that although methods are important, our goal is to solve a real life problems such that clinical practice can benefit from our research.

Jacinto, thank you for adding passion to learn and think in this project and for being always ready to collaborate. In Adelaide, we shared one of the most exciting times when our reinforcement learning work started to take off. Also, my deepest appreciation for your finest inspections of papers, without those small details they would never be complete.

I am also deeply grateful to Luke and Johan for your help to guide my future. It was not an easy decision, but you guys made it look very, very simple. Also, Luke, big thanks for the advice and comments about the radiological background.

I would also like to thank the administrative staff from our school for their great work. Thanks Thuy and Sharyn for preparing our trips; and thanks Stuart, Lenka, Julie, Melissa and Hilary for your help in the front office to facilitate the paperwork.

My deepest gratitude to Emily and Crispin for spreading the word of our work. It is great that our work had such repercussion. My gratitude also to Clare and all the media that covered our detection work.

I am deeply grateful to Kyle and Kristine, from Data Skeptic, for their interest in our work. It was a great and motivating experience to participate in an interview for your podcast.

Special thanks to the Australian Research Council for all of the funding that has made this project feasible. Special thanks also to The University of Adelaide for the facilities provided to develop my research during the PhD, as well as the "Kaurna people, the original custodians of the Adelaide Plains and the land on which the University of Adelaide's campuses at North Terrace is built".

Tutoring has been a great experience to complement this PhD. Thank you Cruz and Pádraig for facilitating my opportunity to teach. Pádraig, it was great sharing the NLP master project with you, I really enjoyed looking at a topic that is related but not the same as my PhD.

Neeraj, thank you for listening and your aim to help me succeed during the time we shared in Adelaide. Thanks for sharing your PhD (and professional) experience with me. Certainly, following your example of how to work on a PhD has been enlightening. However, one of the most important things I learnt from you is the recipe to prepare the Nepalese dumplings!

Thank you Lu for your help and support at the beginning of the PhD. Your advice turned out to be one of the most important, and taught me how to overcome some of the difficulties I encounter during my adaptation to the PhD and life in Adelaide.

Thanks also to Zhibin for being happy to help me with my research and making me feel always welcome even when asking the same question over and over again. Also, thanks to William, with whom I shared a great time in Willunga watching the Tour Down Under, who was always ready to enjoy good times and nice dinners.

Gerard, thank you for bringing fresh air at the right moment of my PhD. From the moment we met, I realized that your different background can add a lot of value to my research. I really enjoyed our cool conversations, your honest opinion, and of course, your good work. I

am very grateful for receiving your feedback to improve. Best of all, it was an amazing and fun time in Granada, including the preparation for the presentation.

I would also like to express my gratitude to my lab mates, with who I have been happy to share some lunches, dinners or coffees: Ming (and our basketball conversations), Shing Fan, Swichaya, Samya, Saroj and Huu (our motivation for the most popular joke... "who?"). Special thanks also to Dzung, Hoa, Cuong (what a great sense of humor) and Toan for making me feel welcome at lunch time.

Ergnoor, thanks for your words in difficult moments, they helped me to change my focus. Also, thank you for sharing your culture, some of your stories, your passion for Albania, and some nice dinners during the time I spent in Adelaide.

Thank you Cosimo, for bringing that special sense of humour to our lab. Even though your visit to Adelaide was short, we still remember and laugh at your jokes from time to time.

Rafa, thanks for your will to be active and share nice moments during our PhDs. I remember our first night out and that (sad) Sunday we went to dance bachata... it turned out to be a part of the butterfly effect! (Caranguejo!). Also, thanks for those mid-day chocolates!

I would also like to thank Ali for our nice discussions and for sharing the will to do cool research. Our conversations pushed me to get keep going at those times where I felt that I could not handle all the things that were happening at the same time. Also, my greatest gratitude to you for your help preparing the podcast interview.

Thanks also to Javi and Ana for the Friday's we spend at the Irish pub, they made my adaptation easier during my first year in Adelaide. My appreciation as well to those Migas and tortillas, they were just amazing!

I would like to express my deepest gratitude also to Mehdi for all the help with doctors, changing accommodation or whatever was needed during my time in Adelaide. Special thanks also for our nice laughs, interesting conversations (including about Persian history) and for teaching me some nice wrestling skills during our breaks. But as expected, my deepest appreciation is for the yummy kebabs we ate together (subways were nice too).

Toan, thanks for all the coffees we drank together to power our days. Thanks also for our lunches, I really enjoyed tasting your delicious Vietnamese food. I also take with me all those great tennis conversations, they were a great way to begin the day with. In general, thanks for all those small but important details such as the good and bad moments of the paper reviewing process, or your visit after my surgery. More importantly, please, do not loose your sense of humour.

Hayden, thanks for choosing to do a PhD. Huge thanks for all your help (in any aspect) at any moment. Thanks for sharing your good values, I keep learning from you every time we

talk, even when you go nuts (which I enjoy too). Although your jokes will never be as funny as my dad jokes, thanks also for your (intelligent – at least sometimes) sense of humour that makes easier going to the lab even when I am tired. Many "instants" come to my mind, but special thanks for discovering those salads, for protecting me from "dangerous" wildlife and for our (not very frequent) basketball games and Mario Karts. Among the special moments, one of the greatest appreciation is for inviting me to such a cool Showdown, I would not be a (Power) fan if it wasn't for you. Although you know it, my deepest appreciation for your feedback and English corrections (including these acknowledgements).

Michelle, thank you for arriving at the front desk at the same time on that 2nd of March, 2015. I cannot imagine my PhD without you. Thanks for your listening and support during my weakest moments. Even more importantly, thanks for our conversations and laughs, we will definitely live longer. I really enjoyed (and I now miss) sharing our stories or just taking a break to talk about random stuff (how many bananas?). I am also deeply grateful that you invited me to the adventure film screening... that made me wonder what I am doing with my life.

Thanks Frank, Sharon and Peter for opening the doors of your home and sharing your outstanding cooking skills with me. Thank you Peter for sharing your knowledge and passion for your hikes.

Attending conferences has been an important time of my PhD, giving me the opportunity to realize where my work really places within the field. Conference days are moments where intensity flourished, not only because I wanted to get the most out of the technical area, but because I also wanted to enjoy the city I was visiting. Luca, thanks for the wonderful time in Québec City, our beers and great meals will stay in the mind of my stomach forever. It was my first big conference and it was great to meet someone honest, open minded and willing share life experiences. Great thanks also to Larissa for the great day we enjoyed out in the outskirts of Québec City and the nice night in Granada.

I will not forget "the bus girl" who put the cherry on top of the cake for MICCAI'18 in Granada. Thank you for making my bus trip to Sevilla interesting, funny and unforgettable; but most importantly, for reminding me that I am from and I belong to Andalucía.

My deepest appreciation to Jethro and Mel for being the best housemates ever. Living with you guys was great, it made my everyday life towards the PhD comfortable and easy. I really felt our place was like a home away from home. I can remember many good moments, but I am sure I won't forget our (long) interesting discussions, Spanish-English translation doubts, laughs, celebrations, pizzas, asados, rosé nights, beach moments, Anthony street café, boiling cabbage, singing nights, or watching TV shows. Of course, thanks to the Finneapples for your constant smiles.

Also, my deepest gratitude to Maryam, Adel, Sushrith and Manjari. I would have not been able to finish the PhD without your amazing help, even when I only gave very little in return. Whenever I could not cook due to the surgery or because I was extremely busy during the weeks in Adelaide, you guys always "fed" and support me. Of course, thanks for our great laughs and conversations that made me finish everyday with a smile. However, I have to suggest that you guys taste again my outstanding soups. Also, thanks to Mr. Tim T. for the wonderful and sweet memories.

Maryam, Adel, thank you very much for your great will to help and please all the time. Thank you for such comfortable moments in Adelaide. Certainly our deep conversations (also in the car) are one of the highlights of my time in Adelaide, I really enjoy listening and discussing with people that have different points of views. As expected, my warmest appreciation to your salads and kebabs, and also for such a great invitation to my parents.

Manjari, thanks a lot for our cool talks (yes, that includes you teasing me), for taking such a good care of me every day, and for trying to listen and understand me, always seeing the best in me. Thanks for such an amazing description of India, you can be sure I will be visiting you there. I really wish you the best in the near future, I know you will get over these difficult times.

Sushrith, I am sure there will be no droughts with you in Adelaide, rainwater will always flow down the river. Seriously, your chai was the best one I have ever tried, great to warm our "man-talks". Thanks for your sense of humour and for making me improve mine. I take from you that no matter what the situation is, a positive sense of humour should be present. By the way, it is getting a bit cold here... probably too many many fans...

I would also like to express my gratitude to Rhonda for making my arrival in Adelaide easier, and for hosting me at several times during my PhD. I always felt welcome and comfortable! Also, my appreciation to you for treating my parents so well, trying to make the most of their visit. In addition, my greatest thank you for your support in my most difficult times, our conversation corrected my path towards this PhD.

I would particularly like to thank Kristina for sharing such cool 6 months in "our" apartment (where I learnt about the TV shows broadcasted in Australia :p), and Adriana for joining us in such an "adventurous" trip. Special thanks to Amin and Tayana (and Nishant) for waiting at home for me to finish the day with tasty cakes and chai (with ice) (and an occasional dance...).

I would like to offer my special thanks to everyone who shared travels with me. The motivation to do the PhD came from my will to discover Australia, and there is no better way to discover Australia than being surrounded by amazing and interested people. Thank you to

Bettina, Svenia, Seb, Sanne (one of my coolest days in Australia), and Concha, Fernando and Leticia (see Spanish acknowledgements) for being part of my trips.

Niki, thank you for your passion to let people know about your country. You made me really enjoy being in the outback, appreciating every part of it. Thanks also for being open for our little adventure in Adelaide going to Second Valley.

Thank you Flo for such a night under the moving stars. Thank you Valerie for so many laughs in the back of that "bus". Thank you Patrick, Cedric, David for being the authentic 49ers ("maaamaaaaaa" probably summarizes everything). Thank you Damien for your passion transmitting the aboriginal culture and thank you all for such an incredible 14 days, I loved every moment of it. Thank you for the table tennis games, for the long days in the gorges that finished playing cards with a beer in the middle of nowhere, for our Christmas celebration in the middle of another somewhere, and for the new year's night swimming on the beach. My experience in Australia would not have been the same without these incredible moments.

Lili, thanks for being one of the best discoveries of my time in Australia. I wish you had stayed longer, my surfing skills would definitely be stronger. Thanks for that amazing trip, it was just incredible to share those amazing landscapes and beaches with you. Of course, I won't forget our "great" songs, as well as our funny dinners and breakfasts! Also huge thank you for those mantecados, they arrived when I most needed them to sweeten my stress.

Borja, gracias por tus palabras cuando peor lo pasábamos. Muchas gracias, porque sin ello no hubiera conseguido el premio de haber disfrutado de Australia. Fue muy enriquecedor compartir nuestro difícil momento desde la distancia, creo que no sería las misma persona sin haber superado esos baches. Pero gracias también por esos grandes momentos cuando visité España, por la visita a San Sebastián y el (no) surfing en Zarautz, por la tapita en Madrid y por la alegría que siento siempre que hablamos.

Chechu, gracias por ser alguien diferente, analítico y frío. Gracias por recordarme que en España sí se puede y por mostrarme tu interés en la cultura. Aunque la diferencia horaria haga mella, las conversaciones sobre deporte donde no prima el fanatismo son oro puro para mí. Por supuesto, aunque no haga falta mencionarlo, muchas gracias por tu sapiencia, tus razonamientos y tu ayuda en los momentos complicados.

Edu, gracias por esos skype que siempre me suben el ánimo. Da gusto hablar contigo y ver la motivación que transmites. Siempre tras nuestras conversaciones me doy cuenta de que tengo que ser más pragmático. Gracias por estar siempre disponible para escuchar y subir ese ánimo desde la distancia. Y estoy seguro de que tú también vas a conseguir tu objetivo.

Por supuesto, muchas gracias Ana por escuchar e intentar sacarme del pozo en los peores momentos. Al final, la decisión correcta fue seguir con este proyecto.

Blanca y Antonia, fuisteis mi mayor motivación para este proyecto. Blanca, gracias por muchas cosas, pero en especial, por haberme ayudado a cumplir mi sueño y por transmitirme tu espontaneidad y vitalidad. Gracias por haber sido mi modelo a seguir. Sin duda, haber compartido parte del camino contigo me ha ayudado a saber disfrutar esta aventura.

Sara, gracias por enseñarme el tesón necesario para acabar el PhD. Es verdad que, aunque haya momentos duros, hay que seguir creyendo en el proyecto. Fue una casualidad conocernos, pero ha tenido muy buenas consecuencias para mi estancia en Adelaida.

Gracias Diana, por enseñarme a hacer ese esfuerzo extra para hacer de Adelaida mi casa y entablar buenas amistades a base de reir y de disfrutar.

Mariano, gracias por tus valiosos consejos, por tu apoyo y por espabilarme la mente en el momento oportuno para tomar la decision correcta. Pero ante todo, recordaré muchas buenas experiencias: tus invitaciones y esa forma tan rica de hacer el asado, tu ingenio y sentido del humor, el enseñarme como (no) pescar calamares, el kayak, la música, la tienda de campaña y la nevera que me acompañaron en uno de los mejores viajes que hecho, etc... Sin duda, los momentos que hemos compartido han hecho mi estancia en Australia mejor y más valiosa.

Exequiel, muchas gracias (y a tu familia también) por las invitaciones a tu casa donde pude (y mis padres también) aprender algo sobre las artes culinarias en Chile. Además, fue un placer compartir almuerzos y cafés contigo. Pero ante todo, muchas gracias por tu interés en temas que no son "mainstream" y por las conversaciones y reflexiones sobre política que ayudan a aprender a pensar. Y muchas gracias por compartir el "Survival Day" conmigo.

Álvaro, muchas gracias por tu ayuda a pensar durante el doctorado y por tus consejos de como escribir y/o revisar un paper. Gracias también por las risas que nos hemos echado juntos y las cenas y celebraciones que hemos compartido. Por supuesto, muchas gracias también por tu generosidad al dejarme utilizar tu coche para descubir un poco más los alrededores (y no tan alrededores) de Adelaida. De hecho, mi técnica surfera sería (todavía) peor si no hubiera podido utilizar tu coche.

Juan, Tanit, gracias por esos momentos y conversaciones compartidas. Gracias por vuestra confianza, cariño y ayuda. Juan, muchas gracias a ti también por abrirme la mente en el momento oportuno. También, gracias por ser un refugio donde poder compartir ideas y estar al corriente de lo que pasa en España. Quizás no te lo creas, pero gracias por ser valiente, ojalá haya aprendido de ello y lo pueda poner en práctica ahora. Tanit, gracias por escuchar en los días grises, por siempre mostrarte abierta e interesada en salir y hacer cosas, y por tu animosa visita durante la convalecencia. Gracias chicos por ese intenso fin de semana en Flinders que terminó con una gran araña negra bajo mi almohada.

Gracias también, Laura y Benedikt por uniros en la aventura a Flinders. Aunque sé de vuestra gran afición a los juegos, yo os voy a recordar por vuestras (increíbles) dotes para cocinar. Las cenas con vosotros han sido muy parecidas una buena comida en España.

Silvia fue maravilloso compartir contigo tardes, historias y, muchas risas. De los días que coincidimos, recordaré las buenas vibraciones que siempre transmites, siempre de agradecer tras un largo día en el trabajo. Por cierto, aunque quizás suene extraño, tengo que decir que me quedo para mí con tu receta de la paella, simplemente espectacular.

Ghalia, gracias por hacerme recordar el poder de una sonrisa. Y sobre todo, muchas gracias por esos originales sitios para un "brunch" en domingo. Prometo enseñarte buenos sitios de tapas en España. Echaré de menos las conversaciones contigo y con Tanit.

Guisella, gracias por desafiar mis ideas. Me lo pase genial contigo. Gracias por esas conversaciones abiertas, donde tu honestidad y confianza es capaz de ensanchar mi mente. Quizás, la característica más importante en la investigacion es dudar de todo, hacerse buenas preguntas y no ser temeroso a lo desconocido. Y tus conversaciones son una gran ayuda, no sólo en lo profesional, sino también en lo personal. Gracias también por las tarde-noche en el Fringe, algunos bailes, o los momentos en la playa y paseos por el río.

Leti, Fer y Concha, fue una maravilla embullirme en la Australia tropical y compartir unos días con vosotros llenos de risas y bienestar. Por supuesto, no me olvidaré esos cacahuetes en Alice Springs tras una larga caminata o de lo genial que ha sido compartir un poquito Madrid entre tapa y tapa. ¡Sí se puede!

Gracias a mis tíos y primos que me habéis recibido con tanto cariño en mi visita a mitad del doctorado y ahora al final. Muchas gracias por esas comilonas y risas, y sobre todo, por esas sobremesas tan largas y saludables. Gracias, Pablo, por enseñarme mis primeros acordes con la guitarra, no te defraudaré. Y me quiero acordar también de mi tía Juana, que falleció después de un cáncer de mama el día que envié uno de los trabajos presentados en esta tesis.

Gracias Papá y Mamá (con mayúsculas) por compartir esta aventura conmigo en el mayor espectro posible. Gracias por vuestra paciencia, por saber aguantar el chaparrón la primera vez que estuvisteis en Adelaida y por saber disfrutar la segunda. Me encantaron vuestras visitas a Australia y, sobre todo, que os gustasen tanto. Al viajar con vosotros consegui cumplir un sueño, que es que conociérais un sitio especial al venir a visitarme. Creo que nuestras conversaciones en aquel julio de 2015 nunca las olvidaré. Pero tampoco el viaje por el Kimberley (foto en el baobab incluida), el cielo estrellado de Kangaroo Island, o el salir de trabajar e ir a cenar con vosotros. ¿A dónde será el próximo viaje?

Finally, I would like to acknowledge all the help I received during those complicated months of 2015, as well as all during the surgery and visits to doctors.

Finalmente, agradecer toda la ayuda recibida en aquellos meses complicados de 2015, así como durante la operación y las visitas al médico.

Australia, a time full of big moments.

Chapter 1

Introduction

Breast cancer is one of the most dangerous pathologies affecting women[1–4] across the world, regardless of how developed the country is [5–7]. For example, in countries such as Spain and Australia it is the most commonly diagnosed tumour in women [8, 9] and it is the most deadly cancer in Spain [10] and the second in Australia [9]. It is estimated that of all new breast cancer cases and deaths for 2018, 99% occur in females as opposed to 1% in males [1].

Several studies have shown that a key factor to reduce mortality from breast cancer is the early diagnosis of tumours [7, 11–16], as higher survival rates are correlated with the detection of small size tumours. Aiming at localizing small tumours, healthcare institutions [9, 17–21] are implementing breast screening programs focusing on the finding and diagnosis of asymptomatic lesions [22, 23] through different imaging modalities.

These population-based screening programs are generally based on mammography [9, 17, 21, 24]. However, the sensitivity of such screening process tends to decrease for the diagnosis of women with denser breast tissue [25–28] because dense tissue appears with a similar intensity to tumours in mammography [29, 30]. High risk patients for breast cancer (those whose first or second degree relatives have suffered from breast cancer or women with the BRCA1 or BRCA2 gene mutations [31, 32]) are recommended to begin screening programs at an earlier age [31, 33]. Nonetheless, younger women tend to have denser breast tissue [25, 29], limiting the effectiveness of mammography. As a result, women at high-risk are recommended magnetic resonance imaging (MRI) exams [31, 34–37] that can increase the sensitivity of the screening process [22, 35, 38]. If MRI is not suitable, additional screening with ultrasound can be recommended [37, 39], but ultrasound tends to have a low positive predictive value (i.e. the probability that a positive diagnosis is a true positive is low) [28]. The most common MRI modality for such assessment is dynamically contrast-

Fig. 1.1 Pre-hoc pipeline for breast screening from DCE-MRI. In the first stage, suspicious regions, possibly including malignant and benign lesions as well as false positive detections, are localized in the image. In the second stage, the suspicious ROIs are classified to obtain a malignancy score for the diagnosis of the breast DCE-MRI volume.

enhanced MRI (DCE-MRI) [32, 40–42], but other modalities such as DWI or T2-weighted are also used to improve the diagnosis accuracy [43–46].

The reading of DCE-MRI images consists of analysing a sequence of 3D volumes acquired over time. Interpreting such a large amount of data is tedious [42, 47] and time consuming [48], resulting in observer errors [47, 49, 50] and inter-reader variability [51–55]. Similarly to mammography [56–60], computer-aided diagnosis (CAD) systems can interpret breast DCE-MRI data as a second reader to aid radiologists [61, 62]. Studies have shown that the use of CAD systems for breast MRI analysis can increase the sensitivity [53, 63] and specificity [64–68] of the diagnosis, as well as increasing the efficiency in image interpretation [47, 63, 69]. However, designing such CAD systems to aid radiologists in the diagnostic process is challenging due to the large variability in the location, morphology [47, 70], size [70, 71], contrast enhancement patterns [50, 72, 73], and the low signal-to-noise ratio [74] present in lesions.

Traditional fully automated CAD systems [14, 75] tackled the problem of breast cancer diagnosis from DCE-MRI following a pre-hoc approach, where the problem of diagnosis is split into two consecutive stages: 1) the detection of suspicious regions, representing the benign and malignant lesions, followed by 2) the classification of these lesions into malignant or not malignant. The classification probabilities of suspicious regions present in each breast are then combined to produce the probability that such breast contains a malignant lesion. See Fig. 1.1 for a diagram of the pre-hoc pipeline.

This traditional pre-hoc pipeline suffers from two main deficiencies: the sub-optimality of hand-crafted features used for the detection and classification steps, and the high computational complexity of the search approaches used for lesion detection, such as the exhaustive search [75], or the mean-shift clustering [14]. These limitations can be addressed with deep

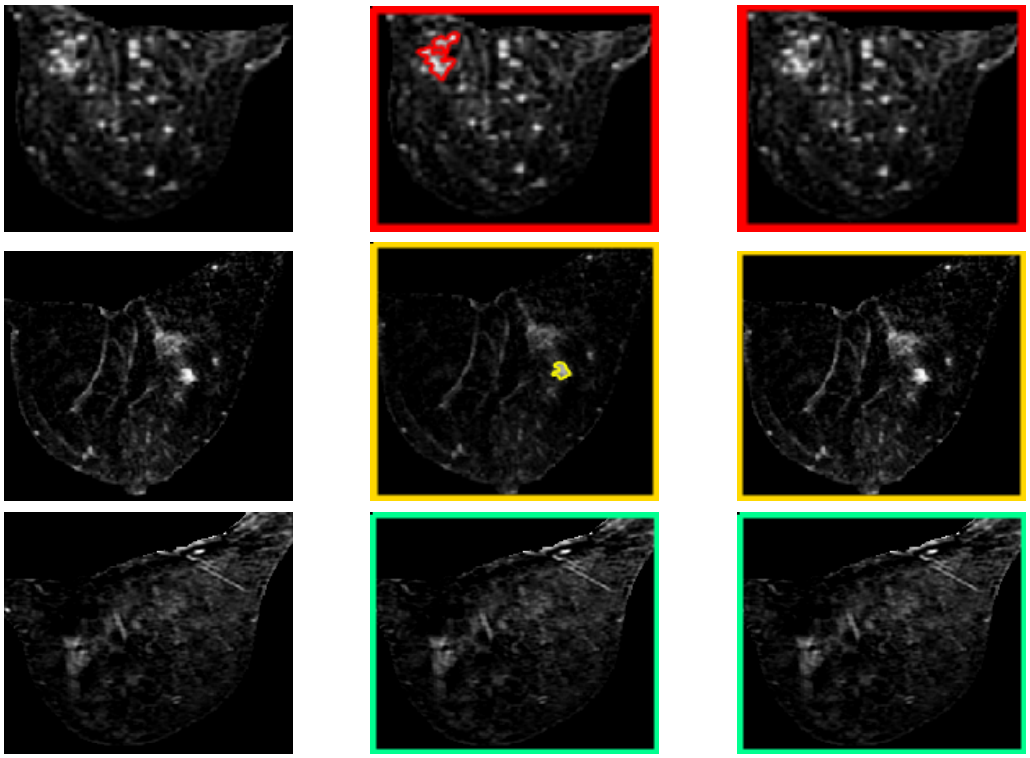


Fig. 1.2 Example of volume annotations. Left column is one slice of a DCE-MRI volume. Middle column corresponds to its strong annotation, i.e. the voxelwise labelling of each lesion together with the characterization of each lesion. Right column corresponds to its weak annotation, i.e the volume-level characterization of the volume in the left column. First, second and third rows correspond to a malignant, benign and healthy cases respectively.

learning methods [76–79]. For instance, the running time complexity for lesion detection can be reduced with the use of attention models [80] (note that [80] is a contribution of this thesis) or by efficiently computing a segmentation map [81]. The sub-optimality of hand-crafted features can also be overcome with deep learning methods that automatically learn the optimal features for the detection and classification problems [81, 82].

Despite the improvements brought by deep learning methods, typical pre-hoc methods remained sub-optimal due to the independent optimization of each of the detection and classification stages. The joint optimisation of detection and classification models has been explored by computer vision researchers [83–85], and could, in principle, be applied in medical image analysis. However, the successful application of these methods requires large datasets with strong annotations [83–86], which are generally unavailable for medical image analysis applications, and in particular for breast MRI analysis.

Strongly annotated datasets, i.e. the voxelwise labelling of each lesion (see Fig. 1.2 for examples of types of annotations), are costly to obtain. Even with the use of semi-automated

annotation tools (e.g., region growing algorithms) that can be used to reduce the time needed for an annotator, expert interaction is still required to trace the segmentation maps [14]. In addition, such datasets are inevitably noisy due to two main reasons. Firstly, the delineation of the lesion boundary on an MRI image can only be done with a limited degree of certainty, even when this is done by an expert annotator. Secondly, the characterization of each lesion into benign or malignant can introduce additional noise into the annotation because each lesion can be assigned a wrong label. For example, in breasts where multiple lesions are present, only the most suspicious is normally confirmed with a biopsy, whereas the labels for the rest of the lesions are inferred by the radiologist inspecting the volumes, leading to possible wrong annotations given the variability present in lesions. Consequently, strongly annotated datasets to train detection and classification models should be used considering the fact that they are likely to be noisy and relatively small.

On the other hand, weakly annotated datasets, i.e. datasets where annotations are at the volume level (see Fig. 1.2 for examples of this type of annotation), are cheap to obtain and are less noisy, resulting in more reliable and larger training sets. The less amount of noise stems from the facts that such datasets no longer contain lesion delineations, and that breast-level classification labels are represented by the characterization of the most suspicious lesion in the breast: a breast is labelled malignant if it contains at least one malignant lesion, benign if it contains only benign lesions or healthy if there are no lesions. As the most suspicious lesion in each breast volume is labelled according to a biopsy, the amount of noise in the labels is likely to be reduced. The workload needed to annotate breast volumes is also reduced as there is no need for the voxelwise annotation of each lesion and the desired weak labels are normally available from Picture Archiving and Communication Systems (PACS) in hospitals

Post-hoc CAD systems [87, 88] benefit from weakly annotated datasets and perform the diagnosis based on the classification of the entire breast volume, which is important given tumours can generate appearance and structural changes in the whole breast [89]. Diagnosis in post-hoc CAD systems can accurately be solved with state-of-the-art classification models based on deep learning methodologies [90–92], where the main advantage lies in the end-to-end training with weakly labelled training sets. However, two main drawbacks arise: 1) the need of large datasets to train such classifiers to accurately perform diagnosis, and 2) the lack of lesion localization output to explain the malignancy decision of the classifier [93], which is critical to facilitate the translation of such system into clinical practice [94–96]. See Fig. 1.3 for a diagram of the post-hoc pipeline.

In this thesis, we firstly propose a pre-hoc CAD system that has a more efficient inference runtime complexity for its lesion localization stage, compared to its competitors. In addition, we show that our proposed pre-hoc CAD system can be trained with relatively small strongly

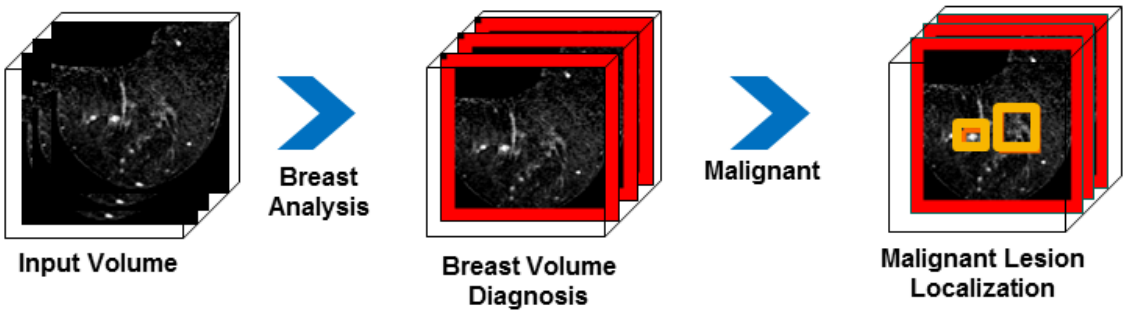


Fig. 1.3 Post-hoc pipeline for breast screening from DCE-MRI. Conversely to the pre-hoc pipeline, firstly diagnosis is performed by analysing the whole breast volume. If the outcome of the diagnosis is positive, malignant lesions are localized in the input volume to provide an interpretation of the diagnosis.

annotated breast DCE-MRI datasets [80, 97]. Secondly, we develop a post-hoc CAD system that can perform a relatively accurate diagnosis after being trained from a small weakly annotated DCE-MRI dataset [98] and localize lesions that explain such diagnosis [99]. Finally, we provide a systematic comparison between our proposed pre-hoc and post-hoc CAD systems for the problem of breast cancer diagnosis and malignant lesion localization from breast MRI.

We evaluate our methods in a dataset containing DCE-MRI volumes of 117 patients. Results show that our post-hoc CAD system trained with a small weakly labelled dataset achieves an AUC of 0.91, setting the new state-of-the-art performance for breast cancer diagnosis from breast DCE-MRI. The post-hoc system additionally can competitively localize malignant lesions, although the pre-hoc localization methods trained with strong annotations yield better performance for malignant lesion localization.

1.1 Motivation

Fully automated CAD systems for breast screening from DCE-MRI traditionally followed a pre-hoc approach based the localization of suspicious lesions and their subsequent classification. Although the inefficiencies of this pipeline could be addressed with the application of state-of-the-art computer vision techniques [83–85] that can be trained end-to-end to perform the diagnosis, they are not suitable to analyse breast DCE-MRI volumes because of the large strongly annotated datasets they require. Such impediment led us to design two lesion localization methodologies that can be trained with small strongly annotated training sets. In addition, we also address the inefficient inference time for the lesion detection approach [80, 97]. Firstly, we propose an efficient lesion segmentation approach that can be

applied after the detection step and that can significantly improve the segmentation accuracy. Secondly, we propose an attention model for lesion detection that is able to detect lesions at a state-of-the-art (SOTA) accuracy while significantly reducing inference times. Detected lesions with the proposed attention model [80] are classified with a SOTA classifier [90] to build the complete pre-hoc system. Diagnosis is performed by combining the probability of malignancy of each detected region in the breast [100].

Post-hoc CAD systems perform the diagnosis by analysing the whole breast volume, where the main advantage lies in the use of weakly labelled datasets during the training phase, which can reduce the effort in the preparation of training sets and reduce the noise present in the annotations. However, to be able to achieve a relatively successful performance, post-hoc systems require large training datasets that are not readily available for breast DCE-MRI analysis. Such issue motivated us to design a new training approach for the diagnosis module of post-hoc systems that is capable of performing state-of-the-art diagnosis when trained from small weakly labelled training sets [98].

One of the main difficulties of the adoption of post-hoc CAD systems for breast cancer screening lies in the absence of a method that can explain their diagnosis decision. In our set-up, the positive diagnosis of a breast as malignant is accompanied by the localization of the malignant lesions that can explain such diagnosis. Consequently, we propose a 1-class saliency detector [99] that localizes malignant lesions in positively classified breast DCE-MRI volumes to interpret the decisions of post-hoc approaches.

The proposed pre-hoc and post-hoc approaches solve the problem of performing breast diagnosis and malignant lesion localization with different pipelines that have different annotations requirements. However, there is no comparison between approaches when applied to solve each of the problems. In order to fill this gap, and using the methods proposed in this thesis, we present a systematic comparison between pre-hoc and post-hoc approaches for breast cancer diagnosis from DCE-MRI [100].

1.2 Contributions

The contributions of the thesis lie in the development of pre-hoc and post-hoc approaches for breast screening from DCE-MRI CAD systems and their systematic comparison regarding their diagnosis and lesion localization performance.

In terms of pre-hoc CAD systems, we focus on reducing the inference time of the suspicious region localization phase:

1. We present a globally optimal lesion segmentation methodology that can be applied after any detection algorithm. We propose the global minimization of an energy that

incorporates a segmentation prior from a deep learning model. Our method achieves the best results in the field at a significant smaller inference time [97]. See Chapter 3 for more details

2. We develop an attention model that significantly reduces the inference time while maintaining state-of-the-art accuracy for the lesion detection phase of pre-hoc systems [80]. Based on deep reinforcement learning, the model learns how to evolve a large initial bounding box until it tightly finds a lesion. See Chapter 4 for more details.

In terms of the post-hoc CAD systems:

3. We propose a training scheme for diagnosis systems based on the analysis of the whole volume and that can be trained from small weakly annotated breast DCE-MRI datasets. Our proposed method, based on meta-training with curriculum learning, establishes a new state-of-the-art performance for this problem, improving over several previous baselines [98].See Chapter 5 for more details.

4. We present a 1-class saliency detector to interpret the classification decisions of post-hoc diagnosis systems. Our proposed method assures that detected regions in the volume correspond to malignant lesions in positively classified breast DCE-MRI volumes. Experiments show that we establish a new state-of-the-art accuracy for the problem of weakly supervised lesion detection from breast DCE-MRI [99]. See Chapter 6 for more details.

Finally, due to the developments in 1–4, and in particular the advances proposed to mature the post-hoc systems, we propose a systematic comparison between both pipelines for breast DCE-MRI CAD systems:

5. We compare the diagnosis and lesion localization performance of pre-hoc and post-hoc CAD systems under the same dataset and evaluation metrics. We select the systems described in this thesis as representatives for pre-hoc and post-hoc approaches. Results show that the post-hoc system trained from a weakly labelled dataset achieve better diagnosis performance. However, the pre-hoc approach localize lesions more accurately [100], possibly benefiting from the strong annotations used during the training phase . See Chapter 7 for more details.

1.3 Outline

The thesis is organised as follows:

- Chapter 2 includes the literature review discussing pre-hoc and post-hoc approaches for CAD systems.
- Chapter 3 introduces a globally optimal lesion segmentation method to significantly improve the accuracy and reduce the inference time of lesion segmentation methods in pre-hoc CAD systems trained from small breast DCE-MRI datasets.
- Chapter 4 introduces an attention model to significantly reduce the inference time while maintaining the accuracy of the lesion detection step of pre-hoc systems trained from small breast DCE-MRI datasets.
- Chapter 5 introduces a novel training method to improve the training of post-hoc diagnosis methods trained from small weakly labelled breast DCE-MRI datasets.
- Chapter 6 develops a 1-class saliency detector to localize lesions in positively classified breast DCE-MRI volumes, interpreting the decisions provided by the model presented in Chapter 5.
- Chapter 7 presents a comparison in terms of lesion detection and diagnosis performance between pre-hoc and post-hoc CAD systems for breast screening from DCE-MRI.
- Chapter 8 concludes the thesis and summarizes its contributions. It also presents the limitations of this study and future research directions.

Chapter 2

Literature Review

CAD systems can be grouped into pre-hoc or post-hoc systems depending on the order of the stages of their diagnosis pipeline. Pre-hoc methods generally involve a first stage that detects suspicions regions and a second stage that classifies those regions. Post-hoc methods, on the other hand, reverse these stages. They firstly classify the whole volume for its diagnosis, and secondly localise the malignant lesions if the input volume is positively. We review pre-hoc and post-hoc approaches in Sec. 2.1 and Sec. 2.2 respectively. Sec. 2.3 motivates the need for a comparison between both pipelines. Finally, Sec. 2.4 summarizes the literature review of the thesis.

2.1 Pre-Hoc Systems

Pre-hoc approaches perform diagnosis in two sequential stages. Firstly, region(s) of interest (ROIs) are localized in the input volume. In this context, localization can be interpreted as a detection task, where the aim is to place a bounding box as tightly as possible around the lesions, or as a segmentation task, where the aim is to estimate the contour of the lesion. Secondly, previously localized ROIs are classified into true positives or false positive detections to perform diagnosis. Since suspicious regions localized in the first step will include malignant and benign lesions and possibly false positive localizations that correspond to normal tissue, the classification step of breast screening systems will distinguish malignant lesions (considered positive) from false positives and benign lesions (considered negative).

Initial CAD systems aided radiologists in the diagnosis of manually detected suspicious ROIs by displaying measurements of hand-crafted features based on contrast enhancement [45, 93, 101–103]. CAD systems evolved to perform the automated classification of manually or semi-automatically localized ROIs to estimate the probability of malignancy for each ROI. The localization of ROIs was performed either with manual detection [104, 105],

manual segmentation [68], or semi-automated segmentation [106–116] methods based on the interaction of the system with a radiologist that is in charge of placing a seed on a suspicious area of enhancement or selecting an initial ROI that contains the lesion as the initialization for the lesion segmentation algorithm. Hand-crafted features such as:

- Morphology and/or dynamic and/or texture [65, 68, 104, 107, 109–117]
- Pharmacokinetic [108]
- Multifractal [105]
- Texture from wavelet transforms [118]
- Radiomics [119]

were computed from each each ROI and used to characterise it to perform its classification and compute its probability of malignancy using methods such as:

- Support vector machines [68, 105, 110, 112, 113, 119]
- Random Forests [110, 112, 114–116]
- Artificial neural networks [65, 120]
- Logistic regression [108, 109, 111]
- Linear discriminant analysis. [118]

We refer the reader to [121] for a systematic review of types of features and classifiers used for the classification of lesions from breast DCE-MRI.

Research interests shifted from semi-automated to fully automated pre-hoc systems [14, 67, 75, 122] that include the automated localization of lesions. The main idea behind automating the full pipeline is that reducing user intervention is important to minimize the amount of ROIs that have to be processed [123]. Initial automated lesion localization strategies relied on methods such as thresholding based on the enhancement [67, 124] that could not capture the variability present in lesions. More complex strategies involved the extraction of hand-designed features with computationally expensive lesion localization strategies such as exhaustive search [75, 122, 125] or unsupervised clustering followed by structured learning [126]. The classification of automatically localized ROIs into positive or negative diagnosis was performed by characterizing each ROI with region-wise hand-designed features that were used by classifiers such as neural networks [67], random forests [14, 75] or naive Bayes [122] to estimate the diagnosis.

The fully automated methods mentioned above suffer from: 1) the expensive computational resources required by the complex strategies used to localize suspicious regions, 2) the sub-optimality of hand-designed features that are employed in both localization and classification stages, and 3) the lack of optimality of the full diagnosis pipeline due to the independent optimization of each of the detection and classification stages. The introduction of recently developed deep learning methods in medical image analysis [127, 128] can help overcome these limitations [129].

Two main ideas can improve the efficiency of the localization step: attention models [130] that can focus on the relevant areas of the image, and the rapid computation of lesion segmentations maps. Both of these approaches have been explored for breast lesion localization from DCE-MRI: Maicas *et al.* [80] employed an attention model based on deep reinforcement learning to detect lesions (note that this work [80] is a contribution of this thesis), and Dalmics *et al.* [81] and Zhang *et al.* [131] compute the segmentation of lesions using a U-net [132].

There are other deep learning methods developed by the computer vision community that could be applied to find the segmentation maps of breast lesions from DCE-MRI. For example, fully convolutional networks [133], its medical image extension V-net [134] that uses the Dice coefficient as the loss function, or the inclusion of a shape prior in deep learning models for segmentation [135–137] have shown to produce accurate results. However, the use of such methods is challenging for breast lesion segmentation from DCE-MRI due to the small training sets available for breast DCE-MRI analysis and the large amount of training data required by these methods. The limitation of the size of the dataset has been addressed by Zhao *et al.* [138], where a patch based training was explored before training the model with the shape prior as explained by Zheng [135].

Feature sub-optimality due to the use of hand-crafted features has been addressed for both the localization and diagnosis stages of pre-hoc systems. Regarding the localization step, feature sub-optimality has been tackled by doing exhaustive search with a deep model trained to distinguish whether each patch of the image belongs to a lesion [139]. Deep learning methods have also been explored for the diagnosis of lesions from breast DCE-MRI, allowing the learning of optimal features to classify previously localized lesions [81, 82, 140]. Although the training of these models from scratch has been shown to be relatively successful for breast lesion classification from DCE-MRI [81, 82, 140], it is worth noting that the use of transfer learning has been suggested to improve the performance of medical image analysis deep models [141, 142]. In particular, the use of pre-trained models may be interesting for the classification of lesions from breast DCE-MRI because they can improve the accuracy of models where not enough training data is available. However, evidence provided by Amit

et al. [143] showed a better performance of a model trained from scratch compared to a pre-trained model.

Despite the individual optimality of each stage of the diagnosis process [81], the optimality of the end-to-end pipeline for diagnosis is not assured. Recent state-of-the-art computer vision methods that have proposed end-to-end training of the localization and classification stages, and thus their joint optimization, could potentially address this issue. For example, Faster R-CNN [83] (and its extension Mask R-CNN [85]) and Yolo [84] methods propose the joint optimization of localization and classification steps of pre-hoc systems. The main difference between Faster R-CNN and Yolo methods is that while Faster R-CNN methods are based on the classification bounding boxes of different aspect ratios at each location, Yolo regresses ROIs bounding box locations and outputs their class probabilities. Similarly, another scheme proposes the joint optimization of the subsequent steps of detection, segmentation and classification [144] for instance segmentation and classification. Although these approaches are relatively successful in terms of classification accuracy and inference time, their application is challenging for some medical image analysis problems due the large amount of strongly annotated training data needed to achieve good performance in the test data. For example, the application of Faster R-CNN has been shown to be successful when a large training set was available [86, 145], but challenging for relatively small training sets [146].

Aiming to reduce the burden of the large amount of strong annotations required to successfully train deep models, the medical image analysis community is also exploring semi-supervised learning methods [147–150] that can combine strongly annotated samples with other weakly labelled samples. Two main techniques have been proposed to incorporate weakly supervised data into the training process: 1) the modification of the loss function employed depending on the type of annotation of each sample [148, 149], and 2) the proposal of training a deep model in a two-stage process using alternative updates [151] or based on expectation maximization [147, 150]. However, although these methods can reduce the number of strongly annotated samples required for training, still relatively large datasets (which are currently not available for breast DCE-MRI) are needed to successfully train the methods [149].

2.2 Post-Hoc Systems

Post-hoc systems reverse the order of the stages of the pre-hoc pipeline (see Fig. 2.1 for a block diagram of pre-hoc and post-hoc pipelines). Initially, the system performs diagnosis by analysing the entire volume, possibly considering the information from tissue surrounding

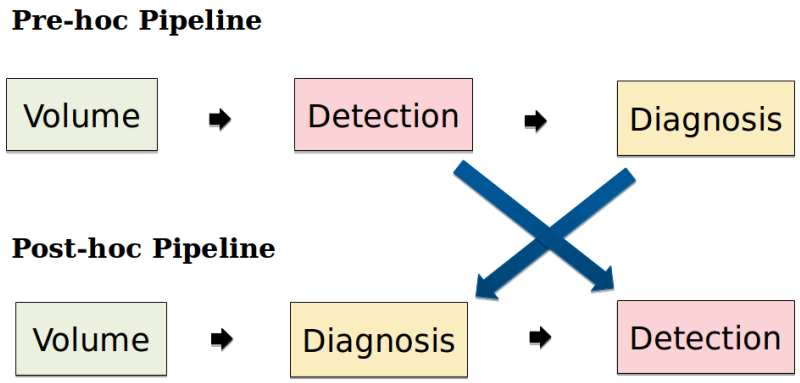


Fig. 2.1 Block diagram of pre-hoc and post-hoc pipelines for breast screening from DCE-MRI. Pre-hoc systems localize suspicious regions in the input volume. These regions are subsequently classified to perform diagnosis. On the contrary, post-hoc systems initially perform diagnosis based on the whole breast volume and, if the outcome is positive, malignant lesions are localized in the input volume.

the lesions (not only from lesions as in pre-hoc approach) that may be relevant to assess malignancy [89]. If the outcome of the diagnosis is positive, the system proceeds to localize the malignant regions in the image that can explain such positive diagnosis. Conversely to pre-hoc models, post-hoc approaches are based on the training with weakly annotated datasets, reducing the burden of the voxelwise annotation of each lesion present in the dataset.

The diagnosis stage is formulated as a classical supervised image classification problem, and thus the model can be optimally trained and implemented using standard computer vision classification architectures [78, 90, 152–155]. For example, Wang *et al.* [87], Rajpurkar [156] *et al.* and Gundel *et al.* [157] based their work in such well studied models. Alternatively, Geras *et al.* [158] argued that standard computer vision models are not suitable for medical image analysis because images have to be heavily downsampled, loosing important details. In their work, they proposed a multi-view deep learning model to combine information from several high-resolution views obtained at a single mammogram exam . Similarly to Geras *et al.* [158], other contributions [159, 160] focused on extending standard models to solve particular medical image analysis problems. For instance, Yao *et al.* [159] proposed to train a system by using information that can be extracted from the dependencies among the labels in the training set. Guan *et al.* [160] proposed a method such that the classification of the whole volume is guided by a deep learning model that functions as an attention model.

The diagnosis methods described above are becoming widely employed due the increasing availability of large weakly labelled datasets [87, 145, 158, 161–166] that have the potential to allow for a robust modelling of deep learning classifiers. Unfortunately, even though such large datasets are available for some medical image analysis problems, they are not

available for breast screening from DCE-MRI. In order to improve the learning from smaller datasets, Zhu *et al.* [91] presented a deep multiple instance learning framework to fine-tune a pre-trained model, while Xue *et al.* [92] suggested multitasking to train a model in different but related tasks such that the training of each task benefits from the training of the others. Although results from the methods above [91, 92] show improved performance (when compared with classic machine learning training approaches), they still need relatively large training datasets.

A key limitation for the deployment of post-hoc diagnosis systems in medical practice lies in the need for localizing the relevant regions in the volume that led to the diagnosis. Such visualization of relevant regions can make the system more intelligible for radiologists and facilitate the adoption of post-hoc systems in the clinical workflow. For example, in breast screening from DCE-MRI, a positive diagnosis should be accompanied by the localization of the malignant lesion(s) present in the volume. However, this is a challenging problem due to the absence of lesion localization ground truth in the training phase.

The interpretation of deep learning models, where the aim is to detect the visual class that is relevant for the output of a classification stage, is receiving substantial attention by the computer vision community. A relevant early approach is the class activation maps (CAM) [167] proposed by Zhou *et al.*, where authors weighted the activations involved in the classification decision to produce a low resolution visualization of the regions involved in the classification of the image. This idea inspired the medical image analysis community in several contributions [87, 156, 168–170]. However, even though this approach worked reasonably well when visual classes did not contain large variability in shape and appearance, and the size of the target class is relatively large, it lacks the precision required for tumour detection from breast DCE-MRI because of the low resolution of the feature maps employed.

Several approaches try to alleviate the inner limitation of the low resolution of CAMs. For example, Guided Grad-CAM [171] and Respond-CAM [172] proposed to use the gradient of the target class with respect to a particular layer of the model, Cai *et al.* [173] and Yao *et al.* [174] considered information of the deep model at different resolutions, and Dubost *et al.* [175] proposed to generate a high resolution attention map based on regressing the number of lesions. Although these approaches improve the resolution with respect to CAMs, we argue that such methods are not suitable for the problem of lesion detection from breast DCE-MRI due to two reasons: 1) if the diagnosis is negative, there should be no regions detected in the volume, and 2) there is no guarantee that the detected regions represent lesions, as the model might be learning other features that are not clinically relevant. An interesting approach towards satisfying these conditions can by inspired by Dabkowski and Gal [176], where the computer vision classes that should be localized in the image are explicitly defined.

2.3 Evaluation - Comparison Pre-Hoc and Post-Hoc Approaches

The pre-hoc and post-hoc approaches presented in Sec. 2.1 and Sec. 2.2 represent two different pipelines with different annotations requirements for malignancy diagnosis and malignant lesion localisation. However, establishing a comparison between both types of pipelines to determine their suitability for the problem of breast screening from DCE-MRI is not straightforward due to [100]: 1) the use of private breast DCE-MRI datasets that prevent methods from sharing the same training, validation and testing data splits for a fair comparison, 2) the criterion to establish whether a lesion assigned a BI-RADS = 3 is understood as a benign or as a malignant lesion [75], 3) the decision of whether to manually remove false positive detections corresponding to normal tissue from the diagnosis stage [14] of pre-hoc systems, and 4) the different overlap criterion adopted to establish whether a detected region is a true positive. For example, previous works [75, 126] define a detection to be true positive if at least one pixel is detected, but we believe that such criterion should be more ambitious, requiring at least a minimum Dice coefficient of 0.2 between the detected region and the ground truth [80, 97, 99, 100].

2.4 Conclusion

Initial fully automated pre-hoc CAD systems for breast screening from DCE-MRI [75, 126] suffered from the computationally expensive lesion localization strategies and the sub-optimal hand-crafted features employed for both lesion localization and classification stages. Both limitations can be addressed with deep learning methodologies that can speed-up the inference time as well as computing optimal features for each of the stages [81]. However, such methodologies remain sub-optimal for volume diagnosis since optimal detection does not assure optimal classification. Diagnosis optimality following a pre-hoc pipeline could be achieved with state-of-the-art computer vision methods [83, 84, 86], where localization and classification are trained end-to-end, but these methods require large strongly annotated datasets that are not available for breast DCE-MRI.

On the other hand, post-hoc systems [87, 91, 92] can be optimally trained for breast volume diagnosis using weakly labelled datasets. However, large datasets are required to successfully train a classifier to achieve a relatively accurate diagnosis. Additionally, aiming to provide an interpretation of the diagnosis, post-hoc models should provide the localization of malignant lesions in cases where the system outputs a positive diagnosis for breast cancer. Although some methods have been proposed [172, 173, 175] by the computer vision and

medical image analysis communities, none of them can guarantee that localized regions represent lesions.

In this thesis, we firstly propose to reduce the inference time requirements of the lesion localization stage in pre-hoc systems by proposing:

1. an attention model for lesion detection that progressively focuses on lesions in a breast DCE-MRI volume [80], and

2. a globally optimal segmentation algorithm that includes a shape prior from a deep learning model [97].

Secondly, we focus on post-hoc systems and propose:

3. the design of a novel method for training post-hoc diagnosis systems from small weakly labelled datasets [98], and

4. a 1-class weakly-supervised saliency detector designed for localizing lesions in positively diagnosed volumes [99], where we aim to associate the positive diagnosis to breast containing lesions.

Finally, given the improvements in 1–4 above, we propose a systematic comparison between our proposed pre-hoc and post-hoc systems for breast screening from DCE-MRI [100], addressing the difficulties explained in the Sec. 2.3.

Chapter 3

Globally Optimal Breast Mass Segmentation from DCE-MRI Using Deep Semantic Segmentation as Shape Prior

The work contained in this chapter has been published as the following paper:

Gabriel Maicas, Gustavo Carneiro, Andrew P. Bradley. Globally optimal breast mass segmentation from DCE-MRI using deep semantic segmentation as shape prior. IEEE International Symposium on Biomedical Imaging (ISBI), 2017.[1] **Oral Presentation.**

DOI: https://doi.org/10.1109/ISBI.2017.7950525

[1] We refer to mass and non-mass-like lesions throughout the paper, not just masses

Statement of Authorship

Title of Paper	Globally Optimal Breast Mass Segmentation from DCE-MRI using Deep Semantic Segmentation as Shape Prior
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Gabriel Maicas, Gustavo Carneiro, Andrew P. Bradley,. Globally Optimal Breast Mass Segmentation from DCE-MRI using Deep Semantic Segmentation as Shape Prior. International Symposium on Biomedical Imaging (ISBI), 2017.

Principal Author

Name of Principal Author (Candidate)	Gabriel Maicas Suso		
Contribution to the Paper	- Developed the idea of the paper - Coded the proposed algorithm - Designed experiments to validate the algorithm - Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	27 July 2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Developed the idea of the paper - Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	03-08-18

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	- Supervised the development of the work - Refined the manuscript		
Signature		Date	27 July 2018

GLOBALLY OPTIMAL BREAST MASS SEGMENTATION FROM DCE-MRI USING DEEP SEMANTIC SEGMENTATION AS SHAPE PRIOR

Gabriel Maicas[†] *Gustavo Carneiro*[†] *Andrew P. Bradley*[⋆] [*]

[†] ACVT, School of Computer Science, The University of Adelaide
[⋆] School of ITEE, The University of Queensland

ABSTRACT

We introduce a new fully automated breast mass segmentation method from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). The method is based on globally optimal inference in a continuous space (GOCS) using a shape prior computed from a semantic segmentation produced by a deep learning (DL) model. We propose this approach because the limited amount of annotated training samples does not allow the implementation of a robust DL model that could produce accurate segmentation results on its own. Furthermore, GOCS does not need precise initialisation compared to locally optimal methods on a continuous space (e.g., Mumford-Shah based level set methods); also, GOCS has smaller memory complexity compared to globally optimal inference on a discrete space (e.g., graph cuts). Experimental results show that the proposed method produces the current state-of-the-art mass segmentation (from DCE-MRI) results, achieving a mean Dice coefficient of 0.77 for the test set.

***Index Terms*—** breast cancer, deep learning, energy-based segmentation, shape prior, breast mass segmentation, breast MRI, global optimization.

1. INTRODUCTION

Breast screening based on dynamically contrast-enhanced magnetic resonance imaging (DCE-MRI) is particularly useful for patients with dense breasts [1], given that for this cohort, DCE-MRI allows an increase in sensitivity, compared to mammograms [2, 3]. Due to the necessity of interpreting 4D images (3D volumes over time), analysing DCE-MRI images is a complex task that requires medical expertise and is prone to large inter-user reading variability. As a result, computer assisted detection (CAD) systems are being developed to assist radiologists in this task [4]. The analysis used in these systems can be divided into the detection, segmentation and classification of masses where the main contribution of this paper lies in the segmentation of masses. Furthermore, we also propose a novel multimodal detection approach to allow the implementation of a fully automated segmentation methodology.

The particular problem of mass segmentation is challenging due to the variable size, appearance and shape of tumours [5], and the relatively low signal to noise ratio of the masses in DCE-MRI. In addition, fully automated methodologies need to address the usually inaccurate alignment of the initial region of interest (ROI) for the segmentation. State-of-the-art segmentation methods mostly rely on the development of hand-crafted features [6, 7] and methods based on globally optimal inference on a discrete space [7].

In this paper, we propose GOCS-DLP, a new breast mass segmentation methodology from DCE-MRI based on *globally optimal inference* on a *continuous space* that relies on a *shape prior based on the semantic segmentation computed from a deep learning* (DL) model (see Fig. 1). The method is inspired by a recent work [8] that explores *locally optimal inference* on continuous space and uses a shape prior based on the semantic segmentation computed from a DL model for the problem of left ventricle segmentation from MRI. We extend this method with the use of globally optimal inference, which shows robustness to the initialisation of the inference process [9]. Compared to the work by Cremers et al. [9], the main novelty lies in the use of a DL model as a shape prior. In order to make the segmentation fully automated, we extend the breast mass detection methodology (from mammograms) proposed by Dhungel et al. [10] with a model that is formed by a cascade of multimodal deep learning classifiers. While we employ DCE-MRI for segmentation, we combine T1-weighted, T2-weighted and DCE-MRI for detection .

We test our proposed methodology using a breast multimodal MRI dataset, containing 117 cases, with 141 annotated masses, with 46 being benign and 95 malignant, where 58 patients are for training and 59 for testing. We compare, in terms of the mean Dice coefficient ($\bar{D}$), the segmentation results produced by our proposed method ($\bar{D} = 0.77$) with several baselines: globally optimal inference on a discrete space [7] (GODS) ($\bar{D} = 0.74$), globally optimal inference on a continuous space without a DL shape prior (GOCS-MS) ($\bar{D} = 0.73$), locally optimal inference on a continuous space using DL shape priors [8] (LOCS) ($\bar{D} = 0.62$), and semantic segmentation from convolutional neural network (CNN) [11] ($\bar{D} = 0.68$). These results show that our method is significantly more accurate than the competition for the fully automated problem (i.e. using automated mass detection) and for the semi-automated problem (i.e., with manual mass detection).

1.1. Literature Review

The segmentation of breast masses from DCE-MRI has been addressed in several ways. Jayender et al. [6] use hand-crafted features in a voxel-wise classification, where the disadvantage lies in the relatively poor segmentation accuracy produced due to the lack of a shape model that can smooth the result, which is an issue that also affects region-growing segmentation methods [12]. Relying on hand-crafted features is another issue of the method above [6], which has recently been addressed by the computer vision community with the use of deep learning models that automatically learn features for particular classification/regression tasks [13]. More accurate results can be obtained with segmentation methods based on global infer-

[*]This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP140102794). Prof. Bradley is the recipient of an Australian Research Council Future Fellowship(FT110100623)

Globally Optimal Breast Mass Segmentation from DCE-MRI Using Deep Semantic Segmentation as Shape Prior

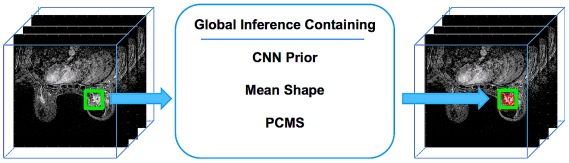


Fig. 1. Breast mass segmentation from DCE-MRI using a global minimisation on a continuous space based on the following energy functional terms: DL (CNN) prior [11], mean shape and the terms inherited from the piece-wise constant Mumford Shah [19] model (PCMS).

ence on discrete spaces [7] (i.e., graph cut [14]), which has high memory complexity that is circumvented with the use of techniques to reduce the number of graph nodes (e.g., superpixels [15]).

Deep learning models have been explored for semantic segmentation [11], which in general need large amounts of annotated training data and produce relatively poor results in terms of segmentation accuracy due to the lack of shape models. The use of shape prior models within deep learning [16] has addressed the accuracy issue at the expense of complex learning methods that require even larger annotated training sets. Given that in medical image analysis it is rare to come across problems with large annotated training sets, recent developments focus on the use of the shape produced by deep learning models as a weak shape prior that is combined with other segmentation cues, such as: strong edges, homogeneous grey-value intensities, contour smoothness, etc. In essence, this involves the combination of level set methods [17] and deep learning shape priors, which has been recently studied for the problem of left ventricle segmentation from MRI [8]; here the issue is the strong dependence on an accurate intialisation because level set produces a locally optimal segmentation result. Level set optimisation problems have been relaxed [18] to transform it into globally optimal inference, which in turn has been adapted to work with (non-deep learning) shape priors [9]. Therefore, the main novelty of our paper is the combination of the globally optimal inference on continuous space with the use of a deep learning-based shape prior.

For the mass detection problem, we extend the methodology proposed by Dhungel at al. [10], which consists of a cascade of deep learning and random forest classifiers (where the random forest classifiers use hand-crafted features) to detect masses on mammograms. This methodology currently holds the state-of-the-art results in a few publicly available mammogram datasets. Our proposed extension reduces the complexity of Dhungel at al.'s approach [10] by reducing the number of cascade stages and avoids the use of hand-designed features by automatically learning features with DL. Furthermore, similarly to [7], we also consider a multimodal approach for mass detection.

2. METHODOLOGY

In this section we explain the deep learning model used to produce the shape prior, the globally optimal inference on a continuous space that uses this deep learning shape prior, and the breast mass detection approach. Hereafter, let $\mathcal{D} = \{\mathbf{v}_i, \mathbf{y}_i\}_{i=1}^{|\mathcal{D}|}$ be the annotated dataset, where each DCE-MRI volume is represented by $\mathbf{v} : \Omega \subset \mathbb{R}^3 \rightarrow [0,1]$ (where mass-like voxels have values closer to 1) and the corresponding breast mass annotation is denoted by $\mathbf{y} : \Omega \subset \mathbb{R}^3 \rightarrow \{0,1\}$, where 0 represents background and mass is denoted by 1.

2.1. Deep Learning Model

The deep learning shape prior is produced by a convolutional neural network (CNN) that outputs a semantic segmentation [11] of the breast mass from a DCE-MRI. The CNN is defined by:

$$f(\mathbf{v}, \theta) = \mathbf{y}^{*,CNN} = f_{out} \circ f_L \circ ... \circ f_2 \circ f_1(\mathbf{v}(0)), \quad (1)$$

where $\mathbf{v}(0) = \mathbf{v}$ (i.e., the original DCE-MRI volume), $\circ$ denotes the composition operator, $\mathbf{y}^{*,CNN} \in [0,1]$, and θ represents the CNN parameters (i.e., weights and biases). Note from (1) that the output $\mathbf{y}^{*,CNN}$ estimates a binary map with the mass segmentation. Each layer in (1) contains a set of filters, defined by

$$\mathbf{v}(l) = f_l(\mathbf{v}(l-1)) = \sigma(\mathbf{W}_l^\top \mathbf{v}(l-1) + \beta_l), \quad (2)$$

where $\sigma(.)$ represents a non-linearity [13], and the convolutional filters are represented by the weight matrix $\mathbf{W}_l$ and bias vector β_l. The modelling of the CNN is performed with a supervised learning process, where the goal is to approximate the annotation by minimising the following per-pixel binomial logistic loss:

$$L = \sum_{i=1}^{|\mathcal{D}|} \sum_{x \in \Omega} \log\left(1 + e^{(-\mathbf{y}_i(x) \times \mathbf{y}_i^{*,CNN}(x))}\right), \quad (3)$$

where x indexes the volume lattice Ω.

2.2. Globally Optimal Inference on a Continuous Space using a Deep Learning Shape Prior

The locally optimal inference on a continuous space (i.e., level set method [17]) can denote the segmentation function in various ways [18], such as the zero level set of a signed distance function, or as one of the two regions of a binary function. We assume the latter representation, with the level set function denoted by $\tilde{\mathbf{u}} : \Omega \rightarrow \{0,1\}$, where the final segmentation is obtained with $\mathbf{y}^{*,LO} = \tilde{u}$, where LO stands for local optimisation. The energy functional of our approach extends the piece-wise constant Mumford-Shah (PCMS) [19] (Fig. 2):

$$E(\tilde{\mathbf{u}}) = \beta \int_\Omega (\bar{\mathbf{y}}(x) - \tilde{\mathbf{u}}(x))^2 dx + \alpha \int_\Omega (\mathbf{y}^{*,CNN}(x) - \tilde{\mathbf{u}}(x))^2 dx + \lambda \int_\Omega (\mathbf{v}(x) - \tilde{\mathbf{u}}(x))^2 dx + |\nabla \tilde{\mathbf{u}}(x)| \quad (4)$$

where $\bar{\mathbf{y}}(x) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbf{y}_i(x)$ represents the mean shape prior [9] at position $x \in \Omega$, $\mathbf{y}^{*,CNN}(.)$ denotes the DL shape prior from (1). While the role of the DL shape prior is to capture mass variability, the mean shape prior attemps to constrain its translation and scale. Finally, the last two terms (from PCMS) penalise differences between $\mathbf{v}$ and $\tilde{\mathbf{u}}$ and large segmentation perimeters. The minimisation of the energy functional in (4) finds the steady state solution of the gradient flow by iteratively computing the solution of the equation $\frac{\partial \tilde{\mathbf{u}}}{\partial t} = -\frac{\partial E}{\partial \tilde{\mathbf{u}}}$, where the $\frac{\partial E}{\partial \mathbf{u}}$ denotes the Gâteaux derivative of $E(\tilde{\mathbf{u}})$.

The energy functional proposed in (4) is not convex because although the functional $E(\tilde{\mathbf{u}})$ is convex, the domain of optimisation is a non-convex set of functions. Following the approach by Chan et al. [18], we relax $\tilde{\mathbf{u}} : \Omega \rightarrow \{0,1\}$ to $\mathbf{u} : \Omega \rightarrow [0,1]$ so it can represent a convex set of functions (i.e., the domain of optimisation

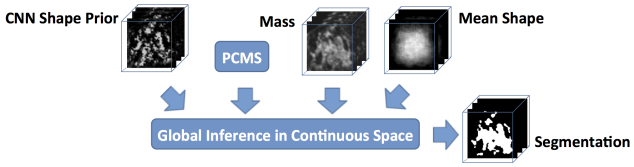


Fig. 2. Energy functional terms of the proposed method: shape prior from the CNN semantic segmentation (1), PCMS appearance and shape terms, and mean shape from training annotations. The final segmentation is estimated with a global inference on a continuous space.

is now convex), and as a result transform (4) into a convex optimisation problem. Theorem 1 in [18] assures the existence, but not uniqueness, of a global minimiser of the original problem reached by thresholding $\mathbf{u}(.)$:

$$\mathbf{y}^{*,GO} = \mathbf{1}_{\Sigma(x)}, \text{ with } \Sigma = \{x \in \Omega \mid \mathbf{u}(x) > \tau\}, \qquad (5)$$

where $\tau \in [0, 1]$ and $\mathbf{1}_{\Sigma(x)}$ denotes an indicator function that returns 1 if $x \in \Sigma$. Therefore, the inference based on the minimisation of $E(\mathbf{u})$ represents a globally optimal inference in the continuous space, where the main advantage lies in the use of arbitrary initialisation.

2.3. Breast Mass Detection based on a Cascade of Deep Learning Models

Initial mass regions are found automatically. Firstly, we follow [7] to obtain the breast region in the volume by applying Hayton's algorithm [20]. Then, the mass detection method extends the approach proposed by Dhungel at al. [10] that finds breast masses in mammograms. The proposed approach consists of a cascade of CNNs (1), where the first stage is a pixel-wise detection of mass candidates, using as input the 3-D data of the different MRI modalities: T1-weighted, T2-weighted and DCE-MRI. The second stage uses connected component analysis to merge the mass candidate voxels to form region candidates. Finally, these region candidates are fed into a cascade of two CNNs that process those candidates sequentially. The activations from the last layer of each CNN are concatenated and passed into a random forest classifier [21] to produce a final region classification. This approach improves Dhungel et al.'s approach [10] by reducing the complexity of the cascade in terms of the number of stages and by including multimodal image data.

3. EXPERIMENTS

This section introduces the dataset and experimental setup, the methods used in the comparison, the details of our proposed method, and the results.

3.1. Dataset

The dataset used to assess the accuracy of our methodology contains breast MRI studies of 117 patients, where the mean age is 48 $\pm$ 12 with age range between 22 and 84 years. Three image modalities are used in this study: DCE-MRI scans for segmentation and ROI detection, and T1-weighted anatomical and T2-weighted anatomical scans for ROI detection only. All images were acquired on a 1.5T GE Signa HDxt scanner, with the patient in prone position. T1-weighted anatomical volumes acqusition was performed axially (acquisition matrix of 512×512) and without fat suppression. T2-weighted anatomical volumes were obtained axially (acquisition matrix of 320×224) and with fat suppression. In order to obtain the DCE-MRI volumes, four or five volumes were acquired axially with fat suppression. The first (pre-contrast) volume is obtained before a contrast agent is injected to the patient. Then, several acquisitions are obtained at different time points (post-contrast volumes), where subtraction volumes are obtained by subtracting pre-contrast and post-contrast volumes. The acquisition matrix is 360x360 for these DCE-MRI volumes. All images for each patient are registered to the first post-contrast volume. There is at least one breast mass present in each DCE-MRI study, where the total number of masses is 141 (46 benign and 95 malignant) that were cyto- or histopathology confirmed. All masses were annotated by a radiographer using a region growing algorithm on the subtraction volumes [22].

The dataset was randomly divided into training and testing sets. In contrast to [7], where the training and testing data consisted of 35 (41 lesions) and 85 patients (93 lesions) respectively, our training set contains studies from 58 patients, with 72 lesions (23 benign and 49 malignant), and testing has the studies from 59 patients, with 69 lesions (23 benign and 46 malignant). Segmentation accuracy is assessed with the mean and median Dice coefficient on the training and testing sets.

3.2. Experimental Setup

We evaluate our segmentation methodology on the first subtraction image of DCE-MRI. As the initial ROI, we use both automated (as explained in Sec 2.3) and manual detection. A lesion is correctly detected when the Dice coefficient between the manual annotation and the ROI is at least 0.4, yielding a true positive rate (TPR) of 0.85 at 3.66 false positive regions per patient. T1-weighted, T2-weighted and the first two subtraction volumes are used for all deep learning models during the detection phase. In the case of the manual set-up, the initial region is the bounding box of the ground truth augmented by three voxels in each direction.

For the segmentation baseline methods, we use the results of the globally optimal inference on a discrete space (GODS) method [7] that holds the current state-of-the-art results for the dataset above. We re-implemented the locally optimal inference on a continuous space (LOCS) [8] that uses a DL shape prior model on a distance regularised level set method [23]. In addition, we also implemented the segmentation for globally optimal inference on a continuous space with a mean shape prior (GOCS-MS) [9]. Finally, we also implemented the CNN semantic segmentation [11] for comparison.

For our methodology, we use the training set to estimate the mean shape $\bar{\mathbf{y}}$ in (4), the weights and biases of the CNN in (1), and the weights of the terms in the energy functional (4). The CNN consists of 3 convolutional layers, with linear activation functions, and an output layer with two channels of size $30 \times 25 \times 18$ representing the probability of background (channel 1) or mass (channel 2). This fixed output size requires that the annotations are resized to fit that output layer during training. The first layer has 10 filters of size $5 \times 5 \times 3$, while the second and third layers contain 20 filters of size $3 \times 3 \times 3$. The learning rate is 0.1 and the input layer is a volume of size $30 \times 25 \times 18$, where the input volume is resized with cubic interpolation to fit this input layer. The CNN structure and its weights and biases in (1) are estimated using exclusively the training set, which is sub-divided into training (with 45 patients, containing 57 lesions) and validation (with 13 patients and 15 lesions) sets, for model selection. We augment the training data by flipping

	Mean Dice		Median Dice			
	Train	Test	Train	Test	Detection	Inference Time
GOCS-DLP(Ours)	0.80 ± 0.11	$\mathbf{0.77 \pm 0.14}$	0.82	**0.82**	Auto	7.78 ± 20.69 s
GOCS-DLP(Ours)	0.79 ± 0.13	$\mathbf{0.77 \pm 0.13}$	0.81	0.80	Manual	5.95 ± 18.07 s
GODS [7]	-	0.74 ± 0.12	-	0.76	Auto	-
LOCS [8]	0.64 ± 0.16	0.62 ± 0.15	0.65	0.64	Auto	14.37 ± 41.15 s
LOCS [8]	0.61 ± 0.15	0.59 ± 0.17	0.64	0.61	Manual	12.90 ± 28.13 s
GOCS-MS [9]	0.76 ± 0.18	0.73 ± 0.21	0.81	0.79	Auto	7.61 ± 21.44 s
GOCS-MS [9]	0.75 ± 0.18	0.72 ± 0.22	0.80	0.79	Manual	5.83 ± 15.02 s
CNN	0.69 ± 0.16	0.68 ± 0.19	0.70	0.74	Auto	0.12 ± 0.16 s
CNN	0.66 ± 0.16	0.66 ± 0.18	0.68	0.69	Manual	0.11 ± 0.03 s

Table 1. Mean, median and standard deviation for training/testing Dice coefficients and inference time (per lesion) for each methodology.

ROIs in each of the three possible axes. Finally, the weights in the energy functional in (4) are estimated (using the training set) over a grid of possible values for each term, where the estimated values are $\lambda = 55, \alpha = 2.5, \beta = 1.5$. We label our approach as globally optimal inference on a continuous space using DL shape prior (GOCS-DLP). Note that for the LOCS and GOCS-MS method, we run a similar method to estimate these weights.

For the inference of GOCS-DLP, GOCS-MS and LOCS the initial segmentation consists of a rectangular centered prism of 90% of the ROI volume. When the optimisation process has converged, we threshold the solution of the relaxed problem at the value of $\tau = 0.75$ as defined in (5).

The same training, validation and testing sets are employed to automatically detect masses. For the first stage, we use a three-scale CNN. The network architecture in every scale is composed of 4 convolutional layers, a fully connected layer and a softmax operation to normalize probabilities of being normal and mass tissue. The multi-scale CNN produces an average of 30 false positive ROIs per image. Such candidates are fed into a cascade of two CNNs, each containing two convolutional - max pooling layers, two convolutional layers, a fully connected layer and a softmax layer. Activations from both last layers are concatenated together to form the input to a random forest classifier to produce the final classification.

3.3. Results

We compare our proposed GOCS-DLP with GODS, LOCS, GOCS-MS, and the CNN semantic segmentation for the training and testing sets in Table 1, where the last column refers to the average time employed for the segmentation of one lesion. We measure the statistical significance of the results for both automated and manual detection in Table 1 with the Wilcoxon signed-rank test, and the results from the proposed GOCS-DLP are significant compared to all others, assuming a significance level of 0.01. In particular, p-values of fully automated GOCS-DLP with respect to fully automated GODS and GOCS-MS are 0.0081 and 0.0005 respectively. Fig.3 shows examples of mass segmentations achieved with our proposed GOCS-DLP after an automated detection.

4. DISCUSSION AND CONCLUSION

The experimental resuts show that the segmentation accuracy produced by our proposed GOCS-DLP is significantly better ($p < 0.01$) than the baselines GODS, LOCS and GOCS-MS and CNN semantic segmentation for the problem of breast mass segmentation from DCE-MRI. Note that the segmentation accuracy using the manual ROI detection is not better than its automated counterpart because masses that are not automatically detected are not passed to the segmentation stage, and these masses turn out to be the most challenging ones to be segmented.

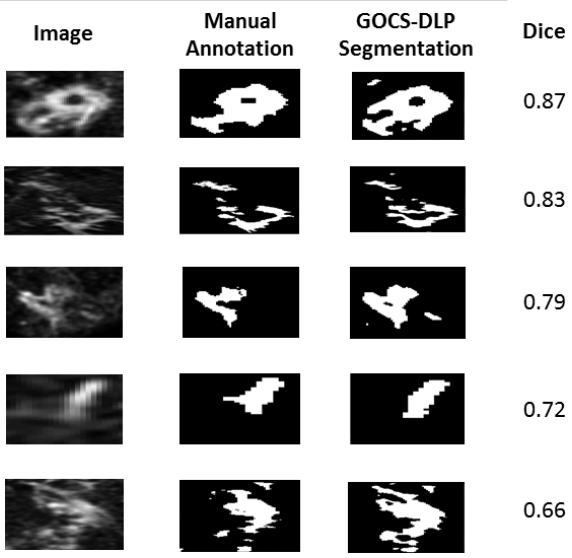


Fig. 3. Examples of mass segmentations produced by our proposed GOCS-DLP. GODS results not available

In Sec. 1, we hypothesise that the semantic segmentation from CNN would not be robust enough because of the small training set, and the evidence in Table 1 provides support for that hypothesis. Table 1 also shows that the inclusion of the DL shape prior significantly improves the quality of the segmentation (GOCS-DLP vs GOCS-MS). However, we expect that for large training sets, the CNN alone will be able to produce accurate segmentation on its own. In addition, one of the advantages of globally optimal inference compared to the locally optimal inference is the independence with respect to the initialisation: global methods produce significantly better segmentation than the local method. This advantage is particularly effective to avoid large errors in segmentation after an automated detection of the lesion, which might yield a misaligned ROI. In fact, even a more precise initialisation (the bounding box of the manual annotation) for LOCS does not achieve an accurate segmentation ($\bar{D} = 0.68$ for the test set) compared to the global methods.

The inference time in Table 1 shows that the global methods converge much faster. Note that the large variability in the inference time is due to the variation of shape and size of masses. Even though we do not have the training and inference time results for GODS [7], the comparison in terms of "Mean Dice Test" and "Median Dice Test" shows evidence of the disadvantage of using superpixels to decrease the memory complexity that implicitly assumes appearance homogeneity, which may not be correct for this problem. Finally, the visual results in Fig. 3 show that our proposed GOCS-DLP produces quite precise segmentation results for different types of masses.

In this work, we introduce a new segmentation method that combines global inference in the continuous space with deep learning for the problem of breast mass segmentation from DCE-MRI. Our results show a significant improvement over the state-of-art for this problem, where we also present results produced by several baseline methods based on DL alone, discrete global optimisation and continuous global optimisation. We intend to apply the proposed methodology in other medical imaging segmentation problems.

5. REFERENCES

[1] Albert L Siu, "Screening for breast cancer: Us preventive services task force recommendation statement," *Annals of internal medicine*, 2016.

[2] Per Skaane, Andriy I Bandos, Randi Gullien, Ellen B Eben, Ulrika Ekseth, Unni Haakenaasen, Mina Izadi, Ingvild N Jebsen, Gunnar Jahr, Mona Krager, et al., "Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program," *Radiology*, vol. 267, no. 1, pp. 47–56, 2013.

[3] Debbie Saslow, Carla Boetes, Wylie Burke, Steven Harms, Martin O Leach, Constance D Lehman, Elizabeth Morris, Etta Pisano, Mitchell Schnall, Stephen Sener, et al., "American cancer society guidelines for breast screening with mri as an adjunct to mammography," *CA: a cancer journal for clinicians*, vol. 57, no. 2, pp. 75–89, 2007.

[4] C Dromain, B Boyer, R Ferre, S Canale, S Delaloge, and C Balleyguier, "Computed-aided diagnosis (cad) in the detection of breast cancer," *European journal of radiology*, vol. 82, no. 3, pp. 417–423, 2013.

[5] Yading Yuan, Maryellen L Giger, Hui Li, Kenji Suzuki, and Charlene Sennett, "A dual-stage method for lesion segmentation on digital mammograms," *Medical physics*, vol. 34, no. 11, pp. 4180–4193, 2007.

[6] Jagadaeesan Jayender, Sona Chikarmane, Ferenc A Jolesz, and Eva Gombos, "Automatic segmentation of invasive breast carcinomas from dynamic contrast-enhanced mri using time series analysis," *Journal of Magnetic Resonance Imaging*, vol. 40, no. 2, pp. 467–475, 2014.

[7] Darryl McClymont, Andrew Mehnert, Adnan Trakic, Dominic Kennedy, and Stuart Crozier, "Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph," *Journal of Magnetic Resonance Imaging*, vol. 39, no. 4, pp. 795–804, 2014.

[8] Tuan Ngo and Gustavo Carneiro, "Fully automated non-rigid segmentation with distance regularized level set evolution initialized and constrained by deep-structured inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3118–3125.

[9] Daniel Cremers, Frank R Schmidt, and Frank Barthel, "Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–6.

[10] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, 2015, pp. 1–8.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[12] Ali Qusay Al-Faris, Umi Kalthum Ngah, Nor Ashidi Mat Isa, and Ibrahim Lutfi Shuaib, "Computer-aided segmentation system for breast mri tumour using modified automatic seeded region growing (bmri-masrg)," *Journal of digital imaging*, vol. 27, no. 1, pp. 133–144, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[16] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[17] Stanley Osher and James A Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[18] Tony F Chan, Selim Esedoglu, and Mila Nikolova, "Algorithms for finding global minimizers of image segmentation and denoising models," *SIAM journal on applied mathematics*, vol. 66, no. 5, pp. 1632–1648, 2006.

[19] David Mumford and Jayant Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on pure and applied mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[20] Paul Hayton, Michael Brady, Lionel Tarassenko, and Niall Moore, "Analysis of dynamic mr breast images using a model of contrast enhancement," *Medical image analysis*, vol. 1, no. 3, pp. 207–224, 1997.

[21] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] Antoine Rosset, Luca Spadola, and Osman Ratib, "Osirix: an open-source software for navigating in multidimensional dicom images," *Journal of digital imaging*, vol. 17, no. 3, pp. 205–216, 2004.

[23] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox, "Distance regularized level set evolution and its application to image segmentation," *Image Processing, IEEE Transactions on*, vol. 19, no. 12, pp. 3243–3254, 2010.

Chapter 4

Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI

The work contained in this chapter has been published as the following paper:

Gabriel Maicas, Gustavo Carneiro, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid. Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.

DOI: https://doi.org/10.1007/978-3-319-66179-7_76

Statement of Authorship

Title of Paper	Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Gabriel Maicas, Gustavo Carneiro, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid. Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.

Principal Author

Name of Principal Author (Candidate)	Gabriel Maicas Suso		
Contribution to the Paper	- Developed the idea of the paper - Coded the proposed algorithm - Designed experiments to validate the algorithm - Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	27 July 2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Developed the idea of the paper - Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	03-08-18

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	- Refined the manuscript		
Signature		Date	27 July 2018

Name of Co-Author	Jacinto C. Nascimento		
Contribution to the Paper	- Supervised the development of the work - Refined the manuscript		
Signature		Date	23/07/2018

Name of Co-Author	Ian Reid		
Contribution to the Paper	- Supervised the development of the work - Refined the manuscript		
Signature		Date	13/8/18

Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI ⋆

Gabriel Maicas[†] Gustavo Carneiro[†] Andrew P. Bradley[‡]
Jacinto C. Nascimento[‡‡] Ian Reid[†]

[†]ACVT, School of Computer Science, The University of Adelaide
[‡]School of ITEE, The University of Queensland, Australia
[‡‡]Institute for Systems and Robotics, Instituto Superior Tecnico, Portugal

Abstract. We present a novel methodology for the automated detection of breast lesions from dynamic contrast-enhanced magnetic resonance volumes (DCE-MRI). Our method, based on deep reinforcement learning, significantly reduces the inference time for lesion detection compared to an exhaustive search, while retaining state-of-art accuracy.
This speed-up is achieved via an attention mechanism that progressively focuses the search for a lesion (or lesions) on the appropriate region(s) of the input volume. The attention mechanism is implemented by training an artificial agent to learn a search policy, which is then exploited during inference. Specifically, we extend the deep Q-network approach, previously demonstrated on simpler problems such as anatomical landmark detection, in order to detect lesions that have a significant variation in shape, appearance, location and size. We demonstrate our results on a dataset containing 117 DCE-MRI volumes, validating run-time and accuracy of lesion detection.

Keywords: deep Q-learning, Q-net, reinforcement learning, breast lesion detection, magnetic resonance imaging

1 Introduction

Breast cancer is amongst the most commonly diagnosed cancers in women [1, 2]. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) represents one of the most effective imaging techniques for monitoring younger, high-risk women, who typically have dense breasts that show poor contrast in mammography [3]. DCE-MRI is also useful during surgical planning once a suspicious lesion is found on a mammogram [3]. The first stage in the analysis of these 4D (3D over time) DCE-MRI volumes consists of the localisation of breast lesions. This is a challenging task given the high dimensionality the data (4 volumes each containing $512 \times 512 \times 128$ voxels), the low signal to noise ratio of the dynamic sequence and the variable size and shape of breast lesions (see Fig. 1). Therefore, a computer-aided detection (CAD) system that automatically localises breast lesions in DCE-MRI data would be a useful tool to facilitate radiologists. However, the high dimensionality of DCE-MRI requires computationally efficient methods for lesion detection to be developed to be viable for practical use.

⋆ Supported by Australian Research Council through grants DP140102794, CE140100016 and FL130100102.

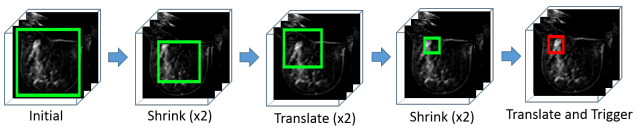


Fig. 1: Example of the detection process of breast lesions from DCE-MRI with DQN. Depth transformation are not shown for simplicity.

Current approaches to lesion detection in DCE-MRI rely on extracting handcrafted features [4, 5] and exhaustive search mechanisms [4–6] in order to handle the variability in lesion appearance, shape, location and size. These methods are both computationally complex and potentially sub-optimal, resulting in false alarms and missed detections. Similar issues in the detection of visual objects have motivated the computer vision community to develop efficient detectors [7, 8], like the Faster R-CNN [7]. However, these models need large annotated training sets making their application in medical image analysis (MIA) challenging [9]. Alternatively, Caicedo and Lazebnik [8] have recently proposed the use of a deep Q-network (DQN) [10] for efficient object detection that allows us to deal with the limited amount of data. Its adaptation to MIA applications has to overcome two additional obstacles: 1) the extension from visual object classes (e.g., animals, cars, etc.) to objects in medical images, such as tumours, which tend to have weaker consistency in terms of shape, appearance, location, background and size; and 2) the high dimensionality of medical images, which presents practical challenges with respect to the DQN training process [10]. Ghesu et al. [11] have recently adapted DQN [10] to anatomical landmark detection, but did not address the obstacles mentioned above because the visual classes used in their work have consistent patterns and are extracted from fixed small-size regions of the medical images.

Here, we introduce a novel algorithm for breast lesion detection from DCE-MRI inspired by a previously proposed DQN [10, 8]. Our main goal is the reduction of run time complexity without a reduction in detection accuracy. The proposed approach comprises an artificial agent that automatically learns a policy, describing how to iteratively modify the focus of attention (via translation and scale) from an initial large bounding box to a smaller bounding box containing a lesion, if it exists (see Fig. 2). To this end, the agent constructs a deep learning feature representation of the current bounding box, which is used by the DQN to decide on the next action, i.e., either to translate or scale the current bounding box or to trigger the end of the search process. Our methodology is the first DQN [10] that can detect such visually challenging objects. In addition, unlike [11] that uses a fixed small-size bounding box, our DQN utilises a variable-size bounding box. We evaluate our methodology on a dataset of 117 patients (58 for training and 59 for testing). Results show that our methodology achieves a similar detection accuracy compared to the state of the art [6, 5], but with significantly reduced run times.

Table 1: Summary of results from previous approaches.

	Evaluation Criteria	**Time**
Vignati at al. [12]	0.89 TPR @ 12.00 FPI	7.00 min
Renz et al. [13]	0.96 Sensitivity @ 0.75 Specificity	-
Gubern-Merida et al. [4]	0.89 TPR @ 4.00 FPI	-
McClymont et al. [5]	1.00 TPR @ 4.50 FPI	$\mathcal{O}(60)$min
Maicas et al. [6]	0.80 TPR @ 2.80 FPI	2.74 min

2 Literature Review

Automated approaches for breast lesion detection from DCE-MRI are typically based on exhaustive search methods and hand-designed features [12, 13, 4, 5]. Vignati at al. [12] proposed a method that thresholds an intensity normalised DCE-MRI to detect voxel candidates that are merged to form lesion candidates, from which hand-designed region and kinetic features are used in the classification process. As shown in Tab. 1, this method has low accuracy that can be explained by the fact that this method makes strong assumptions about the role of DCE-MRI intensity and does not utilise texture, shape, location and size features. Renz et al. [13] extended Vignati at al.'s work [12] with the use of additional hand-designed morphological and dynamical features, showing more competitive results (see Tab. 1). Further improvements were obtained by Gubern-Merida et al. [4], with the addition of shape and appearance hand-designed features, as shown in Tab. 1. The run-time complexity of the approaches above can be summarised by the mean running time (per volume) shown by Vignati at al.'s work [12] in Tab. 1, which is likely the most efficient of these three approaches [12, 13, 4]. McClymont et al. [5] extended the methods above with the unsupervised voxel clustering for the initial detection of lesion candidates, followed by a structured output learning approach that detects and segments lesions simultaneously. This approach significantly improves the detection accuracy, but at a substantial increase in computational cost (see Tab. 1). The multi-scale deep learning cascade approach [6] reduced the run-time complexity, allowed the extraction of optimal and efficient features, and had a competitive detection accuracy as shown in Tab. 1.

There are two important issues regarding previously proposed approaches: the absence of a common dataset to evaluate different methodologies and the lack of a consistent lesion detection criterion. Whereas detections in [12, 13] were visually inspected by a radiologist, [4, 5] considered a lesion detected if a (single) voxel in the ground truth was detected. In [6] a more precise criterion (minimum Dice coefficient of 0.2 between ground truth and candidate bounding box) was used - in the experiment, we adopt this Dice > 0.2 criterion and use the same dataset as a few previous studies [5, 6].

3 Methodology

In this section, we first define the dataset, then the training and inference stages of our proposed methodology, shown in Fig. 2.

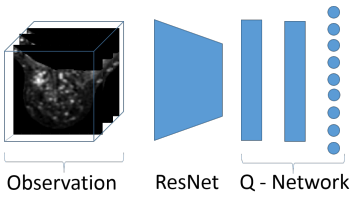


Fig. 2: Block diagram of the proposed detection system.

3.1 Dataset

The data is represented by a set of 3D breast scans $\mathcal{D} = \left\{ \left(\mathbf{x}, \mathbf{t}, \{\mathbf{s}^{(j)}\}_{j=1}^{M} \right)_i \right\}_{i=1}^{|\mathcal{D}|}$, where each $, \mathbf{x}, \mathbf{t} : \Omega \rightarrow \mathbb{R}$ denotes the first DCE-MRI subtraction volume and the T1-weighted anatomical volume, respectively, with $\Omega \in \mathbb{R}^3$ representing the volume lattice of size $w \times h \times d$; $\mathbf{s}^{(j)} : \Omega \rightarrow \{0, 1\}$ represents the annotation for the j^{th} lesion present, with $\mathbf{s}^{(j)}(\omega) = 1$ indicating presence of lesion at voxel $\omega \in \Omega$. The entire dataset is patient-wise split such that the mutually exclusive training and testing datasets are represented by $\mathcal{T}$,$\mathcal{U} \subset \mathcal{D}$, where $\mathcal{T} \bigcup \mathcal{U} = \mathcal{D}$.

3.2 Training

The proposed DQN [10] model is trained via interactions with the DCE-MRI dataset through a sequence of observations, actions and rewards. Each observation is represented by $\mathbf{o} = f(\mathbf{x}(\mathbf{b}))$, where $\mathbf{b} = [b_x, b_y, b_z, b_w, b_h, b_d] \in \mathbb{R}^6$ (where b_x, b_y, b_z represent the top-left-front corner and b_w, b_h, b_d denotes the lower-right-back corner of the bounding box) indexes the input DCE-MRI data $\mathbf{x}$, and $f(.)$ denotes a deep residual network (ResNet) [14, 15], defined below. Each action is denoted by $a \in \mathcal{A} = \{l_x^+, l_x^-, l_y^+, l_y^-, l_z^+, l_z^-, s^+, s^-, w\}$, where l, s, w represent translation, scale and trigger actions, with the subscripts x, y, z denoting the horizontal, vertical or depth translation, and superscripts $+, -$ meaning positive or negative translation and up or down scaling. The reward when the agent chooses the action $a = w$ to move from $\mathbf{o}$ to $\mathbf{o}'$ is defined by:

$$r(\mathbf{o}, a, \mathbf{o}') := \begin{cases} +\eta, & \text{if } d(\mathbf{o}', \mathbf{s}) \geq \tau_w \\ -\eta, & \text{otherwise} \end{cases}, \tag{1}$$

where $d(.)$ is the Dice coefficient between a map formed by the bounding box $\mathbf{o} = f(\mathbf{x}(\mathbf{b}))$ and the segmentation map $\mathbf{s}$, $\eta = 10$ and $\tau_w = 0.2$ (these values have been empirically defined - for instance, we found that increasing η to 10.0 from 3.0 used in [8] helped triggering when finding a lesion). For the remaining of the actions in $\mathcal{A} \setminus \{w\}$, the rewards are defined by:

$$r(\mathbf{o}, a, \mathbf{o}') := \text{sign}(d(\mathbf{o}', \mathbf{s}) - d(\mathbf{o}, \mathbf{s})). \tag{2}$$

The training process models a DQN that maximises cumulative future rewards with the approximation of the following action-value function: $Q^*(\mathbf{o}, a) =$

$\max_\pi \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... \mid \mathbf{o}_t = \mathbf{o}, a_t = a, \pi]$, where r_t denotes the reward at time step t, γ represents a discount factor per time step, and π is the behaviour policy. This action-value function is modelled by a DQN $Q(\mathbf{o}, a, \theta)$, where θ denotes the network weights. The training of $Q(\mathbf{o}, a, \theta)$ is based on experience replay memory and the target network [10]. Experience replay uses a dataset $\mathcal{E}_t = \{e_1, ..., e_t\}$ built with the agent's experiences $e_t = (\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1})$, and the target network with parameters θ_i^- computes the target values for the DQN updates, where the values θ_i^- are held fixed and updated periodically. The loss function for modelling $Q(\mathbf{o}, a, \theta)$ minimises the mean-squared error of the Bellman equation, as in:

$$L_i(\theta_i) = \mathbb{E}_{(\mathbf{o},a,r,\mathbf{o}')\sim U(\mathcal{E})}\left[\left(r + \gamma \max_{a'} Q(\mathbf{o}', a'; \theta_i^-) - Q(\mathbf{o}, a; \theta_i)\right)^2\right]. \quad (3)$$

In the training process, we follow an ϵ-greedy strategy to balance exploration and exploitation: with probability ϵ, the agent explores, and with probability 1-ϵ, it will follow the current policy π (exploitation) for training time step t. At the beginning of the training, we set $\epsilon = 1$ (i.e., pure exploration), and decrease ϵ as the training progresses (i.e., increase exploitation). Furthermore, we follow a modified guided exploration: with probability κ, the agent will select a random action and with probability $1-\kappa$, it will select an action that produces a positive reward. This modifies the guided exploration in [8] by adding randomness to the process, aiming to improve generalisation. Finally, the ResNet [14, 15], which produces the observation $\mathbf{o} = \mathbf{x}(\mathbf{b})$, is trained to decide whether a random bounding box $\mathbf{b}$ contains a lesion. A training sample is labelled as positive if $d(\mathbf{o}, \mathbf{s}_j) \geq \tau_w$, and negative, otherwise. It is important to notice that this way of labelling random training samples can provide a large and balanced training set, extracted at several locations and scales, that is essential to train the large capacity ResNet [14, 15]. In addition, this way of representing the bounding box means that we are able to process varying-size input bounding box, which is an advantage compared to [11].

3.3 Inference

The trained DQN model is parameterised by θ^* learned in (3) and is defined by a multi-layer perceptron [8] that outputs the action-value function for the observation $\mathbf{o}$. The action to follow from the current observation is defined by:

$$a^* = \arg\max_a Q(\mathbf{o}, a, \theta^*). \quad (4)$$

Finally, given that the number and location of lesions are unknown in a test DCE-MRI, this inference is initialised with different bounding boxes at several locations, and it runs until it either finds the lesion (with the selection of the trigger action), or runs for a maximum number of 20 steps.

4 Experiments

The **database** used to assess our proposed methodology contains DCE-MRI and T1-weighted anatomical datasets from 117 patients [5]. For the DCE-MRI, the

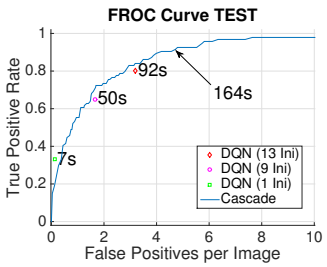
first volume was acquired before contrast agent injection (pre-contrast), and the remaining volumes were acquired after contrast agent injection. Here we use only one volume represented by the first subtraction from DCE-MRI: the first post-contrast volume minus pre-contrast volume. The T1-weighted anatomical is used only to extract the breast region from the initial volume [5], as a pre-processing stage. The training set contains 58 patients annotated with 72 lesions, and the testing set has 59 patients and 69 lesions to allow a fair comparison with [6]. The detection accuracy is assessed by the proportion of true positives (TPR) detected in the training and testing sets as a function of the number of false positives per image (FPI), where a candidate lesion is assumed to be a true positive if the Dice coefficient between the candidate lesion bounding box and the ground truth annotation bounding box is at least 0.2 [16]. We also measure the running time of the detection process using the following computer: CPU: Intel Core i7 with 12 GB of RAM and a GPU Nvidia Titan X 12 GB.

The **pre-processing** stage of our methodology consists of the extraction of each breast region (from T1-weighted) [5], and separate each breast into a resized volume of ($100 \times 100 \times 50$) voxels. For training, we select breast region volumes that contain at least one lesion, but if a breast volume has more than one lesion, one of them is randomly selected to train the agent. For testing, a breast may contain none, one or multiple lesions. The **observation o** used by DQN is produced by a **ResNet** [14] containing five residual blocks. The input to the ResNet is fixed at ($100 \times 100 \times 50$) voxels.We extract 16K patches (8K positives and 8K negatives) from the training set to train the ResNet to classify a bounding box as positive or negative for a lesion, where a bounding box is labelled as positive if the Dice coefficient between the lesion candidate and the ground truth annotation is at least 0.6. This ResNet provides a fixed size representation for $\mathbf{o}$ of size 2304 (extracted before the last convolutional layer).

The **DQN** is represented by a multilayer perceptron with two layers, each containing 512 nodes, that outputs nine actions: six translations (by one third of the size of the corresponding dimension), two scales (by one sixth in all dimensions) and a trigger (see Sec. 3.2). For training this DQN, the agent starts an episode with a centred bounding box occupying 75% of the breast region volume. The experience replay memory $\mathcal{E}$ contains 10K experiences, from which 100 mini-batch samples are drawn to minimise the loss (3). The DQN is trained with Adam, using a learning rate of 1×10^{-6}, and the target network is updated after running one episode per volume of the training set. For the ϵ-greedy strategy (Sec. 3.2), ϵ decreases linearly from 1 to 0.1 in 300 epochs, and during exploration,the balance between random exploration and modified guided exploration is given by $\kappa = 0.5$. During inference, the agent follows the policy in (4), where for every breast region volume, it starts at a centred bounding box that covers 75% of the volume. Then it starts at each of the eight non-overlapping (50,50,25) bounding boxes corresponding to each of the corners. Finally, it is initialised at another four (50,50,25) bounding boxes centred at the intersections of the previous bounding boxes. The agent is allowed a maximum number of 20 steps to trigger, otherwise, no lesion is detected.

4.1 Results

We compare the training and testing results of our proposed DQN with the multi-scale cascade [6] and structured output approach [5] in the Table in Figure 3. In addition, we show the Free Response Operating Characteristic (FROC) curve in Fig. 3 comparing our approach (using a varying number of initialisations that lead to different TPR and FPI values) with the multi-scale cascade [6]. Finally, we show examples of the detections produced by our method in Fig. 4.



	Test		
	TPR	**FPR**	**Time**
DQN (Ours)	**0.80**	**3.2**	**92± 21s**
Ms-C [6]	0.80	2.8	164± 137s
SL [5]	1.00	4.50	$\mathcal{O}(60)$ m

Fig. 3: FROC curve showing TPR vs FPI and run times for DQN and the multi-scale cascade [6] (left) TPR, FPR and mean inference time per case (i.e. per patient) for each method (right). Note run time for Ms-C is constant over the FPI range.

We use a paired t-test to estimate the significance of the inference times between our approach and the multi-scale cascade [6], giving $p \leq 9 \times 10^{-5}$.

5 Discussion and Conclusion

We have presented a DQN method for lesion detection from DCE-MRI that shows similar accuracy to state of the art approaches, but with significantly reducing detection times. Given that we did not attempt any code optimisation, we believe that the run times have the potential for further improvement. For example, inference uses several initialisations (up to 13), which could be run in parallel as they are independent, decreasing detection time by a factor of 10. The main bottleneck of our approach is the volume resizing stage that transforms the current bounding box to fit the ResNet input - currently representing 90% of the inference time. A limitation of this work is that we do not have an action to change the aspect ratio of the bounding box, which may improve detection of small elongated lesions. Finally, during training, we noted that the most important parameter to achieve good generalisation is the balance between exploration and exploitation. We observed that the best generalisation was achieved when $\epsilon = 0.5$ (i.e. half of the actions correspond to exploration and half to exploitation of the current policy). Future research will improve run-time performance via learning smarter search strategies. For instance, we would like to avoid revisiting regions that have already been determined to be free from lesions with high probability. At present we rely on the training data to discourage such moves, but there may be more explicit constraints to explore. We would like to acknowledge NVIDIA for providing the GPU used in this work.

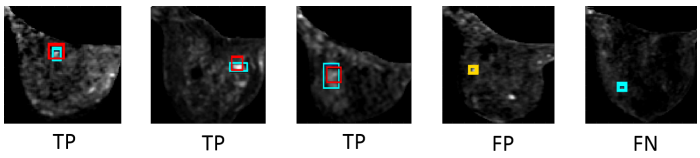


Fig. 4: Examples of detected breast lesions. Cyan boxes indicate the ground truth, red boxes detections produced by our proposed method and yellow false positive detections.

References

1. Smith, R.A., Andrews, K., Brooks, D., et al.: Cancer screening in the United States, 2016: A review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians (2016)
2. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: A Cancer Journal for Clinicians **66**(1) (2016) 7–30
3. Siu, A.L.: Screening for breast cancer: US preventive services task force recommendation statement. Annals of internal medicine (2016)
4. Gubern-Mérida, A., Martí, R., Melendez, J., et al.: Automated localization of breast cancer in DCE-MRI. Medical image analysis **20**(1) (2015) 265–274
5. McClymont, D., Mehnert, A., Trakic, A., et al.: Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph. JMRI **39**(4) (2014) 795–804
6. Maicas, G., Carneiro, G., Bradley, A.P.: Globally optimal breast mass segmentation from DCE-MRI using deep semantic segmentation as shape prior. In: 14th International Symposium on Biomedical Imaging (ISBI), IEEE (2017) 305–309
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. (2015) 91–99
8. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: CVPR. (2015) 2488–2496
9. Akselrod-Ballin, A., Karlinsky, L., Alpert, S., et al.: A region based convolutional network for tumor detection and classification in breast mammography. In: DLMIA 2016: Deep Learning and Data Labeling for Medical Applications, Springer (2016)
10. Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540) (2015) 529–533
11. Ghesu, F.C., Georgescu, B., Mansi, T., et al.: An artificial agent for anatomical landmark detection in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2016) 229–237
12. Vignati, A., Giannini, V., De Luca, M., et al.: Performance of a fully automatic lesion detection system for breast dce-mri. JMRI **34**(6) (2011) 1341–1351
13. Renz, D.M., Böttcher, J., Diekmann, F., et al.: Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast MRI. JMRI **35**(5) (2012) 1077–1088
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
15. Huang, G., Sun, Y., Liu, Z., et al.: Deep networks with stochastic depth. In: ECCV, Springer (2016) 646–661
16. Dhungel, N., Carneiro, G., Bradley, A.P.: Automated mass detection in mammograms using cascaded deep learning and random forests. In: DICTA, IEEE (2015)

Chapter 5

Training Medical Image Analysis Systems like Radiologists

The work contained in this chapter has been published as the following paper:

Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Training Medical Image Analysis Systems like Radiologists. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018. **Oral Presentation.**

DOI: https://link.springer.com/chapter/10.1007/978-3-030-00928-1_62

Statement of Authorship

Title of Paper	Training Medical Image Analysis Systems like Radiologists
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Training Medical Image Analysis Systems like Radiologists. International Conference on Medical Imaging Computing and Computer Assisted Intervention (MICCAI). 2018

Principal Author

Name of Principal Author (Candidate)	Gabriel Maicas Suso
Contribution to the Paper	- Developed the idea of the paper - Coded the proposed algorithm - Designed experiments to validate the algorithm - Wrote and refined the manuscript
Overall percentage (%)	50%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date: 27 July 2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Andrew P. Bradley
Contribution to the Paper	- Refined the manuscript
Signature	Date: 27 July 2018

Name of Co-Author	Jacinto C. Nascimento
Contribution to the Paper	- Refined the manuscript
Signature	Date: 23/07/2018

Name of Co-Author	Ian Reid		
Contribution to the Paper	- Supervised the development of the work - Refined the manuscript		
Signature		Date	13/8/18

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Developed the idea of the paper - Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	03-08-18

Training Medical Image Analysis Systems like Radiologists ⋆

Gabriel Maicas[†] Andrew P. Bradley[‡] Jacinto C. Nascimento[‡‡]
Ian Reid[†] Gustavo Carneiro[†]

[†]Australian Institute for Machine Learning, The University of Adelaide
[‡]Science and Engineering Faculty, Queensland University of Technology
[‡‡]Institute for Systems and Robotics, Instituto Superior Tecnico, Portugal

Abstract. The training of medical image analysis systems using machine learning approaches follows a common script: collect and annotate a large dataset, train the classifier on the training set, and test it on a hold-out test set. This process bears no direct resemblance with radiologist training, which is based on solving a series of tasks of increasing difficulty, where each task involves the use of significantly smaller datasets than those used in machine learning. In this paper, we propose a novel training approach inspired by how radiologists are trained. In particular, we explore the use of meta-training that models a classifier based on a series of tasks. Tasks are selected using teacher-student curriculum learning, where each task consists of simple classification problems containing small training sets. We hypothesize that our proposed meta-training approach can be used to pre-train medical image analysis models. This hypothesis is tested on the automatic breast screening classification from DCE-MRI trained with weakly labeled datasets. The classification performance achieved by our approach is shown to be the best in the field for that application, compared to state of art baseline approaches: DenseNet, multiple instance learning and multi-task learning.

Keywords: meta-learning, curriculum learning, multi-task training, breast image analysis, breast screening, magnetic resonance imaging.

1 Introduction

Radiologists are exceptionally trained specialists who play a crucial role interpreting and assisting other doctors and specialists in diagnosing and treating diseases. Their training program typically requires the trainee to solve tasks of increasing difficulty [1], where each task contains a relatively small number of "training images". Such a program bears little resemblance to the training of medical image analysis systems based on machine learning that are modeled to solve narrowly defined, but complex classification problems [2], requiring large training sets. Once trained, these models cannot be easily adapted to new problems – they must be re-trained with new large training sets. The use of pre-trained models [3] as a way of initializing a model is the first step towards

⋆ Supported by Australian Research Council through grants DP180103232, CE140100016 and FL130100102.

a more similar approach to the training program of radiologists. However, pre-training does not train a model to be able to learn new tasks – instead it is a "trick" to improve convergence and generalization. Meanwhile, machine learning researchers have developed more effective *learning to learn* approaches [4] – such approaches are motivated by the ability of humans to learn new tasks quickly and with limited "training sets". The optimization in such approaches penalizes classification loss and inefficient learning on new tasks (i.e., classification problems) by using a training scheme that continuously samples new tasks, mimicking the human training process. Our hypothesis is that medical machine learning methods could benefit from such an radiologist's style training process.

In this paper, we introduce an improved model agnostic meta-learning [4] (MAML) as a way of pre-training a classifier. The training process maximizes the ability of the classifier to adapt to new tasks using relatively small training sets. We also propose a technical innovation for MAML [4], by replacing the random task selection with teacher-student curriculum learning as an improved way for selecting tasks [5]. This task selection process is based on the model's performance on the tasks, trying to mimic radiologists' training. Our improved MAML is tested on weakly-supervised breast screening from DCE-MRI, where samples are globally annotated with classes (i.e. volume-level labels): *no findings*, *benign lesions* and *malignant lesions*, but these samples do not have lesion delineations. Note that the use of weakly-labeled datasets is becoming increasingly important for medical image analysis as this is the data available in clinical practice [2].

We test our proposed approach on a dataset of dynamic contrast enhanced MRI for the breast screening classification. Results show that our proposed approach improves the area under the ROC curve (AUC), outperforming baselines such as DenseNet [6], which holds the state-of-the-art (SOTA) for many classification problems; multiple-instance learning [7], which holds SOTA for breast screening in mammography; and multi-task learning [8]. Our learning approach produces an AUC of 0.90, which is better than the best result from the baseline methods that achieves an AUC of 0.85.

2 Literature Review

Breast screening from DCE-MRI aims at early detection of breast cancer in women at high-risk [9]. Currently, this screening process is mostly done manually, where its success depends on the radiologist's abilities [10]. An automated breast screening system working as a second reader can help radiologists reduce variability and increase the sensitivity and specificity of their readings. Traditionally, such systems rely on classifiers trained with large-scale strongly labeled datasets (i.e., containing lesion delineation and global classification) [11–15]. The non-scalability of this process (due to costs related to the annotation process) motivated the development of learning methods that can use weakly-labeled training sets [7] (i.e., samples contain only global classification). However, these methods still follow traditional machine learning approaches, which means that they still need large-scale training sets, even when the model has been pre-trained from other classification problems [3].

Contrasting with traditional machine learning algorithms, humans excel at learning new skills and new "classification" problems, where new learning tasks

often require fewer training samples than the ones before. This *learning to learn* ability has inspired the development of a new generation of machine learning algorithms. For example, multi-task learning uses an optimization function that is trained to simultaneously minimize the loss of several different, but related classification problems [8], helping the regularization of the training procedure. Nevertheless, multi-task learning does not address the issue of making a model effective at learning new classification problems with small datasets. This issue is addressed by *meta-learning* [4], which has been designed to solve the *few-shot* learning problem, where the classifier is trained to train for new classification problems with previously unseen classes containing a small number of images. In meta-learning for few-shot classification, the model is *meta-trained* to solve classification problems for many randomly sampled tasks (i.e., the tasks are not fixed as in multi-task learning). Then the model is *meta-tested* by classifying unseen classes after being able to adapt using *few* training images of such unseen classes.

We explore the potential to improve the meta-learning process using a more useful (i.e., non random) task sampling procedure. For example, formulating the task sampling as a multi-armed bandit problem has been shown to produce faster convergence and better generalization [16]. Similarly, Matiisen *et al.* [5] proposed a new form of curriculum learning [17] that selects new tasks based not on their performance but on their performance improvement. However, these task sampling approaches have been applied in traditional machine learning problems, such as supervised and reinforcement learning problems, which means that our proposed application of curriculum learning for task selection in meta-learning is novel, to the best of our knowledge [1].

3 Methodology

Our methodology consists of three stages (see Fig. 1). We first **meta-train** the model using different tasks (each containing relatively small training sets) to find a good initialization that is then used to **train** the model for the breast screening task (i.e., the healthy and benign versus malignant task). The **inference** is performed using previously unseen test data. Below, we define the dataset and describe each stage.

3.1 Dataset

Let the dataset be represented by $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{t}_i, b_i, d_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{v} : \Omega \rightarrow \mathbb{R}$ is the first subtraction DCE-MRI volume (Ω denotes the volume lattice), $\mathbf{t} : \Omega \rightarrow \mathbb{R}$ is the T1-weighted volume, $b \in \{\text{left}, \text{right}\}$ indicates if this is the left or right breast of the patient, $d_i \in \mathbb{N}$ denotes patient identification, and $y \in \mathcal{Y} = \{0, 1, 2\}$ is the volume label ($y_i = 2$: breast contains a malignant lesion, $y_i = 1$: breast contains at least one benign and no malignant findings, and $y_i = 0$: no findings). We divide $\mathcal{D}$ using the patient identification into the training set $\mathcal{T}$, validation set $\mathcal{V}$ and testing set $\mathcal{S}$, with no overlap between these sets.

[1] While writing the final draft of this paper, we noticed a recent approach by Sharma et al. [18]. However, they sample tasks for the problem of multi-task learning. In addition, sampling tasks is not based on the improvement of performance, but tasks where the performance is worse.

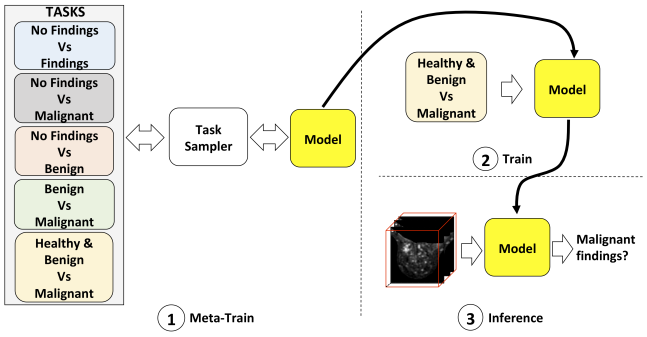


Fig. 1: The model is first meta-trained using several tasks containing relatively small training sets. The meta-trained model is then used to initialize the usual training process for breast screening (i.e., healthy and benign versus malignant). The probability of malignancy is estimated from a forward pass during the inference process.

For the meta-training phase, we use the meta-training set defined by $\{\mathcal{D}_j\}_{j=1}^{5}$ where each meta-set $\mathcal{D}_j \subseteq \mathcal{T}$ contains the relevant volumes for the classification task K_j, defined as follows: 1) K_1 classifies volumes that contain any findings (benign or malignant); 2) K_2 discriminates between volumes with no findings and malignant findings; 3) K_3 discriminates between volumes with no findings and benign findings; 4) K_4 discriminates volumes with benign findings against malignant findings; and 5) K_5 addresses breast screening, i.e. finding volumes that contain malignant findings.

3.2 Model

We **meta-train** a model across a number of tasks so that it can be quickly trained to new unseen tasks from few images, or fine-tuned to become more effective at one of the tasks used in the meta-training phase. See algorithm 1 for an overview of the methodology.

Algorithm 1 Overview of the meta-training procedure

procedure META-TRAIN($\{K_1 \dots K_5\}$, $\{\mathcal{D}_1 \dots \mathcal{D}_5\}$, model f_θ)
  Initialise model parameters θ
  for $m = 1$ *to* M **do** ▷ Meta-update Loop
    Create meta-batch $\mathcal{K}_m$ by sampling $|\mathcal{K}_m|$ tasks from $\{K_1 \dots K_5\}$
    for each task $K_j \in \mathcal{K}_m$ **do**
      Adapt model with (1) using samples from $\mathcal{D}_j$ ▷ Adaptation
    Update model parameters with (2) ▷ Meta-update

Let f_θ be the model parameterized by θ. For each meta update, the model adapts to the multiple tasks using the meta-batch set $\mathcal{K}_m$. The tasks included

in $\mathcal{K}_m$ are sampled according to one of the methods described below in Sec. 3.3. For each task $K_j \in \mathcal{K}_m$, we sample from $\mathcal{D}_j$ a training set $\mathcal{D}_j^{tr}$ and a validation set $\mathcal{D}_j^{val}$ with N^{tr} and N^{val} volumes, respectively. The model parameter θ **adaptation** is performed with the following gradient descent at time step t:

$$\theta_j'^{(t)} = \theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}, \tag{1}$$

where α denotes the adaptation learning rate, and $\mathcal{L}_{K_j}\left(f_\theta\left(\mathcal{D}_j^{tr}\right)\right)$ is the cross-entropy loss to train for the classification task K_j. Finally, given the adapted models $f_{\theta_j'^{(t)}}$ for each task $K_j \in \mathcal{K}_m$, the model parameter θ is **meta-updated** from the error on the validation volumes $\mathcal{D}_j^{val}$ of the task w.r.t. the initial parameters $\theta^{(t)}$:

$$\theta^{(t+1)} = \theta^{(t)} - \beta \sum_{K_j \in \mathcal{K}_m} \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)}{\partial \theta}, \tag{2}$$

where β denotes the meta-learning rate. In summary, the **meta-training** phase consists of updating the parameters of the model based on the error in validation images after being *adapted* to a task using few images. This is equivalent to the following optimization:

$$\min_\theta \sum_{K_j \in \mathcal{K}_m} \mathcal{L}_{K_j} f_{\theta_j'^{(t)}}(\mathcal{D}_j^{val}) = \min_\theta \sum_{K_j \in \mathcal{K}_m} \mathcal{L}_{K_j}\left(f_{\theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}}(\mathcal{D}_j^{val})\right) \tag{3}$$

The resulting model f_θ obtained after the completion of the meta-training process is then fine-tuned using the cross entropy loss for the breast screening binary classification problem. This process consists of the **training phase**, where we use the training set $\mathcal{T}$ for training and validation set $\mathcal{V}$ for model selection. The final model is tested during the **inference phase** by feeding testing volumes from $\mathcal{S}$ through the network to estimate their probability of malignancy.

3.3 Task Sampling

The sampling process to select $|\mathcal{K}|$ tasks from $\bigcup_{j=1}^{5} K_j$ (step 4 of Alg. 1) is currently based on random sampling [4]. However, we consider this to be a crucial step in that algorithm, and therefore propose four sampling methods for step 4 of Alg. 1. In particular, we study the following sampling methods: 1) **Random:** randomly sample all tasks with replacement [4]; 2) **All-task:** sample all $|\mathcal{K}| = 5$ tasks exactly once; 3) **Teacher-Student Curriculum Learning (CL)** [5]: sample tasks that can achieve a higher improvement on their performance. This is formalized by a partially observable Markov decision process (POMDP) parametrized by the *state*, which is the current parameter vector $\theta^{(t)}$; the next *action* to perform, which is the task K_j to train on; the *observation* O_{K_j}, consisting of the AUC improvement after adapting the parameters from $\theta^{(t)}$ to $\theta'^{(t)}$ for task K_j; and the *reward* R_{K_j}, which is computed from the AUC

Baseline	AUC
DenseNet [6]	0.83
MIL [7]	0.85
Multi-task [8]	0.85

Table 1: Baseline AUC for classifiers trained on breast screening.

		AUC per sampling method			
Model	$\|\mathcal{K}\|$	**Random**	**All-task**	**MAB**	**CL**
BSML	3	0.86	N/A	0.88	**0.90**
BSML	5	0.85	0.89	0.89	**0.90**
BSML-NS	4	0.85	0.88	0.87	0.89

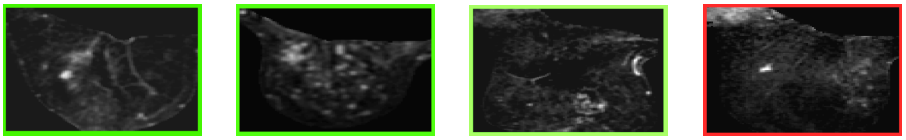
Table 2: AUC for our proposed models depending on the meta-batch size and task sampling methods and trained for breast screening.

improvement of the current observation O_{K_j} minus the AUC improvement obtained from the last time the task K_j was sampled. The goal of the sampling algorithm is to maximize the score of all tasks, which is solved based on reinforcement learning using Thompson sampling. More specifically, a buffer $\mathcal{B}_j$ stores the last B rewards for task K_j, and at sampling time, a *recent reward* is randomly chosen from each of the buffers $\mathcal{B}_j$. The next task for the meta-training is the one associated with the buffer that produced the highest absolute valued *recent reward*. This procedure chooses to lean a task until its improvement stabilizes, and then different tasks will be sampled and so on. Note that by sampling according to the absolute value, tasks where the performance is decreasing will tend to be sampled again; and 4) **Multi-armed bandit (MAB)** [16]: sample in the same way as the CL approach above, but the observation O_{K_j} is stored in the buffer instead of the reward R_{K_j}. Also, the next task is selected based on the highest valued *recent observation* (not its absolute value).

4 Experiments and Results

We assess our methodology on a breast DCE-MRI dataset containing 117 patients, divided into a training set with 45 patients, a validation with 13 and a test set with 59 patients [19, 15]. Each sample for each patient in this dataset contains T1-weighted and dynamically-contrast enhanced MRI volumes. Given the current interest in decreasing the number of scans [12, 15], only the first subtraction volume is used. Although all patients contain at least one lesion (benign or malignant, confirmed by biopsy), not all breasts contain lesions. The T1-weighted volume is only used to automatically segment and extract the left and right breasts into volumes of size $100 \times 100 \times 50$ [15] and assign separate labels to them, where the label of a breast can be "no-finding", "malignant" (if it contains at least one malignant lesion), or "benign" (if all lesions are benign). All evaluations below are based on the area under the ROC curve (AUC).

The model f_θ, implemented in 3D, is based on the DenseNet [6], which currently holds the best classification performance in several computer vision applications. The model architecture and hyper-parameters are selected based on the highest AUC for the breast screening problem in the validation set. The architecture is composed of five dense blocks of two dense layers each and is trained with a learning rate of 0.01 and a batch size of 2 volumes. For our proposed methodology (labeled as BSML), the number of meta-updates is $M = 3000$, the meta-learning rate $\beta = 0.001$, the number of training and validation volumes selected for task K_j from the meta-set $\mathcal{D}_j$ is $N^{tr} = N^{val} = 4$, the number of



(a) (b) (c) (d)

Fig. 2: Classification examples. Image (2a) shows a correct negative classification of a volume containing a benign lesion, images (2b) and (2c) show a correct positive classification of a volume containing a malignant lesion, and image (2d) shows an false negative classification of a volume containing a small malignant lesion.

gradient descent updates is 5, and the adaptation learning rate $\alpha = 0.1$. We check the influence of the meta-batch size $|\mathcal{K}| \in \{3, 5\}$. Also, we evaluate the influence of all task sampling approaches listed in Sec. 3.3. Finally, we also run experiments to check the performance of our model when the task of breast screening is not used for meta-training (BSML-NS). This means that the training process has to learn an unseen task starting from the initialization achieved in the meta-training step. In this case, we use $|\mathcal{K}| = 4$ and test the influence of the different task sampling approaches.

Our proposed model is compared against the following baselines: 1) a *DenseNet* trained for the breast screening binary task; 2) the pre-trained DenseNet (1) fine-tuned using a multiple-instance learning framework *(MIL)* [7] – this approach holds the SOTA for the breast screening problem in mammography; and 3) a DenseNet trained with a *multi-task* loss [8] using the 5 tasks defined in sec. 3.1.

Tables 1 and 2 contain the AUC for baselines and experiments detailed above. Figure 2 shows examples of the classification produced by our methodology.

5 Discussion and Conclusion

We presented a methodology to train medical image analysis systems that tries to mimic the process of training a radiologist. This is achieved by meta-training the model with several tasks containing small meta-training sets, followed by a subsequent training to solve the particular problem of interest. We established a new SOTA for the weakly supervised breast screening problem when compared to several baselines such as DenseNet [6], a multi-task trained DenseNet [8] and a DenseNet fine-tunned in a MIL framework [7]. Note that the MIL setup does not achieve a large improvement as reported in the original paper [7]. We believe that this is due to the use of DenseNet, which tends to show better classification results than Alexnet [7]. Also, it is worth mentioning that our proposed method has not shown any false positive classification in the test set.

As reflected in the experiments, the sampling of the tasks to meta-train is an important step of our proposed methodology. In particular, the CL sampling showed more accurate classification than random sampling, which yields similar results to the baselines. The MAB sampling improved over random selection, but it is still not as competitive as curriculum learning. We conjecture that sampling according to the best performance (i.e., MAB) keeps selecting more

often the tasks that produce the highest reward, while CL samples tasks with a larger margin for improvement because they can achieve a larger slope in the learning curve. Consequently, CL aims at improving the reward for ALL tasks. Also, the meta-batch size does not appear to have much influence in the results. Furthermore, the BLML-NS results in Tab. 2 show that our proposed methodology can be successfully trained for breast screening even when this task is not included in the meta-training phase. In particular, notice that the AUC is competitive, being 1 point smaller than our best result (that includes breast screening in meta-training), but between 4 and 6 points better than the baselines.

References

1. The Royal Australian and New Zealand College of Radiologists: Training in Clinical Radiology
2. Wang, X., Peng, Y., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. (2017)
3. Bar, Y., Diamant, I., et al.: Deep learning with non-medical training used for chest pathology identification. In: Medical Imaging: Computer-Aided Diagnosis. (2015)
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. (2017)
5. Matiisen, T., Oliver, A., Cohen, T., Schulman, J.: Teacher-student curriculum learning. arXiv preprint arXiv:1707.00183 (2017)
6. Huang, G., Liu, Z.: Densely connected convolutional networks. In: CVPR. (2017)
7. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: MICCAI. (2017)
8. Xue, W., Brahm, G., et al.: Full left ventricle quantification via deep multitask relationships learning. Medical image analysis (2018)
9. Smith, R.A., Andrews, K.S., et al.: Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians (2017)
10. Vreemann, S., Gubern-Merida, A., et al.: The frequency of missed breast cancers in women participating in a high-risk mri screening program. Breast cancer research and treatment (2018)
11. Gubern-Mérida, A., Martí, R., et al.: Automated localization of breast cancer in dce-mri. Medical image analysis (2015)
12. Dalmış, M.U., Vreemann, S., et al.: Fully automated detection of breast cancer in screening mri using convolutional neural networks. Journal of Medical Imaging (2018)
13. Amit, G., Hadad, O., et al.: Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning. In: MICCAI. (2017)
14. Jäger, P.F., Bickelhaupt, S., et al.: Revealing hidden potentials of the q-space signal in breast cancer. In: et al. (2017)
15. Maicas, G., Carneiro, G., et al.: Deep reinforcement learning for active breast lesion detection from dce-mri. In: MICCAI. (2017)
16. Gutiérrez, B., Peter, L., et al.: A multi-armed bandit to smartly select a training set from big medical data. In: MICCAI. (2017)
17. Bengio, Y., Louradour, J., et al.: Curriculum learning. In: ICML. (2009)
18. Sahil Sharma, Ashutosh Kumar Jha, P.S.H.B.R.: Learning to multi-task by active sampling. ICLR (2018)
19. McClymont, D., Mehnert, A., et al.: Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph. JMRI (2014)

Chapter 6

Model Agnostic Saliency for Weakly Supervised Lesion Detection from Breast DCE-MRI

The work contained in this chapter is under review as the following paper:

Gabriel Maicas, Gerard Snaauw, Andrew P. Bradley, Ian Reid, Gustavo Carneiro. Model Agnostic Saliency for Weakly Supervised Lesion Detection from Breast DCE-MRI. Under Review at IEEE International Symposium on Biomedical Imaging (ISBI), 2019.

Statement of Authorship

Title of Paper	Model Agnostic Saliency for Weakly Supervised Lesion Detection from Breast DCE-MRI
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Gabriel Maicas, Gerard Snaauw, Andrew P. Bradley, Ian Reid, Gustavo Carneiro. Model Agnostic Saliency for Weakly Supervised Lesion Detection from Breast DCE-MRI. Submitted to International Symposium on Biomedical Imaging (ISBI) , 2019.

Principal Author

Name of Principal Author (Candidate)	Gabriel Maicas Suso		
Contribution to the Paper	- Developed the idea of the paper - Coded the proposed algorithm - Designed experiments to validate the algorithm - Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	27 July 2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gerard Snaauw		
Contribution to the Paper	- Ran experiments to validate the algorithm - Refined the manuscript		
Signature		Date	25 July 2018

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	- Refined the manuscript		
Signature		Date	27 July 2018

Name of Co-Author	Ian Reid		
Contribution to the Paper	- Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	13/8/18

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Developed the idea of the paper - Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	03/08/18

MODEL AGNOSTIC SALIENCY FOR WEAKLY SUPERVISED LESION DETECTION FROM BREAST DCE-MRI

Gabriel Maicas[†] *Gerard Snaauw*[†‡] *Andrew P. Bradley*[††] *Ian Reid*[†] *Gustavo Carneiro*[† *]

[†] Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide
[‡]Faculty of Applied Sciences, Delft University of Technology
[††] Science and Engineering Faculty, Queensland University of Technology

ABSTRACT

There is a heated debate on how to interpret the decisions provided by deep learning models (DLM), where the main approaches rely on the visualization of salient regions to interpret the DLM classification process. However, these approaches generally fail to satisfy three conditions for the problem of lesion detection from medical images: 1) for images with lesions, all salient regions should represent lesions, 2) for images containing no lesions, no salient region should be produced,and 3) lesions are generally small with relatively smooth borders. We propose a new model-agnostic paradigm to interpret DLM classification decisions supported by a novel definition of saliency that incorporates the conditions above. Our model-agnostic 1-class saliency detector (MASD) is tested on weakly supervised breast lesion detection from DCE-MRI, achieving state-of-the-art detection accuracy when compared to current visualization methods.

***Index Terms*—** saliency, weakly supervised detection, model interpretability, diagnosis explanation, breast lesion localization, breast magnetic resonance imaging.

1. INTRODUCTION

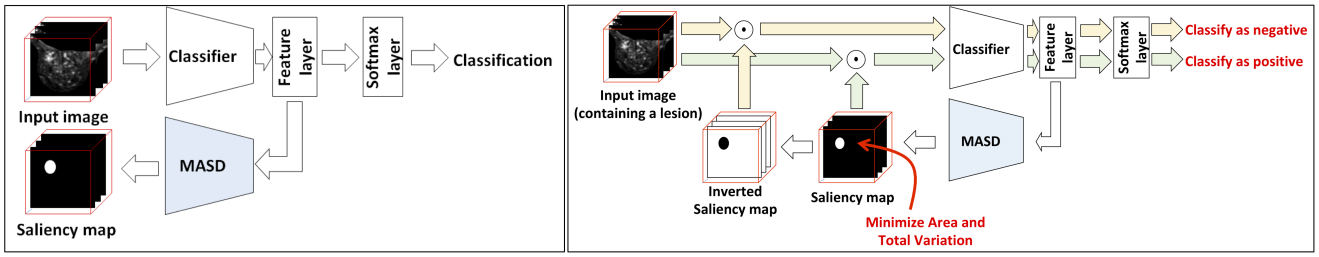
There is growing debate concerning the interpretation of classifications made by deep learning models (DLM) [1], particularly in medical diagnosis systems that can directly influence treatment decisions [2]. The clinical acceptance of DLMs depends, among other factors, on a reliable explanation of the model outcomes [3]. A popular approach that can "explain" DLM predictions relies on a salient region detector [4]. Such weakly supervised DLMs are trained to perform binary classification (negative: no lesion, positive: lesions) and produce salient regions that are assumed to highlight the regions responsible for the positive classification [4]. However, this assumption is unwarranted for two reasons. For positive classifications, there is no guarantee that salient regions represent lesions, and for negative volumes, salient regions have an unclear meaning. We argue that these issues stem from the fact that saliency is poorly defined for weakly supervised DLMs, where the training set contains images with global annotations (i.e image-level labels), but no lesion delineation.

We propose a new paradigm for explaining DLM classifications supported by a novel saliency definition for the problem of lesion detection. We define saliency as an image region that is responsible for a positive classification as opposed to previous saliency definition that was simply assumed to be active image regions during classification. Our new model-agnostic 1-class saliency detector (MASD) is explicitly trained to detect lesions from volumes that have been classified by a separate DLM (see Fig. 1 - note that MASD is independent of the DLM classifier, which is the reason why we call it model-agnostic). This goal is achieved by explicitly defining salient regions as follows: 1) they only appear when the volume in positively classified; 2) they have a small area and a relatively smooth boundary; 3) when used to mask a positively classified volume, it remains positively classified; and 4) when the inverted salient regions are used to mask a positively classified volume, it becomes negatively classified. The design of MASD is adapted from recent saliency detectors [5], [6] to become a 1-class saliency detector by incorporating our saliency definition above. We test our approach on two weakly supervised lesion detection problems from breast dynamic contrast-enhanced magnetic resonance volumes (DCE-MRI): 1) (benign and malignant) lesion detection, and 2) malignant lesion detection. We show that our results are more accurate than the ones produced by the following state-of-the-art (SOTA) saliency detectors: CAM [7], a popular approach for weakly supervised detection; Grad-CAM and Guided-GRAD-CAM [8], which are two extensions that improve upon CAM.

2. LITERATURE REVIEW

DCE-MRI is recommended as a complementary imaging modality for screening patients at high-risk for breast cancer [9]. Computer-aided detection methods have been designed [10], [11], [12], [13] and trained with strong (voxel-wise) annotations to assist radiologists. As strong annotations are time consuming and noisy, this approach does not scale to the large datasets necessary for deep learning systems. An alternative approach is to use the large weakly labeled datasets that are more readily available. The main challenge behind this approach is the requirement that the

*Supported by Australian Research Council through grants DP180103232, CE140100016 and FL130100102.



a) Inference **b) Training (for images containing lesions)**

Fig. 1: During inference (a), the weakly trained DLM produces a classification for the presence of a breast lesion – if the classification is positive, our proposed MASD outputs a saliency map that highlights the regions containing lesions. The training for volumes containing lesions in (b) is based on a loss function that penalizes: 1) saliency maps that are not smooth and contain large regions, 2) negative classification from the input volume filtered by the saliency map, and 3) positive classifications from the input volume filtered by the inverse of the saliency map. The training process for volumes that do not contain lesions penalizes the presence of any active region in the saliency map.

classifier should not only classify the case, but also highlight the region containing the lesion for the positive cases.

The medical imaging community has addressed this challenge by exploring saliency detectors [14], [15]. These approaches are based on highlighting regions of an image that are involved in the classification of each visual class by looking at the activations [7]. However, they do not work well when searching for tumors in breast DCE-MRI (given their inconsistency in shape and appearance [16]). Other approaches addressed these issues [17], [8], but none of them consider the major weakness of such saliency detection – the assumption that **salient regions represent lesions**.

Dabkowski and Gal [5] extended the work by Fong et al. [6] by guaranteeing that salient regions represented the visual class associated with the classification by explicitly defining saliency. They introduced a saliency loss function that finds a saliency mask such that: 1) the classification confidence is not perturbed when the image is masked and 2) the classification confidence is reduced when masked image regions are removed. However, there is no definition for tumors and lesions that address the characteristics of saliency in medical images. We propose a 1-class saliency detector that is able to interpret the classification of breast DCE-MRI volumes to detect lesions. Our main contribution lies on the explicit definition of saliency based on the definition of a breast lesion that then allows the definition of an appropriate loss function.

3. METHODOLOGY

3.1. Dataset

The DCE-MRI dataset is defined by $\mathcal{D} = \left\{\left(\mathbf{x}_i, \mathbf{t}_i, \{\mathbf{s}_i^{(j)}\}_{j=1}^{S_i}, b_i, y_i\right)\right\}_{i=1}^{N}$, where $\mathbf{t}_i : \Omega \rightarrow \mathbb{R}$ represents the T1-weighted volume (with Ω being the volume lattice), $\mathbf{x}_i : \Omega \rightarrow \mathbb{R}$ denotes the DCE-MRI first subtraction volume, the segmentation map $\mathbf{s}^{(j)} : \Omega \rightarrow \{0, 1\}$ is a binary volume indicating the presence or absence of lesion at each voxel for one of patient i's S_i lesions (note that we use this annotation only for testing our approach – *not for training*), $b_i \in \{\text{left}, \text{right}\}$ indicates if this is the left or right breast of the patient, and $y_i \in \mathcal{Y} = \{0, 1, 2\}$ denotes the breast label ($y_i = 2$: breast contains a malignant lesion, $y_i = 1$: breast contains at least one benign and no malignant findings, and $y_i = 0$: no findings). We consider two scenarios: 1) **lesion detection**, where labels $y \in \{1, 2\}$ are joined into the positive class, and $y = 0$ represents the negative class; and 2) **malignant lesion detection**, where labels $y \in \{0, 1\}$ are joined into the negative class, and $y = 2$ represents the positive class. We divide $\mathcal{D}$ in a patient wise manner into training, validation and testing with no overlap between sets.

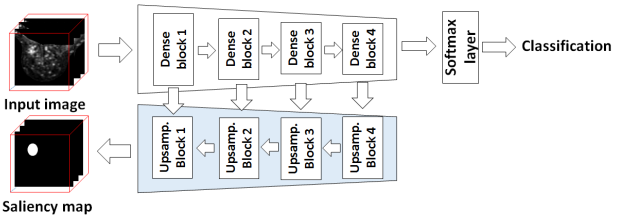


Fig. 2: MASD model diagram.

3.2. Model Agnostic 1-class Saliency Detector (MASD)

Our proposed MASD model adapts the saliency detector of Dabkowski and Gal [5] to work as a 1-class detector. The saliency detector in [5] uses a encoder-decoder structure with skip connections. The encoder is fixed and trained to produce the classification of the input image, and the decoder generates a mask of the same size of the input and is trained with a loss function that implements the saliency loss described in the last paragraph of Sec. 2. In [5], the encoder and decoder are connected through a class selector at the lowest resolution, indicating the visual classes of the salient regions.

Our model (see Fig. 2) uses a similar structure but with no class selector. The encoder consists of a four-block DenseNet [18] trained for the corresponding weakly

supervised classification problem. The decoder consists of four blocks, with each block comprising a feature map resize, a convolution layer, a batch normalization layer and ReLU activation [19] – the decoder outputs the saliency mask $\mathbf{m} : \Omega \rightarrow [0, 1]$. Our contribution lies in the removal of the class selector and modification of the loss function to become a 1-class saliency detector that produces salient regions with the lesion properties defined in Sec 1 (2^{nd} paragraph). The loss function used during **training** for each sample i is:

$$\begin{aligned}\ell_i(\mathbf{m}) =& \lambda_1 \ell_{TV}(\mathbf{m}) + \lambda_2 \ell_A(\mathbf{m}) - y_i \lambda_3 \ell_P(\mathbf{m}, \mathbf{x}_i) \\ &+ y_i \lambda_4 \ell_D(1 - \mathbf{m}, \mathbf{x}_i),\end{aligned} \tag{1}$$

where $\ell_{TV}(\mathbf{m})$ denotes the total variation of the mask and encourages detected regions to be relatively smooth, $\ell_A(\mathbf{m})$ computes the area of salient regions and penalizes large regions in the mask, $\ell_P(\mathbf{m}, \mathbf{x}_i) = \log P(y = 1|\phi(\mathbf{m}, \mathbf{x}))$ aims to maximize the positive classification when the input volume is masked with the saliency mask (i.e. $\phi(\mathbf{m}, \mathbf{x})$), and $\ell_D(1 - \mathbf{m}, \mathbf{x}_i) = P(y = 1|\phi(1 - \mathbf{m}, \mathbf{x}))$ aims to minimize the positive classification when the regions of the mask are removed from the volume. The loss function (1) is designed for the model to become a 1-class saliency detector and respond only to a single class (i.e lesions). It minimizes the number and area of regions in both negative and positive volumes and masks only positive volumes (avoiding the class selector), which are the ones containing salient regions. Note that $y_i = 0$ switches off ℓ_P and ℓ_D losses so that volumes classified as negative will not produce any salient regions.

During **inference**, the saliency mask is generated with a forward pass of the whole architecture, and we threshold the output mask $\mathbf{m}$ and form $\mathbf{m}^{(\tau)} : \Omega \rightarrow \{0, 1\}$, such that $\mathbf{m}^{(\tau)}_{ijk} = 1$, if $\mathbf{m}_{ijk} > \tau$, and $\mathbf{m}^{(\tau)}_{ijk} = 0$, otherwise.

4. EXPERIMENTS AND RESULTS

The dataset used to assess MASD contains 117 DCE-MRI and T1-weighted volumes (one DCE-MRI and one T1-weighted volume per patient) [12], [11]. The training, validation and test set contain 45, 13, and 59 patients [11] respectively. The number of lesions for each of the above sets is 57 (38 malignant (m) and 19 benign (b)), 15 (11 m and 4 b), and 69 (46 m and 23 b). The T1-weighted volume is used to automatically extract the left and right breasts into separate volumes of $100 \times 100 \times 50$ voxels [11], and the DCE-MRI volume is used for the classification and lesion detection approaches.

For the MASD model, the localization is performed only on positively classified samples, defined by a probability of being positive higher than the equal error rate (EER) of the classifier. Note that EER is computed using the validation set. We perform experiments on the two problems defined in Sec. 3.1: **lesion detection** and **malignant lesion detection**. We train a DenseNet as the base classifier for each of the problems. The parameters, estimated with the validation set using grid search, in (1) are $\lambda_1 = 0.1$, $\lambda_2 = 2$, $\lambda_3 = 0.3$, and $\lambda_4 = 2$ for lesion detection and $\lambda_1 = 0.1$, $\lambda_2 = 3$, $\lambda_3 = 1$, and $\lambda_4 = 2.5$ for malignant lesion detection problem. We evaluate our methodology in terms of true positive rate (TPR) and the number of false positive detections (FPD) per patient, which are plotted as a free response operating characteristic curve (FROC) (obtained by thresholding the mask produced by MASD at different values in $[0, 1]$). For each of the two problems above, we present results for two different scenarios: 1) **All**: the TPRs are computed using the total number of lesions (problem 1: lesion detection) and the total number of malignant lesions (problem 2: malignant lesion detection) in the dataset, and the FPDs per patient rates are computed using the total number of patients with lesions (problem 1), and the number of patients that have malignant lesions (problem 2); and 2) **C+**: the TPRs and FPDs are computed as above, but we disregard all volumes classified as negative by the DenseNet classifier. The second scenario above isolates the performance of MASD, which is the main contribution of this paper. A true positive is considered to be a detection if it has Dice ≥ 0.2 [11]. Finally, we provide a comparison with current SOTA weakly supervised region detectors: CAM [7], GRAD-CAM [8] and Guided-Grad-CAM [8].

Figure 3 shows the FROC curves for each of the problems detailed above, and Fig. 4 shows MASD results for the **lesion detection** problem in test images.

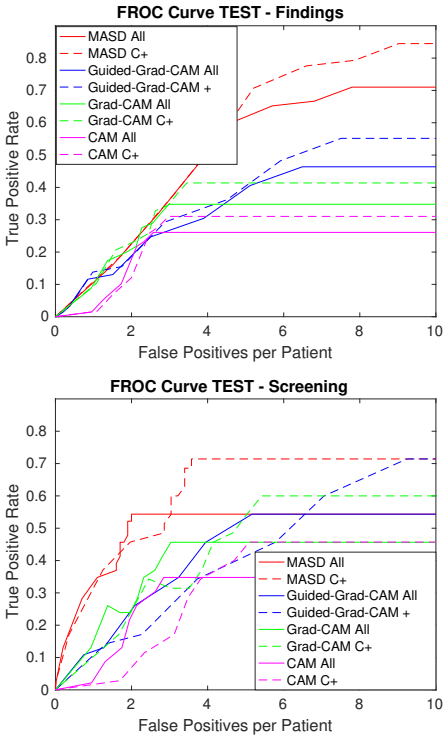
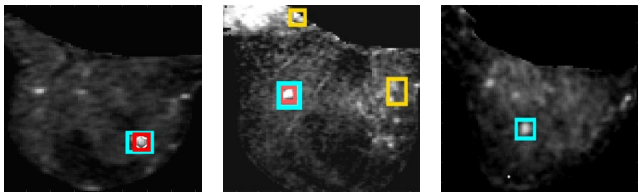


Fig. 3: FROC curves showing TPR vs FPD for **lesion detection** (Top) and **malignant lesion detection** (Bottom) for the **All** (solid curves) and **C+** (dashed) experiments.

The results in Fig. 3 show that our proposed MASD approach is more accurate for both problems of lesion



(a) (b) (c)

Fig. 4: Detections obtained with MASD for **lesion detection**. Red boxes mark true positive detections, yellow boxes false positive detections, and cyan the ground truth boxes.

and malignant lesion detection than current SOTA saliency visualization methods [7], [8], where only Guided-Grad-CAM [8] achieves a relatively competitive result. We believe that the reason behind this Guided-Grad-Cam result lies in the use of guided backpropagation that enables it to capture the fine details of the salient regions, but the mistakes made in low resolution are also carried over, increasing the FPD. Using the DenseNet classification result to filter out the negatively classified samples, all methods present slightly improved FROC curves, where the TPR curve tends to be consistently higher for the same rates of FPDs per patient. Compared to strongly supervised lesion detection approaches [11], [12], which show a TPR=0.8 for FPD=3 and TPR=0.9 for FPD=4, the MASD approach still has room for improvement – for instance, a TPR=0.8 happens only at FPD=8. We conjecture that more competitive results can be achieved with significantly larger datasets, but the empirical confirmation of this supposition is left for future work.

We tested the influence of each term in the loss function (1) and we found that the terms that produced the largest variation in our results were $\ell_D(.)$ and $\ell_A(.)$ because they allowed a significant reduction in the number and size of salient regions. The $\ell_{TV}(.)$ and $\ell_P(.)$ terms mainly influenced the lesion detection problem by increasing the FPDs.

5. CONCLUSION

We proposed MASD, a model agnostic 1-class saliency detector that can localize lesions in weakly supervised classification problems from breast DCE-MRI. By designing a loss function that explicitly incorporates terms that define a lesion (e.g. size, masked volume classification performance, absence in negative images), we demonstrate that the detected salient regions are more likely to represent the lesions that explain the decision process of deep learning classifiers. We believe that explaining the decision process of weakly-supervised classifiers will become a dominating aspect in the field because it is likely that doctors will require an explanation that can justify a DLM classification [3].

6. REFERENCES

[1] Zachary C Lipton, "The mythos of model interpretability," *arXiv*, 2016.

[2] Bryce Goodman and Seth Flaxman, "European union regulations on algorithmic decision-making and a" right to explanation"," *arXiv*, 2016.

[3] Rich Caruana, Yin Lou, and et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *KDD*, 2015.

[4] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv*, 2017.

[5] Piotr Dabkowski and Yarin Gal, "Real time image saliency for black box classifiers," in *NIPS*, 2017.

[6] Ruth C Fong and Andrea Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*, 2017.

[7] Bolei Zhou, Aditya Khosla, and et al., "Learning deep features for discriminative localization," in *CVPR*, 2016.

[8] Ramprasaath R Selvaraju, Michael Cogswell, and et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[9] Robert A Smith, Kimberly S Andrews, and et al., "Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening," *CA*, 2017.

[10] Albert Gubern-Mérida, Robert Martí, and et al., "Automated localization of breast cancer in dce-mri," *MedIA*, 2015.

[11] Gabriel Maicas, Gustavo Carneiro, and et al., "Deep reinforcement learning for active breast lesion detection from dce-mri," in *MICCAI*, 2017.

[12] Darryl McClymont, Andrew Mehnert, and et al., "Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph," *JMRI*, 2014.

[13] Guy Amit, Omer Hadad, and et al., "Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning," in *MICCAI*, 2017.

[14] Xinyang Feng, Jie Yang, and et al., "Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules," in *MICCAI*, 2017.

[15] Xiaosong Wang, Yifan Peng, and et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017.

[16] Arnau Oliver, Jordi Freixenet, and et al., "A review of automatic mass detection and segmentation in mammographic images," *MedIA*, 2010.

[17] Zhe Wang, Yanxin Yin, and et al., "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *MICCAI*, 2017.

[18] Gao Huang, Zhuang Liu, and et al., "Densely connected convolutional networks," in *CVPR*, 2017.

[19] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

Chapter 7

Pre and Post-hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI

The work contained in this chapter is under review as the following paper:

Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Pre and Post-hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI. Under Review at Medical Image Analysis.

Statement of Authorship

Title of Paper	Pre and Post hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, Gustavo Carneiro. Pre and Post hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI. Submitted to Medical Image Analysis (MedIA), 2018

Principal Author

Name of Principal Author (Candidate)	Gabriel Maicas Suso
Contribution to the Paper	- Developed the idea of the paper - Coded the proposed algorithms - Designed experiments to validate the algorithm - Wrote and refined the manuscript
Overall percentage (%)	50%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date: 27 July 2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Andrew P. Bradley
Contribution to the Paper	- Developed the idea of the paper - Refined the manuscript
Signature	Date: 27 July 2018

Name of Co-Author	Jacinto C. Nascimento
Contribution to the Paper	- Wrote and refined the manuscript
Signature	Date: 23/07/2018

Name of Co-Author	Ian Reid		
Contribution to the Paper	- Refined the manuscript		
Signature		Date	13/8/18

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Developed the idea of the paper - Supervised the development of the work - Wrote and refined the manuscript		
Signature		Date	03-08-18

Pre and Post-hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI

Gabriel Maicas[a,1,*], Andrew P. Bradley[b,1], Jacinto C. Nascimento[c,1], Ian Reid[a,2], Gustavo Carneiro[a,1]

[a] *Australian Institute for Machine Learning, The University of Adelaide, Australia*
[b] *Science and Engineering Faculty, Queensland University of Technology, Australia*
[c] *Institute for Systems and Robotics, Instituto Superior Tecnico, Portugal*

Abstract

We propose a new method for breast cancer screening from DCE-MRI based on a post-hoc approach that is trained using weakly annotated data (i.e., labels are available only at the image level without any lesion delineation). Our proposed post-hoc method automatically diagnosis the whole volume and, for positive cases, it localizes the malignant lesions that led to such diagnosis. Conversely, traditional approaches follow a pre-hoc approach that initially localises suspicious areas that are subsequently classified to establish the breast malignancy – this approach is trained using strongly annotated data (i.e., it needs a delineation and classification of all lesions in an image). Another goal of this paper is to establish the advantages and disadvantages of both approaches when applied to breast screening from DCE-MRI. Relying on experiments on a breast DCE-MRI dataset that contains scans of 117 patients, our results show that the post-hoc method is more accurate for diagnosing the whole volume per patient, achieving an AUC of 0.91, while the pre-hoc method achieves an AUC of 0.81. However, the performance for localising the malignant lesions remains challenging for the post-hoc method due to the weakly labelled dataset employed during training.

*Corresponding author
Email address: gabriel.maicas@adelaide.edu.au (Gabriel Maicas)

[1]This work was partially supported by the Australian Research Council project (DP180103232).

[2]IR acknowledges the Australian Research Council: ARC Centre for Robotic Vision (CE140100016) and Laureate Fellowship (FL130100102)

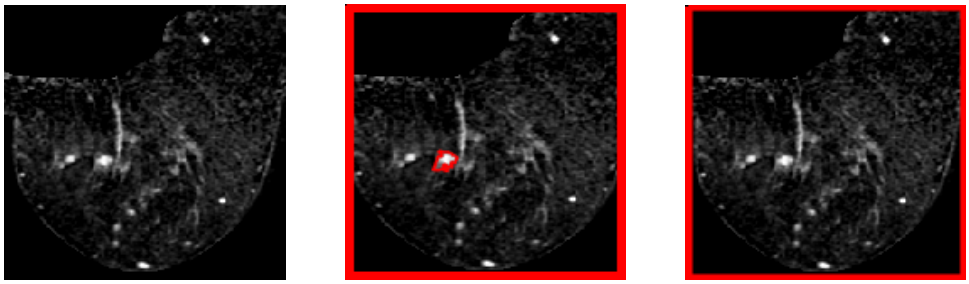
Keywords: magnetic resonance imaging, breast screening, diagnosis, meta-learning, weakly supervised learning, strongly supervised learning, model interpretation, lesion detection, deep reinforcement learning.

1. Introduction

Breast cancer is amongst the most diagnosed cancers (AIHW, 2007; Siegel et al., 2017) affecting women worldwide (DeSantis et al., 2015; Torre et al., 2015). One of the most effective ways of increasing the survival rate for this disease is based on early detection (Saadatmand et al., 2015; Welch et al., 2016). Screening programs aim to provide such early detection by diagnosing at-risk, asymptomatic patients, allowing for an early intervention and treatment. The most widely employed image modality for population-based breast screening is mammography. High risk patients are also recommended to undergo screening with dynamically contrast enhanced magnetic resonance imaging (DCE-MRI) (Mainiero et al., 2017; Smith et al., 2017). DCE-MRI is known to increase the sensitivity, compared to mammography, especially in young patients that have denser breasts (Kriege et al., 2004).

However, the diagnosis and interpretation of DCE-MRI is a challenging and time consuming task that involves the interpretation of large amounts of data (Behrens et al., 2007) and is prone to high inter-observer variability (Grimm et al., 2015; Lehman et al., 2013). Computer-aided diagnosis (CAD) systems are designed to reduce the analysis time (Gubern-Mérida et al., 2016; Wood, 2005), increase sensitivity (Vreemann et al., 2018) and specificity (Meinel et al., 2007), and serve as a second (automated) reader (Shimauchi et al., 2011). Designing such systems is challenging due to the variability in location, appearance (Levman et al., 2009), size and shape (Song et al., 2016), and the low signal-to-noise ratio (Kousi et al., 2015) of lesions. In general, such CAD systems can be categorised as pre-hoc or post-hoc, depending on how the processing stages are organised, as explained below.

Fully automated pre-hoc CAD methods for breast screening (Amit et al., 2017b; Dalmış et al., 2018; Gubern-Mérida et al., 2015) from DCE-MRI compute the confidence score of malignancy of a breast using the following two-stage sequential approach: 1) detection of suspicious lesions, and 2) classification of the detected lesions. During detection (i.e., first stage), the algorithm localises benign and malignant lesions, and possibly false positive detections, in the image, which are then classified as malignant or non-malignant in the



(a) (b) (c)

Figure 1: Example of a DCE-MRI breast image and annotation types. Image (1a) shows a slice of a breast DCE-MRI volume. Image (1b) shows the same slice with the strong annotations: lesion delineation classification as malignant. Image (1c) shows the weak annotation (i.e., whole image) of the same breast volume as malignant.

second stage. Four important challenges arise with this pre-hoc approach. Firstly, the modelling of the detector requires strong labels, i.e., precise voxel-wise annotation of lesions (see Fig. 1 for an example of different types of annotations). Strong annotation is expensive because it requires experts to label a relatively large number of training volumes; in addition, given the difficulties involved in such manual labelling process, this annotation may contain noise (this happens partly because experts are generally not trained to provide such precise annotations in regular practice). Secondly, the classifier may be trained using incorrect manually annotated lesion class labels. Such manual annotation is usually produced by biopsy analysis, but if there are benign and malignant lesions jointly present in the same breast, this analysis may not determine the correct association. Thirdly, apart from rare exceptions that need large annotated training sets (Ribli et al., 2018), pre-hoc diagnosis systems are generally trained in a two-stage process (Gubern-Mérida et al., 2015; Mcclymont, 2015). This pipeline is not the optimal way to maximise classification diagnosis performance because the final classification depends on the detection, but the detection optimality does not warrant classification optimality. Finally, the fourth challenge is that the classification accuracy is limited by the detector performance, where it is impossible for the classifier to recover from a missing lesion detection because it can not be classified.

An alternative approach that is starting to gain traction (Esteva et al., 2017; Maicas et al., 2018a; Wang et al., 2017a) reverses these stages. The first stage aims to classify the whole breast scan directly, followed by a second

stage that localizes regions in the scan that can explain the classification – for instance, if the first stage outputs a malignant diagnosis, then the second stage aims to find malignant lesions in the scan. We term this a *post-hoc* approach. This approach is of special interest for the problem of breast screening from DCE-MRI because the whole-scan diagnosis can, for example, analyse regions other than lesions that may contain relevant information for the diagnosis (Kostopoulos et al., 2017). The main advantage of these systems compared to pre-hoc systems is the possibility of using scan-level labels (referred to as weak labels in the rest of the paper). Such labels are already present in many Picture Archiving and Communication Systems (PACS) or can be automatically extracted from radiology reports (Wang et al., 2017a), eliminating almost completely the effort needed for the manual annotation described above for the pre-hoc approach. Also, the use of scan-level labels overcomes the limitations in annotations required by pre-hoc approaches. Firstly, there is no need for lesion delineation avoiding such costly process. Secondly, the incorrect labelling of lesions explained above is reduced as the most likely lesion to be malignant is biopsied and therefore the label is more likely to be correct –there is no need to associate labels with lesions). The main challenge of post-hoc systems resides in highlighting the scan regions that can justify a particular classification (e.g., in the case of a malignant classification, it is expected that the regions represent the malignant areas of the scan), given that such manual annotation is not available. This challenge is important for the deployment of post-hoc systems in clinical practice (Caruana et al., 2015).

In this paper, we propose a new post-hoc method and a systematic comparison between pre-hoc and post-hoc approaches for breast screening from DCE-MRI. We aim to answer the following research questions: 1) which approach should be chosen if the goal is to optimally classify a whole scan in terms of malignant or non-malignant findings, and 2) how accurate is the localisation of malignant lesions produced by post-hoc approaches when compared with the localisation of malignant lesions produced by pre-hoc methods. The pre-hoc system considered in this paper is based on our recent detection model (Maicas et al., 2017b) that achieves state-of-the-art (SOTA) lesion localisation, while reducing the inference time needed by traditional exhaustive search methods. For the post-hoc system, we rely on our recently proposed approach based on meta-learning (Maicas et al., 2018a) that holds the SOTA performance for the problem of breast screening from DCE-MRI. Decision interpretation is based on our recent 1-class saliency de-

tector (Maicas et al., 2018b), especially designed for the weakly supervised lesion localisation problem after performing volume diagnosis. See Fig. 2 for an overview of the pre-hoc and post-hoc pipelines.

Experiments on a breast DCE-MRI dataset containing 117 patients and 141 lesions show that the post-hoc system achieves better malignancy classification accuracy than the pre-hoc method. In terms of lesion localisation, the post-hoc approach shows less accurate performance compared to the pre-hoc system, which we infer that is mostly due to the weak annotation used in the training phase of the post-hoc method.

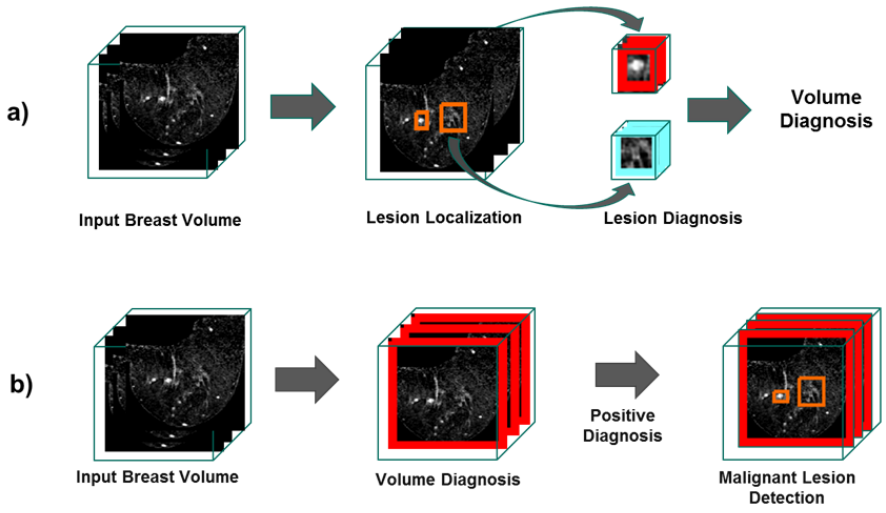


Figure 2: Pre-hoc and post-hoc approaches for breast screening. **a)** The pre-hoc approach first localises lesions in the input breast volume (e.g., detections in orange), and then these lesions are classified to decide about their malignancy (e.g., red indicates positive and blue means negative malignancy classification). Finally, the breast volume is diagnosed according to the classification scores of the lesions. **b)** The post-hoc approach first diagnoses the input breast volume (e.g., red means positive malignancy classification). If the diagnosis is positive, then malignant lesions are localised in the breast (e.g., detections in orange).

2. Literature Review

2.1. Pre-hoc Approaches

Pre-hoc approaches are assumed to contain two sequential stages: 1) detection of regions of interest (ROI) containing suspicious tissue, and 2)

classification of ROIs into malignant or not malignant (benign and/or false positive) tissue.

Traditional pre-hoc approaches for breast screening from breast DCE-MRI were based on manual (Agner et al., 2014; Gallego-Ortiz and Martel, 2015; Mus et al., 2017; Soares et al., 2013) or semi-automated (Chen et al., 2006; Dalmış et al., 2016; Meinel et al., 2007; Milenković et al., 2017; Platel et al., 2014) ROI detection. In addition, the classification in these traditional approaches was based on support vector machine (SVM), random forest, or artificial neural network models, using hand-designed features (e.g., dynamic, morphological, textural or multifractal) (Dalmış et al., 2016; Meinel et al., 2007; Milenković et al., 2017; Platel et al., 2014).

Aiming at reducing user intervention to reduce the number of ROIs (Liu et al., 2017), pre-hoc systems evolved to be fully automated. Such automated pre-hoc approaches generally employed an exhaustive search method or clustering to detect ROIs in the scan using hand-designed features (Gubern-Mérida et al., 2015; Mcclymont, 2015; Renz et al., 2012; Wang et al., 2014). The classification of ROIs into false positive, benign or malignant findings is then performed with a new set of hand-designed features extracted from the ROIs (Gubern-Mérida et al., 2015; Mcclymont, 2015; Renz et al., 2012; Wang et al., 2014). These fully automated methods generally suffer from two issues: 1) the sub-optimality of hand-designed features needed at both ROI localization and ROI classification, and 2) the high computational cost of the exhaustive search to detect ROIs.

Both limitations have been addressed after the introduction of deep learning methodologies (Krizhevsky et al., 2012) in the field of medical image analysis. Initially, feature sub-optimality was addressed either for ROI detection (Maicas et al., 2017a,b) or classification (Amit et al., 2017a,b; Rasti et al., 2017), but it was recently solved for both detection and classification (Dalmış et al., 2018). Dalmış et al. (2018) also reduced the inference time of the exhaustive search by directly computing a segmentation map from the scan using a U-net (Ronneberger et al., 2015).

Although each step of the pipeline has been individually optimized, there is no guarantee that the full pipeline is optimal in terms of classification accuracy. This was addressed with the formation of large datasets that has enabled the use of SOTA one-stage detection and classification computer vision techniques, such as Faster R-CNN (Ren et al., 2015) or Mask RCNN (He et al., 2017). The main advantage of these methods lies in the optimality of the end-to-end training, effectively merging the detection and classification

tasks (Dalmış et al., 2018). For example, Ribli et al. (2018) applied Faster R-CNN to detect tumours from mammograms and they showed that this approach is quite efficient in terms of inference time. However, Faster R-CNN generalises poorly, which means that the training set must contain a large annotated set of ROIs and, at the same time, be rich enough to comprise all possible lesion variations. Besides the need for large datasets, which are difficult to acquire for DCE-MRI breast screening, these systems suffer from the need for strong annotations (i.e., the accurate delineation of the lesions). Li et al. (2018) partially addressed this issue by developing a semi-supervised system, alleviating the need of lesion annotations. However, a large number of annotated images (880) is still required to train the system.

2.2. Post-hoc Approaches

Post-hoc systems aim to overcome the need for strong annotations by training models with only scan-level labels (i.e., weak labels). This is especially useful for the problem of breast screening, where the analysis of adjacent regions to lesions may be important (Kostopoulos et al., 2017). In addition, the classification accuracy of post-hoc systems are not constrained by the lesion detection, which is the case in pre-hoc systems.

Several post-hoc systems have been proposed (Wang et al., 2017a; Zhu et al., 2017). For instance, Wang et al. (2017a) use a deep learning model to produce classification scores from whole scans and Zhu et al. (2017) propose a deep multiple instance learning. However, these approaches still require large datasets to achieve good performance. This issue was addressed by Maicas et al. (2018a), who proposed a new meta-learning methodology to learn from a small number of annotated training images. Their work established a new SOTA classification accuracy for breast screening from DCE-MRI.

The main challenge for post-hoc models arises from the fact that they do not use manually annotated ROIs for training, which makes the ROI localisation (and delineation) a hard task. Such ROI localisation is important for explaining the classification made by the CAD system in clinical settings (e.g., for a scan classified as malignant, doctors are likely to know where the lesions are located). Solving this lesion localisation problem is a research problem that is being actively investigated in the field (Dubost et al., 2017; Feng et al., 2017; Maicas et al., 2018b; Wang et al., 2017b; Yang et al., 2017). The approach proposed by Maicas et al. (2018b) achieves SOTA detection performance by properly defining saliency for the problem of weakly

supervised lesion localisation, which assures that salient regions represent malignant lesions in the image.

However, the literature does not provide any studies comparing pre and post-hoc diagnosis approaches. The main reason for this absence of comparison among the methods described in this literature review is that such analysis is not straightforward due to (Maicas et al., 2017b): 1) the lack of publicly available datasets that can be used to compare new approaches to the current state-of-the-art, 2) the criteria to decide if an ROI is a true positive detection, and 3) the criteria to decide if lesions labelled as the challenging BIRADS=3 should be included into the benign category (Gubern-Mérida et al., 2015). In addition, not all assessments of pre-hoc fully automated methodologies consider false positives in the diagnostic stage as they only differentiate between benign and malignant (Mcclymont, 2015). We propose to compare both types of automated approaches for the problem of breast screening from breast DCE-MRI. With the use of a common dataset and well-defined criteria to satisfy the issues described above, we investigate which approach performs better for breast diagnosis and lesion localisation.

3. Methods

This section provides a formal description of the dataset in Sec. 3.1, the pre-hoc method in Sec. 3.2, and the post-hoc approach in Sec. 3.3.

3.1. Dataset

Let $\mathcal{D} = \left\{ \left(\mathbf{b}_i, \mathbf{x}_i, \mathbf{t}_i, \{\mathbf{s}_i^{(j)}\}_{j=1}^{M}, \{\mathbf{l}_i^{(j)}\}_{j=1}^{M}, \mathbf{y}_i \right)_i \right\}_{i \in \{1,\dots,|\mathcal{D}|\}, \mathbf{b}_i \in \{\text{left,right}\}}$ denote the 3D DCE-MRI dataset, where $\mathbf{b}_i \in \{\text{left}, \text{right}\}$ specifies the left or right breast of the i^{th} patient; $\mathbf{x}_i, \mathbf{t}_i : \Omega \rightarrow \mathbb{R}$ represent the first 3D DCE-MRI subtraction volume and the T1-weighted MRI volume used for preprocessing, respectively, with $\Omega \in \mathbb{R}^3$ representing the volume lattice of size $w \times h \times d$; $\mathbf{s}_i^{(j)} : \Omega \rightarrow \{0, 1\}$ is the voxelwise annotation of the j^{th} lesion present in the breast $\mathbf{b}_i$ ($\mathbf{s}_i^{(j)}(\omega) = 1$ indicates the presence of lesion in voxel $\omega \in \Omega$, and $\mathbf{s}_i^{(j)}(\omega) = 0$ denotes the absence of lesion); $\{\mathbf{l}_i^{(j)}\}_{j=1}^{M} \in \{0, 1\}$ indicates the classification of lesion j as benign or malignant, respectively; and $\mathbf{y}_i$ is a scan-level label with the following values: $\mathbf{y}_i = 0$ if there is no lesion in breast $\mathbf{b}_i$, $\mathbf{y}_i = 1$ if all the lesion(s) in breast $\mathbf{b}_i$ are benign or $\mathbf{y}_i = 2$ if there is at least one malignant lesion. The dataset is patient-wise split into train $\mathcal{T}$, validation $\mathcal{V}$ and test $\mathcal{U}$ sets, such that images of each patient only belong to one of the sets. Note that the

voxelwise lesion annotations $\{\mathbf{s}_i^{(j)}\}_{j=1}^{M}$ and $\{\mathbf{l}_i^{(j)}\}_{j=1}^{M}$ are not employed during the training of the post-hoc system – they are only used to train and test the pre-hoc system and in the quantification of the results for both systems. Finally, the motivation behind the use of the first subtraction image $\mathbf{x}$ lies in the reduction of cost and time for image acquisition and analysis (Gilbert and Selamoglu, 2018; Mango et al., 2015).

3.2. Pre-hoc Method

Our proposed pre-hoc approach is based on the following steps:

1. **Lesion detection** (Sec. 3.2.1): an attention mechanism based on deep reinforcement learning (DRL) (Mnih et al., 2015) searches for lesions using a method that analyses large portions of the breast volume and iteratively focuses the search on the appropriate regions of the input volume.
2. **Lesion diagnosis** (Sec. 3.2.2): a state-of-the-art deep learning classifier (Huang et al., 2017) analyses the lesions detected in the previous step in order to classify them as malignant or non-malignant (note that non-malignant regions are represented by benign lesions or normal tissue, i.e. false positive detections). The confidence score of malignancy for the breast volume is defined as the maximum probability of malignancy among the detected lesions.

3.2.1. Lesion Detection

We propose an attention model that is capable of reducing the inference time of previous methods for lesion detection (Gubern-Mérida et al., 2015; McClymont et al., 2014) in pre-hoc systems. This attention mechanism searches for lesions by progressively transforming relatively large initial bounding volumes (BV) (*i.e.* sub-regions of the *MRI volume*) into smaller regions containing a more focused view of potential lesions (Maicas et al., 2017b). The transformation process is guided by a policy π that indicates how to optimally change the current BV to detect a lesion. The policy is represented by a deep neural network, called deep Q-net (DQN), that receives as input an embedding vector $\mathbf{o} \in \mathbb{R}^O$ of the current BV and outputs a measurement (i.e., the Q-value (Q)), representing the optimality associated with each of the possible transformations to find a lesion. See Figure 3 for a block diagram of this process. The aim of the learning phase is to model such policy, i.e., find the optimal parameters of the DQN. The inference exploits the policy to detect the lesions present in a breast DCE-MRI volume

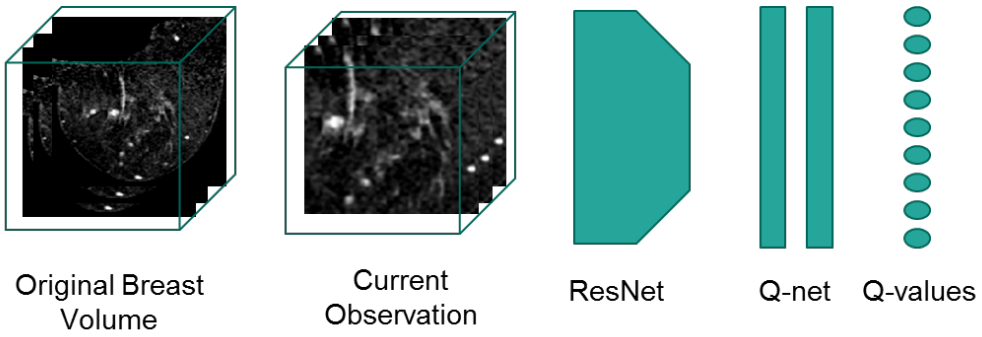


Figure 3: Overview of the proposed lesion detection method. The bounding volume of the current observation is extracted from the input breast volume and fed to the 3D ResNet to obtain the embedding of the observation. The embedding is then forwarded through the Q-net to obtain the Q-values for each of the actions.

The training process of the DQN follows that of a traditional Markov Decision Process (MDP), which models a sequence of decisions to accomplish a goal from an initial state. At every time step, the current BV, represented by the observations $\mathbf{o}$, will be transformed by an action a, yielding a reward r – this reward indicates the effectiveness of the the chosen transformation for detecting a lesion. The goal is to learn what actions should be applied to transform the current observation to another one with larger Dice coefficient measured with respect to the target lesion. In an MDP set-up, this translates into choosing the action that maximizes the expected sum of discounted future rewards (Mnih et al., 2015): $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$, where $\gamma \in (0, 1)$ is a discount factor.

Let $Q^\star(\mathbf{o}, a)$ be the optimal Action-Value Function representing the expected sum of discounted future rewards by choosing action a to transform the observation $\mathbf{o}$. The optimal Action-Value function follows the policy π, as in:

$$Q^\star(\mathbf{o}, a) = \max_{\pi} \mathbb{E}[R_t | \mathbf{o}_t = \mathbf{o}, a_t = a, \pi]. \quad (1)$$

Intuitively, $Q^\star(.)$ represents the *quality* of performing the action a given the current observation $\mathbf{o}$ to achieve the final goal. Therefore, the goal of the training process is to learn $Q^\star(.)$, which maximizes the commutative sum of expected discounted rewards.

The optimal $Q^\star(\mathbf{o}, a)$ can be computed iteratively using the *Bellman* equa-

tion and the Q-Learning algorithm (Sutton and Barto, 1998):

$$Q_{i+1}(\mathbf{o}_t, a_t) = \mathbb{E}_{\mathbf{o}_{t+1}}\Big[r_t + \gamma \max_{a_{t+1}} Q(\mathbf{o}_{t+1}, a_{t+1}) | \mathbf{o}_t, a_t\Big]. \tag{2}$$

However, since it is impractical to compute $Q(\mathbf{o}_t, a_t)$ due to the large size of the observation-action space, a DQN function approximator, represented by $Q(\mathbf{o}, a, \boldsymbol{\theta})$, can be used. The weights $\boldsymbol{\theta}$ of the DQN $Q(\mathbf{o}_t, a_t, \boldsymbol{\theta}_t)$ can be learned by minimizing the mean square error of the Bellman equation:

$$L(\boldsymbol{\theta}_t) = \mathbb{E}_{(\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1})\ \sim\ U(\mathcal{E})}\Big[\Big(\underbrace{r_t + \gamma \max_{a_{t+1}} Q(\mathbf{o}_{t+1}, a_{t+1}; \boldsymbol{\theta}_t^-)}_{\text{target}} - Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}_t)\Big)^2\Big], \tag{3}$$

where $\boldsymbol{\theta}_t$ are the parameters of the DQN at iteration t, $\boldsymbol{\theta}_t^-$ are the weights of the target network (defined below) used to compute the target value at iteration t, and $U(\mathcal{E})$ is a batch of experiences uniformly sampled from the experience replay memory $\mathcal{E}_t$ (also defined below). The target network is used to compute the target values for each update of the weights of the DQN. The architecture of this target network is the same as that of the DQN and its parameters $\boldsymbol{\theta}_t^-$ contain the weights of the DQN at a previous iteration of the optimization process. The weights $\boldsymbol{\theta}_t^-$ are updated after every iteration through the entire training set from the parameters $\boldsymbol{\theta}_t$ at the iteration $t-1$ and maintained constant between updates: $\boldsymbol{\theta}_t^- = \boldsymbol{\theta}_{t-1}$. The experience-replay memory $\mathcal{E}_t = \{e_1, ..., e_t\}$ stores previous experiences denoted by $e_t = \{\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1}\}$, where each e_t is collected at time step t by choosing the action a_t to transform from $\mathbf{o}_t$ into $\mathbf{o}_{t+1}$, yielding the reward r_t. We describe in the next paragraphs how to obtain the observations, to choose the actions and to compute the reward function.

The embedding $\mathbf{o}$ of the current BV is computed as:

$$\mathbf{o} = f_{ResNet}(\mathbf{x}(\mathbf{b}), \theta_{\mathbf{ResNet}}) \tag{4}$$

where $\mathbf{b} = [b_x, b_y, b_z, b_w, b_h, b_d] \in \mathbb{R}^6$ is a bounding volume, with the triplets (b_x, b_y, b_z) and (b_w, b_h, b_d) denoting the top-left-front and the lower-right-back corners of the bounding volume, respectively; the DCE-MRI data is represented by $\mathbf{x}$; and $f_{ResNet}(.)$ represents a 3D Residual Network (ResNet) (He et al., 2016). The training of the 3D ResNet in (4) relies on a binary loss function that differentiates between input bounding volumes with and with-

out lesions. The dataset to train this 3D ResNet is built by sampling random BVs that are labelled as positive if the Dice Coefficient with a ground truth lesion is larger than 0.6, and negative otherwise. Note that the training of the 3D ResNet with a potentially infinite number of BVs from different scales, sizes and locations allows us to obtain a rich collection of BVs without the need for a large training set.

The set $\mathcal{A} = \{l_x^+, l_x^-, l_y^+, l_y^-, l_z^+, l_z^-, s^+, s^-, w\}$ represents the actions to modify the current BV, where $\{l, s, w\}$ represent the translation, scale and trigger (to terminate the search for lesions) actions, respectively; the subscripts $\{x, y, z\}$ denote the horizontal, vertical and depth translation, and the superscripts $\{+, -\}$ represent the positive/negative translation or up/down scaling.

The reward function depends on the improvement in the lesion localisation process after selecting a specific action. For action $a \in \mathcal{A} \setminus \{w\}$, we measure the improvement in terms of the variation of the *Dice coefficient* after applying action a to transform the observation $\mathbf{o}_t$ to $\mathbf{o}_{t+1}$:

$$r(\mathbf{o}_t, a, \mathbf{o}_{t+1}) = sign(d(\mathbf{o}_{t+1}, \mathbf{s}) - d(\mathbf{o}_t, \mathbf{s})), \tag{5}$$

where $d(.)$ is the Dice coefficient between the bounding volume $\mathbf{o}$ and the ground truth $\mathbf{s}$. The intuition behind (5) is that the reward is positive if the Dice coefficient from observation $\mathbf{o}_t$ to observation $\mathbf{o}_{t+1}$ increases, and the reward is negative otherwise. The quantization in (5) avoids a deterioration of the training convergence due to small changes in $d(.)$ (Caicedo and Lazebnik, 2015).

The reward for the trigger action, $a = w$, is defined as:

$$r(\mathbf{o}_t, a, \mathbf{o}_{t+1}) = \begin{cases} +\eta & \text{if } \ d(\mathbf{o}_{t+1}, \mathbf{s}) \geq \tau_w \\ -\eta & \text{otherwise} \end{cases} \tag{6}$$

where $\eta > 1$ encourages the trigger action to finalize the search for lesions if the Dice coefficient with the ground truth $\mathbf{s}$ is larger than a pre-defined threshold τ_w.

Actions during the training process are selected according to a modified ϵ-greedy strategy to balance *exploration* and *exploitation* (Maicas et al., 2017b): with probability ϵ, a random action will be explored, and with probability $1 - \epsilon$, the action will be chosen from the current policy. During *exploration*, with probability κ, a random action is selected, and with probability $1 - \kappa$, a

random action from the actions that will produce a positive reward is selected. During *exploitation*, the action is selected according to the current policy: $a_t = \arg\max_{a_t} Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}_t)$. The training process starts with $\epsilon = 1$, which decreases linearly, transitioning from pure exploration to mostly exploitation following the current policy as the model learns to detect lesions.

During **inference**, we exploit the learned policy to detect lesions. In practice, we propose several initial bounding volumes covering different relatively large portions of the DCE-MRI volume. Each initialization is processed independently and is iteratively transformed according to the action $a_t^\star$ indicated by the optimal action-value function:

$$a_t^\star = \arg\max_{a_t} Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}^\star). \tag{7}$$

where $\boldsymbol{\theta}^\star$ represents the parameter vector of the trained DQN model learned with (3).

We define the set of detected lesions as $\mathcal{D}^{pre} = \{\mathcal{D}_i^{pre}\}_{i=1}^{|\mathcal{D}^{pre}|}$, where $\mathcal{D}_i^{pre}$ represents the i^{th} bounding volume, when the trigger action is selected to stop the inference process. If the trigger action is not selected after 20 iterations, the search for a lesion is stopped yielding no detection.

3.2.2. Lesion diagnosis

The detected lesions in $\mathcal{D}^{pre}$, formed during the lesion localization stage, are classified in terms of their malignancy. This binary classification is performed with a 3D DenseNet (Huang et al., 2017), trained using the detections from the training set to differentiate normal tissue and benign lesions (i.e., negative diagnoses) from malignant lesions (positive diagnosis). During inference, each detection $\mathcal{D}_i^{pre}$ is fed through the 3D DenseNet to obtain its probability of malignancy. Finally, the confidence score of malignancy of a breast is defined as the maximum of the malignancy probabilities computed from all the detected regions in such breast. The confidence score of malignancy for the breast volume with no detections is set to zero.

3.3. Post-hoc Method

Our proposed post-hoc approach is characterised by the following steps:

1. **Diagnosis** (Sec. 3.3.1): the classifier outputs the probability that a breast DCE-MRI volume contains a malignant lesion. Given the small training dataset, the model is first meta-trained with a teacher-student

curriculum learning strategy to learn to solve several tasks. Then, the classifier is fine-tuned to solve the breast screening diagnosis task.

2. **Lesion Localization** (Sec. 3.3.2): the detector is weakly-trained to localise malignant lesions on breast DCE-MRI volumes that have been positively classified in the diagnosis stage above. This lesion localisation process can be used to interpret the decision from the diagnosis stage.

3.3.1. Breast Volume Diagnosis

Meta-training aims to learn a model that can solve new given tasks (classification problems) as opposed to traditional classifiers that solve a specific classification problem. Traditionally, models for solving new tasks have been achieved by fine-tuning pre-trained models (Tajbakhsh et al., 2016). However, these pre-trained models are rarely available for 3D volumes and large datasets are still required. These limitations can be overcome by including a meta-training phase before training, where the model is presented with several classification tasks that need to be solved, where each task has a small training set. Eventually, the model learns to solve new tasks that contain small training sets.

As noted in our previous work (Maicas et al., 2018a), the order in which to present classification tasks during meta-training influences the ability of the model to solve new tasks. Therefore, we propose to use the teacher-student curriculum learning strategy (Matiisen et al., 2017) that has been shown to outperform other strategies (Maicas et al., 2018a).

We propose to meta-train the model to solve several related classification tasks, each containing a relatively small number of training images instead of training a classifier to distinguish volumes with any malignant findings from others containing no malignant lesions. Firstly, during the meta-training phase, our model learns to solve different tasks that are formed from our breast DCE-MRI datasets. The tasks to be presented to the model are selected via the teacher-student curriculum learning strategy and contain a small training set. Secondly, the training phase is similar to that of any traditional classifier and solves the breast screening task using the samples available from the training set. The difference in our approach lies in the employment of the meta-trained model as the initialization for the training process. As a result, when the meta-trained model is fine-tuned on the breast screening task with the small training set, it is able to efficiently and effectively classify previously unseen volumes containing malignant find-

ings (Maicas et al., 2018a). Finally, the inference phase (or breast diagnosis) consists of feeding the input volumes to the classifier to estimate the probability that they contain a malignant finding. See Figure 4 for an overview of the volume diagnosis process.

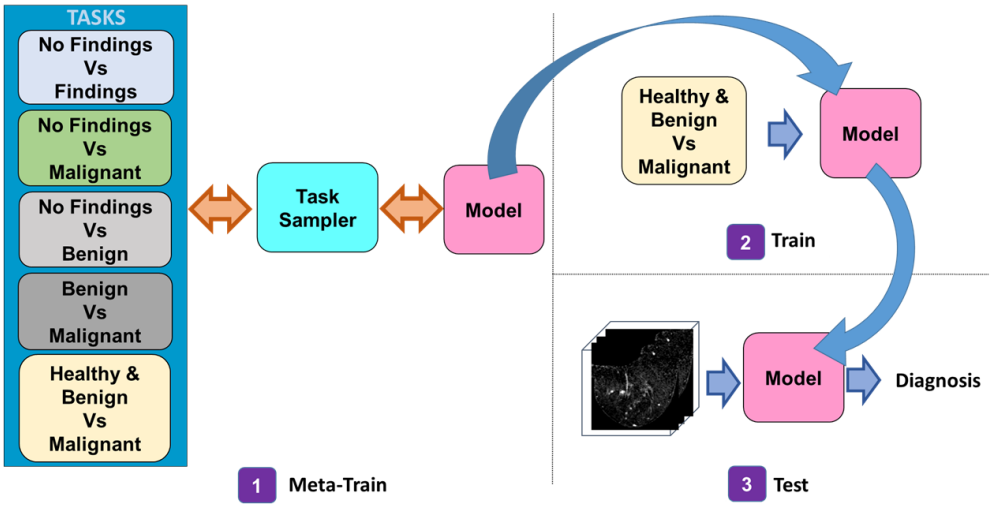


Figure 4: Volume diagnosis process. Firstly, the model is meta-trained on several related classification tasks. Secondly, the model is trained in the breast screening task. Finally, the model is tested on the breast screening task.

During **meta-training**, the model is meta-trained to solve the following five classification tasks:

1. K_1 : findings (lesions) versus no findings,
2. K_2 : malignant findings versus no findings,
3. K_3 : benign findings versus no findings,
4. K_4 : benign findings versus malignant findings,
5. K_5 : malignant findings versus no malignant findings (i.e., breast screening).

Let $K = \cup_{i=1}^{5} K_i$, where each task K_i is associated with a dataset $\mathcal{D}_i$ that contains the volumes from the training set that are relevant for the task K_i. We define the meta-training set $\mathcal{D} = \cup_{i=1}^{5} \mathcal{D}_i$.

Let the model to be meta-trained be defined by g_θ and the meta-update step be indexed by t. For each meta-update, a meta-batch $\mathcal{K}_t$ of tasks is

sampled and contains $|\mathcal{K}_t|$ tasks from K (see bellow for a description of the task sampling method). For each of the tasks $K_j \in \mathcal{K}_t$, $N = N^{tr} + N^{val}$ volume-label samples are sampled from the corresponding meta-training set $\mathcal{D}_j$ to form $\mathcal{D}_j^{tr}$. Let $\mathcal{D}_j^{tr}$ contain N^{tr} samples that will be used as training set and $\mathcal{D}_j^{val}$ contain N^{val} samples that will be used as validation set during the t^{th} meta-update for the j^{th} task.

For every task $K_j \in \mathcal{K}_t$ in the meta-batch, the model is trained with $\mathcal{D}_j^{tr}$ to adapt to the task by performing several gradient descent updates. For simplicity, the adaptation of the model with one gradient descent update is defined by:

$$\theta_j'^{(t)} = \theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(g_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}, \tag{8}$$

where $\theta^{(t)}$ are the parameters of the model at meta-iteration t, $\mathcal{L}_{K_j}(g_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right))$ is the cross-entropy loss computed from $\mathcal{D}_j^{tr}$ for task K_j, α is the learning rate for model adaptation, and $\theta_j'^{(t)}$ are the adapted parameters after performing model adaptation for task K_j.

The adapted models $g_{\theta_j'^{(t)}}$ are subsequently evaluated with the validation pairs $\mathcal{D}_j^{val}$ of the corresponding task. The loss produced by the validation set on each of the tasks is used to compute the meta-gradient associated to each task. Finally, the model parameters θ are updated using the average of the meta-gradients associated to each of the tasks in the meta-batch:

$$\theta^{(t+1)} = \theta^{(t)} - \beta \sum_{K_j \in \mathcal{K}_m} \frac{\partial \mathcal{L}_{K_j}\left(g_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)}{\partial \theta}, \tag{9}$$

where β is the meta-learning rate and $\mathcal{L}_{K_j}\left(g_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)$ is the cross entropy loss of the validation volumes in $\mathcal{D}_j^{val}$ for task K_j. This procedure is repeated for M meta-iterations, as shown in Alg. 1.

The breast screening **training** process is initialised by the meta-trained model. Using the entire training set $\mathcal{T}$, the model adapts to the breast screening task by performing several gradient descent updates, similarly to the training of a traditional deep learning classifier. We use the validation set $\mathcal{V}$ for model selection. The **inference** of the model is similar to that of any standard classifier and consists of feeding the testing volume through the network to obtain the probability of malignancy of each of the input

Algorithm 1 Overview of the meta-training procedure presented in (Maicas et al., 2018a)

procedure META-TRAIN($\{K_1 \dots K_5\}$, $\{\mathcal{D}_1 \dots \mathcal{D}_5\}$, model g_θ)
  Initialise parameters θ from g_θ
  for $t = 1$ *to* T **do**
    Sample meta-batch $\mathcal{K}_t$ by sampling $|\mathcal{K}_t|$ tasks from $\{K_1 \dots K_5\}$
    for each task $K_j \in$ meta-batch $\mathcal{K}_t$ **do**
      Adapt model using (8) with samples from $\mathcal{D}_j^{tr}$
      Evaluate adapted model using with samples from $\mathcal{D}_j^{val}$
    Meta-update model parameters with (9)

volumes. The confidence score of malignancy corresponds to the probability of malignant output by the classifier.

During the **meta-learning process**, the **task sampling** process to form a meta-batch of tasks depends on the past observed performance improvements of the model in each of the tasks. This has been shown to outperform other alternative approaches (Maicas et al., 2018a). A partially observable Markov decision process (POMDP) solved using reinforcement learning with Thompson Sampling can model such an approach. A POMDP is characterized by observations, actions, and rewards. In our set-up, we define an observation O_{K_j} as the variation in the area under the receiving operating characteristic curve (AUC) of the adapted model $\theta_j'^{(t)}$ compared to the initial AUC before the model $\theta^{(t)}$ was adapted to the task $K_j \in \mathcal{K}_t$ – in both cases, the AUC is measured using the sampled validation set $\mathcal{D}_j^{val}$. The actions correspond to sampling a particular task. The reward is defined as the difference between the current and previous observations during the last time that the task was sampled. The goal is to decide which action to apply, i.e. which task should be sampled for the next meta-training iteration. We use Thompson sampling to decide the next task to be sampled, which allows us to balance between sampling new tasks, and sampling tasks for which the improvement of performance is currently higher (similar to the exploration-exploitation dilemma in reinforcement learning) (Matiisen et al., 2017).

Let $\mathcal{B}_j$ be a buffer of recent rewards for task K_j – this buffer stores the last B rewards for this task. To perform Thompson sampling, a random recent reward $R_{\mathcal{B}_j} \in \mathcal{B}_j$ is uniformly sampled. The next task K_j to be included in the meta-batch $\mathcal{K}_t$ of iteration t is selected with $j = \arg\max_i |R_{\mathcal{B}_i}|$. This process is repeated for $|\mathcal{K}_t|$ times to form $\mathcal{K}_t$. The intuition behind this is

that for tasks where performance is increasing rapidly (i.e. yielding higher rewards) they will be sampled more frequently until mastered (i.e. the reward will tend to zero as the variation in AUC after adaptation will tend to be smaller in consecutive iterations). Then, a different task will be sampled more frequently. However, if the model reduces the performance in the previously mastered task, it will be sampled again more frequently because the absolute value of the reward will tend to be higher again.

3.3.2. Malignant Region Localization

A breast volume is diagnosed as malignant in the previous step if its confidence score of malignancy is higher than the equal error rate (EER) of the proposed classifier on the validation set. The EER as threshold is chosen to avoid any preference between sensitivity and specificity. For positively classified volumes, we aim to generate a saliency map represented by a binary mask indicating the localization of lesions that can explain the decision made by the classifier; while for negatively classified volumes, no salient region is produced. Therefore, we propose a 1-class saliency detector (Maicas et al., 2018b) that has been specifically designed to satisfy these conditions.

Our 1-class saliency detector is modelled with a weakly-supervised training process to detect salient regions in positively classified volumes, where these regions denote malignant lesions. The detector follows an encoder-decoder architecture that generates a mask $\mathbf{m} : \Omega \rightarrow [0, 1]$ of the same size as the input volume, where this mask localizes the most salient regions of the input volume that are involved in the positive classification. The encoder is the classifier from Sec. 3.3.1, which produces the diagnosis. The decoder upsamples the output from the encoder to the original resolution from the lowest resolution feature maps by concatenating four blocks of feature map resize, convolution layer, batch normalization layer and ReLU activation (Zeiler and Fergus, 2014). Skip connections are used to connect corresponding layers of the same resolution in the encoder and decoder. During training, the parameters of the encoder are fixed and the parameters of the decoder are updated using the gradient corresponding to the following loss for each volume $\mathbf{x}_i$:

$$\ell_i(\mathbf{m}) = \lambda_1 \ell_{TV}(\mathbf{m}) + \lambda_2 \ell_A(\mathbf{m}) - y_i \lambda_3 \ell_P(\mathbf{m}, \mathbf{x}_i) + y_i \lambda_4 \ell_D(1 - \mathbf{m}, \mathbf{x}_i), \quad (10)$$

where ℓ_{TV} measures the total variation of the mask forcing the boundary of salient regions to be relatively smooth, ℓ_A measures the area of the salient regions and aims to reduce the total area of regions, ℓ_P measures the con-

fidence in the classification of the input volume $\mathbf{x}_i$ masked with $\mathbf{m}$, and ℓ_D measures the confidence in the classification of the input volume $\mathbf{x}_i$ masked with the inverse of the generated mask, i.e $(1 - \mathbf{m})$.

By **training** the mask generator model with the loss function (10), there is an explicit relationship between saliency and malignant lesions (Maicas et al., 2018b). By setting $y_i = 0$ for negative volumes, they are forced to have no salient regions. For positives volumes, salient regions are forced to have the following characteristics: 1) be small and smooth, 2) when used to mask the input volume, the classification result is positive; and 3) when its inverse is used to mask the input volume, the classification result is negative. During **inference**, volumes diagnosed as positive are fed forward through the decoder to produce a mask, where each voxel has values in $[0, 1]$. This mask is thresholded at ζ to obtain the malignant lesions.

4. Experiments

In this section, we describe the dataset and experimental set-up used to assess the proposed methods for the problems of breast screening and malignant lesion detection.

4.1. Dataset

Our methods are evaluated with a dataset containing MRI scans from 117 patients. The dataset is patient-wise split into training, validation and test sets using the same split as previous approaches (Maicas et al., 2017b, 2018a,b). The training set contains scans from 45 patients, where these scans show 38 malignant lesions and 19 benign lesions – the scans also show that 29 of the patients have at least one malignant lesion while 16 only have benign lesion(s). The validation set has scans from 13 patients, with 11 malignant and 4 benign lesions – these scans show that 9 of the patients have at least one malignant lesion while 4 patients have only benign lesion(s). The test set contains scans from 59 patients, with 46 malignant and 23 benign lesions – the scans show that 37 of the patients have at least one malignant lesion while 22 have only benign lesion(s). The characterization of each lesion was confirmed with a biopsy. Every patient has at least one lesion, but not every breast contains lesions. There are 42, 13, and 58 breasts with no lesions in the training, validation and testing sets, respectively. Likewise, 18, 4, and 22 breasts contain only benign lesions (i.e. are considered "benign") and 30, 9, and 38 contain at least one malignant lesions (i.e. are considered

"malignant"). For the breast screening problem, "Malignant" breasts are considered positive while "benign" and breasts with no lesions are considered negative. The MRI dataset (McClymont et al., 2014) contains T1-weighted and two dynamic contrast enhanced (pre-contrast and first post-contrast) volumes for each patient acquired with a 1.5 Tesla GE Signa HDxt scanner. The T1-weighted anatomical volumes were acquired without fat suppression and with an acquisition matrix of 512×512. The DCE-MRI images are based on T1-weighted volumes with fat suppression, with an acquisition matrix of 360×360 and a slice thickness of 1 mm. Firstly, a pre-contrast volume was acquired before a contrast agent was injected. The first post-contrast volume was acquired after a delay of 45 seconds after the acquisition of the pre-contrast. The first subtraction volume is formed by subtracting the pre-contrast volume to the first post-contrast volume. Both T1-weighted and DCE-MRI were acquired axially.

The dataset was preprocessed using the T1-weighted volume to segment the breast region from the chest wall using Hayton's method (Hayton et al., 1997; McClymont et al., 2014). This involves removing the pectoral muscle which may produce false positive detections. In addition, The breast region was divided into left and right breasts by splitting the volume in halves, as the breast region was initially centred. Each breast volume was resized to a size of $100 \times 100 \times 50$ voxels. Note that we operate the proposed methods breast-wise.

4.2. Experimental Set-up

The aim of the experiments is to assess our pre-hoc and post-hoc approaches in terms of their performance for diagnosing malignancy and localising malignant lesions from breast DCE-MRI. Firstly, we individually evaluate the components of our proposed pre-hoc and post-hoc methods. Secondly, we compare the performance of both approaches in terms of diagnosis accuracy and malignant lesion localisation. Note that in every localisation evaluation we consider a region to be true positive if the Dice coefficient measured between a candidate region and the ground truth lesion is at least 0.2 (Maicas et al., 2017b, 2018b).

4.2.1. Pre-hoc System

The lesion detection step in the pre-hoc approach is evaluated in terms of the free response operating characteristic (FROC) curve measured patient-wisely (as in previous detection works (Gubern-Mérida et al., 2015; Maicas

et al., 2017b, 2018b)), which compares the true positive rate (TPR) against the number of false positive detections per patient (FPP). We also measure the inference time in a computer with the following configuration: Intel Core i7, 12 GB of RAM and a GPU Nvidia Titan X 12 GB. As in previous diagnosis work (Maicas et al., 2017b), the diagnosis step in the pre-hoc method is evaluated in terms of the area under the receiving operation characteristic curve (AUC), which compares true positive diagnosis rate against false positive diagnosis rate. The AUC is measured breast-wise in two different scenarios: 1) all the breasts in the testing set are considered, and 2) only breasts with at least one detected region are considered.

The **lesion detection** uses a 3D ResNet trained from scratch with random bounding volumes sampled from the training volumes. More specifically, we sample 8000 positive and 8000 negative patches that are resized to $100 \times 100 \times 50$ (the input size to the 3D ResNet). The choice of the input size of the ResNet is $100 \times 100 \times 50$ so that every lesion is visible – some tiny lesions disappear at finer resolutions. The architecture of the 3D ResNet (He et al., 2016) comprises 5 Residual Blocks (Huang et al., 2016), each of them preceded by a convolutional layer. After the last residual block, the model contains two additional convolutional layers and a fully connected (FC) layer. The embedding of the observation "$\mathbf{o}$" is the output of the second to last convolutional layer, before the FC layer and it has 2304 dimensions.

The DQN is a 2-layer multi-layer perceptron, with each layer containing 512 nodes. It outputs the Q-value for 9 actions: translation by one third of the size of the observation in the positive or negative direction on each of the dimensions (i.e. 6 actions), scaling by one sixth of the size of the observation and is applied in every dimension (i.e. 2 actions) and the trigger action. The reward value for the trigger action has been empirically defined as $\eta = 10$ if $\tau_w = 0.2$ (i.e., the Dice coefficient is at least 0.2 during the trigger action), and the discount factor is $\gamma = 0.9$. The DQN is trained with batches of 100 experiences from the experience replay memory $\mathcal{E}$, which can store 10000 experiences. We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-6}.

During training, the model is initialized with one centred large observation covering 75% of the input breast volume. During inference, the lesion detection algorithm is launched from 13 different initializations in order to increase the chances of finding all possible lesions present in a breast. In addition to the same initialization used during training, eight initializations are placed in each of the eight $50 \times 50 \times 25$ corner volumes, and four $50 \times 50 \times 25$

initializations are placed centred between the previous 8 initializations. The balance between exploration and exploitation during training is given by ϵ, which decreases linearly from $\epsilon = 1$ to $\epsilon = 0.1$ after 300 epochs, and by $\kappa = 0.5$.

Detected regions are resized to $24 \times 24 \times 12$, which is the median value of the size of all detections in the training set. The **lesion diagnosis** uses a 3D DenseNet (Huang et al., 2017) composed of three dense blocks of two dense layers each. Each dense layer comprises a batch normalization, ReLu and a convolutional layer. In the particular DenseNet implementation used in this paper, we use a compression of 0.5 and a growth rate of 6. Global average pooling of $6 \times 6 \times 3$ is applied after the last dense block and before the fully connected layer. The DenseNet is optimized with stochastic gradient descent with a learning rate of 0.01. The dataset used to train the 3D DenseNet is composed of all detections obtained from the training set. Model selection is performed using the detections from the validation set based on the breast-wise AUC for breast screening. Note that detections that correspond to malignant lesions are labelled as positive while detections that correspond to benign lesions or false positives are labelled as negative

4.2.2. Post-hoc System

The **diagnosis** step in the post-hoc approach is evaluated with the breast-wise AUC. The malignant lesion localization step in the post-hoc approach is evaluated in terms of FROC curve patient-wise under two different scenarios: 1) all the patients in the test set are considered to compute the FROC (A), and 2) only the number of patients that had at least one breast diagnosed as malignant (+), such that the performance of the 1-class saliency detector can be isolated.

The breast volume diagnosis meta-training algorithm uses as the underlying model a 3D DenseNet (Huang et al., 2017). The architecture was decided based on the optimization of a 3D DenseNet (trained with the training set $\mathcal{T}$) to achieve the best results for the breast screening task on the validation set $\mathcal{V}$ and consists of 5 dense blocks with 2 dense layers each. Each dense layer comprises a batch normalization, ReLu and convolutional layer, where compression was 0.5 and growth rate 6. No data augmentation or dropout were used since they did not improve the performance of this 3D model. For meta-training, the learning rate is $\alpha = 0.01$ and the meta-learning rate is $\beta = 0.001$. The number of gradient descent steps during adaptation is 5 and the number of meta-iterations is $M = 3000$. The meta-batch size con-

	Inference Time Per Patient
DQN (13 Initializations)	$92 \pm 21s$
MS-SL	$164 \pm 137s$
Cascade	$\mathcal{O}(60)min$

Table 1: Inference time per patient of our proposed pre-hoc detection method (DQN using 13 initializations per breast), the MS-SL (mean-shift structured learning), and the multi-scale cascade baselines.

tained $|\mathcal{K}_t| = 5$ tasks, where each task had $N^{tr} = 4$ samples for training and $N^{val} = 4$ for validation. Each buffer $\mathcal{B}_j$ stored 40 recent rewards.

The localisation of **malignant lesions** in positively classified volumes is achieved by thresholding the generated saliency map at $\zeta = 0.8$ – this threshold was decided based on the detection performance in the validation set. The parameters for training the 1-class saliency detector in (10) are: $\lambda_1 = 0.1$, $\lambda_2 = 3$, $\lambda_3 = 1$, and $\lambda_4 = 2.5$.

4.2.3. Comparison Between Pre- and Post-Hoc

Using the set-up described above for pre-hoc and post-hoc approaches, we compare the performance of both methods. We evaluate diagnosis breast-wisely and patient-wisely in terms of the area under the receiving operation characteristic curve (AUC). We also evaluate the performance of malignant lesion localisation for each approach patient-wisely using the FROC curve as in previous works (Gubern-Mérida et al., 2015; Maicas et al., 2017b, 2018b). Note that the TPR for malignant lesion localization breast-wisely is the same as patient-wisely, while the FPR breast-wisely is the same as the one for patient-wisely divided by two. For the post-hoc method, we also plot the two scenarios (A) and (+), explained above.

4.2.4. Experimental Results for the Pre-hoc System

We compare the performance of our **lesion detection** step against an improved version of exhaustive search, namely a multi-scale cascade based on deep learning features (Maicas et al., 2017a), and a mean-shift clustering method followed by structured learning (McClymont et al., 2014) (note that only one operating point is available for this approach), which is evaluated on the same dataset using a different training and testing data split. Figure 5 shows the FROC curve with the detection results and Table 1 contains the inference times per patient needed by each of the methods.

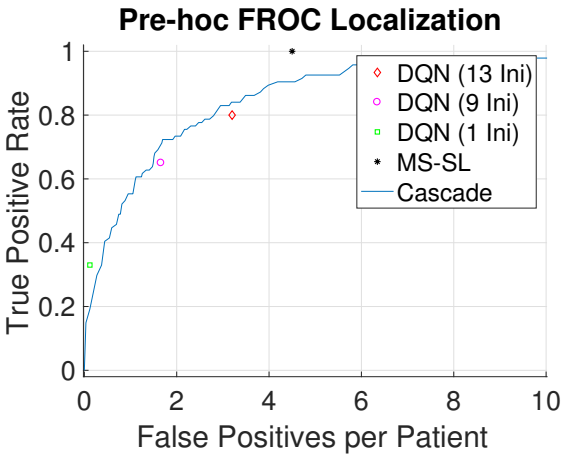


Figure 5: FROC curve per patient for for the lesion detection step of our pre-hoc method, labelled as DQN, where the information in brackets refers to the number of initialisations per breast used during the inference process. MS-SL refers to the mean-shift structured learning approach, and Cascade denotes the multi-scale cascade method based on deep learning features.

The **diagnosis** of breast volumes, based on the classification of the detected regions, achieves an AUC of 0.85, if all volumes in the test dataset are considered.

4.2.5. Experimental Results for the Post-hoc System

We evaluate the performance of our post-hoc diagnosis against three state-of-the-art classifiers. The first baseline is the 3D DenseNet (Huang et al., 2017) that has been optimized to solve the breast screening problem (as explained in Sec. 4.2.2). The second baseline is the same 3D DenseNet fine-tuned using a multiple instance learning (MIL) set-up (Zhu et al., 2017), which holds the state-of-the-art for the breast screening problem from mammography. Finally, we compare against a 3D DenseNet trained from scratch using multi-task learning (Xue et al., 2018), such that the model is jointly trained to solve all the tasks defined in Sec. 3.3.1. See Table 2 for the AUC diagnosis results.

Our 1-class saliency detector specially designed to **detect malignant lesions** in positively classified volumes is compared against the following baselines: CAM (Zhou et al., 2016), and Grad-CAM and Guided Grad-CAM (Selvaraju et al., 2017). Figure 6 shows the FROC curves for our proposed methods and baselines in each of the two scenarios (A) and (+).

Baseline	**AUC**
Meta-Training**(Ours)**	**0.90**
Multi-task (Xue et al., 2018)	0.85
MIL (Zhu et al., 2017)	0.85
DenseNet (Huang et al., 2017)	0.83

Table 2: Breast-wise AUC for diagnosis in post-hoc systems. Our proposed post-hoc diagnosis method based on meta-training is labelled as Meta-Training, while the baseline based on multiple instance learning is labelled as MIL and the one based on multi-task learning is denoted as Multi-task.

	Pre-Hoc	Post-Hoc
Breast-wise	0.85	**0.90**
Patient-wise	0.81	**0.91**

Table 3: AUC comparing the diagnosis performance between pre-hoc and post-hoc measured breast-wise and patient-wise.

4.2.6. Experimental Results for the Comparison Between Pre- and Post-Hoc

Table 3 contains the AUC for the malignancy diagnosis measured breast-wise and patient-wise for the pre-hoc and post-hoc approaches. Figure 7 shows the ROC curves used in the computation of the AUC in Table 3. Figure 8 shows the FROC curves for malignant lesion detection of pre-hoc and post-hoc ((A) and (+)) methods. Figures 9 , 10, and 11 display examples of breast diagnosis and lesion localizations obtained from the proposed pre-hoc and post-hoc methods, where both methods correctly performed diagnosis (Fig. 9), only the pre-hoc method correctly diagnosed the breast (Fig. 10), and only the post-hoc method correctly diagnosed the breast (Fig. 11).

5. Discussion

The localization step in the pre-hoc method achieves similar accuracy to the baseline methods. As shown in Figure 5, the TPR and FPR directly depends on the number of initializations used by the reinforcement learning algorithm. In addition, the performance of our localization step is very similar to the baseline based on a multi-scale cascade using exhaustive search with deep features. However, multi-scale cascade (164s) and clustering+structure learning (several hours) methods require large inference times compared to our attention model (92s) as shown in Table 1.

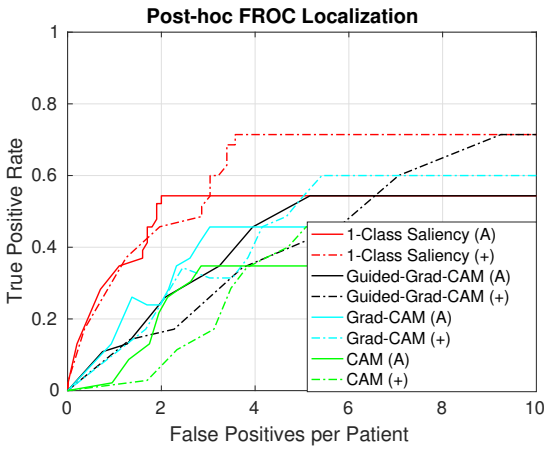


Figure 6: Patient-wise FROC curves for post-hoc malignant lesion detection, where our method is denoted as 1-Class Saliency. Baselines are denoted as CAM (Zhou et al., 2016), and Grad-CAM and Guided Grad-CAM (Selvaraju et al., 2017). For each method, we present two scenarios: (A) all the volumes in the test set are considered to compute the FROC, and (+) only positively classified volumes are considered.

The post-hoc diagnosis step improves over several baseline methods, as shown in Table 2. These baseline methods are based on a DenseNet (Huang et al., 2017), specifically optimised for the breast screening classification, and on extensions derived from multiple instance learning (Zhu et al., 2017) and multi-task learning (Xue et al., 2018). These results show that meta-training the model to solve tasks with small training sets is an important step to improve the learning of methods when only small datasets are available. Baseline approaches (Huang et al., 2017; Xue et al., 2018) only show a limited improvement over the DenseNet baseline.

The localization step in our post-hoc method benefits from our definition of saliency, as shown in Figure. 6. In contrast, baseline methods show activations that do not correlate well with the target classification. In addition, baseline methods, such as CAM (Zhou et al., 2016) and Grad-Cam (Selvaraju et al., 2017), suffer from the low resolution of the activation feature maps, despite the improvement achieved by Guided Grad-Cam (Selvaraju et al., 2017). Measuring results only on positively classified volumes ((+) curves in Figure 6) discounts the mistakes made by the diagnosis step and provides an evaluation that isolates the lesion detection ((A) curves in Figure 6). Note that there is no straightforward comparison between the localization steps

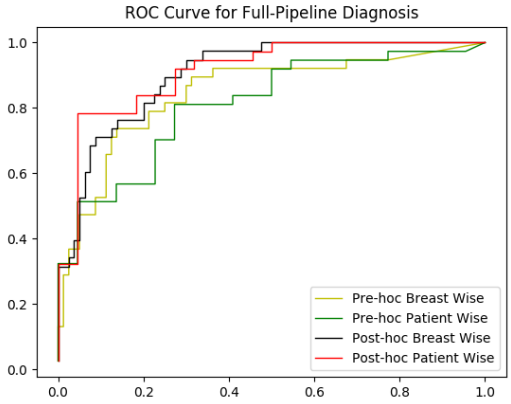


Figure 7: ROC curves for malignancy diagnosis of pre-hoc and post-hoc full pipelines measures breast and patient-wise.

in post-hoc methodologies (that only detects malignant lesions) and the localization step in pre-hoc methodologies (that detects benign and malignant lesions). Such malignant lesion detection comparison only makes sense in terms of the full pre-hoc and post-hoc pipelines, which is detailed below.

In terms of the full pipeline, we observe in Table 3 and Figure 7 that the post-hoc system has a higher classification AUC than the pre-hoc for breast screening from breast DCE-MRI. The difference between these two methods is higher when measured patient-wise compared to breast-wise. It seems reasonable to think that the reason behind such discrepancy is an effect of the missed detections in pre-hoc. In difficult (small and low contrast lesions) cases with missed detections, the confidence score of malignancy of a breast is considered 0. While this effect is smaller when the AUC is measured breast-wise (as there are 118 samples of breasts), it is larger when measured patient-wise (59 samples of patients). Furthermore, the better results of the post-hoc method suggest that the analysis of the whole image allows it to find indications for malignancy that are located in other areas of the image (Kostopoulos et al., 2017).

Regarding the localization of malignant lesions, the pre-hoc system achieves better accuracy, compared with the post-hoc. This suggests that the strong annotations used to train the pre-hoc method gives it an advantage for the localisation of lesions, when compared with the weak annotation used to train the post-hoc approach. This issue is exemplified in Figure 9 (Row 1), where

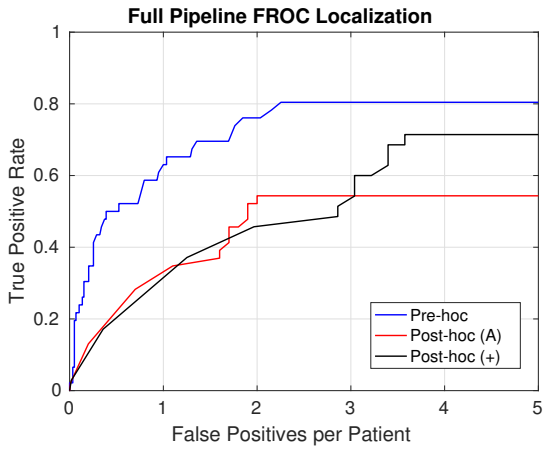


Figure 8: Patient-wise FROC curve for malignant lesion detection of pre-hoc and post-hoc full pipeline methods. For the post-hoc method, we present two scenarios the two scenarios (A) and (+)

although both approaches present a correct diagnosis, the post-hoc method yields a higher number of false positive malignant lesion detections. A similar behaviour can be seen in Figure 10 (Row 2), where the post-hoc produces an incorrect diagnosis and additionally yields two false positive detections. In addition, the detection step for the pre-hoc system is mainly designed to achieve good performance when only a small training set is available. On the contrary, the malignant lesion localization step in the post-hoc approach is not particularly focused on being able to perform well from a small dataset. This difference in design focus is likely to be influencing the detection results too. Finally, it is worth noting that the FROC for the post-hoc approach (Post-hoc (A) curve in Fig. 8) is affected by the diagnosis process. However, if we remove the effect of the diagnosis step and consider the performance of malignant lesion localization in positively classified volumes, we observe a closer performance compared to the pre-hoc method, even though the post-hoc system is trained with weak annotations.

6. Limitations and Future Work

The main limitation of our work comes from the small dataset available. In addition to a larger test set, we aim to increase our dataset to include patients where no lesions are found in order to better recreate the scenario

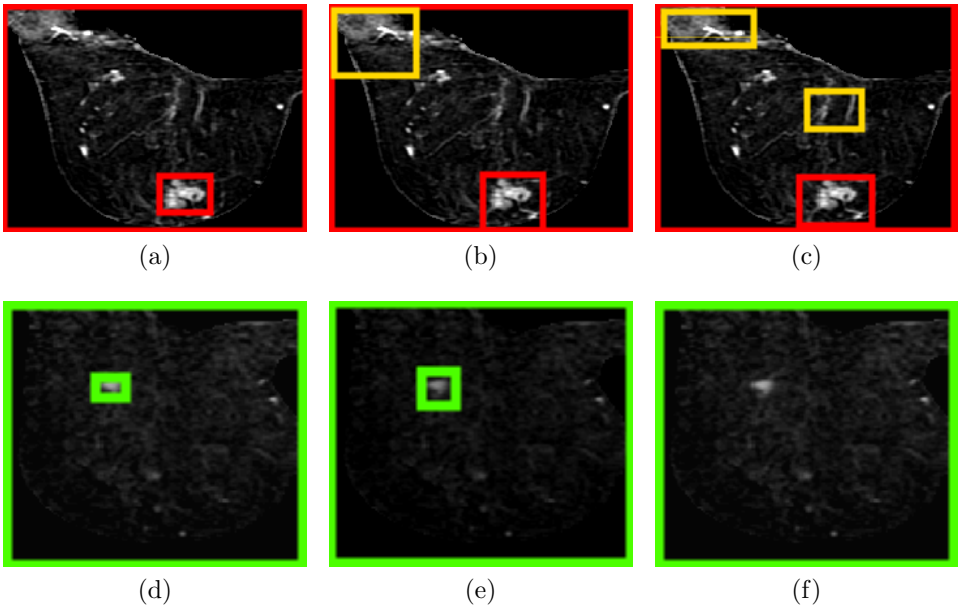


Figure 9: Example of two correct diagnosis by both pre-hoc and post-hoc full pipeline methods. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicates TP malignant detections, yellow detections indicate FP malignant detections, detections in green indicate benign lesions. **First row:** pre-hoc and post-hoc correct positive diagnosis with the malignant lesion detected. **Second row:** pre-hoc and post-hoc correct negative diagnosis where the pre-hoc method correctly classified as negative a detected benign lesion and the post-hoc method did not localize any malignant lesion

of a screening population. Ideally, this dataset will contain scanners from different vendors too. Another limitation of this work involves the lack of cross-validation experiments. This decision is justified to allow a fair comparison with other works (Maicas et al., 2017a,b, 2018a,b; McClymont et al., 2014) on the same dataset.

Future work involves the improvement of the malignant lesion localization in post-hoc methodologies by designing a new method specifically for the small training set available. We believe that the Lesion localization step in pre-hoc approaches could also be improved in terms of inference time and accuracy. As noted in (Maicas et al., 2017b), improvements in running time can be achieved by running the different initializations of the detection algorithm in parallel and by optimizing the resizing operation of the current

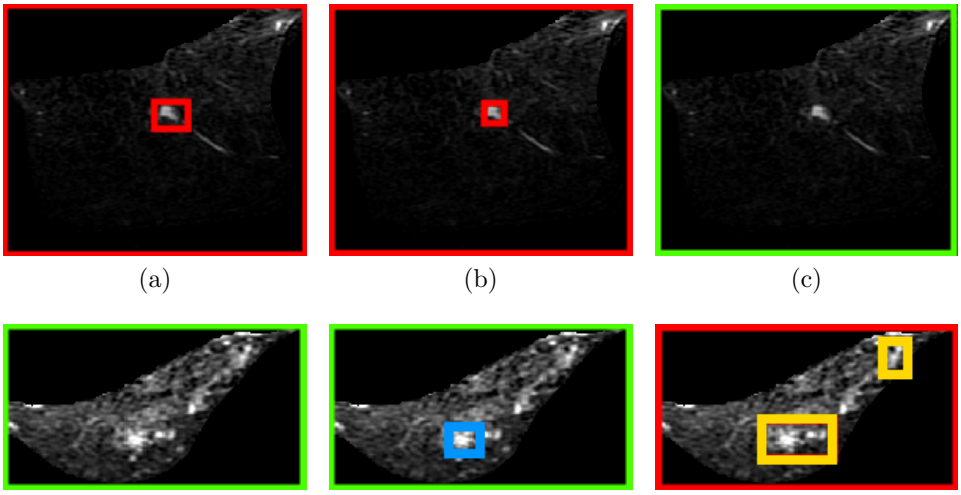


Figure 10: Example of two correct diagnosis by the pre-hoc system, but wrongly diagnosed by the post-hoc method. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicate TP malignant detections, yellow detections indicate FP malignant detections, detection in blue indicates a ROI detection correctly classified as negative (non-malignant). **First row:** correct positive diagnosis by the pre-hoc method with the malignant lesion correctly detected but incorrect non-malignant diagnosis by the post-hoc method. **Second row:** correct negative diagnosis by the pre-hoc method, but incorrect positive diagnosis by the post-hoc system – yielding the potential malignant regions in the rectangles shown in yellow.

bounding volume. In addition, the use of a U-net (Ronneberger et al., 2015) would allow the implementation of a faster segmentation map maintaining the detection accuracy. Finally, it would be interesting to design a method that could diagnose based on the combined analysis of MRI and mammography.

7. Conclusion

We introduced and compared two different approaches for breast screening from breast DCE-MRI: pre-hoc and post-hoc methods. The pre-hoc method localizes suspicious regions (benign and malignant lesions) using an attention model based on deep reinforcement learning. Detected regions were subsequently classified into malignant or non-malignant lesions using

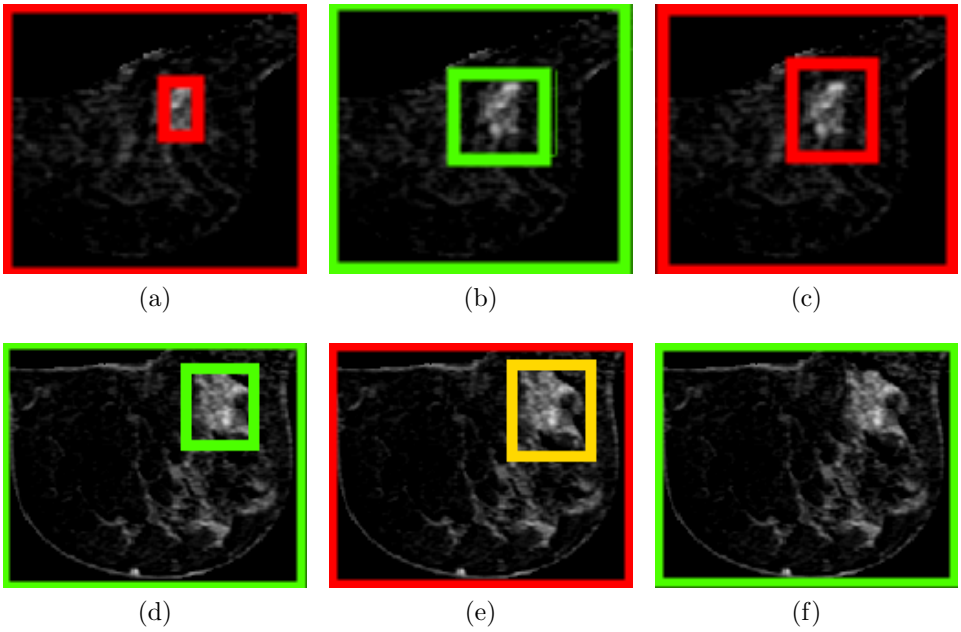


Figure 11: Example of two incorrect diagnosis by the pre-hoc, but correctly diagnosed by the post-hoc method. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicate TP malignant detections, yellow detections indicate FP malignant detections, detection in green indicates a benign ROI detection. **First Row:** the pre-hoc system incorrectly diagnoses as negative, while post-hoc system correctly diagnoses as positive and yields the malignant lesion. **Second row:** the post-hoc method correctly diagnoses as negative, but pre-hoc incorrectly diagnoses as positive due to the wrong positive classification of a detected lesion

a 3D DenseNet. The post-hoc method diagnoses a DCE-MRI breast volume using a classifier that, before being trained to solve the breast screening task, has been meta-trained to solve several breast-related tasks where only small training sets are available. Malignant regions are then localized with a 1-class saliency detector specifically designed for post-hoc systems that perform diagnosis. Results showed that the post-hoc method can achieve better performance for malignancy diagnosis, whereas the pre-hoc method could more precisely localize malignant lesions. However, this improvement of the pre-hoc detection method relies on the employment of strong annotations during the training process. On the other hand, post-hoc methods only use weak labels during the training phase and outperforms pre-hoc methods in

diagnosis, which is the main aim of a breast screening system. In conclusion, we believe that future research should focus on the development and improvement of post-hoc diagnosis methods.

References

Agner, S.C., Rosen, M.A., Englander, S., Tomaszewski, J.E., Feldman, M.D., Zhang, P., Mies, C., Schnall, M.D., Madabhushi, A., 2014. Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced mr images: a feasibility study. Radiology 272, 91–99.

AIHW, 2007. Cancer in Australia 2017. Technical Report. The Australian Institute of Health and Welfare.

Amit, G., Ben-Ari, R., Hadad, O., Monovich, E., Granot, N., Hashoul, S., 2017a. Classification of breast mri lesions using small-size training sets: comparison of deep learning approaches, in: Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 101341H.

Amit, G., Hadad, O., Alpert, S., Tlusty, T., Gur, Y., Ben-Ari, R., Hashoul, S., 2017b. Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 594–602.

Behrens, S., Laue, H., Althaus, M., Boehler, T., Kuemmerlen, B., Hahn, H.K., Peitgen, H.O., 2007. Computer assistance for mr based diagnosis of breast cancer: present and future challenges. Computerized medical imaging and graphics 31, 236–247.

Caicedo, J.C., Lazebnik, S., 2015. Active object localization with deep reinforcement learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2488–2496.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1721–1730.

Chen, W., Giger, M.L., Bick, U., 2006. A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced mr images. Academic radiology 13, 63–72.

Dalmış, M.U., Gubern-Mérida, A., Vreemann, S., Karssemeijer, N., Mann, R., Platel, B., 2016. A computer-aided diagnosis system for breast dce-mri at high spatiotemporal resolution. Medical physics 43, 84–94.

Dalmış, M.U., Vreemann, S., Kooi, T., Mann, R.M., Karssemeijer, N., Gubern-Mérida, A., 2018. Fully automated detection of breast cancer in screening mri using convolutional neural networks. Journal of Medical Imaging 5, 014502.

DeSantis, C.E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B.O., Jemal, A., 2015. International variation in female breast cancer incidence and mortality rates. Cancer Epidemiology and Prevention Biomarkers 24, 1495–1506.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W.J., Vernooij, M., De Bruijne, M., 2017. Gp-unet: Lesion detection from weak labels with a 3d regression network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 214–221.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature .

Feng, X., Yang, J., Laine, A.F., Angelini, E.D., 2017. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 568–576.

Gallego-Ortiz, C., Martel, A.L., 2015. Improving the accuracy of computer-aided diagnosis for breast mr imaging by differentiating between mass and nonmass lesions. Radiology 278, 679–688.

Gilbert, F., Selamoglu, A., 2018. Personalised screening: is this the way forward? Clinical radiology 73, 327–333.

Grimm, L.J., Anderson, A.L., Baker, J.A., Johnson, K.S., Walsh, R., Yoon, S.C., Ghate, S.V., 2015. Interobserver variability between breast imagers

using the fifth edition of the bi-rads mri lexicon. American Journal of Roentgenology 204, 1120–1124.

Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B., 2015. Automated localization of breast cancer in dce-mri. Medical image analysis 20, 265–274.

Gubern-Mérida, A., Vreemann, S., Martí, R., Melendez, J., Lardenoije, S., Mann, R.M., Karssemeijer, N., Platel, B., 2016. Automated detection of breast cancer in false-negative screening mri studies from women at increased risk. European journal of radiology 85, 472–479.

Hayton, P., Brady, M., Tarassenko, L., Moore, N., 1997. Analysis of dynamic mr breast images using a model of contrast enhancement. Medical image analysis 1, 207–224.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q., 2016. Deep networks with stochastic depth, in: European Conference on Computer Vision, Springer. pp. 646–661.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kostopoulos, S.A., Vassiou, K.G., Lavdas, E.N., Cavouras, D.A., Kalatzis, I.K., Asvestas, P.A., Arvanitis, D.L., Fezoulidis, I.V., Glotsos, D.T., 2017. Computer-based automated estimation of breast vascularity and correlation with breast cancer in dce-mri images. Magnetic resonance imaging 35, 39–45.

Kousi, E., Borri, M., Dean, J., Panek, R., Scurr, E., Leach, M.O., Schmidt, M.A., 2015. Quality assurance in mri breast screening: comparing signal-to-noise ratio in dynamic contrast-enhanced imaging protocols. Physics in Medicine & Biology 61, 37.

Kriege, M., Brekelmans, C.T., Boetes, C., Besnard, P.E., Zonderland, H.M., Obdeijn, I.M., Manoliu, R.A., Kok, T., Peterse, H., Tilanus-Linthorst, M.M., et al., 2004. Efficacy of mri and mammography for breast-cancer screening in women with a familial or genetic predisposition. New England Journal of Medicine 351, 427–437.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lehman, C.D., Blume, J.D., DeMartini, W.B., Hylton, N.M., Herman, B., Schnall, M.D., 2013. Accuracy and interpretation time of computer-aided detection among novice and experienced breast mri readers. American Journal of Roentgenology 200, W683–W689.

Levman, J.E., Causer, P., Warner, E., Martel, A.L., 2009. Effect of the enhancement threshold on the computer-aided detection of breast cancer using mri. Academic radiology 16, 1064–1069.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision, in: Proceedings of the IEEE conference on computer vision and pattern recognition.

Liu, H., Zheng, Y., Liang, D., Tang, P., Ren, F., Zhang, L., Zhao, Z., 2017. Total variation based dce-mri decomposition by separating lesion from background for time-intensity curve estimation. Medical physics 44, 2321–2331.

Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G., 2018a. Training medical image analysis systems like radiologists, in: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Maicas, G., Carneiro, G., Bradley, A.P., 2017a. Globally optimal breast mass segmentation from dce-mri using deep semantic segmentation as shape

prior, in: International Symposium on Biomedical Imaging, IEEE. pp. 305–309.

Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I., 2017b. Deep reinforcement learning for active breast lesion detection from dce-mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 665–673.

Maicas, G., Snaauw, G., Bradley, A.P., Reid, I., Carneiro, G., 2018b. Model agnostic saliency for weakly supervised lesion detection from breast dce-mri. arXiv preprint arXiv:1807.07784 .

Mainiero, M.B., Moy, L., Baron, P., Didwania, A.D., Green, E.D., Heller, S.L., Holbrook, A.I., Lee, S.J., Lewin, A.A., Lourenco, A.P., et al., 2017. Acr appropriateness criteria® breast cancer screening. Journal of the American College of Radiology 14, S383–S390.

Mango, V.L., Morris, E.A., Dershaw, D.D., Abramson, A., Fry, C., Moskowitz, C.S., Hughes, M., Kaplan, J., Jochelson, M.S., 2015. Abbreviated protocol for breast mri: are multiple sequences needed for cancer detection? European journal of radiology 84, 65–70.

Matiisen, T., Oliver, A., Cohen, T., Schulman, J., 2017. Teacher-student curriculum learning. arXiv preprint arXiv:1707.00183 .

Mcclymont, D., 2015. Computer assisted detection and characterisation of breast cancer in mri .

McClymont, D., Mehnert, A., Trakic, A., et al., 2014. Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph. JMRI 39, 795–804. URL: http://dx.doi.org/10.1002/jmri.24229, doi:10.1002/jmri.24229.

Meinel, L.A., Stolpen, A.H., Berbaum, K.S., Fajardo, L.L., Reinhardt, J.M., 2007. Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system. Journal of magnetic resonance imaging 25, 89–95.

Milenković, J., Dalmış, M.U., Žgajnar, J., Platel, B., 2017. Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast dce-mri protocol. Medical physics .

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518, 529.

Mus, R.D., Borelli, C., Bult, P., Weiland, E., Karssemeijer, N., Barentsz, J.O., Gubern-Mérida, A., Platel, B., Mann, R.M., 2017. Time to enhancement derived from ultrafast breast mri as a novel parameter to discriminate benign from malignant breast lesions. European journal of radiology 89, 90–96.

Platel, B., Mus, R., Welte, T., Karssemeijer, N., Mann, R., 2014. Automated characterization of breast lesions imaged with an ultrafast dce-mr protocol. IEEE transactions on medical imaging 33, 225–232.

Rasti, R., Teshnehlab, M., Phung, S.L., 2017. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. Pattern Recognition 72, 381–390.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Renz, D.M., Böttcher, J., Diekmann, F., Poellinger, A., Maurer, M.H., Pfeil, A., Streitparth, F., Collettini, F., Bick, U., Hamm, B., et al., 2012. Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast mri. Journal of Magnetic Resonance Imaging 35, 1077–1088.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Scientific reports 8, 4165.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Saadatmand, S., Bretveld, R., Siesling, S., Tilanus-Linthorst, M.M., 2015. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj 351, h4901.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 618–626.

Shimauchi, A., Giger, M.L., Bhooshan, N., Lan, L., Pesce, L.L., Lee, J.K., Abe, H., Newstead, G.M., 2011. Evaluation of clinical breast mr imaging performed with prototype computer-aided diagnosis breast mr imaging workstation: reader study. Radiology 258, 696–704.

Siegel, R.L., Miller, K.D., Jemal, A., 2017. Cancer statistics, 2017. CA: A Cancer Journal for Clinicians .

Smith, R.A., Andrews, K.S., Brooks, D., Fedewa, S.A., Manassaram-Baptiste, D., Saslow, D., Brawley, O.W., Wender, R.C., 2017. Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians 67, 100–121.

Soares, F., Janela, F., Pereira, M., Seabra, J., Freire, M.M., 2013. 3d lacunarity in multifractal analysis of breast tumor lesions in dynamic contrast-enhanced magnetic resonance imaging. IEEE Transactions on Image Processing 22, 4422–4435.

Song, J.L., Chen, C., Yuan, J.P., Sun, S.R., 2016. Progress in the clinical detection of heterogeneity in breast cancer. Cancer medicine 5, 3475–3488.

Sutton, R., Barto, A.G., 1998. Reinforcement learning: An introduction. volume 2. MIT press.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A., 2015. Global cancer statistics, 2012. CA: a cancer journal for clinicians 65, 87–108.

Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., Mann, R.M., 2018. The frequency of missed breast

cancers in women participating in a high-risk mri screening program. Breast Cancer Research and Treatment URL: https://doi.org/10.1007/s10549-018-4688-z, doi:10.1007/s10549-018-4688-z.

Wang, L., Harz, M., Boehler, T., Platel, B., Homeyer, A., Hahn, H.K., 2014. A robust and extendable framework towards fully automated diagnosis of nonmass lesions in breast dce-mri, in: International Symposium on Biomedical Imaging, IEEE. pp. 129–132.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017a. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., Wang, X., 2017b. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 267–275.

Welch, H.G., Prorok, P.C., OMalley, A.J., Kramer, B.S., 2016. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. New England Journal of Medicine 375, 1438–1447.

Wood, C., 2005. Computer aided detection (cad) for breast mri. Technology in cancer research & treatment 4, 49–53.

Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S., 2018. Full left ventricle quantification via deep multitask relationships learning. Medical image analysis 43, 54–65.

Yang, X., Wang, Z., Liu, C., Le, H.M., Chen, J., Cheng, K.T.T., Wang, L., 2017. Joint detection and diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 426–434.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: ECCV.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

Zhu, W., Lou, Q., Vang, Y.S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Chapter 8

Conclusion

In this thesis, we propose pre-hoc and post-hoc methods to perform diagnosis and localize malignant lesions from breast DCE-MRI. Firstly, we present two approaches that can be trained from small strongly labelled datasets and that significantly decrease the inference time of the lesion localization stage of pre-hoc systems. Secondly, we focus on the design of a post-hoc system. We introduce a novel training method from relatively small weakly labelled datasets for post-hoc diagnosis systems that accurately diagnoses breast volumes. We additionally propose a new method to detect the malignant lesions that led the post-hoc system to a positive diagnosis. Finally, we compare our proposed pre-hoc and post-hoc approaches in terms of the type of annotation required in the training phase and their performance for diagnosis and malignant lesion localization. Results show that the post-hoc system trained from a weakly labelled dataset outperforms the pre-hoc system for whole volume diagnosis. On the other hand, the pre-hoc system achieves better malignant lesion localization due to the benefits of using strongly labelled datasets during the training phase.

In this chapter, we briefly summarize the contributions of this thesis in Sec. 8.1 and describe the limitations and future work in Sec. 8.2.

8.1 Summary of Contributions

In summary, the main contributions of this thesis are:

1. A globally optimal segmentation method that can be trained from small strongly labelled datasets for the lesion localization stage of pre-hoc systems [97]. Our method is based on the global minimization of an energy functional that incorporates a shape prior from a deep learning model. Experiments show that our methodology sets the

new state-of-the-art lesion segmentation accuracy while significantly reducing the inference time when compared to multiple baselines [126, 177, 178].

2. An attention model that progressively focuses on lesions to speed up the lesion detection stage of pre-hoc systems [80]. Our method is based on deep reinforcement learning and can be trained from small strongly labelled datasets. Results show that our proposed method achieves similar lesion detection accuracy to state-of-the-art methods [75, 126] while significantly decreasing the inference time.

3. A novel method to train a classifier to perform diagnosis in post-hoc systems where only a small weakly labelled training set is available. We propose to meta-train the classifier following a curriculum learning strategy to learn to solve classification problems with small training sets. The classifier is then trained with a small training set to precisely perform diagnosis. Experiments show that our proposed method outperforms state-of-the-art training methodologies such as multi-tasking [92] and deep multiple instance learning [91].

4. A 1-class saliency detector, trained from a weakly labelled training set, that can interpret the decisions of a post-hoc diagnosis system by localizing malignant lesions in positively diagnosed volumes [99]. Our 1-class saliency detector tries to assure that detected regions in the image correspond to lesions by explicitly defining them in the training loss function. Experiments show that our method achieves the new state-of-the-art detection performance among weakly supervised post-hoc lesion detectors [167, 171].

5. Finally, we present a systematic comparison between our proposed pre-hoc and post-hoc approaches for breast screening from DCE-MRI [100]. The pre-hoc approach is built by: 1) detecting lesions with the proposed attention model based on deep reinforcement learning [80], and 2) classifying the detected regions to perform diagnosis using the state-of-the-art classifier DenseNet [90]. The post-hoc approach is built by: 1) performing diagnosis with a classifier that has been meta-trained with curriculum learning to learn to solve problems where only a small training set is available [98], and 2) for positively diagnosed volumes, malignant lesions are localized in the volume with the 1-class lesion detector [99]. Experiments show that the post-hoc approach trained with weak labels achieves better performance at volume diagnosis. However, the pre-hoc approach can better localize malignant lesions, probably due to the use of strongly annotated data during the training process.

8.2 Limitations and Future Work

Results presented in this study are limited by the small training, validation and testing sets employed to assess the proposed methods – 45, 13, and 59 patients respectively. Even though the purpose of this thesis is not to perform the clinical validation of the methods presented, the first step towards incorporating such algorithms into clinical practice should be the training and evaluation of the methods on larger datasets.

The composition of the dataset suffers from four limitations. Firstly, the dataset only contains patients with at least one lesion (benign and/or malignant) confirmed with biopsy. In the interest of validating the proposed methods, it would be desirable to include a larger proportion of patients with no lesions that can better represent a population-based breast screening setting. It is unclear how the results presented in this thesis would translate into a set-up where there is a large class imbalance between healthy and non-healthy cases. Secondly, the dataset employed in this thesis contains images from only one scanner. Aiming at deploying our proposed methods in clinical practice at different hospitals, the algorithms should be evaluated in MRI scanners from different vendors and further research should also focus on adapting and evaluating the methods on multi-scanner and multi-centre [179] datasets. Thirdly, our methods only consider the first DCE-MRI subtraction volume aiming to reduce the acquisition times and cost [180, 181]. It would be interesting to evaluate whether this choice is limiting the performance of our algorithms. Finally, our dataset does not contain DWI volumes. Since DWI-MRI is non-invasive imaging modality, we suggest that further research focuses on the extension of our methods to breast DWI-MRI datasets to assess their potentiality in comparison to breast DCE-MRI [44].

Regarding the experimental section of this thesis, note that all experiments use the same training, validation and testing set to allow a fair comparison between the proposed methods [80, 97–100] and previous work [14, 126]. However, we believe that further experiments should be performed using cross-validation to allow an analysis of the variance and stability in the performance of our methods.

Regarding the methods presented in this thesis, we believe that our proposed localization approaches for breast screening from DCE-MRI should be benchmarked against the popular U-net [132] that can quickly produce segmentation maps. Additionally, regarding the 1-class saliency detector, we believe that a better performance can be achieved if the training phase is specially designed for small training sets. We specifically suggest to meta-train the 1-class saliency detector to learn to localise lesions in tasks where only small training sets are available [98]. Note that improving the accuracy of weakly supervised lesion localization methods is a critical step to facilitate the adoption of these systems in clinical settings.

Finally, although the proposed algorithms have been evaluated in the problem of breast screening from DCE-MRI, we believe that further research should focus on adapting, extending and evaluating our methods in other types of cancer and disease diagnosis where it is not possible to build large datasets.

References

[1] Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2018. CA: A Cancer Journal for Clinicians **68**(1) (2018) 7–30

[2] Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2017. CA: A Cancer Journal for Clinicians (2017)

[3] Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: A Cancer Journal for Clinicians **66**(1) (2016) 7–30

[4] Smith, R.A., Andrews, K., Brooks, D., et al.: Cancer screening in the United States, 2016: A review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians (2016)

[5] Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: a cancer journal for clinicians **65**(2) (2015) 87–108

[6] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: a cancer journal for clinicians **61**(2) (2011) 69–90

[7] DeSantis, C.E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B.O., Jemal, A.: International variation in female breast cancer incidence and mortality rates. Cancer Epidemiology and Prevention Biomarkers **24**(10) (2015) 1495–1506

[8] Galceran, J., Ameijide, A., Carulla, M., Mateos, A., Quirós, J., Rojas, D., Alemán, A., Torrella, A., Chico, M., Vicente, M., et al.: Cancer incidence in spain, 2015. Clinical and Translational Oncology **19**(7) (2017) 799–825

[9] Australian Institute of Health and Welfare: Cancer in Australia 2017. Cancer series no.101. Cat. no. CAN 100. Canberra: AIHW . (2017)

[10] Sociedad Española de Oncología Médica: Las cifras del cáncer en españa (2018)

[11] Carter, C.L., Allen, C., Henson, D.E.: Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer **63**(1) (1989) 181–187

[12] Elkin, E.B., Hudis, C., Begg, C.B., Schrag, D.: The effect of changes in tumor size on breast carcinoma survival in the us: 1975–1999. Cancer **104**(6) (2005) 1149–1157

[13] Miller, K.D., Siegel, R.L., Lin, C.C., Mariotto, A.B., Kramer, J.L., Rowland, J.H., Stein, K.D., Alteri, R., Jemal, A.: Cancer treatment and survivorship statistics, 2016. CA: a cancer journal for clinicians **66**(4) (2016) 271–289

[14] Mcclymont, D.: Computer assisted detection and characterisation of breast cancer in mri. (2015)

[15] Saadatmand, S., Bretveld, R., Siesling, S., Tilanus-Linthorst, M.M.: Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj **351** (2015) h4901

[16] Welch, H.G., Prorok, P.C., O'Malley, A.J., Kramer, B.S.: Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. New England Journal of Medicine **375**(15) (2016) 1438–1447

[17] Ponti, A., Anttila, A., Ronco, G., Senore, C., Basu, P., Segnan, N., Tomatis, M., Žakelj, M.P., Dillner, J., Fernan, M., Elfström, K.M., Lönnberg, S., Soerjomataram, I., Sankaranaryanan, R., Vale, D.: Cancer screening in the european union. report on the implementation of the council recommendation on cancer screening. Cancer screening in the European Union (2017). Report on the implementation of the Council Recommendation on cancer screening. (2017)

[18] Nelson, H.D., Tyne, K., Naik, A., Bougatsos, C., Chan, B.K., Humphrey, L.: Screening for breast cancer: an update for the us preventive services task force. Annals of internal medicine **151**(10) (2009) 727–737

[19] Karsa, L.v., Anttila, A., Ronco, G., Ponti, A., Malila, N., Arbyn, M., Segnan, N., Castillo-Beltran, M., Boniol, M., Ferlay, J., et al.: Cancer screening in the european union. report on the implementation of the council recommendation on cancer screening. Cancer screening in the European Union. Report on the implementation of the Council Recommendation on cancer screening. (2008)

[20] Zahl, P.H., Strand, B.H., Mæhlen, J.: Incidence of breast cancer in norway and sweden during introduction of nationwide screening: prospective cohort study. Bmj **328**(7445) (2004) 921–924

[21] Ohuchi, N., Suzuki, A., Sobue, T., Kawai, M., Yamamoto, S., Zheng, Y.F., Shiono, Y.N., Saito, H., Kuriyama, S., Tohno, E., et al.: Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the japan strategic anti-cancer randomized trial (j-start): a randomised controlled trial. The Lancet **387**(10016) (2016) 341–348

[22] Lauby-Secretan, B., Scoccianti, C., Loomis, D., Benbrahim-Tallaa, L., Bouvard, V., Bianchini, F., Straif, K.: Breast-cancer screening — viewpoint of the iarc working group. New England Journal of Medicine **372**(24) (2015) 2353–2358 PMID: 26039523.

[23] Autier, P., Boniol, M.: Mammography screening: A major issue in medicine. European Journal of Cancer **90** (2018) 34–62

[24] JG, E., K, A., CD, L., SW, F.: Screening for breast cancer. JAMA **293**(10) (2005) 1245–1256

[25] Carney, P.A., Miglioretti, D.L., Yankaskas, B.C., Kerlikowske, K., Rosenberg, R., Rutter, C.M., Geller, B.M., Abraham, L.A., Taplin, S.H., Dignan, M., et al.: Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Annals of internal medicine **138**(3) (2003) 168–175

[26] Buist, D.S., Porter, P.L., Lehman, C., Taplin, S.H., White, E.: Factors contributing to mammography failure in women aged 40–49 years. Journal of the National Cancer Institute **96**(19) (2004) 1432–1440

[27] Maskarinec, G., Pagano, I., Chen, Z., Nagata, C., Gram, I.T.: Ethnic and geographic differences in mammographic density and their association with breast cancer incidence. Breast cancer research and treatment **104**(1) (2007) 47–56

[28] Kuhl, C.K., Strobel, K., Bieling, H., Leutner, C., Schild, H.H., Schrading, S.: Supplemental breast mr imaging screening of women with average risk of breast cancer. Radiology **283**(2) (2017) 361–370

[29] Freer, P.E.: Mammographic breast density: impact on breast cancer risk and implications for screening. Radiographics **35**(2) (2015) 302–315

[30] Siu, A.L.: Screening for breast cancer: US preventive services task force recommendation statement. Annals of internal medicine (2016)

[31] Saslow, D., Boetes, C., Burke, W., Harms, S., Leach, M.O., Lehman, C.D., Morris, E., Pisano, E., Schnall, M., Sener, S., et al.: American cancer society guidelines for breast screening with mri as an adjunct to mammography. CA: a cancer journal for clinicians **57**(2) (2007) 75–89

[32] Mann, R.M., Kuhl, C.K., Kinkel, K., Boetes, C.: Breast mri: guidelines from the european society of breast imaging. European radiology **18**(7) (2008) 1307–1318

[33] Smith, R.A., Andrews, K.S., Brooks, D., Fedewa, S.A., Manassaram-Baptiste, D., Saslow, D., Brawley, O.W., Wender, R.C.: Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians **67**(2) (2017) 100–121

[34] Kriege, M., Brekelmans, C.T., Boetes, C., Besnard, P.E., Zonderland, H.M., Obdeijn, I.M., Manoliu, R.A., Kok, T., Peterse, H., Tilanus-Linthorst, M.M., et al.: Efficacy of mri and mammography for breast-cancer screening in women with a familial or genetic predisposition. New England Journal of Medicine **351**(5) (2004) 427–437

[35] Sardanelli, F., Boetes, C., Borisch, B., Decker, T., Federico, M., Gilbert, F.J., Helbich, T., Heywang-Köbrunner, S.H., Kaiser, W.A., Kerin, M.J., et al.: Magnetic resonance imaging of the breast: recommendations from the eusoma working group. European journal of cancer **46**(8) (2010) 1296–1316

[36] Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., Zackrisson, S., Cardoso, F.: Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. Annals of oncology **26**(suppl_5) (2015) v8–v30

[37] Mainiero, M.B., Moy, L., Baron, P., Didwania, A.D., Green, E.D., Heller, S.L., Holbrook, A.I., Lee, S.J., Lewin, A.A., Lourenco, A.P., et al.: Acr appropriateness criteria® breast cancer screening. Journal of the American College of Radiology **14**(11) (2017) S383–S390

[38] Orel, S.G., Schnall, M.D.: Mr imaging of the breast for the detection, diagnosis, and staging of breast cancer. Radiology **220**(1) (2001) 13–30

[39] Monticciolo, D.L., Newell, M.S., Moy, L., Niell, B., Monsees, B., Sickles, E.A.: Breast cancer screening in women at higher-than-average risk: Recommendations from the acr. Journal of the American College of Radiology (2018)

[40] Kuhl, C.K., Klaschik, S., Mielcarek, P., Gieseke, J., Wardelmann, E., Schild, H.H.: Do t2-weighted pulse sequences help with the differential diagnosis of enhancing lesions in dynamic breast mri? Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **9**(2) (1999) 187–196

[41] Gillman, J., Toth, H.K., Moy, L.: The role of dynamic contrast-enhanced screening breast mri in populations at increased risk for breast cancer. Women's Health **10**(6) (2014) 609–622

[42] Clauser, P., Mann, R., Athanasiou, A., Prosch, H., Pinker, K., Dietzel, M., Helbich, T.H., Fuchsjäger, M., Camps-Herrero, J., Sardanelli, F., et al.: A survey by the european society of breast imaging on the utilisation of breast mri in clinical practice. European radiology **28**(5) (2018) 1909–1918

[43] Zhang, L., Tang, M., Min, Z., Lu, J., Lei, X., Zhang, X.: Accuracy of combined dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging for breast cancer detection: a meta-analysis. Acta Radiologica **57**(6) (2016) 651–660

[44] Partridge, S.C., Nissan, N., Rahbar, H., Kitsch, A.E., Sigmund, E.E.: Diffusion-weighted breast mri: Clinical applications and emerging techniques. Journal of Magnetic Resonance Imaging **45**(2) (2017) 337–355

[45] Mus, R.D., Borelli, C., Bult, P., Weiland, E., Karssemeijer, N., Barentsz, J.O., Gubern-Mérida, A., Platel, B., Mann, R.M.: Time to enhancement derived from ultrafast breast mri as a novel parameter to discriminate benign from malignant breast lesions. European journal of radiology **89** (2017) 90–96

[46] Gallego-Ortiz, C., Martel, A.L.: Using quantitative features extracted from t2-weighted mri to improve breast mri computer-aided diagnosis (cad). PloS one **12**(11) (2017) e0187501

[47] Gubern-Mérida, A., Vreemann, S., Martí, R., Melendez, J., Lardenoije, S., Mann, R.M., Karssemeijer, N., Platel, B.: Automated detection of breast cancer in false-negative screening mri studies from women at increased risk. European journal of radiology **85**(2) (2016) 472–479

[48] Behrens, S., Laue, H., Althaus, M., Boehler, T., Kuemmerlen, B., Hahn, H.K., Peitgen, H.O.: Computer assistance for mr based diagnosis of breast cancer: present and future challenges. Computerized medical imaging and graphics **31**(4-5) (2007) 236–247

[49] Pages, E.B., Millet, I., Hoa, D., Doyon, F.C., Taourel, P.: Undiagnosed breast cancer at mr imaging: analysis of causes. Radiology **264**(1) (2012) 40–50

[50] Yamaguchi, K., Schacht, D., Newstead, G.M., Bradbury, A.R., Verp, M.S., Olopade, O.I., Abe, H.: Breast cancer detected on an incident (second or subsequent) round of screening mri: Mri features of false-negative cases. American Journal of Roentgenology **201**(5) (2013) 1155–1163

[51] Warren, R., Hayes, C., Pointon, L., Hoff, R., Gilbert, F.J., Padhani, A.R., Rubin, C., Kaplan, G., Raza, K., Wilkinson, L., et al.: A test of performance of breast mri interpretation in a multicentre screening study. Magnetic resonance imaging **24**(7) (2006) 917–929

[52] Dorrius, M.D., Jansen-van der Weide, M.C., van Ooijen, P.M., Pijnappel, R.M., Oudkerk, M.: Computer-aided detection in breast mri: a systematic review and meta-analysis. European radiology **21**(8) (2011) 1600–1608

[53] Lehman, C.D., Blume, J.D., DeMartini, W.B., Hylton, N.M., Herman, B., Schnall, M.D.: Accuracy and interpretation time of computer-aided detection among novice and experienced breast mri readers. American Journal of Roentgenology **200**(6) (2013) W683–W689

[54] Grimm, L.J., Anderson, A.L., Baker, J.A., Johnson, K.S., Walsh, R., Yoon, S.C., Ghate, S.V.: Interobserver variability between breast imagers using the fifth edition of the bi-rads mri lexicon. American Journal of Roentgenology **204**(5) (2015) 1120–1124

[55] Grimm, L.J., Saha, A., Ghate, S.V., Kim, C., Soo, M.S., Yoon, S.C., Mazurowski, M.A.: Relationship between background parenchymal enhancement on high-risk screening mri and future breast cancer risk. Academic radiology (2018)

[56] Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Metz, C.E., Giger, M.L., Doi, K.: Improving breast cancer diagnosis with computer-aided diagnosis. Academic radiology **6**(1) (1999) 22–33

[57] Dinnes, J., Moss, S., Melia, J., Blanks, R., Song, F., Kleijnen, J.: Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. The Breast **10**(6) (2001) 455–463

[58] Gromet, M.: Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. American Journal of Roentgenology **190**(4) (2008) 854–859

[59] Jalalian, A., Mashohor, S.B., Mahmud, H.R., Saripan, M.I.B., Ramli, A.R.B., Karasfi, B.: Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clinical imaging **37**(3) (2013) 420–426

[60] Taylor-Phillips, S., Jenkinson, D., Stinton, C., Wallis, M.G., Dunn, J., Clarke, A.: Double reading in breast cancer screening: cohort evaluation in the co-ops trial. Radiology (2018) 171010

[61] Shimauchi, A., Giger, M.L., Bhooshan, N., Lan, L., Pesce, L.L., Lee, J.K., Abe, H., Newstead, G.M.: Evaluation of clinical breast mr imaging performed with prototype computer-aided diagnosis breast mr imaging workstation: reader study. Radiology **258**(3) (2011) 696–704

[62] Maxwell, A., Lim, Y., Hurley, E., Evans, D., Howell, A., Gadde, S.: False-negative mri breast screening in high-risk women. Clinical radiology **72**(3) (2017) 207–216

[63] Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., Mann, R.M.: The frequency of missed breast cancers in women participating in a high-risk mri screening program. Breast Cancer Research and Treatment (2018)

[64] Lehman, C.D., Peacock, S., DeMartini, W.B., Chen, X.: A new automated software system to evaluate breast mr examinations: improved specificity without decreased sensitivity. American Journal of Roentgenology **187**(1) (2006) 51–56

[65] Meinel, L.A., Stolpen, A.H., Berbaum, K.S., Fajardo, L.L., Reinhardt, J.M.: Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system. Journal of magnetic resonance imaging **25**(1) (2007) 89–95

[66] Meeuwis, C., van de Ven, S.M., Stapper, G., Gallardo, A.M.F., van den Bosch, M.A., Willem, P.T.M., Veldhuis, W.B.: Computer-aided detection (cad) for breast mri: evaluation of efficacy at 3.0 t. European radiology **20**(3) (2010) 522–528

[67] Renz, D.M., Böttcher, J., Diekmann, F., Poellinger, A., Maurer, M.H., Pfeil, A., Streitparth, F., Collettini, F., Bick, U., Hamm, B., et al.: Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast mri. Journal of Magnetic Resonance Imaging **35**(5) (2012) 1077–1088

[68] Agner, S.C., Rosen, M.A., Englander, S., Tomaszewski, J.E., Feldman, M.D., Zhang, P., Mies, C., Schnall, M.D., Madabhushi, A.: Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced mr images: a feasibility study. Radiology **272**(1) (2014) 91–99

[69] Wood, C.: Computer aided detection (cad) for breast mri. Technology in cancer research & treatment **4**(1) (2005) 49–53

[70] Song, J.L., Chen, C., Yuan, J.P., Sun, S.R.: Progress in the clinical detection of heterogeneity in breast cancer. Cancer medicine **5**(12) (2016) 3475–3488

[71] Mann, R.M., Balleyguier, C., Baltzer, P.A., Bick, U., Colin, C., Cornford, E., Evans, A., Fallenberg, E., Forrai, G., Fuchsjäger, M.H., et al.: Breast mri: Eusobi recommendations for women's information. European radiology **25**(12) (2015) 3669–3678

[72] Jansen, S.A., Fan, X., Karczmar, G.S., Abe, H., Schmidt, R.A., Giger, M., Newstead, G.M.: Dcemri of breast lesions: Is kinetic analysis equally effective for both mass and nonmass-like enhancement? Medical physics **35**(7Part1) (2008) 3102–3109

[73] Levman, J.E., Causer, P., Warner, E., Martel, A.L.: Effect of the enhancement threshold on the computer-aided detection of breast cancer using mri. Academic radiology **16**(9) (2009) 1064–1069

[74] Kousi, E., Borri, M., Dean, J., Panek, R., Scurr, E., Leach, M.O., Schmidt, M.A.: Quality assurance in mri breast screening: comparing signal-to-noise ratio in dynamic contrast-enhanced imaging protocols. Physics in Medicine & Biology **61**(1) (2015) 37

[75] Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B.: Automated localization of breast cancer in dce-mri. Medical image analysis **20**(1) (2015) 265–274

[76] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4) (1989) 541–551

[77] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324

[78] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105

[79] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)

[80] Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I.: Deep reinforcement learning for active breast lesion detection from dce-mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 665–673

[81] Dalmış, M.U., Vreemann, S., Kooi, T., Mann, R.M., Karssemeijer, N., Gubern-Mérida, A.: Fully automated detection of breast cancer in screening mri using convolutional neural networks. Journal of Medical Imaging **5**(1) (2018) 014502

[82] Amit, G., Hadad, O., Alpert, S., Tlusty, T., Gur, Y., Ben-Ari, R., Hashoul, S.: Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 594–602

[83] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99

[84] Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 6517–6525

[85] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2980–2988

[86] Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Scientific reports **8**(1) (2018) 4165

[87] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 3462–3471

[88] Yates, E., Yates, L., Harvey, H.: Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. Clinical radiology (2018)

[89] Kostopoulos, S.A., Vassiou, K.G., Lavdas, E.N., Cavouras, D.A., Kalatzis, I.K., Asvestas, P.A., Arvanitis, D.L., Fezoulidis, I.V., Glotsos, D.T.: Computer-based automated estimation of breast vascularity and correlation with breast cancer in dce-mri images. Magnetic resonance imaging **35** (2017) 39–45

[90] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4700–4708

[91] Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 603–611

[92] Xue, W., Brahm, G., et al.: Full left ventricle quantification via deep multitask relationships learning. Medical image analysis (2018)

[93] Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annual review of biomedical engineering **15** (2013) 327–357

[94] Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM (2016) 1135–1144

[95] Chartrand, G., Cheng, P.M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C.J., Kadoury, S., Tang, A.: Deep learning: A primer for radiologists. RadioGraphics **37**(7) (2017) 2113–2131

[96] Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)

[97] Maicas, G., Carneiro, G., Bradley, A.P.: Globally optimal breast mass segmentation from dce-mri using deep semantic segmentation as shape prior. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE (2017) 305–309

[98] Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Training medical image analysis systems like radiologists. In: Submitted to International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2018)

[99] Maicas, G., Snaauw, G., Bradley, A.P., Reid, I., Carneiro, G.: Model agnostic saliency for weakly supervised lesion detection from breast dce-mri. arXiv preprint arXiv:1807.07784 (2018)

[100] Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Pre and post-hoc diagnosis and interpretation of malignancy from breast dce-mri. arXiv preprint arXiv:1809.09404 (2018)

[101] Giger, M.L., Chan, H.P., Boone, J.: Anniversary paper: History and status of cad and quantitative image analysis: the role of medical physics and aapm. Medical physics **35**(12) (2008) 5799–5820

[102] Liu, Y.H., Xu, L., Liu, L.H., Liu, X.S., Hou, Z.Y., Hou, D.L., Chen, Z.Q., Li, W.W., Huang, Y.: 3.0 t mr-cad: Clinical value in diagnosis of breast tumor compared with conventional mri. Journal of Cancer **5**(7) (2014) 585

[103] Wu, H.: Automatic Computer Aided Diagnosis of Breast Cancer in Dynamic Contrast Enhanced Magnetic Resonance Images. PhD thesis (2016)

[104] Jansen, S.A., Shimauchi, A., Zak, L., Fan, X., Karczmar, G.S., Newstead, G.M.: The diverse pathology and kinetics of mass, nonmass, and focus enhancement on mr imaging of the breast. Journal of magnetic resonance imaging **33**(6) (2011) 1382–1389

[105] Soares, F., Janela, F., Pereira, M., Seabra, J., Freire, M.M.: 3d lacunarity in multifractal analysis of breast tumor lesions in dynamic contrast-enhanced magnetic resonance imaging. IEEE Transactions on Image Processing **22**(11) (2013) 4422–4435

[106] Chen, W., Giger, M.L., Bick, U.: A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced mr images1. Academic radiology **13**(1) (2006) 63–72

[107] Chen, W., Giger, M.L., Li, H., Bick, U., Newstead, G.M.: Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. Magnetic resonance in medicine **58**(3) (2007) 562–571

[108] Chang, Y.C., Huang, Y.H., Huang, C.S., Chang, P.K., Chen, J.H., Chang, R.F.: Classification of breast mass lesions using model-based analysis of the characteristic kinetic curve derived from fuzzy c-means clustering. Magnetic resonance imaging **30**(3) (2012) 312–322

[109] Huang, Y.H., Chang, Y.C., Huang, C.S., Wu, T.J., Chen, J.H., Chang, R.F.: Computer-aided diagnosis of mass-like lesion in breast mri: differential analysis of the 3-d morphology between benign and malignant tumors. Computer methods and programs in biomedicine **112**(3) (2013) 508–517

[110] Gallego-Ortiz, C., Martel, A.L.: Classification of breast lesions presenting as mass and non-mass lesions. In: Medical Imaging 2014: Computer-Aided Diagnosis. Volume 9035., International Society for Optics and Photonics (2014) 90351Z

[111] Wang, T.C., Huang, Y.H., Huang, C.S., Chen, J.H., Huang, G.Y., Chang, Y.C., Chang, R.F.: Computer-aided diagnosis of breast dce-mri using pharmacokinetic model and 3-d morphology analysis. Magnetic resonance imaging **32**(3) (2014) 197–205

[112] Platel, B., Mus, R., Welte, T., Karssemeijer, N., Mann, R.: Automated characterization of breast lesions imaged with an ultrafast dce-mr protocol. IEEE transactions on medical imaging **33**(2) (2014) 225–232

[113] Yang, Q., Li, L., Zhang, J., Shao, G., Zheng, B.: A new quantitative image analysis method for improving breast cancer diagnosis using dce-mri examinations. Medical physics **42**(1) (2015) 103–109

[114] Gallego-Ortiz, C., Martel, A.L.: Improving the accuracy of computer-aided diagnosis for breast mr imaging by differentiating between mass and nonmass lesions. Radiology **278**(3) (2015) 679–688

[115] Dalmış, M.U., Gubern-Mérida, A., Vreemann, S., Karssemeijer, N., Mann, R., Platel, B.: A computer-aided diagnosis system for breast dce-mri at high spatiotemporal resolution. Medical physics **43**(1) (2016) 84–94

[116] Milenković, J., Dalmış, M.U., Žgajnar, J., Platel, B.: Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast dce-mri protocol. Medical physics (2017)

[117] Chen, W., Giger, M.L., Bick, U., Newstead, G.M.: Automatic identification and classification of characteristic kinetic curves of breast lesions on dce-mri. Medical physics **33**(8) (2006) 2878–2887

[118] Tzalavra, A., Dalakleidi, K., Zacharaki, E.I., Tsiaparas, N., Constantinidis, F., Paragios, N., Nikita, K.S.: Comparison of multi-resolution analysis patterns for texture classification of breast tumors based on dce-mri. In: International Workshop on Machine Learning in Medical Imaging, Springer (2016) 296–304

[119] Drukker, K., Anderson, R., Edwards, A., Papaioannou, J., Pineda, F., Abe, H., Karzcmar, G., Giger, M.L.: Radiomics for ultrafast dynamic contrast-enhanced breast mri in the diagnosis of breast cancer: a pilot study. In: Medical Imaging 2018: Computer-Aided Diagnosis. Volume 10575., International Society for Optics and Photonics (2018) 105753U

[120] Nie, K., Chen, J.H., Hon, J.Y., Chu, Y., Nalcioglu, O., Su, M.Y.: Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast mri. Academic radiology **15**(12) (2008) 1513–1525

[121] Fusco, R., Sansone, M., Filice, S., Carone, G., Amato, D.M., Sansone, C., Petrillo, A.: Pattern recognition approaches for breast cancer dce-mri classification: a systematic review. Journal of medical and biological engineering **36**(4) (2016) 449–459

[122] Wang, L., Harz, M., Boehler, T., Platel, B., Homeyer, A., Hahn, H.K.: A robust and extendable framework towards fully automated diagnosis of nonmass lesions in breast dce-mri. In: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, IEEE (2014) 129–132

[123] Liu, H., Zheng, Y., Liang, D., Tang, P., Ren, F., Zhang, L., Zhao, Z.: Total variation based dce-mri decomposition by separating lesion from background for time-intensity curve estimation. Medical physics **44**(6) (2017) 2321–2331

[124] Vignati, A., Giannini, V., De Luca, M., Morra, L., Persano, D., Carbonaro, L.A., Bertotto, I., Martincich, L., Regge, D., Bert, A., et al.: Performance of a fully automatic lesion detection system for breast dce-mri. Journal of Magnetic Resonance Imaging **34**(6) (2011) 1341–1351

[125] Chang, Y.C., Huang, Y.H., Huang, C.S., Chen, J.H., Chang, R.F.: Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced mri. Magnetic resonance imaging **32**(5) (2014) 514–522

[126] McClymont, D., Mehnert, A., Trakic, A., Kennedy, D., Crozier, S.: Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph. Journal of Magnetic Resonance Imaging **39**(4) (2014) 795–804

[127] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42** (2017) 60–88

[128] Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19** (2017) 221–248

[129] Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Scientific reports **6** (2016) 24454

[130] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (MIDL). (2018)

[131] Zhang, J., Saha, A., Zhu, Z., Mazurowski, M.A.: Breast tumor segmentation in dce-mri using fully convolutional networks with an application in radiogenomics. In: Medical Imaging 2018: Computer-Aided Diagnosis. Volume 10575., International Society for Optics and Photonics (2018) 105750U

[132] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241

[133] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440

[134] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 565–571

[135] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1529–1537

[136] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., et al.: Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging **37**(2) (2018) 384–395

[137] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4) (2018) 834–848

[138] Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y.: A deep learning model integrating fcnns and crfs for brain tumor segmentation. Medical image analysis **43** (2018) 98–111

[139] Wu, H., Gallego-Ortiz, C., Martel, A.: Deep artificial neural network approach to automated lesion segmentation in breast. In: Proceedings of the 3rd MICCAI Workshop on Breast Image Analysis. (2015) 73–80

[140] Rasti, R., Teshnehlab, M., Phung, S.L.: Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. Pattern Recognition **72** (2017) 381–390

[141] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging **35**(5) (2016) 1299–1312

[142] Hadad, O., Bakalo, R., Ben-Ari, R., Hashoul, S., Amit, G.: Classification of breast lesions using cross-modal deep learning. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE (2017) 109–112

[143] Amit, G., Ben-Ari, R., Hadad, O., Monovich, E., Granot, N., Hashoul, S.: Classification of breast mri lesions using small-size training sets: comparison of deep learning approaches. In: Medical Imaging 2017: Computer-Aided Diagnosis. Volume 10134., International Society for Optics and Photonics (2017) 101341H

[144] Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3150–3158

[145] Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. arXiv preprint arXiv:1710.01766 (2017)

[146] Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., Barkan, E.: A region based convolutional network for tumor detection and classification in breast mammography. In: Deep Learning and Data Labeling for Medical Applications. Springer (2016) 197–205

[147] Carneiro, G., Nascimento, J.C.: Incremental on-line semi-supervised learning for segmenting the left ventricle of the heart from ultrasound data. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 1700–1707

[148] Pesce, E., Ypsilantis, P.P., Withey, S., Bakewell, R., Goh, V., Montana, G.: Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv preprint arXiv:1712.00996 (2017)

[149] Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Li, F.F.: Thoracic disease identification and localization with limited supervision. In: CVPR. (2018)

[150] Zhu, W., Vang, Y.S., Huang, Y., Xie, X.: Deepem: Deep 3d convnets with em for weakly supervised pulmonary nodule detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2018)

[151] Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 253–260

[152] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. (2015)

[153] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1–9

[154] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778

[155] Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals. In: International Conference on Learning Representations. (2017)

[156] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. (2017)

[157] Guendel, S., Grbic, S., Georgescu, B., Zhou, K., Ritschl, L., Meier, A., Comaniciu, D.: Learning to recognize abnormalities in chest x-rays with location-aware dense networks. arXiv preprint arXiv:1803.04565 (2018)

[158] Geras, K.J., Wolfson, S., Shen, Y., Kim, S., Moy, L., Cho, K.: High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047 (2017)

[159] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)

[160] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y.: Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927 (2018)

[161] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3) (2015) e1001779

[162] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al.: Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. In: Medical Imaging with Deep Learning (MIDL). (2017)

[163] Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A.P., Palmer, L.J.: Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv preprint arXiv:1711.06504 (2017)

[164] Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Development and validation of deep learning algorithms for detection of critical findings in head ct scans. arXiv preprint arXiv:1803.05854 (2018)

[165] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639) (2017) 115

[166] Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering **2**(3) (2018) 158

[167] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2921–2929

[168] Yang, X., Wang, Z., et al.: Joint detection and diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2017)

[169] Feng, X., Yang, J., et al.: Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2017)

[170] Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M.: Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: International Conference on Machine Learning in Medical Imaging (MLMI 2018) In conjunction with MICCAI. (2018)

[171] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. (2017) 618–626

[172] Zhao, G., Zhou, B., Wang, K., Jiang, R., Xu, M.: Respond-cam: Analyzing deep models for 3d imaging data by visualizations. (2018)

[173] Cai, J., Lu, L., Harrison, A.P., Shi, X., Chen, P., Yang, L.: Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2018)

[174] Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K.: Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint arXiv:1803.07703 (2018)

[175] Dubost, F., Bortsova, G., et al.: Gp-unet: Lesion detection from weak labels with a 3d regression network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (2017)

[176] Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: Advances in Neural Information Processing Systems. (2017) 6967–6976

[177] Anh Ngo, T., Carneiro, G.: Fully automated non-rigid segmentation with distance regularized level set evolution initialized and constrained by deep-structured inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 3118–3125

[178] Cremers, D., Schmidt, F.R., Barthel, F.: Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–6

[179] Lakhani, P., Prater, A.B., Hutson, R.K., Andriole, K.P., Dreyer, K.J., Morey, J., Prevedello, L.M., Clark, T.J., Geis, J.R., Itri, J.N., et al.: Machine learning in radiology: applications beyond image interpretation. Journal of the American College of Radiology **15**(2) (2018) 350–359

[180] Gilbert, F., Selamoglu, A.: Personalised screening: is this the way forward? Clinical radiology **73**(4) (2018) 327–333

[181] Mango, V.L., Morris, E.A., Dershaw, D.D., Abramson, A., Fry, C., Moskowitz, C.S., Hughes, M., Kaplan, J., Jochelson, M.S.: Abbreviated protocol for breast mri: are multiple sequences needed for cancer detection? European journal of radiology **84**(1) (2015) 65–70