

# **Building an Intelligent System to Detect Respiratory Insufficiency**

Vitor Daisuke Tamae

CAPSTONE PROJECT  
PRESENTED TO THE DISCIPLINE  
MAC0499

Supervisors:  
Prof. Dr. Alfredo Goldman  
Prof. Dr. Marcelo Finger  
MSc. Renato Cordeiro Ferreira

São Paulo, December of 2022

# Contents

<b>List of Figures</b>	<b>v</b>
<b>1 Introduction</b>	<b>3</b>
<b>2 Concepts</b>	<b>4</b>
2.1 Intelligent Systems . . . . .	4
2.1.1 Use cases . . . . .	4
2.1.2 MLOps . . . . .	5
2.2 Microservices Architecture . . . . .	5
2.3 Reactive Systems . . . . .	6
2.4 Layered Architecture . . . . .	7
2.5 Hexagonal Architecture . . . . .	8
2.6 Progressive Web Application (PWA) . . . . .	9
<b>3 Preliminary Implementation</b>	<b>10</b>
3.1 Context . . . . .	10
3.2 Challenges . . . . .	10
3.3 Lessons learned . . . . .	11
3.3.1 Well-defined interfaces . . . . .	11
3.3.2 Dependency Management . . . . .	11
3.3.3 Automated Tests . . . . .	12
3.3.4 MLOps . . . . .	12
<b>4 Architecture</b>	<b>13</b>
4.1 Objectives . . . . .	13
4.2 Context . . . . .	13
4.3 Model Registry . . . . .	14
4.3.1 Description . . . . .	14
4.3.2 Implementation . . . . .	15
4.4 Message Service . . . . .	16
4.4.1 Description . . . . .	16
4.4.2 Implementation . . . . .	18
4.5 Storage . . . . .	19
4.5.1 Description . . . . .	19
4.5.2 Implementation . . . . .	20

4.6	Back-end Services . . . . .	20
4.6.1	API . . . . .	21
4.6.2	Model Server . . . . .	21
4.7	UI . . . . .	22
<b>5</b>	<b>API</b>	<b>23</b>
5.1	Core . . . . .	23
5.2	Web Server . . . . .	24
5.3	Authentication . . . . .	25
5.4	Database . . . . .	27
5.4.1	Description . . . . .	27
5.4.2	Implementation . . . . .	28
5.5	Inference Requests . . . . .	29
5.6	Deployment . . . . .	31
<b>6</b>	<b>Model Server</b>	<b>32</b>
6.1	Registry Adapter . . . . .	32
6.2	Inference Messages . . . . .	33
6.3	Deployment . . . . .	34
6.3.1	ML Pipeline . . . . .	34
6.3.2	Microservice . . . . .	35
<b>7</b>	<b>UI</b>	<b>36</b>
7.1	User Access . . . . .	36
7.2	Inference Form . . . . .	37
7.3	Inference List . . . . .	37
<b>8</b>	<b>Tests</b>	<b>39</b>
8.1	TDD . . . . .	39
8.2	Test Hierarchy . . . . .	39
8.2.1	Unit Tests . . . . .	40
8.2.2	Integration Tests . . . . .	40
8.2.3	System Tests . . . . .	41
8.3	Multi-Stage Build . . . . .	42
8.4	CI Pipeline . . . . .	42
<b>9</b>	<b>Conclusion</b>	<b>43</b>
9.1	Results . . . . .	43
9.2	Next steps . . . . .	43
9.2.1	Model feedback . . . . .	44
9.2.2	NATS improvements . . . . .	44
9.2.3	Load balancer . . . . .	44
9.2.4	UI improvements . . . . .	44
9.2.5	Training pipeline . . . . .	44

<b>Bibliography</b>	<b>45</b>
---------------------	-----------

# List of Figures

2.1	Intelligent System applications. . . . .	4
2.2	Origins of the MLOps set of practices . . . . .	5
2.3	Reactive systems benefits diagram . . . . .	6
2.4	<b>Hexagonal Architecture diagram.</b> The application contains the business logic and it communicates with the adapters through the ports. . . . .	8
3.1	<b>Timeline of the SPIRA project:</b> this research aims to build an inference system for SPIRA. Icons created by Freepik - Flaticon. . . . .	10
4.1	Architecture of the SPIRA intelligent system . . . . .	14
4.2	<b>MLFlow UI.</b> The UI can be used to track the list of models in production, the staging and production versions, and also past experiments. . . . .	16
4.3	<b>NATS topic diagram.</b> The API sends inference messages to each topic . . . . .	18
4.4	<b>NATS UI.</b> The UI can be used to perform server health-checks, or check live connections and subscriptions. . . . .	19
4.5	<b>MinIO Server UI.</b> The audio files are stored in a bucket and can be consulted in the directory named with the corresponding inference id . . . . .	20
4.6	<b>Diagram showcasing the logos of the tools used to implement the inference system.</b> . . . . .	21
5.1	<b>Architecture of the API service.</b> . . . . .	23
5.2	<b>User creation sequence diagram.</b> . . . . .	25
5.3	<b>User authentication sequence diagram.</b> . . . . .	27
5.4	<b>Mongo Express UI.</b> The UI can be used to directly interact with the database. . . . .	29
5.5	<b>Database sequence diagram for inference requests.</b> Once inference requests are received, the web server calls the core, which then calls the database port to store the entity in the Mongo DB. . . . .	29
5.6	<b>File storage sequence diagram for inference requests.</b> Audio files are stored in MinIO server to be retrieved by Model Servers later. . . . .	30
5.7	<b>Message service sequence diagram for inference requests.</b> Published message will be received by a Model Server subscribed to the model topic. . . . .	30
5.8	<b>Docker Hub image repositories for the API service.</b> Images are automatically pushed to Docker Hub upon new tag creations in GitHub. . . . .	31
6.1	<b>Architecture of the Model Server.</b> . . . . .	32
6.2	<b>Model Server retrieving audio files from MinIO server</b> . . . . .	34

6.3	<b>Model Server using the model to make predictions</b> . . . . .	34
6.4	<b>Docker Hub repositories of the Model Server and the model registry</b> . . .	35
7.1	<b>Inference App installed in a mobile device.</b> . . . .	36
7.2	<b>Inference list page for desktop users</b> . . . . .	37
7.3	<b>Inference form page.</b> Users fill the form before making the audio recordings. . . .	38
7.4	<b>Inference audio recording page.</b> Users need to follow the instructions in the screen to record the audios. . . . .	38
7.5	<b>Sign In page in the inference app.</b> Users should have an account to access the system. . . . .	38
7.6	<b>Inference form page for mobile users.</b> The application changes the layout of the page to a cards display instead of a table. . . . .	38
9.1	<b>MLOps milestones achieved in the research</b> . . . . .	43

# Abstract

Respiratory insufficiency is a symptom caused by the inadequate gas exchange performed by the respiratory system. **SPIRA** is a research project created during the COVID-19 pandemic to detect respiratory insufficiency via speech analysis based on Machine Learning models. The project is currently preparing to train a new generation of models that will be validated in hospitals with the help of medical personnel. Due to the demand for validation, one of the steps of this preparation phase is to build a new system that applies these models. This monograph describes the planning, implementation and deployment of an intelligent distributed inference system that allows medical personnel to perform a respiratory insufficiency pre-diagnosis using the models created by SPIRA. The research shows the advantages in responsiveness and resilience obtained by adopting a reactive microservices architecture. Moreover, it emphasizes the importance of MLOps in modern Machine Learning Engineering through the lessons learned from the preliminary system. The impacts on quality obtained by following these principles are highlighted with the implementation of a pipeline and a registry to automate the deploy of new models in the final version of the inference system.

**Keywords:** SPIRA, Respiratory Insufficiency, Machine Learning Engineering, Intelligent Systems, Reactive Microservices, Hexagonal Architecture, MLOps

# Resumo

Insuficiência respiratória é um sintoma causado pela troca inadequada de gases feita pelo sistema respiratório. **SPIRA** é um projeto de pesquisa criado durante a pandemia de COVID-19 para detectar insuficiência respiratória via análise de fala baseada em modelos de Machine Learning. O projeto está atualmente se preparando para treinar uma nova geração de modelos que serão validados em hospitais com o acompanhamento de equipes médicas. Devido à demanda por validação, uma das etapas desta fase de preparação consiste em construir um novo sistema que aplicará estes modelos. Esta monografia descreve o planejamento, a implementação e a implantação de um sistema de inferência distribuído e inteligente que permite que equipes médicas realizem um pré-diagnóstico de insuficiência respiratória utilizando os modelos criados pelo SPIRA. A pesquisa mostra as vantagens em responsividade e resiliência obtidas através da adoção de uma arquitetura reativa de microsserviços. Além disso, é enfatizada a importância do MLOps na Engenharia de Machine Learning moderna através das lições aprendidas a partir do sistema preliminar. Os impactos em qualidade obtidos seguindo esses princípios são ressaltados com a implementação de um pipeline e um registro para automatizar a implantação de novos modelos na versão final do sistema de inferência.

**Palavras-chave:** SPIRA, Insuficiência Respiratória, Engenharia de Machine Learning, Sistemas Inteligentes, Microsserviços Reativos, Arquitetura Hexagonal, MLOps



# Chapter 1

## Introduction

Respiratory insufficiency or failure is a well-known medical symptom resultant from an inadequate gas exchange performed by the respiratory system (Cam65). It may cause severe symptoms such as hypoxemia, a reduction in the concentration of oxygen in bloodstream, and hypercapnia, a rise in arterial carbon dioxide levels (RK03). These conditions may lead to cough, tiredness, shortness of breath, and in extreme cases, death. This symptom may be caused by various diseases. Examples of them are the flu, severe asthma and heart conditions (SLK09; CCN<sup>+</sup>00; DS94).

In 2020, respiratory insufficiency was identified as one of the severe symptoms of COVID-19 infection (SVdB22). The symptom is aggravated by a condition called *silent hypoxia*, low blood oxygen concentration without breath shortness (CF20)(TLJ20), which particularly hinders its diagnosis.

There were several attempts to obtain early detection of *silent hypoxia* (Teo20) (RTA<sup>+</sup>21). SPIRA is a research project created during the COVID-19 pandemic to identify respiratory insufficiency via speech analysis based on Machine Learning models (CGC<sup>+</sup>21). It involves doctors, speech therapists, linguistics, and computer scientists from different fields (artificial intelligence, computer music, software engineering) and was idealized as a tool for doctors and nurses to pre-diagnose this symptom.

During the first stage of the project, the authors of (CGC<sup>+</sup>21) created a ML model that could be used to identify respiratory insufficiency with reasonable accuracy, enabling a new phase of the project focused on the development of a more sophisticated generation of models.

This new generation now should be validated by carrying out inference experiments with patients from different hospitals under the supervision of medical personnel. These professionals are not expected to have familiarity with ML related coding tools, therefore a system is needed to properly provide inference features for them. Nevertheless, SPIRA does not have a system capable of providing this functionality yet.

**The main goal of this research is to allow physicians, nurses and other related hospital personnel to perform a pre-diagnosis of respiratory insufficiency using SPIRA models.** Milestones to achieve this objective are described in more detail later at Chapter 4.

This monograph describes the planning, development and implementation of the inference system created to serve this new generation of SPIRA models and it is subdivided as follows: Chapter 2 describes the theoretical basis, Chapter 3 shows a preliminary implementation of the system, Chapter 4 describes the architectural features of the inference system, Chapter 5, Chapter 6 and Chapter 7 describe the implementation details of each component, Chapter 8 describes how the system was tested and Chapter 9 shows the results obtained.

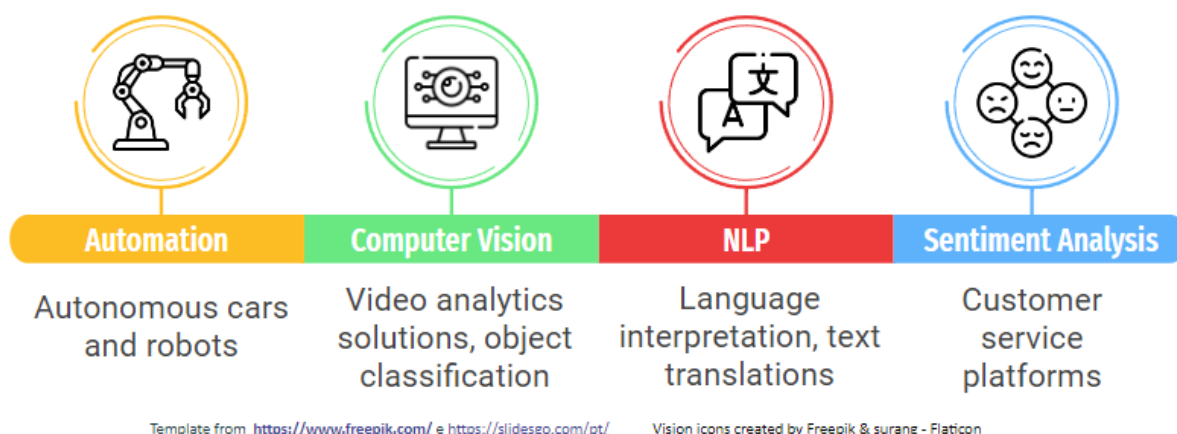
# Chapter 2

## Concepts

### 2.1 Intelligent Systems

#### 2.1.1 Use cases

An intelligent system is a software system that uses artificial intelligence, often produced via machine learning, offering support to it to achieve an objective (Hul18).



**Figure 2.1:** *Intelligent System applications.*

Intelligent systems are appropriate when either there is no analytical solution for a certain problem or the analytical alternative is too expensive given limited computational resources. In these cases an empiric solution via machine learning is more suitable and affordable.

Examples of such problems are:

- problems that require a considerable amount of work.
- open-ended problems, which continue to grow over time.
- time-changing problems, where the right answer changes over time.
- problems that are intrinsically hard.

Detecting respiratory failure is an intrinsically hard problem (TLJ20). The symptom may manifest differently depending on the disease associated with it, such as in the case of *Silent hypoxia*. This is also an indication that the diagnosis is susceptible to changes over time.

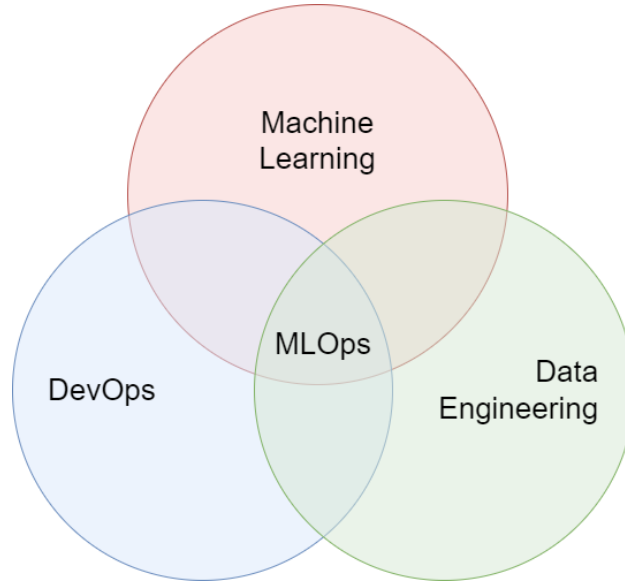
As described in Chapter 1, the use of machine learning models showed to be a good solution for pre-diagnosing respiratory insufficiency. Therefore, an intelligent system is necessary to provide the inferences and manage the life cycle of the SPIRA models.

### 2.1.2 MLOps

In the field of software engineering, DevOps practices are used to shorten the development life cycle, providing continuous delivery to deploy new features more efficiently (EGHS16).

Unlike general software features, an ML model is an entity whose business logic is dependent on the data consumed during the training stage. This leads to many problems such as the lack of reproducibility and difficult deployments (Hut18; TOS<sup>+</sup>20).

For this reason, DevOps culture in the Machine Learning field has originated a new set of practices called MLOps, which aims to deploy and maintain machine learning models reliably and efficiently (TOS<sup>+</sup>20).



**Figure 2.2:** *Origins of the MLOps set of practices*

Even though its behavior is determined by a particular dataset, a ML model is no more than an artifact. Therefore, it is possible to store it, keep track of its version and reproduce it when needed, which is a particularly relevant feature for the model validations that will be conducted in the SPIRA project through the inference system.

In DevOps, Continuous Delivery pipelines, also known as CD pipelines, are created to automatically deploy new software. Likewise, an ML pipeline can be created to automate the deployment of ML models. A ML pipeline not only shortens model creation time span considerably, but it is also a way to standardize model development procedures. Hence, it may further improve the overall productivity for the production of new models inside the SPIRA project.

Nowadays, there are several MLOps tools such as MLFlow, KubeFlow and Comet that can be used to implement each of the practices presented previously. Besides storing model artifacts and creating an ML pipeline, they can also be used to run experiments with the models and monitor their performance in production.

Preliminary implementations not adopting such an MLOps approach are described in [Chapter 3](#). These experiences provided a motivation to search for a more sophisticated procedure to deploy SPIRA models in the current research and its planning will be further described in [Chapter 4](#).

## 2.2 Microservices Architecture

A microservice is a small autonomous service that performs a single functionality. A microservices architecture is an architectural style where the system is composed by microservices (Bon16).

Implementing a microservices architecture offers many benefits to the system:

- **Modularity:** contrary to the monolithic multi-functional paradigm, if a service contains a set of well-defined functionalities, the architecture can help modularizing them into self-contained microservices and avoid turning the system into a "big ball of mud" (FY97).
- **Horizontal Scalability:** a service may be scaled according to its load, without the need to change the remaining system (DLT<sup>+</sup>17). Hence, resources may be distributed more efficiently to optimize the performance of the application.
- **Feature Delivery:** with each microservice having its own isolate deploy, features can be implemented more frequently, because developers will not interfere with one another while working in different microservices.
- **Poliglotism:** isolation between services also allows each of them to use the technology that is the most suitable for their functionality, as long as they respect the established communication protocols.

Certain trade-offs need to be taken into account when adopting a microservices architectural style. In a distributed system, such as one using a microservices architecture, additional uncertainties are added due to the lack of a consistent and reliable shared memory between nodes. Hence, services are not able to fully rely on one another to make a decision.

This uncertainty is very present in the concept of time. Timestamps and clocks are unique for each service, therefore a distributed system has no universal concept of "now" that can be used to synchronize them. One solution is to accept this uncertainty and develop strategies to cope with it, such as handling eventual consistency and making sure the communication protocols are associative (batch-insensitive), commutative (order-insensitive), and idempotent (duplication-insensitive). As it will be presented later in Chapter 4, these properties are important for planning the architecture, but also for choosing the right tools to implement the system.

The inference system will contain multiple intelligent microservices, each containing a SPIRA model, all of them communicating with another microservice responsible for providing an inference API that will receive requests and store results. The next section shows how the trade-offs of the microservices architecture will be handled in this research.

## 2.3 Reactive Systems

A system is reactive when it is built based on the *Reactive Principles* (real4). More precisely, a reactive system relies on asynchronous message-driven communication between components in an attempt to solve some of the inherent problems of distributed systems such as the ones presented for the microservices architecture.

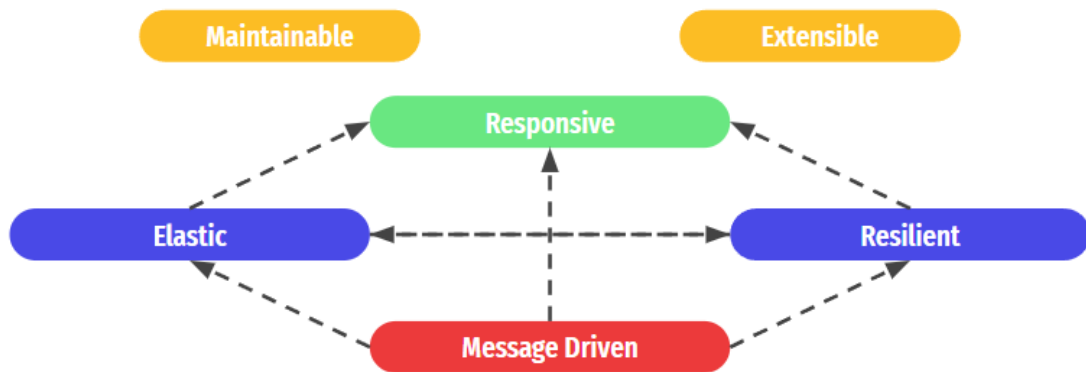


Figure 2.3: *Reactive systems benefits diagram*

The key of a reactive system is to manage the existent uncertainty between components directly in the architecture level. That is, to design autonomous components that expose their protocols

and clearly define what they promise, what behavior will trigger, and how the data model should be used (rea14).

Moreover, a reactive system should give each component the freedom to make its own business logic decisions. When a node sends a call to another one, that node must have the ability to communicate momentary degradations (such as overloads) and must have the ability to not respond when that is appropriate. This requires the protocol between them to be asynchronous and event-driven, because component-level failures can only be contained if the communication protocol between the services anticipates the possibility of unsuccessful or late responses. Neither should interrupt their internal functioning due to these failures (rea14).

Asynchronous communication also leads to temporal decoupling, because one node does not require the availability of another node to communicate. Components become even more autonomous and the system gets more reliable (Bon16).

Reactive systems also aim to create space decoupling by making use of network communication to avoid allocating all its components in one specific location. This reduces the dependency on specific hardware, which are susceptible to malfunctioning and inaccessibility, and allows the nodes to be distributed in different locations to compose the system. Such freedom makes the system elastic and scalable, as it can dynamically control its resources to satisfy both momentary and permanent increases in demand for a service (Bon16).

The inference system developed in this research will be used by multiple hospitals to validate new models created by the SPIRA project. Thus, a certain degree of volatility in the demand for each model is expected both upon their creation and also in case they outperform other models. Therefore, the intention is to follow the concepts presented above to develop a reactive system that is capable of adapting to these changes in demand.

## 2.4 Layered Architecture

The layered architecture is an architectural style in which a system is designed in layers, each containing components related to a specific role within the application.

While the number of layers and their roles are flexible, architectural patterns implementing a layered architecture usually contain the following three layers:

- **Presentation Layer:** responsible for abstracting inner data to the user and vice-versa. Note that this layer has no concerns regarding how the data is treated and retrieved in the application.
- **Business Layer:** responsible for executing business rules of the application. Likewise, this layer is neither involved in how the data is displayed to the user nor the origin of the data.
- **Database Layer:** solely responsible for storing application relevant data.

Besides the responsibilities of each layer, there are additional constraints used to maintain the service organization: a layer only interacts with two other layers and data-flow is bidirectional, i.e., either from the user to the system or the other way around.

While the layered architecture offers topological constraints to organize responsibilities in the service level, there are inherent pitfalls of the architectural style such as undesired dependencies between layers and infiltration of business logic in user interface code.

In the example above, the presentation layer depends on the business layer, which then depends on the database layer, violating the Dependency Inversion Principle (DIP) (Mar96), as business rules depend on database technologies.

Some of the known consequences of those pitfalls are:

- developers are often not able to isolate and properly test the business logic, and
- the application is highly coupled with human-driven use, precluding the possibility of batch-running the system.

The violation of the DIP motivated the creation of other architectural patterns where the layers were organized in an onion-like structure with the business logic contained in the innermost layer of the design. Examples of them are the *clean architecture* (MGB<sup>+</sup>18), the onion architecture (jef08), and the hexagonal architecture (Coc).

## 2.5 Hexagonal Architecture

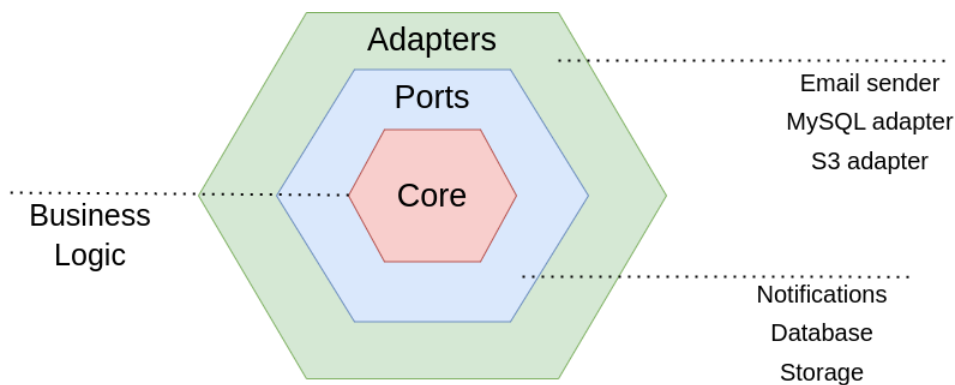
The hexagonal architecture is a layered architectural pattern described by Alistair Cockburn in an attempt to deal with recurrent structural problems of object-oriented programming and layered architectures (Coc).

As commented in the previous section, communication between layers in a layered architectural pattern must be properly designed and managed to avoid logic infiltration. While this is crucial for the system maintainability, the layered architectural style neither specifies how this communication should be done nor how the abstractions should be converted. The hexagonal architecture solves this problem by implementing a new layer above the business layer solely responsible for managing this communication.

The architecture is then divided in three main layers:

- **the core** contains all the business logic of the system. It has a functional nature represented in the form of use cases, the methods of the business logic.
- **the adapters** compose the outer layer of the architecture. They are responsible for the execution of business logic, signaling the proper technology to take action. Examples of adapter components are database clients and API endpoints.
- **the ports** layer is responsible for managing the communication between the core and the adapters. The ports are primarily responsible for establishing the contract between the two other layers. They also convert the abstraction of an use case to the abstraction of an adapter method and vice-versa.

Ports are also divided in two categories: primary and secondary. The primary ports, also called driving ports, are the ones that drive the core calling the business logic. Example of it are REST API adapters. The secondary ports, also called driven ports, are the ones that are called by the business logic to execute some action. Examples of it are database adapters.



**Figure 2.4: Hexagonal Architecture diagram.** The application contains the business logic and it communicates with the adapters through the ports.

It is important to note two aspects of this design:

- the core is agnostic to the implementation of the adapters and their technology, and
- the ports are the boundaries of the business logic.

both bring immediate benefits to the system. If there is a need to change the technology of a component in the system, the core code can be reused and one can simply create the new adapter and connect it to the port. Additionally, ports also become the entry point for testing the business logic, since the ports layer naturally isolates it.

The main difference from many architectural patterns to the hexagonal architecture is that most of them do not fully isolate the business logic from external dependencies. The MVC pattern, for example, implements this idea for the controller layer, but not others.

This architecture also facilitates test automation with the use of mock adapters, which further helps asserting the correctness of the system as it will be discussed in [Chapter 8](#).

The system developed in this research requires many distinct technologies for its implementation, hindering the use of an architectural pattern that does not fully isolate the business logic from the technologies. For this reason, the hexagonal architecture will be adopted.

## 2.6 Progressive Web Application (PWA)

Progressive Web Applications are cross-platform compatible web applications that are able to work offline and can be added to device home screens. The term was first used in 2015 ([Rus15](#)) to describe apps that took advantage of modern browser features that enable users to convert web apps into native apps of their operating system.

Examples of such features are service workers, which are web workers used by PWAs as a proxy that checks the availability of the remote server and stores cache content for the application. This cache stores the UI content, which enables a more native-like user experience without the dependence on web connectivity. It also allows content to be loaded progressively from a initial layout provided by local storage, hence the name Progressive Web Application.

As it will be described in [Chapter 2](#), using a PWA as the front-end tool to collect inferences has many advantages over native solutions.



## Chapter 3

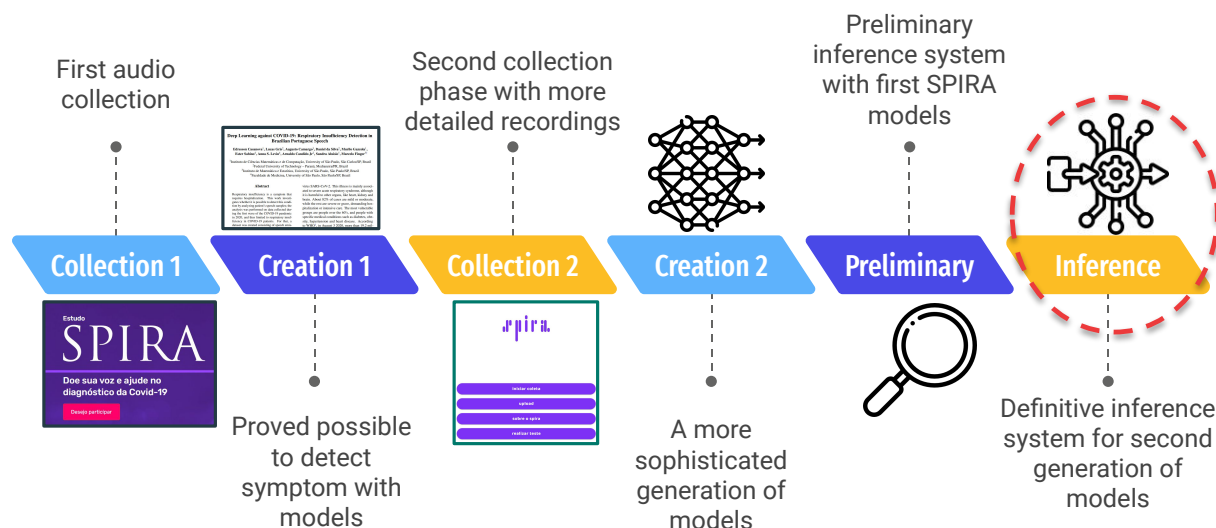
# Preliminary Implementation

### 3.1 Context

The SPIRA project started with the first voice audio recording collection phase, as described in Figure 3.1. These recordings were then used by researchers to produce their first machine learning model and prove the feasibility of symptom detection (CGC<sup>+</sup>21). Currently, a second collection phase is being conducted and the new data will be used to train a new generation of SPIRA models.

In parallel to that, there were attempts to encapsulate SPIRA models inside a service and make it available for inferences requests. While the system was successfully developed and deployed, there were many challenges faced in this phase that could either be avoided or better handled to facilitate its implementation.

For the purposes of this research, these first attempts will be referred as the preliminary system, while the current project will be referred as the definitive system from now on. The definitive system was heavily based on the lessons learned from the mistakes made in this preliminary implementation.



**Figure 3.1: Timeline of the SPIRA project:** this research aims to build an inference system for SPIRA. Icons created by Freepik - Flaticon.

### 3.2 Challenges

The model was developed in an preliminary process where the training source code was not based on any specific design pattern or convention. Rather than maintainability, feasibility was its primary goal.



The main consequence of that was the absence of a well-defined interface of the model. Hence, developers had to do create a model server adapted to create an entry point for this single model. Particularly, input and output data processing were responsible for the great majority of the changes.

After the adjustments, developers and researchers needed to validate the model produced by the new pipeline. The goal was to ensure its predictions were coherent with the results presented in (CGC<sup>+</sup>21) and the validation was divided in three stages:

- **First stage:** the first goal was to check if all the dependencies were compatible with the ones used in the article. This was necessary because many machine learning libraries are under constant change due to heavy development and can change their behavior considerably between versions.
- **Second stage:** subsequently, automated tests were created to isolate errors in the code. As stated previously, input and output processing methods were the ones which underwent most changes. Therefore, special attention was given to ensure their correctness.
- **Third stage:** finally, the third step was to do a more detailed analysis of results. Unfortunately, true reproducibility was impossible, since the researchers did not use any MLOps process while creating their preliminary model. Therefore, there was no way to regenerate the exact trained model used in the experiments. Consequently, the disparity in metrics was minimized, but not fully eliminated.

After the validation process, the deployed system was able to process inference calls via HTTP protocol and retrieve prediction history. All results were stored in a local database and they were used for a more thorough analysis of the model performance afterward.

### 3.3 Lessons learned

The challenges faced during the preliminary implementation showed critical points of the development process and the system design that needed to be addressed for the definitive version. The sections below present the main problems and the proposed solution for them.

#### 3.3.1 Well-defined interfaces

It was clear after the preliminary attempt that the presence of an interface between the model and the model server was crucial for the scalability of the SPIRA Inference system. This is because, without defining the boundaries of the roles of the model server and the model class, it would be necessary to adapt the model to fit the necessities of each new trained model. This would make system scaling infeasible when the number of models became considerable, besides all the maintenance problems that each entry point would cause.

With such an interface, the intention was to delegate the processing of input and output data for the model class, rather than the model server. Models should directly consume inference-related metadata and process it according to its own particularities before using it for prediction. The same should happen to the output value, that should also be converted back to the inference result abstraction.

#### 3.3.2 Dependency Management

After the validation process, dependency management needed more attention than initially planned. This step is also crucial during the containerization of the system, as dependencies are all installed in compilation time.

Another concern comes with library updates, where one might update a dependency and disregard any changes necessary to sub dependencies, which are a potential cause of system disruption. Ideally, this process should not be done manually for the sake of long-term maintainability.

### 3.3.3 Automated Tests

During the preliminary attempt, tests showed to be a useful tool to detect failures in the code. They are, undoubtedly, necessary for system maintenance. Despite such effectiveness, the development process of the definitive version should follow the procedures described in the TDD section to better track and test the system requirements.

Another important consideration is that the preliminary version was an encapsulation of the model with no reactive features in it. As a consequence, test hierarchy was not of much importance as there was no need to test the integration of the components of the system. With the new inference system architecture, nonetheless, test hierarchy does matter, as there are many levels in which the system needs to be tested to ensure its integrity, from the internal functioning of each of its components to the integration between them. Details about the strategies used for tests are described later on [Chapter 8](#).

### 3.3.4 MLOps

Reproducibility of model behavior was infeasible during the preliminary attempt and a solution was certainly needed for model storage, versioning and monitoring. As cited in [Chapter 2](#), there are many solutions that can be used to manage the life cycle of a model and track its prediction history. This is an important tool to audit the system by enabling true reproduction of a model.

# Chapter 4

## Architecture

### 4.1 Objectives

The SPIRA inference system is a reactive microservices intelligent system that pre-diagnoses respiratory insufficiency. To describe this architecture, it is important to describe this project's research goals. As described in [Chapter 1](#), the goal of this research is to allow medical personnel to perform a pre-diagnosis of respiratory insufficiency using SPIRA models. To accomplish this objective, the research will be divided in the following milestones:

- **Create a solution to store new models**

Currently, there is no default procedure to store newly trained models. The inference system needs such a solution to encapsulate SPIRA models with a server and expose them to the system.

- **Turn models accessible**

After a model is trained and evaluated, it needs to be deployed through a service that will wrap it and make it accessible to process inference requests. Currently, a new generation of models is under development, therefore the serving system should be scalable enough to keep up with this process. It is an even greater challenge to build a model serving infrastructure that can wrap all variants of SPIRA models.

- **Create a system to manage models and inferences**

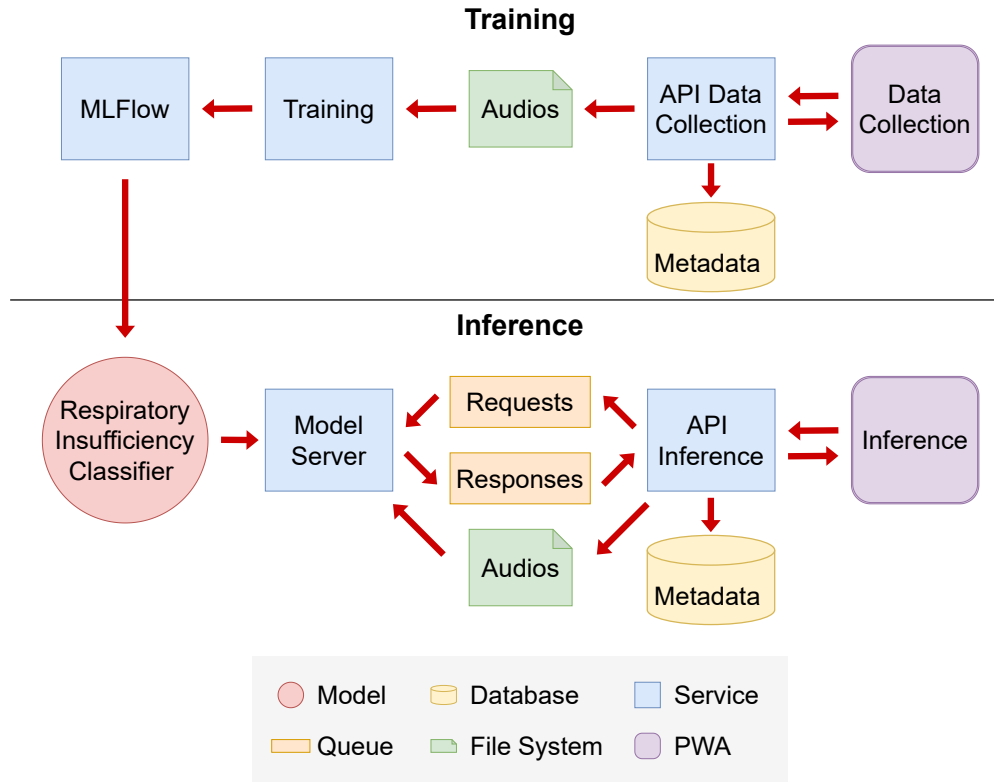
Once the Microservices carrying the models are ready, an orchestrator is needed to redirect inference requests, process predictions, and organize the pre-diagnoses. The system should store all the data regarding inference metadata and results, while being resilient enough to cope with the time needed to process the inferences.

- **Provide an App to be used by medical personnel to access the models**

At last, the models ought to be used by medical personnel to determine the pre-diagnosis. Therefore, a new App is needed to provide them with this access.

### 4.2 Context

The architecture of the entire SPIRA system is described in [Figure 4.3](#).



**Figure 4.1:** Architecture of the SPIRA intelligent system

The upper half of the diagram represents the training sub-system and it includes the audio data collection and storage processes, model training, model creation and finally model storage. The collection is done through a PWA designed to record voice audios of patients speaking specific sentences according to a data collection protocol.

The data sets will subsequently be used by the data science team to create new machine learning models. These models are then stored in a server where they become available for use by the rest of the system.

The lower half of the diagram is the definitive inference sub-system being implemented in this research. It is composed by a PWA that sends inference requests to the back-end service also developed in this project, specifying the desired model created with the data from the training sub-system. The inference request is then processed by an orchestrator API and subsequently sent to the corresponding microservice carrying the target model. The pre-diagnosis generated by the prediction is then sent back to the API and becomes available for the users.

From now on, the terms "API" and "API Service" will be used to refer to the orchestrator API microservice, while the term "Model Server" will be used to refer to the microservice that carries a ML model and makes processes inferences with it.

Since the objective of the inference system is to provide access to all SPIRA models, there will be multiple model servers deployed, each containing their own model retrieved from the model registry, enabling the API to send prediction requests to any of them.

## 4.3 Model Registry

### 4.3.1 Description

As observed during [Chapter 3](#), one of the main challenges found during development was the reproducibility of results, since models used in (CGC<sup>+</sup>21) could not be replicated. Therefore, a model registry is crucial to follow MLOps practices during the development of the next generation of SPIRA models.

The chosen solution was the MLFlow tracking server ([mlf22](#)). MLFlow is a platform created by Databricks to manage the lifecycle of Machine Learning models. It is agnostic to ML libraries, making it compatible with any technology used to develop Machine Learning models, including PyTorch, the most commonly used library inside the SPIRA project, and other libraries such as Keras and TensorFlow. Therefore, if developers decide to change the technology used to create models, MLFlow will be equally compatible with the new libraries of choice.

The model registry is composed by three distinct containers:

- **MLFlow server:** The server hosting the MLFlow model registry.
- **Database:** The database is used by the server to store all metadata related to experiments and model versioning.
- **Storage Unit:** It is used by the server as an artifact store for model files. The storage unit will be explained in more detail in [Section 4.5](#).

Model Servers connect to the model registry during initialization to retrieve one of the models stored in it. Details on how this retrieval is done will be explained in [Subsection 4.6.2](#).

While the goal of this research is to use this component as a model storage solution, the registry is actually agnostic to the existence of the inference system. Developers may also use it to retrieve models for any other purpose besides inferences such as for model performance comparisons or for running independent experiments. As the server can keep track of a production version of each model, it is possible to change the tag of a model to production whenever it is ready to be deployed to the system.

### 4.3.2 Implementation

The registry used is a docker container with a python image which downloads the MLFlow CLI and initializes a server using it and was first developed by ([reg22](#)). In the development repository ([mod22](#)), where the image is defined, the Dockerfile is described below:

```
1 FROM python:3.8 as mlflow_base
2
3 ENV POETRY_VERSION=1.1.13
4
5 WORKDIR /project
6
7 RUN apt-get update
8 RUN pip install --upgrade pip
9 RUN pip install "poetry==$POETRY_VERSION"
10 RUN poetry config virtualenvs.create false
11
12 COPY pyproject.toml poetry.lock .tool-versions ./
13
14 RUN poetry install --no-interaction --no-ansi --no-root
15
16 EXPOSE 5000
```

The file first pulls a python image and installs poetry package. Poetry then uses the poetry.lock file, where dependencies for the registry are declared, and uses it to install them. Finally, the internal port 5000 of the container is exposed, which is the default port used by MLFlow server.

The command to initialize the server itself is specified directly in the docker-compose file as follows:

```

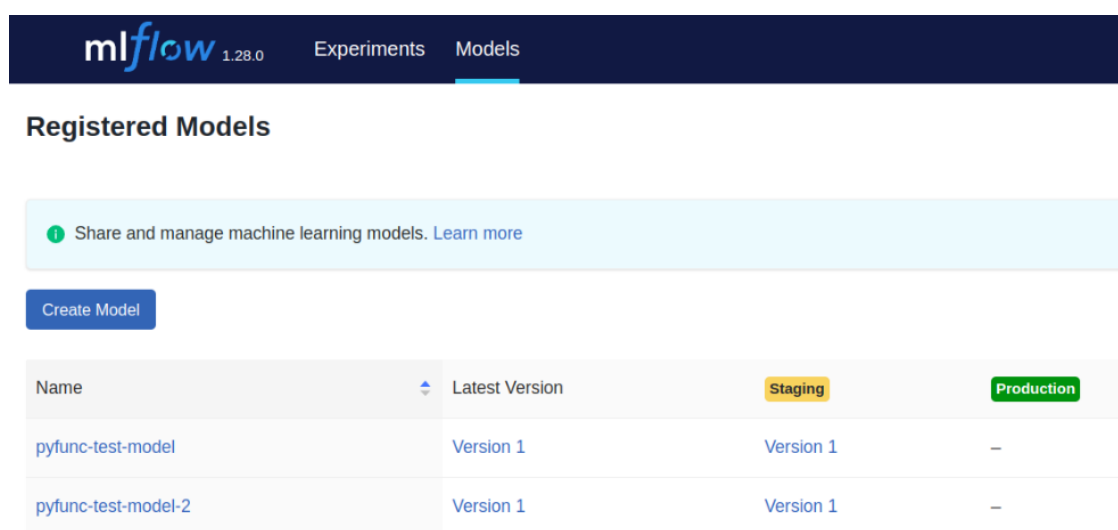
1 'mlflow server \
2   --backend-store-uri \
3   mysql+pymysql://${MYSQL_USER}:${MYSQL_PASSWORD}@db:3306/${MYSQL_DATABASE} \
4   --default-artifact-root s3://mlflow/ --host 0.0.0.0'

```

It is necessary to specify both the metadata and the artifact storages. The metadata storage used is a MySQL relational database container, whereas the artifact storage is a MinIO bucket storage system. While the bucket is of exclusive use for MLFlow, the same MinIO instance is used by the server to store audio files.

Afterwards, the docker image can be uploaded to DockerHub and downloaded back in production to initialize the server. This process was automated with a CD Pipeline using GitHub Actions. Details on how the pipeline was setup will be described in [Subsection 6.3.1](#).

Models stored in the server can be revisited using the MLFlow server UI as shown in [Figure 4.4](#).



**Figure 4.2: MLFlow UI.** The UI can be used to track the list of models in production, the staging and production versions, and also past experiments.

The UI is exposed in the same port as the MLFlow server, hence in a private port. Developers with access to the production machine might then use it to control the staging and production versions of each model and also create new models directly from it, despite this not being the recommended method to register new models. The server supports programmatic registration of models and the ML pipeline setup to do that will be described in [Section 6.3](#).

## 4.4 Message Service

### 4.4.1 Description

As described in [Chapter 2](#), there are numerous benefits in using event-driven communication. Most importantly, the system becomes reactive, therefore more resilient and horizontally scalable. The two mostly utilized solutions to implement such asynchronous communication are message and queue services.

While the current plan is to maintain one communication stream for each model deployed in production, giving margin to implement a queue service, there are three reasons for not using queues for this system:

- **Harder Model Server setup:** Each newly created model needs to be registered and a new server to be deployed with it. Using a queue service would also require a new queue to be deployed with each server, which would make this setup more complicated. Using a message service brings the benefit of using a single broker to manage all communication stream, therefore there is no need for deploying a new infrastructure for new models introduced in the system.
- **1:N requests:** Although the system was implemented to send inferences to a single model for each inference request, it might be a future feature request to send inference requests to multiple models for comparing results from the same input. A message service provides a simplified 1:N communication, as many listeners can subscribe to the same stream, while queue services would require the API to send messages proportionally to the number of models.
- **N:1 updates:** Message services also enable all Model Servers to send their results to a single topic. If a queue service were to be used instead, each deployed model would also require an additional queue solely for sending their own updates. The orchestrator API would also need to scale the number of threads listening to incoming messages from the queues, which would further complicate its implementation.

Once an inference request gets to the API, the request metadata is stored in a NoSQL database and subsequently encoded in the form of a message. The API is responsible for sending this message to a broker in a specific channel and delegates the inference for the consumer on the other end. Once the prediction is complete, the results may be sent encoded as another message. The API recovers this result and update the database, making it available for retrieval in the UI.

The message service to be used is NATS ([nat22](#)), an open-source cloud-native solution that is easily scalable. NATS maintains communication streams through a topics system. A producer may publish messages in a topic to a broker and consumers willing to listen to the given topic may subscribe to it to start receiving published messages.

In the inference system, there is a single broker with each model subscribed to an exclusive topic. Once the API receives a request, it retrieves the correct channel to be used in the database and then sends the message in the given topic. The Model Server subscribed to it then proceeds with the prediction.

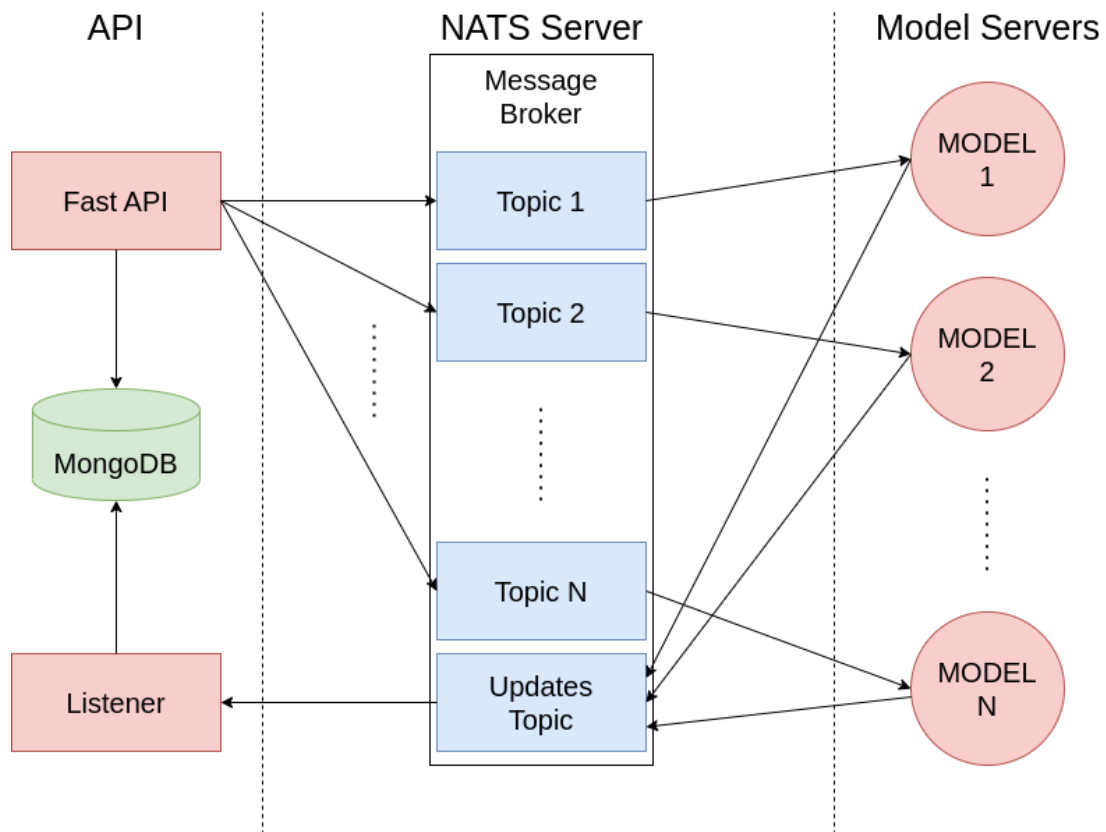
Updates are all sent to an updates topic. The API subscribes to this topic and all Model Servers publish messages in it.

Handling NATS message-driven communication requires a producer inside one service and a consumer on another one. In the context of the hexagonal architecture, a producer can be abstracted in the form of a message service adapter and the consumer as a running thread adapter. The latter will be called the listener adapter from now on.

The message service adapter is the point of communication between the servers and the broker. It is responsible for encoding and sending inference requests to a certain topic. This section of the adapter is called by the use cases triggered by the web-server. The incoming inference requests schedule the messages as shown in [Figure 4.3](#) and a successful status is returned to the front-end in case the message is successfully sent.

The same adapter is also responsible for receiving and decoding messages of a given topic. Receiving a message on push mode, i.e. actively wait for messages in a given topic, depends on having a process exclusively for calling the receiver method. This process is created by the listener adapter.

The listener calls the business logic use case of receiving a message in a given topic, the message service adapter is then called to wait for a message and, once arrived, the received message triggers the database update in the orchestrator API and the model prediction in the Model Server. Once completed, the method ends and the listener restarts this procedure.



**Figure 4.3: NATS topic diagram.** The API sends inference messages to each topic

#### 4.4.2 Implementation

The NATS broker or server is a docker container pulling the NATS image. The service in the docker-compose file is described below:

```

1 nats:
2   image: nats
3   restart: always
4   ports:
5     - '${NATS_CLIENT_PORT}:${INNER_NATS_CLIENT_PORT}'
6     - '${NATS_UI_PORT}:${INNER_NATS_UI_PORT}'

```

NATS requires two distinct ports. By default, 4222 is the one used by consumers and producers to communicate with the broker and 8222 is a monitoring UI where connections and messages can be checked. These ports are exposed in the internal Docker network of the system. They are also mapped to private ports in the local machine (localhost), where they become available for users with access to production.





**Figure 4.4: NATS UI.** The UI can be used to perform server health-checks, or check live connections and subscriptions.

NATS also exposes a third port 6222 for clustering support, however this feature will not be implemented in this research and may be considered as a future improvement for the messaging system.

## 4.5 Storage

### 4.5.1 Description

The file storage has two distinct roles in the system:

- **Audio recordings:** Inference requests contain audio recording files that are crucial for result prediction. As explained previously, the Model Server and the orchestrator API communicate via message broker. NATS, as many other message services, have a message size limit of 1MB, which is not enough space for the necessary audio recordings. Therefore, a file storage solution where the API is able to share recording with the Model Servers is necessary.
- **MLFlow Artifacts:** MLFlow server delegates data storage functionalities to other components. One of the delegations is for artifact storage, which require a file storage service.

MLFlow documentation recommends an AWS S3 bucket or similar file system like service as the default artifact storage. Therefore the solution should be preferably compatible with the S3 service. It is also of the interest of the SPIRA project to maintain all audio recordings stored on premise, as these recordings may also be used for model training or data auditing conducted through the privacy related regulations ([lgp18](#)).

Due to these requirements, the chosen solution was MinIO simple storage ([min22](#)). The MinIO solution is compatible with the Amazon S3 cloud service and stores all its contents in a container volume mounted in the local file system.

Like the S3 simple storage, MinIO organizes the stored files as objects inside buckets. SPIRA contains two buckets. The first bucket is used by the API to store the incoming audio files in directories referenced by the inference ID. Model Servers can then access these files using MinIO client. The second bucket is used as the default artifact storage of the model registry. This bucket is not accessed directly by any part of the system other than the registry itself.

Likewise the message service, a simple storage adapter is used by the API and the Model Server to interact with MinIO. The API requires a method to store files and the Model Servers only require a method to retrieve them.

Storing should happen before the inference message is sent to NATS server. This way, it is guaranteed that the Model Server will find the necessary files in MinIO by the time it receives the request and starts the prediction.

### 4.5.2 Implementation

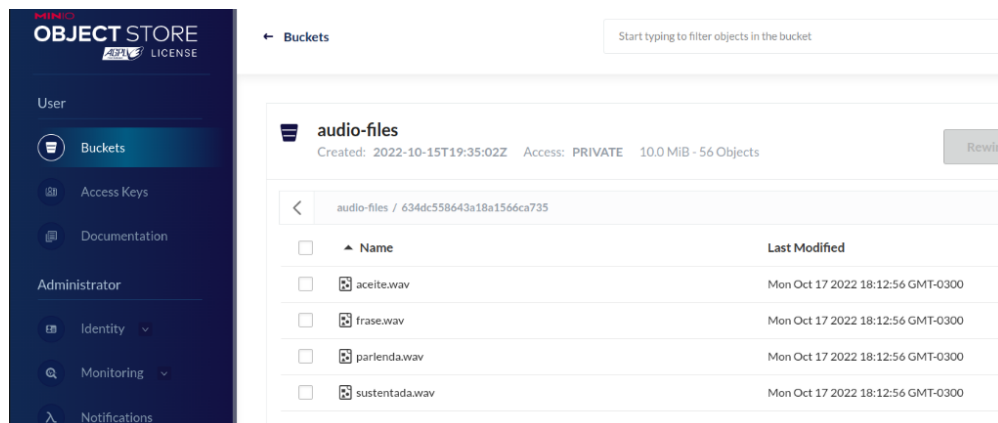
The file storage server is a docker container pulling the MinIO image. The service in the docker-compose file is described below:

```

1 minio:
2   image: minio/minio
3   restart: always
4   ports:
5     - '${MINIO_CLIENT_PORT}:${INNER_MINIO_CLIENT_PORT}'
6     - '${MINIO_UI_PORT}:${INNER_MINIO_UI_PORT}'
7   command:
8     - server /data
9     --console-address ':${INNER_MINIO_UI_PORT}'
10    --address ':${INNER_MINIO_CLIENT_PORT}'
11   volumes:
12     - minio_data:/data
13   env_file:
14     - .env

```

MinIO Server requires two distinct ports. The first port exposes a file management client that is used by other services to manipulate the files, while the second port exposes a UI that can be used to revisit stored objects in the server, as shown in [Figure 4.5](#). Both ports are private, which means that only users with access to the production server should have access to it.



**Figure 4.5: MinIO Server UI.** The audio files are stored in a bucket and can be consulted in the directory named with the corresponding inference id

MinIO stores all its contents in /data directory by default, which is a docker volume minio\_data in the local machine. Hence, the data is preserved even if the container is deleted.

## 4.6 Back-end Services

The implemented services, i.e. the Model Server and the orchestrator API, were implemented in Python. There were trade-offs to choose this language.

The development of an intelligent system requires a language with great support for Machine Learning, as it will be constantly dealing with model inferences. Python is currently one of the

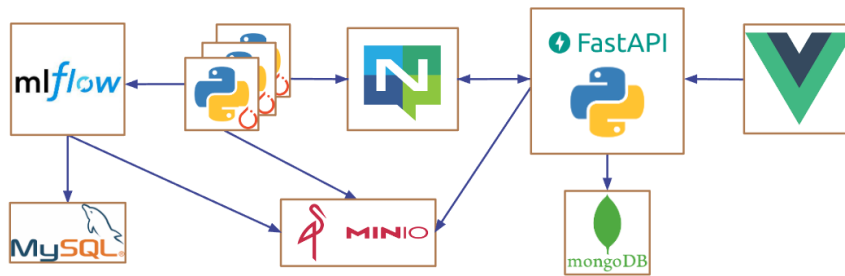
most popular language for ML, making it a great choice for the project due to its large community support and wide variety of frameworks and tools related to the area.

Particularly, the SPIRA models are developed with PyTorch library, which is exclusive for Python language. Therefore, the use of Python is unavoidable to some extent, since the interface between the model and the inference system needs to be defined.

In addition, building a project in the same language not only guarantees that there will not be compatibility issues when interacting with the model, but also provides a variety of functionalities from the dependencies that would otherwise not be available using another language.

Thus, the ML compatibility benefits of using Python far out-weight the cons, as they can be easily handled to not jeopardize the implementation of the services.

The final tools used in the implementation are summarized in [Figure 4.6](#).



**Figure 4.6:** *Diagram showcasing the logos of the tools used to implement the inference system.*

#### 4.6.1 API

The inference management API service is the central component of the inference system. It is responsible for receiving incoming inference requests from the PWA. It serves as an orchestrator, sorting requests and sending each of them to the desired Model Server via broker. The system may then be used to compare the performance of each model, validate them, and compare the models for further improvements.

The service uses two threads. The first thread is a web server which will be publicly exposed, from where the UI client will communicate with the back-end system. The second thread is a listener loop which is subscribed to a NATS topic. This topic is where all the incoming results from inference models will be published. The listener thread will receive the messages and update the database with the corresponding predictions for each inference.

The API also contains a database to control user, model and inference related data. The database solution to be used for the project is MongoDB. The reason for this choice is the fact that MongoDB is one of the most widely used NoSQL database solutions. It provides a simple way to store metadata and evolve it over time without the need to use rigid schemas such as the ones used in a relational database.

Further details about the implementation of the API will be given in [Chapter 5](#).

#### 4.6.2 Model Server

The component responsible for retrieving a model from the MLFlow server and making inferences with it is the Model Server. As explained in [Section 4.2](#), SPIRA will contain multiple Model Servers, each wrapping a model from the registry.

The Model Server first retrieves the ML model from the registry and subscribes to a specific NATS topic in initialization time. The API publishes messages in this given topic whenever it processes inference requests for that particular model. The Model Server, which continuously listens for messages incoming from the NATS server, receives the published message, processes the inference using the retrieved model instance, and then publishes another message containing the pre-diagnosis in the updates topic in which the API subscribed.

As explained in [Chapter 2](#), eventual consistency should be properly handled in order to guarantee that inferences are being correctly processed by the system. One of the ways to cope with it is to ensure that the operations are associative (batch-insensitive), commutative (order-insensitive), and idempotent (duplication-insensitive). In the current research, operations are inferences made by Model Servers.

Note that if there are multiple instances of a Model Server carrying the same ML model and, therefore, listening to the same message topic, each Model Server instance will process the prediction and send the corresponding result for an inference, referenced by its ID. Since both the model and the input are equal to all instances, any result is valid. Therefore, the last update in the inference result is the one which will be persisted in the database. Hence, the operation is batch-insensitive.

The model ID is part of the primary key of the inference entity, which means that an inference relates to exactly one model of the system. Therefore, no particular order of inference requests will give different final persisted data in the database. Hence, the operation is commutative.

Finally, if the same inference is processed multiple times, e.g., an inference request somehow triggers multiple messages for the same Model Server, the effect is similar to having two distinct instances of the same model, in which case the final result is still valid. Hence, the operation is also idempotent.

With those three properties, we may ensure that, if the model prediction is correct, the system will also correctly inform the result to the user.

Further details about the implementation of the Model Server will be given in [Chapter 6](#).

## 4.7 UI

The UI Application is the PWA to be used by medical personnel. Users will send inference requests through the UI and will also visualize inference results through it.

Currently, there exists a collection PWA which is being used for the new collection phase of the SPIRA project. From now on, this App will be explicitly referenced as the "collection PWA" while the App being developed in this research will be referred as the "Inference PWA", "new PWA" or simply "PWA".

The planning of the collection PWA passed through UX reviews and is already being used by medical personnel. Since the collection and inference processes are identical to each other. Using the same UI structure for the inference app not only helps with the adaptation of users to the new app and also guarantees that no bias is added to the inference process due to changes in the UI layout.

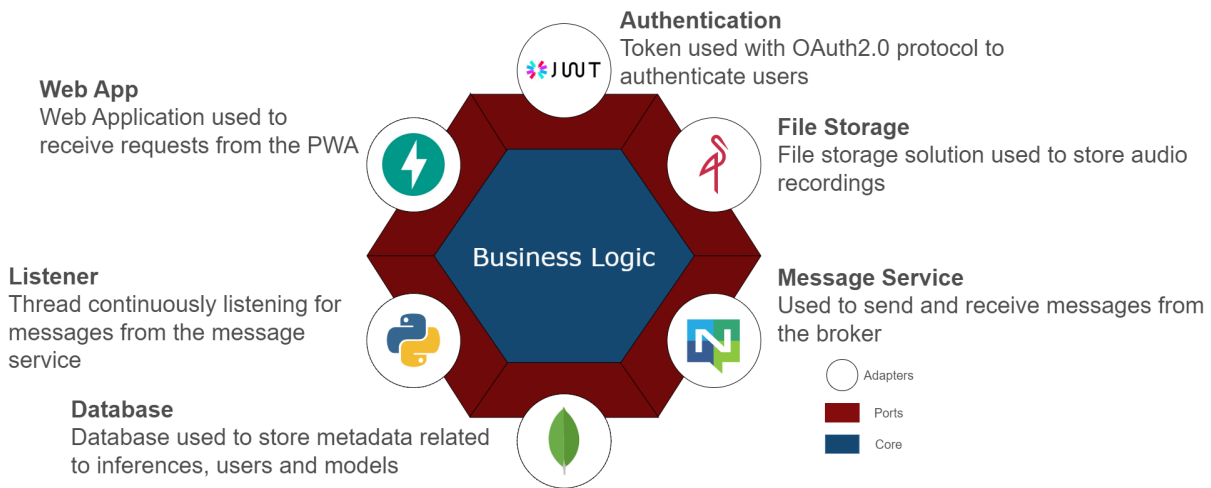
To further preserve the data flow of the collection PWA, the inference App also maintains the same connections with collection back-end server to retrieve metadata necessary to build the inference form.

The only existing data flow that needs to be redirected to the new orchestrator API is the form submission. While the collection App uses it to store data in the collection server, the same dataflow will be converted to a stream of inference requests. Therefore, it is possible to extend the collection PWA so that it can be deployed in either collection or inference mode. This way, the same project can be re-utilized for the development of the new PWA.

# Chapter 5

## API

The architectural pattern used in the API is the Hexagonal Architecture as shown in [Figure 5.1](#). The adapters implemented in it and the core business logic are explained in more detail in the following sections.



**Figure 5.1:** *Architecture of the API service.*

### 5.1 Core

The core of the API is where the business logic is contained and it is composed of two components: the entities and the services.

The entities are the definition of classes used inside the business logic. They were created using Pydantic, a data validation Python library that enforces the use of type-hints. This package will be used by the use cases to ensure that types are being followed correctly during run time.

The services, or use cases, should be immutable to the external dependencies ([Coc](#)). Therefore, the core does not use environment variables. They are rather all used inside the adapters. Hence, use cases are reliant solely on the inputs provided by the primary ports via Dependency Injection to understand the circumstances of the call. Through this reasoning, the functional paradigm becomes more suitable to them rather than the object-oriented programming. Thus, the core was implemented following the functional paradigm.

With the aforementioned decisions, environment changes do not interfere with business rules, which facilitates the deployment of the system and promotes space decoupling, as promoted by the reactive principles. Note that a functional core also ensures that a change in one of its use cases will not be disruptive to other sections of the business logic, which is beneficial for future feature requests directly related with the core.

The following example is a use case called by the endpoint showed previously for user creation:

```

1  def create_new_user(
2      authentication_port: AuthenticationPort,
3      database_port: DatabasePort,
4      user_form: UserCreationForm,
5      token: Token,
6  ) -> None:
7      try:
8          if not authentication_port.validate_token(token):
9              raise DefaultExceptions.credentials_exception
10
11         _validate_new_user(database_port, user_form)
12
13         new_user = UserCreation(
14             username=user_form.username,
15             email=user_form.email,
16             password=user_form.password,
17         )
18         new_user.password = authentication_port.get_password_hash(new_user.password)
19         database_port.insert_user(new_user)
20
21     except LogicException:
22         raise
23     except:
24         raise LogicException(
25             "Could not create new user", status.HTTP_500_INTERNAL_SERVER_ERROR
26         )

```

The method first authenticates the call and proceeds with the user form validation. Once the form is validated, the new user is then inserted in the database. If for any reason a failure happens during this process, the exception is caught and informed to the client through the HTTP status code in the response. Likewise, each endpoint in the router calls their respective service.

## 5.2 Web Server

The framework used to implement the web server was FastAPI, due to its better performance in comparison with Flask. FastAPI also has built-in support for some of useful technologies, such as the OAuth2.0 protocol that is used in the authentication system.

Additionally, FastAPI is able to automatically convert request body contents as Pydantic Base-Models. This feature is particularly useful for POST requests such as user and inference creations.

Additionally, users are only allowed to perform inference operations within their own scope, that is, `/users/:user_id`. Any request from a third-party user to an endpoint out of this scope will generate a Forbidden Operation response.

The endpoints are abstracted in the Hexagonal Architecture as an adapter. Each endpoint is declared as a python function in FastAPI. The router then calls the core method that will execute the corresponding service. Following the same example of user creation, the implementation of the `/users` endpoint is described below:

```

1  @router.post("/")
2      def create_user(
3          user_form: UserCreationForm,
4          token_content: str = Depends(oauth2_scheme)
5      ):

```

```

6      try:
7          create_new_user(
8              authentication_port,
9              database_port,
10             user_form,
11             Token(content=token_content),
12         )
13     except LogicException as e:
14         raise HTTPException(e.error_status, e.message)
15     return {"message": "user registered!"}

```

A POST request in `/users` will call the method above, which will then call the `create_new_user` method inside the core. The core method will add the entity in the database through the database port and then the MongoDB adapter and will acknowledge the web server regarding the success of the operation. The sequence of calls is described in more detail in the sequence diagram of Figure 5.2.

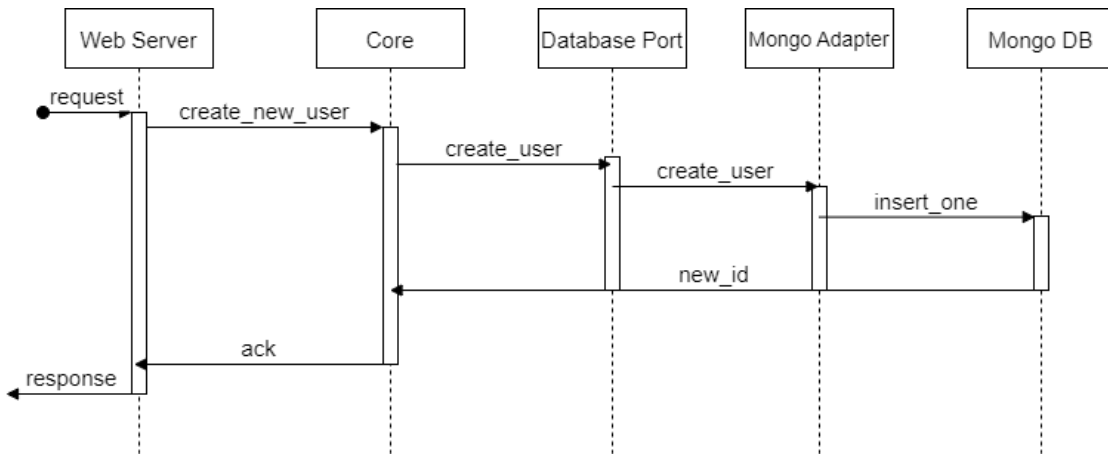


Figure 5.2: *User creation sequence diagram.*

If the request is processed successfully, a response is returned with an HTTP status code 200 (Ok), otherwise, a `LogicException` object is thrown by the Core and the router converts it to an HTTP response.

The implemented server supports requests at the endpoints described in Table 5.1.

## 5.3 Authentication

In contrast with the training sub-system, inference related operations using SPIRA models are not designed to be publicly available. The intended use of the system is restricted to medical personnel. Therefore, an authentication mechanism is necessary in order to satisfy this requirement.

The protocol to be used is OAuth 2.0. This protocol requires users to send their credentials to an especial endpoint `/users/auth`, where they will be validated and responded with a token. The users may then use this token to access the inference system and make requests using it.

The format of the token may vary according to projects decisions. The token pattern to be used in this research is the JSON Web Token, mostly known as JWT ([jwt22](#)). JWT is composed of three distinct strings. The first part is the header, where details regarding the algorithm used to encode the token are stored. The second part is the payload, that may include any data relevant for the application. The last is the signature, which is the most relevant part for authentication. The server uses a private key alongside the header and the payload to encode the signature when creating the token for a user. In case a malicious request is sent with a fake token, the server is able to validate

Request	Endpoint	Description
GET	/users/:user_id	Retrieves the user entity from DB by the user ID
POST	/users	Creates a new user
POST	/users/auth	Authenticates user credentials
POST	/models	Creates a new model
GET	/models/:model_id	Retrieves the model entity from DB by the model ID
GET	/models	Retrieves a list with all models
GET	/users/:user_id/inferences/:inference_id	Retrieves the inference entity from DB by the inference ID
POST	/users/:user_id/inferences	Creates a new inference
GET	/users/:user_id/inferences	Retrieves the list of inference entities of a user
GET	/users/:user_id/inferences/:inference_id/result	Retrieves the result of an inference

**Table 5.1:** *Endpoints supported by the orchestrator API.*

by checking if this signature is correct given its private key, which is supposedly not accessible to the client.

In accordance with the architecture pattern adopted, the authentication mechanism is implemented as an adapter. The business logic is then responsible for implementing the OAuth2.0 protocol, while being completely agnostic to the JWT implementation. The adapter on the other hand, is solely responsible for correctly implementing JWT authentication, while also being ignorant of the fact that the OAuth2.0 protocol is being used. As a result, the protocol and the token pattern may be changed for newer and more sophisticated versions in the future without interfering with each other.

The core method to generate the token upon an authentication request is as follows:

```

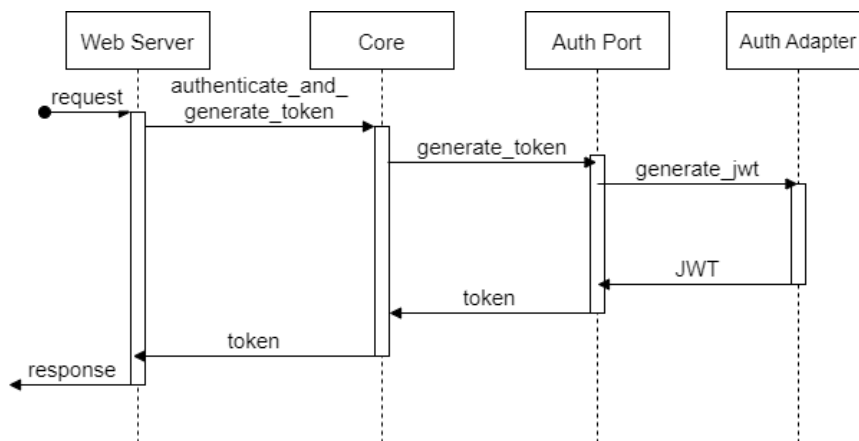
1  def authenticate_and_generate_token(
2      authentication_port: AuthenticationPort,
3      database_port: DatabasePort,
4      username: str,
5      password: str,
6  ) -> Tuple[str, Token]:
7      try:
8          user: User = _authenticate_user(
9              authentication_port, database_port, username, password
10         )
11         if user is None:
12             raise
13
14         token = authentication_port.generate_token(
15             data=TokenData(username=user.username)
16         )
17     except:
18         raise DefaultExceptions.user_form_exception
19
20     return user.id, token

```

Users credentials are first authenticated and, in case they are valid, a token is generated through the authentication port and adapter. The method proceeds returning the token. Note that in any



moment the core is informed that the token is a JWT. The sequence of calls is detailed in [Figure 5.3](#).



**Figure 5.3:** *User authentication sequence diagram.*

## 5.4 Database

### 5.4.1 Description

The database stores information regarding four main entities: Users, Models, Inferences and Results.

The Users collection stores the hashed passwords of the accounts. The hashed passwords are then compared with login credentials to authenticate the access and generate the JWT token that will be used in subsequent requests of that account.

The Models collection is related to the models that are available for use in the system. It also contains the NATS topic in which the API will send inference request messages for that particular model. Model Servers retrieving this a particular model should then subscribe to this topic in order to receive the requests properly.

The Inferences collection contains all metadata related to inferences. The entity is currently defined as a Pydantic BaseModel inside the Core as:

```

1 class Inference(BaseModel):
2     id: str
3     status: str
4     user_id: str
5     created_in: str
6     model_id: str
7     rgh: str
8     mask_type: str
9     gender: str
10    covid_status: str
11    local: str
12    age: Optional[int]
13    cid: Optional[str]
14    bpm: Optional[str]
15    respiratory_frequency: Optional[str]
16    respiratory_insufficiency_status: Optional[str]
17    location: Optional[str]
18    last_positive_diagnose_date: Optional[str]
19    hospitalized: Optional[str]
20    hospitalization_start: Optional[str]
  
```

```
21     hospitalization_end: Optional[str]
22     spo2: Optional[str]
```

While the definition of inference is already set for the current phase of the SPIRA project, this definition is still subject to change. The current definition of an inference is actually an improved version of the preliminary implementation, which implies that future versions may also include additional data.

This volatility in the definition is the main reason why a NoSQL database is more suitable for the purposes of this research. A relational database would require constant back-fills of missing information each time the definition of an inference is changed, while a NoSQL database, such as MongoDB, is able to support all the definitions at once.

The Results collection contains the results of an inference and its entities contain the output of the model along with the final pre-diagnosis given by it. They are stored in a different table, because despite the current implementation mapping an inference request to only one result, the system may be adapted in the future to trigger multiple model predictions with one inference request. In this case, it is useful to maintain the separation between the two entities so that this feature can be implemented more easily.

### 5.4.2 Implementation

The database service is a docker container pulling the NongoDB image. The image does not provide means to visualize the data stored in the DB such as with NATS and MinIO. However, there are third-party solutions that can be used to provide the direct access to MongoDB. The solution used in this case was a new container pulling the mongo-express image, which exposes a UI in a private port of the local machine to visualize the database, as show in [Figure 5.4](#).

```
1  mongo:
2    image: mongo
3    restart: always
4    env_file:
5      - .env
6
7  mongo-express:
8    image: mongo-express
9    restart: always
10   depends_on:
11     - mongo
12   env_file:
13     - .env
14   ports:
15     - '${ME_PORT}:${ME_INNER_PORT}'
```

The UI not only provides a visualization of the data, but also a way to manually add and change entities in the collections of the database.

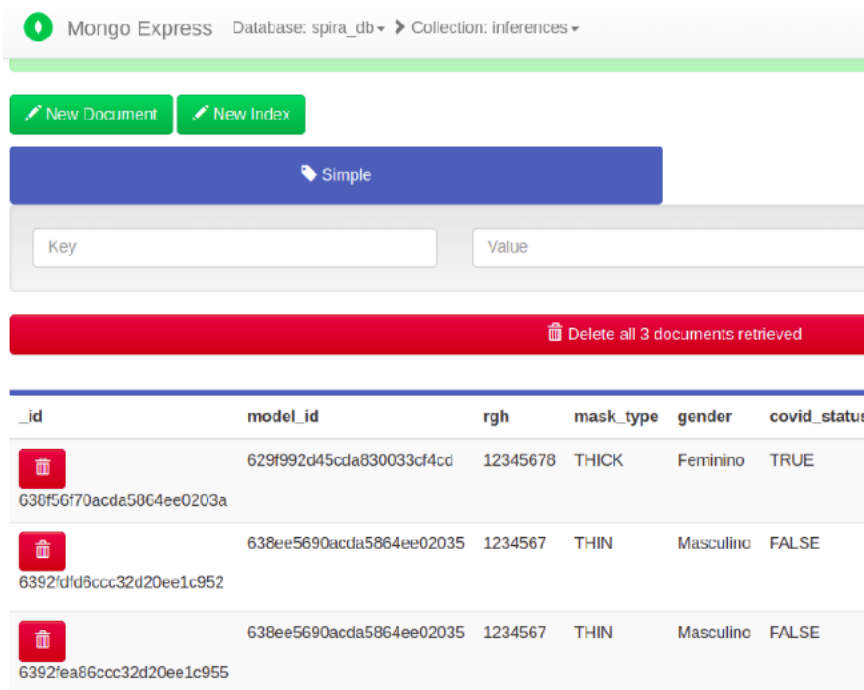


Figure 5.4: *Mongo Express UI. The UI can be used to directly interact with the database.*

## 5.5 Inference Requests

The main data flow of the API are the inference requests. Upon receiving a POST request at `/users/:user_id/inferences`, the API will receive the request and store a new inference entity in the database, following the sequence diagram described in Figure 5.5.

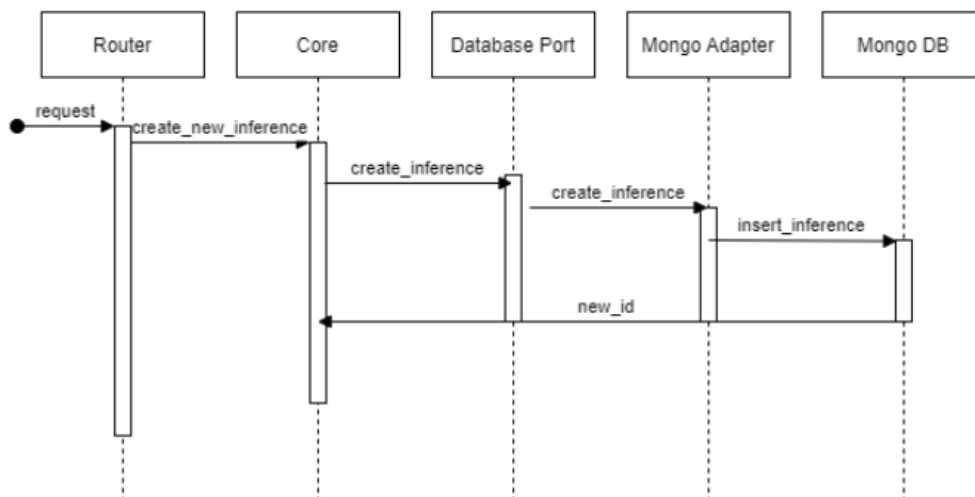
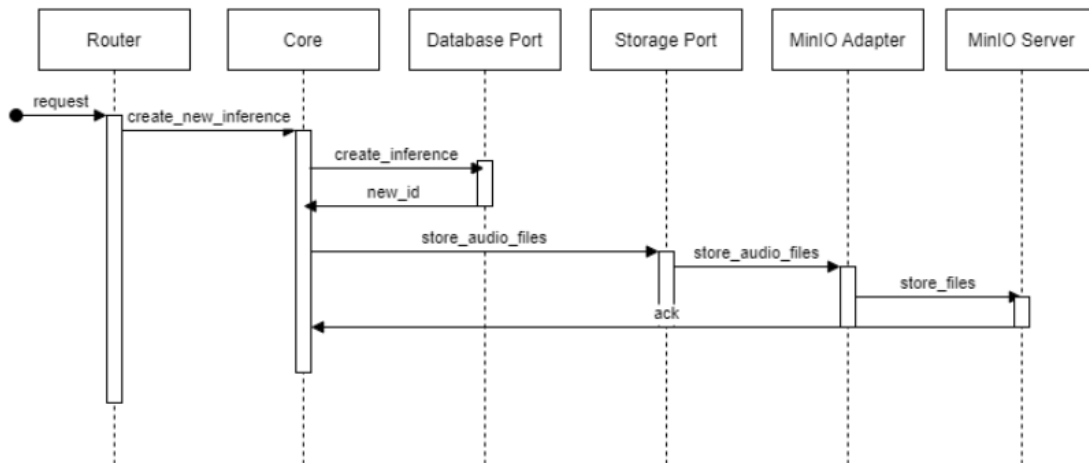


Figure 5.5: *Database sequence diagram for inference requests. Once inference requests are received, the web server calls the core, which then calls the database port to store the entity in the Mongo DB.*

Once the entity is inserted in the database, the new inference ID is returned to the core and the method will proceed storing the audio recording files from the body of the request in the MinIO server, following the sequence described in Figure 5.6.



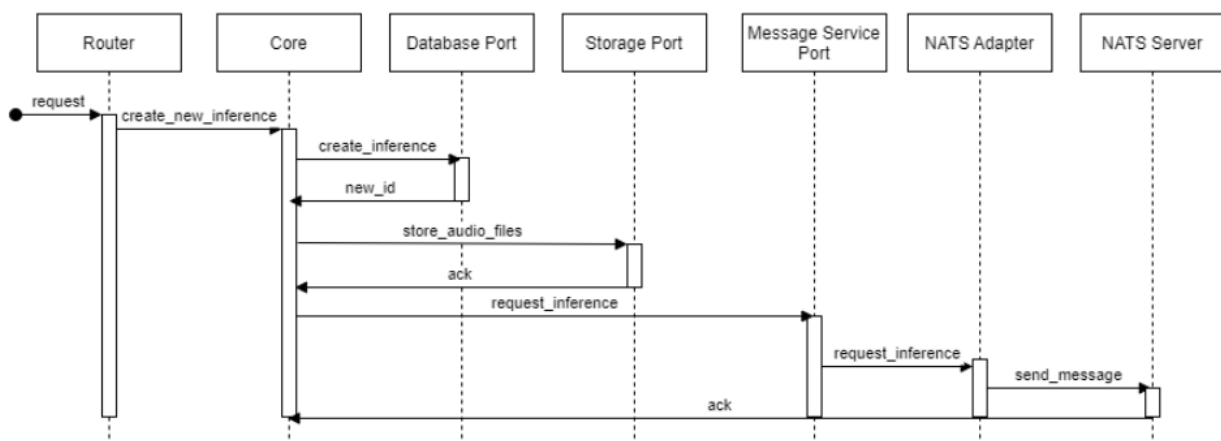
**Figure 5.6:** *File storage sequence diagram for inference requests.* Audio files are stored in MinIO server to be retrieved by Model Servers later.

The inference request is currently composed of four distinct audio recordings:

- **Acceptance:** verbal acceptance of the terms of service.
- **Sustained:** sustained vowel for as long as the subject can hold on.
- **Rhyme:** a nursery rhyme presented by the app.
- **Sentence:** a sentence presented by the app.

The four audio files are stored in the *audio-files* bucket, at a directory named with the ID obtained from the database insertion.

Finally, once the storage port acknowledges the core that the files were inserted, the method proceeds publishing a new message in the topic of the requested model. The sequence is described in [Figure 5.7](#). The message body contains the encoded inference entity described in [Section 5.4](#).



**Figure 5.7:** *Message service sequence diagram for inference requests.* Published message will be received by a Model Server subscribed to the model topic.

As explained in [Section 2.3](#), a microservice has no information regarding the health of another microservice. While the Model Server is able to decide whether to process a request or not due to momentary degradations, the API is prepared to receive a late or unsuccessful response from the Model Server. For this reason, the API sends a successful response to the client once the message

is successfully sent to the NATS broker, without waiting for an acknowledgment that the inference was received or processed. The API then *reacts* to incoming responses when a Model Server does manage to process an inference request, but does not stop its execution in case no response is obtained.

## 5.6 Deployment

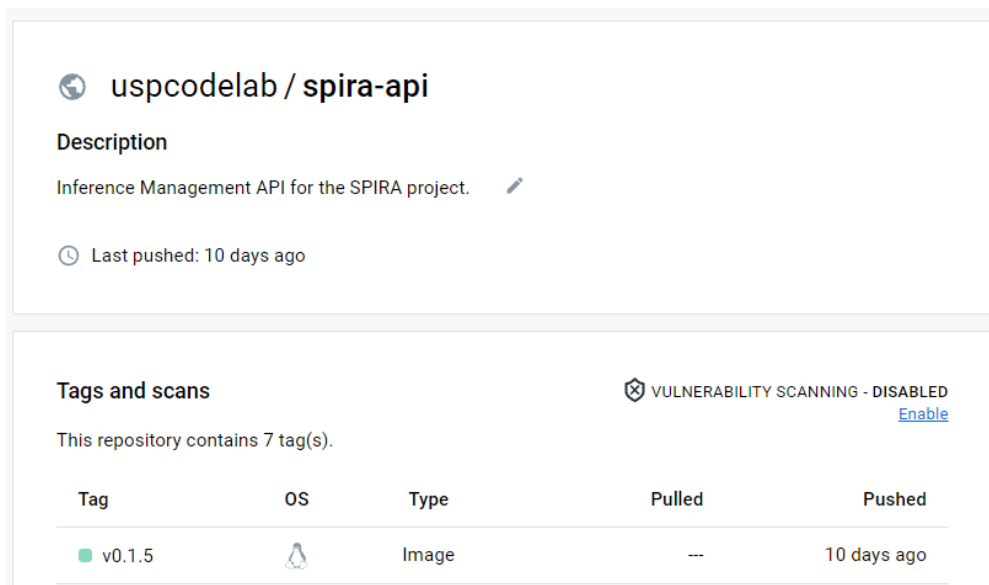
The docker image of the API is built through a Dockerfile that pulls a Python image. The resulting image can then be uploaded to DockerHub ([doc22](#)) as shown in [Figure 5.8](#), from where it can be pulled to assemble the production system.

The production repository is located at <https://github.com/spirabr/SPIRA-Inference-System>. It contains two docker-compose files, one solely for the model-server service and another containing all the rest of the back-end system. This separation is necessary so that Model Servers can be added and removed from the inference system without the need to restart the entire system.

The deploy is hosted in an on-premise server. The production repository was cloned in the machine and a `.service` file was created to run the API service as a systemd unit.

A continuous delivery pipeline was set up with GitHub Actions to automatically upload images to DockerHub upon a tag creation in the development repository.

GitHub Actions are configured through `.yaml` files in the `.github/workflows` directory inside the repository. The tags of the images were configured to be the same as the pushed tags in the repository.



**Figure 5.8: Docker Hub image repositories for the API service.** Images are automatically pushed to Docker Hub upon new tag creations in GitHub.

In order to prevent man-in-the-middle attacks between the UI client and the API, an HTTPS connection was setup between the two. The TLS certification was setup using [Caddy Server](#) as a reverse proxy to the orchestrator API.

The main advantage of using Caddy instead of other services such as [Nginx](#) is that Caddy has TLS automation features which are considered safe and modern by the literature ([ABC<sup>+</sup>19](#)), thus facilitating the configuration of a reverse proxy.

## Chapter 6

# Model Server

The architecture pattern used for the Model Server is the Hexagonal Architecture and is represented in Figure 6.1. Similarly to the API service, the core of the Model Server is purely functional, while ports and adapters are *singletons*.

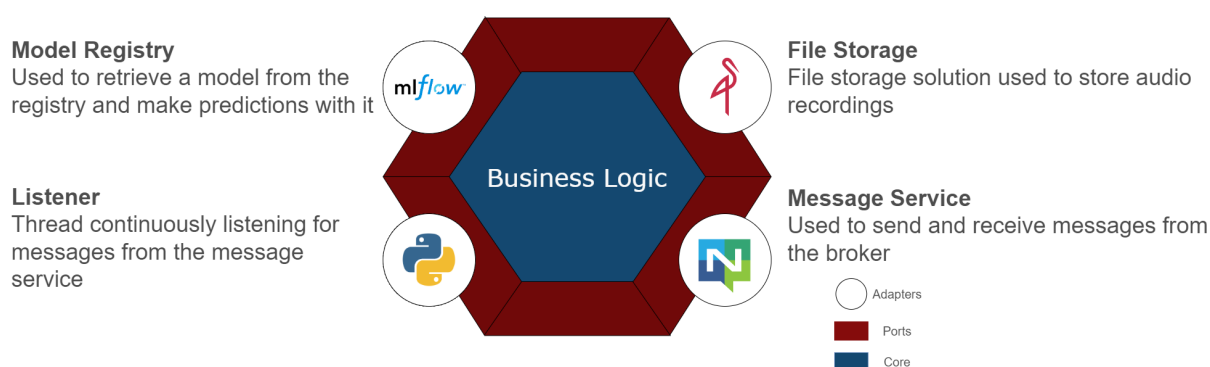


Figure 6.1: *Architecture of the Model Server.*

### 6.1 Registry Adapter

The model is recovered from the MLFlow server during Model Server initialization. More precisely, this retrieval happens inside the MLFlow adapter as presented below:

```
1 class MLFlowAdapter:
2     def __init__(self, conn_url, model_path):
3         logging.info("setting mlflow adapter.")
4         self._wait_for_server_connection()
5         mlf.set_registry_uri(conn_url)
6         logging.info("connected to mlflow server.")
7         self._model = mlf.pyfunc.load_model(model_uri=model_path)
8         logging.info("model loaded successfully.")
9
10    def predict(
11        self, inference: Inference, inference_files: InferenceFiles
12    ) -> Tuple[List[float], str]:
13        return self._model.predict([inference.dict(), inference_files.dict()])
```

When the adapter is first instantiated, it tries to connect to the MLFlow server. Once the connection is established, the correct model is loaded through a path passed as an environment variable to the adapter. The model instance is stored and will be subsequently used to make predictions.

As cited in Chapter 3, it is important to have a well-defined interface for prediction between the model instances and the MLFlow adapter.

The built-in standardized file system for Python models used by MLFlow is called Pyfunc. Along with the file system conventions, PyFunc also provides the abstract class `PythonModel` that should be inherited by the models in order to store them in the MLFlow server. SPIRA models will then need to implement the following template:

```

1  from mlflow.pyfunc import PythonModel
2
3  class ModelTemplate(PythonModel):
4
5      def load_context(self, context) -> None:
6          """
7              This method is called as soon as the model is instantiated.
8
9              The same context will also be available during calls to predict,
10             but it may be more efficient to override this method and load
11             artifacts from the context at model load time.
12
13             :param context: A :class:`~PythonModelContext` instance containing artifacts
14             that the model can use to perform inference.
15          """
16          pass
17
18      def predict(self, context, model_input) -> Tuple[List[float], str]:
19          """
20              Evaluates the inference files and metadata and returns a prediction.
21
22              :param context: A :class:`~PythonModelContext` instance containing artifacts
23              that the model can use to perform inference.
24              :param model_input: A list of two dictionaries, the first containing inference
25              metadata and the second containing audio file byte arrays.
26
27              returns: A Tuple where the first element is a list of numbers and the second is
28              the diagnosis string ("positive", "negative", "inconclusive").
29          """
30          pass

```

The `load_context` method is called during model initialization time and it is an optional method for custom initial setups.

The `predict` method is the one used to make inferences with the model. The expected input is a dictionary containing both the inference metadata and the byte arrays of the audio recordings, while the expected output is a Tuple where the first element is the model output, i.e., a list of numbers, and the second element is the diagnosis given by the model.

The adapter will then make use of this template to format the return value of the prediction into the result message for the API.

## 6.2 Inference Messages

The Model Server contains only one primary port given by a listener thread. The thread continuously calls the NATS adapter to look for incoming messages at the NATS server in a topic specified among the environment variables. Once a message is received, the core will call the storage port to retrieve the audio files stored by the API in the MinIO server, as described in [Figure 6.2](#).

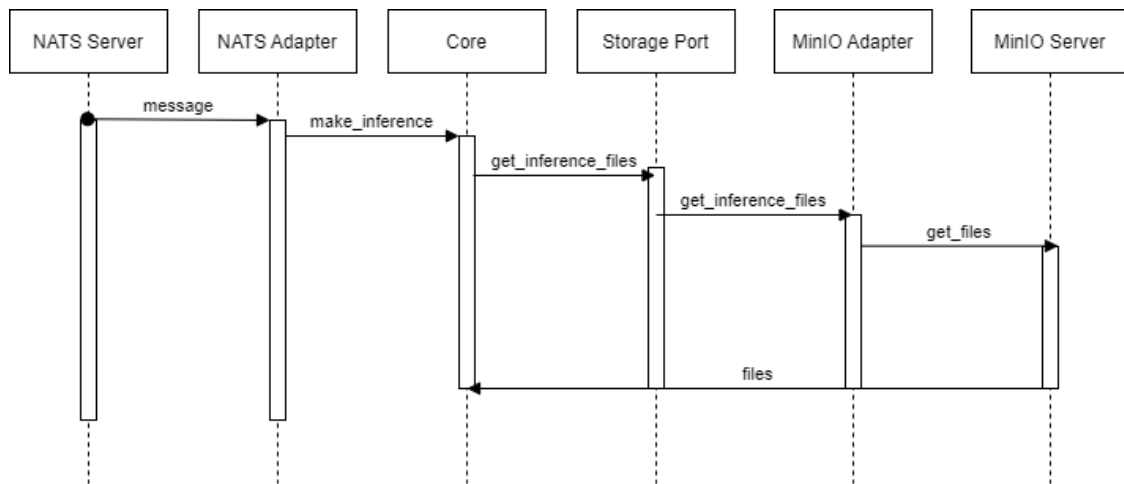


Figure 6.2: *Model Server retrieving audio files from MinIO server*

With the audio files and the inference metadata obtained from the received NATS message, the core is able to build the input for the ML model stored in the MLFlow adapter. The sequence of calls follows Figure 6.3. Once the model returns the prediction result, the core encodes it in a NATS message and publishes it in the *updates* topic being consumed by the orchestrator API.

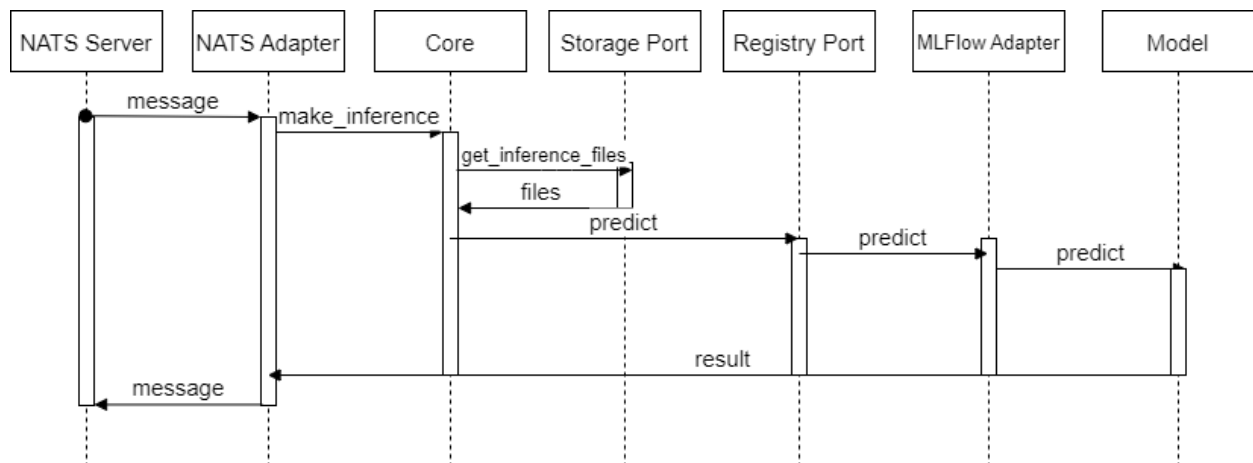


Figure 6.3: *Model Server using the model to make predictions*

## 6.3 Deployment

### 6.3.1 ML Pipeline

While MLFlow allows model creation directly through the UI, there are additional steps that are necessary for the deployment of the model besides registering it in MLFlow server. The deployment of a new model in the inference system can be summarized in the following steps:

1. **Model entity insertion in database:** A new model entity needs to be inserted in the database with its name and topic to be used in the message service broker to receive inference request messages.
2. **Model registration in MLFlow server:** The model should be loaded to a Python script which connects to the MLFlow server and registers it in the server.
3. **Deployment:** Once the model is available in the MLFlow server, the production repository ([inf22](#)) needs to be cloned to a new directory and the respective environment variables pointing to the model path in the MLFlow server should be set. The service is then ready to be started.



4. **Versioning:** New versions of the same model can also be uploaded to the registry. The staging and production versions of each model can be selected through the MLFlow UI shown in Figure 4.4.

The deployment is similar for all models in the system, therefore a pipeline executing the steps above was developed to automate it at <https://github.com/spirabr/Inference-System-ML-Pipeline>. Although it is still possible to make manual deploys to the system, the new pipeline is the recommended procedure to make models available in production.

The same pipeline can also be executed to upload new model versions to production. This can be done by specifying the name of an existent model, which is guaranteed to be unique in the database by the API service.

### 6.3.2 Microservice

The development repository (`mod22`) contains both the implementations of the Model Server and the MLFlow server. Similarly to the API, a CD pipeline was created with GitHub Actions to automatically upload both images to Docker Hub whenever a new tag is pushed to the repository.

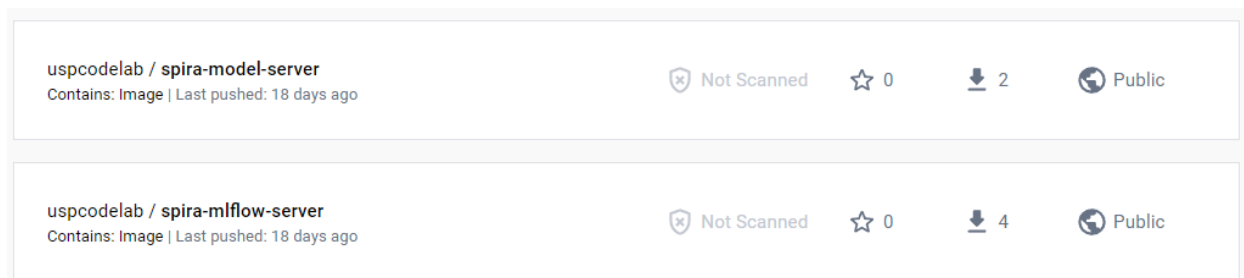


Figure 6.4: *Docker Hub repositories of the Model Server and the model registry*

As described in Chapter 5, the repository used in production is at <https://github.com/spirabr/SPIRA-Inference-System>. The MLFlow server is deployed in the same docker-compose file as the API service, thus they are part of the same systemd unit. The Model Server, on the other hand, is deployed as a separate unit so that Model Servers can be added, updated and deleted without interfering with the rest of the system.

# Chapter 7

## UI

As described in [Section 4.2](#), a collection PWA was developed inside SPIRA ([PWA22](#)). The collection App was designed to guarantee that the layout of the application does not interfere with the reading process of the patients.

To prevent additional bias from being added to the audio recording process, the new app was developed as a different deploy mode of the collection app, thus sharing a great part of the UI layout. The project repository is at <https://github.com/spirabr/PWA-App/tree/feature/inference-app>.

The Inference App is hosted inside [Netlify](#). The deployment requires a branch from the project repository which will serve as a production branch. Netlify then sets an automatic CD pipeline that updates the application deploy whenever a new push is made to this branch, without the need to use custom Webhooks.

Although the application was developed to be used mainly in mobile platforms, the PWA is also compatible with desktop platforms. Users can also choose between using the application directly from the browser or installing the app in the device as showcased in [Figure 7.1](#).



**Figure 7.1:** *Inference App installed in a mobile device.*

Despite being a deploy mode of the collection app, to work as intended, the new PWA required additional features that were not present in the collection PWA. The features implemented are described in the following sections.

### 7.1 User Access

While the API provides the mechanisms for authentication, as seen in [Section 5.3](#), the UI needs to react accordingly blocking the access of unauthorized users to the application and forbidding users from requesting inference data that they do not own.

The new application contains a new *Sign In* page that users must pass through to access the remaining UI. The credentials from the page are sent to the API where they are validated. Once the client is authenticated, the application proceeds storing the token locally to send it in the header of all subsequent requests until it expires or the user signs out of the system. In these cases, the user is redirected to the Sign In page. This new page is presented in [Figure 7.5](#).

## 7.2 Inference Form

As cited before, inference and collection processes are similar regarding input generation. Inferences require one additional information regarding the model to be used to make the prediction.

In accordance to that, the new form contains a new select field that retrieves the list of models from the API and the final request now includes the ID of the selected model as shown in [Figure 7.3](#).

Once the form is filed, the user is redirected to make the voice recordings, as shown in [Figure 7.4](#). At the end of the recording process, the user may send the inference request to the API.

## 7.3 Inference List

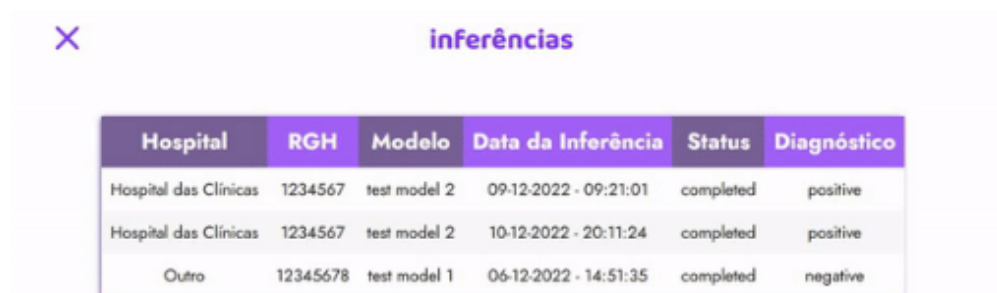
After the inference is registered in the back-end, the user may revisit them and check their status through the inferences page. Medical personnel utilize RGH as a unique identifier to the patients. In order to identify an inference, users need to be provided with the following information along with the final diagnosis:

- **Hospital:** Hospital from where the inference was sent. This column is necessary in case inference requestors send requests from different hospitals.
- **Patient RGH:** RGH is the unique patient identifier used by a hospital.
- **Model:** Inferences are intended to be directed to a model. Users should be provided with the model that they chose for a particular inference.
- **Inference Date:** As many inferences may be requested for the same user with the same model, the timestamp of the inference request is also provided for users to identify the inferences they are looking for.
- **Status:** As cited, users need information regarding the stage of the inference request.
- **Diagnosis:** The diagnosis is needed to show the result of the inference once it is completed or it halts due to an error.

Users can identify an inference with the first four columns and infer the current situation of it with the last two columns. The final desktop layout of the page is presented in [Figure 7.2](#). It is possible to sort the inferences in ascending or descending order with respect to any of the parameters by clicking in the headers of the table.

Due to the width of the mobile screens, a different layout was necessary to show the results for mobile users. The mobile display of the same page is presented in [Figure 7.6](#)

Similarly to the desktop version, users are able to sort the inferences by selecting the field at the top of the page.



Hospital	RGH	Modelo	Data da Inferência	Status	Diagnóstico
Hospital das Clínicas	1234567	test model 2	09-12-2022 - 09:21:01	completed	positive
Hospital das Clínicas	1234567	test model 2	10-12-2022 - 20:11:24	completed	positive
Outro	12345678	test model 1	06-12-2022 - 14:51:35	completed	negative

Figure 7.2: Inference list page for desktop users

**Figure 7.3: Inference form page.** Users fill the form before making the audio recordings.

**Figure 7.4: Inference audio recording page.** Users need to follow the instructions in the screen to record the audios.

**Figure 7.5: Sign In page in the inference app.** Users should have an account to access the system.

**Figure 7.6: Inference form page for mobile users.** The application changes the layout of the page to a cards display instead of a table.

# Chapter 8

## Tests

### 8.1 TDD

Tests were a major step in the development process of the inference system to ensure it works correctly. Therefore, development techniques that are strongly related to tests were adopted, especially the TDD (Bec03).

TDD, or Test Driven Development, is a technique that uses short development and testing cycles to implement the requirements of an application. While TDD was not used for the whole development of the research, a substantial part of the features was implemented using this method.

The development cycle consists of the following steps:

1. Write a test that covers a requirement of the system.
2. Develop the program to pass in the tests.
3. Refactor the code.
4. Repeat the process until all requirements are satisfied.

This way, tests written by the developer are highly related to the requirements of the system without being biased by the implementation details to achieve them. The technique also encourages developers to write more tests as they become a standard procedure for the implementation of new features in the project (EMT05).

While a certain overhead is added to the development process due to the necessity of writing new tests, the return of investment obtained from adopting TDD is highly correlated with Quality Assurance benefits (MP03). While the current project aims to provide a *definitive* solution for SPIRA model inferences, this research assumed that the advantages of TDD will outweigh the disadvantages in the long term.

### 8.2 Test Hierarchy

Tests made can be subdivided in categories, according to their scope:

- **Unit Tests:** Responsible for testing the behavior of classes and methods.
- **Integration Tests:** Responsible for testing the connection between components.
- **System Tests:** Responsible for testing major system data flows.

### 8.2.1 Unit Tests

Unit tests were made with **Pytest**. They were used to test the behavior of classes and methods inside the model server and the API service. The three main sections of unit tests are divided in tests for the adapters, ports, and the core. In the case of adapters and ports, tests were used to verify the behavior of the classes public methods, whereas for the services located inside the core, the methods themselves were directly tested, as the use cases are purely functional.

The following example is a unit test to check whether the adapter is retrieving a model properly from the database.

```
1 def test_get_model_by_id(database_adapter: MongoAdapter):
2     model = database_adapter.get_model_by_id("629f992d45cda830033cf4cd")
3
4     assert model == {
5         "_id": ObjectId("629f992d45cda830033cf4cd"),
6         "name": "fake_model",
7         "publishing_channel": "fake_channel_2",
8     }
```

The unit test above is responsible for testing whether the behavior of the database adapter is correct. To fully isolate the adapter and only test its behavior, both ends were substituted by mocks. On one hand, the MongoDB client used by the adapter instance is mocked to return false values upon calls to the database. On the other hand, the calls made to the adapter, which should come from the ports layer, is also faked by directly calling the method inside the test, further isolating it from the business logic of the microservice. Thi way, the test can check whether the behavior is correct without the interference of the database connection or the behavior of other layers inside the microservice.

With each layer of the hexagonal architecture being isolated through the mocks, the origin of an error can be more easily identified, as unit tests will point out the layer and class responsible for the faulty behavior.

### 8.2.2 Integration Tests

Integration tests were used to check the connection between the containers of the system. Unlike unit tests where connections are mocked, integration tests require the components of the system to be lifted in order to perform them. While Pytest was still used to make the tests, a special tester service was created containing both the microservice source code and the test code, so that after the containers are started, the tests can be executed from within the microservice container. More details on how the tester service was set up are given in [Section 8.3](#).

In the same example using database tests, the same test was made *without mocking the MongoDB client* used by the adapter instance. Hence, the test now depends on the connection between the adapter and the database being properly set up to succeed.

Note that repeating the same test without the mocked database allows one to diagnose the root of an error within the system more efficiently. Failing in the unit test first would imply that the behavior of the method itself is faulty, while passing in the unit test but not in the integration test implies that the adapter behavior is correct, but the connection with the database is not. The procedure was then repeated with other components in the system.

Another set of tests was also used to test the integration between the core, ports and adapters. While they do not check the connection between components, they are useful to check the cohesion between the layers of the main services in the system.

### 8.2.3 System Tests

System tests aim to test the application from end to end. In this case, the entire project is started and requests or messages are artificially sent *from outside* the docker network to test the expected behaviors of the system. Therefore instead of using python scripts, tests were created using bash scripts that are called from outside the containers. The following example is a bash script testing the retrieval of inferences, including expected behaviors on malformed, unauthorized, and forbidden cases.

```

1  #!/bin/bash
2  # some details were ommited for the sake of visualization
3  RESPONSE_STATUS="$(curl \
4      --write-out '%{http_code}' --silent --output /dev/null \
5      'localhost:3000/v1/users/639686c4ba1604f1387a6c00 \
6      /inferences/638f56f70acda5864ee0203a' \
7      --header "Authorization: Bearer $TOKEN")"
8
9  ...
10
11  FORBID_RESPONSE_STATUS="$(curl \
12      --write-out '%{http_code}' --silent --output /dev/null \
13      'localhost:3000/v1/users/639686c4ba1604f1387a6c01 \
14      /inferences/638f56f70acda5864ee0203b'\
15      --header "Authorization: Bearer $TOKEN")"
16
17  # tests a successfull request
18  if [ "$RESPONSE_STATUS" -eq 200 ]
19  then
20      echo 'PASSED';
21  else
22      echo 'FAILED';
23      exit 1;
24  fi;
25
26  ...
27
28  # tests a forbidden request
29  if [ "$FORBID_RESPONSE_STATUS" -eq 403 ]
30  then
31      echo 'PASSED';
32  else
33      echo 'FAILED';
34      exit 1;
35  fi;

```

Combined with unit and integration tests, system tests are useful to inspect problems in the general data flow without the need to concern with the behavior of particular methods, as they were already tested in unit tests, or the connection between containers inside the docker network, since they were also checked in integration tests.

They are also helpful to preserve the behavior of the system during feature requests, which further ensures that the back-end will behave as expected for outside components such as the UI client.

## 8.3 Multi-Stage Build

Unit tests of single methods can be done by directly running python test files. However, integration tests that are responsible for testing the connection between the components of the project require a more sophisticated approach.

One way to solve this problem is to use Docker multi-stage builds. Multi-Stage builds is a feature of Docker used to define various build stages in a single Dockerfile. Developers can then define the stage in which the system should stop building and use the image for specific purposes such as tests and deployments. The Dockerfile of the current project contains three different build stages:

- **base:**  
base is responsible for configuring the environment to be used by the next stages. It pulls the Python image and installs **Poetry** to manage the dependencies afterwards. It is also responsible for setting the environment variables.
- **dev:**  
dev is the stage that copies the source code and the test code to the image. This image is used to run the tests with the containerized system.
- **prod:**  
prod is the stage that only copies the source code and its necessary dependencies. The image is built to run in production.

For each test subroutine at all levels of the hierarchy, the whole system is lifted and stopped before the next subroutine. Particularly, both the API and the model server containers are built with their respective dev images, which contain both source and test code. Unit tests and connection tests are then performed from inside these central containers.

In contrast, system tests are performed with the prod image, where no test code is contained in the project and tests are execute from outside the system.

## 8.4 CI Pipeline

Similarly to the CD pipeline, the CI Pipeline was implemented in the project with GitHub Actions. GitHub Actions provide a runner to run custom jobs given a certain event. For both development repositories, it was decided that the pipeline would run upon pull requests to the main branch.

The execution of tests was automated using Makefiles. They were particularly useful due to the possibility of organizing them in test sections using makefile targets. Each section can then be called separately during development, and a central target was made to run all sections, starting from unit tests, then integration tests and, finally, system tests.

In total, around **170 tests** were created for the model server and the API service combined, including unit, integration and system tests.



# Chapter 9

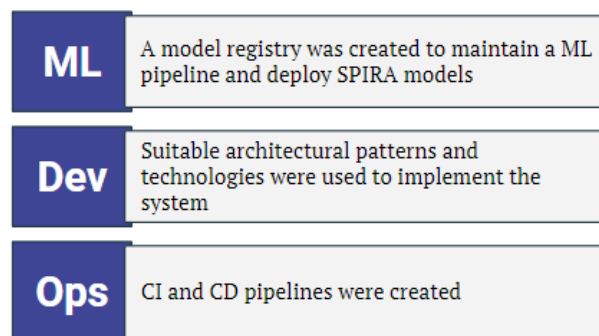
## Conclusion

### 9.1 Results

An intelligent reactive microservices system was implemented to process respiratory insufficiency inference requests. The system is fully deployed and medical personnel are able to make use of it once the next generation of SPIRA models are deployed in the system.

To achieve this objective, the current research was divided in 4 main stages. [Chapter 2](#) was the research for theoretical concepts that were relevant for the planning of the system. [Chapter 3](#) focused on the lessons learned from a preliminary version of the inference system. [Chapter 4](#) focused on the architectural decisions made based on the acquired concepts and lessons learned. Finally, [Chapter 5](#), [Chapter 6](#) and [Chapter 7](#) showed the implementation process of each component of the final inference system.

Lessons learned from [Chapter 3](#) showed that MLOps practices were crucial for the quality assurance of the system. The adoption of these practices was made through the creation of a ML pipeline to automate the deploy of new models to the inference system, the creation of CI/CD pipelines to automate the testing and deploy of new features, and the adoption of appropriate patterns to implement the architecture and design of the microservices.



**Figure 9.1:** *MLOps milestones achieved in the research*

Applying MLOps principles solved the main challenges faced in the preliminary implementation: reproducibility and ease of deploy.

Finally, an [article](#) describing some of the experiences of this research was written and published ([FGT<sup>+</sup>22](#)).

### 9.2 Next steps

While the inference system was successfully implemented, there are still improvements to be made to the project.

### 9.2.1 Model feedback

While the system offers inference services as initially planned, the inference data stored in the system may be used as feedback input to further train the models and continuously improve their performance. Nevertheless, this feedback process should be carefully planned, as a poorly implemented pipeline may leave the model exposed to possible adversarial attacks to worsen its performance, which would be extremely harmful for the project.

### 9.2.2 NATS improvements

As the system demand and the number of model servers increase, the load of messages managed by NATS broker is only set to grow. In that case, adding clustering support for NATS would be a great addition for the sake of horizontal scalability of the system.

### 9.2.3 Load balancer

While the definitive system is fully deployed, there is only one API service instance running in the host machine. If the load of the service grows, the ideal solution is to create more instances of the services and a load balancer to distribute the loads between them. This can be accomplished with the use of Docker Stacks, which were not used in the current research.

### 9.2.4 UI improvements

The current UI does satisfy all the requirements needed. However, there are currently bug-fix requests for the collection App that also affect the inference App, as they share a great part of its code base.

### 9.2.5 Training pipeline

The MLFlow server was a great addition to the inference system to implement the best practices of MLOps. However, SPIRA models are now required to follow a defined template and adapt to the correct I/O formats in order to be stored in MLFlow server. A pipeline can be built to make trained models compatible with the Model Server.

# Bibliography

- [ABC<sup>+</sup>19] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, et al. Let's encrypt: an automated certificate authority to encrypt the entire web. In **Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security**, pages 2473–2487, 2019. 31
- [Bec03] Kent Beck. **Test-driven development: by example**. Addison-Wesley Professional, 2003. 39
- [Bon16] Jonas Bonér. **Reactive Microservices Architecture**. O'Reilly Media, Inc., 2016. 5, 7
- [Cam65] EJM Campbell. Respiratory failure. **British medical journal**, 1(5448):1451, 1965. 3
- [CCN<sup>+</sup>00] Jason W Chien, Russell Ciufo, Ronald Novak, Mary Skowronski, JoAnn Nelson, Albert Coreno, and ER McFadden Jr. Uncontrolled oxygen administration and respiratory failure in acute asthma. **Chest**, 117(3):728–733, 2000. 3
- [CF20] Jennifer Couzin-Frankel. The mystery of the pandemic's 'happy hypoxia', 2020. 3
- [CGC<sup>+</sup>21] Edresson Casanova, Lucas Gris, Augusto Camargo, Daniel da Silva, Murilo Gazzola, Ester Sabino, Anna Levin, Arnaldo Candido Jr, Sandra Aluisio, and Marcelo Finger. Deep learning against COVID-19: Respiratory insufficiency detection in Brazilian Portuguese speech. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pages 625–633, Online, August 2021. Association for Computational Linguistics. 3, 10, 11, 14
- [Coc] Alistair Cockburn. **Hexagonal architecture**. 8, 23
- [DLT<sup>+</sup>17] Nicola Dragoni, Ivan Lanese, Stephan Thordal Larsen, Manuel Mazzara, Ruslan Mustafin, and Larisa Safina. Microservices: How To Make Your Application Scale. **arXiv e-prints**, page arXiv:1702.07149, February 2017. 6
- [doc22] Spira images. <https://hub.docker.com/u/uspcodelab>, 2022. Accessed: 2022-10-26. 31
- [DS94] Joseph V DiCarlo and James M Steven. Respiratory failure in congenital heart disease. **Pediatric Clinics of North America**, 41(3):525–542, 1994. 3
- [EGHS16] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. Devops. **Ieee Software**, 33(3):94–100, 2016. 5
- [EMT05] Hakan Erdogmus, Maurizio Morisio, and Marco Torchiano. On the effectiveness of the test-first approach to programming. **IEEE Transactions on software Engineering**, 31(3):226–237, 2005. 39
- [FGT<sup>+</sup>22] Renato Cordeiro Ferreira, Dayanne Gomes, Vitor Tamae, Francisco Wernke, and Alfredo Goldman. Spira: Building an intelligent system for respiratory insufficiency detection.

- In **Anais do II Workshop Brasileiro de Engenharia de Software Inteligente**, pages 19–22. SBC, 2022. 43
- [FY97] Brian Foote and Joseph Yoder. Big ball of mud. **Pattern languages of program design**, 4:654–692, 1997. 6
- [Hul18] Geoff Hulten. **Building Intelligent Systems**. Apress, 2018. 4
- [Hut18] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018. 5
- [inf22] Spira inference system. <https://github.com/spirabr/SPIRA-Inference-System>, 2022. Accessed: 2022-10-26. 34
- [jef08] The onion architecture : part 1. <https://jeffreypalermo.com/2008/07/the-onion-architecture-part-1/>, 2008. Accessed: 2022-12-18. 8
- [jwt22] Jwt. <https://jwt.io/>, 2022. Accessed: 2022-11-29. 25
- [lgp18] Lei nº 13.709, de 14 de agosto de 2018. [http://www.planalto.gov.br/ccivil\\_03/\\_Ato2015-2018/2018/Lei/L13709.htm](http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm), 2018. Accessed: 2022-12-23. 19
- [Mar96] Robert C Martin. The dependency inversion principle. **C++ Report**, 8(6):61–66, 1996. 7
- [MGB<sup>+</sup>18] Robert C Martin, James Grenning, Simon Brown, Kevlin Henney, and Jason Gorman. **Clean architecture: a craftsman’s guide to software structure and design**. Number s 31. Prentice Hall, 2018. 8
- [min22] Minio. <https://min.io/>, 2022. Accessed: 2022-11-06. 19
- [mlf22] Mlflow tracking server. <https://www.mlflow.org/docs/latest/tracking.html>, 2022. Accessed: 2022-11-06. 15
- [mod22] Spira inference service. <https://github.com/spirabr/SPIRA-Inference-Service>, 2022. Accessed: 2022-12-29. 15, 35
- [MP03] Matthias M Müller and Frank Padberg. About the return on investment of test-driven development. In **Edser-5 5 th international workshop on economic-driven software engineering research**, page 26, 2003. 39
- [nat22] Nats. <https://nats.io/>, 2022. Accessed: 2022-11-06. 17
- [PWA22] Spira pwa app. <https://github.com/spirabr/PWA-App/>, 2022. Accessed: 2022-10-18. 36
- [rea14] Reactive manifesto. <https://www.reactivemanifesto.org/>, 2014. 6, 7
- [reg22] Mlflow server. [https://github.com/Fernando-Freire/MLFlow\\_docker\\_compose\\_template](https://github.com/Fernando-Freire/MLFlow_docker_compose_template), 2022. Accessed: 2022-10-30. 15
- [RK03] Charis Roussos and A Koutsoukou. Respiratory failure. **European Respiratory Journal**, 22(47 suppl):3s–14s, 2003. 3
- [RTA<sup>+</sup>21] Ahsab Rahman, Tahani Tabassum, Yusha Araf, Abdullah Al Nahid, Md Ullah, Mohammad Jakir Hosen, et al. Silent hypoxia in covid-19: pathomechanism and possible management strategy. **Molecular biology reports**, 48(4):3863–3869, 2021. 3
- [Rus15] Alex. Russell. "progressive web apps: Escaping tabs without losing our soul". retrieved june 15, 2015. <https://infrequently.org/2015/06/progressive-apps-escaping-tabs-without-losing-our-soul/>, 2015. 9

- [SLK09] Meghna R Sebastian, Rakesh Lodha, and SK Kabra. Swine origin influenza (swine flu). **The Indian Journal of Pediatrics**, 76(8):833–841, 2009. 3
- [SVdB22] Deeks JJ Dinnes J Takwoingi Y Davenport C Leeﬂang MMG Spijker R Hooft L Emperador D Domen J Tans A Janssens S Wickramasinghe D Lannoy V Horn SR A Struyf, T and A Van den Bruel. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19. **Cochrane Database of Systematic Reviews**, (5), 2022. 3
- [Teo20] Jason Teo. Early detection of silent hypoxia in covid-19 pneumonia using smartphone pulse oximetry. **J. Med. Syst.**, 44(8):134, June 2020. 3
- [TLJ20] Martin J Tobin, Franco Laghi, and Amal Jubran. Why COVID-19 silent hypoxemia is baffling to physicians. **Am. J. Respir. Crit. Care Med.**, 202(3):356–360, August 2020. 3, 4
- [TOS<sup>+</sup>20] Mark Treveil, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, and Lynn Heidmann. **Introducing MLOps**. O’Reilly Media, 2020. 5