A Pattern for Arguing the Assurance of Machine Learning in Medical Diagnosis Systems

Chiara Picardi, Richard Hawkins, Colin Paterson, and Ibrahim Habli

Assuring Autonomy International Programme, The University of York, York, U.K.
{Chiara.Picardi, Richard.Hawkins, Colin.Paterson,
Ibrahim.Habli}@york.ac.uk

Abstract. Machine Learning offers the potential to revolutionise healthcare with recent work showing that machine-learned algorithms can achieve or exceed expert human performance. The adoption of such systems in the medical domain should not happen, however, unless sufficient assurance can be demonstrated. In this paper we consider the implicit assurance argument for state-of-the-art systems that uses machine-learnt models for clinical diagnosis, e.g. retinal disease diagnosis. Based upon an assessment of this implicit argument we identify a number of additional assurance considerations that would need to be addressed in order to create a compelling assurance case. We present an assurance case pattern that we have developed to explicitly address these assurance considerations. This pattern may also have the potential to be applied to a wide class of critical domains where ML is used in the decision making process.

Keywords: Machine Learning · Assurance · Assurance Cases · Clinical Diagnosis.

1 Introduction

Machine Learning (ML) offers the potential to create health care applications that can perform as well as, or better than, human clinicians for certain tasks [17]. This could help address major societal challenges, including the shortage of clinicians to meet the demands of an ageing population and the inadequate access to health care services in poor parts of the world [28]. For example, the prevalence of sight-threatening diseases has not been matched by the availability of ophthalmologists with the clinical expertise to interpret eye scans and make the appropriate referral decisions [3]. ML has the potential to address this shortage and augment, and in certain cases improve, existing clinical practices by giving clinicians more time to care for patients. [26].

However, clinical diagnosis is a critical activity, the failure of which could compromise the safety and quality of the overall care process. As such, the introduction of clinical diagnosis technologies for augmenting or replacing human expertise has to undergo the necessary rigorous evaluation of the system in its intended context and the assurance of the processes by which the system is developed, evaluated and maintained [24]. For ML-based systems, this includes

performance characteristics, e.g. hit and false alarm rates, and the appraisal of the quality and appropriateness of the data and processes by which the system is trained and tested.

Because of their critical nature, clinical diagnosis systems require assurance. Assurance is defined as justified confidence in a property of interest [13], often the property of interest is safety. The assurance of a system is typically communicated in the form of an assurance case, capturing *"a reasoned and compelling argument, supported by a body of evidence, that a system, service or organisation will operate as intended for a defined application in a defined environment"* [1].

This paper proposes an assurance argument pattern that provides a structured, clear and reusable basis for justifying, as part of an assurance case, the use of Machine Learnt models (MLM) in clinical diagnosis systems. This includes reasoning about the performance of the models and the means by which they are trained and tested. The argument pattern can be used to support the development of holistic assurance cases, potentially utilising further evidence for clinical effectiveness and patient safety from randomised control trials and pilot clinical deployments. The generation of a compelling assurance case will both guide development of MLM, as well as facilitating the necessary dialogue between ML developers, clinical users and independent assessors (e.g. regulators).

The rest of the paper is organised as follows. In Section 2, we motivate the need for an assurance argument pattern by focusing on a significant machine-learnt system for retinal diagnosis and referral [5]. We construct an explicit assurance argument for this system and examine the assurance factors that have to be demonstrated prior to the adoption of such a system. In Section 3, we propose an assurance argument pattern that addresses the assurance factors highlighted in the previous section. This considers, in an integrated manner, the performance of the MLM and the means by which these models are trained and tested. In Section 4 we discuss the argument pattern and consider its applicability in the wider domain, e.g. for non-healthcare industries, noting that generalisability would require a similarly detailed analysis in other domains. This is identified in Section 5 as one of the areas for future work.

2 Motivating Case Study

The pattern introduced in this paper arose through the consideration of the implicit assurance arguments for three major deep learning models covering the following clinical areas:

- Retinal disease diagnosis and referral [5];
- Optimal treatment strategies for sepsis in intensive care [16];
- Arrhythmia detection and classification [12].

In this paper, we focus on the first study by Fauw and colleagues [5], because of the significance and richness of the published results. The study describes a system able to examine three-dimensional Optical Coherence Tomography (OCT) scans and make referral recommendations on a range of sight-threatening

retinal diseases. Figure 1 shows how the system is composed of two parts represented by two different deep neural networks: segmentation and classification.

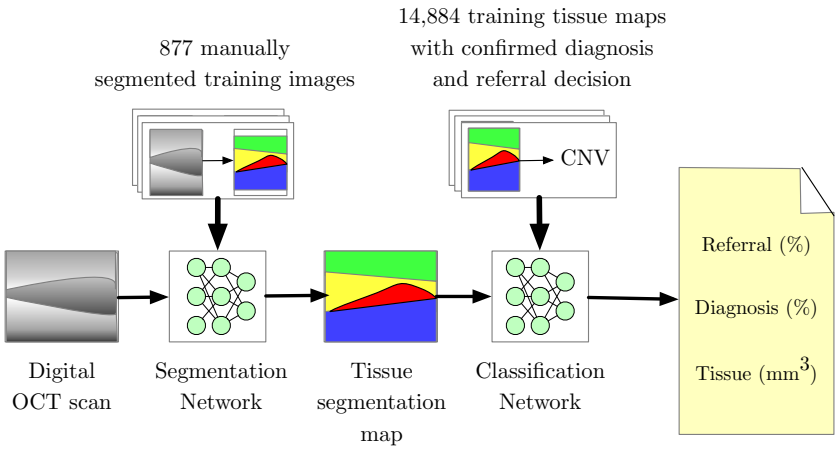

Fig. 1. Automated Retinal Disease Diagnosis and Referral System (Adapted from [5]).

The segmentation network, which is trained using 877 images manually segmented by trained ophthalmologists, takes as input OCT scans and creates a detailed device-independent tissue-segmentation map (used for identifying clinical features in scans for diagnosis). The map created is then given as input to the classification network in order to provide one of the four referral suggestions in addition to the presence or absence of multiple retinal pathologies. The classification network is trained using 14884 tissue maps labelled by four retina specialists and four optometrists with the diagnosis and the referral decision. The two neural networks represent the two MLM of the system. In this section we report our interpretation of the implicit assurance argument contained in the published study and discuss the additional assurance considerations needed to support a potential deployment of the technology.

2.1 Understanding the Implicit Assurance Argument

We represent here the assurance argument structures for the segmentation and classification neural networks which we have extracted from the information in the published study. This implicit argument has been represented explicitly using the Goal Structuring Notation (GSN) [1]. GSN is a graphical notation for explicitly capturing the different elements of an argument (claims, evidence and contextual information) and the relationships between these elements. GSN is a generic argument structuring language that is widely used in the safety-critical domain [18].

Figure 2 shows the graphical elements that we use in this paper. In GSN, the claims of the argument are documented as *Goals* and the evidence is cited in *Solutions*. Additional information, in the form of *Contexts*, are also provided.

Fig. 2. GSN Graphical notation

The notation includes two types of links that can be used to document the relationships between elements: *SupportedBy* (represented as lines with solid arrowheads) indicates inferential or evidential relationships; *InContextOf* (represented as lines with hollow arrowheads) declares contextual relationships. The additional elements shown in Figure 2 are provided to support patterns and are introduced in Section 3. The reader is advised to consult the publicly available GSN standard [1] for a more detailed description of the notation.

The assurance arguments for the neural networks are shown in Figures 3 and 4 (abstracted from the detailed assurance arguments in [20]). The main claim is that the neural network achieves or exceeds the intended performance (i.e. in tissue-segmentation, diagnosis and referral). This claim is supported by the performance results reported in the study. In addition to this claim and supporting evidence, the study provides a number of items of contextual information:

- description of the clinical setting (Moorfields Eye Hospital which is the largest eye hospital in Europe and North America);
- description of the neural networks used;
- description of the benchmark against which the performance of the neural networks is judged, including the profiles of the clinical experts;
- description of the data used.

The data is divided into three different sets: training, validation and test sets. The training set is used to find the best model; the validation set is used to choose the hyperparameters of the model in order to avoid overfitting; and the test set is used to verify the model with data never seen before. The type of the data, and the amount included in each set, are described as context to the main claim.

It is important to highlight that the arguments reported above represent our interpretation of the implicit argument contained within the published study, which required several review iterations of the results, including the rich supplementary material. We could characterise the structure of the implicit arguments for the neural networks as being of the form depicted in Figure 5. That is, the performance claim is directly supported by evidence. Importantly, this evidential relationship is established with clear links to the machine learnt network, the clinical context, the data used and the benchmark against which the acceptability of the performance is judged.

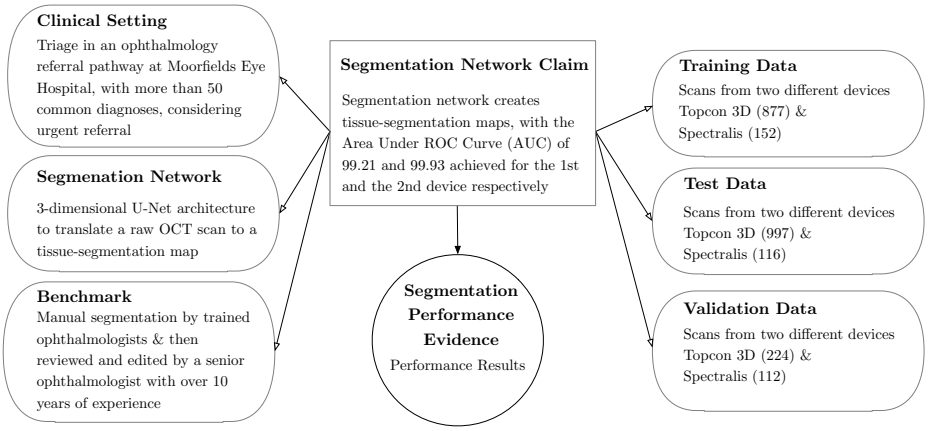

Fig. 3. Segmentation Neural Network Assurance Argument

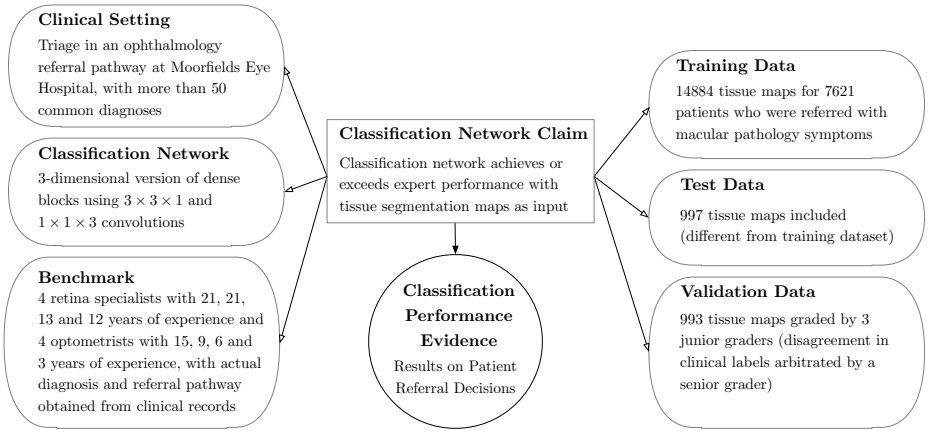

Fig. 4. Classification Neural Network Assurance Argument

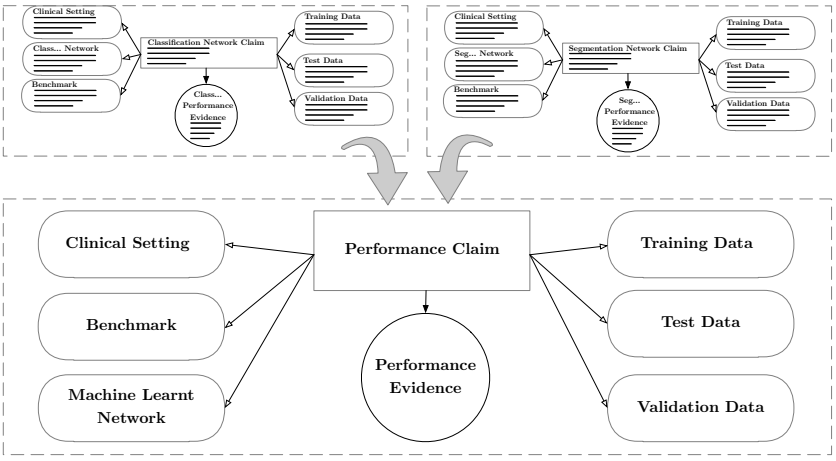

Fig. 5. The Structure of the Implicit Assurance Arguments for the ML Networks

2.2 Review of the Implicit Assurance Argument

Having identified the implicit argument shown in Figure 5, we evaluated this argument from the point of view of an assessor who is seeking to make a decision on whether to permit the use of the system as part of real clinical diagnosis. In doing so we identified a number of additional assurance considerations that would need to be addressed in order for use of the system to be approved. It is important to note that the issues we identify are not deficiencies in the published study as they are beyond the scope of the reported results. However, they do represent requirements for a potential assured deployment of the system. The assurance considerations we identified are summarised below. They were identified by performing a systematic review of the argument structure in Figure 5, following the *staged argument review process* in [1] [15], by considering the sufficiency of each of the elements in turn with respect to the confidence they provide.

1. **Clinical Setting**: In order to assure the learnt model, the context in which that model will be used must be fully and clearly understood. If the model is used in a manner for which it was not developed then there is little confidence that the model will perform as required. The clinical setting is described in the published paper, but there is no evidence to support the sufficiency of this description with respect to how the model will be used in practice. In addition, the impact of possible changes or variations in the clinical setting is not clearly considered. For example, is the model still assured if used in a hospital other than Moorfields? Is there anything in particular about this setting that is significant from an assurance perspective? An assurance case for the neural network would need to justify the validity of the clinical setting description.

2. **Benchmark**: If a judgement is to be made on the safety of the network in clinical diagnosis, then a target against which the performance of the network can be judged must be defined. The benchmark is identified in the case study as the gold standard obtained from clinical records of the final diagnosis and optimal referral pathways determined by experts. The profile of the experts involved in the diagnosis are described. The published study does not make clear how the experts were chosen: how was it decided how many years of experience are enough? What specialty is considered appropriate for the benchmark? An assurance case would need to explain why the benchmark is considered sufficient to indicate that the output of the model is acceptable.
3. **Machine Learnt Model**: Whilst the problem domain restricts the choice of the MLM which may be employed, the number of model types and variants which can be used to tackle a problem is still typically large. Selecting a model type and variant has a significant impact on model performance and is typically performed with reference to previous domain experience. The choice of model should also be undertaken with respect to a wider set of requirements, such as the need for explainability, or with consideration of the operating environment. An argument should therefore be constructed to explain the choice of model with reference to the system level requirements. In the case study the model form is clearly shown, i.e. a convolutional network, and the performance demonstrated with respect to the classification and segmentation tasks. If an assurance case were to be created for this network, the wider impact of this choice, and explicit justification for the decisions made would be required.
4. **Training and Validation Data**: The data collected for the training of MLM is a key assurance consideration as the knowledge encoded within the model is derived directly from this data. The data should be sufficient to represent all relevant aspects of the clinical setting. An assurance argument will need to consider both the relevance and completeness of the data used for training the model. The case study gives specific details on the setting in which data was gathered, i.e. 32 clinic sites serving an urban, mixed socioeconomic and ethnicity population centered around London, but does not supply explicit justification for the relevance or coverage that this data provides.
5. **Test Data**: Whilst every effort is made to ensure that the training and validation data captures the features present in the clinical setting, evidence is required to verify that the model will continue to perform as expected when deployed for real world diagnosis. To provide such assurances requires the test data to be both representative of the clinical setting and independent of the training data and learning process. The size of the test data set is provided in the case study, however, details of independence are implicit. To form a compelling assurance case a justification of the decisions concerning the collection of test data should be presented.
6. **ML Process**: The development strategy has a profound impact on the performance of the MLM and as such an argument should be made about the choices which underpin the design strategy. Typically this will concern the

validation strategies used to evaluate model performance, the hyperparameters used to control the training process and the methods employed to select and tune these hyperparameters. In the case study the authors give details of the process undertaken (e.g. the segmentation network was trained five times with different order of inputs and random initialised weights) with reference to previous work which demonstrated the effectiveness of such approaches. Further explicit justification of decisions taken during the development process are required for a more compelling case (discussed in Section 3).

Importantly, it is how issues such as those described above are addressed that would be of most interest to an independent assessor e.g. representing a regulatory authority; the performance evidence alone would not be considered to provide sufficient confidence, particularly when the assurance case is extended to cover safety. This is analogous to how conventional safety-related software requires an understanding of the implementation of the software in addition to black-box testing. In forming this view we have been fortunate to be able to interact with a number of assessors from the medical domain including representatives from NHS Digital. It would also be necessary to show how the MLM provides other desired features such as explainability or robustness. In the next section we propose an argument pattern that explicitly addresses these issues.

3 Making an Explicit and Compelling Assurance Argument for ML Decision Making

Figures 6 and 7 show a pattern that documents a reusable assurance argument structure that can be instantiated to create arguments for MLMs. The argument pattern is represented using the pattern language of GSN [1]. Figure 2 showed the *to be developed* and *to be instantiated* symbols that can be used to create abstract argument structures that can then be re-used as approapriate. *To be developed* attached to an element indicates that the element must be further developed as appropriate for the target system (through provision of specific argument and evidence). *To be instantiated* attached to an element indicates that some part of the element's content is a variable that requires instantiation. Variables are declared as part of the argument structure using curled braces, such as {MLM} in Figure 6. These variables can be substituted for references to specific instances relevant to the system of application (for example a reference to the actual MLM that has been created).

The pattern extends the argument extracted from the published study in Figure 5 such that the additional assurance considerations identified in Section 2.2 can be addressed. In particular, the pattern makes use of Assurance Claim Points (ACPs) [14], indicated by the black squares in the pattern. These ACPs represent points in the argument at which further assurance is required through the provision of a more detailed assurance argument focusing specifically on how confidence can be demonstrated (referred to as a confidence argument [14]). It should be noted that although the argument could be made without using

ACPs we feel that it is more clear and effective to do so. The advantages of separating confidence and risk arguments within an assurance case are discussed in detail in [14]. It should be noted that the undeveloped claims in Figures 6 and 7 will require further development when instantiated for a specific application; all claims must eventually be supported by evidence.

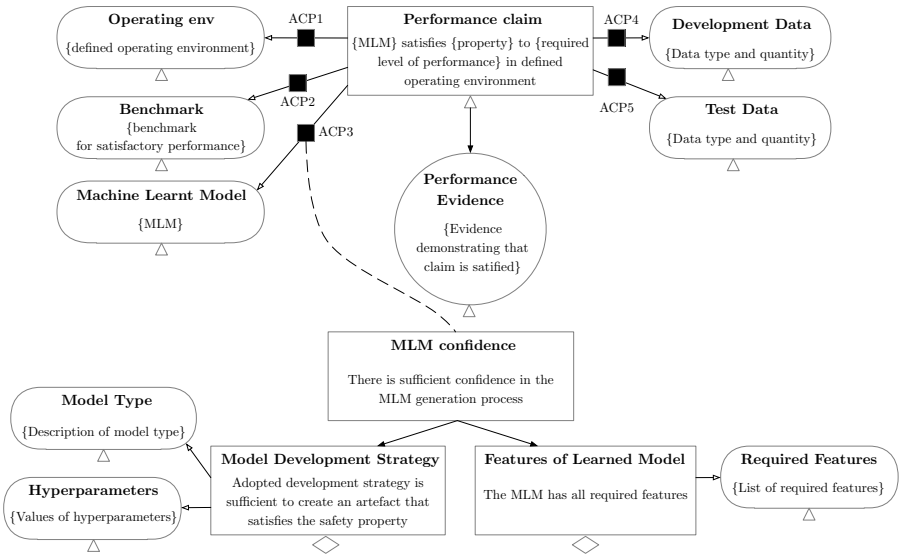

Fig. 6. Assurance Argument Pattern for Machine Learning in Medical Diagnosis

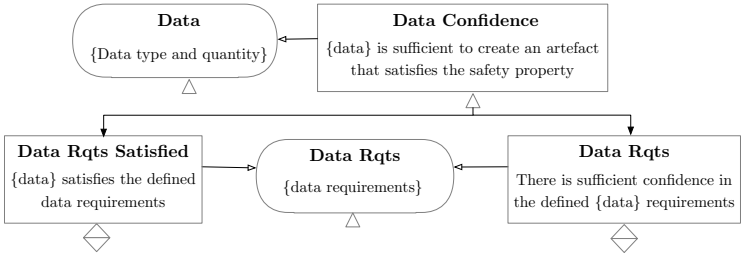

Fig. 7. Data Confidence Argument Pattern for ACP4 and ACP5

The pattern in Figure 6 retains the performance claim, supported by performance evidence, and is made in the context of the defined operating environment, the performance benchmark, and the MLM. We have used 'operating environment' rather than 'clinical setting' as this represents the more general case for the context that defines where and how MLM may be used. The data, that had

previously been split into training, test and validation, has now been split into just development data and test data. This represents the fact that there are multiple ways in which development data may be used. Whether separate validation data is selected (as in the published study) depends upon the chosen validation strategy. This representation therefore provides a more general case. Each of the items of context has an associated ACP (ACP1 to ACP5).

As can be seen in Figure 6 the pattern includes the structure for the confidence argument for ACP3 to demonstrate that there is sufficient confidence in the process used to generate the MLM. This is done through consideration of the development strategy adopted, including the choice of the model type and the respective hyperparameters, and the required features such as explainability or robustness that the learned model possesses. A pattern is also presented for the arguments at ACPs 4 and 5 to demonstrate confidence in the data. This pattern is shown in Figure 7. It can be seen that, althought the particular details of the argument will be different (as discussed later), the same general approach can be taken to argue about both the development and the test data. Therefore a standard pattern can be created for these data types.

The argument pattern presented in Figures 6 and 7 has been constructed to explicitly address the six assurance considerations identified in section 2.2. Here we explain how the argument pattern addresses each:

Considerations 1 and 2 are addressed at ACP1 and ACP2 respectively, where arguments will be provided to justify that the operating environment and benchmark are correctly defined for the application of the MLM as part of the diagnosis system. The sufficiency of the environmental definition and the benchmark that is used cannot be assessed through consideration of the MLM alone. The sufficiency of both can only be assessed within the broader context of the diagnosis and referral pathway. As such these issues would be addressed as part of the broader assurance case for the diagnosis system of which this argument forms a part [11]. Further discussion of this is beyond the scope of this paper.

Consideration 3 concerning the machine learnt model is addressed at ACP3 through focusing on confidence in the machine learnt model. The structure of this argument is shown in Figures 6. Selecting a suitable model type will typically be undertaken with reference to the category of problem being addressed by machine learning (e.g. classification or regression), type and quantity of development data available [2, 22] and in light of personal experience. The choice of model also affects a number of criteria which may impact assurance claims such as the explainability [7] or the ability of the model to be transferred between operating contexts [19]. In addition, features of the artefact produced may influence assurance arguments. Where this is the case, it should be made explicit. Reusing convolutional layers in a neural network may improve performance and training times for example, but introduce the risk of 'backdoors' [10].

Consideration 4 concerning the development data is addressed at ACP4 using the data confidence argument pattern shown in Figure 7. It is important for the argument to consider firstly, what the requirements on the training data are. These requirements should reflect the property of interest (e.g. correct diagnosis),

and the defined operating environment in which that must be achieved. Two characteristic which are of particular interest are relevance and completeness. In order to construct an argument concerning the relevance of data used in training, one should be able to demonstrate that the data is representative of the intended operational environment. In practice, collecting this data may be difficult due to safety, security or financial concerns. In such cases, it may be necessary to synthesise data sets [21] or reuse data from similar domains [23]. Even when data can be collected directly from the operating environment it is unlikely to be complete due to the complexity of most real world environments. Indeed defining completeness in many environments is a difficult task. Consider the task of photographing an injury from a single patient for use in a classification task. The lighting and position of the camera with respect to the patient will lead to a large number of possible images. A clear argument therefore needs to be presented about how the data is captured and how much data is required to adequately characterise the features of interest with the operational environment. In addition, rare cases may be known to exist but difficult to gather in practice thus leaving holes in the data set. Finally, labelling of images is a non-trivial task and experts may differ in the diagnosis offered for a given patient. In such cases, the process of labelling should be clearly stated as part of the data preparation task with conflicts and resolutions clearly stated. The supplementary information in [5] provides a detailed case of how such a task could be rigorously performed.

Consideration 5 concerning the test data is addressed at ACP5, again using the data confidence argument pattern. The central challenge of machine learning is to ensure that the trained model performs well on new, previously unseen, inputs (this is known as generalisation [9]). It is vital therefore that the test set is both representative of the operating environment and independent of the training process. It is common in machine learning to have a single data collection process and set aside a portion (usually 20%) of the data for testing. Whilst this may be suitable in some contexts, it may be more appropriate to have a collection team designated to collect testing data since the collection process itself may introduce bias into the data sets (i.e. similar to the independence requirement between the development and verification teams in the aerospace guidance DO178C [8]). Humans are very good at spotting patterns and unusual features in a data set and, if the developers have sight of the test set, the temptation to engineer features of the training set to improve training may invalidate the assumed test set independence. For the case study for example, it may be possible to collect scans from a different hospital which uses the same hardware. It is also common in traditional software engineering for the test team to check edge cases; similar tactics may be employed in the testing of MLM with rare, or complex, cases over represented in the test set.

Consideration 6 concerning the ML process is also addressed as part of ACP3 when focusing on the development strategy. Having selected a suitable artefact type, the machine learning strategy tunes parameters of the artefact to optimise an error function. The aim of the function is to quantify the performance of the artefact. In order to make such an assessment, the development team must choose

a validation strategy during training. Typically this involves strategies such as cross-validation which allow the developer to reason about the artefacts ability to generalise to unseen data. This ability to generalise is important in all but the most simple domains and as such the validation strategy should be provided as part of the assurance evidence. The model training process itself is controlled through the selection of hyperparameters which, in turn, control the performance of the artefact produced. The choice of hyperparameters should, therefore, be explicitly stated to support any assurance argument. Hyperparameters such as early stopping [9] or dropout [25], for example, may be used to control overfitting of the model to training data. Once initial values for the hyperparamters are selected, these are tuned by repeatedly training the models and updating the hyperparamters through the analysis of model performance.

In this section we have presented a pattern that we have developed for arguing the assurance of MLM, based on our review of machine-learnt models for clinical diagnosis. In the next section, we discuss the benefits and implications of using such a pattern to help assure similar systems.

4 Discussion

The assurance argument pattern presented in the previous section is intended to be used to guide developers of MLM for use in clinical diagnosis systems. It identifies how to create a compelling assurance case for the MLM that is sufficient to support a decision regarding approval to deploy the models as part of a diagnosis system. The argument pattern identifies the nature of the claims that must be made about the MLM, but also importantly helps to identify where evidence is required (testing, analysis, validation, review etc.) to support those claims. As such, practitioners who make use of the pattern will be guided towards performing a particular set of assurance activities that are required to make an assurance case for their system. In this way, the pattern should help to improve processes and practices for the utilisation of ML in clinical diagnosis.

ML is often seen as essentially an optimisation problem [27]. One thing that this paper has particularly highlighted is the fact that when ML is being used in critical applications such as clinical diagnosis, although optimisation of the learnt model remains important, other aspects of the ML process and associated contextual assumptions take on a much more critical role. It should be noted that many of these additional considerations highlighted in this paper are things that ML developers are already addressing to some extent (see the excellent supplementary information in [5]), however there has been little consideration, in the ML community, for their role in a justification for the system.

It is important to emphasise that this paper has considered only the machine learnt aspects of a larger overall system that deals with the entire retinal disease diagnosis function. The arguments discussed in this paper would therefore form part of a larger assurance case that considered the safety of the entire system. One approach to decomposing a system such as this is to consider the system as an agent characterised by a need to sense the environment of operation (Sensing),

to understand the information that is sensed by interpreting it in the context of the system and to create a useful model of the real-world (Understanding), to make decisions based upon that model (Deciding), and to perform actions that implement that decision (Acting). Each of these elements, as well as the interactions between them, must be considered as part of the system assurance case along with an understanding of the requirements of the system as a whole. The neural networks considered in this paper would form part of the Understanding and Deciding elements of the overall system (e.g. tissue segmentation, classification and referral for retinal disease). In other work we are investigating the form of the holistic assurance argument, but the details of this are outside of the scope of this paper.

Although this paper has focused on medical diagnosis, it is likely that the principles that have been extracted from studying these systems and that have been captured in the argument pattern are more broadly applicable, both to other medical applications, but potentially more broadly to other types of critical system that make use of MLM. Demonstrating this will require further case studies in other domains, however our experience shows that the techniques and processes applied in developing MLM for medical diagnosis are the same techniques that are often used for developing models for other domains, e.g. object detection and classification in autonomous driving [4]. The nature of the requirements and operational context will of course be unique to the application, and may bring unique challenges that must be addressed, but we hope that the general approach reported here will still be valid. This is one of our ongoing areas of research.

5 Conclusions and Future Work

Machine learning promises to revolutionise the way many tasks are performed and recent years has seen a growth in the application of ML to domains where failure would compromise the safety of critical processes. One such area is medical diagnosis where the benefits offered could address major societal challenges. However, the adoption of ML will require a change in the way machine learnt models are developed. Where ML, and the models generated by ML processes, are intended for use in these critical domains, there is a need for explicit assurance.

In this paper, we presented a reusable assurance case pattern that can be used to create arguments for machine learnt models in medical diagnosis systems and, as such, informs ML development teams of the key issues to be considered. The pattern reflects current ML practice as applied in medical diagnosis systems, and addresses identified assurance considerations. This includes the explicit justification of choices made during the development process including the nature of the data used. As part of our overall validation of the approach, we have presented our work to a wide clinical safety audience [6] and have received positive feedback on the utility of our approach. We believe that the pattern may also be applicable in a wide range of critical application contexts that make use of

MLMs, however demonstrating this will require a similarly detailed analysis of multiple case studies to be conducted across a number of different domains. The focus of our future work will be to carry out such an evaluation, and to update and improve the pattern based upon this experience.

6 Acknowledgements

This work is funded by the Assuring Autonomy International Programme (https://www.york.ac.uk/assuring-autonomy).

References

1. Assurance Case Working Group [ACWG]. Goal Structing Notation Community Standard version 2. `https://scsc.uk/r141B:1?t=1`, 2018. Accessed on 11/13/2018.
2. Azure-Taxonomy. How to choose algorithms for Azure Machine Learning Studio. `https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice`, 2019. Accessed on February 2019.
3. Rupert RA Bourne, Seth R Flaxman, Tasanee Braithwaite, Maria V Cicinelli, Aditi Das, Jost B Jonas, Jill Keeffe, John H Kempen, Janet Leasher, Hans Limburg, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888–e897, 2017.
4. Simon Burton, Lydia Gauerhof, and Christian Heinzemann. Making the Case for Safety of Machine Learning in Highly Automated Driving. In *International Conference on Computer Safety, Reliability, and Security*, pages 5–16. Springer, 2017.
5. Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan ODonoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.
6. NHS Digital. Digital Health Safety Conference 2019. `https://digital.nhs.uk/news-and-events/events/2019-events/digital-health-safety-conference-2019`, 2019. Accessed on 30/05/2019.
7. Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 210–215. IEEE, 2018.
8. RTCA SC-205 EUROCAE WG-12. *Software Considerations in Airborne Systems and Equipment Certification*. EUROCAE and RTCA, 2012.
9. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
10. T. Gu, B. Dolan-Gavitt, and S. Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv*, 1708.06733, 2017.
11. Ibrahim Habli, Sean White, Mark Sujan, Stuart Harrison, and Marta Ugarte. What is the safety case for health IT? A study of assurance practices in England. *Safety Science*, 110:324–335, 2018.

12. Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.
13. Richard Hawkins, Ibrahim Habli, Tim Kelly, and John McDermid. Assurance cases and prescriptive software safety certification: A comparative study. *Safety science*, 59:55–71, 2013.
14. Richard Hawkins, Tim Kelly, John Knight, and Patrick Graydon. A new approach to creating clear safety arguments. In *Advances in systems safety*, pages 3–23. Springer, 2011.
15. Tim Kelly. Reviewing assurance arguments-a step-by-step approach. In *Workshop on assurance cases for security-the metrics challenge, dependable systems and networks (DSN)*, 2007.
16. Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.
17. Thomas M Maddox, John S Rumsfeld, and Philip RO Payne. Questions for Artificial Intelligence in Health Care. *Jama*, 2018.
18. University of York. Goal Structuring Notation. `impact.ref.ac.uk/casestudies/CaseStudy.aspx?Id=43445`, Nov 2014. Accessed: 03 Jan 2019.
19. Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
20. Chiara Picardi and Ibrahim Habli. Perspectives on assurance case development for retinal disease diagnosis using deep learning. In *17th Conference on Artificial Intelligence in Medicine*, 2019.
21. German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.
22. scikit Taxonomy. scikit - Choosing the right estimator. `https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html`, 2019. Accessed on February 2019.
23. Michael Shneier, Tommy Chang, Tsai Hong Hong, Geraldine S Cheok, Harry Scott, Steven Legowik, and Alan Lytle. Repository of sensor data for autonomous driving research. In *Unmanned Ground Vehicle Technology V*, volume 5083, pages 390–396. International Society for Optics and Photonics, 2003.
24. Edward H Shortliffe and Martin J Sepúlveda. Clinical Decision Support in the Era of Artificial Intelligence. *Jama*, 320(21):2199–2200, 2018.
25. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
26. Eric Topol. The Topol Review: Preparing the healthcare workforce to deliver the digital future. `https://topol.hee.nhs.uk/`, 2019. Accessed: 27 Feb 2019.
27. Kiri Wagstaff. Machine Learning that Matters. *arXiv preprint arXiv:1206.4656*, 2012.
28. World Health Organisation (WHO). Health workforce. `https://www.who.int/gho/health_workforce/en`, 2019. Accessed: 27 Feb 2019.