Materials and Methods

This section is separated into three parts which correspond to the three axes of the projects. Each axis is standalone and can be separated from the others, although they all interconnect in various ways. All three parts of the projects intertwine: the first part sets up tools and interpretability methods for ML models that are already put in place, the second part establishes a baseline for the results of these models in an effort to validate their use and provide a baseline for other projects on similar dataset, and the third part is focused on making sure different moving parts of the same project are well soldered into one organized, coherent platform.

Interpretability:

Interpretability of Machine Learning models is a key issue in the field of ML and remains one of the keys to its use in the medical field. In order to be able to understand the issue at hand, time was spent reviewing the literature.

First, the importance of model explainability and interpretability was studied. In the medical field especially, end-users want to be able why a model acts the way it does. Understanding the edge cases, and in which cases the model might be biased is paramount to a fair use of this kind of technique.

It is important to remind ourselves that while some domains can benefit hugely from machine learning by solely applying the perfectly-tuned algorithms to their problems, others such as medicine and clinical diagnosis require a more applied approach. Doctors and clinicians need to be involved in the process of designing such an algorithm so that it contains the right metrics and is optimized and trained in a way that is fitting to the context. Sensitivity might be more important than specificity in some cases, and it might not make sense to “reach” for the upper left corner of the ROC curve. These particularities are absolutely context-dependent and should be thought about in association with an expert in the field of medicine.

There exists a tradeoff between interpretability and model complexity. With a more simple model, its behavior can be better understood – but it lacks the predicting power to deal with complex data. A more complex model can achieve great prediction score at the cost of a more complex architecture and poorly understood behavior.

Interpretability is also important for error tracking. When the algorithm misclassifies or outputs a wrong regression output, the user should be able to understand why. Understanding the model’s behavior generally, locally, or on a single data instance can give insight on its behavior, and some possible confounding factors in the data input used.

One of the keys to building trust in machine learning algorithms in the medical field is to focus on uncertainty in the model and results. A result should be associated to its probability of prediction so that predictions can be cut out according to a threshold. The algorithm should only communicate predictions if it these are trustworthy / probable enough. Clinicians should be able to “set” a threshold at which they want to have predictions be trustworthy. When looking at a specificity vs. sensitivity curve, the tradeoff can vary based on different models and datasets. If in a specific context, there is no way to maximize both specificity and sensitivity, the clinician should be allowed to choose which one would be more suited to the context. For example in a case of epidemics, the clinician would not want to miss out on any case – at the cost of treating patients that are falsely positive (high sensitivity but low specificity) (CHECK THIS OR IS IT THE OTHER WAY AROUND). In a low-resource context however, if there is only enough medication to treat a certain number of patients, a clinician would want to be sure to treat patients that are certain to be positive – maximize specificity but disregard sensitivity. The algorithm should be able to leave this amount of choice to the end-user as it can have a strong impact on results.

Confidence intervals on the results should also be constructed. A list of other statistical tools such as p-values will be included as well. The uncertainty in a model is split between aleatoric and epistemic

Second, a list of methods which help resolve some of the issues aforementioned were looked into.

Pertaining to interpretability of machine learning models – especially “black-boxes”, several tools have be found to be useful and efficient. The tools which have been investigated are LIME, Shapley values, SHAP, influential instances, outlier investigation, PDP plots, ALE plots, and some other tools (MENTION). Another key component of this approach is to understand the model itself and the ways in which it might be appropriate or not for the data – and problem – at hand. Although doctors are very well suited in describing what is needed for an algorithm to be trusted and useful, the understanding of the algorithm’s internal mechanisms and output might be difficult to understand without prior explanation. Since ML understanding is as important as understanding of the clinical context, the knowledge should be shared in both directions. In an effort to simply the considerable amount of knowledge required to fully understand these models, a guide was created. This guide takes a less mathematical, and more overviewing approach of the machine learning topic and its application in the medical field. Images will be used to exemplify the use of these tools, and a table will be added in order to compare interpretation models on a global level. The table includes fields such as “name”, “output”, an image of its output on a dataset, “usability”, etc… which will help identify if the method interprets a single result, a row of result, a feature, if it is a robust statistic or more of an intuition, etc… The general applicability, computational complexity and overall time to run will also be added into the table.

Some visualization tools were thought about in order to convey the maximum information in a concise and clear way. In medical contexts, just like many applications of ML, the models might deal with a considerable amount of features and data points. Therefore an emphasis was put on the selection of data to be visualized and the way data can be selected in order to convey the most useful information. The plotting technique used are all based on PDP plots.

Another aspect of this was to design a how-to guide related to the platform use. The guide is split into different parts. A first part pertains to navigating the platform and being able to use all of its functionalities; which amounts to creating a step-by-step user guide. A second part is focused on the explanation of the features and their context and use. Any end-user with a basic background in medical and mathematical sciences should be able to get the most out of the platform , therefore there will be effort put into creating a knowledge base – mostly ML and mathematics oriented, but applied to medical contexts – which can be used in parallel with platform’s algorithms. This guide should help the end-user understand the models but also perhaps understand which model is best suited to what kind of dataset and how some models contain design flaws and should be interpreted considering this. A third and final part will be the creation of medical user cases. These user cases should be focused on the aspects previously discussed. Another aspect of this guide is to have end-users be able to understand their model outputs. Therefore the guide should contain information on what results to expect, what can go wrong, what types of models are known to fail in which cases, etc… Visualization tools and metrics will also be explained and their use will be detailed.

Methods and model types will also be explained, with an image of how they generally divide the probabilistic space; for example for an SVM, have an image of how the classifier behaves so end-users can get an intuition of what type of dataset it can be used on, and what type of dataset could potentially make it behave oddly. Regularization techniques such as Lasso, and feature choice techniques such as Boruta will also be explained with both texts and images.

Third, the methods mentioned beforehand were implemented in code, using Python and libraries: pyplot, etc (ADD). A standalone python code base was set-up in order to make sure that the analysis tools to be included in the platform were both efficient and not too computationally expensive before integrating them.

Fourth, in a parallel manner to the previous points, a survey was written and conducted on (SPECIFY) a certain number of clinicians and doctors with various backgrounds in statistics and machine learning – or none. This approach is necessary because of the need of machine learning to be interwoven with the medical context, and because the platform on which the data can be used and train needs (RUN ON) to be well integrated into the workflow of clinicians (REMEMBER ITS ALSO MOSTLY FOR ON-SITE). There are many machine learning tools that have already been developed but few are tailored to the needs of the end-users in the medical field. The way the questionnaire was built was to start from reading papers related to clinicians’ needs in the field of ML, then figuring out which of these needs can be addressed via the platform. Once a first version of the questionnaire was constructed, it was given to several doctors and clinicians with different backgrounds in both mathematics, and machine-learning. Feedback from different doctors was reported. Since the questionnaire focuses on the use of the platform, a demonstration of its functionalities and an explanation of its use was performed before giving out the questionnaire. During this whole process, an emphasis was put on the fact that the platform should be suited to clinicians – therefore features should be added or removed with this in mind. Once the feedback was collected, an analysis on which features to collect was done. These chosen features will be discussed and an analysis of their feasibility and interest will be done before potentially integrating them into the platform over time.

The design of the questionnaire is done in a through way, and the choice of people answering the survey was also through. The goal of the survey is to make sure that the prioritization of the team’s concerns is correct according to the doctors. The reporting of results was done in a tabloid manner, with open-ended questions’ answers categorized into different groups. The categorization was made in order to be able to regroup answers and gain insight in a more general manner. However, specific results were still conserved, to be make sure the granularity of answers is preserved.

Model validation:

Validation of the ML algorithms used in the platform is a key step in the direction of building trust for the application of ML in the medical field.

In this context, a part of the project focused on showing that the algorithms used performed over a certain baseline established by other, fine-tuned algorithms, on the same dataset. Much criticism has been drawn over ML models in medical field being evaluated on a specific dataset which is unavailable and without having been benchmarked against any other model or dataset. In an effort to go against this trend and provide end-users with a certain notion of reliability of the model, the task of validating the platform algorithms was undertaken.

First, literature review was conducted in order to determine adequate and robust ways of comparing ML models on a similar task.

Second, an investigation on the importance of this task was done. It is important to note that this validation study is fully linked to the first part of the project which focuses on interpretability and explainability of the platform’s models. It is vital to be able to know if the models used perform well on “standard” datasets before using them on any other dataset.

The dataset used is an Ebola-dataset which was produced and curated by the IDDO (Infectious Diseases Data Observatory). The cleaning of the dataset was done by Ridha and Jonathan (perhaps). Ridha also created several baseline models to verify the dataset itself – before it is used to validate the platform models. The creation of new models using the same data-preprocessing pipeline as the platform was done. These models were tuned and the results obtained were compared to those of the platforms in order to establish a baseline of what should be expected: do the algorithms provided perform well enough compared to fully specialized and fine-tuned algorithms on the same dataset for the same task. The goal of this validation is to show that the algorithms can be used on this type of dataset with a guarantee of being not too different at a certain threshold from benchmarking algorithms.

Platform management:

A team of four people are currently working on project related to the platform. The platform management’s main goals were to oversee the different projects and their future integration to the platform, discuss the importance of various features with respect to their clinical applicability and how they would fit into the platform, how they would be understood by the end-user, etc.. , as well as sharing information and communicating on advances and setbacks with respect to each platform and any concerns or improvements related to the platform. The information shared revolved around four main topics: platform debugging, platform improvements, communication intra-team and between members and outside organizations (IDDO, clinicians, etc…) and lastly coordination of projects (making sure nothing is done twice or nothing is done in spite of someone else’s project.

Another key point of the management effort was evaluating the short term and long term objects of the platform and their feasibility. For example, the addition of some functionality modules (interpretation, imputation, clustering, sentinel/surveillance tool, multiclass labeling), some totally different functionalities (ability to support different input formats such as images and sounds) and making sure that the platform remains true to its aim. Although the aim might change with time, the platform should always be concerned with its usability and explainability, and should not get bogged down with unnecessary features. Part of the goal being the ease-of-use of the platform, an effort will be made to keep the platform “streamlined”, therefore the different team members’ projects should address this.