A review on the need for better explainability with increasing reliance on machine generated medical diagnostics

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Eric Armstrong
Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _12/9/20__
Eric Armstrong

Approved _____ Date _12/9/20__
Anil Vullikanti, Department of Computer Science

A review on the need for better explainability with increasing reliance on machine generated medical diagnostics

Eric Armstrong
Computer Science
University of Virginia
Charlottesville VA USA
era9dv@virginia.edu

Eric Armstrong. 2020. A review on the need for the development of explainability with increasing reliance on machine generated medical diagnostics. University of Virginia. *Charlottesville, VA, USA, 5 pages.*

ABSTRACT

I will be writing a literature review on the topic of machine learning particularly around the explainability of models in machine predicted diagnostics. I will also address legal concerns that could arise around how to fairly incorporate these algorithms into society especially due to the difficulty of addressing biases in models. With the rise of machine learning in generating medical diagnostics, learning models could foreseeably surpass humans in predicting these diagnostics, therefore changing the medical landscape to trust models' predictions more than doctors due to standard of care definitions and its consequences on medical malpractice. This means society would have to trust black box diagnostics over historically trusted human diagnostics. I will be discussing potential flaws about how machine learning could help or hurt and potential future steps to further the conversation on this topic. Current models are identifying melanoma in 95% of images where current doctors are only able to identify 86.6%. Additionally, around 5% of adults receive misdiagnosis from doctors, which leads to additional treatment being necessary. The significance of these findings is that doctors may soon be required to make a diagnosis along with a Machine learning model to better validate the accuracy of both predictions. This could ultimately lead to doctors relying more and more on machines' predictions that are harder to understand due to the complex way that their predictions are made, which is why the explainability of models and predictions is a crucial corequisite in this societal development.

ACM Reference format:

1 Introduction

Medical malpractice occurs when a hospital, doctor or other health care professional, through a negligent act or omission, causes an injury to a patient. The negligence might be the result of errors in diagnosis, treatment, aftercare or health management. As machine learning diagnostics continue to improve and are consistently able to outperform medical professionals, medical malpractice law could cause the new standard of care to require that doctors make diagnoses alongside a high-quality machine learning diagnosis [1]. However, medical malpractice goes beyond just diagnosis. If the injury was due to the medical professional's error in providing a treatment, it could still be considered malpractice. Consequently, if machine learning models are too difficult to be understood by doctors, they may not know how or why a certain diagnosis was made and may neglect to provide the best treatment possible. Since treatment strategies are often not as effective when deployed in clinical practice compared to preliminary evaluation, there could be a decrease in the quality of care, which is the primary purpose of the medical malpractice law to begin with [1]. In order for machine learning diagnostics to be readily accepted, more transparency is needed for medical professionals to understand which factors were most significant in determining the output and why. There are many different approaches to make models and predictions explainable but in order for them to be explainable, from a legal perspective, for instance, a set of rules must be defined,

particularly with what the model should guarantee (e.g. trustworthiness of the model and comprehensibility for medical professionals [2]. There are many important aspects of these black box problems, but it can be better understood by breaking them up into different categories and analyzing solutions to each of these categories separately [2].

2 Related Work

Much of this review is focused on the need for more transparency in machine learning algorithms for deployment of machine assisted diagnostic predictions. We must be able to understand how a predictor decided on a particular diagnosis to ensure the right features are being used to draw conclusions. The focus for opening a black box is primarily on explaining predictions with some mention of the importance of explaining the model. The main predictor being analyzed is a model agnostic local predictor LIME and its global variant SP-LIME.

Related research covers model specific, versus agnostic, methods, and analyzes how different algorithms for explainability are better-suited for different medical predictions. Other black box explainability research can be found with the intention of explaining bias in machine learning applications, which is very important in all aspects of machine learning where humans are involved, to ensure all patients are receiving the same standard of care regardless of socio-economic background.

3 System Design

3.1 Taxonomy of Literature

3.1.1 *Focus on introduction of Machine learning in the medical field*

Both "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning" and "Explainable deep learning models in medical image analysis" greatly discuss the introduction of machine learning into the medical domain. The former is more focused on the social concerns with legal and financial challenges around the rise of superior proprietary software over open source, and the potential for over reliance on these prediction systems to cause major problems down the road. The latter is focused on different explainability techniques for specific medical diagnoses in addition to the challenges of deploying them in a clinical setting. While ""Why Should I Trust You?" Explaining the Predictions of Any Classifier" discusses this less, it does discuss an experiment done with their model with medical experts such as would be done with Human in the Loop (HITL) systems with feature engineering and how this greatly improved performance and trust with the user.

3.1.2 *Focus on Explainability techniques for Black Box models*

The papers ""Why Should I Trust You?" Explaining the Predictions of Any Classifier" and "A Survey of Methods for Explaining Black Box Models" are more focused on the problem of "explainability" and try to quantify and define exactly what this means by breaking it into Model explainability, outcome explainability, and model inspection. The former is very focused on a specific model-agnostic algorithm (SP)-LIME in the context of model explainability and outcome explainability, while the latter generalizes much more and provides analysis on many algorithms in the problems of opening black box problems defined above.

3.2 Detail of Literature

3.2.1 *Why do we need explainable AI?*

In order to frame the problem for the need of explainable AI, it is important to understand the current challenges of introducing predictive algorithms into the medical profession. It is important to first realize that doctors are used to making decisions on their own based on what they know, and can very clearly explain why they chose a particular diagnosis or treatment based on the information they have available to them. Whether they come to the best conclusion is not guaranteed, but they are able to reasonably explain and defend themselves if necessary, which is not the case for many current AI's. As mentioned earlier, 5% of patients require additional treatment due to misdiagnosis and treatment. AI is already superior than doctors in diagnosis is many areas but particularly cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others [1]. If AI continues to demonstrate it can diagnose patients better than doctors, than by

definition of the standard of care, it could very likely be required that AI is required to assist doctors in diagnosis. Further, if a doctor decides to stray from the AI recommended treatment without a good justification, and misdiagnoses the patient, they could be at risk of a malpractice liability. The potential of solely machine run diagnostics is also addressed as a scenario but is criticized heavily. If changes in research arise to diagnose differently or machines start only getting diagnosis data from their own predictions, in the long run they could start to perform differently or unexplainably, therefore causing the opposite of what the standard of care laws set out to do.

In conclusion of the paper "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning" it is decided that the best solution is a Human-In-The-Loop (HITL) approach, where the medical professional is in control of the prediction process and, therefore, has a better understanding of how a prediction was reached. However, in order for the medical professional to be able to diagnose patients in real-time, predictions will have to be more explainable and interpretable.

3.2.2 *How do we explain black box problems?*

Before we can determine if a model is interpretable, a definition for interpretability needs to be defined. As defined in reference [2], interpretability is "the ability to explain or to provide the meaning in understandable terms to a human", where the human in our case is the medical professional. Meaning the output should most likely be in medical terms that the type of doctor would be able to understand at a glance and that gets the relevant information across. It is also important to consider the complexity of the explanation and how long it might take for the doctor to review the information versus how much time they have to make a decision.

In addition to understanding the output format for a prediction of a particular model, it is important the doctor knows how a particular outcome was reached in order to trust their final decision. This is defined as outcome explanation. The outcome explanation is a human understandable explanation derived using an interpretable local predictor around the black box in question and the input the black box predicted upon [2]. The other part of explaining black boxes focuses on explaining the black box as a whole, otherwise known as a global interpretation. This is essentially an interpretable and transparent representation to try and understand how the model acts overall with varying inputs. While this may not be useful to a medical professional who wants to know how a particular diagnosis was derived, it is useful for the doctor to build trust in the model. In order to present this explanation for a single instance or set of instances, it is common in literature for models to use decision trees and the rules for them to a given instance as the explanation(s) [2]. Recent research, however, has been focused on agnostic explanators for interpreting any black box, versus model specific explanators. Two methods that will be examined in particular are LIME, which is an agnostic method for outcome explanation, and its variant SP-LIME which is an agnostic method for model explanation.

3.2.3 *LIME as a solution for outcome explanation*

LIME (Local Interpretable Model-agnostic Explanations) is an agnostic, local explanator for individual predictions, with the goal of building trust for the user. LIME builds trust for its user by interpretably explaining how a prediction was made so that a user is able to validate if the prediction was reasonably constructed based on prior knowledge related to the prediction.

The main struggle of LIME is the tradeoff between interpretability and local fidelity, where interpretability could come from limiting the number of weights from a linear model displayed to the user (lowering the complexity), which would conversely lower the local fidelity, or trust, due to some features being removed. The formula for a prediction is as follows.

$$\xi(x) = L(f, g, \pi_x) + \Omega(g)$$

Where the explanation ξ is the minimum of the unfaithfulness $L(f, g, \pi_x)$, and complexity $\Omega(g)$, and g is some interpretable model. Additionally, f(x) is the probability that x belongs to a specific class, and π_x is a measure of the locality of an instance with respect to class x. The paper uses sparse linear models for explanations; therefore, the complexity would be the count of non-zero weights.

In order to build the local explanation, LIME uses perturbation, on the domain of the interpretable representation, on uniformly distributed samples around the instance being explained, weighted by the distance from the instance, to obtain a function for the

labels of the samples. Perturbation is a way to analyze the effect of changing the input features on the output of an AI model [4], in this case, by masking the features not in the interpretable domain and running a forward pass to see what output each of the surrounding samples produce. Using this to optimize the linear explanation. The interpretable result is then output as the weights of each of the feature which in the case of text, the user can read very quickly. It is also important to note that this whole process is very fast even for a large number of features (which is probably not reasonable for medical text of symptoms, for example), due to it being a linear model optimization. This can also be used to determine trust for a particular black box model or prediction. If you are a doctor and you see the most important features for diagnosing a patient with the flu are things that are medically uncorrelated with the flu, you would know that the model is either predicting poorly for this task, trained on bad data, or that there are certain features in that data that may have to be removed. This explanation is locally faithful and is fairly formidable to sampling noise due to the weights applied based on locality.

3.2.4 SP-*LIME as a solution for model explanation*

SP-LIME (Submodular Pick- LIME) is a model agnostic approach for giving a global understanding of the model through a careful selection of B explanations, chosen based on their importance while also minimizing redundancy, where B is the number of instances the user is willing to look at to judge the trustworthiness of a model (potentially based on some time constraint). The idea behind minimizing the redundancy is to show the user as many different features as possible; if one selection is redundant, most of the features explaining that selection have already been used explaining another selection. The goal is to get the most diverse selection of explanations.

3.2.5 *Methods for medical image analysis*

Two methods for medical image analysis are attribution-based methods and novel methods that are architecture or domain specific, however the focus will be on attribution-based methods since these are the focus of the paper being analyzed [4]. The goal of an attribution method is to determine how any given input feature contributes to an output neuron. These contributions can be arranged in the shape of the input image to form a heatmap, or specifically, an attribution map. This process can be broken down further into perturbation methods and backpropagation methods. Perturbation as mentioned above, is much easier to implement and allows you to see how changing features in the input changes the output, leading to the benchmark for any image attribution study, Occlusion. Occlusion is when you cover parts of the image with a grey square, essentially hiding it from view, to show which features of an image are positively or negatively responsible for the prediction. This helps identify if your model is overfitting and learning irrelevant features, particularly with adversarial images, which are images that would appear to be very obvious to a human but have minor, visibly unnoticeable changes to cause a model to identify it as something else. However, occlusion often takes hours to run per image due to the amount of computation from running a forward pass on every perturbation. Backpropagation methods are much faster due to only needing to do one forward and backward pass to learn all the attributions, however, have a weaker relationship between the outcome and the variation of the output. These methods often perform better based on the type of medical image that needs to be analyzed and the type of output that helps experts the most.

4 Conclusion

4.1 Common Themes

One of the most important aspects of determining whether machine learning algorithms are suitable for medical practice is their transparency, however there is no well accepted standard for what this means or how to measure it, which is what many of these blackbox-oriented papers try to address. Models are typically ranked by their accuracy on a test set; however, this is not always representative of real-world data and in critical predictions such as diagnosis and treatment, the doctor must know how an outcome was reached in order to trust their final decision. To do this, it is unanimously agreed upon that a Human-In-The-Loop system will be absolutely necessary for trust and for addressing potential differences in real world data. Additionally, in order for the doctor to understand these results with little knowledge of machine learning, the explanation will have to be interpretable. Interpretability of explanations is another unclear definition that these papers start to define. A common ground for interpretability with regards to the papers reviewed is that for something

to be interpretable it must be easily understood by the target audience and take into account any limitations or constraints such as the time needed to make a diagnosis.

Medical professionals must be able to make decisions in real time and a large concern over the introduction of machine predictions into this process is the time it will take for the model to determine its prediction or for the doctor to interpret the results of the prediction. While "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning" addresses the need for real time assistance, it fails to give solutions to this beyond batch training instead of stream learning. However, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier" provides an approach to this problem with the method, LIME, that takes into account the amount of time a professional might have to make a decision by limiting the number of explanations shown to the user. One final commonality that is found is the rise of commercial interest in explainability methods. While this is good for the development of these methods, it could lead to very high costs for hospitals if they will need to be paying to use the proprietary models while also needing to pay doctors, not taking into account that doctors may also need to receive additional training to learn how to work with the new technology.

4.2 Disagreements

While the majority of the papers agree on most paths forward and how black boxes need to be explained in order to be useful, there are minor conflicts. The LIME algorithm is featured in "A Survey of Methods for Explaining Black Box Models" as one of the forerunners in local explainability, however has weak points due to the required transformation of any type of data in a binary format that is claimed to be human interpretable in the output of the model. Additionally, the explanations are only provided through linear models and their features' importance which may not be optimal for very complex, non-linear datasets. A new algorithm, LORE, is proposed as a better alternative for performance and clarity of explanations, however is not discussed in detail.

4.3 Course Topics

Machine learning - The majority of information in all 4 papers reviewed is around the topic of machine learning. This ranges from how machine learning is advancing to being able to diagnose better than medical professionals, the black box problem of machine learning, and the machine learning solution of building a model to interpret a black box. Understanding how machine learning works, especially neural nets and decision trees, is crucial to understanding the black box problem, and the need for explainability around predictions when models could be drawing conclusions on features that doctors would not normally use to make predictions.

Databases - The topic of databases are used to supplement part of the background needed for understanding how data is stored for machine learning algorithms. Databases are also mentioned in part of the paper "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning" when discussing solutions to the problem of data becoming overwhelmingly machine generated or machine influenced. Understanding how databases store data and how accessing separate databases to get different parts of data affects efficiency is crucial to evaluate this solution.

5 Future Work

The overarching theme mentioned in all the medical related literature is that Human in the Loop systems will be required in order to introduce machine learning into the medical landscape. From a legal and ethical standpoint, this is needed to make sure conclusions are being drawn from the correct features, and so that doctors can trust these decisions knowing they played a part in them, however, how this will be implemented between the cost of use and doctor acceptance and comfort using the new technology needs to be examined more thoroughly. Additionally, all 4 papers mention the need to better define explainability and interpretability metrics. Furthermore, some papers state the need for these to be graded or quantified in some way to determine if the chosen algorithms are suitable for use in medical practice. Future work mentioned exclusively in individual papers includes the need for cheap alternatives to the probable rise in proprietary-dominant solutions due to medical malpractice deeming worse models as below the standard of care, and proprietary models are still outperforming open

source models in performance and accuracy. Further work also must be done on non-attribution methods for image analysis, especially at the global level. Concept vectors are mentioned as a potentially better way to do this, however, this method is still emerging and there is lots of work to be done.

REFERENCES

[1] Froomkin, A. Michael and Kerr, Ian R. and Pineau, Joelle. 2019. When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning. 61 Ariz. L. Rev. 33 (2019), University of Miami Legal Studies Research Paper No. 18-3, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3114347

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93 (August 2018), 42 pages. https://doi.org/10.1145/3236009

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier, University of Washington Seattle, WA, USA. https://arxiv.org/pdf/1602.04938.pdf

[4] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. Explainable deep learning models in medical image analysis, Department of Systems Design Engineering, University of Waterloo, Ontario, Canada. https://arxiv.org/pdf/2005.13799.pdf