\chapter{Introduction}

\label{chap:backgroud}

% background **is arguing for a gap in the research**

% spec and sens are important, therefore it makes sense to have a platform using these to make the important choices!

\section{Background}

\paragraph{Infectious Disease}

Infectious diseases are responsible for 30\% of all deaths in Africa. **[WHO but check source]**

The World Health Organization (WHO) describes these as “caused by pathogenic microorganisms, such as bacteria, viruses, parasites or fungi; the diseases can be spread, directly or indirectly, from one person to another **[WHO website].**

These diseases, among which HIV, tuberculosis, Malaria, Hepatitis and neglected tropical diseases (NTD) have caused 4.3 million deaths in 2016, down from 5.3 million in 2000. Still, the proportion of people dying from low-income countries is much greater than the rest, with reports showing that Africa and South East Asia are the most **impacted [ WHO report 2019 infectious diseases]**

These diseases are complex and so are their evolution and spread. However there are some factors which put certain populations or geographical zones more at risk than others.

Notably, doctor / clinician availability is not the same in all places: low income countries tend to have less than 10 doctors per 10,000 inhabitants while high-income countries don’t (respectively 90\%,5\%). Access to medical care and facilities, the number of clinicians in the population, the state of medical infrastructures (as well as transport infrastructure which allows to get to these facilities) are very different across the globe. Moreover, some clinicians may not be as trained or not trained for as long as in wealthier countries.

Some other factors are climate-related, although humidity, rainfall and temperature are correlated both positively and negatively with different infectious diseases. However, some diseases are mostly transmitted through insects such as mosquitoes, and this species benefits greatly from humid seasons to reproduce. Climate change will most likely push climate variations to new extremes and could potentially increase the risk for propagation of infectious diseases.

There is also a strong economic factor that comes into play. With more financial resources, diseases can be more adequately treated, not just because health infrastructures are more present and up-to-date but also thanks to a higher number of health workers and to the ability for both the state and the individual to pay for more expensive – and often more efficient – treatments.

Another key aspect to infectious diseases is its transmission. In fast-growing countries, there are both very rural places where contact to animals is frequent and health and safety procedures with regards to animal contact are minimal. Animals, and insects which accompany them, being one of the main factors of transmission of diseases to humans, this behavior can lead to greater risks of outbreaks. These same countries also possess fast growing cities and urban centers with high population densities. This is the case for most countries, yet wealthier countries often possess more health infrastructures. If infectious diseases start to spread in these centers, they might be harder to stop.

Another key aspect of the problem at hands is epidemics and pandemics. The Covid-19 outbreak is a great example of how impactful on the short and long term these disease outbreaks can be. Recent history has shown that pandemics and epidemics thrive in countries that are less wealthy. The Spanish flu epidemic has affected India very strongly in terms of fatalities. However, Covid-19 seemingly has affected wealthier countries more. This can be in part explained by different government responses and co-morbidities such as age and obesity which are less frequent in less wealthy countries. Although countries in Africa and Asia have response and watch/sentinel protocols put in place, the lack of resources still means they would be disproportionately affected. Detecting patients in real time or at least soon enough is very difficult, especially if there are not enough resources at your disposition. Also they already have a lot to deal with epidemics-wise…

The use of drugs if the last aspect of infectious disease we will cover. It has been shown that many antibiotics are used on patients that are not infected, which leads to both a poor attribution of often scarce resources and a rampant rise in antibiotic resistance in microbes. This is exacerbated by the fact that doctors tend to err on the side of caution, meaning that if an antibiotic is available for a patient, even if the patient has a very slight chance of having a disease, it will be prescribed. Drug resistance has become a clear issue in the 21st century and leads to most medications becoming ineffective. When these become useless against a disease, the alternatives are few and more expensive. Moreover, alternatives are not infinite and the pool of antibiotics to choose from shrinks considerably with time.

\paragraph{Example: Malaria}

*[****CDC Barriers to Developing a Malaria Vaccine]***

* No vaccine available: not much financial incentives / market and also generally hard to develop (is a parasite not a virus)
* Why tough to fight: you can be infected for months and only have symptoms months after! Plus you never really get full immunity after being infected
* Current solutions: you can get antimalarial drugs: will 1) make sure it doesn’t get worse 2) stop you from spreading elsewhere
* Diagnosis: symptoms are mostly shared with other diseases: fever lol. The best way to solve this: use tests (rapid diagnostic test RDT), but obviously if you don’t have money / personnel it is harder to do that. In these same countries: malaria is not the only problem, you can’t shift your whole attention and means of production / money towards malaria!
* Other infectious diseases in malaria-endemic world: pneumonia, diarrhea, tuberculosis, HIV/AIDS
* Different types of malaria: uncomplicated and severe. Uncomplicated: fever, chills, sweats, headache, muscle pains, nausea, vomiting. Severe: confusion, coma, focal neurologic signs, severe anemia, respiratory difficulties. For mild malaria, you can stay at home (outpatient) but if severe: require hospitalization.
* There are different types of malaria because different parasites. Most severe usually caused by plasmodium falciparum
* It used to be: if you have fever and under 5 years old: malaria medication. But what if it wasn’t malaria….now: more RDTs and more personnel. Now WHO (2010) says any treatment should be confirmed beforehand with RDT.
* RDTs: antigen detection. Quick results (about 20 ish minutes). Easier than using a microscope! Still need microscopy sometimes, why: if the infection is still sort of beginning (low number of parasites) the RDT might not get the positive results. So should have both if possible (one and then the other…. But yea then why need RDTs in the first place? Then microscopes would detect false negatives … still though). Not all RDTs are equal, some do better. Transport and storage can be tricky… high temperature and high humidity don’t do them any favors. Still too expensive if you want to test an entire population…
* Drug resistance: two parasites already resistant: plasmodium falciparum and p vivax to some drugs (all?)

\paragraph{e-CDAs, e-POCTs}

Making a diagnosis based on the symptoms of the patient is difficult in a limited amount of time. This problem is currently addressed in part by the WHO. A guideline (IMCI) has been written and put in place in order to give some information to clinicians on possible diagnostics. Since these guidelines are static, they are constantly outdated. Moreover, they are very broad and general, and since disease usually have a specific geographic and climate based context, are often not optimal.

IMCI guidelines fall in the CDA category, along with many different and new algorithms.

In an effort to palliate these issues, electric CDAs have been put in place. However, most of these are still static and not geography-specific.

While the use of e-CDAs is a step forward, there is ample room for improvement with the use of ML, data science and data collection. A new technique, e-POCT has been put forward, with the general idea being to collect data during point-of-care tests and to use it to gather information not only on a patient but on a population and its evolution with time. This provides much needed granularity of population analyses in both the geographic and the temporal dimensions.

\section{ Medalai}

The goal of the Medalai project is to help the fight against infectious diseases by providing an accessible method of data-analysis.

For a specific disease, any clinician can gather basic knowledge on how to act from IMCI guidelines, and some context knowledge from years of expertise in their own clinics or from their collaborators. The idea of Medalai is to allow clinicians to gain more insight on their patients and the context that is specific to their location.

The overarching idea is to use data science to address clinicians’ needs. Machine learning algorithms will be used to predict diagnostics at a certain threshold – based on the accuracy of other algorithms already put in place, or based on the resource consumption of other algorithms – or even improve overall prediction accuracy. A parallel goal is to give insight to the end-user on ways to improve the accuracy, specificity or another chosen metric of results.

\section{Diagnosis using Machine Learning}

\paragraph{Data sets}

There is uncertainty in the data collected. Many values are often missing, values are not normalized or standardized across clinics or regions, some features useful to the ML prediction are disregarded when conducting point-of-care testing, etc… Data pre-processing also adds a potential source of errors. There is also general uncertainty in the data itself – meaning the process that is being measured, which is called aleatoric uncertainty. While uncertainty in the data collection process can be improved, aleatoric uncertainty is inherent to the data itself and is informative of the differences in the data. In general, data is dirty and data scientists have to deal with that.

Datasets usually contain different types of data (categorical / numerical / binary) which poses a problem. What is the best way to decide if a feature is categorical or numerical? A threshold can be placed on unique values to decide, but the automatization of this can be tricky…

The data sets used usually are collections of different data sources, for example different clinics or hospitals. There are many different features, and not all are harmonized during data collection.

If all features are included in the model training, the results might become confusing (more than a hundred features with a good number of them containing more than acceptable levels of missing values. Therefore the platform has already implemented a mechanism to take out some of the features without even trying to include them in the model training. To take this one step further, the platform can also apply feature reduction techniques such as applying Lasso or Boruta. Lasso uses an L1 penalty factor on feature coefficients during the training process. Cross-validation can be used to choose this parameter in an optimized fashion. Since the platform is not using optimization for now, the best method would be to use Boruta. Boruta effectively tests a feature column’s effect to the overall model result against the same column with values mixed up within. In the latter case the model is trained using a feature column that is “randomized”, and the model results will show a decrease in metric scores if that feature was actually important to the model prediction.

After having discussed the dataset specificities, one of the conclusions to be drawn regarding the manner to investigate the model explanations. For the end-user to analyze results, many features need to be discarded first. Missing value thresholds and general knowledge about the problem at hand should be the two leading criteria to make this cut.

\paragraph{Machine Learning Task}

There is uncertainty that is specific to the model being used which does not stem from the data, which is called epistemic uncertainty. This uncertainty should be reduced as much as possible.

There is also uncertainty in the results.

This uncertainty can be mitigated using confidence intervals on the statistics outputted, or using interpretation and explanation methods that are suited to the model and data type. Confidence intervals are paramount in domains such as medical diagnosis, since most classification predictions are probability-based and should be considered with their associated probability. Using bootstrapping can yield robust estimates of a population’s parameter of interest.

it is now important to discuss the machine learning problem at hand. The main task appears to be disease diagnostics. This is a classification type problem. The advantage of disease diagnostic is that there are diseases that can be diagnosed with high efficacy point-of-care tests. Therefore, the dataset’s labels are accurate at an acceptable threshold, which is not always the case in medical data sets. An important point to mention is that when dealing with not just binary classification, but multi-class labeling, new problems might arise. For a binary outcome; does a patient have a disease or not, the data is either pushing towards the presence of a disease or towards the absence of it. Different diseases have specificities, and symptoms that make them identifiable. However, many have symptoms in common. If a model has to choose between five different diseases, with the possibility of a patient potentially having multiple diseases, the model training becomes complicated. Interpretation tools an help verify if the model is using the “known” symptoms of a disease for its classification.

The mathematical logic behind some models should make it possible to solve this problem, however a model is only good as its data. This is where the real issue is. In the case where more than one disease needs to be diagnosed, a dataset with about 3000 data points (patients), which already took some time to be collected, will not be enough – especially if some portion of the dataset contains missing values. Since different geographic locations have different specificities, as discussed previously, it can be hard to merge data sets together to solve this issue. An emphasis should be put on data collection. Class imbalance is the other issue, especially for a disease like malaria which has different variants.

\paragraph{End Users}

The testing phase of the platform is mainly done by users with a strong background in statistics, computer science, or even machine learning. Even when trying to rely only on the platform’s functionalities, it is hard not to include personal knowledge and previous experience into play. This is why it is paramount to show the platform to end-users with a weaker background in those domains, to make sure that it is still accessible to people with less intuition on what to do and what results actually mean. This is one of the hardest issues to reconcile within the platform. The previous points call for as many functionalities and detail as possible, while the overall goal of the platform calls for simplicity and ease-of-use. This is why the interpretation methods should follow a logical flow.

***Move this to end of background***

*\paragraph{Thesis Description}*

*This thesis aims to better the interpretability of Machine Learning (ML) models – including black box models – in order to generate results and behaviors that are more easily understood by clinicians and therefore generalize, widen and give more credibility to the use of ML models in the medical field (especially in the precise field of medical diagnostic).*

*Interpretability of ML models will be explored at different levels; at the data level, at the model level and at the output level and will be done through different interpretability and visualization techniques (exploring probabilistic space of the model, intelligent descriptive statistics, confidence intervals for results).*

*This thesis will also provide a validation test of the Medalai platform through the use of a dataset pertaining to Ebola infection cases in Tanzania. The models used will be carefully analyzed in order to find insights on a list of important (non-exhaustive) questions) and therefore validate the platform. The models and results themselves will be validated and bench-marked using another set of models and results.*

*The Medalai platform is a web-based platform that allows users to upload tabular data in order to output clinical decision algorithms.*

*The end goal is for clinicians, the end-users, to be able to give criticism concerning the platform and alter it to evolve in that direction.*

***Move this to OUR specific dataset presentation (methods: dataset):***

*One of the possible labels to detect is the patients being malaria positive or negative. There are many missing values in the dataset and only a subset of features make sense to be included in the detection of malaria. One example is that of the geographical location of the clinic visit. In theory, this information would be greatly helpful to detect cases of disease spreading. However, the platform for now does not seek to detect evolution of diseases over time. Moreover, the number of missing values is too high to be able to recover a sub-dataset that is large enough for network training.*