uc3m | Universidad Carlos III de Madrid

Master's Degree in Computer Science and Technology
2022-2023

*Master's Thesis*

# Continual Adaptation of Machine Learning Models applied to a Healthcare Platform for the Detection of Tuberculosis

Simón E. Sánchez Viloria

Jesús Carretero PhD.
Lara Visuña Perez
Leganés, September 2023

# ABSTRACT

**Keywords:** tuberculosis, machine learning, data analysis, era4tb, healthcare

# DEDICATION

Special thanks to my advisor, for guiding me and supporting me at every step in the making of this thesis...

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

As health information becomes increasingly digitized, machine learning (ML) algorithms play a crucial role in deriving meaningful insights from complex, multi-modal data that is often difficult to interpret by humans. Conventional machine-learning approaches, however, may be inadequate in the face of rapidly evolving health technology, shifting patient needs, and the increasing availability of large amounts of uncertain and noisy data that is deemed too unreliable for use in clinical settings.

Modern techniques aim to address these limitations by refining models in a continual loop, prioritizing data acquisition, labeling, and feedback from experts to dynamically adapt to the data stream reliably. This work explores the potential of these techniques in the context of healthcare applications.

In particular, we incorporate the use of continual and active learning methods in the design of an end-to-end platform for the diagnosis of *tuberculosis* using computer vision techniques. Tuberculosis (TB) is an infectious pulmonary disease that has affected humankind for well over 4,000 years [1] and still affects millions of people worldwide. According to the World Health Organization, until the arrival of the Covid-19 pandemic, TB was the leading cause of widespread death from a single infectious agent, ranking even above HIV/AIDS [2].

The platform is meant to be integrated into a Healthcare and Data Portal that is being developed as part of an ongoing European project for the research of Tuberculosis, and it will serve as a tool to test and validate the use of machine-learning models for the diagnosis of Tuberculosis.

As such, in an effort to research the competence of continual adaptation methods in that context, the platform incorporates a system that is designed to continually adapt to new data as it becomes available and to ease labeling efforts by prioritizing the acquisition of the most informative data samples. Including a front-end interface for the visualization of the detection, labeling, and interaction with the models.

We evaluate the system's performance on a real dataset and compare it to a baseline model that doesn't use adaptation techniques. Our results show that the use of these methods outperforms the baseline in terms of robustness and sample efficiency while maintaining a similar level of accuracy on the test set. Furthermore, the system proposed can automatically adapt to new data and improve its performance over time.

These results demonstrate the potential of incorporating these techniques into the design of real-world systems for healthcare applications and other high-stake domains. The final chapter of this work includes a lengthy discussion about possible future work and improvements to the system that might facilitate its integration into other projects and applications.

### 1.1. Context and Motivation

This work is done in the context of the European Regimen Accelerator for Tuberculosis (ERA4TB). ERA4TB is a public-private initiative that started in 2020 and aims to create an open European platform to accelerate the development of new treatment regimens for tuberculosis (TB). The project is integrated by over 31 organizations from the European Union and the United States, including academic institutions, research centers, non-profit organizations, and other public and private entities [3].

ERA4TB's mission statement aligns itself with (and is in response to) the United Nation's (UN) Sustainable Development Goals (SGD). The project's website reads 'The goal of ERA4TB is to deliver an innovative and differentiated combination regimen for the treatment of TB, which can play a key role in the TB elimination agenda' [3]. This is in line with the UN's SGD target to end the TB epidemic by 2030 [4].

To address some of the challenges of TB drug development and clinical trial design, one of the project's objectives from deliverable 1.15 of ERA4TB's agenda is the development of a data-science-specific platform to enable the efficient use of machine learning methods from the collaborative platform. The platform is meant to be used by researchers and other project stakeholders to facilitate the use, development, and evaluation of machine-learning models that can aid in the research of TB.

Thus, the motivation behind this work comes from the idea of incorporating novel methods into this data-science-specific platform to support researchers and collaborators in their mission to end TB by 2030. The techniques described in this work are designed to improve the performance of supervised models while prioritizing their overall robustness and reliability, aspects that are crucial in the context of healthcare applications.

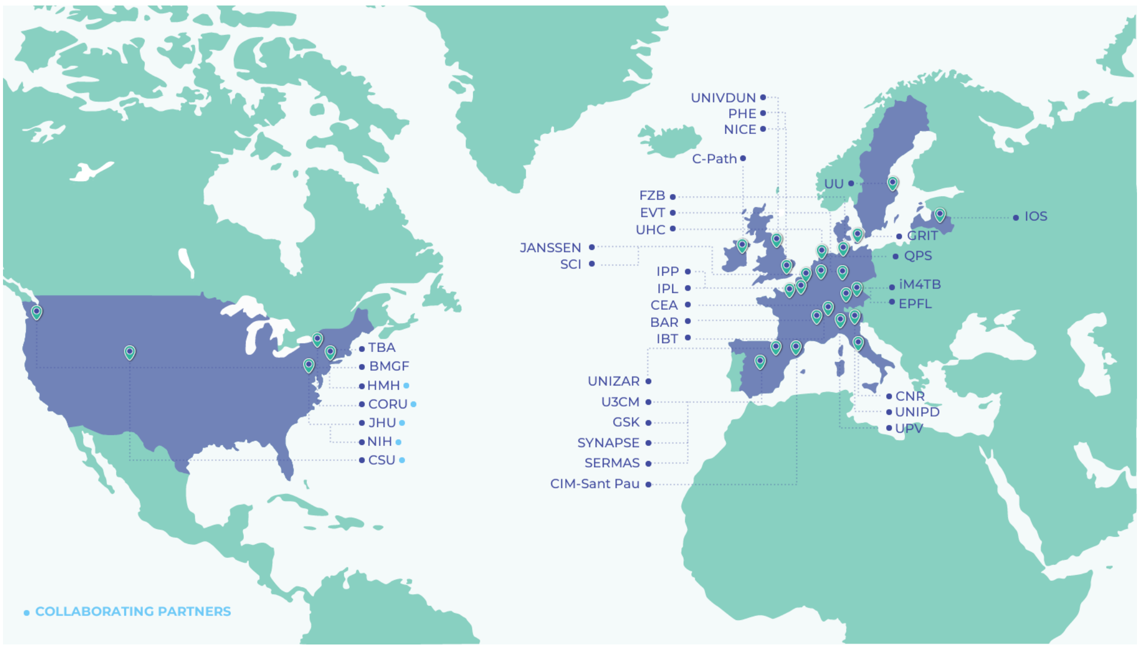


Fig. 1.1. ERA4TB's Consortium of Partners and Collaborators. Source: [3]

## 1.2. Objectives

### Main objective

The main objective of this work is to research the most relevant techniques on continual adaptation methods in Machine Learning, make a comparative analysis of the most relevant techniques, and their relevance for health applications, and design and implement a system for the diagnosis of tuberculosis that incorporates these techniques into its design that can be integrated into the ERA4TB platform. The platform should allow its users to incorporate test machine-learning models into the platform and facilitate the use and collaboration between researchers and other stakeholders.

### Specific objectives

Auxiliary to the main objective, the following specific objectives have been defined to guide the development of the work and evaluate its success:

1. The system should be capable of automatically triggering the continual learning process when new data is available or when the model's performance degrades based on a predefined metric.
2. The platform must implement a feedback loop between the data annotation process and model training that prioritizes the acquisition of the most informative data samples to improve the model's performance (Active Learning).
3. Develop a front-end interface that allows users to interact with the machine-learning models by selecting or submitting new data samples and visualizing the model's predictions.
4. Consider the limitations of the proposed system and the ethical implications of its use and present a well-founded outline of necessary future work to address these limitations or improve the system in a way that aligns with the project's mission statement.
5. Evaluate possible future research directions that could be explored in the area of continual and dynamic adaptation in Machine Learning, highlighting the contributions that have a higher potential for impact in healthcare or other high-stake domains.

## 1.3. Background Concepts

The following section provides a brief introduction to some of the concepts and techniques that are relevant to this work. It is meant to provide the reader with the necessary high-level information to understand the context and motivation behind the ideas proposed that were used to inform every design decision and experiment conducted. For a more in-depth overview of the same concepts, the reader is referred to the literature review in Chapter 2.

### 1.3.1. Supervised Machine-Learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that studies the design of algorithms that can learn from data and make predictions based on it. ML algorithms are fed a set of data samples (often called the training set) and learn a function that maps the input data to a desired output. The goal is to learn a function that can generalize well to unseen data and make accurate predictions.

Supervised learning is a machine-learning paradigm where the goal is to fit a function $f : x \rightarrow \hat{y}$ that maps a given input, $x$, to a 'prediction' output, $\hat{y}$, based on an available (finite) set of input-output pairs $(x_i, y_i)$ that are passed to a learnable model as 'training' data. The function is learned by minimizing a loss function $L(y, \hat{y})$ that measures the difference between the 'predicted' output and the actual one and returns a value that represents the error of the prediction and is then used to update the model's parameters.

The most common supervised learning tasks are classification and regression, where the goal is to predict discrete and continuous outputs, respectively. Conversely, the aforementioned loss function is often defined based on the task at hand and the type of data available (e.g., cross-entropy loss for classification tasks, mean squared error for regression tasks, etc.).

The presence of a known output (or label) for each input is what makes this paradigm 'supervised'. For example, in healthcare, supervised learning might be used to predict the presence of a disease or condition based on a set of features extracted (or learned) from the patient's data. A model can be trained to predict the presence or severity of a disease based on a set of symptoms, clinical history, or imaging data (like the case described in this work).

Besides its many benefits, the most significant disadvantage of supervised learning is that it often needs large amounts of labeled data to produce accurate and robust results [5], which is especially challenging in the healthcare field where the data is scarce, expensive to obtain, or difficult to label.

Approaches such as *semi-supervised learning* aims to address this issue by leveraging both labeled and unlabeled data to train the model. It often works by using unlabeled data to learn an intermediary representation of the data from which features can be extracted and used to train a supervised model.

### 1.3.2. Machine Learning for Clinical Diagnosis

### 1.3.3. Tuberculosis Treatment and Diagnosis

Tuberculosis is caused by the bacillus (bacteria) *Mycobacterium tuberculosis* (MTB), which is transmitted through the air when people who are sick with TB expel the bacteria into the air by coughing, sneezing, or spitting. The disease is preventable with the administration of a vaccine, and curable with the use of antibiotics over a significant period of time (although drug-resistant strains of the bacteria are becoming increasingly common) [6].

Nevertheless, the disease is often underdiagnosed and undertreated, especially in low-resource settings (i.e., developing countries, rural areas, and marginalized/impoverished communities), where the disease is more prevalent, calling for the development of more efficient and cost-effective methodologies to diagnose and treat the disease [2], [6], [7].

Some of the diagnosis techniques to detect TB include chest X-rays, sputum smear microscopy, and molecular tests. *Chest X-rays* are a common technique to diagnose tuberculosis due to the wide availability of the imaging devices required to perform the procedure, but they are often inconclusive and require expert radiologists to interpret the results.

*Sputum smear microscopy* is another widely-available technique that requires a trained clinician to identify the bacteria under images taken with a microscope of a patient's sputum (a mixture of saliva and mucus from the respiratory tract). This technique is relatively inexpensive but requires a high concentration of bacteria in the sample to be effective [8]. Additionally, studies argue that in conditions with limited resources and a large number of samples, there have been reports of poor sample observation and quality control measures, which can result in false-negative results [8].

Molecular tests based on *nucleic acid amplification* (NAATs), similar to the ones popularized to detect Covid-19, are another technique to diagnose TB [9]. These tests identify the presence of bacilli by amplifying the genetic material of the bacteria in a patient's sputum sample (if any is present) and using a chemical solution to react to the material. NAATs are by far the most reliable method to diagnose TB and have the advantage of being rapid and fully automated. However, they are also the most expensive to produce and require specialized equipment, making them less accessible in low-resource settings [10], [11]. A study conducted in 2018 showed that the ratio of smear microscopy tests to NAATs in countries with a high burden of TB was 6:1 [11], [12].

Each of these techniques presents its own benefits and limitations. Recently, there have been efforts to develop machine-learning models that can aid in the diagnosis of TB by reducing the need for expert clinicians in the process (or providing a first/second opinion) and improving the speed and/or accuracy of the diagnosis [1].

[1]More on this is presented in section **??**

### 1.3.4. Continually Adaptive Systems

## 1.4. Main Contributions

## 1.5. Structure of the Work

This work is divided into five chapters, including this introduction. Chapter 1 describes the context and motivation of this work and objectives, and provides the necessary background information to understand the concepts and techniques used.

Chapter 2 describes relevant work and state-of-the-art techniques. It also provides a literature review of related work in the field of machine learning methods for continual systems and their applications, highlighting the most important contributions and what their limitations are.

Chapter 3 describes the methodology used to design and implement the proposed system. It gives a detailed description of the data, models, techniques, and specific tools used in the implementation of the platform, experiments conducted, and relevant metrics to evaluate the system's performance.

Chapter 4 presents the results of the experiments, analyzes the performance of the system, and compares the results with the baseline metrics. Finally, Chapter 5 presents the conclusions and future work. It also discusses the limitations of the proposed system and the implications of the results obtained in the context of the project.

# 2. STATE OF THE ART

## 2.1. Relevant Techniques

While it is important to consider the design of robust machine learning models from their training phase, and part of this work will discuss techniques to optimize for that goal, our primary concern is to provide a model-agnostic framework for maintaining reliability *after* a model has been already deployed.

To accomplish this, we consider a set of novel techniques proposed in the literature to address the problems related to deploying supervised and semi-supervised machine learning models in real-world scenarios. The following section describes these techniques and their relevance to the problem at hand.

Note that while the focus is on applying these techniques to healthcare applications, it is easy to show that these methods can be generalized to other domains.

### 2.1.1. Continual Learning

Continual learning refers to the concept of constantly updating a model as new information arrives, allowing them to adapt to changing data [13]. This constant change in the data distribution requires models to be updated periodically to maintain their accuracy and relevance, prevent model stagnation and extend their relevance to the application.

We can consider two frameworks for implementing continual learning in a machine-learning system. We refer to the first as *offline learning*, where the model is updated periodically using a batch of data collected over time. Offline learning can be additionally subdivided into two categories based on how data is monitored prior to updating a model: *passive* and *active* continual learning.

The model is updated passively when the data is collected over a fixed period of time (e.g., every six months) or after a fixed number of instances have been processed and labeled (e.g., every 1000 new instances). Conversely, the model is updated actively when we update only when the model's performance drops below a certain threshold or when the system detects a significant change in the data distribution [13].

Online learning is the second framework for implementing continual learning. The difference between online and offline learning is that, with the former, the model is updated as soon as new data arrives, with every new instance - or small batch of instances - being used to update the model. Because this approach tends to be more computationally expensive than offline learning and suffers from well-known problems such as catastrophic forgetting [13], it is often used in conjunction with offline learning to improve the model's performance.

Continual learning is well-suited for healthcare applications where data arrives in a stream (e.g., wearable health monitors, electronic medical records, and imaging systems). It is also useful in terms of

#### 2.1.2. Active Learning

Active learning strategies selectively acquire data based on their informativeness or uncertainty to the model. Its value comes from allowing the model to guide its own data acquisition process, thus potentially reducing the need for vast - or unnecessary - amounts of pre-labeled data before a model is trained or updated [13]–[15]. This is particularly important in health applications, as the availability of annotated medical data is often limited due to privacy concerns, expert time constraints, and the complexity of pathological findings.

Active learning enables the development of accurate models using significantly less labeled data, paving the way for more efficient and cost-effective machine-learning platforms. Some of the most common active learning strategies include uncertainty sampling, query by committee, and expected model change.

#### 2.1.3. Adversarial Learning

Adversarial learning is a technique that aims to make models more robust to adversarial examples. Adversarial examples are inputs intentionally designed to be misclassified by a model. They are created by adding small perturbations to the input data, which are imperceptible to humans but can cause the model to make a wrong prediction. Adversarial examples are a major concern in healthcare applications, as they can mislead models into making incorrect diagnoses, misrepresenting the patient's condition, and leading to incorrect treatment decisions [16].

#### 2.1.4. Adaptive Policies from Human Feedback

In healthcare applications, feedback from experts is a very important resource, essential for improving the accuracy of the model. This technique aims to explode human feedback as much as possible to obtain model improvements. The goal is to construct policies that adapt the model based on feedback provided by the expert.

One recent example of this technique is OpenAI's ChatGPT, which used Reinforcement Learning with Human Feedback (RLHF) to instruction fine-tune their GPT-3 language models [17], [18] to align them better with human preferences. RLHF consists of training a reward function that maximizes the probability of the model's predictions matching the expert's feedback. This reward function is then used to train or fine-tune the model using reinforcement learning.

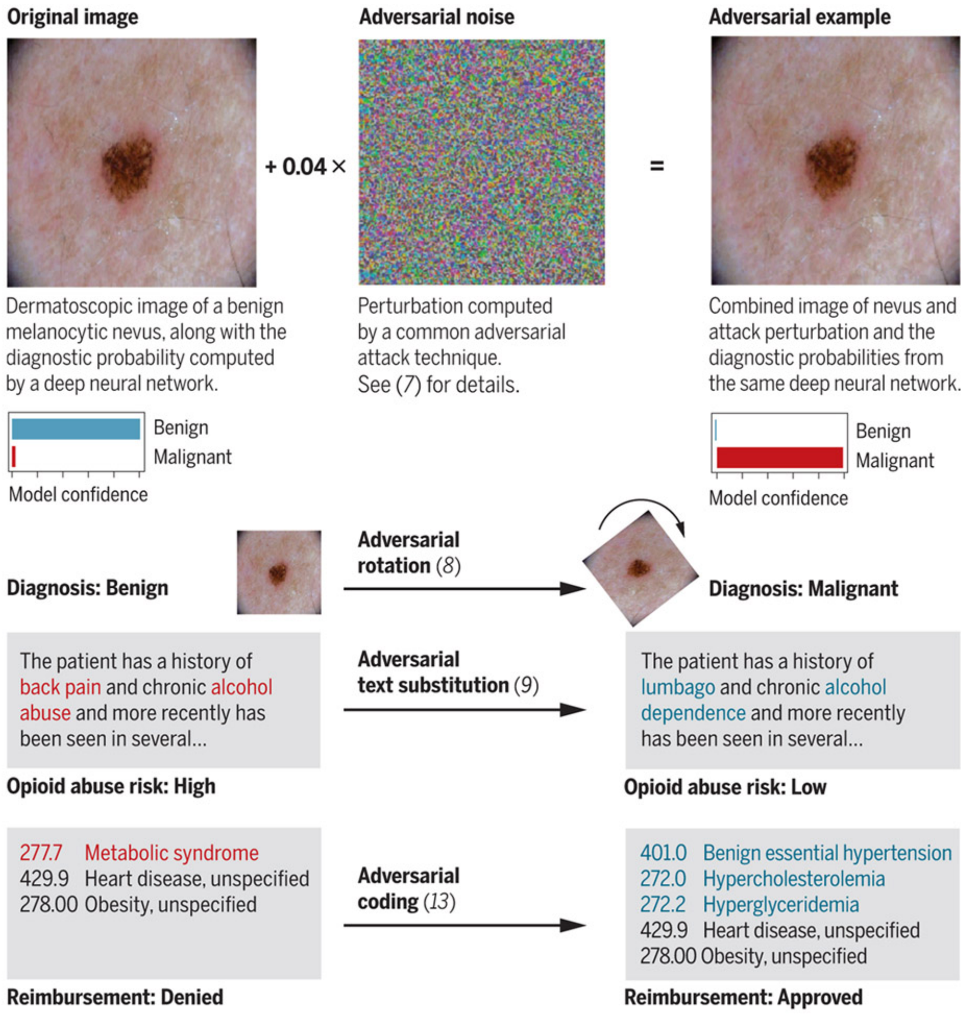


Fig. 2.1. Adversarial examples in medical imaging from Finlayson et al. (2019) [16].

## 2.2. Related Work

# 3. DESIGN OF THE SOLUTION

This work is driven by the problem of designing platforms primarily used to perform Machine-Learning inference (only obtaining the output predictions from a model) while ensuring that the obtained outputs are as reliable and robust as possible when encountering new data. This is important in scenarios where a model is deployed in high-stake environments where their output causes significant downstream impact, and inaccurate predictions may be costly to the relevant stakeholders.

We consider that these types of platforms have the following characteristics:

- The (already trained) models are uploaded to the platform primarily to obtain predictions from new data. We refer to the bag of models available to the platform as the *model repository*.
- The data used to train and evaluate the models is available. This means the platform can have prior knowledge about the data distribution the models were trained on.
- The platform has enough hardware resources available to perform inference with the models that it has available.
- The platform constantly receives new data for inference, which is assumed to be independent of the data used to train the models.
- Labeling capabilities are limited. That is, annotating new data is costly and unfeasible in large amounts. This is common in real-world applications, notably in the healthcare sector, where the labeling task is performed by professional workers, making it a costly and time-consuming task [14], [15], [19].

Furthermore, for the purpose of this work, we consider that the inference environment is deployed alongside an *experimental* environment. This environment is assumed to be detached from the training environment and is set up to continuously evaluate the models' robustness on new data, evaluate its performance after retraining using the aforementioned techniques, and compare the results with the deployed model. If the experimental model outperforms the deployed model, the latter is updated with the new model. This process is repeated for the entire lifetime of the application.

The role of the human annotator in designing machine-learning models is also an important aspect of this work...

# 4. RESULTS

In this chapter, we present a rundown of the results obtained from evaluating all the experiments. First, an analysis of the results of testing our system...

## 4.1. Analysis of the results

.

## 4.2. Comparison of the different methods

## 4.3. Discussion

# 5. CONCLUSIONS

Discussion about the results obtained and implications in the context of the project, limitations of the proposed system, future work, etc.

## 5.1. Main Implications

## 5.2. Limitations of the System

## 5.3. Future Work

## 5.4. Other Research Directions

## 5.5. Final Remarks

# REGULATORY FRAMEWORK

# SOCIO-ECONOMIC ENVIRONMENT

## Budget

The estimated costs of the realization of this project include those related to the human labor and material costs associated with it.

In terms of human resources, both the student author of the work and the advisor of the thesis are considered to have put hours of labor into the making of this project. We have assumed the salary of the student to be equivalent to the one of a junior engineer of about €15.00 per hour of labor, and the one of the advisor to be equivalent to that of a senior engineer of around €35.00 per hour.

An estimate of xxx hours...

| | # of Hours | Salary per hour (€) | Total Salary (€) |
|---|---|---|---|
| Author | xxxx | yy | zzzz |
| Advisor | xxx | yy | zzzz |
| | | **Total Cost:** | **€...** |

Table 5.1. ESTIMATED COSTS OF HUMAN RESOURCES

In terms of non-human resource costs, the only relevant ones are those associated with the material costs of the hardware utilized throughout the work. This is because only open-source software tools (Python, R, LaTeX) were utilized.

...

## Socio-Economic Impact

## Ethical Considerations

...

# BIBLIOGRAPHY


[1] CDCTB, *World TB Day History*, en-us, Feb. 2023. [Online]. Available: https://www.cdc.gov/tb/worldtbday/history.htm (visited on 08/19/2023).

[2] WHO, *Global Tuberculosis Report 2022*, en, 2022. [Online]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 (visited on 08/19/2023).

[3] *European Accelerator of Tuberculosis Regime Project*, en-US. [Online]. Available: https://era4tb.org/ (visited on 04/24/2023).

[4] World Health Organization. Regional Office for Europe, "Tuberculosis: Fact sheet on Sustainable Development Goals (SDGs): Health targets," en, World Health Organization. Regional Office for Europe, Tech. Rep. WHO/EURO:2017-2388-42143-58059, 2017, number-of-pages: 8. [Online]. Available: https://apps.who.int/iris/handle/10665/340885 (visited on 08/19/2023).

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," en, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Number: 7553 Publisher: Nature Publishing Group. DOI: 10.1038/nature14539. [Online]. Available: https://www.nature.com/articles/nature14539 (visited on 05/06/2022).

[6] WHO, *Tuberculosis (TB)*, en, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis (visited on 08/20/2023).

[7] IMI, *IMI Innovative Medicines Initiative | ERA4TB | European regimen accelerator for tuberculosis*, en, Jan. 2020. [Online]. Available: http://www.imi.europa.eu/projects-results/project-factsheets/era4tb (visited on 08/20/2023).

[8] P. Desikan, "Sputum smear microscopy in tuberculosis: Is it still relevant?" *The Indian Journal of Medical Research*, vol. 137, no. 3, pp. 442–444, Mar. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3705651/ (visited on 08/19/2023).

[9] C. C. for Disease Control and Prevention, Governmental, *TB Diagnostic Tool: Xpert MTB/RIF Assay Fact Sheet | TB | CDC*, en-us, Aug. 2016. [Online]. Available: https://www.cdc.gov/tb/publications/factsheets/testing/xpert_mtb-rif.htm (visited on 08/19/2023).

[10] H. Albert *et al.*, "Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: What lessons have we learnt and how can we do better?" eng, *The European Respiratory Journal*, vol. 48, no. 2, pp. 516–525, Aug. 2016. DOI: 10.1183/13993003.00543-2016.

[11] E. MacLean *et al.*, "Advances in Molecular Diagnosis of Tuberculosis," *Journal of Clinical Microbiology*, vol. 58, no. 10, 10.1128/jcm.01582–19, Sep. 2020, Publisher: American Society for Microbiology. DOI: 10.1128/jcm.01582-19. [Online]. Available: https://journals.asm.org/doi/10.1128/jcm.01582-19 (visited on 08/19/2023).

[12] D. Cazabon *et al.*, "Market penetration of Xpert MTB/RIF in high tuberculosis burden countries: A trend analysis from 2014 - 2016," eng, *Gates Open Research*, vol. 2, p. 35, 2018. DOI: 10.12688/gatesopenres.12842.2.

[13] C. Huyen, *Designing Machine Learning Systems*, en. O'Reilly Media, Inc., May 2022, ISBN: 9781098107963. [Online]. Available: https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/ (visited on 04/24/2023).

[14] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," en, *Journal of Biomedical Informatics*, vol. 58, pp. 11–18, Dec. 2015. DOI: 10.1016/j.jbi.2015.09.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415002038 (visited on 04/24/2023).

[15] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, Feb. 2012. DOI: 10.1186/1472-6947-12-8. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8 (visited on 04/24/2023).

[16] S. G. Finlayson *et al.*, "Adversarial attacks on medical machine learning," *Science (New York, N.Y.)*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019. DOI: 10.1126/science.aaw4399. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7657648/ (visited on 04/25/2023).

[17] P. Christiano *et al.*, *Deep reinforcement learning from human preferences*, arXiv:1706.03741 [cs, stat], Feb. 2023. DOI: 10.48550/arXiv.1706.03741. [Online]. Available: http://arxiv.org/abs/1706.03741 (visited on 04/25/2023).

[18] L. Ouyang *et al.*, *Training language models to follow instructions with human feedback*, arXiv:2203.02155 [cs], Mar. 2022. DOI: 10.48550/arXiv.2203.02155. [Online]. Available: http://arxiv.org/abs/2203.02155 (visited on 04/25/2023).

[19] A. Yakimovich, A. Beaugnon, Y. Huang, and E. Ozkirimli, "Labels in a haystack: Approaches beyond supervised learning in biomedical applications," en, *Patterns*, vol. 2, no. 12, p. 100 383, Dec. 2021. DOI: 10.1016/j.patter.2021.100383. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921002506 (visited on 04/24/2023).