I would like to thank my father who always supported me in everything I do.

Contents

Chapter 1

Introduction

This final paper aims to illustrate machine learning for medicine as both an enhancement and a potential solution to racial and gender discrimination in the process of medical diagnosis. To do so it bases itself on current literature and an empirical study. It follows a structure, which is described in the following sentences.

The Theory chapter introduces the reader to the advances of machine learning and the ethical issues that surround it. It uses definitions and examples to describe and understand the presence of demographic biases in machine learning models. Moreover, it suggests different approaches to reduce these disparities, most of which are based on already existing studies. A more in-depth look of these issues is taken in the use of machine learning in medicine and, specifically, in the process of medical diagnosis. Lastly, the Theory chapter presents the research questions, intended as a metric to evaluate the results of the paper.

The Methods chapter aims to enrich the paper on detection and reduction of biases in machine learning for medicine by introducing the reader to an empirical study - an algorithm predicting the occurrence of acute liver failure of a patient. It provides a strategy to develop a well-performing model for medical diagnosis and detect the potential representation biases. Moreover, it suggests a possible approach to reduce them. To achieve these goals it provides a framework with tools to answer the previously stated research questions.

The Data section presents the results of the empirical study before and after the attempt to mitigate potential representation biases. These insights are evaluated and interpreted in the Discussion section, where the research questions are answered and the severity of the observed discrimination is discussed. Moreover, this section transparently communicates the limitations of the study.

Lastly, in the Conclusions section, the results of the models are compared to the initial goals of the final paper and the current state of knowledge.

Chapter 2

Theory

2.1 Machine learning and its biases

Born from the idea that computers can improve their performance automatically through experience, machine learning is a subset of artificial intelligence, in which models are trained to make estimations and predictions based on what they learned from data. At the same time, it is also one of the most common buzzwords of the century.[1] The increase in computational power, together with the boom of huge amounts of valuable data of various types, enables the high-speed processing of datasets resulting in improved decision-making. Machine learning is the technology behind many current inventions such as Netflix recommendation systems, predictive maintenance algorithms, and, what will be discussed in-depth later, automated medical diagnosis.

The main distinction between different machine learning models is based on the characteristics of the data provided. It can be split into supervised and unsupervised learning models. Supervised learning is characterized by the prediction of targets based on labeled data fed to the machine. Instead, unsupervised models infer the inherent structure of the data without having labeled outputs.

The most common tasks in supervised learning are regression and classification. The difference between the two is that regression has a continuous/numerical dependent variable, while classification has a categorical one. In this article, the focus will be strictly on classification tasks. Their goal is to determine the value of a variable Y given the set of independent variables X. Two very common applications are the prediction of an event, where the occurrence is a binary dependent variable, and image classification, in which the algorithm determines whether a certain object/phenomena is contained in the image. An example of a successful method for image classification is convolutional neural networks(CNN), which use a kernel to slide over the input image, and in this way improve the generalization abilities, which were not as good when using a standard artificial neural network(ANN).

The models used to solve these tasks learn the outcome by approximating a function

that takes features as an input and minimizes a cost function to obtain the correct result. They are trained, in the sense of determining good values for all the weights and the bias from labeled examples.[2] In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss.

Loss is the penalty for a bad prediction. The loss is zero when the model's prediction is perfect. The more imperfect the predictions become, the higher the loss is. The result of successfully training a model is that the weights and biases that have the lowest loss on average are found. For example, Figure 2.1 presents two models: one with a high loss(on the left) and one with a low loss(on the right). The loss is represented by the arrows, while the blue line illustrates predictions.[2]

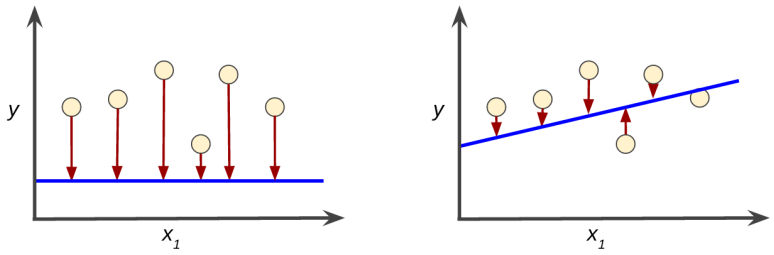


Figure 2.1: Loss Function

Different metrics capturing the ability of a model to classify observations correctly can be used to assess the performance. They are not always the same as the ones used to train the model. For example, one might use accuracy to give a score to the performance of a classification model, but use cross-entropy as a loss function to train it. Additional metrics capturing misclassification such as accuracy, sensitivity, specificity and precision can be added as a constraint on the optimization problem. The choice of a metric strictly depends on the context and aim of the model.

In the last decades, various researchers show that the performance of machine learning often exceeds the one of experts in various fields despite its lack of long working experience. The productivity of computers exceeds the one of humans. They manage to provide an otherwise unreachable level of analytical capabilities replacing intuition with data and statistics.[3]

2.1.1 Types of biases

With this widespread adoption of machine learning in our daily lives, it is inevitable to consider the ethical issues it may bring. Machine learning exceeds humans in its analytical abilities, but there are some aspects in which both sides fail. One of them is the avoidance of biases. When talking about disparities in our daily life we mean disproportionate weight in favor of or against an idea or thing, usually in a way that is

closed-minded, prejudicial, or unfair.[4] It can be innate or learned. Similarly, biases are also observed in machine learning, where they are defined as the phenomena of observing results that are systematically prejudiced due to faulty assumptions.[5] The two types of disparities have a very similar impact on society: the creation of socioeconomic and demographic inequalities. However, their sources are different. While for humans it is their own beliefs, in machine learning such events happen mostly due to "faulty assumptions", which are usually in one way or another linked to the human involvement in different activities throughout the technological processes.

Initially, this disparity can be split into two types, the first one being the dataset bias, and the second one being the representation bias, which is the one giving rise to unfair judgment learned by the model.[6] In this paper, the focus will be on the latter.

There are 3 main types of representation biases: inductive, unfair and legally prohibited. Inductive bias is fundamental for machine learning to sustain its performance. If it were not for it, the No Free Lunch Theorem states that all classifiers would give the same results in expectation like the ones reached by a random classification[7]. It is incurred in several ways - when the scientist chooses the cost function to be minimized during the training of the model, hence, setting the search stage. Else, it could come from the context, purpose, or level of adequacy of the dataset.[5] It becomes clear that the inductive bias in machine learning is needed for its functionality; hence, one should not aim to eliminate it, but rather be upfront about its presence.

However, the other two types of bias (the unfair and the illegal one) are not necessarily needed for the overall performance of the models, yet their presence could harm society. The most common source for their existence lays in the data reflecting historical examples, which are full of social disparities, stereotypes and wrongful beliefs towards certain social groups or minorities. Being fed by this data, the algorithm‘s only way of being effective and efficient is by mirroring these prejudice.

To consider bias as discriminatory and claim that it shall be legally prohibited, one has to first determine if it is justified and harmful.[5] For example, when Amazon launched their Prime Same Day Delivery service the company used an algorithm to determine which neighborhoods it should start the launch from.[8] In many big cities, it turned out that ZIP codes from areas with a majority of white population were much more likely to benefit from the upgrade. Nonetheless, Amazon argued that it does not consider race, ethnicity and other demographic characteristics in its model. Instead, they base their decisions on the density of people owning a prime account in the neighborhood. On this basis, even though harm might have been done, this bias could not be considered against law, and it remains only unfair.[8] On the other side, when the National Institute of Standards and Technology tested Facial Recognition Softwares it found that they produce a much higher rate of false positives for people of color, and even more so for Afro-American women.[9] This racial differential can

be very harmful, especially when such systems are used for Security Control, as many false accusations would follow for certain social groups. Therefore, this bias is often considered illegal, and the use of this technology by the police department has been banned at least temporarily in many big states like California and New Jersey.

2.1.2 Sources of bias in machine learning

To examine how socioeconomic and demographic disparities are adopted by algorithms this paper will analyze a simplified version of the cycle of machine learning, illustrated in Figure 2.2.[6] The focus will be on the potential sources of bias at each step.

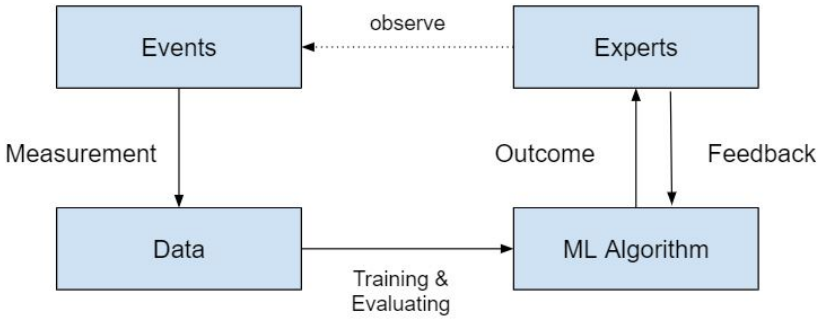


Figure 2.2: Life Cycle of machine learning

Measurement

Information is produced constantly from the everyday activities of people, businesses, organizations, and society at large. They reflect all of the prejudice, which still affect our world. This information is not very valuable to scientists in its raw form. The process of transforming it to fit into rows, columns and values, that allows for it to be further processed and analyzed is called measurement. In this stage, some assumptions may have to be made and some subjective decisions have to be taken. Such a choice is, for example, deciding how to structure demographic characteristics into categories.

A recent study authored by The New York Times titled “Even with Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago”, evaluates the evolution of socioeconomic disparities, showing that a lot of actions still have to be taken so that equal opportunity for minorities is reached.[10] The results clearly demonstrate that the increase in students from underrepresented groups is incomparable to the increase in the total share of population of the same segments. Nonetheless, the reliability of these results is not granted, as there are other aspects to be considered, such as the category used for the aim of the study. It turns out that the study lacks a “multiracial” category, which can be a cause for biased results. With the increase of children from mixed marriages and the lack of corresponding

categorization, more people are forced to choose a single race to identify with. It could be that many students who have a Hispanic and a White parent, identify themselves as White due to fear from prejudice. Hence, the imbalance stated in the article might be rather overestimated.[10] Similarly, with the increase of social acceptance of alternative genders, it could be that the standard division of gender into males and females is no longer sufficient.

Another important and challenging choice during the measurement phase is the one of a target variable. For example, if we were to choose being a "good student" as our binary dependent variable, we might then take into account the engagement in extracurricular activities offered by the university. At the same time, hypothetically, it could also be the case that many students coming from minorities attend the university thanks to a scholarship, and have to take a part-time job to cover their expenses due to their lower income. Hence, they could likely lack the time to engage in extracurricular activities. Moreover, the model might take into account a teacher's evaluation of good behavior, which, given the long history of discrimination, might be biased as well. The two described variables are examples of a potential proxies in the model. They are strongly correlated with the race characteristic of the student and account for it even if the demographic variable itself is not fed to the algorithm.

Other factors to consider during the collection of data are missing values, measurement errors, and imbalance of the sample. Missing data has many origins, it could be missing completely at random, missing at random, and missing not at random.[10] If values are missing completely at random, they are a random subset of all of the observations. If values are missing at random, they could be systematically different than the observed values, but the latter can be explained by other observed variables. All types affect negatively the performance of the model, but the ones that could become a reason for social disparities are the missing not at random. These types of missing values occur, for example, when one has to share very personal information, which one considers sensitive. This could be their sexuality, mental health state, relationship with their family, or even income. An example of missing values not at random would be if people with lower incomes were more reluctant to share it than people with higher income. The problem occurs in data processing, as the model cannot be fed with a dataset that includes missing information. Therefore, many techniques are used to estimate and fill these empty spots. A very common way to do it is by replacing them with the mean or mode of the population. The choice between the two mathematical transformations depends on the type of each variable. If a group, say the one with low income, is underrepresented, and the person who refuses to share their income belongs to this same group, then the missing information gets replaced by the population's mean, the imbalance becomes even stronger, and the underlying bias is further enhanced. A similar trend is observed in the treatment of outliers.

Training

In the next step the data potentially reflecting demographic bias is fed to the machine learning model, which learns from it. If no additional actions to reduce disparities are taken, the model will surely preserve and use them to reach a higher accuracy. The algorithm will extract harmful stereotypes in the very same way in which it extracts any other information.[6] One of the main aspects to consider in this step is the balance of the sample. It is clear that our society is not evenly separated into different social groups, but it is often the case that minorities are even more underrepresented in some studies than they are in reality. A good example are online surveys. Suppose that a supermarket wants to understand the price sensitivity of its customers. A very common shopper is represented by the mother in a family.[11] This type of customer is more likely to not own a smartphone or laptop, if she is coming from a low-income family[12]; thus, she is less likely to fill in the survey. There are two problems with this. Firstly, the sample will not be representative of the population, and the analysis will fail. Secondly, the size of the underrepresented group will be so small, that even if all of the results for it were wrong, the overall performance of the model might still be high but biased. Therefore, if no further actions are taken, the model would generalize results based on majority groups only, which could be really harmful when the decisions that are taken based on the technology affect people's lives.

Once the model is trained it produces outcomes, which help experts in taking actions. If the predictions of the model are biased, the expert should be able to notice and understand it. However, this is rarely the case. Machine learning leverages association in the form of correlation, but not in the form of causality. As a result, the interpretability of outcomes is not straightforward.

Feedback loops

If an action is taken based on a biased prediction, it might reinforce the disparity forming a feedback loop.[6] If the previously mentioned prediction of being a "good student" is taken into account in the admission process of universities, then fewer people from minorities will be accepted, which would further underrepresent them, and would empower the already existing bias. The reinforcement of stereotype perpetuation and cultural denigration could be the main obstacle for the further adoption and development of machine learning, which could otherwise bring huge value to society. Therefore, we have to make the algorithms sensitive to such actions, and enable them to refuse to impose a disadvantage to certain demographic groups.

2.1.3 Reduction of bias in machine learning

Algorithms can be seen as the solution to the existence of representation bias. In statistics, a method called BLUE, short for best linear unbiased estimator already exists exactly for this purpose. Nevertheless, the reduction of such disparities in machine learning can be a very challenging task. The dataset imabalances reflect the difficulty experienced when trying to get a correct representation of the real world. Despite these limitations, there are some actions that can be taken to de-bias the models.

Deletion of representation feature

Often the first thing that comes to mind is the deletion of variables that contain demographic information. It seems like a very simple solution, and intuitively it should work. However, there is more to social biases than just the variables representing gender, race, or belonging to any other group. Our society has such a strong history of prejudices that their effects are much broader. Even if we were to eliminate the exact variables, the dataset contains a large number of proxies some of whose correlation with the omitted problematic characteristics are not obvious. A direct evidence for this are many AI-supported recruiting systems, which include independent variables like "good performance at the previous job"[6]. At first sight, this should not be an issue, but it proves to be one, as often the environment in companies is hostile towards people of color and women. In conclusion, even though intuitive such a solution is not effective.

Constraint on fairness criteria

Instead, the first step in the reduction of representation biases should be their clear identification and examination. The first thing we could do in the case of binary classification is to examine whether the outcome for each data point changes when we change the sensitive demographic characteristic such as race or gender. If it does, then we are talking about disparate treatment. In this case, if one does not want these characteristics to influence the outcome, the usual solution is to omit this variable.[5] Instead, if building, for example, a predictor for the likelihood of a person becoming a victim of crime, then the inclusion of these features is still needed.

Otherwise, our model might be subject to Disparate Mistreatment, which is observed when the model performs worse for a certain social group compared to the other(s). This depends very much on which metric is being used in the training and evaluation of the model. Very often accuracy is not the best criterion, as the different types of misclassification might have different levels of harmfulness. For example, when using Facial Recognition for Security Systems it would be much worse to imprison an innocent person, at the same time leaving the real criminal free, than to just not identify the guilty.

Suppose that this setting is being examined. Then the fairness criterion should be the specificity score, which captures precisely the false positive classifications. This metric should be compared for each racial and gender groups. Many such systems show that even if the accuracy score is similar for the different categories, the specificity observed on the Afro-American segment is much lower, showcasing a typical representation bias. Mathematically, what we will be examining here would be whether:

$$P(output_{predicted} != output_{actual} | output_{actual} = 0, race = white) == P(output_{predicted} != output_{actual} | output_{actual} = 0, race = black)$$

Krishna Gummadi, studied the controversial recidivism prediction tool called COMPAS to address such issues.[6] As a result of his research, he understood that adding a constraint to the optimization on the false negative and false positive rate, he managed to achieve similar levels of misclassification for the two racial groups at the cost of a small reduction in accuracy. He showed that the difference between false positive rate for white and black people approached zero, while only 0.8% of the accuracy dropped.

Balancing the sample sizes of social groups

In other situations the issue cannot be solved only through the introduction of an additional constraint. For example, when working with datasets in which one of the groups is significantly underrepresented, one might want to balance the distribution first. This delicate step in the process can be handled through oversampling, which is the increase in the number of samples of a minority class to match the number of samples in the majority class. If it is performed before training the model, the results could be unrealistically optimistic. A solution is to balance the training set during the cross validation. Cross validation is used to assess the generalization properties of machine learning models and discard models that overfit. This is done by splitting the data randomly into training and testing samples. The training set is used to estimate the parameters, while the test set is used to validate the performance. The prediction error of the estimated model is computed based on the comparison between the predicted and the actual values of the test sample. Similarly, oversampling should be performed through the splitting into training and validation folds, on each of which the transformation is applied. Of course, the fitting is performed only on the training set.

Even if we do manage to rid our systems of the above-mentioned biases, there is still potential for bias to appear over time. Once algorithms are trained, perform well and are put into production, it is important to continuously monitor how the algorithm is updated. It might be good to have someone who is responsible for ensuring that two years after the adoption the system is still doing well, and can determine what it means for the system to be doing well.[5]

2.2 Machine learning in medicine

2.2.1 Medical diagnosis

Medical diagnosis is the attempt of explaining which disease or condition is the reason for a patient's symptoms and health state.[13] It is a crucial process in the correct treatment of a patient. To understand its complexity it is helpful to look at a cyclical model provided by the National Academies of Sciences, Engineering and Medicine in 2015, which illustrates the reliance of physicians' decision-making on the correct information gathering, and the centrality of the patient.[13]

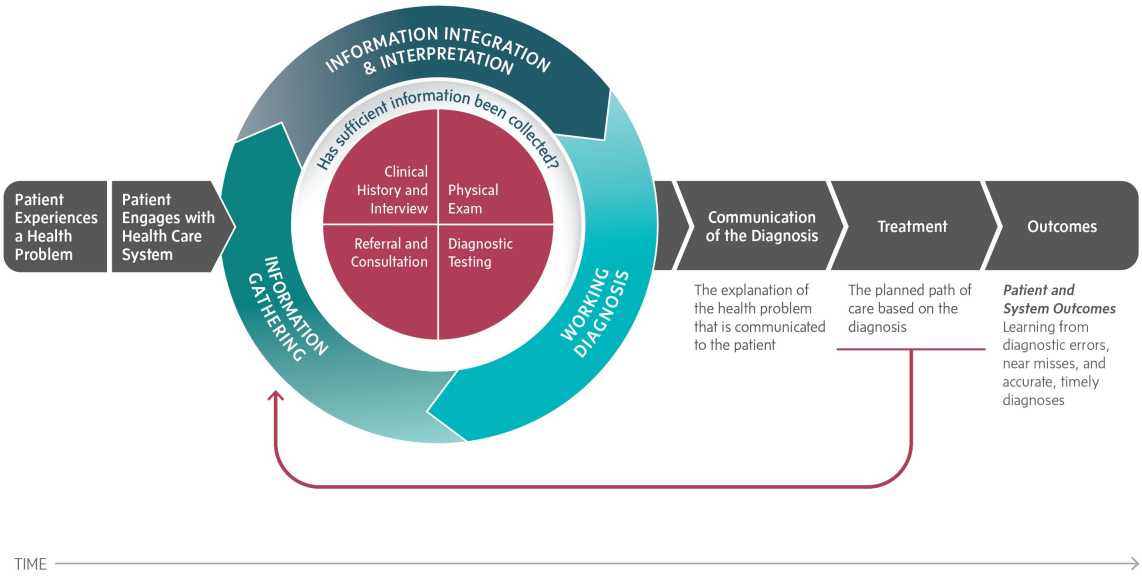


Figure 2.3: Cyclical Model of Medical Diagnosis

The length of the timeline of this process depends on many factors and can vary in every case. The first step is the occurrence of a problem in the health of a patient, followed by him/her seeking health care assistance. After this, the diagnostic cycle begins, and the physician starts to gather information, looking at the patient's health history, conducting physical exams and diagnostic tests, and might consult with fellow physicians to interpret correctly the information and be able to formulate a working diagnosis. The continuous process of information gathering, integration, and interpretation involves hypothesis generation and updating prior probabilities as more information is learned.[13] During this stage, it is also very important to always evaluate whether the information collected is sufficient to accept or refuse a certain diagnosis, and with what confidence level. After the loop has been exited, the diagnosis is communicated with the patient, who then receives a treatment plan, the outcome of which should be continuously tracked by the physician.

Looking at the diagnostic cycle, one can understand how delicate it is and how easily a mistake can be done. A human error in the detection of a disease can result in delayed

testing and treatment, leading to potential complications of the overall health of the patient. The World Health Organization states that "Diagnostic errors are relatively common in primary care and most people will likely experience a diagnostic error in their lifetime." Moreover, evidence has shown that tests showing abnormal results are not acted upon in the case of cancer identification, which contributes to 39% of the diagnosis delays.[14] Delays in the treatment of this class of diseases are both harmful and costly.

Diagnosis errors and biases

Many of the factors leading to diagnosis errors are connected to human mistake: poor communication, low availability of specialists, lack of learning and feedback from errors, poor teamwork, suboptimal trainings and distraction are all among the common reasons for wrongful diagnosis. Despite the technological advances and the popularity of topics regarding diversity, one of the key reasons for misdiagnosis and mistreatment remain the existence of socioeconomic disparities.

Historically, many of the medications, which were invented a long time ago, and are still used today without being re-evaluated, have been tested only on men. For example, even though 70% of people who suffer from chronic pain are women, 80% of the pain medications have been tested exclusively on men.[15] In Western Medicine women are often said to have a "complicated biology" and, as a result, are frequently not even considered in trials.[15] Hence, the effects of drugs on approximately half of the population are less known.

Moreover, racial biases are frequently observed in the decisions of physicians throughout the diagnostic process. On one hand, the problem is that the established symptoms and trails are based mainly on the study of samples of the white population. As a result, it has been observed that black women are much more likely to experience serious complications during childbirth compared to white women.[15] This event is very closely linked to the racial prejudice that black women are able to resist on higher levels of pain.

On the other hand, as Tobias Baes explains in his book "Understand, Manage and Prevent Algorithmic Bias", it is also connected to the cognitive biases observed among people, who in this context are represented by doctors. As a psychologist and data scientist, Tobias explains that the human brain works thanks to these disparities, as they provide shortcuts for people just like they do for algorithms. According to him, there are three competing objectives in human nature: accuracy, speed and energy efficiency. Since conscious thinking is much more consuming than the effortless subconscious reasoning, the brain often relies on the latter in the decision-making process. This is an efficient characteristic, but it introduces all kinds of disparities. Similarly, physicians often find themselves in situations where they have to make an accurate diagnosis for a

very limited time, as the condition of the patient might be getting worse every second. At this moment, the already created by the subconscious part of the reasoning shortcuts come in handy and result in actions being taken on the basis of unfair prejudice.

2.2.2 Machine learning in medical diagnosis and its biases

Machine learning algorithms play an essential role in diagnosing diseases, recommending treatments and ensuring the engagement and adherence of the patient. They differ from traditional medical technologies, as they have the ability to gain, process and interpret information based on which they are able to make faster decisions at just a fraction of the cost. The output of the trained algorithm is easily interpretable by humans, even the ones not having medical education, which allows for the complete automation of the technology in the well-known form of websites and mobile applications. Moreover, machine learning is crucial in the medical imaging field, enabling computer-assisted diagnosis, image segmentation, image registration, image fusion, image-guided therapy, image annotation, and image database retrieval.[16]

Medical diagnosis is the process of determining the nature of a disease or disorder and distinguishing it from other possible conditions.[17] Hence, statistically speaking, the diagnosis of a condition includes running a set of classification tests with binary dependent variables. This is one of the simplest machine learning problems. It considers assigning an individual to one of two categories, by measuring a series of attributes.[18] In this case, the two categories will be disease vs. no disease. The goal of the classification is to learn a function in such a manner as to minimize the occurrence of misclassified data points. Mathematically, it aims to minimize the probability $P(y*F(x) < 0)$, where y is the class label taking value -1 when the result is negative and + 1 when the result is positive. Commonly used algorithms to solve a problem of this essence include K-Nearest Neighbors, Logistic Regression, Multilayer Perceptron, ensemble methods such as Boosting and Random Forest and some Deep Learning methods based on Convolutional Neural Networks. The latter is usually used when leveraging computer vision and working with huge unstructured datasets.

Performance

Due to the mathematical essence of medical diagnosis, many algorithms proved to have a better performance than the majority of experienced physicians. Thanks to medical imaging our society enjoys an improved detection of colonic polyps, cerebral microblading, diabetic retinopathy and many other conditions.[16] The models also help predicting in-hospital mortality and prolonged length of stay more accurately, improving the overall management of the institution, which directly affects the wellbeing of individuals.

A research done by Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau and Sebastian Thrun titled "Dermatologist-level classification of skin cancer with deep neural networks", examines the potential automation of classification of the most common human malignancy.[19] They used deep convolutional neural networks and a dataset of 129,450 clinical images consisting of 2,032 different diseases to compare the performance of their algorithm in classifying the two most common and the two most critical types of skin cancer to the one of 21 board-certified dermatologists. The CNN performed the diagnosis better than 2 of them, while on average its results were on par with the ones of the remaining 19 experts. The research concludes that Artificial Intelligence shows dermatologist-level capability in predicting skin cancer. One might wonder what is so revolutionary in these outcomes. Not only do machines take just a small fraction of the usually incurred cost, but they also make the method deployable on a mobile app, where a user can upload a picture and understand whether he/she should take further treatment actions. The broadening of accessibility of primary care, in this case, is very impactful because the early detection of skin cancer is critical. As the research states, the 5-year estimated survival rate for melanoma drops from over 99% if detected in an early stage to approximately 14% if detected later. Obtaining a benefit from the use of machine learning in medical diagnosis is clearly realistic.

However, it comes to attention that this research did not examine the potential representation biases, which are very common in computer vision algorithms, as in the previously mentioned Facial Recognition Systems, which tend to have a bias discriminating both black and female people, the combination of which is the group with the highest misclassification rate. Given these facts, it is not irrational to question the existence of such disparities in the above-described Skin Cancer study. If such a phenomenon is observed and not acted upon, this could prevent the implementation of an otherwise beneficial for society technology.

Biases in machine learning for medicine

One of the promises of these algorithms is the avoidance of biases such as gender or racial inequalities, which are often observed in traditional healthcare, and can have serious consequences like misdiagnosis and mistreatment. If in good hands, machine learning algorithms used in healthcare can be a powerful weapon to fight these disparities. However, if scientists do not work on the identification and reduction of representation bias, the same algorithms can enhance and contribute to the existing social disparities in medicine, creating a self-fulfilling prophecy.

An article by the Washington Post titled "Racial bias in a medical algorithm favors white patients over sicker black patients" provides insights regarding the unfairness of an algorithm used by numerous hospitals to support them in deciding which patients

would benefit most from additional care.[20] It turns out that the sickest black patients' needs are undervalued. The researchers estimated that if they were to eliminate the disparity, this would double the number of black people who would get extra care. The most interesting aspect of this study is that the algorithm did not even include a race variable, however, there was a proxy, a variable showing how much the patient would cost to the health care system in the future, which was strongly correlated with the demographic feature.[20] In this case, the solution to the problem of reducing the bias was quite intuitive: instead of predicting future cost to the health care system, the scientists predicted the severity of future health conditions of the patients. This gives hope for the capability of machine learning to detect and reduce bias in a measurable and actionable way, something which is not so easily implementable in reality, as the discrimination can often be rather intangible.

Reduction of biases in machine learning for medical diagnosis

In some other cases, however, the reduction of bias in machine learning for medical diagnosis is not so intuitive. Literature suggests that one way to do it is by implementing an interdisciplinary approach and ensuring the continuous involvement of physicians to conduct follow up studies and ensure the meaningfulness of results. It suggests that doctors should be supported by these algorithms, so they should be properly trained to understand, interpret, and detect potential biases in the technology. This is also the "hybrid approach" discussed by Tobias Baer, according to which there should be a balance between algorithm and human judgment.

One possibility to reduce bias would be to build and test models in environments that provide the required socioeconomic diversity to avoid unbalanced datasets, in which one or more of the segments is underrepresented. They could ensure correct data engineering, splitting the data into test and train set according to their target population. They could also keep into account that this algorithm is going to be fed by new data created by the facility in which the technology is deployed. Hence, feedback loops can be implemented to continuously monitor the outputs, their validity and the potential introduction of disparities, even if the initial model did not show any such issues. To enable these processes, scientists could include representative features such as gender and race in their dataset when appropriate. When biases are detected similar measures to the ones discussed in the subsection "Machine learning and its biases" can be applied.

2.3 Research questions

This paper studies the topic of bias in machine learning for medicine, because its controversy goes beyond the question of rightfulness, ethics and law. In this application

of the technology, society is no longer concerned only about the justifiability of the existence of a disparity. Instead, the main concern regards the correct treatment and the well-being of entire communities, nations, races and genders. As already mentioned above, traditional medicine failed to abstract from such prejudice. Hence, it is in the hands of machine learning to detect, reduce, and maybe even eliminate biases to ensure an equally accurate and accessible treatment for all people around the globe. To attempt to prove the potential of this solution, an empirical study introduced in the next section will answer the following research questions in an exemplary case:

1. Are biases present in medical diagnosis algorithms?
2. How can one detect representation biases in machine learning models for medicine?
3. What is the reason for the existence of representation biases in machine learning algorithms for medicine?
4. How can representation biases in machine learning for medicine be reduced?

Chapter 3

Methods

3.1 Model design

This section aims to enrich the paper on detection and reduction of biases in machine learning for medicine by introducing the reader to an empirical study - an algorithm predicting the occurrence of acute liver failure of a patient. It provides a strategy to develop a well-performing model for medical diagnosis, detect the potential representation biases and choose an approach to reduce them. Figure 3.1 illustrates the methods used to find an answer to each of the research questions introduced in the Theory section of this paper.

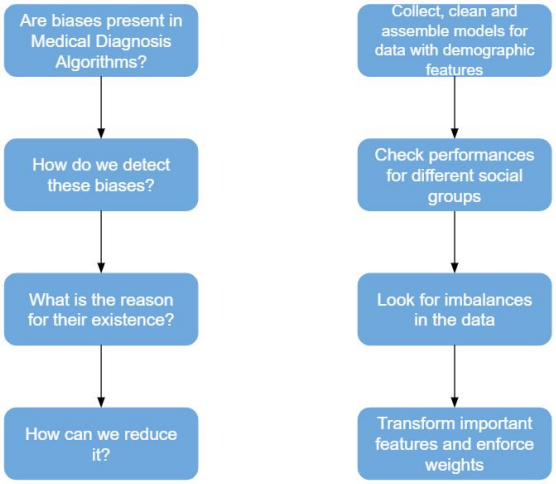


Figure 3.1: Methods to answer research questions

3.1.1 Gender bias in acute liver failure prediction

Acute liver failure (ALF) is the term frequently applied as a generic expression to describe patients presenting with or developing an acute episode of liver dysfunction, referring to a highly specific and rare syndrome. This means an acute abnormality of

liver blood tests in an individual without underlying chronic liver disease.[21] ALF is less common than chronic liver failure, which develops more slowly. This disease occurs rapidly, develops only within days or weeks, and causes serious complications, being life-threatening with a death rate of 40%. Therefore, the potential implementation of a machine learning technology to predict the risk of patients being diagnosed with this condition can have a huge positive impact for society, since measures to prevent it can be taken only at early stages.

As an input, this model will take a mixture of features such as ones accounting for the history of diseases of the patients and their family, their physical characteristics like weight, height, age and gender. The latter independent variable suggests that the aim of this algorithm will not only be to accurately diagnose the disease, but also to inspect for the potential presence of gender bias in the process. As mentioned before, the problem can be represented as a binary classification task. In our case, the binary dependent variable will be "ALF", which is equal to 1 when the condition was observed and 0 when it was not.

Data collection and dataset description

This paper uses a dataset already available on Kaggle called "Acute Liver Failure: To predict the liver failure over demographic variables". The initial data consists of 8785 data points representing different adults who are citizens of India and participated in the surveys conducted by the JPAC Center for Health Diagnosis and Control in the years 2008-2009 and 2014-2015. The institution used direct interviews, examinations, and blood samples to gather a demographically diverse dataset, with a total of 30 features and a 7.73% Acute Liver Failure ratio.[22]

The features can be divided into three main groups: pathologies, demographical data and physical features and habits. In the first segment, there are characteristics such as the dependent variable(ALF), obesity, dyslipidemia, peripheral vascular disease, poor vision, hypertension(individual's or in the family), diabetes(individual's or in the family), hepatitis(individual's or in the family) and chronic fatigue. The presence of these features enables the analysis of a correlation of other diseases with the examined condition, which could provide physicians with more specific risk groups. The demographic variables consist of age, gender, region, marital status, income, education and source of care. This group is crucial for the study, as it not only enables the creation of risk groups and takes into account relevant factors, but also will be the basis on which the detection and reduction representation biases will be built. The last group of features provides an overview of the lifestyle and overall health condition of the patients. It consists of weight, height, BMI, waist, cholesterol (good, bad and total), physical activity and alcohol consumption. The sensitivity analysis of these can not only indicate the predisposition of certain groups but can also be used for a correct treatment recommendation. Moreover,

the variables can be split according to their types, being binary, categorical or numerical. The following figures describe them according to this separation method and the specific feature characteristics each group brings.

Variable	Description	% of 1 out of non-missing values	Missing Values
Obesity	1 if BMI is greater than 30	**31.15%**	206
Education	Years of education of the subject: 1 if above the threshold	**43.66%**	15
Unmarried	Marital status: 1 if unmarried	**37.50%**	301
Income	Income of the subject: 1 if above the threshold	**42.09%**	792
Alcohol consumption	1 if the subject consumes alcohol	**30.41%**	0
Dyslipidemia	Abnormal amount of lipids in the blood: 1 if above the threshold	**10.58%**	0
PVD	Peripheral Vascular Disease, 1 if the subject is affected	**3.77%**	0
Poor Vision	1 if the subject has Poor Vision	**5.99%**	376
Hypertension / Family HT	1 if the subject has High Blood Pressure / 1 if a family member has the condition	**39.92% / 23.02%**	53 / 0
Diabetes / Family DB	1 if the subject has Diabetes / 1 if a family member has the condition	**11.11% / 31.07%**	1 / 0
Hepatitis / Family HP	1 if the subject has Hepatitis / 1 if a family member has the condition	**6.51% / 2.02%**	13 / 3
Chronic Fatigue	1 if the subject has Chronic Fatigue Syndrome / 1 if a family member has the condition	**2.83%**	26

Figure 3.2: Binary Features

Variable	Description	N° categories	Missing Values
Region	Region of Asia from which the patient comes from: North, South, West, East	**4**	0
Physical Activity	Categories defining the degree of physical activity of the patient: 1, 2, 3, 4	**4**	8
Source of care	Type of consulted facility: Private Hospital, Clinic, Never Consulted, Government Hospital	**4**	0
Gender	Gender of the subject: Male, Female	**2**	0

Figure 3.3: Categorical Features

Variable	Description	Mean	Variance	Min Value	Max Value	Missing Values
Max Blood Pressure	Systolic Blood Pressure, measured in mmHg	125.57	21.07	72.00	233.30	206
Min Blood Pressure	Diastolic Blood Pressure, measured in mmHg	71.55	12.45	10.00	132.00	252
Good Cholesterol	HDL (high-density lipoprotein, measured in mg/dL)	51.82	15.82	8.00	160.00	8
Bad Cholesterol	LDL (low-density lipoprotein, measured in mg/dL)	152.26	42.60	33.00	560.00	8
Total Cholesterol	LDL + HDL Cholesterol	204.08	42.27	75.00	606.00	6
Age	Age of the subject, the only discrete numerical variable	49.15	18.81	20.00	85	0
Weight	Weight of the subject, measured in Kg	79.06	19.57	33.70	193.30	133
Height	Height of the subject, measured in cm	67.11	10.15	130.40	200.10	139
BMI	Body Mass Index: Weight in kg divided by Height in m squared	28.24	6.20	14.42	66.44	206
Waist	Waist of the subject, measured in cm	96.79	15.20	58.60	173.4	215

Figure 3.4: Numerical Features

Data cleaning

It immediately comes to attention that this dataset contains a large number of missing values. In fact, initially, only 4322 of the data points have all the information available for each feature. Out of these missing values, there are 2875, which miss also the dependent variable ALF. In this study they will be omitted from the training set, as they are not too valuable in the supervised learning model. An alternative approach would be not to omit them and to use them for estimating means/medians for missing variables. This would leave a total of 6000 observations. The variables containing more than 10% omitted values, will be dropped, which is relevant only for the "Income" feature. After the creation of dummies for Source of Care, Gender, Region, Physical Activity, and the deletion of one category to avoid the well-known dummy trap, the dataset will have a total of 34 variables.

For pathologies, habits and demographic factors, which are not of primary interest, a replacement of the missing values with the mean or mode will be reasonable. However, when it comes to variables that could be correlated with representative characteristics, such as in this case gender, a different approach will be needed, as to not further enhance potential biases. Therefore, to deal with the missing values of height, weight and waist, it is reasonable to only drop the observations missing the value for at least two out of

the three features, while the other data points' missing values will be predicted with a regression. Each of the mentioned variables' missing values are predicted using the set of features except from ALF as independent variables. Since body mass index and obesity can be mathematically derived from weight and height, and are in the same time not linearly correlated with the latter, the newly predicted values for weight and height can be used to replace the gaps in these features. Body mass index is a person's weight in kilograms over their height in meters squared. Instead, obesity is a binary feature set equal to one when a person's body mass index is greater than 30, and zero otherwise. After this treatment the dataset will contain 5331 observations with no missing values out of the initial 6000.

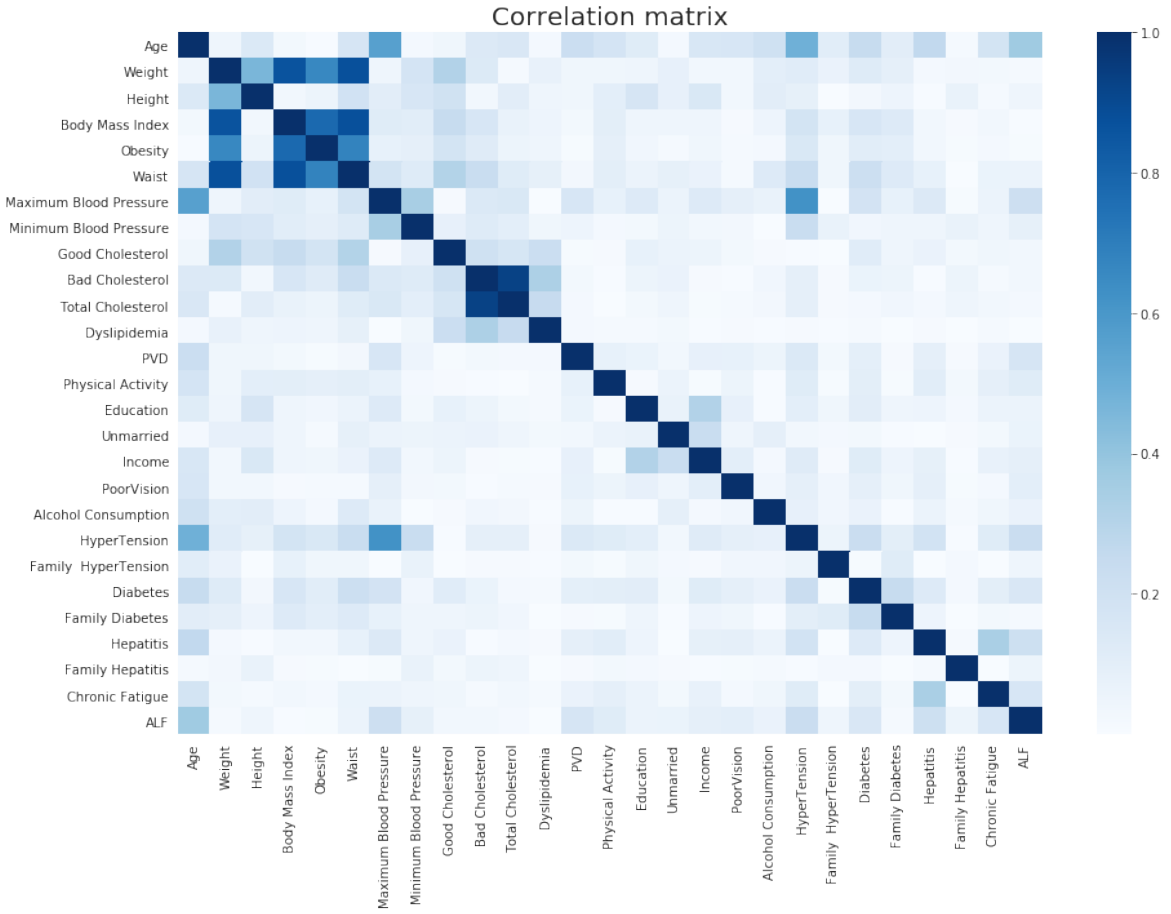


Figure 3.5: Correlation Matrix

As the correlation between total and bad cholesterol is around 93%, it will be required that total cholesterol is dropped, letting bad and good cholesterol control for it. One can see that Hypertension is very correlated with both age and maximum blood pressure, but since the relationship is not linear and is easily explainable from a medical point of view, and the above-discussed features are also the ones which are most associated with the target variable, dropping them is not necessary. Interestingly, even though the causes of acute liver failure remain unknown for nearly 15% of patients, its high positive correlation with vascular diseases (hypertension, maximum blood pressure

and PVD) and hepatitis are consistent with medical records describing them as possible causes of the disease of interest.

The dataset proves to contain a significant number of outliers and some variables skewed to the right like weight, waist, body mass index and others. The presence of outliers has an effect on the mean and standard deviation, as well as the accuracy and overall performance of the models. Moreover, they can become a reason for the existence of bias. Hence, this paper uses Isolation Forest to remove them.[23] Isolation Forest is different from other popular outlier detection methods because it explicitly identifies anomalies instead of profiling normal data points. Like any tree ensemble method, it is built on the basis of decision trees. In these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature. The path length represents the number of splittings that an algorithm makes for a certain instance. Intuitively, an outlier will have a shorter path length. As with other outlier detection methods, an anomaly score is required for decision making. In the case of Isolation Forest, it is defined as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where h(x) is the path length of an observation x, c(n) is the average path length of unsuccessful search in a Binary Search Tree and n is the number of external nodes.[23]

Model assembly

As mentioned earlier, medical diagnosis is a classification problem, which in the case of a single disease has a binary dependent variable. For this reason, this paper will use the following models to achieve an accurate prediction:

- Logistic Regression: as the name suggests, it uses a logistic function to predict a binary dependent variable
- Polynomial Logistic Regression: to run a Logistic Regression with polynomial features, as a result, capturing nonlinear features
- Random Forest Classification: an ensemble technique, which uses multiple decision trees at a training time, and provides the mode of predictions as an output
- Gradient Boosting Classification: again an ensemble technique, which leverages Decision Trees, but allows for the improvement of each sequential Tree; In this way, it transforms weak separate predictions into a strong ensemble classifier
- K Nearest Neighbors: a non-parametric classifier, which assigns to the observation the class corresponding to the most commonly observed one of its k neighbors

- Voting Classifier: a model which trains on the ensemble of numerous models, in this case, the ones mentioned before

For the optimization of the hyperparameters, GridSearchCV maximizing the sensitivity is utilized. The latter score can be derived from the confusion matrix, also known as an error matrix, which helps visualize the performance of the classification. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class(or vice versa).[24] An example of a confusion matrix is shown in Figure 3.6. In this way, the True Positives (the number of correctly predicted observations of class 1), False Positives (the number of incorrectly predicted observations of class 0), False Negatives (the number of incorrectly predicted observations of class 1) and True Negatives (the number of correctly predicted observations of class 0) rates are derived. The metrics commonly used in machine learning to evaluate models, are derived from these rates in the manner shown in Figure 3.7.

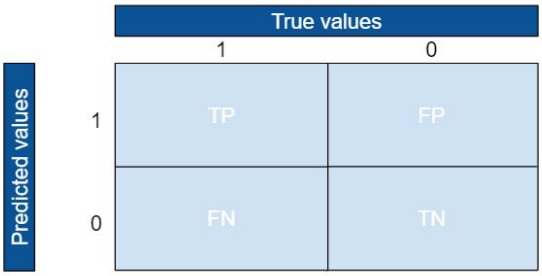


Figure 3.6: Confusion Matrix

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Misclassification rate (1 – Accuracy)	$\frac{FP+FN}{TP+TN+FP+FN}$
Sensitivity (or Recall)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision (or Positive Predictive Value)	$\frac{TP}{TP+FP}$

Figure 3.7: Performance Metrics

Checking performance for different social groups

When it comes to medical diagnosis, it is most harmful for the patients if they are told that they do not have a disease, which they actually do, as it would reduce their chances of survival significantly. This means that a minimization of false negative classifications is needed. Such an aim can be achieved equivalently by the optimization of sensitivity. Hence, in this paper the standard objective functions is used when doing the fitting,

but then sensitivity is maximized during the optimization of the hyperparameters of each model. Moreover, sensitivity will be considered as a fairness criterion, and the detection and reduction of gender bias will be done according to it.

To validate the performance of the model a prediction of the classes of the data points of the test set will be compared to their actual values. The importance of sensitivity, which this study will aim to maximize, should not be at the cost of a strong decrease of accuracy; thus, monitoring the values of the latter will be reasonable.

Moreover, the aim of this prediction is to not only perform well on the entire population but also to not present any biases towards one of the genders. Hence, the dataset will be split into "Male" and "Female" subsets to enable observing the performance on them separately. However, the models will not be fitted on the subsets, as to illustrate the bias one would incur when ignoring such issues. If there is a significant difference in the sensitivity of the two groups(using a minimum of 5% as a rule of thumb), further actions will be taken to understand the reasons for this phenomenon, and how it can be mitigated.

Checking for imbalances in the data

A problem observed in the dataset is its imbalance in terms of acute liver failure occurrences. As mentioned before, the failure ratio is approximately 8%, which would make the overall accuracy too optimistic, while the recall score(sensitivity), which is the most important metric in medicine might be too low, if no other action is taken.

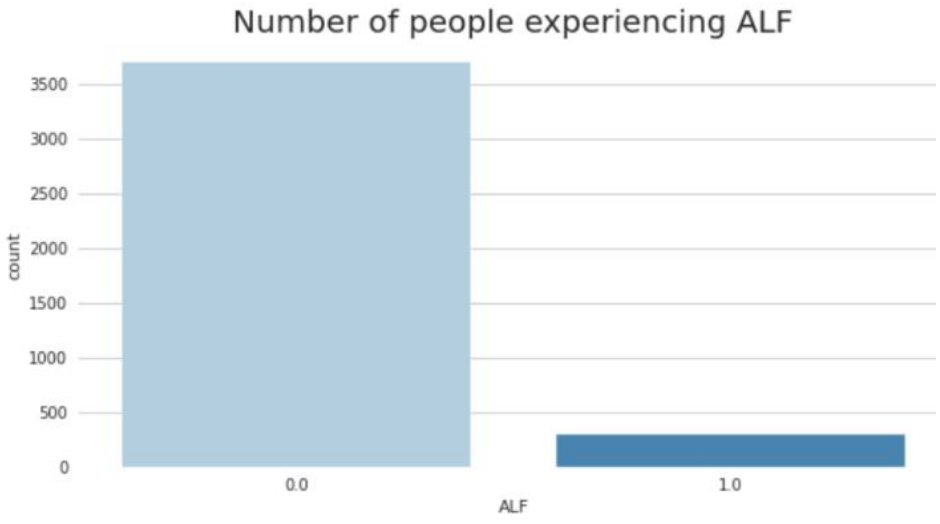


Figure 3.8: ALF Distribution

For this reason, this paper will use the approach of upsampling the underrepresented class during the cross validation to avoid overfitting. Using the imblearn pipeline, the data will be split into training and validation folds, oversampling the minority class in each fold. The classifier is trained on the training folds and validated on the remaining

folds. One of the benefits is that the pipelines are compatible with GridSearchCV, so looping over parameters manually will not be required.[25]

Moreover, the female patients represent approximately 47% of the dataset after dropping the rows missing an ALF value and have a lower percentage of people experiencing ALF than men do. This could become the basis for a representation bias. To further evaluate it a feature importance and a sensitivity analysis will be performed. In the latter, the focus will be on the odds ratio of the gender variable.

Transforming important variables and enforcing weights

If such a social disparity in performance is observed, actions need to be taken to reduce it. To do so this paper takes several approaches. Firstly, it evaluates whether the most significant features obtained through a feature importance analysis are differently distributed for men and women. Interestingly, Age, which is the most significant positively correlated with the dependent variable characteristic, is higher on average for women than for men. Hence, in the case of disparity between the genders, this variable can be transformed as a remedy. On such occasions, age can be standardized separately for the two social groups by subtracting the mean of both and dividing by their standard deviations. In this way, both genders' ages would have the same mean (0) and the same variance (1).

The second approach consists of adjusting the sample_weight parameter in the fitting of the best performing on average models, which still show unfair treatment towards one of the genders. Intuitively, the data points representing the discriminated category are given a larger weight in the training, the optimization of the hyperparameters and the evaluation of results.

Chapter 4

Data

4.1 Results of the models

Initially, a set of random classifiers are run, and it is observed that the sensitivity and the accuracy vary between 40-50%. After the models are trained and the hyperparameters are optimized, the results in Table 4.1 are observed.

Metric/Model	LogReg	Poly LogReg	Random Forest	XGB	KNN	Voting CLF
Accuracy	77.5%	77.5%	81.9%	83.1%	76.4%	81.6%
Sensitivity	87.1%	87.1%	82.3%	79.0%	85.5%	82.3%
Specificity	76.9%	76.9%	81.9%	83.3%	76.1%	81.6%

Table 4.1: Performance of Models on entire set

The feature importance analysis, which leverages the Random Forest algorithm to understand the effect of the independent variables on the dependent variable, shows that the 10 most important features are the ones described in Figure 4.1.

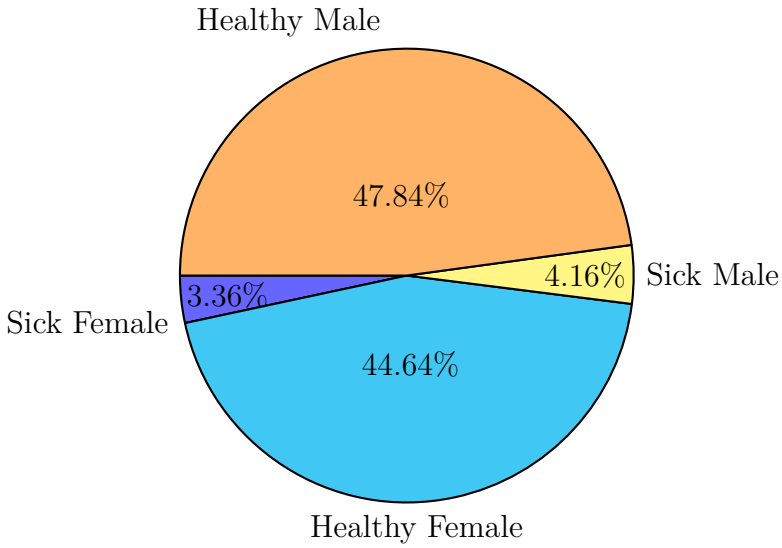
features	importance
Age	0.577046
HyperTension	0.205239
Region_south	0.074758
Waist	0.029181
Family HyperTension	0.027990
Hepatitis	0.022001
Physical Activity_3.0	0.011041
Source of Care_Private Hospital	0.010023
Maximum Blood Pressure	0.008431
Region_north	0.007058

Figure 4.1: Feature Importance

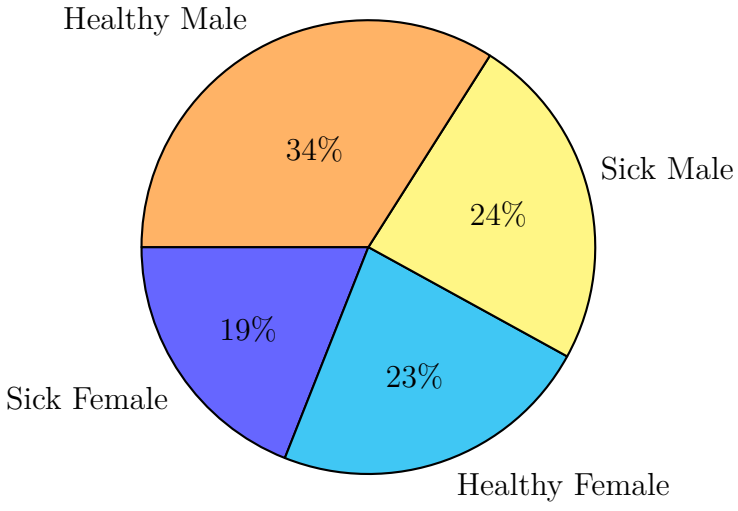
4.2 Bias detection

To discuss demographic bias, one has to first examine the representativeness of the characteristic in the dataset and its distribution according to the dependent variable.

Before proceeding with the upsampling of positive cases of acute liver failure, the gender representation is the one illustrated in the pie chart below.



SMOTE is the technique used by this paper to oversample the underrepresented group. It is short for synthetic minority oversampling technique. It uses the k nearest neighbors of each observation of the underrepresented group to link them through lines. It then generates all of the new data points across these lines. After upsampling has been performed using a SMOTE technique, the distribution becomes different.



After the models are trained the dataset is separated into a female and a male subset to predict the outcomes on these two groups, which allows to observe the performance of the models for them separately. The following scores are obtained:

Metric/Model	LogReg	Random Forest	XGB	KNN	Voting CLF
Accuracy	78.8%	82.0%	84.3%	78.2%	82.5%
Sensitivity	80.0%	77.1%	77.1%	82.3%	77.1%
Specificity	78.7%	82.3%	84.7%	77.9%	82.8%

Table 4.2: Performance of Models for Male Segment

Metric/Model	LogReg	Random Forest	XGB	KNN	Voting CLF
Accuracy	76.1%	81.8%	81.6%	74.8%	80.7%
Sensitivity	96.3%	88.9%	81.5%	88.9%	88.9%
Specificity	75.0%	81.4%	81.6%	74.0%	80.3%

Table 4.3: Performance of Models for Female Segment

The results of Logistic Regression are derived from the confusion matrices in Figure 4.2. Moreover, leveraging the properties of the this model, one can understand that the odds ratio of a man being diagnosed with acute liver failure to a woman being diagnosed is around 1.23.

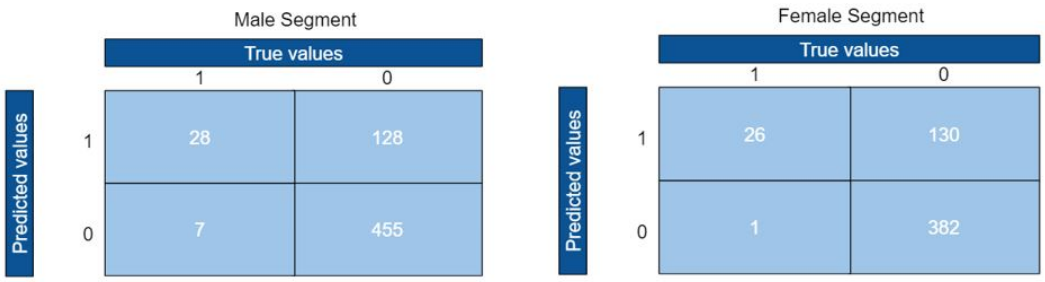


Figure 4.2: Confusion Matrices per Gender

4.3 Bias reduction

Bias reduction is model specific. To perform it this paper focuses on two models - K Nearest Neighbors and Logistic Regression. Following the first approach of transformation of the most important features, which differ among the two genders, a standardization of the age is performed separately for the two genders. The changes in performance resulting from this transformation on the entire dataset are illustrated in Table 4.4, while the ones on the two gender subsets are presented in Table 4.5 and Table 4.6.

Metric/Model	LogReg	KNN
Accuracy	77.8%	75.7%
Sensitivity	86.6%	85%
Specificity	77.3%	75.2%

Table 4.4: Results on the entire dataset with age standardized separately

Metric/Model	LogReg	KNN
Accuracy	79.9%	77.3%
Sensitivity	78.1%	81.3%
Specificity	80%	77.1%

Table 4.5: Results on the male subset with age standardized separately

Metric/Model	LogReg	KNN
Accuracy	75.4%	73.5%
Sensitivity	96.4%	92.9%
Specificity	74.2%	72.4%

Table 4.6: Results on the female subset with age standardized separately

In the second approach to bias reduction, age is standardized as the rest of the continuous variables but each observation is assigned a weight. There are four groups of weights based on a patient's gender and occurrence of acute liver failure. The weight is larger than one for the data points belonging to the discriminated groups(the more discriminated a group is, the larger the importance given to it), and smaller than one for the groups that enjoy better performance of the models. They are used both in the confusion matrix computation and in the GridsearchCV hyperparameter optimization. Four sets of sample weights that satisfy this logic are chosen through a randomized approach. With the weights 1.8 for sick men, 0.96 for healthy men, 1.2 for sick women, 0.64 for healthy women, the results are:

Metric/Sample	All	Female	Male
Accuracy	54.7%	53.4%	55.8%
Sensitivity	98.3%	100.0%	98.9%
Specificity	52.5%	51.2%	53.6%

Table 4.7: Result with sample weights (Example1)

With the weights 1.2 for sick men, 0.64 for healthy men, 1.8 for sick women, 0.96 for healthy women, the results are:

Metric/Sample	All	Female	Male
Accuracy	55.3%	54.8%	55.8%
Sensitivity	100.0%	100.0%	100.0%
Specificity	53.2%	52.8%	53.4%

Table 4.8: Result with sample weights (Example2)

With the weights 2 for sick men, 0.9 for healthy men, 1.32 for sick women, 1.2 for healthy women, the results are:

Metric/Sample	All	Female	Male
Accuracy	62.2%	64.9%	60.3%
Sensitivity	98.2%	100.0%	96.8%
Specificity	60.5%	62.8%	58.3%

Table 4.9: Result with sample weights (Example3)

With the weights 1.9 for sick men, 0.9 for healthy men, 0.84 for sick women, 1.2 for healthy women, the results are:

Metric/Sample	All	Female	Male
Accuracy	70.2%	73.3%	67%
Sensitivity	91.1%	92.3%	90.6%
Specificity	69.2%	72.4%	65.8%

Table 4.10: Result with sample weights (Example3)

The odds ratio of a man being diagnosed with acute liver failure to a woman being diagnosed after the introduction of the last set of sample weights drops to around 1.16.

Chapter 5

Discussion

This section aims to evaluate the machine learning models used for the prediction of an acute liver failure, to determine whether a representation bias exists and what its sources are, and to evaluate the methods for reduction of bias, choosing the best one.

5.1 Models evaluation

This subsection evaluates the results obtained by fitting six classification models. The best models are chosen by balancing the trade-off between maximizing sensitivity and not compromising accuracy.

All of the models' hyperparameters have been optimized in a manner to maximize sensitivity. The results on the entire dataset in Table 4.1 will be used to evaluate their performance. As mentioned in the Methods section, the aim is to maximize sensitivity but not at the cost of decreasing accuracy too much. While none of the models presents an accuracy that is too low, none of them meets the desired sensitivity level either.

Logistic Regression and Polynomial Logistic Regression present the same results and can be discussed together, since during the GridSearch optimization of hyperparameters the polynomial degree was set to 1. This makes them equivalent. Moreover, they are the ones with the highest sensitivity, followed closely by the K Nearest Neighbors classifier. All of the models based on tree methods including Random Forest and Gradient Boosting perform significantly worse. The ensemble method does not manage to improve the performance above the one of the single models. Therefore, this paper focuses primarily on Logistic Regression and K Nearest Neighbors to further optimize performance, and evaluate and reduce potential representation biases.

5.2 Bias existence and sources

This subsection aims to answer the research questions regarding the existence and the reasons for the existence of representation bias in machine learning for medicine.

The empirical study of acute liver failure prediction demonstrates clear and significantly large gender bias in favor of the female subset. For Logistic Regression, the disparity amounts to approximately 16.3% difference in sensitivity. In K Nearest Neighbors the model results in a 6% higher sensitivity score for women than for men. The same ratio is 11% higher for women using the Random Forest classifier and the Voting classifier, and 4% higher for the same social group according to Gradient Boosting.

The direction of the bias is unexpected because theory suggests that the underrepresented segment is most commonly the one that is harmed by a disparity. Instead, in this study women represent less than half of the dataset. Moreover, men can be seen as a risk group for two reasons. Firstly, the training set shows a higher density of "sick" male patients than female ones. As a result, during the oversampling of ALF more male data points are generated, and the difference in the percentage of ALF positiveness of the two genders becomes even larger than before. Secondly, the odds ratio, which represents the probability of men suffering from an acute liver failure over the probability of the same event for women, is greater than one, confirming the predisposition of men to the disease. Logically, this should increase the number of true positives of the male segment, and, since true positives are part of the nominator of sensitivity, it should be higher as well. The nonobviousness of the source of the bias is not surprising, because, as mentioned previously, such social inequalities are often subtle.

Nonetheless, age, which, as the feature importance analysis in Figure 4.1 shows, is the most significant positively correlated with ALF variable, is on average 1.6 years higher for women than for men. Hence, according to the distribution of age in the dataset, women are more likely to get acute liver failure.

From the two confusion matrices of Logistic Regression in Figure 4.2, which correspond to the two genders' results, one can understand that the source of the bias is not the true positive rate of men being lower, but the false negative rate being higher. It could be that the algorithms make a generalization of men being "sicker" on average; thus, predicting better for the positive class, and not as well for the negative one, resulting in an increased level of false negatives, which, in turn, decreases the overall sensitivity of the gender.

5.3 Bias reduction

This subsection aims to answer the research question regarding the reduction of representation biases in machine learning for medicine. As mentioned in the Data section, the bias reduction is model specific and has been attempted only for K Nearest Neighbors and Logistic Regression in this paper. The reason is that they present the highest overall performance and have a significant gender bias.

The first approach to mitigating inequalities in performance consisted of the stan-

dardization of age separately for the two genders. As mentioned before, the distribution of this variable across genders is different and could be contributing to the poorer performance of the models for men compared to women. Nonetheless, the results show the transformation of age reduces the overall sensitivity of Logistic Regression by 0.5% and increases the bias of the same model by 2%. The same tendency is observed with K Nearest Neighbors - the sensitivity drops by 0.5%, while the bias rises by 5%. Hence, the first approach proved to have exactly the opposite of the desired effect and will not be applied.

The second approach, based on the adjustment of sample weights, is performed on the Logistic Regression model only, as the K Nearest Neighbors classifier does not support it. The Data section presents four different sets of weights and the results from their implementation both on the entire dataset, as well as on the two gender subsets. All of them put the highest importance on the sick males as a technique to optimize true positive classifications, which is a common tool to minimize false negative misclassifications(known to cause the difference of sensitivities across the genders).

With the first set of weights, an increase of 11% in overall sensitivity is accompanied by a reduction of the bias by more than 13%. However, their implementation leads also to a decrease of around 23% in accuracy. The negative effect of this method on accuracy is worrisome because its values become very close to the ones of the random classifier(40-50%). This suggests that another solution to the bias mitigation should be found.

The second set of weights completely eliminates the bias in sensitivity, which increases to 100%, and has accuracy slightly higher than the one achieved by the application of the first set of weights. However, this accuracy is still very low and comparable with the one of the random classifier. It becomes clear that one might be willing to allow a certain percentage of bias to improve the overall accuracy.

The third set of weights achieves this goal. The bias is reduced from 16.3% to approximately 3.2% in favor of women. The increase in sensitivity from the initial models is 11%, while the decrease in accuracy is 15%, which is 8% less than with the other sets of weights. The accuracy is now higher than the one achieved by a random classifier. Nonetheless, the drop in accuracy is still quite big and suggests that there is place for improvement of the solution.

The last set of weights has a different effect on the performance of the model. It causes the smallest decrease in accuracy(5.3%) and still manage to reduce the bias (from 16.3% to 1.7%). To achieve this it increases the overall sensitivity by 4% but it also decrease the sensitivity for the female segment by 4%. Nonetheless, it proves to have the best trade off between mitigating the bias, maximizing sensitivity and not decreasing accuracy too much; thus, it is the one that this paper suggests as a final solution.

5.4 Type of observed representation bias

This subsection aims to classify the observed representation bias according to the three types described in the Theory section(productive, unfair and illegal).

If the models prior to the mitigation of bias are used, the disparity they would cause would unfair, as it would prevent male patients from getting an early treatment in approximately 16% of the cases, while the same event would occur for women only in around 3% of the cases. This also proves that the disparity would be harmful, which is one of the criteria to be able to legally prohibit the application of the models. However, as usual, the justifiability criterion is much more ambiguous, and can be satisfied only in the case, in which no actions are taken to reduce the inequality and the demographic variable is included in the dataset.

After the mitigation of bias, the remaining disparity according to the last set of weights is 1.7% in favor of women. I would classify it as productive, because accuracy decreases by 14.9% when the bias is reduced to zero(using the second set of sample weights).

5.5 Limitations

While all of the research questions were answered successfully and an overall high sensitivity of 91.1% was achieved, the desired level of accuracy was not reached, especially in three out of the four proposed solutions. In the Methods section, it was specified that the latter metric shall not decrease too much after the mitigation of the bias, but the second best solution lowers it to 62.24%, which is too little compared to the previous results(77.5%). The chosen solution does a bit better but still reduces the accuracy by 5.3%. Thus, in future research it would be beneficial to find a solution, which mitigates the bias without decreasing the accuracy.

As mentioned in the Theory section, feedback loops should be implemented to continuously monitor the outputs, their validity and the potential introduction of disparities even when the initial model did not show any such issues. Such a solution is not provided in this paper, as it is strictly dependent on the environment in which the models are deployed.

The empirical study of acute liver failure dealt with the imbalance of the data, but the results of the study could become even more representative of reality, if data is gathered in a better way, achieving a natural balance between sick and healthy patients, and between the two genders and their "sickness".

Chapter 6

Conclusions

The aim of this paper was to determine whether representation biases exists in machine learning for medicine, why they exists, and how they can be reduced. To do so, it leveraged current literature and an empirical study consisting of the prediction of acute liver failure with a focus on the potential differences in its performance for the two genders.

Literature suggests that representation bias is present in algorithms, as it is in the judgment of people. On one hand, this final paper confirms the existence of such disparities in machine learning for medicine through an exemplary case, in which there is a significant bias with an ambiguous source and a surprising direction. On the other hand, it also proves that the enhancement of such prejudice by machines is preventable, as the inequality in performance for different social groups can be mitigated with the help of technological tools.

The solution offered is the one of giving a weight to each data point based on their demographic characteristic of interest and their dependent variable value. The weights should be given according to the fairness and optimization criterion, which, in the case of medicine, is sensitivity. It is important to note that, differently by common beliefs, it is not the most underrepresented group that should necessarily be given highest importance but rather the discriminated one. In the study, this group is the one of sick men. The weights given to data points according to their dependent variable value can be determined by the type of values optimized by the chosen score. In sensitivity this is the true positives; thus, patients with acute liver failure should be given more importance. This method manages to not only mitigate the bias but also improves the overall sensitivity.

Unfortunately, these results are achieved at the cost of the reduction of other performance metrics such as accuracy. The chosen solution is the one that harms this score the least but also manages to mitigate the discrimination. The rest of the proposed solutions have sensitivity close to 100% but it does not mean that they perform well. For example, one can also get trivially to 100% sensitivity when classifying everything

as positive but the resulting model would not be reliable. The final solution has an overall sensitivity of 91.1%, a 1.7% bias benefiting women and reduces accuracy by 5.3%(much less than the other proposed solutions).

Because of the limitations of this study it might be best if a hybrid approach is used in the deployment of this technology. According to it, there should be a balance between algorithm and human judgment. The ways to develop it depend on the different applications of the models. If embraced by a Private Clinique, the models could be used only to determine whether further examinations should be done to confirm the diagnosis rather than whether a treatment should be started. Moreover, physicians should be trained to work side-by-side with the algorithm to achieve optimal use of the technology.

Instead, if the algorithm is deployed on an app, it might be best to be transparent with the users about the performance of the model. They should be aware that, while there is a low probability that they are told to be healthy, but are actually not, there still is a significant possibility that they are told to be sick, but are actually healthy. Hence, the algorithm shall be used as a mechanism to suggest visiting a doctor or scheduling further examinations.

Bibliography

[1] SAS Institute, *Machine Learning: What it is and why it matters, https://www.sas.com/en_us/insights/analytics/machine-learning.html* [last access: 22/06/20].

[2] Google Developers. *Descending into ML: Training and Loss | Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss*[last access: 25/06/20].

[3] Sergeenkov, A. (2020, March 23), *Artificial Intelligence is Becoming Better than Human Expertise, https://hackernoon.com/artificial-intelligence-is-becoming-better-than-human-expertise-16903f4fc3c0* [last access: 22/06/20].

[4] Welsh, M. & Begg, S. (2016), *What have we learned? Insights from a decade of bias research. https://www.publish.csiro.au/aj/AJ15032* [last access: 24/06/20]

[5] Jaspreet. (2019, April 7), *Understanding and Reducing Bias in Machine Learning, https://towardsdatascience.com/understanding-and-reducing-bias-in-machine-learning-6565e23900ac* [last access: 22/06/20].

[6] Barocas, S., Hardt, M., & Narayanan, A. (2019), *Fairness and machine learning, https://fairmlbook.org/* [last access: 22/06/20].

[7] D. H. Wolpert & W. G. Macready, (Dec. 2005) *"Coevolutionary free lunches," in IEEE Transactions on Evolutionary Computation, vol. 9, no. 6, pp. 721-735, doi: 10.1109/TEVC.2005.856205* [last access: 25/06/20] .

[8] Ingold, D., & Soper, S. (2016, April 21), *Amazon Doesn't Consider the Race of Its Customers. Should It?, https://www.bloomberg.com/graphics/2016-amazon-same-day/* [last access: 22/06/20].

[9] NIST. (2020, January 9), *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software, https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software* [last access: 22/06/20].

[10] Ashkenas, J., Park, H., & Pearce, A. (2017, August 24) *Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than*

35 Years Ago, https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html [last access: 22/06/20].

[11] Ball, J., Balogh, E., & Miller, B. T. (2015), *Improving diagnosis in health care, Washington, DC: The National Academies Press* [last access: 22/06/20].

[12] PEW RESEARCH CENTER, (2019, April), *Demographics of Mobile Device Ownership and Adoption in the United States, https://www.pewresearch.org/internet/fact-sheet/mobile/* [last access: 24/06/20].

[13] Ball, J., Balogh, E., & Miller, B. T. (2015), *Improving diagnosis in health care, Washington, DC: The National Academies Press* [last access: 22/06/20].

[14] Pat Croskerry, K. C. (2017, September 19), *Diagnosis: Interpreting the Shadows, https://www.taylorfrancis.com/books/e/9781315116334* [last access: 22/06/20].

[15] Pley, C., Dhatt, R., & Keeling, A. (2019, September 11), *Gender bias in Health AI - prejudicing health outcomes (or getting it right!), https://www.womeningh.org/single-post/2019/09/11/Gender-bias-in-Health-AI—prejudicing-health-outcomes-or-getting-it-right* [last access: 22/06/20].

[16] Zhou, L., Wang, L., Wang, Q., & Shi, Y. (2015), *Machine learning in medical imaging: 6th International Workshop, Mlmi 2011, held in conjunction with Miccai 2015, Munich, Germany, October 5, 2015: proceedings. Cham: Springer* [last access: 22/06/20].

[17] Rakel, R. (2018, November 23), *Diagnosis, https://www.britannica.com/science/diagnosis* [last access: 25/06/20].

[18] Parmigiani, G. (2001). *Decision Theory: Bayesian, International Encyclopedia of the Social & Behavioral Sciences, https://www.sciencedirect.com/topics/computer-science/binary-classification* [last access: 25/06/20].

[19] Esteva, A., Kuprel, B., Novoa, R. a., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017, January 25), *Dermatologist-level classification of skin cancer with deep neural networks* [last access: 22/06/20].

[20] Johnson, C. Y. (2019, October 24), *Racial bias in a medical algorithm favors white patients over sicker black patients, https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/* [last access: 22/06/20].

[21] EASL-The Home of Hepatology. (2017), *Journal of Hepatology 2017 vol. 66, Acute Liver Failure Guidelines, https://easl.eu/publication/acute-liver-failure-guidelines/* [last access: 22/06/20].

[22] Kumar, R. (2018, September 15), *Acute Liver Failure, https://www.kaggle.com/rahul121/acute-liver-failure#ALF_Data.xlsx* [last access: 22/06/20].

[23] Lewinson, E. (2019, September 26), *Outlier Detection with Isolation Forest, https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e* [last access: 22/06/20].

[24] Powers, D. (2020, January), *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation* [last access: 25/06/20].

[25] Martin, D. (2019, May 20), *How to do cross-validation when upsampling data, https://kiwidamien.github.io/how-to-do-cross-validation-when-upsampling-data.html* [last access: 22/06/20].