Cloud-Based Architectures for Scientific Health Systems: Case Studies Using AWS

Kayo Henrique de Carvalho Monteiro[1], Sebastião Rogerio da Silva Neto[1], Igor Vitor Teixeira[1], Elisson da Silva Rocha[1,3], Cleber Matos de Morais[1,2], Patricia Takako Endo[1,3]

[1] dotLAB Brazil
Caruaru – PE – Brazil

[2]Universidade Federal da Paraíba
João Pessoa – PB – Brazil

[3]Universidade de Pernambuco (UPE)
Recife – PE – Brazil

{srsn, ivt}@ecomp.poli.br

cleber.morais@academico.ufpb.br

{kayo.henrique, elisson.rocha, patricia.endo}@upe.br

Abstract. *The growing availability of large-scale health data has created new opportunities for applying machine learning in public health systems. However, processing these datasets requires scalable computational infrastructures. This paper presents an experience report on the development of cloud-based architectures for scientific health systems using Amazon Web Services (AWS). We analyze three platforms: VALERIA for arboviral disease diagnosis, ANGELS for gestational monitoring, and IAra for malaria forecasting. These systems rely on large epidemiological datasets and machine learning pipelines executed using AWS services such as EC2, S3, ECR, Lambda, and SageMaker. Results show that cloud infrastructures enable scalable data processing, reproducible experimentation, and operational deployment of intelligent healthcare applications.*

Resumo. *A crescente disponibilidade de dados clínicos e epidemiológicos tem ampliado as oportunidades de aplicação de aprendizado de máquina em sistemas de saúde pública. Entretanto, o processamento desses dados requer infraestruturas computacionais escaláveis. Este artigo apresenta um relato de experiência sobre o desenvolvimento de arquiteturas em nuvem utilizando Amazon Web Services (AWS). São analisadas três plataformas: VALERIA, para apoio ao diagnóstico de arboviroses; ANGELS, para monitoramento gestacional; e IAra, para previsão epidemiológica de malária. Os sistemas utilizam serviços como EC2, S3, ECR, Lambda e SageMaker para processamento de dados e treinamento de modelos, demonstrando o papel da computação em nuvem no suporte a aplicações de saúde baseadas em dados.*

1. Introduction

The increasing digitalization of healthcare systems has resulted in the continuous generation of large volumes of clinical and epidemiological data. Health information systems,

electronic medical records, and disease surveillance platforms collect detailed information about patients, clinical events, laboratory tests, and demographic characteristics. This growing availability of health data has created new opportunities for the application of machine learning and data-driven approaches to support medical decision-making and public health management [Rajkomar et al. 2019, World Health Organization 2020].

Machine learning techniques have been widely explored in the biomedical domain to identify patterns in complex datasets and support predictive analysis. These techniques have been applied to a wide range of healthcare problems, including disease diagnosis, prognosis prediction, patient risk stratification, and epidemiological surveillance [Esteva et al. 2019, Miotto et al. 2018]. For instance, machine learning models have been used to predict disease progression and mortality risks based on demographic, clinical, and laboratory attributes extracted from large healthcare databases [Barros et al. 2021].

Recent studies highlight the potential of machine learning techniques to support clinical prognosis and decision making in infectious diseases. By analyzing patient attributes and historical records, predictive models can estimate the probability of adverse outcomes and assist healthcare professionals in selecting appropriate treatment strategies [Rajkomar et al. 2019]. Such approaches are particularly relevant in public health contexts where early identification of risk factors may improve patient outcomes and optimize resource allocation.

However, the practical implementation of machine learning models in healthcare systems presents several challenges. Healthcare datasets are typically large, heterogeneous, and often contain incomplete or inconsistent records. Data preprocessing, feature selection, model training, and hyperparameter optimization require substantial computational resources, particularly when multiple algorithms and experimental scenarios are evaluated [Miotto et al. 2018].

Another important challenge lies in the integration of predictive models into operational healthcare platforms. While many studies focus on the development of predictive models, fewer works address the design of scalable systems capable of supporting the full lifecycle of data analysis, including data ingestion, model training, deployment, and real-time inference.

Cloud computing has emerged as a key enabling technology for addressing these challenges. Cloud platforms provide scalable storage and computational resources that allow researchers to process large datasets and execute complex machine learning workflows without the need for dedicated high-performance infrastructure [Monteiro et al. 2018].

This paper presents an experience report on the development of cloud-based architectures for scientific health systems using Amazon Web Services (AWS). The study analyzes three operational platforms developed for public health applications: VALERIA, a clinical decision support system for arboviral diseases; ANGELS, an intelligent gestational follow-up system; and IAra, a malaria forecasting system. These systems illustrate how machine learning models can be integrated with scalable cloud infrastructure to support large-scale health data analysis and operational decision support.

2. Background

2.1. Machine Learning in Healthcare

Machine learning has become an important tool for analyzing biomedical data and supporting healthcare decision-making. These techniques enable the identification of patterns in complex datasets, supporting tasks such as disease diagnosis, prognosis prediction, risk assessment, and treatment recommendation [Rajkomar et al. 2019, Esteva et al. 2019].

In infectious disease research, machine learning models have been applied to clinical and epidemiological datasets to identify risk factors, predict disease outcomes, and support public health decision-making [Barros et al. 2021]. More recently, deep learning approaches have expanded these capabilities by enabling automatic feature extraction from raw biomedical data, reducing the need for manual feature engineering and improving the analysis of large healthcare datasets [LeCun et al. 2015, Miotto et al. 2018].

2.2. Connected Health Systems and Data-Driven Healthcare

Advances in mobile computing, wearable sensors, and Internet of Things (IoT) technologies have enabled connected health systems capable of continuously monitoring patient conditions through real-time collection of physiological and behavioral data [Monteiro et al. 2018]. However, due to the limited computational and storage capabilities of IoT devices, modern e-health architectures integrate these devices with fog and cloud infrastructures, where edge components perform preliminary processing while cloud platforms provide scalable resources for large-scale data analysis and machine learning [Monteiro et al. 2018].

2.3. Cloud Computing for Scientific Health Systems

Cloud computing provides scalable infrastructure for processing large healthcare datasets and executing computationally intensive workloads such as machine learning training and large-scale data analysis [Miotto et al. 2018].

In scientific health systems, cloud platforms support multiple stages of the data analytics pipeline, including data storage, preprocessing, model training, and deployment of predictive services. These capabilities enable reproducible experiments, collaborative research, and scalable deployment of data-driven healthcare applications.

3. Cloud-Based Architecture using AWS

Scientific health systems that rely on large-scale epidemiological and clinical datasets require computational infrastructures capable of supporting the entire lifecycle of machine learning applications. This lifecycle includes data ingestion, storage, preprocessing, model training, model versioning, and deployment of predictive services.

Cloud computing platforms provide scalable and flexible infrastructures that allow researchers to execute computationally intensive workloads while ensuring reproducibility and operational reliability. In this work, the analyzed systems were implemented using Amazon Web Services (AWS), which offers services for distributed storage, high-performance computing, container orchestration, and serverless execution.

The proposed architecture follows a modular design composed of four main layers: *Data Sources (Ingestion & Storage)*, *Processing (Preparation & Training)*, *Model Management (Containerization & Registry)*, and *Operational Inference (Serverless & Applications)*. Figure 1 illustrates the conceptual architecture adopted.

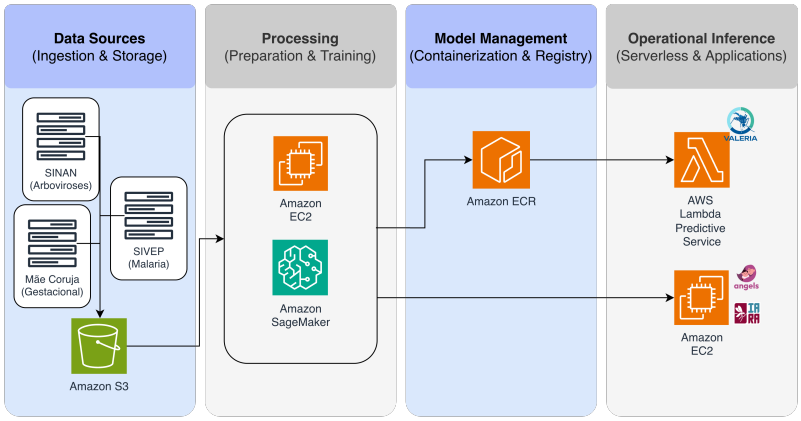

Figura 1. Cloud-based architecture for scientific health systems using AWS.

3.1. Data Sources (Ingestion & Storage)

The data sources layer is responsible for data ingestion and persistent storage of heterogeneous health datasets. As illustrated in Figure 1, this layer integrates multiple data sources, including epidemiological systems such as SINAN (arboviruses), SIVEP-Malaria, and maternal health databases such as *Mãe Coruja*.

These datasets are ingested and consolidated into a centralized storage repository using Amazon Simple Storage Service (S3). S3 provides high durability, scalability, and efficient access to large volumes of data, enabling the storage of raw data, preprocessed datasets, model artifacts, and experiment logs. This layer ensures data availability and supports reproducibility by maintaining versioned datasets that can be reused across multiple experimental pipelines.

3.2. Processing (Preparation & Training)

The processing layer is responsible for data preparation, feature engineering, and model training. This stage involves transforming raw data into structured inputs suitable for machine learning workflows.

As shown in Figure 1, Amazon EC2 instances are used to execute computationally intensive tasks such as data preprocessing, feature extraction, and large-scale model training. EC2 enables dynamic allocation of computing resources, allowing scalability according to dataset size and experiment complexity.

Additionally, Amazon SageMaker is used to manage structured machine learning pipelines, including training, evaluation, and experiment tracking. SageMaker facilitates reproducibility and supports iterative experimentation with multiple models and configurations.

3.3. Model Management (Containerization & Registry)

The model management layer is responsible for ensuring reproducibility, versioning, and deployment readiness of machine learning models.

In this architecture, machine learning pipelines and application components are encapsulated using Docker containers. These containers are stored and versioned in Amazon Elastic Container Registry (ECR), as illustrated in Figure 1.

This approach ensures consistent execution across development, testing, and production environments, while enabling modular deployment and collaboration among research teams. Containerization also simplifies the integration between training pipelines and inference services.

3.4. Operational Inference (Serverless & Applications)

The operational inference layer is responsible for deploying trained models and delivering predictions to end users through applications and services. As illustrated in Figure 1, inference is primarily implemented using AWS Lambda, enabling serverless execution with automatic scaling and reduced infrastructure management. In some cases, EC2-based services are used to host APIs and application backends supporting systems such as VALERIA, ANGELS, and IAra. This layer enables the integration of machine learning models into real-world healthcare applications, including dashboards and decision-support systems, ensuring that predictions are accessible and usable in operational environments.

4. Experience Report: Scientific Health Systems Using AWS

This section reports the experience of developing three scientific health systems supported by AWS-based cloud infrastructures. Although the systems address different public health problems, they share a common architectural rationale: all of them rely on large and heterogeneous health datasets, demand extensive preprocessing and feature engineering, and require repeated training and evaluation of machine learning models under multiple experimental configurations. In this context, AWS resources were adopted not merely as deployment infrastructure, but as an enabling computational layer for scalable scientific experimentation.

4.1. VALERIA: Clinical Decision Support for Arboviral Diseases

VALERIA[1] is a clinical decision support platform developed to assist healthcare professionals in the differential diagnosis of arboviral diseases, particularly Dengue and Chikungunya. The system integrates machine learning models trained on epidemiological and clinical data in order to support decision making in the public health system.

VALERIA provides an AI-based diagnostic assistant capable of supporting clinical evaluation through probabilistic classification of arboviruses, contributing to faster and more reliable diagnosis within the Brazilian Unified Health System (SUS) [da Silva Neto et al. 2022]. The application was developed using a rigorous machine learning methodology that includes benchmarking of multiple algorithms, feature selection techniques, and hyperparameter optimization [da Silva Neto 2024].

[1] valeria.dotlabbrazil.com.br/

From a data perspective, the system relies on large-scale epidemiological datasets derived from the Brazilian disease surveillance system (SINAN). The original dataset contained more than 13 million records and 118 attributes. After preprocessing and filtering steps, the dataset still contained over 7.6 million records and 56 attributes, including more than 4.3 million Dengue cases and hundreds of thousands of Chikungunya cases [da Silva Neto et al. 2022].

Such a dataset size significantly increases the computational cost associated with model training and experimentation. The training pipeline involves repeated preprocessing operations, feature selection, model benchmarking, and hyperparameter optimization. These processes require scalable computational resources capable of processing millions of records efficiently. For this reason, cloud infrastructure based on AWS was adopted, using services such as Amazon S3 for data storage and Amazon EC2 for executing training workflows.

4.2. ANGELS: Intelligent Gestational Monitoring System

ANGELS[2] is an intelligent gestational monitoring system designed to support maternal and neonatal healthcare through predictive analytics and machine learning models. The system integrates multiple predictive models that analyze clinical and sociodemographic data collected during prenatal care in order to identify potential risks affecting the mother or the newborn.

ANGELS was designed as a modular system capable of integrating several predictive models related to different stages of pregnancy, childbirth, and postpartum monitoring [Rocha et al. 2026]. The system relies on data collected through the Mãe Coruja Pernambucana Program, a public health initiative that provides prenatal monitoring services across more than one hundred municipalities in the state of Pernambuco.

The datasets used in the development of the system present significant scale and complexity. In the congenital syphilis prediction case study, the unified dataset contained 256 attributes and 218,014 records collected from clinical, gestational, and sociodemographic databases [Rocha et al. 2026], with a significant amount of missing or poorly populated data.

Processing these datasets requires preprocessing steps such as table merging, feature engineering, missing-data handling, and attribute filtering prior to model training. Additionally, the platform supports multiple versions of predictive models for each clinical problem, increasing the computational demand for training and validation.

Given this scenario, AWS cloud resources provide the necessary infrastructure for scalable experimentation. Amazon S3 is used for storing datasets and experimental artifacts, while Amazon EC2 instances enable the execution of machine learning pipelines and hyperparameter optimization routines. Containerization using Amazon ECR also supports reproducibility and modular deployment of predictive services.

The system architecture is composed of three main components: Models API, ANGELS API, and Eureka Server. These components orchestrate predictive services, manage communication between modules, and expose functionalities to client applications. These components were deployed on Amazon EC2 instances to provide greater control

[2]angels.dotlabbrazil.com.br/

over the execution environment during development and experimentation. However, the architecture supports more scalable approaches, where some services can be implemented using AWS Lambda, enabling automatic scaling and reduced operational costs.

4.3. IAra: Malaria Forecasting Platform

IAra [3] is a scientific platform designed to support malaria surveillance and epidemic forecasting in Brazil's Legal Amazon region. The system integrates machine learning models with epidemiological data in order to predict future malaria incidence and identify potential outbreak scenarios.

IAra combines artificial intelligence models with epidemiological surveillance tools, enabling public health authorities to analyze disease patterns and forecast malaria cases through interactive dashboards and predictive analytics [de Carvalho Monteiro 2025]. The platform was developed using data from the SIVEP-Malaria system, which contains millions of records related to malaria notifications across the Amazon region.

According to the IAra platform documentation, the malaria dataset used in the system contains more than 6.6 million confirmed records collected between 2003 and 2022 [Monteiro et al. 2025]. In the raw data extraction process, the database originally contained more than 44 million records and dozens of attributes describing epidemiological, laboratory, and geographic information.

The scale of these datasets requires substantial computational resources for preprocessing, spatial clustering, time-series aggregation, and model training. IAra evaluates multiple forecasting models, including Random Forest, Support Vector Regression, ARIMA, and deep learning architectures such as LSTM networks [de Carvalho Monteiro 2025]. Additionally, the experiments involve spatial segmentation of municipalities and temporal aggregation of epidemiological notifications, which multiplies the number of experimental scenarios to be evaluated.

Under these conditions, AWS cloud services provide the computational scalability necessary for training and evaluating predictive models. Amazon S3 supports storage of historical epidemiological data, while Amazon EC2 instances enable the execution of large-scale time-series modeling experiments. Amazon SageMaker can also be used to structure reproducible machine learning pipelines and facilitate deployment of predictive models.

5. Discussion

The development of the three platforms highlights key aspects regarding the design of cloud-based architectures for scientific health systems.

A central finding is the need for scalable infrastructures to support machine learning workflows over large and heterogeneous health datasets. These systems demand substantial resources for preprocessing, feature engineering, and model training. Cloud environments provide the required elasticity, enabling efficient execution by dynamically adjusting computational capacity.

Reproducibility also emerges as a critical factor. Healthcare machine learning experiments typically involve multiple iterations with varying models and configurations.

[3] iara-dotlab.com.br

Cloud storage services, such as Amazon S3, allow centralized management of datasets and artifacts, facilitating version control and consistent experiment replication.

Containerization further ensures consistency across environments. By packaging pipelines and services into container images stored in Amazon ECR, it becomes possible to standardize execution environments and simplify deployment, especially in collaborative scenarios.

Another relevant observation relates to the operationalization of machine learning models. While many research works focus primarily on model development and evaluation, the transition from experimental models to operational decision-support systems requires additional architectural considerations. Serverless computing services such as AWS Lambda enable the deployment of predictive models as scalable inference services, simplifying system maintenance and reducing infrastructure management overhead.

Overall, the VALERIA, ANGELS, and IAra platforms demonstrate that cloud computing is fundamental not only for deployment but also for enabling large-scale scientific experimentation in healthcare.

However, it is important to acknowledge that adopting these technologies involves a non-trivial learning curve. Mastering cloud services, containerization, and distributed architectures requires time and specialized knowledge. Despite this initial barrier, the long-term benefits in scalability, reproducibility, and system robustness justify the investment.

6. Conclusion

This paper presented an experience report on the development of cloud-based architectures for scientific health systems using Amazon Web Services (AWS). The study analyzed three real-world platforms developed for public health applications: VALERIA, a clinical decision support system for arboviral diseases; ANGELS, an intelligent gestational monitoring system; and IAra, a malaria forecasting platform.

The results demonstrate that cloud infrastructures play a fundamental role in supporting the development of data-driven healthcare systems. The use of services such as Amazon S3, EC2, ECR, Lambda, and SageMaker enabled scalable storage, efficient data processing, reproducible experimentation, and the operational deployment of machine learning models.

The experience obtained from these systems highlights the importance of integrating machine learning pipelines with cloud-native architectures in order to support large-scale health data analysis. Cloud platforms provide the computational flexibility required to process large epidemiological datasets, execute complex training workflows, and deliver predictive services to healthcare professionals through operational systems.

Future work includes the integration of real-time health data streams from connected health devices, the incorporation of advanced deep learning models for predictive analytics, and the exploration of cloud-native architectures designed for large-scale epidemiological surveillance. Additionally, further research may investigate the integration of federated learning approaches and privacy-preserving mechanisms to support secure and collaborative health data analysis across institutions.

Statement on the Use of Artificial Intelligence

In accordance with the SBC Code of Conduct, we declare that a generative Artificial Intelligence (AI) tool, specifically ChatGPT (OpenAI), was used solely for text revision purposes, including improving clarity, grammar, and organization. The tool was not used for generating scientific content, data analysis, figures, tables, or methodological development. It is not considered an author of this work, and the authors take full responsibility for the accuracy, originality, and integrity of the content presented.

Referências

Barros, M. H. L. F. d. S., Alves, G. O., Souza, L. M. F., Rocha, E. d. S., Oliveira, J. F. L., Lynn, T., Sampaio, V., and Endo, P. T. (2021). Benchmarking machine learning models to assist in the prognosis of tuberculosis. *Informatics*, 8(2):27.

da Silva Neto, S. R. (2024). *Clinical decision support for arboviral diseases using machine learning models*. PhD thesis, Universidade de Pernambuco.

da Silva Neto, S. R. et al. (2022). Arboviral disease record data - dengue and chikungunya, brazil, 2013–2020. *Scientific Data*, 9:198.

de Carvalho Monteiro, K. H. (2025). *IAra: A Platform to Combat Malaria in Brazil's Legal Amazon*. PhD thesis, Universidade de Pernambuco.

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.

Monteiro, K., Rocha, E., Silva, E., Santos, G. L., Santos, W., and Endo, P. T. (2018). Developing an e-health system based on iot, fog and cloud computing. In *IEEE/ACM International Conference on Utility and Cloud Computing Companion*, pages 17–18. IEEE.

Monteiro, K. H. d. C. et al. (2025). Integrating machine learning and spatial clustering for malaria case prediction in brazil's legal amazon. *BMC Infectious Diseases*, 25:802.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380:1347–1358.

Rocha, É. D. S., De Morais, C. M., Teixeira, I. V., Monteiro, K. H. D. C., Neto, S. R. D. S., Soares, H. R., Saldanha, R., Neto, W. B., and Endo, P. T. (2026). Angels: An intelligent gestational follow-up system. *IEEE Access*.

World Health Organization (2020). Global strategy on digital health 2020–2025. *World Health Organization Report*.