UNISINOS

Graduate Program in

Applied Computing

Academic Doctorate

Felipe André Zeiser

MultSurv: A Multimodal Deep Learning Model for Hospitalized Patients Survival Analysis in the Context of a Pandemic

São Leopoldo, 2024

Felipe André Zeiser

MULTSURV: A MULTIMODAL DEEP LEARNING MODEL FOR HOSPITALIZED PATIENTS SURVIVAL ANALYSIS IN THE CONTEXT OF A PANDEMIC

Dissertation presented as a partial requirement to obtain the Doctor's degree from the Applied Computing Graduate Program of the Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Cristiano André da Costa, PhD

Co-advisor:
Prof. Gabriel de Oliveira Ramos, PhD

São Leopoldo
2024

Z47m Zeiser, Felipe André.
Multsurv : a multimodal deep learning model for hospitalized patients survival analysis in the contexto of a pandemic / Felipe André Zeiser. – 2024.
132 f. : il. ; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2024.
"Orientador: Prof. Dr. Cristiano André da Costa
Coorientador: Prof. Dr. Gabriel de Oliveira Ramos"

1. Artificial Intelligence. 2. Deep Learning. 3. Multimodal Data. 4. Pandemics. 5. Survival Analusis. I. Título.

CDU 004.4

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Silvana Dornelles Studzinski – CRB 10/2524)

To my family.

"The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a simple datum of experience."

— Albert Einstein

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisors, Professor Cristiano André da Costa and Professor Gabriel de Oliveira Ramos, for their continuous support, guidance, and inspiration throughout this journey. I would also like to thank Prof. Andreas Maier for hosting me for an internship at the Friedrich-Alexander University Erlangen-Nürnberg (FAU) in Erlangen.

I would like to thank the following hospitals: Grupo Hospitalar Conceição (GHC), Hospital de Clínicas de Porto Alegre (HCPA), Hospital Ernesto Dornelles (HED), Hospital Moinhos de Vento (HMV), Hospital Universitário de Santa Maria (HUSM), Santa Casa de Misericórdia de Porto Alegre (ISCMPA), and Unimed Central, which were fundamental in the MyDigitalHealth project. Without the participation of these institutions, this research would not have been possible.

I would also like to thank my parents, Darci and Melita, for their unconditional support. You taught me the value of effort, dedication, and persistence from an early age. I would especially like to thank my brother Mateus for his support and critical eye as a mathematician. I also extend my gratitude to my girlfriend, now fiancée, life and travel partner, Mônica, whose patience, encouragement, and love were essential at every stage of this journey. Thank you for believing in me, celebrating each achievement, and supporting me in difficult times.

I would also like to express my gratitude to my partners Ismael, Henrique, and Paulo, who, as research collaborators, contributed invaluably to the development of this work. You were essential as friends and advisors, and I am grateful for your collaboration and commitment to each stage of this process.

I would also like to mention my project colleagues Blanda Mello, Fausto Vanin, Ana Alegretti, Ana Bertoni, and Bruna Donida — for their friendship, team spirit, and mutual support throughout the project.

Finally, I would like to thank the Coordinating Agency for Advanced Training of Graduate Personnel (CAPES) (C.F. 001) and the National Council for Scientific and Technological Development (CNPq) (No. 309537/2020-7) for supporting this work.

ABSTRACT

BACKGROUND: Respiratory infectious diseases represent a major challenge in modern society. We recently faced the most significant public health challenge of the last century. Severe Acute Respiratory Syndrome Coronavirus 2 has overwhelmed almost all health systems worldwide, highlighting pre-existing weaknesses. The heterogeneity of COVID-19 clinical manifestations has made it challenging to manage hospitalized patients, making it crucial to identify those at greatest risk, especially for efficiently allocating vital resources. Unlike past pandemics, hospitalized patients are currently monitored continuously and through different modalities. These data generate large longitudinal and multimodal datasets in health institutions. In this context, data-driven solutions can support clinical decisions and provide new tools for risk management of hospitalized patients during pandemics. **OBJECTIVE:** Therefore, we propose integrating clinical, laboratory, and chest X-ray imaging features into a survival analysis model for hospitalized patients with COVID-19. With the model, we aim to combine multimodal and longitudinal data to capture the dynamic nature of COVID-19 and provide an explainable hazard function. **METHODOLOGY:** The methodology involves the proposition and development of the model. The model is divided into five main components: (i) pre-processing; (ii) feature encoders; (iii) temporal attention; (iv) CheXReport; and (v) multitask networks. The pre-processing component is responsible for data cleaning, outlier removal, variable selection, and image processing. In the feature encoders, the categorical and continuous data are transformed into a vector of embeddings that capture the complex and non-linear relationships between the variables. Then, based on the embeddings up to the current time instant, we extract a temporal context vector using temporal attention. The CheXReport component processes the patient's X-ray images using a fully-transformers architecture, which integrates visual features with the textual elements of the reports. Finally, all feature vectors are concatenated to be processed in the multitask networks, a set of neural networks that allow the model to capture the specific characteristics of each risk. **RESULTS:** To evaluate the model performance, we used an incremental ablation study. We use the public datasets PBC2, MIMIC-CXR, Curated Dataset for COVID-19, and a private dataset. Then, we compare the results of the MultSurv model with the state of the art. The results obtained demonstrate that the MultSurv outperforms all reference architectures, with a C-index of 0.723 ± 0.008 for $t = 1$ and $\Delta t = 1$, and 0.695 ± 0.003 for $t = 7$ and $\Delta t = 7$. **CONCLUSION:** The main scientific contribution of this study is the proposal of a multimodal model for processing dynamic and longitudinal data in survival analysis in the context of COVID-19. Furthermore, the MultSurv model offers a tool to support patient prioritization in pandemic scenarios. Finally, the application of the model can be adapted to different clinical contexts, extending beyond COVID-19.

Keywords: Survival Analysis. Pandemics. Multimodal Data. Deep Learning. Artificial Intelligence.

RESUMO

CONTEXTO: As doenças infecciosas respiratórias representam um grande desafio na sociedade moderna. Recentemente, enfrentamos o maior desafio de saúde pública do último século. A *Severe Acute Respiratory Syndrome Coronavirus 2* provocou uma sobrecarga em quase todos os sistemas de saúde do mundo, evidenciando as fragilidades preexistentes. A heterogeneidade das manifestações clínicas da COVID-19 dificultou o manejo dos pacientes hospitalizados, tornando crucial a identificação daqueles em maior risco, especialmente para a alocação de recursos vitais. Diferentemente de pandemias passadas, os pacientes hospitalizados atualmente são monitorados de forma contínua e por meio de diferentes modalidades. Esses dados geram grandes conjuntos de dados longitudinais e multimodais nas instituições de saúde. Nesse contexto, soluções baseadas em dados podem apoiar as decisões clínicas e fornecer novas ferramentas para a gestão de riscos dos pacientes hospitalizados durante pandemias. **OBJETIVO:** Deste modo, propomos integrar características clínicas, laboratoriais e de imagens de Raio-X do tórax em um modelo de análise de sobrevivência para pacientes hospitalizados com COVID-19. Com o modelo, buscamos combinar dados multimodais e longitudinais para capturar a natureza dinâmica da COVID-19 e fornecer uma função de risco explicável. **METODOLOGIA:** A metodologia envolve a proposição e desenvolvimento do modelo MultSurv. O modelo é dividido em cinco componentes principais: (i) *pre-processing*; (ii) *feature encoders*; (iii) *temporal attention*; (iv) *CheXReport*; e (v) *multitask networks*. O *pre-processing* é responsável pela limpeza de dados, remoção de outliers, seleção de variáveis e processamento de imagens. Nos *feature encoders*, os dados categóricos e contínuos são transformados em um vetor de *embeddings* que captura as relações complexas e não-lineares entre as variáveis. Em seguida, com base nos *embeddings* até o instante de tempo atual, extraímos um vetor de contexto temporal utilizando a *temporal attention*. O *CheXReport* processa as imagens de Raio-X do paciente utilizando uma arquitetura *fully-transformers*, que integra características visuais com os elementos textuais dos laudos. Finalmente, todos os vetores de características são concatenados para serem processados nas *multitask networks*, um conjunto de redes neurais multitarefas que permite ao modelo capturar as características específicas de cada risco. **RESULTADOS:** Para avaliar o desempenho do modelo MultSurv, realizamos um estudo de ablação incremental. Utilizamos os conjuntos de dados públicos PBC2, MIMIC-CXR, *Curated Dataset for COVID-19* e um conjunto de dados privado. Em seguida, comparamos os resultados do modelo MultSurv com o estado da arte. Os resultados obtidos demonstram que o modelo MultSurv superou todas as arquiteturas de referência, com um C-index de 0.723 ± 0.008 para $t = 1$ e $\Delta t = 1$, e 0.695 ± 0.003 para $t = 7$ e $\Delta t = 7$. **CONCLUSÃO:** A principal contribuição científica deste estudo é a proposta de um modelo multimodal para o processamento de dados dinâmicos e longitudinais na análise de sobrevivência no contexto da COVID-19. Além disso, o modelo MultSurv oferece uma ferramenta de apoio à priorização de pacientes em cenários de pandemia. Por fim, a aplicação do modelo MultSurv pode ser adaptada para diferentes contextos clínicos, estendendo-se além da COVID-19.

Palavras-chave: Análise de Sobrevivência. Pandemia. Multimodal data. Aprendizado Profundo. Inteligência Artificial.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI	Artificial Intelligence
AM	Amazonas
ANN	Artificial Neural Network
AUC	Area under the ROC Curve
BLEU	Bilingual Evaluation Understudy
C-index	Concordance index
CIS	Clinical Information Systems
CKDEPI	Chronic Kidney Disease Epidemiology Collaboration
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CoxPH	Cox Proportional Hazards
DL	Deep Learning
EHR	Electronic Health Record
FN	False Negative
FP	False Positive
GLEU	Google BLEU
GHC	Grupo Hospitalar Conceição
GRU	Gated Recurrent Unit
HCPA	Hospital de Clínicas de Porto Alegre
HED	Hospital Ernesto Dornelles
HMV	Hospital Moinhos de Vento
HUSM	Hospital Universitário de Santa Maria
ICU	Intensive Care Unit
IQR	Interquartile Range
ISCMPA	Santa Casa de Misericórdia de Porto Alegre
LIS	Laboratory Information System
LSTM	Long Short-Term Memory
MDH	MyDigitalHealth
MDRD	Modification of Diet in Renal Disease
MERS	Middle East Respiratory Syndrome
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MIMIC-CXR	MIMIC Chest X-ray

ML	Machine Learning
MLP	Multilayer Perceptron
MQ	Main Question
MSA	Multi-head Self-attention
NLP	Natural Language Processing
PACS	Picture Archiving and Communication System
PBC2	Primary Biliary Cirrhosis
PCA	Principal Component Analysis
ReLU	Rectified Linear Units
RJ	Rio de Janeiro
RLR	Rapid Literature Review
RNN	Recurrent Neural Network
RO	Rondônia
ROC	Receiver Operating Characteristic
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RS	Rio Grande do Sul
RSNA	Radiological Society of North America
RT-qPCR	Quantitative Reverse Transcription Polymerase Chain Reaction
SARI	Severe Acute Respiratory Infections
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SC	Santa Catarina
SGD	Stochastic Gradient Descent
SHAP	Shapley
SLR	Systematic Literature Review
SP	São Paulo
SQ	Specific Questions
SUS	Sistema Único de Saúde
SW-MSA	Shifted Window Multi-head Self-attention
Tanh	Hyperbolic Tangent
TN	True Negatives
TP	True Positives
UNISINOS	Universidade do Vale do Rio dos Sinos
ViT	Vision Transformer
W-MSA	Window Multi-head Self-attention

CONTENTS

1 INTRODUCTION

We recently faced the most significant healthcare challenge of the last century (WENHAM et al., 2021; JAMES; MENZIES; RADCHENKO, 2021). In December 2019, a cluster of pneumonia cases of unknown cause was epidemiologically associated with a sea market in Wuhan province, China (HUANG et al., 2020). The causative agent, a new betacoronavirus, has its probable origin in the spillover between species from animals to humans (WU et al., 2021). The Coronavirus Disease 2019 (COVID-19), caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), presents flu-like symptoms that can become severe in high-risk individuals (LI et al., 2020).

Over the past two decades, several infectious agents have emerged, including COVID-19, Severe Acute Respiratory Syndrome-associated Coronavirus (SARS-CoV), Middle East Respiratory Syndrome (MERS), and Mpox (PANEL et al., 2023). Some studies suggest that due to globalization, climate change, and greater interactions with animals, the probability of transmission of pathogens between species has increased considerably (BENGIS et al., 2004; OSTERHOLM, 2005; RAHMAN et al., 2020). Infectious diseases, especially zoonotic ones, promise to be one of the biggest public health challenges in the coming decades (BEDFORD et al., 2019).

Despite the challenges posed by these infectious agents, the COVID-19 pandemic has highlighted a series of weaknesses in all types of systems – be they healthcare, food, economy, or governance (LEACH et al., 2021). Furthermore, problems related to disinformation, governance strategies, and lack of coordination between countries in the world worsened the scenario of the COVID-19 pandemic (BURSZTYN et al., 2020; ANSELL; SØRENSEN; TORFING, 2020; LITEWKA; HEITMAN, 2020; ZEISER et al., 2022). In addition, the pandemic has further highlighted the economic discrepancy between countries worldwide. First, developing countries have suffered from a lack of resources for effective SARS-CoV-2 screening and diagnosis. The lack of tracking resulted in a greater need for access to healthcare providers, which, even before the COVID-19 pandemic, showed signs of inability to meet the demands in developing countries. Furthermore, after the emergence of the first vaccines, low-income countries again had late access and insufficient quantities to promote an efficient immunization campaign (ACHARYA; GHIMIRE; SUBRAMANYA, 2021).

Digitalizing healthcare services, especially throughout the 21st century, has allowed the recording of detailed information in Electronic Health Records (EHRs) of patients infected by SARS-CoV-2 (WANG et al., 2021). A patient collects clinical data, vital signs, and laboratory and imaging tests throughout hospitalization to characterize the disease and current health status. In the context of COVID-19, these records made it easier for the scientific community to share information to analyze,

understand, diagnose, treat, and prevent the disease throughout the pandemic more efficiently (GRONVALL, 2020).

However, data related to COVID-19 are heterogeneous and complex, reflecting variability in disease progression and clinical outcomes among different patients (HO et al., 2024). Furthermore, data collection from hospitalized patients generally does not follow a well-defined periodicity, which generates a considerable amount of sparse data (WIEGREBE et al., 2024). This sparseness, combined with the longitudinal nature of data, where information is recorded over time, and the competition of risks, in which multiple health events can occur simultaneously or in rapid succession, presents significant data analysis and interpretation challenges (LEE; YOON; SCHAAR, 2019).

To efficiently extract non-linear relationships from these heterogeneous data, building Artificial Intelligence (AI) models is necessary. However, current literature still has difficulties dealing with the data's temporal nature, relying mainly on the last measurement available for prediction (WIEGREBE et al., 2024). To address these limitations, it is essential to develop techniques that can integrate temporal data capturing the evolution of the disease over time. Additionally, utilizing multimodal data, which includes medical images, vital signs, text records, and other types of data, can provide a more complete picture of a patient's health status (ZHONG et al., 2024).

In this sense, the main focus of this dissertation is to develop a survival analysis model for hospitalized patients with COVID-19 to assist healthcare professionals and provide patients with personalized care. In this sense, we propose a survival analysis model with learning capacity on longitudinal and sparse data collected from EHRs. These data are multimodal, with sociodemographic, clinical, laboratory, and imaging information about the patients. The model can capture the temporal relationships of data through a temporal attention network based on a Recurrent Neural Network (RNN). Information from patients' imaging exams is extracted by a fully Transformer architecture, which can extract information from chest X-ray exams and suggest findings in natural language. Finally, we provide healthcare professionals with explainability information to aid decision-making.

1.1 Motivation

The COVID-19 pandemic has highlighted existing weaknesses in healthcare systems around the world. Under normal conditions, these systems already operate close to the limit of their capacity. Pandemics, such as COVID-19, where the rapid spread of the virus and variability in disease progression among patients exponentially increase the demand for medical care, further expose these vulnerabilities, resulting in an extreme strain on healthcare resources and teams. The

shortage of hospital beds, personal protective equipment, mechanical ventilators, and the physical and mental exhaustion of healthcare professionals are just some of the consequences of this scenario. Furthermore, the need for rapid and accurate diagnoses becomes crucial for effectively managing the disease, highlighting the importance of clinical decision-support tools that can assist in screening and treating patients efficiently and reliably.

In this context, the need for effective and accurate systems to predict patient outcomes becomes a valuable decision-support tool to help define management practices. Traditional survival analysis models, while helpful, often fall short when confronted with the complex and multimodal data generated by modern healthcare systems. Unlike previous pandemics, such as SARS-CoV in 2002 and MERS in 2012, we currently have vast volumes of data stored in EHRs, which include sociodemographic information, clinical records, laboratory tests, and imaging exams. Each data type provides insight into a patient's condition, but most existing models are limited to analyzing a single modality or simplistic combinations of modalities. This limitation prevents the model and, consequently, the outcomes from capturing and representing, respectively, a complete view of the patient's health status.

Furthermore, patients in critical care settings are continuously monitored, with data being collected irregularly over time. For example, not all patients have equivalent imaging data or laboratory records. This disparity contributes to the heterogeneous and longitudinal nature of clinical data. Traditional survival models, such as Cox Proportional Hazards (CoxPH) models, often assume that data are collected at regular intervals or rely on static snapshots of patient data. This assumption does not align with the reality of clinical practice, where the timing of data collection can vary significantly, and missing data are expected. Failure to account for these temporal dynamics can result in models that do not fully capture the progression of a patient's condition, leading to less accurate predictions. Additionally, patients are also subject to multiple health risks throughout their hospitalization.

The current literature has advanced in capturing complex and non-linear relationships by combining the CoxPH model with neural networks (KATZMAN et al., 2018). Furthermore, adopting multitask learning strategies has shown promise in efficiently modeling and capturing the distinct characteristics of the various risks to which a patient may be exposed (FOTSO, 2018). For example, Dynamic-DeepHit leverages RNNs and neural networks to capture temporal dependencies and competing risks in longitudinal data (LEE; YOON; SCHAAR, 2019). However, despite these advances, existing models face significant limitations, particularly their ability to incorporate and process multimodal data effectively.

Another gap identified in our review concerns healthcare data's dynamic and

evolving nature. To accurately assess patient risk, models must capture and analyze temporal and historical relationships that reflect the progression of a patient's health status. This is particularly important in chronic and complex diseases, where a patient's condition can change rapidly. Furthermore, in high-risk clinical settings, implementing explainability mechanisms is crucial. Providing clear and interpretable insights into how models arrive at their predictions can increase the confidence and adoption of these technologies in clinical practice. Without such transparency, healthcare professionals may hesitate to rely on these models for decision-making, limiting their practical utility and impact.

1.2 Research Question

The integration of AI into healthcare, particularly for survival analysis, has shown considerable potential in improving patient outcomes by enabling more accurate predictions of patient risk and prognosis. However, as highlighted in the previous sections, the complexity of healthcare data presents significant challenges that must be addressed to realize this potential fully. While increasingly sophisticated, current models often fall short in handling clinical data's multimodal and longitudinal nature. In light of these gaps, our work aims to answer the following research question:

How to develop a deep learning architecture that leverages dynamic multimodal data to enhance survival analysis predictions while ensuring the explainability of the model's outputs?

The proposed methodology integrates multimodal and longitudinal data to address this issue, offering explainability mechanisms. We combine clinical, laboratory, sociodemographic data, and X-ray images to view the patient's health status comprehensively. This is accomplished using a Deep Learning (DL) architecture that uses embeddings to represent continuous and categorical variables and a fully Transformer model for image analysis. To capture the temporal evolution of patient data, we use an RNN-based temporal attention network. Additionally, we incorporate techniques that enable the interpretation of model predictions, providing insights into the variables that most influence patient survival analysis. Finally, the model architecture is designed to simultaneously deal with multiple health risks, sharing information between networks specialized in different health events. This allows the model to capture complex relationships and improve the accuracy of survival analysis. Together, these aspects aim to overcome the limitations of traditional survival analysis models by providing a robust and interpretable tool to

assist in patient management during pandemics and other critical public health situations.

1.3 Objectives

The general objective of this dissertation is to develop and evaluate a survival analysis model that integrates longitudinal and multimodal data to enhance the diagnosis and prognosis of COVID-19 patients. With the model, we aim to provide healthcare professionals with an explainable, accurate tool for prioritizing patient care and optimizing resource allocation in hospital settings. By incorporating clinical, laboratory, and imaging data, the model seeks to improve the understanding of the influences of different variables on patient survival and support the implementation of personalized treatment strategies. To achieve this objective, we have defined some specific objectives:

- Perform a Rapid Literature Review (RLR) of the state of the art in DL models for multimodal and longitudinal survival analysis;
- Propose a DL model that supports the integration of dynamic multimodal and longitudinal data to fill the identified gaps in the literature;
- Define a mechanism to capture the characteristics of patients' health status evolution over time;
- Evaluate the applicability of embedding techniques to capture complex interactions and relationships between categorical and continuous variables in tabular data;
- Explore explainability techniques to improve the transparent and interpretable predictions of the DL model.

1.4 Scientific Contributions

The main scientific contribution of this dissertation is the proposal of a survival analysis model that can process multimodal and longitudinal data from patients hospitalized with COVID-19. Tabular data is transformed into dense vector representations that can capture complex non-linear relationships. The model processes historical data using an RNN to capture information from different temporal moments and generate a temporal contextual vector. Then, the model extracts features from the X-ray images. A set of multitasking networks processes all these contextual vectors to model multiple health risks simultaneously. The secondary scientific contributions of this dissertation are listed below:

- **State of the art survey.** We selected articles published over the last five years using the RLR protocol (HARKER; KLEIJNEN, 2012; KHANGURA et al., 2012). We analyze aspects of the methods in the current literature, divided into four main topics: current techniques, how the different data fusion techniques impact accuracy and reliability, how the literature handles longitudinal data, and the challenges that involve the analysis of multimodal data with competitive risks.

- **CheXReport architecture.** We propose CheXReport architecture, a novel fully transformer architecture pre-trained for generating radiological reports with Swin Transformer blocks. CheXReport's design integrates local and global image features more effectively. It employs Swin Transformer blocks that dynamically adjust receptive fields, enhancing the model's ability to extract and correlate detailed visual features with report text.

- **Embeddings for categorical and continuous variables.** Our model introduces embedding techniques for categorical and continuous variables, transforming them into dense vector representations. This allows the capture of complex, non-linear relationships within the data, improving the model's ability to integrate and interpret diverse types of patient information for more accurate survival analysis.

- **Multitask learning for competing risks.** The model employs a multitask learning approach to handle multiple competing health risks simultaneously. This enables the identification and differentiation of risk factors, providing a comprehensive assessment of patient prognosis and facilitating treatment strategies.

1.5 Project Context

This dissertation is linked to a larger project called MyDigitalHealth. MDH is a consortium between Universidade do Vale do Rio dos Sinos (UNISINOS) and seven institutions involved in the fight against COVID-19 in Rio Grande do Sul (RS): Grupo Hospitalar Conceição (GHC), Hospital de Clínicas de Porto Alegre (HCPA), Hospital Ernesto Dornelles (HED), Hospital Moinhos de Vento (HMV), Hospital Universitário de Santa Maria (HUSM), Santa Casa de Misericórdia de Porto Alegre (ISCMPA), and Unimed Central. The main objective of MDH is to develop an intelligent information and communication model based on a Blockchain architecture with standardized clinical data to link diverse health service providers and the patients of pandemics, particularly COVID-19. Therefore, this dissertation proposes a multimodal survival analysis model to provide healthcare professionals with a tool for aiding patient care and optimizing resource allocation in hospital settings.

Furthermore, this study was conducted following the Declaration of Helsinki. This research was approved by the Research Ethics Committee (CAAE under number 33540520.6.3004.5327) to develop the project entitled "Modelo Inteligente de Blockchain para Informações de Saúde e Interação com Pacientes no âmbito da COVID-19", called MDH. The project was submitted to Platform Brazil and approved by the GHC[1], ISMCPA[2], HED[3], HCPA[4], HMV[5], HUSM[6] hospital partners and UNISINOS.

1.6 Text Organization

The remainder of the study is divided into seven chapters. Chapter 2 presents the concepts related to the present work, introducing technologies and algorithms used to develop the proposed model. Then, in Chapter 3, the related works are discussed to present the state-of-the-art in the context of survival analysis with multimodal data. Chapter 4 presents the proposed model and design decisions. Chapter 5 presents the validation methodology and metrics for the model. Chapter 6 presents the results, evaluation, and discussions for the proposed model. Finally, Chapter 7 presents the final considerations regarding the findings and future work.

[1]Grupo Hospitalar Conceição <https://www.ghc.com.br/>
[2]Santa Casa de Misericórdia <https://santacasa.org.br/>
[3]Hospital Ernesto Dornelles <https://www.hed.com.br/>
[4]Hospital de Clínicas <https://www.hcpa.edu.br/>
[5]Hospital Moinhos de Vento <https://www.hospitalmoinhos.org.br/>
[6]Hospital Universitário de Santa Maria <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/husm-ufsm>

2 BACKGROUND

The rapid evolution of human society and the interconnectedness of global populations have significantly altered the infectious disease landscape. The terms endemic, epidemic, and pandemic are fundamental to understanding the scope and impact of disease outbreaks. Endemic diseases are localized with predictable transmission rates, while epidemics represent a sudden increase in cases over a wider area. Pandemics, the most worrying, spread globally and cause widespread health, social and economic disruptions. This chapter investigates these definitions and the factors contributing to the emergence and spread of infectious diseases in our modern world.

Additionally, this chapter explores the role of AI and DL in managing and mitigating the impacts of pandemics. AI, first conceptualized in 1956, encompasses computational techniques designed to replicate human cognitive functions. Within AI, Machine Learning (ML) focuses on developing algorithms that allow systems to learn and make data-based decisions. Among them, DL, a subset of ML, employs multi-layer artificial neural networks (often called deep neural networks) to model complex patterns in high-dimensional data. These networks have demonstrated competitive performace in several tasks, including computer vision and Natural Language Processing (NLP), making them contemporary tools for analyzing clinical and radiological data in COVID-19. This section investigates the technical foundations of DL, which are essential for developing diagnostic and prognostic models.

2.1 Endemic, Epidemic, and Pandemic

The definition of endemic, epidemic, and pandemic is directly related to the disease's transmission capacity and geographic scope. An endemic is a local disease transmission with a predictable transmission rate. An epidemic is characterized by an unpredictable increase in sick people over a larger geographic area. Finally, a pandemic causes disease cases globally (GRENNAN, 2019; PIRET; BOIVIN, 2021). In this sense, human evolution from gatherers to hunters, and now with extensive global trade and greater interactions with ecological systems, favor the emergence of new diseases (GRAHAM; SULLIVAN, 2018; BEDFORD et al., 2019; PIRET; BOIVIN, 2021). Finally, while the spread of bacterial or fungal pathogens is a global health concern, its pandemic appears less likely (GRAHAM; SULLIVAN, 2018).

Climate change favors the expansion of zoonotic vectors, such as Aedes Aegypti mosquitoes (CAMINADE; MCINTYRE; JONES, 2019). Furthermore, the constant evolution of viruses such as influenza virus is a critical challenge these days (PIRET;

BOIVIN, 2021). Viruses from zoonotic sources cause most human infectious diseases and pandemics (HUGHES et al., 2010; GRAHAM; SULLIVAN, 2018). The animal pathogen evolution to a human-specialized pathogen requires a combination of multiple variables that are not yet fully understood (MEGANCK; BARIC, 2021). In this sense, developing and identifying therapies and vaccines that can treat unknown zoonotic viruses are necessary for the future to reduce problems for health and economic systems (GRAHAM; SULLIVAN, 2018; SHANG; LI; ZHANG, 2021; MEGANCK; BARIC, 2021).

Proposals for the zoonotic pathogens detection that may migrate to humans can be divided into four categories: (i) surveillance and discovery of pathogens; (ii) development of new reagents for serological tests; (iii) vaccines development for viruses with migration potential; and (iv) vaccines fabrication and evaluation before pathogen migration (GRAHAM; SULLIVAN, 2018). However, given the modern research organization and the lack of an immediate market for a vaccine candidate, it is an obstacle to advancing pandemic preparedness measures (MEGANCK; BARIC, 2021). For example, problems faced with the COVID-19 pandemic have already been predicted in past studies, such as the health systems organization, priorities in access to hospital resources, and how to increase antivirals, masks, and antibiotics productions (OSTERHOLM, 2017). However, during the COVID-19 pandemic, many platforms, protocols, and policies were based on foundations used during pandemics with SARS-CoV, MERS, and avian flu (GRAHAM; SULLIVAN, 2018). In this sense, in the next section, we present the main concepts regarding the behavior and diagnosis of COVID-19.

2.2 SARS-CoV-2

The fear of other viruses emergence that cause infectious diseases with pandemic potential has existed for decades (BENGIS et al., 2004; OSTERHOLM, 2005). Even viruses similar to SARS-CoV-2 were found in bats in 2013 and could infect humans without previous adaptation (GE et al., 2013). However, the SARS-CoV outbreak demonstrated that despite the concern, the world was unprepared for the COVID-19 pandemic (PIRET; BOIVIN, 2021). SARS-CoV-2 was first officially identified in hospitals in Wuhan, Hubei Province, China, using surveillance mechanisms established after the SARS-CoV outbreak in 2003 (ZHU et al., 2020; LI et al., 2020). The outbreak likely originates from a wholesale seafood market in Wuhan (ZHU et al., 2020).

Despite a lower reported mortality rate than SARS-CoV and MERS-CoV, SARS-CoV-2 is responsible for more deaths and cases (ZEISER et al., 2022). The problems with the rapid and deadly spread of SARS-CoV-2 worldwide are related to

the transmission rate and the possibility of transmitting the virus even in asymptomatic individuals (EZHILAN; SURESH; NESAKUMAR, 2021). In SARS-CoV, for example, the peak viral load is reached between 6 and 11 days after the onset of symptoms, facilitating patient identification and isolation (PIRET; BOIVIN, 2021). In MERS-CoV, at least what is known so far, transmission occurs only from animals to humans (EZHILAN; SURESH; NESAKUMAR, 2021).

2.2.1 Symptoms

Despite high rates of asymptomatic cases, COVID-19 can lead to the development of a broad spectrum of diseases, from mild symptoms to life-threatening illnesses (WIERSINGA et al., 2020; ORAN; TOPOL, 2021). Some of the most common symptoms are cough, myalgias, and headache. In addition, loss of taste or smell are symptoms commonly associated with SARS-CoV-2 infection (BRANDAL et al., 2021). However, no specific symptoms or signs can reliably diagnose COVID-19 (STRUYF et al., 2021).

The evolution from mild to severe cases can happen quickly, sometimes in less than a week (COHEN et al., 2020). The most frequent complications involve the development of Severe Acute Respiratory Infections (SARI) (WANG et al., 2020). SARI causes respiratory failure in critically ill patients and is defined by the acute onset of noncardiogenic pulmonary edema, hypoxemia, and the need for mechanical ventilation (MATTHAY et al., 2019). In addition, COVID-19 can cause cardiac, thromboembolic, neurological complications and long-term sequelae (COHEN et al., 2020; WANG et al., 2020; LEISMAN et al., 2020).

2.2.2 Diagnosis

COVID-19 can present a broad spectrum of symptoms, making a differential diagnosis based on clinical features complex due to similarity to other diseases (DONIDA; COSTA; SCHERER, 2021; ZEISER et al., 2022). The gold standard for diagnosis is Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) (PATEL; JERNIGAN et al., 2020). However, throughout the COVID-19 pandemic, at different times, the testing capacity of the health system proved to be limited (BURKI, 2020; ZEISER et al., 2022). This limitation is mainly located in regions far from large urban centers and socially more vulnerable countries (MEHTAR et al., 2020).

One of the most critical findings of COVID-19 is pneumonia (XU et al., 2021). In this sense, radiology exams play a fundamental role in diagnosing and monitoring patients with COVID-19. The main chest X-ray findings are consolidation and

ground-glass opacities, with bilateral, peripheral, and lower lung zone distributions (WONG et al., 2020). Meanwhile, chest computed tomography scans in COVID-19 patients most commonly demonstrate ground-glass opacification with or without consolidated abnormalities, consistent with viral pneumonia (SHI et al., 2020). However, X-ray or computed tomography findings cannot define with complete certainty the COVID-19 infection (ACR; RADIOLOGY et al., 2020). Therefore, the Radiological Society of North America (RSNA) recommends classifying the radiological findings as typical, indeterminate, or atypical for COVID-19 (SIMPSON et al., 2020). Despite this, radiological findings can be identified even before the onset of symptoms and the ability to identify the SARS-CoV-2 in samples from the upper respiratory tract (SHI et al., 2020). Finally, radiological examinations are a key point in the screening and treatment definition of patients (PANWAR et al., 2020).

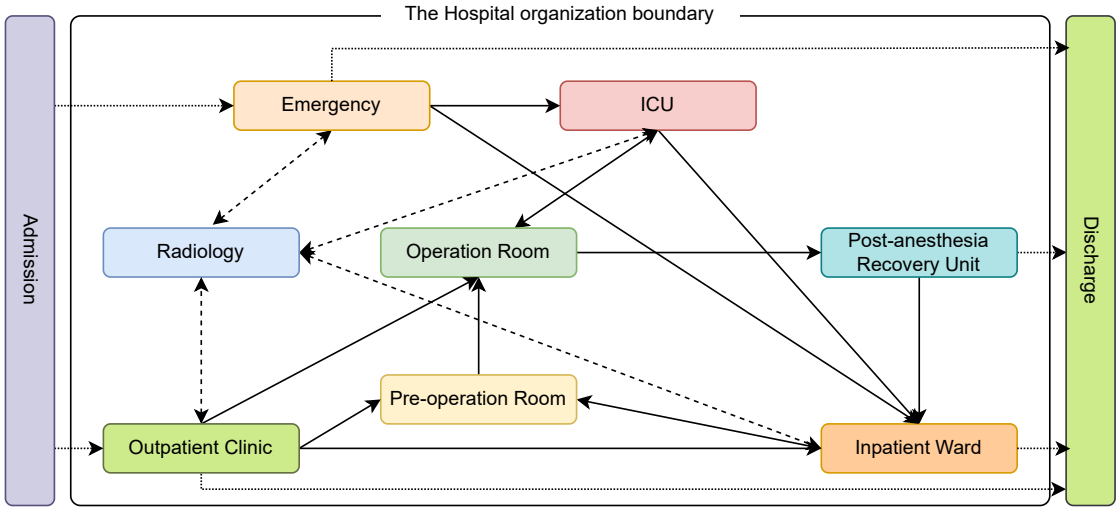
2.2.3 Patient Flow in Hospitals During the Pandemic

The dynamics of the COVID-19 pandemic have overwhelmed most healthcare systems globally. High transmission and hospitalization rates led health institutions to adjust their service, management, and healthcare flows to meet demand (MCCABE et al., 2020). However, this adaptation was challenging, especially in remote and socially vulnerable regions, highlighting the lack of qualified professionals, medical supplies, and supplementary oxygen (ZEISER et al., 2022). Furthermore, several health institutions were not used to caring for patients with high-level isolation needs, which led to adaptations with available resources (PANDEY et al., 2020). In addition to this, chronic patients continued to place demands on healthcare institutions at the same time as they were one of the main risk groups (ZEISER et al., 2022). One of the measures adopted by many countries to reduce hospitalization needs was the cancellation of elective surgeries and the adoption of telehealth measures (MCCABE et al., 2020).

Differences in healthcare systems and the clinical behavior of variants have impacted in-hospital mortality rates (ZEISER et al., 2022). Despite the variations in the scale of healthcare institutions, the flow of care for a patient is similar in all hospitals and can be seen in Figure 1. With the COVID-19 pandemic, each of these sectors needed changes to guarantee care, diagnosis, and treatment for patients and, simultaneously, protect employees of healthcare institutions (PANDEY et al., 2020). These decisions involved the adoption of digital flows to reduce contact, infection control for cleaning facilities, clinical management, such as the moment of intubation and extubation, admission to the Intensive Care Unit (ICU), daily communication with family members and the team, and expansion of bed capacities (GRIFFIN et al.,

2020). Despite a continuous effort to define patient management and treatment guidelines, these were constantly changed throughout the pandemic due to the etiology of the disease and healthcare resources available (ZEISER et al., 2022).

Figure 1 – Flow diagram representing the journey of a patient within a hospital, from admission to discharge. Arrows indicate possible paths between units.



Source: Adapted from Åhlin, Almström and Wänström (2022).

Due to the healthcare system's capabilities, patients admitted to hospitals during the COVID-19 pandemic were mainly severe cases (MACEDO; GONCALVES; FEBRA, 2021). These patients had access to different care and treatment structures due to the unknown nature of the disease. Therefore, the prognosis of patients varies considerably between regions and periods. Finally, management errors like those in the North of Brazil may have worsened hospital mortality rates (ZEISER et al., 2022).

2.3 Artificial Intelligence

AI was first introduced in 1956 by John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude Shannon at the Dartmouth Summer Research Project on Artificial Intelligence (DICK, 2019). Historically, researchers have defined the AI concept in several ways, mainly linked to human performance fidelity and the definition linked to rationality (RUSSELL; NORVIG, 2020). The difficulty in determining the AI concept is related to human intelligence understanding, which, despite advances in recent decades, we are still far from fully understanding the neurobiological mechanisms (BARBEY, 2018; CHOLLET, 2021). In this work, we adopt the definition of intelligence from Legg, Hutter et al. (2007): *"Intelligence measures the agent's ability to achieve goals in a wide range of environments"*.

Early AI techniques were primitive and mostly rule-based. However, these were

only suited to a limited spectrum of tasks. A new concept within AI, ML, emerges from this limitation. ML can be defined as the process of learning a function $f : X \rightarrow Y$ that maps an input X to an output Y (MITCHELL et al., 1997). However, the learning process is not always linear, given the real-world complexity. In the ML context, learning can be carried out in five main ways: supervised, unsupervised, semi-supervised, weakly supervised, self-supervised, and reinforcement learning.

In **supervised learning**, models are constructed from a large number of training examples, with each example containing a label indicating the ground truth (ZHOU, 2018). However, using supervised learning is not always possible in several areas, given the need for problem-domain knowledge and the high costs for data collection (ZHANG et al., 2020). An alternative is to use **unsupervised learning**. In this paradigm, learning occurs without prior data knowledge, so there is no label for each sample. The goal of unsupervised learning is to identify patterns in the input without specific feedback (RUSSELL; NORVIG, 2020).

Semi-supervised and weakly supervised learning can be considered a mixture of supervised and unsupervised learning. For the use of **semi-supervised learning** techniques, only part of the data has labels (ENGELEN; HOOS, 2020). **Weakly-supervised learning** is a problem defined as the learning process based on partial labels, such as image-level labels (ZHANG et al., 2021). In **self-supervised learning**, the model learns to generate its own labels from the data itself by defining pseudo-labels based on the inherent structure or properties of the data (MISRA; MAATEN, 2020). Unlike fully supervised learning, which requires extensive labeled data, self-supervised techniques allow models to learn useful representations from large amounts of unlabeled data. These learned representations can later be fine-tuned with a smaller labeled dataset for specific downstream tasks (JING; TIAN, 2020). Finally, in **reinforcement learning**, the agent learns through rewards and punishments (RUSSELL; NORVIG, 2020). Reinforcement learning techniques can be characterized as agents who learn a policy from reward signs interacting with their environment. The agent aims to find an ideal policy that maximizes its cumulative reward (SUTTON; BARTO, 2018).

Recent advances in computing have allowed intelligent systems development for specific tasks with human-like cognitive capacity (JANIESCH; ZSCHECH; HEINRICH, 2021). The most recent advance, driven primarily by increased computing power and massive amounts of data, has been the evolution of Artificial Neural Networks (ANNs) towards ever-deeper architectures with enhanced capabilities (GOODFELLOW; BENGIO; COURVILLE, 2017). These architectures are part of studying a sub-area of AI defined as DL (LECUN; BENGIO; HINTON, 2015). Concerning this focus, we discuss relevant DL details in the following subsections.

2.3.1 Deep Learning

A key aspect of DL is the ability to extract representations from the data using a general-purpose learning process based on non-linear units (LECUN; BENGIO; HINTON, 2015). The organization of these units takes place in tens or hundreds of layers that learn the representations through artificial neurons (GOODFELLOW; BENGIO; COURVILLE, 2017). Although several concepts were developed based on the human brain understanding, the current mechanisms of DL cannot be considered as artificial brains (CHOLLET, 2021). However, deep neural networks have been used successfully for tasks such as computer vision (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SIMONYAN; ZISSERMAN, 2014; LECUN; BENGIO; HINTON, 2015; HE et al., 2016; CHOLLET, 2017; TAN; LE, 2019; DOSOVITSKIY et al., 2021; LIU et al., 2022a), speech recognition (NASSIF et al., 2019), NLP (OTTER; MEDINA; KALITA, 2020), and reinforcement learning (YU et al., 2021). Next, we cover the theoretical foundations of artificial neurons and ANNs necessary to understand the methods related to this dissertation.

2.3.1.1 Artificial Neural Network

An ANN is organized into layers composed of several units, formally called neurons. The artificial neuron design still follows the design proposed by McCulloch and Pitts (1943). The first neuron was limited to binary inputs. Then, an evolution came with the perceptron-like neuron definition (ROSENBLATT, 1958). The perceptron's main contribution lies in combining the inputs into a weighted sum, and if the sum exceeds a certain threshold, the neuron produces a positive output (RUSSELL; NORVIG, 2020). An activation function determines this limit. So let y be the output, x_i and w_i the input and weight for the sample i, and σ the activation function, we have:

$$y = \sigma\left(\sum_{i=0}^{n} w_i * x_i\right) \tag{2.1}$$

Activation functions determine how the neuron will activate (RUSSELL; NORVIG, 2020). There are some activation functions in the current literature that have different uses. Below, we list some of the most important functions:

- **Step:** *step* function was the first activation function used. Currently, the *step* function is not used in ANNs because it is not differentiable (HAYKIN, 2009).

The *step* function varies between -1 and 1 and is defined by:

$$\alpha(v) = \begin{cases} 1 \text{ if } v \geq 0 \\ -1 \text{ if } v < 0 \end{cases} \tag{2.2}$$

- **Sigmoid:** *sigmoid* outputs values between 0 and 1, making it the preferred for probability prediction problems (RUSSELL; NORVIG, 2020). The *sigmoid* function is defined by:

$$\alpha(v) = \frac{1}{1 + e^{v}} \tag{2.3}$$

- **Tanh:** the Hyperbolic Tangent ($Tanh$) is a sigmoid variation. The main advantage compared to *sigmoid* is related to the function's derivative, which presents larger values that accelerate the finding of the global minimum (LECUN et al., 2012). The $Tanh$ function varies between -1 and 1 and is defined by:

$$\alpha(v) = tanh(v) \tag{2.4}$$

- **ReLU:** e Rectified Linear Units (ReLU's) main advantage over other activation functions is that it usually does not suffer from gradient saturation (FUKUSHIMA, 1969). The ReLU function varies between 0 and 1 and is defined by:

$$\alpha(v) = max(0, v) \tag{2.5}$$

However, the perceptron neuron still had limitations related to the weights optimization and the impossibility of application in nonlinear problems (MINSKY; PAPERT, 1969). To deal with this limitation, the organization of neurons in multilayers was proposed. This structure was called Multilayer Perceptron (MLP). In addition to layering, MLP networks take advantage of three other concepts: (i) nonlinear activation functions, (ii) weight optimization using gradient descent, and (iii) backpropagation (COPPIN, 2010). A MLP network can be formally defined as:

$$y^i = \sigma\left(\sum_{i=0}^{n} W^i y^{i-1} + b^i\right) \tag{2.6}$$

where y^i is output from the layer i, W^i corresponds to the weights of the layer, y^{i-1} the outputs from the previous layer, b^i to the bias of the current layer, and σ the activation function (HAYKIN, 2009).

We can use an error-based methodology in MLP to carry out the weight optimization process. The error is measured at the end of each epoch or iteration using a loss function (RUSSELL; NORVIG, 2020). An epoch is the complete pass of the training set. Meanwhile, an iteration is a partial pass of the training set. Loss

functions measure the distance between the network predictions and the expected outputs (CHOLLET, 2017). It is necessary to define the loss functions according to the problem that the ANN will be proposed to solve (RUSSELL; NORVIG, 2020). This need is linked to weight optimization, which depends on the loss function output (GOODFELLOW; BENGIO; COURVILLE, 2017). Finally, it is essential to note that the objective of the training process is to minimize the loss function output value (CHOLLET, 2017).

Therefore, training an ANN consists of minimizing the loss (GOODFELLOW; BENGIO; COURVILLE, 2017). Currently, gradient descent is used for this process. Considering a scenario with two weights (w_0, w_1), the gradient descent computes the loss function gradient. It moves the weights a little towards the gradient descent, repeating the process until it finds a local or global minimum for the loss (RUSSELL; NORVIG, 2020). Therefore, considering a simple neuron definition with a single input (x, y), we have:

$$h_w(x) = w_1 x + w_0 \tag{2.7}$$

To find the weights optimization direction, the loss function partial derivative is calculated:

$$\partial h_w(x) = w_1 x + w_0 \tag{2.8}$$

Considering a quadratic loss, the minimum is calculated by the partial derivative of the function:

$$\frac{\partial}{\partial w_i} Loss(w) \tag{2.9}$$

Applying the chain rule, we have:

$$\frac{\partial}{\partial w_i}(y - h_w(x)))^2 = 2(y - h_w(x)) \times \frac{\partial}{\partial w_i}(y - h_w(x)) \tag{2.10}$$

Applying Equation 2.10 to each of the weights w_0 and w_1, we have:

$$\begin{aligned} \frac{\partial}{\partial w_0} Loss(w) &= -2(y - h_w(x)) \\ \frac{\partial}{\partial w_1} Loss(w) &= -2(y - h_w(x)) \times x \end{aligned} \tag{2.11}$$

Adding the α learning rate, which is responsible for setting the step size when updating weights (COPPIN, 2010). In this way, we have:

$$\begin{aligned} w_0 &\leftarrow w_0 + \alpha(y - h_w(x)) \\ w_1 &\leftarrow w_1 + \alpha(y - h_w(x)) \times x \end{aligned} \tag{2.12}$$

Finally, in a training set with N examples, we have:

$$\begin{aligned} w_0 &\leftarrow w_0 + \alpha \sum_j (y_j - h_w(x_j)) \\ w_1 &\leftarrow w_1 + \alpha \sum_j (y_j - h_w(x_j)) \times x_j \end{aligned} \tag{2.13}$$

This algorithm is called batch gradient descent (RUSSELL; NORVIG, 2020). Over the years, several extensions have been proposed, such as Stochastic Gradient Descent (SGD) (ROBBINS; MONRO, 1951), Adagrad (DUCHI; HAZAN; SINGER, 2011), RMSProp (HINTON; SRIVASTAVA; SWERSKY, 2012), Adam (KINGMA; BA, 2014), or Nadam (DOZAT, 2016). Despite Adam's popularity, there is still no universal weight optimization method applicable to any network or task (DOGO et al., 2018).

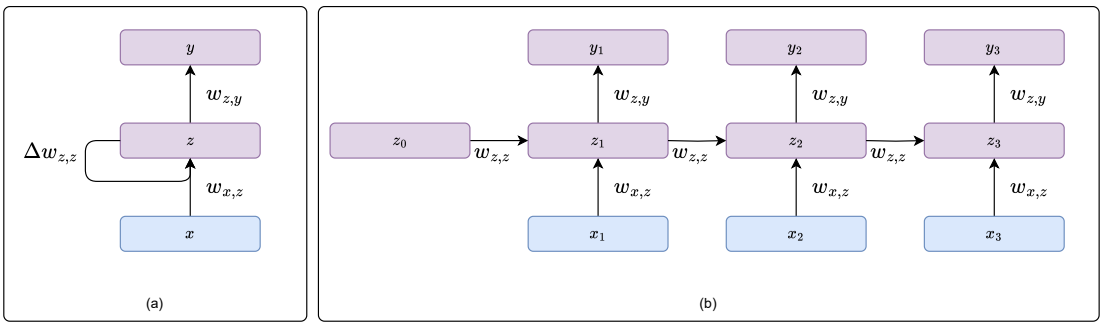
To replicate this optimization process to all the weights of the layers of an ANN, the backpropagation algorithm is used (RUSSELL; NORVIG, 2020). In the backpropagation algorithm, the gradient is applied recursively by the chain rule, updating the values of our weights (COPPIN, 2010). Finally, a problem in optimizing weights in large networks using gradient descent happens when partial derivatives are small or zero at some weights, updates not have significance in the output. This issue is known as vanishing gradient (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.3.1.2 Recurrent Neural Network

Several tasks are sequential, for example, in prediction any input can depend on past inputs. In this sense, RNNs have a mechanism that allows a cycle in the inference process. The loop consists of persisting the previous neuron output in the current neuron inference (RUSSELL; NORVIG, 2020). Figure 2 shows a schematic of the basic mechanism of an RNN.

Over time, several specialized RNNs were proposed (HOCHREITER; SCHMIDHUBER, 1997; CHO et al., 2014; KRAUSE et al., 2016). For example, RNNs have difficulties with long-term dependencies (BENGIO; SIMARD; FRASCONI, 1994). In this sense, Long Short-Term Memory (LSTM) uses a memory cell, which essentially keeps a copy of important information from past entries (HOCHREITER; SCHMIDHUBER, 1997). This cell has several gates, which are artificial neurons

Figure 2 – Unrolled recurrent neural network. (a) Diagram of an RNN with a Δ delay. (b) View of an RNN along three entries.



Source: Adapted from Russell and Norvig (2020).

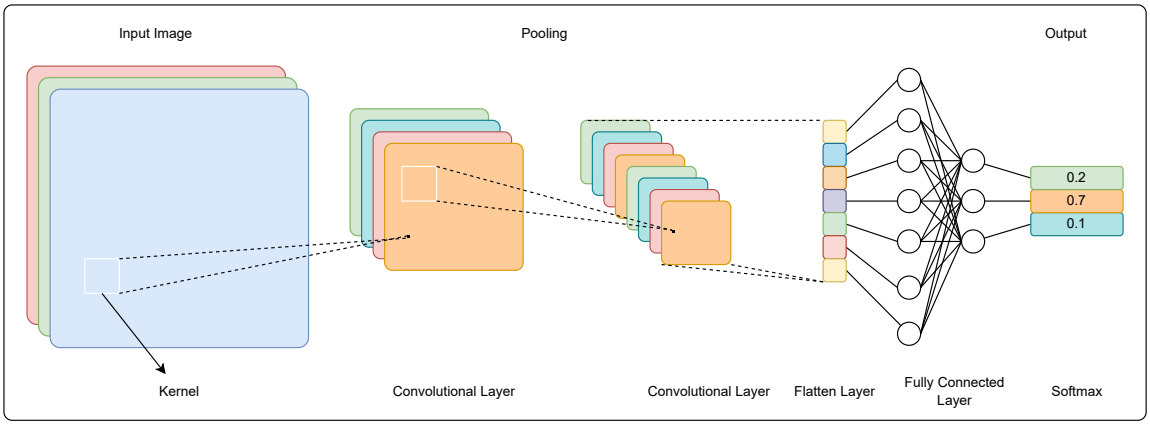
deciding what information is relevant or not for the memory cell (GOODFELLOW; BENGIO; COURVILLE, 2017). A traditional LSTM has four gates: forget, input, cell state, and output gate. The forget gate decides what information should be passed or withheld.

Meanwhile, the input gate receives the previous state and the current input, which are combined using an output multiplication of *sigmoid* and *tanh* functions to update the cell state. The cell state is updated by the sum of the forget and input gates outputs. Finally, the output gate combines the current input, the previous output, and the cell state, determining the LSTM current output (RUSSELL; NORVIG, 2020).

2.3.1.3 Convolutional Neural Network

Human vision can identify patterns and objects in fractions of a second (GEIRHOS et al., 2017). However, for ANNs, pattern detection in images suffers from some limitations. The main one is the need to process an image in vector format, eliminating the dependency relationship between image pixels (RUSSELL; NORVIG, 2020). In this context, Convolutional Neural Network (CNNs) showed significant evolution in the pattern recognition field, especially in image processing (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SIMONYAN; ZISSERMAN, 2014; LECUN; BENGIO; HINTON, 2015; HE et al., 2016; CHOLLET, 2017; TAN; LE, 2019; DOSOVITSKIY et al., 2021; LIU et al., 2022a). Generally, a CNN architecture is made up of several layers. In Figure 3, we present a structure with the main layers of a CNN architecture. In the following subsections, we present the CNN essential components and some key ideas.

Figure 3 – Illustration of the main layers of a Convolutional Neural Network.



Source: Adapted from Rawat and Wang (2017).

2.3.1.3.1 Convolutional Layer

An asterisk formally denotes the convolution operation and is traditionally used in signal processing (GOODFELLOW; BENGIO; COURVILLE, 2017). A convolution operation is defined by:

$$s(t) = (x * w)(t) \tag{2.14}$$

where x is the input, w the weights, t is the time, and $s(t)$ is the output over time. For the CNN domain, x is referred to the input image I, w to the kernel K, and the output to the feature map S. Therefore, we can adapt the convolution process to the following function:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2.15}$$

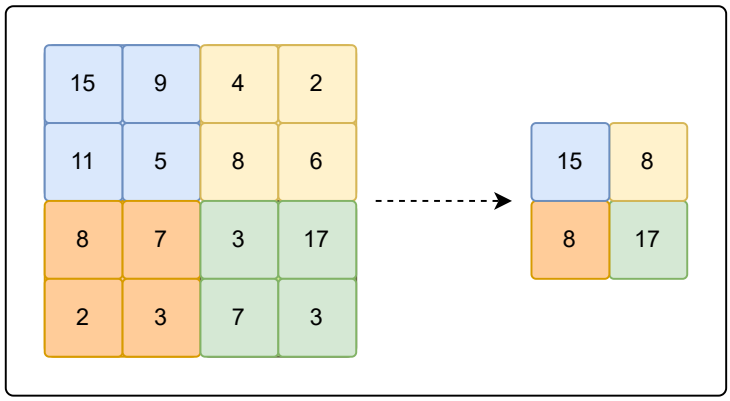
The use of CNNs concerning ANNs is linked to two main aspects: sparse connections and parameter sharing. Sparse connections reduce the connections between layers, keeping the essential characteristics of each region (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Parameter sharing is related to the kernel's ability to identify a specific pattern across the entire image, unlike ANNs where each parameter is used only once (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.3.1.3.2 Pooling Layers

The pooling layer's main objective is to reduce the image spatial dimension and consequently the computation cost (LECUN; BENGIO; HINTON, 2015). The two most popular pooling layers are max and average pooling. The max-pooling selects

the maximum value in a neighborhood. Meanwhile, average pooling selects an average value in the neighborhood (GOODFELLOW; BENGIO; COURVILLE, 2017). In Figure 4, we present an example of the max-pooling process.

Figure 4 – $2x2$ pixel max-pooling process with a 2 pixel stride applied to a $4x4$ pixel image, resulting in a $2x2$ pixel image.



Source: Adapted from Rawat and Wang (2017).

2.3.1.4 Attentions-based Architectures

Attention-based architectures initially emerged for NLP, especially in text translation (BAHDANAU; CHO; BENGIO, 2014). The attention mechanism solved a common problem, the RNN's inability to remember long sentences in the translation process (CHAUDHARI et al., 2021). The main motivation for attention mechanisms is human vision. Our view tends to selectively focus on specific regions, ignoring irrelevant information (XU et al., 2015).

For example, an attention mechanism represents a word's importance concerning others in a sentence. This process is achieved through three arrays: query Q, key K, and value V. An attention function can generally be defined as a mapping of Q and (K, V) to output as a weighted sum of V, where the weight associated with each V depends on Q and K. An Attention layer can be calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.16}$$

where d_k are the dimensions of the queries and keys.

Attention mechanisms were initially proposed for NLP, particularly the translation task (CHAUDHARI et al., 2021). Since then, several architectures have emerged, such as Transformers (VASWANI et al., 2017), and Vision Transformer (ViT) (DOSOVITSKIY et al., 2021). Another benefit of the attention

mechanism is the increased interpretability by explicitly showing how much each input element contributes to the prediction with the attention weights (CHAUDHARI et al., 2021).

2.3.1.5 Dropout

A ML model mainly aims to maintain performance on previously unobserved data. This ability is called generalization (GOODFELLOW; BENGIO; COURVILLE, 2017). In this sense, two key aspects are associated with generalization problems: underfitting and overfitting. Underfitting occurs when the model performs poorly during training, while in overfitting, the model has only learned patterns that apply to the training set and does not perform well on new sets (ZHANG; ZHANG; JIANG, 2019). An alternative to prevent overfitting is the use of dropout layers. The main idea is to turn off a percentage of neurons during the training process, allowing each iteration to use a new neural subnet (SRIVASTAVA et al., 2014). With neurons deactivated randomly, the models are not dependent on a neuron or connection, thus reducing overfitting. In addition, dropout layer processing is computationally cheap (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.3.1.6 Batch Normalization

Batch normalization is applied between layers of ANNs during the training and testing process. Batch normalization accelerates the training process and helps to decrease information loss due to very small weights (IOFFE; SZEGEDY, 2015; RUSSELL; NORVIG, 2020). However, the reasons why it is effective in these processes are still not fully understood and are discussed in several recent articles (SANTURKAR et al., 2018; LABATIE et al., 2021). Since H is a set of feature maps of the layer to be normalized, we have:

$$H' = \frac{H - \mu}{\sigma} \tag{2.17}$$

where μ is the average and σ is the standard deviation of the feature maps of the layer to be normalized. H' are normalized feature maps processed the same way as if the network were processing H without normalization (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.4 Survival Analysis

Survival analysis, also known as time-to-event analysis, is an approach that seeks to analyze the expected duration until one or more events occur (KAPLAN; MEIER,

1958). The events of interest can vary widely depending on the context — ranging from time to death, cancer recurrence, or time to failure of a mechanical component (BRADBURN et al., 2003). Survival modeling assumes that observations come from unknown distributions (COX, 1972). In this sense, the main objectives of survival analysis are to estimate the survival times distribution, compare survival experiences among different groups, and model the relationship between survival time and covariates (BRADBURN et al., 2003).

Survival data are described and modeled in two key functions: the survival function $S(t)$, and the hazard function $\lambda(t)$ (CLARK et al., 2003). The survival function $S(t)$ gives the probability that the event of interest has not occurred by time t. Mathematically, it is expressed as:

$$S(t) = P(T > t) \tag{2.18}$$

where T is a random variable representing the time to event. The survival function is a non-increasing function, and it ranges from $S(0) = 1$ to $\lim_{t\to\infty} S(t) = 0$. In other words, the survival function decreases inversely as t, with the initial value being 1 when t is 0 (WANG; LI; REDDY, 2019).

On the other hand, the hazard function $\lambda(t)$ is also known as the instantaneous death rate. The hazard function gives the rate that an individual under observation at a time t will have an event at that time, given that it has not occurred until time t (CLARK et al., 2003). The hazard function is defined as:

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \tag{2.19}$$

The hazard function is not a probability but rather a rate, and it provides insights into the risk of the event occurring at time t, conditional on survival until that time (WANG; LI; REDDY, 2019). The hazard function is particularly useful in identifying high or low-risk periods. For example, the hazard function $\lambda(t)$ can reveal critical periods when patients are at the highest risk of complications or death (CLARK et al., 2003).

2.4.1 Censoring

A unique challenge in survival analysis is the presence of censored data (WANG; LI; REDDY, 2019). Censorship occurs when the exact time of the event is not observed for all subjects (KLEINBAUM et al., 2012). The most common types of censorship are: (i) a patient has not yet experienced an event; (ii) the patient was not followed until the end of the study; and (iii) the patient experienced an event different from those observed in the study (CLARK et al., 2003).

Censoring impact on the interpretation and analysis of survival data (WANG; LI; REDDY, 2019). Without proper handling, censoring can lead to biased estimates and incorrect conclusions (KLEINBAUM et al., 2012). For example, if right-censored observations are not accounted for, estimated survival times may be artificially short since the analysis would ignore that some individuals survived beyond the observed times (LEE; YOON; SCHAAR, 2019).

Furthermore, the amount and type of censoring in a dataset can influence the precision and reliability of estimates (KLEIN, 2003). High levels of censoring reduce the effective sample size, leading to wider confidence intervals and less precise estimates (WANG; LI; REDDY, 2019). Therefore, careful consideration of censoring mechanisms and appropriate statistical methods is essential for accurate survival analysis (JIANG; GUTERMAN, 2024).

2.4.2 Survival Models

Several models are used to analyze survival data, with the Kaplan-Meier estimator and the CoxPH model being among the most widely employed (WANG; LI; REDDY, 2019). Considering censored data, the Kaplan-Meier estimator provides a non-parametric estimate of the survival function $S(t)$ (KAPLAN; MEIER, 1958). It is defined as:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.20}$$

where d_i is the number of events at time t_i and n_i is the number of individuals at risk just before time t_i.

Kaplan-Meier is a non-parametric estimator that is particularly effective in handling censored data, which is common in survival studies where not all subjects experience the event of interest during the observation period (WANG; LI; REDDY, 2019). However, the Kaplan-Meier estimator assumes that the probability of survival is equal for all subjects (JIANG; GUTERMAN, 2024). This implies that it does not consider possible heterogeneities or covariates that may influence the survival rate, such as age, sex, or clinical condition of the individuals (KLEIN, 2003). Furthermore, the method does not distinguish between different risk groups unless survival curves are estimated separately for each group (CLARK et al., 2003).

On the other hand, CoxPH allows us to investigate the effect of several variables on the time until a specific event, such as death or disease recurrence. CoxPH, a semi-parametric model, seeks to estimate the hazard function $\lambda(t|X)$, which describes the instantaneous risk of the event occurring at time t, given that the subject survived until that moment and is conditioned by a set of covariates $X = (X_1, X_2, \ldots, X_p)$ (COX,

1972). Therefore, the hazard function in the Cox model is expressed as:

$$\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) \tag{2.21}$$

where the hazzard function $\lambda(t|X)$ depends on a set of p covariates $X = (X_1, X_2, \ldots, X_p)$, whose impact is measured by the size of the coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ (BRADBURN et al., 2003).

The central assumption of the CoxPH model is the proportional hazards assumption, which states that hazard rates across individuals are constant over time (WANG; LI; REDDY, 2019). This implies that the effect of a covariate on risk is multiplicative and does not change as time progresses (BRADBURN et al., 2003). However, in many real-world situations, the impact of a covariate on risk can change as time passes (KLEIN, 2003). When this assumption is violated, the results of the CoxPH model can become misleading, leading to incorrect conclusions about the relationship between covariates and survival times (LEE; YOON; SCHAAR, 2019).

2.5 Final Remarks

This chapter presented the fundamental concepts of endemic, epidemic, and pandemic diseases, emphasizing their transmission capabilities and geographic scope. Understanding these distinctions is crucial for the global health community to develop effective surveillance, prevention, and response strategies. The COVID-19 pandemic has highlighted the need for preparedness and adaptability in managing infectious diseases and the importance of continued research and development in zoonotic pathogens.

Finally, we analyze the main aspects related to AI and DL that are involved in defining the model of this dissertation. In this sense, survival analysis models can contribute to health systems in times of overload, providing support mechanisms for patient care. Integrating multimodal data is essential for a complete patient's health status analysis and assertive predictions. In the next chapter, we analyze, through an RLR, the works considered state-of-the-art for survival analysis in multimodal data. We also investigate the current methods used in the literature, challenges, and research opportunities that guided this dissertation.

3 RELATED WORK

The multimodal nature of modern medicine is complex, with unstructured texts, different image modalities, and tabular data, such as laboratory tests. This complexity makes it challenging to extract underlying patterns from healthcare data efficiently. When we analyze data from hospitalized or intensive care patients, this data has even greater degrees of dimensionality. Despite recent architectural advances, particularly in generative AI, the timelessness and longitudinality of data represent considerable challenges for model development. Furthermore, individuals are subject to competing risks, which may lead the individual to experience other outcomes or influence the patient's outcome in a non-linear manner.

Therefore, for most healthcare applications with these data characteristics, the primary goal is to estimate the risk or predict the time to the event of interest for the patient in the future. This type of problem is also known in statistics as survival analysis. Although, over the years, several statistical models have demonstrated significant results for predicting the time of an event of interest, modeling multimodal data is still a challenge for purely statistical models. In this sense, current literature uses models based on ML and DL architectures for survival analysis in multimodal health data.

The primary purpose of this chapter is to provide a comprehensive and structured investigation of survival analysis in multimodal data from hospitalized patients. We demonstrate an overview of traditional statistical models compared to current literature based on ML and DL. Furthermore, we explore how longitudinal and timeless characteristics of data are incorporated into current methods. Finally, we discuss the open challenges and gaps, outlining opportunities for future work.

This chapter is divided into three main parts. The first defines the methodology adopted for the search and selection of articles. Afterward, we discuss and analyze the selected studies, focusing on their applications and challenges. Finally, we analyze the literature's challenges and open research problems, indicating future research directions.

3.1 Methodology

To carry out the analysis of current literature, we used the RLR methodology. The RLR is a simplified approach to synthesizing evidence (HARKER; KLEIJNEN, 2012). The Systematic Literature Review (SLR) is the reference method in healthcare for mapping patient decision aids, clinical practice guidelines, or public policy summaries (TRICCO et al., 2015). Unlike SLRs, which are comprehensive but resource-intensive, RLRs are quicker to conduct and more focused in scope, offering a

timely synthesis of relevant evidence when immediate insights are needed. While not as exhaustive as SLRs, the RLR provides sufficient evidence for informed decision-making in time-sensitive healthcare contexts. It is beneficial when the volume of available literature is manageable or when the required insights are specific and do not require a comprehensive review.

Furthermore, the RLR methodology can help identify research gaps that guide future reviews or original studies, making it a practical complement to SLRs in the evidence synthesis landscape. Therefore, considering the objective of the work, we realized a RLR of the challenges and research opportunities in multimodal survival analysis. The methodology adopted in this bibliographic survey follows the principles and steps defined in previous studies (HARKER; KLEIJNEN, 2012; KHANGURA et al., 2012; TRICCO et al., 2015). Briefly, the steps for executing an RLR are:

- Definition of research questions;
- Selection of keyword and literature databases;
- Article filtering and selection;
- Result analysis and discussion.

3.1.1 Research Questions

We aim to survey the state-of-the-art survival analysis techniques applied to multimodal data fusion in healthcare and biomedical contexts, including time-to-event and competing risk models. Specifically, we are interested in the role of AI, ML, and DL methods in addressing challenges related to hospitalized patients, particularly those hospitalized or in ICUs. To guide our RLR, we divided our research questions into Main Question (MQ) and Specific Questions (SQ). Below, we present our MQ and SQ for this RLR.

- **MQ** - What is the current state of the art on multimodal data survival analysis for hospitalized patients?
- **SQ1** - How are different data fusion techniques used in survival analysis?
- **SQ2** - How do AI-driven models integrate with traditional statistical methods in the analysis of time-to-event data?
- **SQ3** - How does the current state of the art handle the complexities of longitudinal data?
- **SQ4** - How are competing risks incorporated and impact the decision-making for hospitalized patients?

3.1.2 Search Strategy

We need to select a set of relevant studies to identify the answers to the previously proposed questions. In this sense, we defined a set of keywords. The strings were determined based on the author's knowledge and with validation from health professionals. Keywords were grouped using boolean operators, thus forming a search string. At this first moment, we used Google Scholar to search for articles. In Figure 5, we present our search strings. We searched the literature on April 24, 2024.

Figure 5 – Search string used for database query.

Search String

("survival analysis" OR "time-to-event" OR "competing risks")
("multimodal data" OR "data fusion")
("health" OR "healthcare" OR "medical" OR "biomedical")
("artificial intelligence" OR "machine learning" OR "deep learning")
("hospitalized" OR "ICU" OR "hospitalization")

Source: Elaborated by the author.

3.1.3 Article selection

The next step is to define the criteria to filter the raw corpus. The study filtering process comprises a series of steps that aim to select only the works relevant to the RLR objective (ZEISER et al., 2021b). For our survey, the selection criteria were:

- **Publication date:** the corpus should include articles published in the last five years (2019 to 2024) to limit the review to the most recent work in the area;
- **Impurity removal:** removal of duplicate and non-English articles;
- **Filter by title and abstract:** provides an initial screening to remove those not directly related to the RLR;
- **Filter by full text:** complete analysis of the works, selecting only the articles that presented proposals and architectural methods corresponding to our research.

3.2 Findings

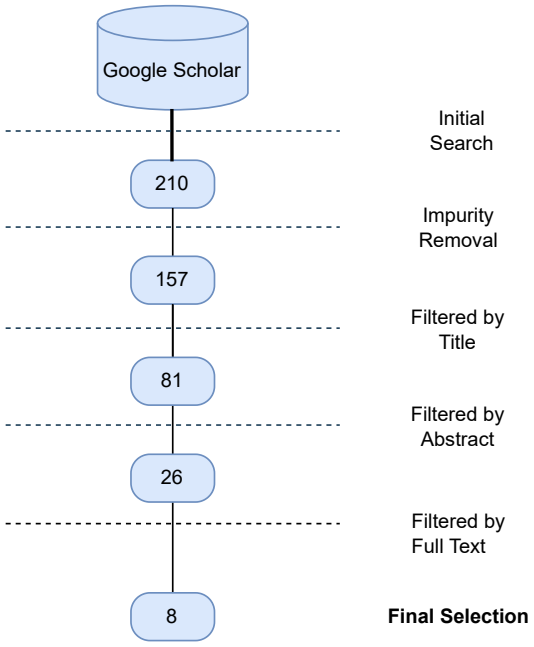
In this section, we present the RLR results and discussion regarding survival analysis techniques. Specifically, we focus on applying AI, ML, and DL to multimodal

data fusion in healthcare and biomedical contexts, particularly involving hospitalized patients or in ICU care. Our findings show how these advanced methodologies manage time-to-event and competing risk analyses.

3.2.1 Article Selection

The process of selecting corpus articles is detailed in Figure 6. Our search returned 210 articles. After removing impurities, 157 articles remained. In the second stage, we analyzed titles and abstracts, removing studies unrelated to the RLR theme and the remaining 26 articles. Finally, we fully analyzed the articles, analyzing the criteria defined in Section 3.1.3. Finally, our final corpus is composed of 8 articles. Table 1 presents our corpus detailing the publication journal/conference and the h5-index.

Figure 6 – Article selection process.



Source: Elaborated by the author.

3.2.2 Literature Analysis

This section analyzes the literature corpus collected using the described methodology. First, we explore studies identifying the current state of the art in multimodal data survival analysis for hospitalized patients, answering MQ1. Then, to address SQ1, we assess the impact of different data fusion techniques on the accuracy

Table 1 – Final corpus of articles published in journals.

Article	Journal/Conference	h5-index
(ZHONG et al., 2024)	IEEE Journal of Biomedical and Health Informatics	91
(SAEED et al., 2024)	ArXiv	-
(TONG; ZHU; LING, 2023)	Heliyon	105
(FU et al., 2023)	Heliyon	105
(YAMGA et al., 2023)	Frontiers in Digital Health	22
(PHILIPP et al., 2022)	Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)	31
(WAN; ZHOU; ZHANG, 2021)	npj Digital Medicine	96
(LEE; YOON; SCHAAR, 2019)	IEEE Transactions on Biomedical Engineering	75

and reliability of survival predictions. For SQ2, we evaluate how AI-driven models integrate with traditional statistical methods in analyzing time-to-event data. Subsequently, for SQ3, we analyze how current methods handle the complexities of longitudinal data. In addressing SQ4, we examine how competing risk models are incorporated and impact decision-making for hospitalized patients. Finally, we highlight the main challenges and open research opportunities in multimodal data survival analysis for hospitalized patients.

3.2.2.1 MQ - What is the current state of the art on multimodal data survival for hospitalized patients?

Historically, survival analysis is based on statistical models, such as the CoxPH model, which assumes a constant hazard rate over time and may not deal with complex nonlinear interactions between covariates (LEE; YOON; SCHAAR, 2019; SAEED et al., 2024). Furthermore, using multimodal data, with interactions between different types of data, is inefficient in traditional statistical models (ZHONG et al., 2024). In this sense, ML models, such as Random Survival Forests and Gradient Boosting Machines, can capture the nonlinear relationships of high-dimensional data with greater efficiency (TONG; ZHU; LING, 2023). However, these ML models may have difficulty dealing with censored data (WANG; LI; REDDY, 2019).

One of the challenges in survival analysis is competing risks. Our corpus deals with competing risks in different ways. The work with greater focus on this issue uses a set of fully connected subnets for each risk and allows the model to learn specific characteristics of each risk (LEE; YOON; SCHAAR, 2019). We also identified hybrid approaches, using DL models for feature extraction and data fusion for processing in Piecewise Exponential Models. This approach allows the model to capture interactions between different risk factors that might influence the likelihood of one event occurring over another (PHILIPP et al., 2022). Adjustments to loss functions to help models deal with competing risks effectively, such as Rank-N-Constrast, which

classifies survival times and applies a contrastive loss function to optimize the correct ordering of these times in the context of competing risks can improve the performance of models (SAEED et al., 2024). A loss function with a similar objective is used in the DynamicDeepHit model (LEE; YOON; SCHAAR, 2019).

Although most ML and DL models can handle temporal data, health data does not have well-defined temporal intervals. In this way, our corpus generally uses the latest observations of specific signals, genomics, or imaging to perform survival analysis. Data integration aims to provide a more holistic view of patient health and disease progression, leading in environmental terms to personalized medicine approaches, where treatment plans are tailored based on patients' individual risk profiles, assessed through survival models (PHILIPP et al., 2022; TONG; ZHU; LING, 2023). However, the loss of context from the patient's health history may decrease the ability to identify the true risk of adverse events and future consequences. The DinamicDeepHit model uses a RNN or LSTM to extract an internal state, allowing the model to maintain an internal state that reflects the historical information of a patient's health status. This state is updated as new data points are processed, allowing the model to dynamically incorporate longitudinal information (LEE; YOON; SCHAAR, 2019).

The interpretability of AI models in healthcare continues to be a significant concern, especially given the complexity of model structures and the critical nature of clinical decision-making. Using clustering techniques, AI models can delineate patient data into distinct clinical phenotypes, thereby increasing inherent insights' transparency and applicability, crucial for developing targeted treatment protocols (YAMGA et al., 2023). Additionally, the integration of Piecewise Exponential Models in DeepPAMM facilitates granular interpretation of hazard assessments at specific time intervals, providing clinicians with a nuanced understanding of risk dynamics over the course of a disease (PHILIPP et al., 2022). Reducing views on survival projections is essential for real clinical applicability. Adopting training mechanisms that isolate the influence of biased environmental characteristics, such as race and gender, from those that are purely predictive can minimize or correct biases (ZHONG et al., 2024). Such methodologies highlight the fundamental attempts to reconcile assistance in stratifying patient risks using AI with the pragmatic needs of medical practice. These techniques aim to ensure that AI tools achieve adequate predictive results and adhere to standards of clarity and practicality required in clinical settings.

3.2.2.2 SQ1 - How are different data fusion techniques used in survival analysis?

To answer this question, we will focus on how articles address the complexity of multimodal data, its heterogeneity, and longitudinal characteristics—most studies in our corpus focus on embedding fusion. Although there is no consensus in the current literature, the way data fusion occurs can be early, intermediate, or late. Different data modalities are aggregated in early fusion before being processed in the models (YAMGA et al., 2023; TONG; ZHU; LING, 2023). The main advantage of this processing is the ability to model correlations and interactions between multimodal data. However, these characteristics also make models susceptible to high dimensional space and, consequently, to more dispersed data.

Another alternative is late data fusion, where expert models extract features, which are aggregated to generate the final decision (SAEED et al., 2024; ZHONG et al., 2024). This method allows the extraction of the most important characteristics of each data domain. Nevertheless, late fusion makes extracting interactions between different data modalities difficult. An alternative is a hybrid or intermediate merger. In this fusion modality, data is processed in separate pipelines for a more abstract representation, which are combined and fed into a final learning model (LEE; YOON; SCHAAR, 2019; WANG et al., 2021; PHILIPP et al., 2022; ZHONG et al., 2024). This method balances between maintaining modality-specific information and taking advantage of intermediate information (LEE; YOON; SCHAAR, 2019).

The effectiveness of data fusion also impacts clinical decision-making by providing clinicians with more detailed and accurate predictions, which are crucial for personalized patient management. For example, integrating genetic data with clinical and imaging data can help predict patient drug responses and disease progression more accurately, leading to better-tailored therapeutic strategies (PHILIPP et al., 2022). In dynamic environments like intensive care units, the ability to integrate and analyze data in real-time can significantly influence patient outcomes. Data fusion techniques incorporating real-time data streams from monitors and sensors into survival models are becoming increasingly important, as demonstrated in studies focused on critical care patient environments (PHILIPP et al., 2022; ZHONG et al., 2024).

3.2.2.3 SQ2 - How do AI-driven models integrate with traditional statistical methods in the analysis of time-to-event data?

In time-to-event analysis, traditional statistical models such as the CoxPH and Kaplan-Meier estimators are applied to datasets where disease progression or treatment efficacy is monitored over time. For example, the CoxPH model can be

used to determine the effect of genetic markers on post-diagnosis survival. In contrast, Kaplan-Meier curves can be used to display the proportion of patients who survive after receiving a new treatment (TONG; ZHU; LING, 2023). Furthermore, the Fine-Gray model is designed to handle cases where different types of events (competing risks) may prevent the primary event of interest from occurring. It provides a way to model subdistribution risk, offering an alternative to the cause-specific risk approach in traditional CoxPH (FU et al., 2023) models.

When we evaluate the integration of these traditional statistical models into our corpus methodologies, the integration of AI-based models occurs mainly as feature encoders (FU et al., 2023; TONG; ZHU; LING, 2023; YAMGA et al., 2023; ZHONG et al., 2024). Feature extraction mainly occurs on high-dimensional data, such as images and patient records (FU et al., 2023; ZHONG et al., 2024). These AI-driven feature extraction techniques can capture complex patterns in data that may not be discernible through manual feature engineering or traditional statistical methods alone. For example, CNNs have been used to extract meaningful features from image data. These are then used as inputs in CoxPH models to more accurately predict patient outcomes (TONG; ZHU; LING, 2023).

Despite advances, the integration of AI and traditional methods presents challenges. Issues such as data heterogeneity, the need for large datasets, and the computational demands of complex models need to be managed (ZHONG et al., 2024). Additionally, there are concerns about the black-box nature of some AI models, which may impede clinical trust and accessibility (TONG; ZHU; LING, 2023). Loss of interpretability of decisions is another significant concern, as clinical adoption requires transparent decision-making processes. Techniques such as Shapley (SHAP) values or partial dependence plots can be integrated to interpret complex ML models (YAMGA et al., 2023). Furthermore, improving the robustness of these models against overfitting, particularly in high-dimensional environments, and validating their performance in diverse situations are crucial steps (LEE; YOON; SCHAAR, 2019).

3.2.2.4 SQ3 - How does the current state of the art handle the complexities of longitudinal data?

One of the main complexities in longitudinal data is the presence of censored or incomplete observations, common in survival analyses (LEE; YOON; SCHAAR, 2019). Most of our corpus does not address the incorporation of longitudinal data, focusing only on the patient's current state (WAN; ZHOU; ZHANG, 2021; FU et al., 2023; YAMGA et al., 2023; ZHONG et al., 2024). This oversight can lead to a lack of dynamic understanding since changes in patient status over time and interactions

between time-varying covariates are crucial for accurate predictions in survival models.

Most statistical methods assume a linear relationship between the covariates and the logarithm of the hazard function (PHILIPP et al., 2022). In models that use DL, this modeling is performed by extending the traditional CoxPH model using a deep neural network to approximate the hazard function in a non-linear way, optimizing the CoxPH partial probability for censored data with regularization techniques to avoid overfitting (KATZMAN et al., 2018). DynamicDeepHit, on the other hand, models survival and event times directly through a shared network architecture that feeds cause-specific subnetworks, employing a loss function that combines log probability for event time and a loss to address competing risks (LEE; YOON; SCHAAR, 2019) effectively.

Furthermore, RNNs and attention mechanisms can be used to model time-varying covariates in survival analysis (LEE; YOON; SCHAAR, 2019). Another alternative is to extend piecewise exponential additive mixed models to handle complex hazard functions and time-varying effects, allowing the model to learn adaptively from longitudinal data with evolving covariate values over time (PHILIPP et al., 2022).

3.2.2.5 SQ4 - How are competing risks incorporated and impact the decision-making for hospitalized patients?

Competing risks in a hospital setting refer to the possibility of a patient experiencing one of several different events, such as death, discharge, or a specific medical complication, with each risk potentially excluding the others (LEE; YOON; SCHAAR, 2019). Understanding these risks is crucial for accurate prognoses and effective patient management (SAEED et al., 2024). Incorporating competing risks into decision-making for hospitalized patients involves understanding and addressing the complexity of potential outcomes that may compete to be the first event to occur (WAN; ZHOU; ZHANG, 2021). This integration has significant implications for prognosis and patient management (LEE; YOON; SCHAAR, 2019).

Techniques such as the cause-specific risks model and the Fine-Gray model are commonly used to estimate the incidence of specific outcomes while considering competing events (FU et al., 2023). These models help understand how different covariates influence the risk of specific outcomes in the presence of competing risks. DL models have introduced sophisticated methods for dealing with competing risks by learning complex data representations that can distinguish between different types of events. For example, models like DeepPAMM use structured additive risk models to flexibly adjust the base risk of each competing event, thereby improving the granularity of risk assessment over time (PHILIPP et al., 2022). Another approach is

to update survival probabilities based on longitudinal data dynamically. This model predicts time-to-event data and adjusts its predictions as patients' health status changes, considering the risk of multiple competing events (LEE; YOON; SCHAAR, 2019).

Incorporating competing risks into survival analysis models has increased their predictive accuracy (LEE; YOON; SCHAAR, 2019). However, integrating competing risks introduces greater complexity in constructing, interpreting, and validating these models (PHILIPP et al., 2022). This complexity can obscure the interpretability of model results, presenting challenges in effectively conveying results to both clinicians and patients, which can, in turn, mitigate the clinical applicability of these analytical tools (YAMGA et al., 2023).

3.2.3 Research Opportunities

Studying survival analysis using multimodal data for hospitalized patients presents challenges and highlights opportunities for future research. One of the main obstacles is integrating diverse data types, such as clinical notes, medical images, and laboratory results, which are often heterogeneous. Effectively incorporating large volumes of unstructured data poses a methodological challenge due to the complexity and variability of the information.

A major concern in this field is the interpretability of models used in multimodal survival analysis. While DL models often achieve high accuracy, their "black box" nature can hinder clinical adoption. Healthcare professionals require models that predict outcomes and provide insights into the underlying mechanisms and factors influencing patient survival, ensuring that predictions are transparent and reliable.

Bias in multimodal datasets is another critical issue that can arise from several sources, including uneven data quality, missing information, and imbalanced patient representation. Addressing these biases is essential to developing fair and reliable predictive models. Current research focuses on developing techniques to mitigate these biases and create robust models to address such issues.

The dynamic nature of patient health data introduces additional challenges. Because a patient's health status can change rapidly, there is a need for models that can update predictions in real-time or near real-time. Incorporating time-varying effects and developing dynamic modeling approaches that adjust predictions as new data become available is crucial for effective clinical decision-making.

Validating and testing multimodal survival analysis models in clinical settings is essential and challenging. The clinical trials or prospective studies required for validation are often time-consuming and expensive. Furthermore, these models must be tested in diverse patient populations and settings to ensure their generalizability

and effectiveness in real-world scenarios. Addressing these challenges is critical to advancing the field and improving patient outcomes.

3.3 Final Remarks

This chapter highlights the complexities and challenges in applying survival analysis to multimodal health data, particularly for hospitalized patients. Traditional models often struggle to integrate multiple data types—such as clinical notes, laboratory results, and imaging—leading to incomplete or less accurate predictions. Furthermore, the dynamic nature of longitudinal patient data, with its irregular time intervals and frequent instances of missing information, further complicates effective modeling.

These challenges, which include integrating multiple data modalities, handling longitudinal information, and ensuring the interpretability of predictive models, are essential to improving the accuracy and utility of survival analysis in clinical settings. Furthermore, interpretability remains a critical concern, as many AI-based models operate as "black boxes," making it difficult for clinicians to understand the logic behind predictions. Thus, this dissertation focuses on the challenges of competing risks with longitudinal and multimodal data to propose an interpretable survival analysis model. The next chapter will explore the architecture in detail, focusing on the design decisions and techniques used to compose the model.

4 MULTSURV MODEL

The COVID-19 pandemic has posed significant challenges to healthcare systems globally (BETTHÄUSER; BACH-MORTENSEN; ENGZELL, 2023). Despite public policy efforts, social isolation practices, and strengthening the capabilities of health systems, medical teams often needed to prioritize patient care (ZEISER et al., 2022). The need to prioritize treatment has challenged ethical and humanitarian principles in broad and profound ways during the pandemic. On several occasions, screening protocols were necessary to optimize resource allocation to patients with a better prognosis and greater potential for maintaining a good quality of life (VERGANO et al., 2020). However, patients' complexity and clinical instability made it extremely difficult to define which patients should be prioritized. Furthermore, the multimodal nature of data in modern healthcare can hide patterns of behavior that are not easily detectable through unimodal analyses.

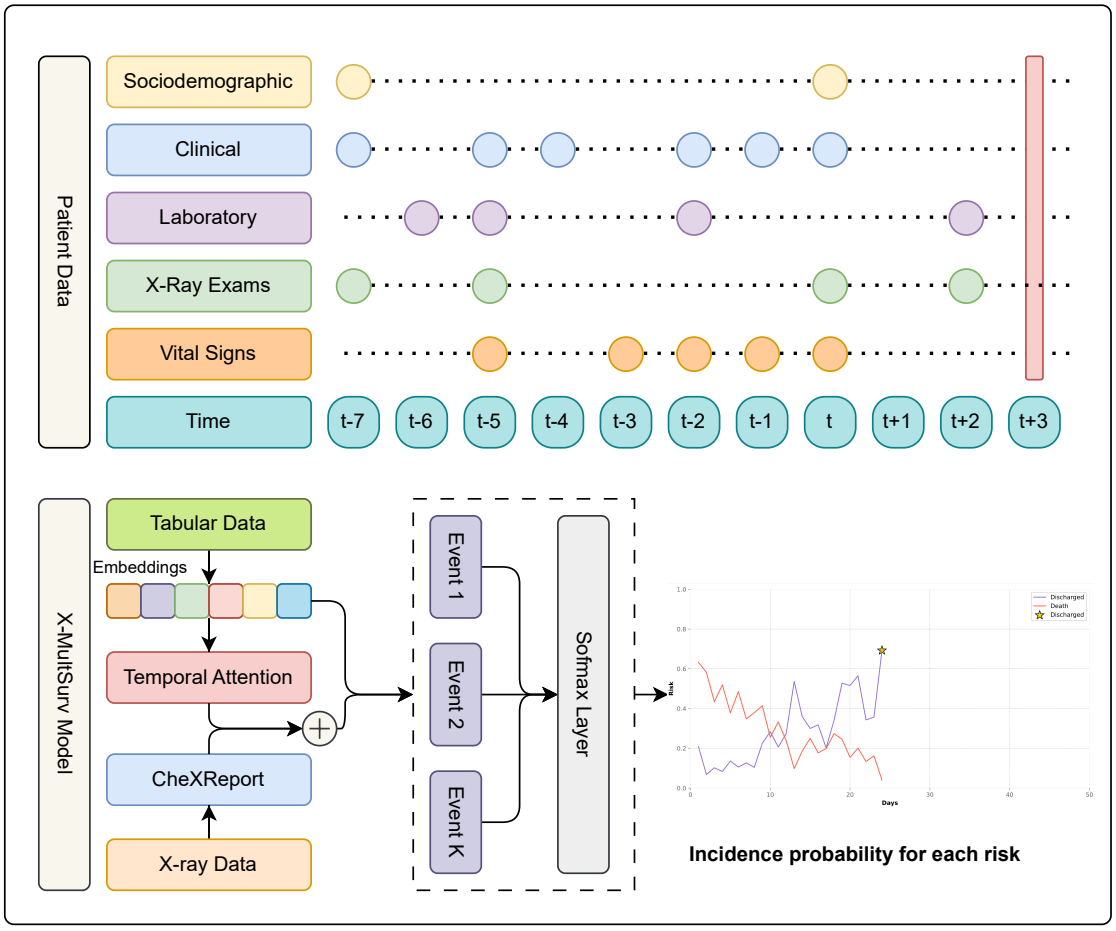
In this context, adopting survival analysis models to define treatment strategies can contribute to identifying patients at higher risk (LEE; YOON; SCHAAR, 2019). Nevertheless, several open research questions are raised in Chapter 3. In this way, the MultSurv model seeks to address the problems of incorporating longitudinal and multimodal data using explainability mechanisms that provide a clearer understanding of the influences of different variables on patient survival. In Figure 7, we present an overview of MultSurv model. It is possible to observe the MultSurv model capability to integrate clinical, laboratory, and imaging data to create a more comprehensive picture of a patient's health status. It allows healthcare professionals to more accurately interpret predictions and underlying risk factors, resulting in management, personalized strategies, and effective treatment.

The name MultSurv was obtained by generating the acronym of the phrase MULtimodal SURVival analysis (MultSurv). The model's design was conducted by observing the research gaps identified in the RLR and the needs identified with partner health institutions. Therefore, the model's main focus is to provide a tool to assist in the screening and follow-up of patients suitable for the hospital routine. This chapter details all the parts that make up the MultSurv model, starting with an overview and explaining each architecture module. Section 4.1 presents the design decisions for the proposed model. Section 4.2 describes an overview and the functionalities of each MultSurv module.

4.1 Project Decisions

One of the current limitations of the traditional survival analysis methods is that they cannot incorporate longitudinal multimodal records. Furthermore, a hospitalized

Figure 7 – MultSurv model overview. The patient can perform different analyses (represented by the colored circles) over the period of interest (t). The red bar represents the patient's outcome. These data may have multimodal natures and are processed in the MultSurv model. Using mechanisms to capture historical context and the current state of tabular and image data to provide risk prediction in a multitask learning architecture.



Source: Elaborated by the author.

patient is constantly monitored for biomarkers and risk factors. In this sense, the design of the MultSurv model considers these factors to incorporate sociodemographic data, biomarkers, laboratory tests, and X-ray images.

Traditional neural network models typically rely on one-hot encoding for categorical variables, treating each category as an independent entity. This approach doesn't capture the inherent relationships or similarities between categories (HUANG et al., 2020). Tabular embeddings, however, allow the model to learn dense vector representations for each category, where similar categories can be positioned closer in the embedding space. This allows the model to implicitly capture relationships and interactions between categories that a traditional neural network model might miss (GORISHNIY et al., 2021). In this sense, for the MultSurv model, we adopted a layer of contextual embeddings for continuous and categorical variables. Furthermore, biomarkers and imaging exams are frequently collected at different periods. To achieve this, MultSurv model has mechanisms to deal with missing data information in the inference process.

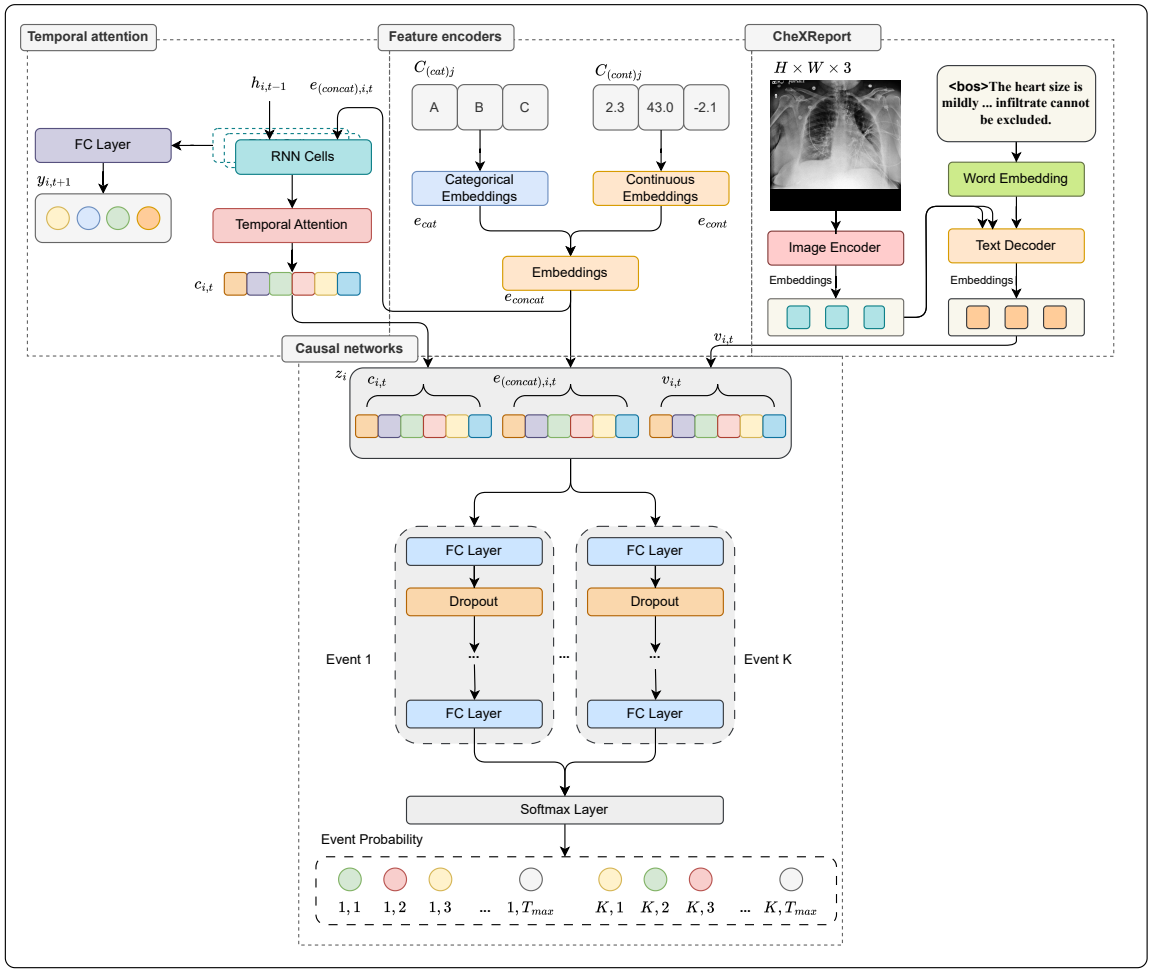
Regarding the challenges of image processing in neural networks, adjustments are needed in three main aspects: (i) image size, (ii) noise, and (iii) annotations. First, neural networks need inputs with a standard size, so in our model, we limit the size of images to 224×224 pixels. Therefore, we resized all images, and to avoid distortions in the images, we added padding to the shortest axis of the original image. Finally, regarding the annotations of the images, we used the X-ray report information already available in the health providers' EHRs or Picture Archiving and Communication System (PACS). In this way, spending human or financial resources to assemble the datasets will not be necessary. In this sense, the MultSurv model findings capability for images will be limited to the information in the X-ray report.

4.2 MultSurv Model Architecture

In this section, we introduce the components of the MultSurv model. In Figure 8, we provide an overview of the MultSurv model components. The MultSurv model is designed to capture information from clinical, laboratory, and imaging data, allowing the temporal and multimodal nature of the information to be captured in the inference process. In Section 4.2.2, we present the definition of the feature encoders used to transform tabular data into dense vector embeddings, which are then used to represent the patient's clinical and demographic information. The temporal attention mechanism, presented in Section 4.2.3, allows the model to focus on the most relevant historical data points, contextualizing the patient's health trajectory. Features from chest X-rays are extracted by the CheXReport architecture (Section 4.2.4). CheXReport uses a fully transformer-based architecture to process

chest X-ray images. This module allows for extracting visual features and integrating them with textual data. The outputs of these components are concatenated and passed through a set of multitask networks (Section 4.2.5). This multitask learning approach allows the model to capture complex relationships between different risk factors. In the following sections, we will explore each component of the MultSurv model, explaining its functionality and role in the overall architecture.

Figure 8 – MultSurv model. Feature encoders process the tabular data, generating a vector of embeddings e_{concat}. Patient i has the samples up to time t processed in Temporal Attention, which generates a contextual vector $c_{i,t}$. The chest X-ray exam is processed by CheXReport, which produces a latent space vector $v_{i,t}$ with the main features identified. The vectors are concatenated z_i and processed in the multitask networks to generate the probability for each time t and event k.



Source: Elaborated by the author.

Finally, we created the Table 2 with the summary of the main notations of each section to support the chapter reading process.

Table 2: Summary of Notations by Section

Notation	Description
	Section 4.2.1 - Pre-processing
$-\infty$	Value assigned to missing data
$B_{a,b}$	Block in CLAHE process
C_i	Set of categorical and continuous covariates for patient i
$J_{(cat)}$	Set of categorical covariates
$J_{(cont)}$	Set of continuous covariates
M_i	Mask representing missing data for patient i
N	Number of patients
T	Maximum time value of the study
X_i	Tuple (C_i, M_i, I_i)
α	Clipping factor in CLAHE
$c_{i,j}$	Patient characteristics of patient i
I_i	Image vectors and report for patient i
t	Time instant
τ	Set of time instants
δ	Event experienced by the patient
γ	Padding size for resizing images
ψ	Clip limit in CLAHE
ρ_i	Vector of survival times for patient i
ω	Dimension of the image after resizing
	Section 4.2.2 - Feature Encoders
Q_j	Set of quantiles for continuous variable j
W_j	Weight matrix for categorical variable j
d	Dimension of the embedding vector
$e_{(cat),j}$	Embedding vector for categorical variable j
$e_{(concat)}$	Concatenated embedding vector
$e_{(per),j}$	Periodic embedding for continuous variable j
$e_{(piece),j}$	Piecewise linear embedding for continuous variable j
$e_{(cont),j}$	Continuous embedding for variable j

Table continues on the next page

Table 2 continued from previous page

Notation	Description
	Section 4.2.3 - Temporal Attention
H_i	Sequence of hidden states over time for patient i
$h_{i,t}$	Hidden state of the GRU at time t for patient i
q	Trainable parameter vector for attention mechanism
$y_{i,t+1}$	Prediction at time $t+1$ for patient i
	Section 4.2.4 - CheXReport
C	Number of channels in the input image
H	Height of the input image
K	Number of tokens in the report
LN	Linear normalization layer
M	Size of non-overlapping windows in Window Multi-head Self-attention
MLP	Multi-layer perceptron
Q, K, V	Query, key, and value matrices in self-attention
S_d	Number of Swin Transformer blocks in the decoder
SW	Shifted Window Multi-head Self-attention
W	Width of the input image
w	Report to be tokenized
y_i	BERT multilanguage embedding of the report tokens
$\hat{x}^l$	Output of W-MSA layer at layer l
x^l	Output of MLP layer at layer l
$\Omega(H_{MSA})$	Computational complexity of traditional Multi-head Self-attention
$\Omega(H_{(S)W-MSA})$	Computational complexity of Window and Shifted Window Multi-head Self-attention
$Attention(Q, K, V)$	Computation of self-attention mechanism
B	Relative position bias matrix
d	Dimension of the query and key in self-attention
$\langle bos \rangle$	Start token for report generation

Table continues on the next page

Table 2 continued from previous page

Notation	Description
	Section 4.2.5 - Multitask Networks
O	Concatenated output of all multitask networks
P	Normalized probability of risk at different time intervals
$e_{(concat),i,t}$	Embeddings from the last collection for patient i at time t
$v_{i,t}$	Latent space vector from CheXReport for patient i at time t
z_i	Input vector to multitask network for patient i
o	Output of multitask network
p	Dropout rate
	Section 4.2.6 - Model Optimization
I_i	Indicator function for event observation for patient i
L_1	Binary cross-entropy loss function for observed events
L_2	Mean squared error for continuous predictions
L_3	Cross-entropy loss for the CheXReport network
$L_{(total)}$	Total loss function
Θ	Model parameters
α_{ctdl}	Attention weights for image regions in CheXReport
$m_{i,t}$	Mask indicating missing variables at time t for patient i
$x_{i,t}$	Predicted value at time t for patient i
$y_{i,t}$	Actual value at time t for patient i

Source: Elaborated by the author.

4.2.1 Pre-processing

Most of the time, an institution's data is stored in several systems. Clinical data is stored in Clinical Information Systems (CIS), images in a PACS, and biomarker data in a Laboratory Information System (LIS). However, the nature and format of data storage varies from system to system and institution to institution. Furthermore, the data collected is subject to different protocols and measurement units.

In this sense, a data pre-processing module is essential for DL models. Concerning the problem addressed in this dissertation, each patient i has a set c of biological marker samples and sociodemographic and clinical characteristics. Furthermore, a patient may have undergone several imaging exams ι and have their respective reports ϵ at different moments of time t. It is essential to highlight that these

collection periods are not regular between each other and among patients. Furthermore, the dataset may be sparse and not have all covariates collected at all time points t. Therefore, for each covariate j missing from patient i, we assign a value $-\infty$.

Therefore, each patient i has a dataset $D_i = \{(X_i, \rho_i, \delta_i)\}_{i=1}^{N}$, where N is the number of patients, X_i is composed of a tuple (C_i, M_i, I_i) collected at an instant of time t, where $t \in \tau = \{1, 2, \ldots, T\}$ and T is the maximum time value of the study. $C_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,j}\}$ is the set of categorical covariates $J_{(cat)} \in J$ and continuous $J_{(cont)} \in J$ variants and statistics over time T. To incorporate information into the MultSurv model regarding missing data, we provide the model with a mask $M = \{m_{i,1}, m_{i,2}, \ldots, m_{i,j}\}$, where:

$$m_{i,j} = \begin{cases} 1, & \text{if } x_{i,j} = -\infty \\ 0, & \text{if } x_{i,j} \neq -\infty \end{cases} \tag{4.1}$$

Furthermore, the model is provided with the set $I = \{\iota, \epsilon\}$, which represent image vectors ι and report ϵ for each patient i with time instants t. For time instants t without image collection $I_{i,t}$ is defined as $-\infty$. Finally, $\rho \in \tau$ is the vector of patient survival times and δ is the event experienced by the patient, where:

$$\delta = \begin{cases} k, & \text{if the individual } ith \text{ was not censored and the event} \\ & \text{occurred due to a cause } k \in K = \{1, 2, \ldots, K\} \\ 0, & \text{if the } ith \text{ individual was censored} \end{cases} \tag{4.2}$$

We adopted a normalization of continuous covariates $J_{(cont)}$ to prevent the model from being susceptible to different magnitude units. In this sense, for each covariate $j \in J_{(cont)}$, we apply normalization by standardization, given by:

$$J_{(cont),z} = \frac{c_z - u_z}{\sigma_z} \tag{4.3}$$

where J_{cont} represents the set of continuous covariates, z represents the variable z of the set of covariates, u_z the mean and σ_z the standard deviation of the covariate.

The set of images I generally comprises ι images of varying dimensions. In this sense, for each image with a height h and width w, we resize the image by a size $\omega \times \omega$. To avoid distortions, we add a padding of size $\gamma = max(h, w) - min(h, w)$ for the axis with the smallest size in pixels. Therefore, the final dimensions h', w' of an image ι are equal to:

$$(h', w') = \begin{cases} (h + \gamma, w) & \text{if } h < w \\ (h, w + \gamma) & \text{if } h > w \\ (h, w) & \text{otherwise} \end{cases} \tag{4.4}$$

Furthermore, collecting chest X-ray images is subject to variations in contrast. The literature commonly applies contrast normalization to images to minimize these variations and assist in the model learning process (ZEISER et al., 2020; JEONG; SUNG, 2022). In this sense, we apply Contrast Limit Adaptive Histogram Equalization (CLAHE), which, first, divides an image ι into smaller blocks B of size $u \times v$, where $u < h$ and $v < w$. For each block $B_{a,b}$, the histogram $H_{a,b}(z)$ is calculated, where z is the pixel intensity level. Then, the clip-limit ψ is calculated:

$$\psi = \alpha \frac{u \times v}{L} \tag{4.5}$$

where, α is a clipping factor and L is the number of image intensity levels.
Then, the histogram is clipped:

$$H'_{a,b}(z) = min(H_{a,b}(z), \psi) \tag{4.6}$$

The pixels are then evenly distributed across:

$$H''_{a,b}(z) = \frac{H'_{a,b}(z) + \sum z(H_{a,b}(z) - \psi)}{L} \tag{4.7}$$

Then a transformation function $F_{a,b}(z)$ is calculated:

$$F_{a,b}(z) = \frac{L-1}{u \times v} \sum_{l=0}^{z} H''_{a,b}(l) \tag{4.8}$$

Finally, we apply the transformation function to each p pixel in the block:

$$p' = F_{a,b}(p) \tag{4.9}$$

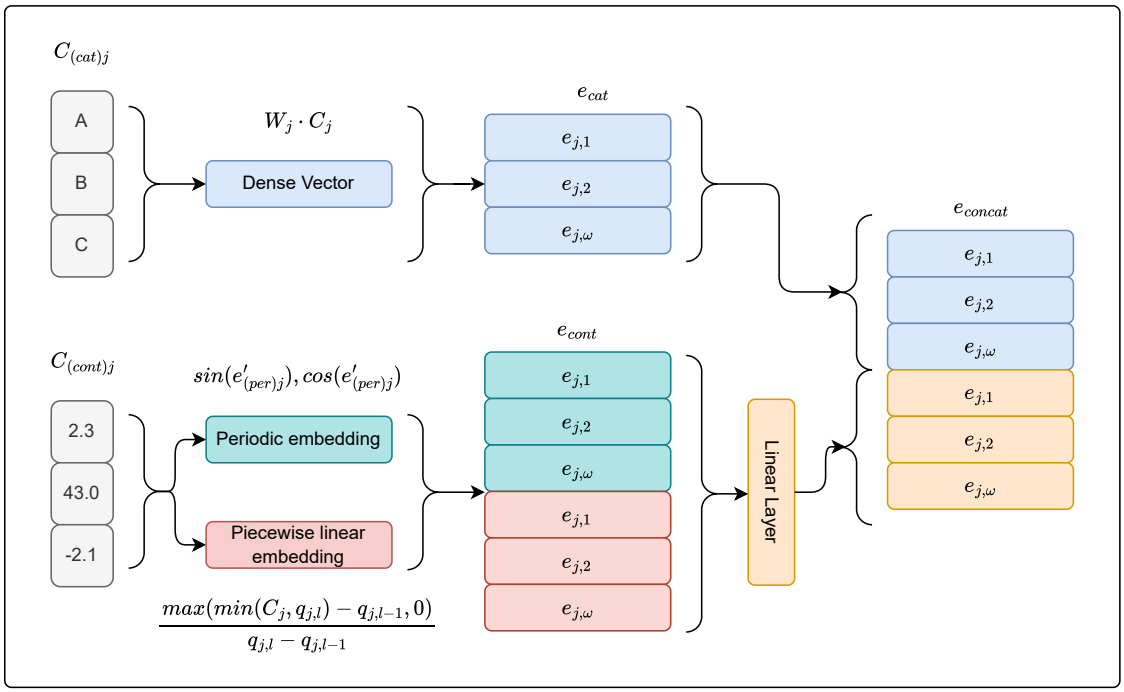
Furthermore, for each pixel p on the edge of the blocks, a bilinear interpolation of the transformed values of adjacent blocks is applied to smooth the transition between blocks.

4.2.2 Feature Encoders

In terms of performance, DL models still struggle to process tabular data compared to tree-based models (HUANG et al., 2020). In this context, we use embeddings to represent categorical and continuous data in the MultSurv model. The use of embeddings allows transforming categorical and continuous variables into dense vector representations and helps capture complex non-linear relationships between variables (HUANG et al., 2020; GORISHNIY et al., 2021). In Figure 9, we present an overview of the embedding generation process for the MultSurv model. The embedding generation process combines two inputs: categorical and continuous

variables. The output of these embeddings are concatenated and passed through a linear layer to produce the final embedding vector.

Figure 9 – Illustration of the MultSurv model embedding generation process for categorical and continuous variables. Each of the variables, categorical $C_{(cat),j}$ or continuous $C_{(cont),j}$, goes through the embedding generation process. These vectors are concatenated and form the input embedding vector e_{concat} of MultSurv model.



Source: Elaborated by the author.

In this sense, for each variable j in the set $J_{(cat)}$, we map a dense vector $e_j \in \mathbb{R}^d$:

$$e_{(cat),j} = W_j \cdot C_j \tag{4.10}$$

where $W_j \in \mathbb{R}^{\omega \times d}$ is the weight matrix for the categorical variable j, ω is the number of unique categories in the variable j and d is the dimension of the embeddings vector.

Meanwhile, we use two embedding techniques for continuous variables: periodic embedding and piecewise linear embedding. For each variable j in the set $J_{(cont)}$ we calculate a periodic embedding $e_{(per)j}$, which is given by:

$$\begin{aligned} e'_{(per)j} &= 2\pi W_{(per)} C_j \\ e_{(per)j} &= \left[sin(e'_{(per)j}), cos(e'_{(per)j}) \right] \end{aligned} \tag{4.11}$$

where $e_{(per)j} \in \mathbb{R}^{N \times 2d}$ and $W_{(per)}$ is the weight matrix of the linear layer.

Additionally, we apply piecewise linear embedding to the set of continuous variables for a set of quantiles $Q_j = \{q_{j,0}, q_{j,1}, \ldots, q_{j,\diamond}\}$, where $\diamond$ is the total of quantiles. To calculate the piecewise linear embedding, it is given by:

$$e_{(piece),j} = \frac{max(min(C_j, q_{j,l}) - q_{j,l-1}, 0)}{q_{j,l} - q_{j,l-1}} \text{ for } l = 1, ..., m \tag{4.12}$$

The embeddings for the continuous variables are then combined using a linear layer:

$$e_{(cont),j} = W_{(cont),j}\left[e_{(piece),j}, e_{(per),j}\right] + b_{(cont)} \tag{4.13}$$

where $W_{(cont)} \in \mathbb{R}^{N \times 3d}$ are the weights and $b_{(cont)}$ is the bias of the linear layer.
Finally, we concatenate the categorical and continuous embeddings into a single vector:

$$e_{(concat)} = \left[e_{(cat)}, e_{(cont)}\right] \in \mathbb{R}^{N \times J \times 4d} \tag{4.14}$$

4.2.3 Temporal Attention

Longitudinal data collected from patients throughout their hospitalization provides historical information on health status over time. In this sense, we adopted a temporal attention network similar to the one proposed in DynamicDeepHit (LEE; YOON; SCHAAR, 2019). In Figure 10, we illustrate the temporal attention mechanism of the MultSurv model. We use a Gated Recurrent Unit (GRU) for the temporal attention mechanism to process the embedding sequences. The hidden state of the GRU is updated at each time step. This way, for each time step $t = 1, ..., T-1$ the GRU cell processes the embedding vector $e_{(concat),i,t}$ and the previous hidden state $h_{i,t-1}$, generating a new hidden state $h_{i,t}$:

$$h_{i,t} = GRU(e_{(concat),i,t}, h_{i,t-1}) \tag{4.15}$$

We then compute a context vector from the hidden states generated by the GRU over time. To calculate the context vector, we use an attention mechanism to weigh each hidden state's importance. The sequence of hidden states over time is given by $H_i = [h_{i,1}, h_{i,2}, \ldots, h_{i,T}] \in \mathbb{R}^{T \times d_{(hidden)}}$. Therefore, the calculation of the context vector is given by:

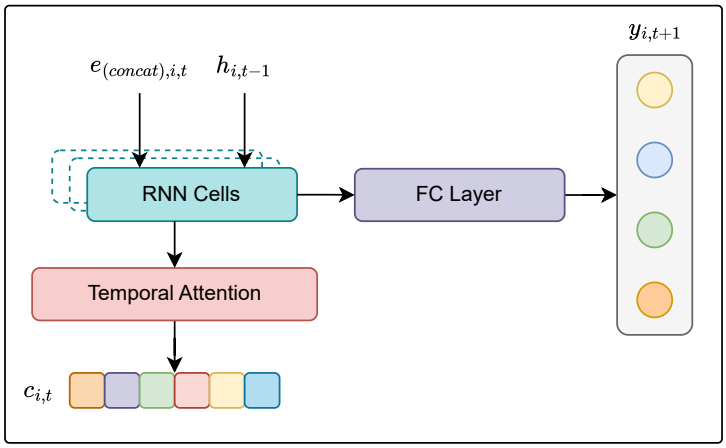
$$C_i = \sum_{t=1}^{T} \varsigma_{i,t} h_{i,t} \tag{4.16}$$

where C_i is the context vector of patient i, $\varsigma_{i,t}$ is given by a softmax function:

$$\varsigma_{i,t} = \frac{e^{(e_{i,t})}}{\sum_{t=1}^{T} e^{(e_{i,t})}} \tag{4.17}$$

and $e_{i,t}$ is a similarity between each hidden state $h_{i,t}$ and a vector of trainable parameters q, given by:

Figure 10 – The MultSurv model temporal attention mechanism for categorical and continuous variables. The embedding vector e_{concat} and the previous hidden state $h_{i,t-1}$ are input into an RNN network to extract an attention-based temporal vector $c_{i,t}$. The current state of the RNN network serves as input to a linear layer that predicts the next value of each variable of interest.



Source: Elaborated by the author.

$$e_{i,t} = q^T h_{i,t} \tag{4.18}$$

The context vector's goal is to provide the MultSurv model with a compact representation of the relevant temporal information. The temporal attention network generates a longitudinal prediction $y_{i,t+1}$. The prediction is generated by:

$$y_{i,t+1} = W c_i + b \tag{4.19}$$

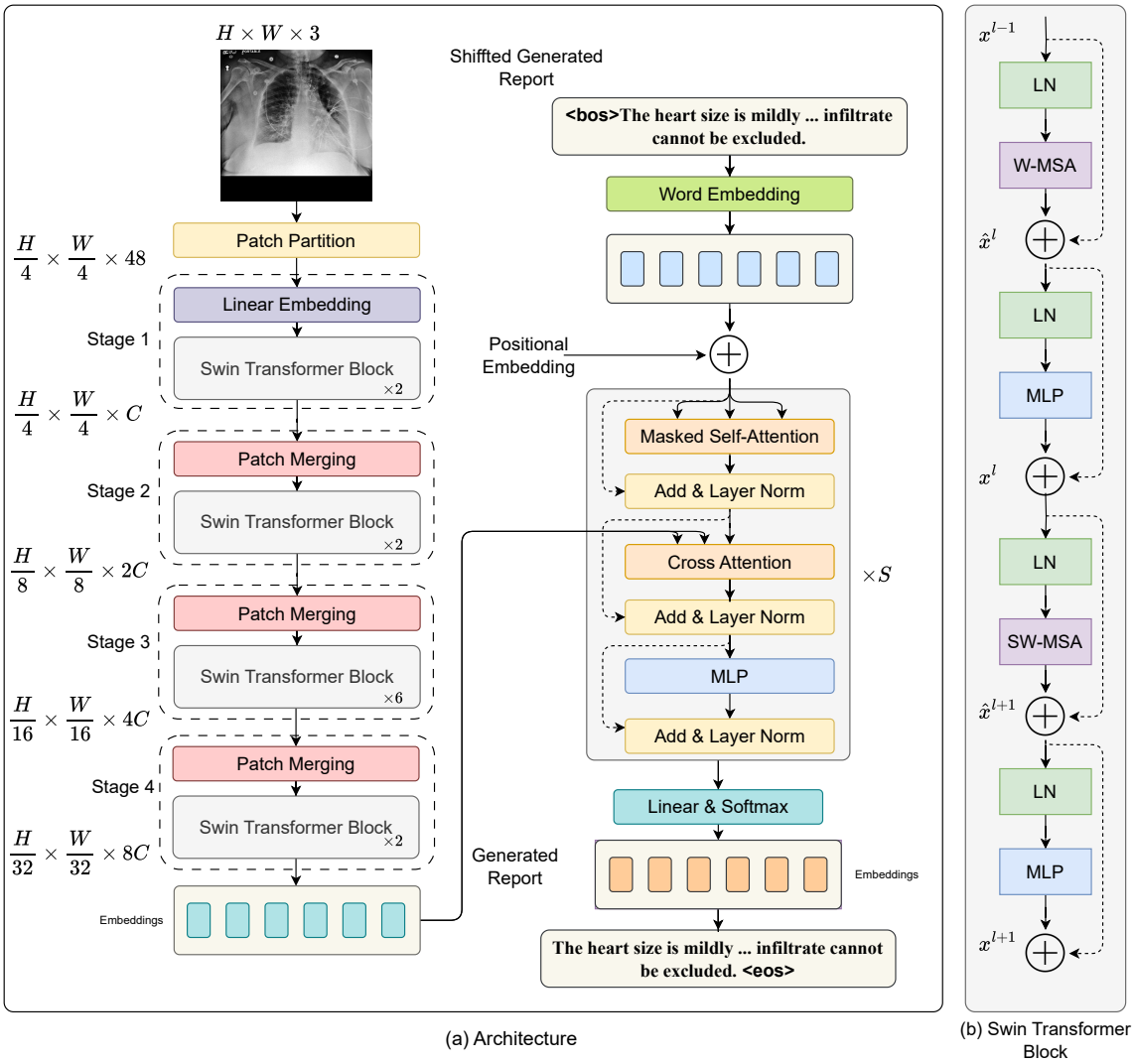
This prediction aims to allow the regularization of temporal network information in a way that keeps relevant information for predictions one step ahead (LEE; YOON; SCHAAR, 2019).

4.2.4 CheXReport

The CheXReport is composed of an encoder-decoder architecture. However, unlike most architectures adopted in the literature, we propose a fully transformer architecture. Adopting a fully transformer architecture with Swin Transformer blocks allows CheXReport to extract better intrinsic visual features and relationships in X-ray images, with improved integration of these visual insights with textual elements. We adopted an encoder with SwinTransformer blocks that extract refined features and intra-relate the features to capture radiological findings from chest X-ray images. The decoder comprises a block of word embeddings connected with S_d Swin Transformer blocks, responsible for merging visual and textual resources. In

Figure 11, we present in detail the CheXReport architecture.

Figure 11 – Overview of the proposed CheXReport network. (a) The text and image encoders are trained together to predict the report suggestion. The X-ray is fed into the Swin Transformer encoder, which extracts the relevant visual features. The text decoder incorporates the visual characteristics, which also receives the text left shifted by one token. (b) We present an organization of the Swin Transformer blocks. Each block is formed by a set of Linear Normalizations (LN), two Multi-Layer Perceptron, one Window Multi-head Self-attention (W–MSA), and one Shifted Window Multi-head Self-attention (SW–MSA).



(a) Architecture

(b) Swin Transformer Block

Source: Elaborated by the author.

4.2.4.1 Encoder

We adopted Swin Transformer blocks (LIU et al., 2021) to reduce the computational cost of an encoder based purely on a ViT model (DOSOVITSKIY et al., 2021). Swin Transformer models can build a set of hierarchical feature maps, merging the intermediate tensors between layers (LIU et al., 2021). The Swin Transformer architecture can progressively reduce the spatial dimension, similar to convolutional networks (LIU et al., 2022b). The reduction is possible by merging patches incorporating a factor s and concatenating the features into a group of $s \times s$ patches (LIU et al., 2021). Compared to ViT's Multi-head Self-attention (MSA) architecture, the spatial constraints added by the Swin Transformer architecture significantly reduce the computational complexity of the layers (LIU et al., 2022b). Traditional MSAs have their computational complexity defined by:

$$\Omega(H_{MSA}) = 4HWC^2 + 2(WC)^2C \tag{4.20}$$

where H, W, and C represent the height, length, and number of channels of the input image. Meanwhile, the Window Multi-head Self-attention (W–MSA) and the Shifted Window Multi-head Self-attention (SW–MSA) (LIU et al., 2021) have their computational complexity defined by:

$$\Omega(H_{(S)W-MSA}) = 4HWC^2 + 2M^2HWC \tag{4.21}$$

Reducing computational complexity is possible for the $(S)W–MSA$ layers by the attention calculation process that is performed locally by dividing a $\mathbb{R}^{H\times W\times C}$ feature map into non-overlapping windows of size $M \times M \times C$. However, this process limits the Swin Transformer to acquire global context over the windows and the loss of connection between the different windows (LIU et al., 2021). To overcome this, the architecture combines alternating $W–MSA$ and $SW–MSA$ layers to connect the feature map's different regions. In $SW–MSA$, the window settings are shifted by $\left(\lfloor\frac{M}{2}\rfloor, \lfloor\frac{M}{2}\rfloor\right)$ pixels of the windows partitioned in $W–MSA$ (LIU et al., 2021). The computation of self-attention is given by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{4.22}$$

where $Q, K, V \in \mathbb{R}^{M^2\times d}$ are the query, key, and value, respectively; d is the dimension of the query and key; M^2 is the number of patches; and $B \in \mathbb{R}^{M^2\times M^2}$ is the relative position bias (LIU et al., 2021).

The final block of a Swin Transformer can be defined as:

$$\begin{aligned} \hat{x}^l &= W\text{–}MSA(LN(x^{l-1})) + x^{l-1} \\ x^l &= MLP(LN(\hat{x}^l)) + \hat{x}^l \\ \hat{x}^l &= SW\text{–}MSA(LN(x^l)) + x^l \\ x^l &= MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1} \end{aligned} \tag{4.23}$$

where $\hat{x}^l$ and x^l are the outputs of $W\text{–}MSA$ and MLP for layer l. LN is a linear normalization layer.

4.2.4.2 Decoder

The decoder receives the chest X-ray report with a start token $\langle bos \rangle$. The report w first goes through a tokenizer that divides the w report into K tokens. We represent one of the report tokens with a BERT multilanguage embedding of 512–D vector, with $y_i \in \mathbb{R}^{512}$ (DEVLIN, 2018). We add a positional embedding passed to a masked self-attention layer. The output of the Masked self-attention layer goes through a layer addition and normalization process to be sent to a cross-attention layer to correlate textual information with the image's visual characteristics, thus producing the report suggestion (LIU et al., 2021). Finally, the last layer of our model is a linear layer with a vocabulary size.

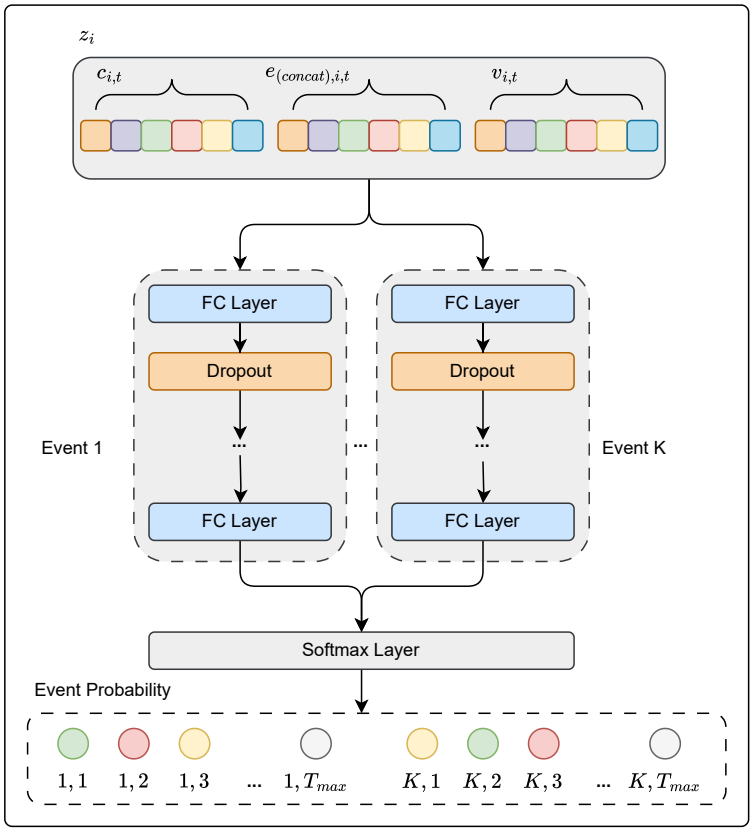
4.2.5 Multitask Networks

We used a multitask learning structure to define the specific risk k that each patient i was subject to at each time point t. Although each network specializes in a particular risk, sharing intermediate information allows the MultSurv model to capture underlying relationships between different risks. Furthermore, each network can specialize in learning specific patterns associated with a type of event and focus on the characteristics most relevant to its event without interference from patterns in other events. In this sense, in the MultSurv model, the multitask networks combine the context vector, the embeddings of the last temporal sample, and a vector from the CheXReport latent space. Each multitask network is a dense neural network that processes this information to estimate the probability of events at different time intervals. We present an overview of how MultSurv model multitask networks work in Figure 12.

The input to each multitask network is the concatenation of the context vector c_i up to time instant t, the embeddings from the last collection $e_{(concat),i,t}$, and a latent space vector from CheXReport $v_{i,t}$, formally represented by:

$$z_i = \left[c_{i,t}, e_{(concat),i,t}, v_{i,t}\right] \tag{4.24}$$

Figure 12 – The multitask networks receive contextual vectors, embeddings, and the CheXReport architecture. Each network predicts a risk k. For each risk k, the MultSurv model generates an extracted output for each time instant t.



Source: Elaborated by the author.

Each multitask network consists of an input layer, several dense hidden layers, and an output layer with nonlinear activation and dropout functions for regularization. The input layer is defined by:

$$h^{(0)} = ReLU(W^{(0)}z_i + b^{(0)}) \tag{4.25}$$

For each $l = 1, \ldots, L$, where L is the number of hidden layers, we have:

$$h^{(l)} = ReLU(W^{(l)}h^{(l-1)} + b^{(l)}) \tag{4.26}$$

After each layer, we apply a dropout for regularization:

$$h^{(l)} = Dropout(h^{(l)}, p) \tag{4.27}$$

where p is the dropout rate. The last layer is given by:

$$o = W^{(L+1)}h^{(L)} + b^{(L+1)} \tag{4.28}$$

The outputs of all multitask networks are then concatenated to form the final output of the MultSurv model.

$$O = [o_1, o_2, \ldots, o_k] \tag{4.29}$$

The concatenated output is then passed to obtain the normalized probabilities of each risk at different time intervals:

$$P = softmax(O) \tag{4.30}$$

4.2.6 MultSurv Model Optimization

To optimize the MultSurv model, we adopt a combination of loss functions. With a composition of loss functions, we can balance the modeling of survival events with the capture of temporal information and chest X-ray images features. In this sense, $L_{(total)}$ is given by the equation:

$$L_{(total)} = L_1 + L_2 + L_3 \tag{4.31}$$

where L_1 is a binary cross entropy function adapted to capture the probability of observed events and is given by:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\left[I_i log_2\left(\sum_{k=1}^{K} m_{i,k}o_{i,k}\right) + (1 - I_i)log_2\left(\sum_{k=1}^{K} m_{i,k}o_{i,k}\right)\right] \tag{4.32}$$

Meanwhile, L_2 is given by a variation of the mean squared error and is given by:

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T} m_{i,t}(1-m_{i,t})(y_{i,t}-x_{i,t})^2 \tag{4.33}$$

For L_2, we adopted a strategy of not considering missing variables in the error calculation to induce the model to understand that a variable with the value $-\infty$ should not influence the extraction of information (LEE; YOON; SCHAAR, 2019).

Finally, L_3 aims to optimize the MultSurv model in relation to feature extraction by the CheXReport network. Therefore, given a sentence $y^*_{1:T}$ with size T and the prediction y^*_t from the model with parameters Θ, we minimize the cross-entropy loss added from double stochastic attention regularization (XU et al., 2015). With the weights $\sum_t \alpha_{ctdl} \approx 1$, we force the model to pay equal attention to each part of the image throughout the generation of chest X-ray report suggestions. Therefore, the loss function can be defined as:

$$L(\Theta) = -\sum_{t=1}^{T} log_2\left(p\Theta(y^*_t|y^*_{t-1})\right) + \sum_{l=1}^{L}\frac{1}{L}\left(\sum_{d=1}^{D}\sum_{i=1}^{M^2}\left(1-\sum_{c=1}^{T}\alpha_{ctdl}\right)\right) \tag{4.34}$$

where D is the number of heads and L is the number of layers.

4.3 Final Remarks

In this chapter, we introduce the MultSurv model. By incorporating longitudinal and multimodal data, the MultSurv model aims to improve the accuracy and interpretability of survival predictions, thereby assisting healthcare professionals in making more informed decisions.

We detail the architecture of the MultSurv model, emphasizing its ability to integrate diverse data types, including clinical, laboratory, and imaging data. The model employs contextual embeddings for tabular data, contrast normalization for images, and a temporal attention mechanism to handle longitudinal data. These components work together to understand a patient's health status over time.

Furthermore, integrating the CheXReport component demonstrates the model's ability to generate detailed and interpretable radiological reports utilizing a fully transformer-based architecture. The next chapter details the implementation and performance evaluations of the MultSurv model.

5 MATERIALS AND METHODS

This chapter presents the methodology used to evaluate the MultSurv model and the technical characteristics and hyperparameters of the developed modules. To evaluate each model, we use public datasets and data from partner institutions. We present in Section 5.1 the details of each dataset. Then, in Section 5.2, the evaluation metrics are detailed. Next, the implementation of the modules and networks are detailed in Section 5.3. Finally, the partial considerations of the chapter are carried out in Section 5.4.

5.1 Materials Description

For training and validation of the MultSurv model, we used the public datasets Curated Dataset for COVID-19, MIMIC Chest X-ray (MIMIC-CXR), and Primary Biliary Cirrhosis (PBC2). In addition, we collected a private dataset at the HCPA. In Table 3, we present an overview of the characteristics of these datasets. Then, in the following subsections, we present a sample and explore each dataset in detail.

Table 3 – Datasets used in the MultSurv model.

Dataset	**Num. of images**	**Finding types**	**Patients**	**Image Annotation**
MDH Dataset	1.066	COVID-19 treatment process, including detailed information on diagnoses, treatments, admissions, ICU admissions, laboratory results, chest X-ray image and report	1,815	Patient and image-level
Curated Dataset for COVID-19 (SAIT et al., 2020)	5,181	Chest X-Ray	-	Image-level
MIMIC-CXR (JOHNSON et al., 2019)	377,110	Chest X-ray and report	65,379	Image-level
PBC2 (MURTAUGH et al., 1994)	-	Primary biliary cirrhosis treatment process, including detailed information on diagnoses, treatments, and laboratory results	362	Patient-level

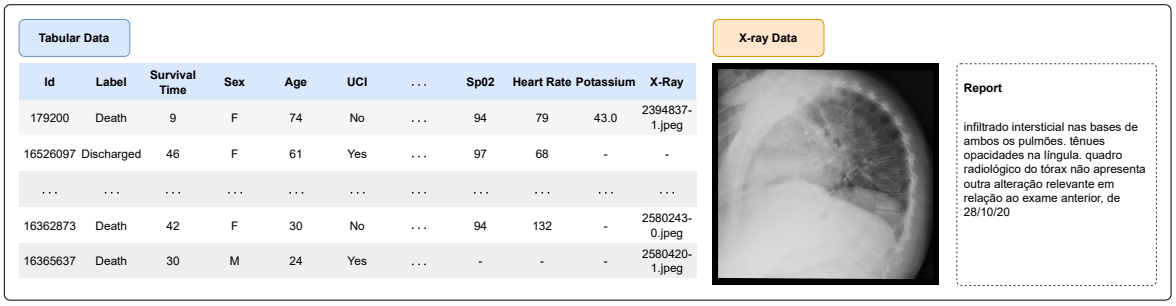
Source: Elaborated by the author.

5.1.1 MyDigitalHealth Dataset

Since the beginning of the COVID-19 pandemic, several datasets with clinical and imaging exams have been publicly available. However, in some situations, this data did not have authorizations for sharing, was shared on social networks, has diagnostic restrictions, or is duplicated in the datasets (ZEISER et al., 2021a). In this sense, MDH focused on building a qualified dataset representing a real pandemic scenario. The MDH dataset was collected at HCPA and covers the period from 03/20/2020 to 06/02/2022. Sociodemographic, clinical, laboratory information, imaging exams, and unstructured electronic medical records were collected from 1,815 patients with COVID-19 confirmed by RT-qPCR.

It is important to highlight that HCPA data collection was conducted retrospectively and longitudinally. In other words, we collected all events of interest to the study throughout the period. Furthermore, as this is a retrospective cohort, we did not control the sampling periods of the variables. Therefore, collection intervals have intra- and inter-patient differences. In Figure 13, we present a sample of data for a study patient. The reports were obtained in a semi-structured format and anonymized. All texts contained only the information from the findings field, without the radiologist's signature or patient's name.

Figure 13 – Ilustration of the MDH survival dataset.



Id	Label	Survival Time	Sex	Age	UCI	...	Sp02	Heart Rate	Potassium	X-Ray
179200	Death	9	F	74	No	...	94	79	43.0	2394837-1.jpeg
16526097	Discharged	46	F	61	Yes	...	97	68	-	-
...	...	...	...	...	...	...	...	...	...	...
16362873	Death	42	F	30	No	...	94	132	-	2580243-0.jpeg
16365637	Death	30	M	24	Yes	...	-	-	-	2580420-1.jpeg

Source: Elaborated by the author.

5.1.2 Curated Dataset for COVID-19

The dataset is a combination of chest X-ray images from different sources. The dataset has X-rays for normal lung, viral pneumonia, bacterial pneumonia, or COVID-19. The authors performed an analysis of all images to avoid having duplicate images. In addition, for the final selection of images, the authors used a CNN to remove images with noise, such as distortions, cropped, and annotations (SAIT et al., 2020). In Table 4, we present the dataset number of images.

Table 4 – Public dataset of chest X-rays used in this article.

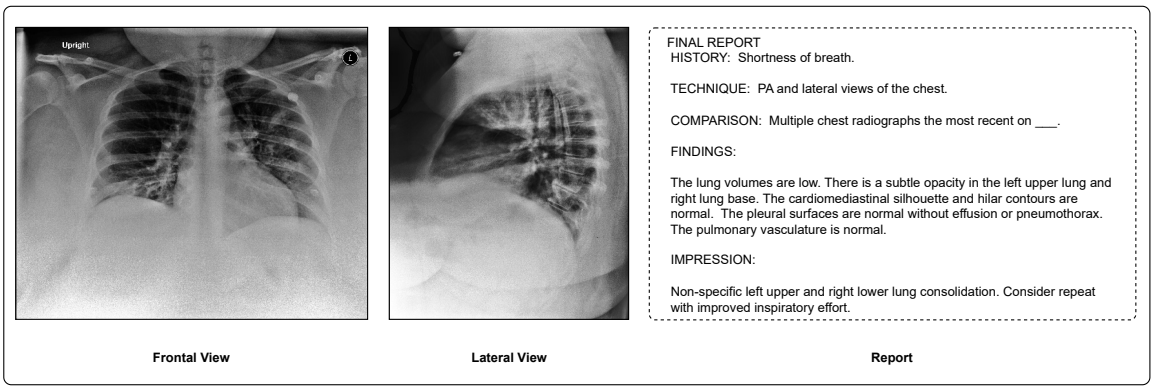
Pathology	Number of images
COVID-19	1,281
Normal	1,300
Viral	1,300
Bacterial	1,300

Source: Sait et al. (2020)

5.1.3 MIMIC-CXR

The MIMIC-CXR dataset is one of the largest public X-ray image datasets. Images were collected at Beth Israel Deaconess Medical Center in Boston, MA. The dataset is anonymized and comprises a set of 65,379 patients with 227,835 studies and 377,110 images and their respective reports. The images were extracted from the Radiology Information System at Beth Israel Deaconess Medical Center in DICOM format. The metadata anonymization process was carried out using Orthanc. The reports made available by MIMIC-CXR were extracted from Radiology Information System in XML format. Areas of the report that corresponded to the institution's header and the professional's signature were excluded from the XML. Furthermore, the free texts of the report were processed to remove sensitive data using a set of regular expressions. In Figure 14, we present a sample of a case from the MIMIC-CXR dataset.

Figure 14 – Example of MIMIC-CXR study with two chest X-ray projections and the report.



Source: Johnson et al. (2019)

5.2 Evaluation Metrics

For the development and evaluation of the MultSurv model, we adopted several performance validations on different learning tasks. Therefore, we present the performance metrics and their characteristics in the following subsections.

5.2.1 Classification Metrics

One of the most common classification model evaluating methods is to organize the predictions in a table format, known as a confusion matrix (SOKOLOVA; LAPALME, 2009). Table 5 presents a confusion matrix example for a binary classification. With the confusion matrix we computed four classes: (i) *True Positives* (TP), number of correctly classified positive examples; (ii) *True Negatives* (TN), number of correctly classified negatives; (iii) *False Positive* (FP), number of wrongly classified negatives; and (iv) *False Negative* (FN), number of misclassified negatives.

Table 5 – Confusion Matrix

		Labeled	
		Positive	*Negative*
Predicted	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Source: Adapted from Sokolova and Lapalme (2009)

We can obtain different performance metrics from these crude rates evaluating different aspects of a classification model (RUUSKA et al., 2018). In Table 6, we present the main metrics derived from the confusion matrix for a binary classification scenario.

The typical radiological findings of COVID-19 are presented in several diseases. Therefore, following the RSNA recommendation to classify the findings into typical, indeterminate, or atypical for COVID-19, it is necessary to use multi-class architectures for detection. In this sense, we can generalize the confusion matrix so that C_k, an array of size $k \times k$, is defined for each class. For each cell, $[i, j]$ represents the frequency of the real class C_i and the inference class C_j. This process will result in a binary confusion matrix for each of the C_k classes (RUUSKA et al., 2018).

A visual way of evaluating the classifier performance is through the Receiver Operating Characteristic (ROC) curve. The ROC curve is obtained by plotting the sensitivity (true positive rate) on the y-axis and the 1-specificity (false positive rate) on the x-axis for a continuous threshold range (HOO; CANDLISH; TEARE, 2017).

Table 6 – Performance metrics derived from the confusion matrix.

Measure	Formula	Goal
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall effectiveness of a classifier
Precision	$\frac{TP}{TP+FP}$	Fraction of positive instances classified correctly
Recall or Sensitivity	$\frac{TP}{TP+FN}$	Effectiveness of a classifier to identify positive labels
Specificity	$\frac{TN}{TN+FP}$	Effectiveness of a classifier to identify negative labels
F1-score	$\frac{2*TP}{2*TP+FN+FP}$	Harmonic measure between precision and sensitivity

Source: Elaborated by the author.

The main objective of the ROC curve is to find an optimal threshold that optimizes sensitivity and specificity. Therefore, the higher and to the left the cut-off point, the better the performance of the classifier (RUUSKA et al., 2018). Finally, we can obtain the Area Under the ROC Curve (AUC) through the ROC curve, which measures the area below the points obtained by the ROC. The value varies between 0 and 1, and the closer to 1, the better the classifier's ability to distinguish the dataset classes (CARTER et al., 2016).

5.2.2 Image Captioning Metrics

To quantitatively evaluate the performance of the CheXReport network, we used the metrics Bilingual Evaluation Understudy (BLEU) (PAPINENI et al., 2002), Metric for Evaluation of Translation with Explicit Ordering (METEOR) (BANERJEE; LAVIE, 2005), Google BLEU (GLEU) (WU et al., 2016), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (LIN, 2004). Due to the radiological report's characteristics, the models' performance is evaluated using the natural language generation metrics. The BLEU-n metrics quantify the similarity between the generated and ground truth reports using n-grams, which are continuous sequences of n words. Firstly, BLEU calculates the precision between the ground truth $\phi = \{\phi_1, \phi_2, \phi_3, \ldots, \phi_k\}$ and predicted $\hat{\phi} = \hat{\phi_1}, \hat{\phi_2}, \hat{\phi_3}, \ldots, \hat{\phi_k}\}$, with k being the length of the ground truth sentence, and w the size of the expected sentence. The calculation of precision between the n-grams of the sequences is given by:

$$P_n = \sum_k \frac{\min(d_k(\hat{\phi}, \max(d_k(\phi)}{\sum_k d_k(\hat{\phi})} \tag{5.1}$$

where $d_k(\cdot)$ is the number of occurrences of an n-gram, and k is the maximum possible n-grams. To prevent short sentences from having high precision and influencing final performance, BLEU adds a brevity penalty:

$$BrevityPenalty = \begin{cases} 1 & if \quad r < c \\ e^{(1-\frac{r}{c})} & if \quad r \geq c \end{cases} \tag{5.2}$$

where c is the cumulative size of the predicted sentence, and r is the set of ground truth. Therefore, the final BLEU equation is given by:

$$BLEU_n = \left(\min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log(p_n) \right) \tag{5.3}$$

where w_n is a constant value, defined as $\frac{1}{N}$.

METEOR considers multiple facets of fact generation quality, including precision, recall, stemming, and synonymy (BANERJEE; LAVIE, 2005). The METEOR calculation is given by:

$$\begin{aligned} METEOR &= F_{mean}(1-p) \\ F_{mean} &= \frac{10PR}{R+9P} \end{aligned} \tag{5.4}$$

where P is the precision, and R is the recall of the unigrams of the predicted sentence in relation to the ground truth. The value p is a penalty for situations where the predicted and ground truth sentences have all unigrams but do not have the same semantic organization. The penalty p is given by:

$$p = 0.5\left(\frac{c}{u_m}\right)^3 \tag{5.5}$$

where c is the number of chunks in the sentence, and u_m is the number of unigrams matched.

Then, we evaluate the performance of CheXReport in terms of GLEU. GLEU measures the quality of the suggestion prediction for chest X-ray reports regarding fluency and grammaticality of the text (WU et al., 2016). GLEU calculates the precision and recall of unigrams and bigrams by penalizing occurrences that contain specific n-grams not found in the text. GLEU can be calculated as:

$$GLEU = \min(R, P) \tag{5.6}$$

with the recall R being calculated as the ratio between the total m_n of predicted n-grams and the total t_n of ground truth n-grams.

$$Recall = \frac{m_n}{t_n} \tag{5.7}$$

Meanwhile, precision P is the ratio between the quantity m_n of n-grams predicted with the total g_n n-grams produced by the model.

$$Precision = \frac{m_n}{g_n} \tag{5.8}$$

In addition, we utilized the ROUGE metric to assess the quality of the generated reports, particularly focusing on the recall aspect of the n-grams, which is critical for capturing the completeness of the generated text in comparison to the ground truth. ROUGE primarily evaluates the overlap of n-grams between the generated and reference texts, thus providing a measure of how much of the relevant content is retained in the generated report (LIN, 2004). The ROUGE-N score is defined as:

$$ROUGE - N = \frac{\sum_{S \in \text{Reference Texts}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \text{Reference Texts}} \sum_{gram_n \in S} Count(gram_n)} \tag{5.9}$$

where $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the generated and reference texts, and $Count(gram_n)$ is the total number of n-grams in the reference text (LIN, 2004). ROUGE, therefore, complements the precision-oriented BLEU metric by ensuring that the generated reports cover all relevant aspects of the ground truth as completely as possible.

5.2.3 Survival Analysis Metrics

Risk prediction for each event monitored in survival analysis models helps healthcare professionals develop personalized treatments for patients (PENCINA; D'AGOSTINO, 2015). In this sense, the Concordance index (C-index) is the most used metric to evaluate the performance of survival analysis models (PARK et al., 2021). The C-index is obtained by the proportion of correct pairs ordered about the event time. The C-index ranges from 0.5 to 1.0, with 0.5 indicating the randomness of the model and 1.0 perfect discrimination. Mathematically, the C-index is obtained by:

$$C - index = \frac{\sum_{i,j} I(h_i > h_j I(t_i < t_j) \varphi_i}{\sum_{i,j} I(t_i < t_j)} \tag{5.10}$$

where, h_i and h_j are the risk values predicted by the model for observations i and j, t_i and t_j are the observed survival times. I is an indicator function, which returns 1 if the condition is true and 0 otherwise. φ_i represents whether the event occurred (1) or was censored (0).

While the C-index is a metric that measures the ability to discriminate between different time points of events, the Brier score can measure the global performance of

the model (PARK et al., 2021). The Brier score measures the mean squared error between probabilistic predictions and expected results. Mathematically, the Brier score is given by:

$$Brierscore = \frac{1}{N}\sum_{n=1}^{N}(p_i, o_i)^2 \tag{5.11}$$

where N is the total number of observations, p_i is the predicted event probability for observation i, and o_i is the observed value (0 or 1) for observation i. The Brier score ranges from 0 to 1, with 0 indicating perfect prediction and 1 indicating random behavior of the model.

5.3 MultSurv Model Study Design

We adopted an ablation study to validate the MultSurv model proposed in Chapter 4. To this end, we trained a series of additional architectures to use throughout model validation. In Table 7, we present the characteristics of each model and architectures used. The ablation process followed an incremental strategy. Firstly, we assessed whether the adoption of embeddings (Model B) for categorical and continuous variables represented a performance gain about the baseline (Model A). Next, we evaluate the best architecture for extracting features from X-ray images. To this end, we evaluated a convolutional classification architecture (Model C) and a chest X-ray report suggestion architecture as a feature extractor (MultSurv model). In the following subsections, we present the characteristics used to train the models.

Table 7 – Main characteristics of the evaluated models.

Model	Embeddings	Image	Temporal	Multitask	Datasets	References
Model A	No	No	Yes	Yes	PBC2 and MHD	(LEE; YOON; SCHAAR, 2019), (MURTAUGH et al., 1994)
Model B	Yes	No	Yes	Yes	PBC2 and MHD	(MURTAUGH et al., 1994)
Model C	Yes	Yes	Yes	Yes	Curated Dataset for COVID-19 and MHD	(SAIT et al., 2020)
MultSurv model	Yes	Yes	Yes	Yes	MIMIC-CXR and MHD	(JOHNSON et al., 2019)

5.3.1 Model A

The choice of DynamicDeepHit (LEE; YOON; SCHAAR, 2019) as a baseline for validation is based on the architecture's ability to deal with censored data and multiple types of events. DynamicDeepHit models the distribution of time until the

event of interest. Through a RNN, the architecture is capable of capturing temporal dependencies of data and aggregating it with patients' current clinical information. Furthermore, the model adopts a missing data handling mechanism. We train the DynamicDeepHit model with two datasets. The first is PBC2 to compare the results from the DynamicDeepHit with the MultSurv model in a public dataset. The second is to assess the performance of DynamicDeepHit with tabular data from MDH dataset. Both models were optimized using a Random Search strategy, following the methodology proposed by the authors (LEE; YOON; SCHAAR, 2019). In Table 8, we present the range of values of the DynamicDeepHit training hyperparameters.

Table 8 – Network hyperparameter search space for Model A.

Hyperparameter	**Value Space**
Batch size	32
Dense layers activation	ReLU, eLU, tanh
RNN activation	ReLU, eLU, tanh
Dense output activation	ReLU, eLU, tanh
Dense dropout	0.2, 0.4, 0.6
RNN dropout	0.2, 0.4, 0.6
RNN cell	GRU, LSTM
RNN hidden size	25, 50, 100, 150, 200
Dense hidden size	25, 50, 100, 150, 200
Dense number layers	1, 2, 3
RNN dense number layers	1, 2, 3
Attention number layers	2, 4, 6, 8, 10
Learning rate	1e-3, 1e-4, 1e-5

Source: Elaborated by the author.

5.3.2 Model B

The model B is composed only of the part of MultSurv model responsible for processing tabular data. This way, we can compare the performance of the models fairly since they were trained on the same dataset. The model B hyperparameter optimization process was carried out using a Random Search. In Table 9, we present the ranges of values used to optimize model B.

5.3.3 Model C

For model C, we evaluated whether adding a convolutional architecture for classifying typical COVID-19 findings would improve the model's performance. In

Table 9 – Network hyperparameter search space for Model B.

Hyperparameter	**Value Space**
Batch size	32
Dense layers activation	ReLU, eLU, tanh
RNN activation	ReLU, eLU, tanh
Dense output activation	ReLU, eLU, tanh
Dense dropout	0.2, 0.4, 0.6
RNN dropout	0.2, 0.4, 0.6
RNN cell	GRU, LSTM
RNN hidden size	25, 50, 100, 150, 200
Dense hidden size	25, 50, 100, 150, 200
Dense number layers	1, 2, 3
RNN dense number layers	1, 2, 3
Attention number layers	2, 4, 6, 8, 10
Learning rate	1e-3, 1e-4, 1e-5
Embeddings Dropout	0.2, 0.4, 0.6
Embeddings size	8, 16, 32, 64

Source: Elaborated by the author.

this sense, we pre-train six convolutional architectures on the Curated Dataset for COVID-19. To train these architectures, we perform pre-processing and data augmentation of the images. All images have been resized to a size of 224×224 pixels. To avoid distortions, we adopted the methodology defined in Section 4.2.1. In this way, a proportional reduction is applied to each axe, and zero padding is added to the smallest dimension. Additionally, we apply CLAHE contrast normalization. For data augmentation, we use horizontal flips, rotations of up to 20 degrees, and shear for the training set.

We use K-fold cross-validation as a method to evaluate the performance of the architectures, with a K=10. Next, we train each of the six convolutional architectures according to the hyperparameters defined in Table 10, with pre-trained weights for the ImageNet dataset. To measure the error of each architecture, we use categorical cross-entropy. We optimize the weights of the architectures with the Adam algorithm.

Then, we aggregate the features of the last dense layer of the best convolutional architecture with the temporal characteristics and embeddings of the tabular data. This set of characteristics serves as input to the MultSurv model's causal networks. Model B training is carried out for 100 epochs using MDH data. To optimize the model, the losses L_1 and L_2 defined in Section 4.2.6 were adopted, and Adam was optimized with a learning rate of $1e^{-4}$.

Table 10 – Parameters used for each of the CNN architectures.

Architecture	Learning Rate	Batch Size	Trainable Params	Non-trainable Params	Depth
DenseNet121 (HUANG et al., 2017)	5×10^{-7}	16	11 149 444	83 648	121
InceptionResNetV2 (SZEGEDY et al., 2017)	5×10^{-7}	4	57 816 420	60 544	572
InceptionV3 (SZEGEDY et al., 2016)	5×10^{-7}	4	26 488 228	34 432	159
MovileNetV2 (SANDLER et al., 2018)	5×10^{-7}	4	7 468 036	34 112	88
ResNet50V2 (HE et al., 2016)	1×10^{-6}	16	31 909 252	45 440	50
VGG16 (SIMONYAN; ZISSERMAN, 2014)	5×10^{-7}	16	48 231 684	38 720	23

Source: Elaborated by the author.

5.3.4 MultSurv Model

Finally, we evaluate MultSurv model with the complete pipeline. In this sense, we trained the MultSurv model architecture using the entire MDH dataset. The hyperparameters for the MultSurv model component that deals with tabular data followed those obtained for model B. For the CheXReport network, we carried out an architecture evaluation process using the MIMIC-CXR dataset. We evaluate pre-trained Swin-T, Swin-S, and Swin-B encoder architectures for the ImageNet dataset. For the encoder, we use a fixed value of 2 as the number of layers N_d. We set the CheXReport set of visual embeddings to 256 for all encoder architectures. BERT multilingual word embeddings were frozen for 10 epochs to ensure that the X-Ray image encoder had already been minimally tuned to capture domain-relevant features. We use the search beam of size five to generate the report suggestions during inference. The generation will stop when the $< eos >$ token is generated, or the sentence reaches the token limit.

Furthermore, to properly evaluate the capability of CheXReport, we compared the performance of traditional feature extraction architectures for image captioning. We compared the ResNet50-v2 and ResNet101-v2 architectures using two different decoder models. The first is composed of the fusion of visual features with BERT textual embeddings. The second decoder is composed of an architecture identical to the CheXReport decoder. These architectures were pre-trained for 100 epochs with a learning rate of $1e^{-4}$ using an Adam optimizer. We reduced the learning rate by 25% every 10 epochs with no improvement for BLEU-4 on the validation set. The batch size was set to 64 images.

We used the Python language, OpenCV, and PyTorch libraries to conduct the experiments. We perform the experiments described in this chapter on a system that contains a 24GB Quadro RTX 6000, Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz,

and 64 GB of RAM.

5.3.5 Baselines

Finally, we evaluate the final performance of the MultSurv model compared with the established architectures in the current literature: CoxTime, CoxCC, DeepSurv, PCHazard, DeepHit, and N-MTLR. CoxTime extends the CoxPH model by incorporating neural networks, allowing it to capture non-linear effects over time (KVAMME; BORGAN; SCHEEL, 2019). On the other hand, CoxCC combines case-control sampling with neural networks to increase both computational efficiency and predictive accuracy (KVAMME; BORGAN; SCHEEL, 2019). DeepSurv uses deep neural networks based on the CoxPH model and is specifically designed to predict the time to occurrence of events (KATZMAN et al., 2018). PCHazard takes a different approach, dividing time into intervals and estimating the hazard function piecewise through neural networks, which facilitates the capture of temporal variation in risk (KVAMME; BORGAN, 2021). Furthermore, DeepHit uses a multi-class approach to directly predict the time-to-occurrence distribution of multiple events (LEE et al., 2018). Finally, N-MTLR employs neural network-based multi-task logistic regression to model the survival function, addressing both temporal dependence and complex interactions between covariates (FOTSO, 2018).

5.4 Final Remarks

This chapter presented the methodology used to evaluate the MultSurv model, detailing the technical characteristics, datasets, and evaluation metrics applied. We describe the specific datasets used. The evaluation metrics section detailed the performance metrics employed to evaluate the model's effectiveness, including classification, image captioning, and survival analysis metrics. These metrics ensure a complete assessment of the MultSurv model's capabilities. Furthermore, we detail the MultSurv model configuration and the incremental ablation study to validate the model performance. In the next chapter, we present the results obtained from the experiments, providing an analysis of MultSurv model's performance.

6 RESULTS AND DISCUSSIONS

This chapter explores the MultSurv model performance, analyzing its quantitative metrics and qualitative observations. We start with a detailed description of the MDH dataset analysis and the specific implementation details of the MultSurv model. In the sequence, we present an ablation study analyzing the impact of different architectural components on the MultSurv model performance. We then present an analysis and comparison with state-of-the-art models. Finally, we provide a qualitative analysis, offering visual and textual comparisons that underscore the MultSurv model capabilities. This comprehensive exploration aims to validate the MultSurv model effectiveness for survival analysis.

6.1 MyDigitalHealth Dataset Analysis

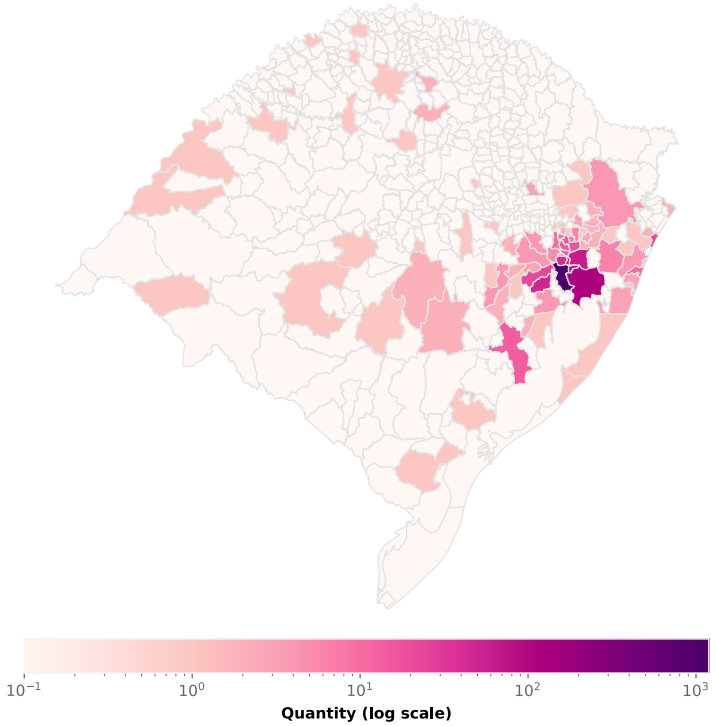
Between March 20, 2020, and June 02, 2022, 1,891 cases, including 1,815 unique patients, were collected from the HCPA. All patients were treated by the SUS and had a positive RT-qPCR test at the time of hospitalization. During hospitalization, 1,266 were admitted to the ICU. Of the 1,891 cases, 36.57% died, and 13.27% were censored. By censoring, we consider patients who abandoned treatment or were transferred to other hospitals.

The main origin of the cases is the Brazilian state of RS (1,884 cases). The other cases are from Santa Catarina (SC) (3 cases), São Paulo (SP), Rio de Janeiro (RJ), Rondônia (RO), and Amazonas (AM) (with 1 case each). Concerning the city of origin, most cases are from Porto Alegre (1,193 cases). This behavior is a characteristic of the healthcare regionalization in Brazil and the HCPA's reference for the RS capital. In Figure 15, we present a stratification by patients' city of origin in RS for the cases collected at HCPA.

Table 11 presents the characterization of our data. The mean population age was 59.15 years (Interquartile Range (IQR) 48.0 - 71.0). Cases were distributed among seven age groups, with a slight predominance of cases in the age group 60 to 69 years (483 (25.54%)). In all groups, the number of males hospitalized by COVID-19 was higher than females. Male patients represented 51.51% of the patients, with 276 deaths and 124 censored cases. In-hospital mortality rates were slightly higher for males than for females during the whole pandemic (28.34% vs. 27.58% of deaths, respectively).

There was a predominance of hospitalized White individuals, representing 82.21% of the patients, followed by the Black population (14.07%). The mortality rate was higher among Asian people (45.46%) and decreased inversely to the educational level. Furthermore, the mortality rate is 15% higher among illiterate people

Figure 15 – Patients stratified by city of origin in the state of Rio Grande do Sul.



Source: Elaborated by the author.

compared to those with a college degree. Most cases were admitted at the Very Urgent level (46.06%). The highest mortality rates were observed for patients admitted as an emergency (45.00%).

Regarding laboratory tests, were made available to only 776 patients. These make up 15,289 different collection records. A collection may contain the results of multiple laboratory tests. The tests with the highest collection percentages are routine blood tests, such as potassium, sodium, serum creatinine, urea, Chronic Kidney Disease Epidemiology Collaboration (CKDEPI), and Modification of Diet in Renal Disease (MDRD). In Table 12, we present a stratification of the percentage of results completed for each laboratory test.

In addition, 1,066 chest X-ray images were collected from 677 patients during hospitalization. Two Carestream DRX-1 (202 images) and DRXPLUS3543C (561 images) detectors were used to collect the chest X-ray exams. For 303 images, there was no detector record in the DICOM metadata. Regarding the collection position, 987 images were collected in the AP position, 66 in the PA position, and 13 in the lateral position. In Figure 16, we present a size distribution of the images.

Regarding the variability of findings, as chest X-ray reports are composed of free text, we used a word cloud to summarize the findings identified by radiologists. In Figure 17, we present the word cloud resulting from the HCPA X-ray reports. It is possible to observe that there are mentions of findings of pulmonary, cardiac, digestive,

Table 11 – Patient characteristics stratified by outcome.

	General	Discharged	Death	Censored
Age, in years				
Min	18	18	21	20
Max	102	95	102	94
Mean	59.15	56.22	65.00	59.94
Median	61	57	67	61
Standard deviation	15.94	15.65	15.12	15.45
Interquartile range	48.0 - 71.0	44.0 - 67.0	57.0 - 75.0	50.0 - 70.0
Age groups, in years				
18 to 29	74 (3.91%)	50 (4.46%)	15 (2.84%)	9 (3.72%)
30 to 39	181 (9.57%)	138 (12.32%)	27 (5.10%)	16 (6.61%)
40 to 49	261 (13.80%)	194 (17.32%)	35 (6.62%)	32 (13.22%)
50 to 59	373 (19.73%)	232 (20.71%)	85 (16.07%)	56 (23.14%)
60 to 69	483 (25.54%)	275 (24.55%)	141 (26.65%)	67 (27.69%)
70 to 79	330 (17.45%)	154 (13.75%)	137 (25.90%)	39 (16.12%)
80 <	189 (9.99%)	77 (6.88%)	89 (16.82%)	23 (9.50%)
General	1 891	1 120 (59.23%)	529 (27.97%)	242 (12.80%)
Sex				
Female	917 (48.49%)	546 (48.75%)	253 (47.83%)	118 (48.76%)
Male	974 (51.51%)	574 (51.25%)	276 (52.17%)	124 (51.24%)
Self-reported race				
Asian	11 (0.58%)	4 (0.36%)	5 (0.95%)	2 (0.83%)
White	1 548 (82.21%)	910 (81.61%)	432 (81.82%)	206 (85.83%)
Brown	59 (3.13%)	38 (3.41%)	15 (2.84%)	6 (2.50%)
Black	265 (14.07%)	163 (14.62%)	76 (14.39%)	26 (10.83%)
Scholarity				
Illiterate	61 (3.23%)	24 (2.14%)	24 (4.54%)	13 (5.37%)
Elementary School	580 (30.67%)	331 (29.55%)	186 (35.16%)	63 (26.03%)
Middle School	407 (21.52%)	236 (21.07%)	115 (21.74%)	56 (23.14%)
High School	497 (26.28%)	341 (30.45%)	101 (19.09%)	55 (22.73%)
College / University	117 (6.19%)	75 (6.70%)	29 (5.48%)	13 (5.37%)
Without information	229 (12.11%)	113 (10.09%)	74 (13.99%)	42 (17.36%)
Condition at admission				
Emergency	20 (1.06%)	10 (0.89%)	9 (1.70%)	1 (0.41%)
Very urgent	871 (46.06%)	589 (52.59%)	182 (34.40%)	100 (41.32%)
Urgent	124 (6.56%)	96 (8.57%)	20 (3.78%)	8 (3.31%)
Less urgent	7 (0.37%)	6 (0.54%)	0 (0.00%)	1 (0.41%)
Not classified	869 (45.95%)	419 (37.41%)	318 (60.11%)	132 (54.55%)

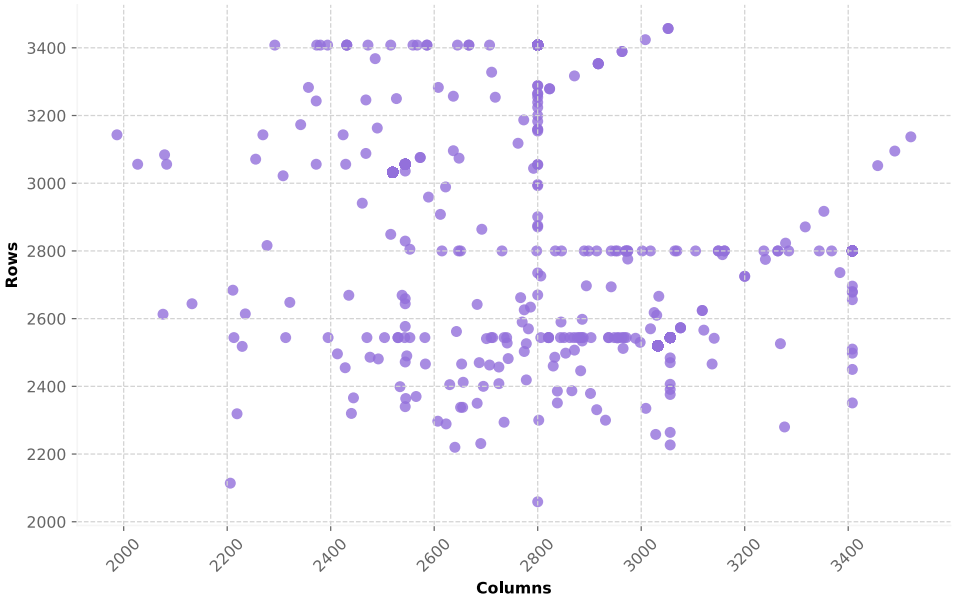
Source: Elaborated by the author.

Table 12 – Percentage of patients exams results available in the HCPA sample.

Exam	%	Exam	%	Exam	%
Potassium	25.51	Sodium	25.84	Serum creatinine	30.37
Urea	30.46	CKD-EPI	30.52	MDRD	30.54
Magnesium	57.96	Leukocytes	68.98	Monocytes %	68.98
Absolute monocytes	68.98	Absolute segmented neutrophils	68.98	Segmented neutrophils %	68.98
Hematocrit	68.98	Hemoglobin	68.98	Absolute lymphocytes	68.98
Lymphocytes %	68.98	Absolute eosinophils	68.99	Absolute basophils	68.99
Basophils %	68.99	Eosinophils %	68.99	MCV Failace	68.99
MCH	68.99	MCHC	68.99	Erythrocytes	68.99
RDW	69.00	Erythroblasts	69.02	C-reactive protein result	72.08
H obs1 blood count	80.70	APTT control	81.39	APTT seconds	81.39
Blood count observation neutrophil bands	83.12	Calcium VR	88.42	Corrected calcium	88.44
Absolute band neutrophils	88.59	Band neutrophils %	88.59	Absolute myelocytes	89.52
Myelocytes %	89.52	H obs2 blood count	89.68	PT control	91.75
PT INR	91.75	PT seconds	91.75	PT activity	91.75
H D-dimers	93.03	Serum chloride	93.09	Direct bilirubin	93.37
Indirect bili result	93.37	Total bilirubin	93.39	CK	93.43
Metamyelocytes %	93.81	Absolute metamyelocytes	93.81	GPT	94.38
GOT/AST result	94.42	Plasma lactate	94.72	H obs3 blood count	95.08
Troponin-T	95.70	LDH VR	96.08	STA compact fibrinogen	97.32
H obs4 blood count	97.90	Albumin	98.16	Observation	98.26
Estimated average glucose	98.79	A1C	98.79	Plasma cells %	98.86
Absolute plasma cells	98.86	Ferri result	99.18	H obs5 blood count	99.29
Absolute promyelocytes	99.30	Promyelocytes %	99.30	Minor/major indicator	99.37
Triglycerides	99.44	Sample creatinine	99.48	Absolute reticulocytes	99.57
Reticulocytes	99.57	Urine sample sodium	99.62	H obs6 blood count	99.67
Biochemical observations	99.67	E170 signal	99.69	Indicator seconds	99.74
Obs	99.76	Sample urea	99.84	CKD-EPI alpha	99.86
Urine sample potassium	99.88	C3 result	99.89	C4 result	99.89
Rheumatoid factor result	99.91	Result	99.93	CD3 value	99.95
H IF CD45/UL	99.95	H CD8 %	99.95	H CD4 %	99.95
CD4/CD8 ratio	99.95	CD4 value	99.95	CD8 value	99.95
H CD3 %	99.95	CSF lactate	99.95	Urine sample chloride	99.95
IgG result	99.95	Mean fluorescence index (MFI)	99.97	Thrombin time numeric	99.97
H obs7 blood count	99.97	IgM result	99.97	CSF ADA support	99.97
Urine volume	99.98	CSF LDH	99.98	Iron	99.99
24h creatinine	99.99	Urine alpha calcium	99.99	Magnesium result	99.99
24h calcium	99.99	ADA support	99.99	Urine alpha creatinine	99.99
CEA result	99.99	Ascites albumin	99.99	Absolute blasts	99.99
Blasts %	99.99	Urine sample calcium	99.99	Urinary urea	99.99
APTT obs	99.99	Activity indicator	99.99	INR indicator	99.99
Total ascites bili 1	99.99	CD4 observation	99.99	Total potassium volume	99.99
Urine alpha sodium	99.99	Serous fluid ADA	99.99	Thrombin time signal	99.99
Biochemical results outside measurement range	99.99	Urine alpha urea	99.99	Urine alpha potassium	99.99
24h urine sodium	99.99	Calcitonin result	100.00	Antithrombin result	100.00
FO urea	100.00				

Source: Elaborated by the author.

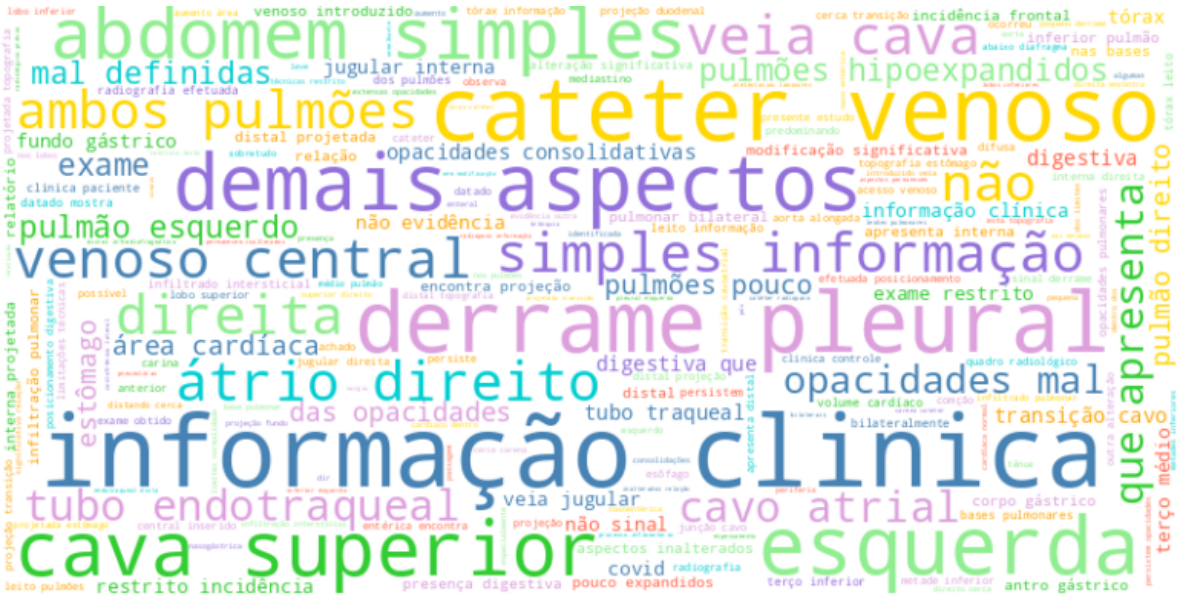
Figure 16 – Size distribution of patients' chest X-ray images.



Source: Elaborated by the author.

and support device changes.

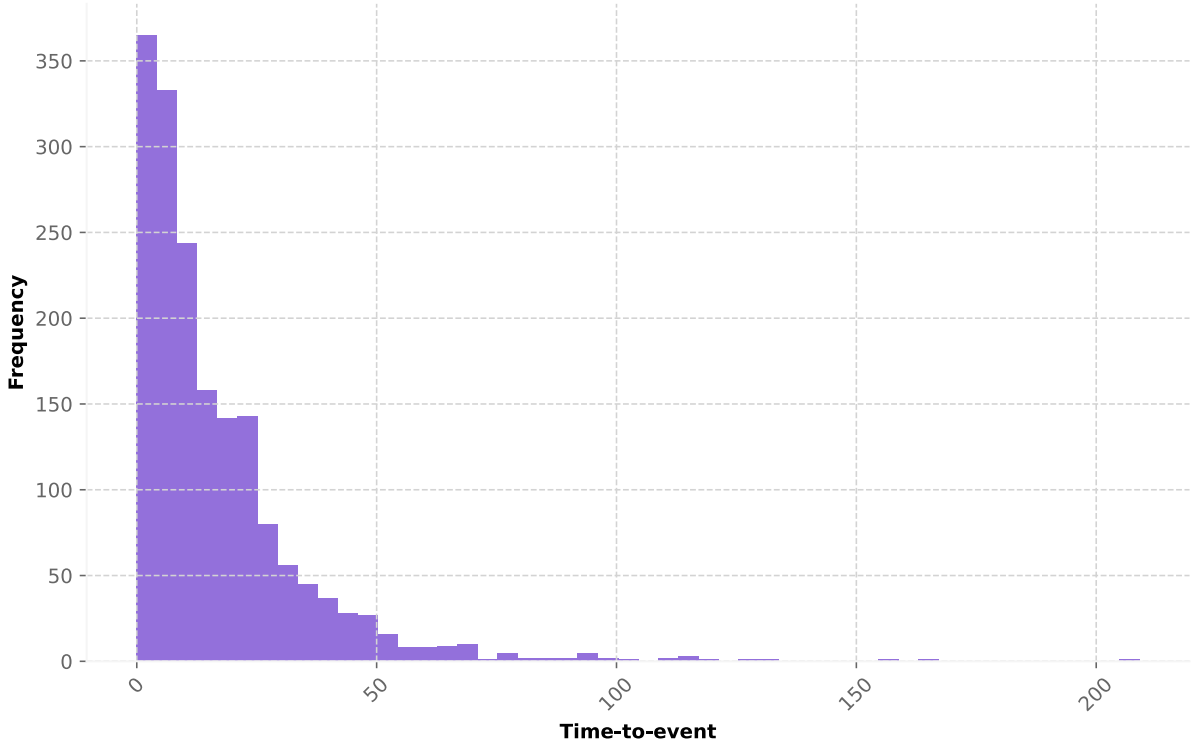
Figure 17 – Most frequent words in the chest X-ray reports.



Source: Elaborated by the author.

We temporally align all patients based on their hospital admission date. Therefore, time instant 0 for all patients was the date of hospitalization. Throughout hospitalization, patients experienced events at different moments in time. The minimum time until the event was 0 days, and the maximum was 209 days (±18.46 IQR [5.0 – 23.0]). Analyzing the frequency of patients' time-to-event, we observed that there are few samples with time-to-event greater than 50 days (Figure 18).

Figure 18 – Time-to-event histogram of patients in the dataset.



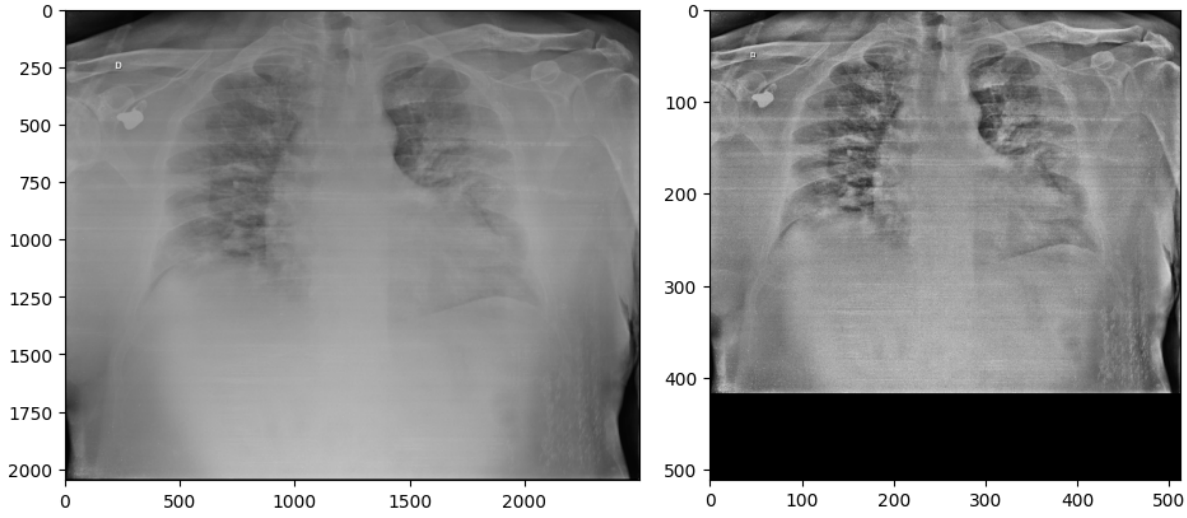
Source: Elaborated by the author.

6.2 Data Preprocessing

We consider the set of possible survivals up to 50 days, with intervals of 1 day. Therefore, T_{max} = 50, based on this filter, we considered 1,655 unique patients with a time-to-event of less than 50 days. We use stratified K-fold cross-validation to evaluate the models, with K=10. With a 10-fold, we train the models 10 times each. At the end of the training, we calculated the mean and standard deviation of the results for the defined metrics. For sociodemographic, clinical, and laboratory data, we only considered data with more than 50% completeness for the final dataset. Columns with continuous data were normalized using the standardization technique. The textual categorical columns were transformed into numerical categories. Finally, we generate the mask for the missing data samples.

We preprocess the X-ray images to enhance the architecture's training (Figure 19). Firstly, we apply CLAHE (ZUIDERVELD, 1994) to highlight the anatomopathological structures projected on chest X-ray images. Then, given the different dimensions of the images, we resized the images to 224×224 pixels. The reduction was proportional to the image width and height, with zero padding for the smallest axis. The use of the pre-trained encoder limited the image resolution. For the MIMIC-CXR and the Curated Dataset for COVID-19, we used the original dataset split suggestion for training and testing.

Figure 19 – (a) Original chest X-ray image. We can see that the dimensions of the image are not proportional. Therefore, in (b), a reduction proportional to the width and height of the image is applied, with zero padding for the shortest axis. Furthermore, in (b), we sought to highlight differences in image contrast with the application of CLAHE.



(a) Original X-ray (b) Preprocessed X-ray

Source: Elaborated by the author.

6.3 Ablation Study

We evaluate each part of the MultSurv model for its purpose and, in general, for its survival analysis capacity. We used temporal prediction points for survival analysis of 1, 3, 5, and 7 days. We evaluated the model's performance over these same time intervals, using the C-index and the Brier Score metrics to measure the discrimination and calibration of predictions, respectively. These time periods were chosen to capture the short-term dynamics of the event of interest, allowing detailed analysis of the model's effectiveness in predicting survival at different temporal stages.

6.3.1 Model A

First, we evaluate the Model A performance based on DynamicDeepHit on the PBC2 public dataset. PBC2 is the databaset available from the DynamicDeepHit repository. DynamicDeepHit is a dynamic survival analysis model that deals with temporal data and simultaneous risk events. We implemented the DynamicDeepHit

following strictly the original implementation[1].

In Tables 13 and 14, we present the performance of Model A in terms of C-index and Brier score at different time points of prediction (t) and evaluation (Δt). We observed that the performance of the C-index varies depending on the prediction and evaluation time. For $t = 52$, the model achieves its best performance at $\Delta t = 12$ with a C-index of 0.975 ± 0.01, indicating the ability to discriminate short-term risks. However, there is a drop at $\Delta t = 36$, followed by a recovery at the $\Delta t = 60$ and $\Delta t = 120$ intervals. For $t = 156$, the initial performance is low at $\Delta t = 12$ but progressively improves, reaching the highest value at $\Delta t = 120$ (0.868 ± 0.04). For $t = 260$, the model shows stable and robust performance across all Δt, with C-index ranging from 0.822 ± 0.03 to 0.842 ± 0.03, reflecting consistent reliability for long-term prediction terms.

Table 13 – Comparison of Model A performance for different prediction and evaluation time points for the C-index (mean and ± standard deviation) in the PBC2 dataset. The bigger, the better.

Prediction Time	$\Delta t = 12$	$\Delta t = 36$	$\Delta t = 60$	$\Delta t = 120$
$t = 52$	0.975 ± 0.01	0.803 ± 0.00	0.910 ± 0.03	0.915 ± 0.03
$t = 156$	0.518 ± 0.00	0.577 ± 0.05	0.639 ± 0.01	0.868 ± 0.04
$t = 260$	0.842 ± 0.03	0.822 ± 0.03	0.825 ± 0.00	0.840 ± 0.01

Source: Elaborated by the author.

Regarding the Brier score (Table 14), which measures the model's calibration, the results indicate that Model A presents a good experience for short predictions, especially at $t = 52$ and $\Delta t = 12$, with a Brier score of 0.016 ± 0.00. However, the score increases as the evaluation interval extends, reaching 0.077 ± 0.00 at $\Delta t = 120$. For $t = 156$, the Brier score values range from 0.047 ± 0.00 at $\Delta t = 12$ to 0.90 ± 0.00 at $\Delta t = 120$. At $t = 260$, performance is relatively constant across all Δt, with scores ranging from 0.089 ± 0.00 to 0.115 ± 0.00, establishing an acceptable clause for long-term predictions.

Next, we evaluate the performance of Model A for the MDH tabular dataset. To ensure the model was optimized for our specific dataset, we performed hyperparameter optimization through RandomSearch over 100 epochs. RandomSearch is an efficient technique for exploring the hyperparameter space, helping to find the combination that maximizes model performance (BERGSTRA; BENGIO, 2012). Table 15 presents the best set of hyperparameters identified for DynamicDeepHit.

This evaluation aims to verify how DynamicDeepHit, with its ability to model the temporal evolution of health data and consider multiple risks simultaneously,

[1] https://github.com/chl8856/Dynamic-DeepHit

Table 14 – Comparison of Model A performance for different prediction and evaluation time points for the Brier score (mean and ± standard deviation) in the PBC2 dataset. The smaller, the better.

Prediction Time	$\Delta t = 12$	$\Delta t = 36$	$\Delta t = 60$	$\Delta t = 120$
$t = 52$	0.016 ± 0.00	0.021 ± 0.00	0.030 ± 0.00	0.077 ± 0.00
$t = 156$	0.047 ± 0.00	0.059 ± 0.00	0.068 ± 0.00	0.90 ± 0.00
$t = 260$	0.089 ± 0.00	0.089 ± 0.00	0.093 ± 0.00	0.115 ± 0.00

Source: Elaborated by the author.

Table 15 – Model A network hyperparameters.

Hyperparameter	**Value**
Batch size	32
Dense layers activation	ReLU
RNN activation	Tanh
Dense output activation	ReLU
Dense dropout	0.2
RNN dropout	0.2
RNN cell	GRU
RNN hidden size	100
Dense hidden size	100
Dense number layers	2
RNN dense number layers	2
Attention number layers	8
Learning rate	1e-3

Source: Elaborated by the author.

behaves in a realistic scenario, such as the one represented by the MDH dataset. This analysis allows us to compare the effectiveness of DynamicDeepHit with the MultSurv model, highlighting the potential improvements provided by the new techniques and approaches incorporated in our proposed model. In this sense, the performance of Model A was evaluated in two main aspects: C-index and Brier score, at different time points of prediction (t) and evaluation (Δt). The following Tables 16 e 17 presents the average results and the respective standard deviations for these metrics.

Table 16 – Comparison of Model A performance for different prediction and evaluation time points for the C-index (mean and ± standard deviation) for the MDH dataset. The bigger, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.666 ± 0.02	0.622 ± 0.02	0.629 ± 0.01	0.648 ± 0.01
$t = 3$	0.584 ± 0.04	0.635 ± 0.03	0.639 ± 0.02	0.633 ± 0.02
$t = 5$	0.605 ± 0.02	0.617 ± 0.03	0.614 ± 0.02	0.611 ± 0.02
$t = 7$	0.584 ± 0.03	0.613 ± 0.02	0.611 ± 0.02	0.618 ± 0.02

Source: Elaborated by the author.

Analyzing the results for the C-index, we can observe that the model's performance varies depending on the time points of prediction and evaluation. The best performance is observed for $t = 1$ and $\Delta t = 1$, with a C-index of 0.666 ± 0.02. As we increase Δt, performance decreases but remains relatively stable. At $t = 3$, the model achieves its best performance at $\Delta t = 3$ with a C-index of 0.635 ± 0.03. For $t = 5$ and $t = 7$, the model presents a more uniform performance between the different Δt.

Table 17 – Comparison of Model A performance for different prediction and evaluation time points for the Brier score (mean and ± standard deviation) for the MDH dataset. The smaller, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.060 ± 0.00	0.092 ± 0.00	0.130 ± 0.00	0.163 ± 0.00
$t = 3$	0.097 ± 0.00	0.136 ± 0.00	0.168 ± 0.00	0.186 ± 0.00
$t = 5$	0.143 ± 0.00	0.174 ± 0.00	0.192 ± 0.00	0.195 ± 0.00
$t = 7$	0.185 ± 0.00	0.202 ± 0.00	0.202 ± 0.00	0.212 ± 0.00

Source: Elaborated by the author.

The results for the Brier score indicate that Model A presents better performance for predictions closer to the initial instant (t=1), with a Brier score of 0.060 ± 0.00 for $\Delta t = 1$. As the prediction time (t) increases, the Brier score values increase, indicating a worsening model calibration. This is especially evident in $t = 7$ and $\Delta t = 7$, where the Brier score reaches 0.212 ± 0.00.

The Tables 16 and 17 results demonstrate that Model A performs satisfactorily in terms of discrimination (C-index) and calibration (Brier score) in short periods of prediction and evaluation. Performance decreases as the prediction and evaluation intervals increase, especially for the Brier score. This suggests the model may be more reliable for short-term predictions.

6.3.2 Model B

Model B is composed only with the MultSurv model components responsible for processing tabular data. In this sense, Model B evaluates the impact of categorical and continuous embeddings. Hyperparameters (Table 18) were selected using RandomSearch (BERGSTRA; BENGIO, 2012).

Table 18 – Model B network hyperparameters.

Hyperparameter	**Value**
Batch size	32
Dense layers activation	ReLU
RNN activation	Tanh
Dense output activation	ReLU
Dense dropout	0.2
RNN dropout	0.2
RNN cell	GRU
RNN hidden size	50
Dense hidden size	50
Dense number layers	4
RNN dense number layers	2
Attention number layers	10
Learning rate	1e-4
Embeddings Dropout	0.4
Embeddings size	64

Source: Elaborated by the author.

We first evaluate Model B for the PBC2 dataset. When we compare the Model A results for the PBC2 dataset (Table 13) with the results obtained with Model B (Table 19) we realize that model B achieves better predictions in short-term, with a C-index of 0.978 ± 0.02 at $t = 52$ and $\Delta t = 12$, surpassing Model A (0.975 ± 0.01). This superiority persists at $\Delta t = 36$, where Model B achieves 0.934 ± 0.02, compared to 0.803 ± 0.00 for Model A. For medium-term predictions ($t = 156$), Model B continues to demonstrate better performance, with a higher C-index across all Δt evaluated. At $t = 260$, both models exhibit a more balanced performance, but Model B maintains a

slight advantage, suggesting more consistent discrimination over time.

Table 19 – Comparison of Model B performance for different prediction and evaluation time points for the C-index (mean and ± standard deviation) in the PBC2 dataset. The bigger, the better.

Prediction Time	$\Delta t = 12$	$\Delta t = 36$	$\Delta t = 60$	$\Delta t = 120$
$t = 52$	0.978 ± 0.02	0.934 ± 0.02	0.886 ± 0.02	0.884 ± 0.02
$t = 156$	0.705 ± 0.01	0.831 ± 0.02	0.885 ± 0.03	0.913 ± 0.00
$t = 260$	0.842 ± 0.01	0.866 ± 0.04	0.847 ± 0.01	0.877 ± 0.03

Source: Elaborated by the author.

Regarding calibration, measured by the Brier score, Model B (Table 20) also presents superior results at various time points. For $t = 52$, the Brier score of Model B at $\Delta t = 12$ is 0.007 ± 0.00, indicating better calibration compared to Model A (0.016 ± 0.00). This advantage remains at $\Delta t = 36$ and $\Delta t = 60$, with Model B showing lower scores and better calibration. In the long run, specifically at $t = 260$, Model B exhibits a slightly better calibration, with scores ranging from 0.081 ± 0.00 to 0.98 ± 0.00, while Model A varies from 0.089 ± 0.00 to 0.115 ± 0.00. These results suggest that Model B offers superior discrimination and a more stable and accurate calibration.

Table 20 – Comparison of Model B performance for different prediction and evaluation time points for the Brier score (mean and ± standard deviation) in the PBC2 dataset. The smaller, the better.

Prediction Time	$\Delta t = 12$	$\Delta t = 36$	$\Delta t = 60$	$\Delta t = 120$
$t = 52$	0.007 ± 0.00	0.011 ± 0.00	0.018 ± 0.00	0.040 ± 0.00
$t = 156$	0.040 ± 0.00	0.038 ± 0.00	0.051 ± 0.00	0.070 ± 0.00
$t = 260$	0.081 ± 0.00	0.085 ± 0.00	0.081 ± 0.00	0.098 ± 0.00

Source: Elaborated by the author.

Next, in Tables 21 e 22, we present the results of model B for the MDH test set. The results for the C-index show that Model B presents superior performance compared to Model A, especially for higher prediction times. For $t = 1$, Model B achieves a C-index of 0.693 ± 0.02 at $\Delta t = 1$, which is slightly lower than Model A (0.666 ± 0.02). Besides, as the prediction time increases, the performance of Model B improves considerably. For $t = 3$, the C-index of Model B is 0.740 ± 0.01 at $\Delta t = 1$, compared to 0.584 ± 0.04 of Model A, demonstrating an advantage in discriminating future events. Furthermore, Model B maintains more consistent performance at longer prediction times, such as $t = 5$ and $t = 7$, where the C-index values are better than those of Model A. This may suggest that embeddings help the model better capture complex relationships between variables, improving the ability to predict events over time.

Table 21 – Comparison of Model B performance for different C-index prediction and evaluation time points (average and ± standard deviation). The bigger, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.693 ± 0.02	0.708 ± 0.02	0.702 ± 0.02	0.701 ± 0.01
$t = 3$	0.740 ± 0.01	0.720 ± 0.02	0.728 ± 0.02	0.714 ± 0.02
$t = 5$	0.722 ± 0.02	0.728 ± 0.01	0.720 ± 0.01	0.706 ± 0.02
$t = 7$	0.694 ± 0.03	0.696 ± 0.02	0.695 ± 0.01	0.688 ± 0.01

Source: Elaborated by the author.

In terms of the Brier score, which evaluates the calibration of predictions, Model B also outperforms Model A. For $t = 1$, Model B has a Brier score of 0.066 ± 0.00 at $\Delta t = 1$, while Model A presents 0.060 ± 0.00. Although Model B presents a slightly higher value, indicating a slightly worse calibration for immediate predictions, it compensates for this difference in longer prediction times. For example, for $t = 3$, Model B's Brier score is 0.094 ± 0.00 at $\Delta t = 1$, compared to 0.097 ± 0.00 for Model A. This trend continues with an increase of t, where Model B consistently presents better Brier scores than Model A. This indicates that Model B can better maintain prediction accuracy over time, benefiting from embeddings to capture and represent variables more effectively.

Table 22 – Comparison of Model B performance for different prediction and evaluation time points for the Brier score (mean and ± standard deviation). The smaller, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.066 ± 0.00	0.074 ± 0.00	0.102 ± 0.00	0.147 ± 0.00
$t = 3$	0.094 ± 0.00	0.141 ± 0.00	0.172 ± 0.00	0.191 ± 0.00
$t = 5$	0.134 ± 0.00	0.194 ± 0.00	0.190 ± 0.00	0.203 ± 0.00
$t = 7$	0.165 ± 0.00	0.200 ± 0.00	0.208 ± 0.00	0.211 ± 0.00

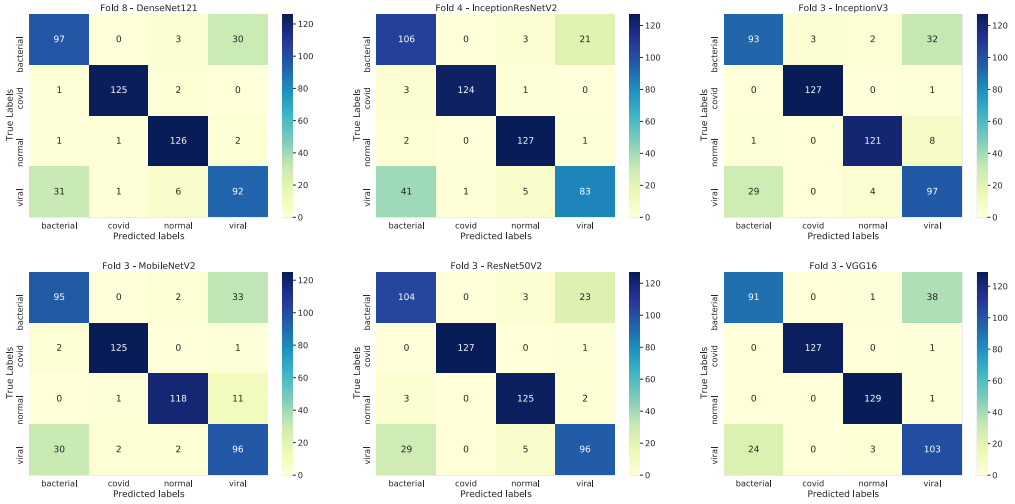
Source: Elaborated by the author.

Analysis of the results shows that using embeddings in Model B provides advantages in terms of performance, especially for longer prediction times. Model B improves event discrimination (C-index) and maintains a more stable calibration (Brier score) over time. These improvements can be attributed to the ability of embeddings to transform categorical and continuous variables into dense vector representations, capturing complex non-linear relationships that traditional tabular analysis methods may not be able to identify. In summary, Model B ablation analysis highlights the importance of incorporating embeddings in tabular data processing, improving the ability to identify underlying patterns and predict risk more accurately.

6.3.3 Model C

One of the objectives of this dissertation was to evaluate the impact of using multimodal data in survival analysis. In this sense, one of the first hypotheses was using architectures for classifying typical findings of pneumonia, COVID-19, and normal findings in chest X-ray images as feature extractors and concatenating them with the tabular embeddings of Model B. In this way, we trained six convolutionary architectures: DenseNet121 (HUANG et al., 2017), InceptionResNetV2 (SZEGEDY et al., 2017), InceptionV3 (SZEGEDY et al., 2016), MovileNetV2 (SANDLER et al., 2018), ResNet50V2 (HE et al., 2016), and VGG16 (SIMONYAN; ZISSERMAN, 2014) by 100 epochs for each fold. We evaluate each model in the validation set at the end of training. The best set of weights was chosen automatically based on the error for the validation set. Figure 20, presents the confusion matrices for each model.

Figure 20 – Confusion matrices of each model for the best test fold set. Dark colors represent a greater number of cases. Light colors represent a smaller amount of cases.



Source: Elaborated by the author.

Analyzing the confusion matrices (Figure 20), we see that all classifiers could correctly classify most cases. We can highlight the tendency to classify cases of viral pneumonia as bacterial and bacterial pneumonia as viral. This trend may indicate that the number of viral and bacterial pneumonia cases was insufficient for an optimized generalization for these two classes. As for the normal cases classification, the ResNet50V2 model had the highest misclassification rate. The classification of pneumonia due to COVID-19 showed similar success rates. The highest false-negative rate for COVID-19 was presented by the InceptionResNetV2 model, with 4 cases. The InceptionV3, ResNet-50, and VGG16 models presented the lowest rate of false negatives, with 1 case.

From the confusion matrix, we can calculate the model's performance metrics

(RUUSKA et al., 2018). Table 23 presents the values obtained for the evaluation metrics in the test fold based on the confusion matrices presented in Figure 20.

Table 23 – Results for the test fold for each model. For each column, the bold values denote the best results.

Model	Accuracy	Sensitivity	Specificity	F1-score	AUC
DenseNet121	81.28±2.27%	81.40±2.23%	81.33±2.26%	81.22±2.32%	0.9620
InceptionResNetV2	84.16±1.42%	83.49±1.52%	84.10±1.47%	84.16±1.42%	0.9707
InceptionV3	83.14±1.01%	83.34±1.09%	83.20±1.00%	83.22±1.04%	0.9704
MovileNetV2	82.04±1.33%	82.55±1.34%	82.10±1.39%	82.21±1.28%	0.9655
ResNet50V2	85.08±1.62%	**85.36**±1.54%	85.12±1.61%	**85.06**±1.60%	0.9748
VGG16	**85.11**±1.30%	85.25±1.27%	**85.16**±1.30%	85.03±1.42%	**0.9758**

Source: Elaborated by the author.

After analyzing the results, it is clear that there was relative stability in the performance metrics for each model. The largest standard deviation for accuracy was ±2.27%, and the largest difference between the models was 3.83% (VGG16 and DenseNet121). These results indicate an adequate generalization of each model for detecting pneumonia due to COVID-19. As for sensitivity, which measures the ability to classify positive classes correctly, the models differ by 3.96%. The maximum variation between models for specificity was 3.83%.

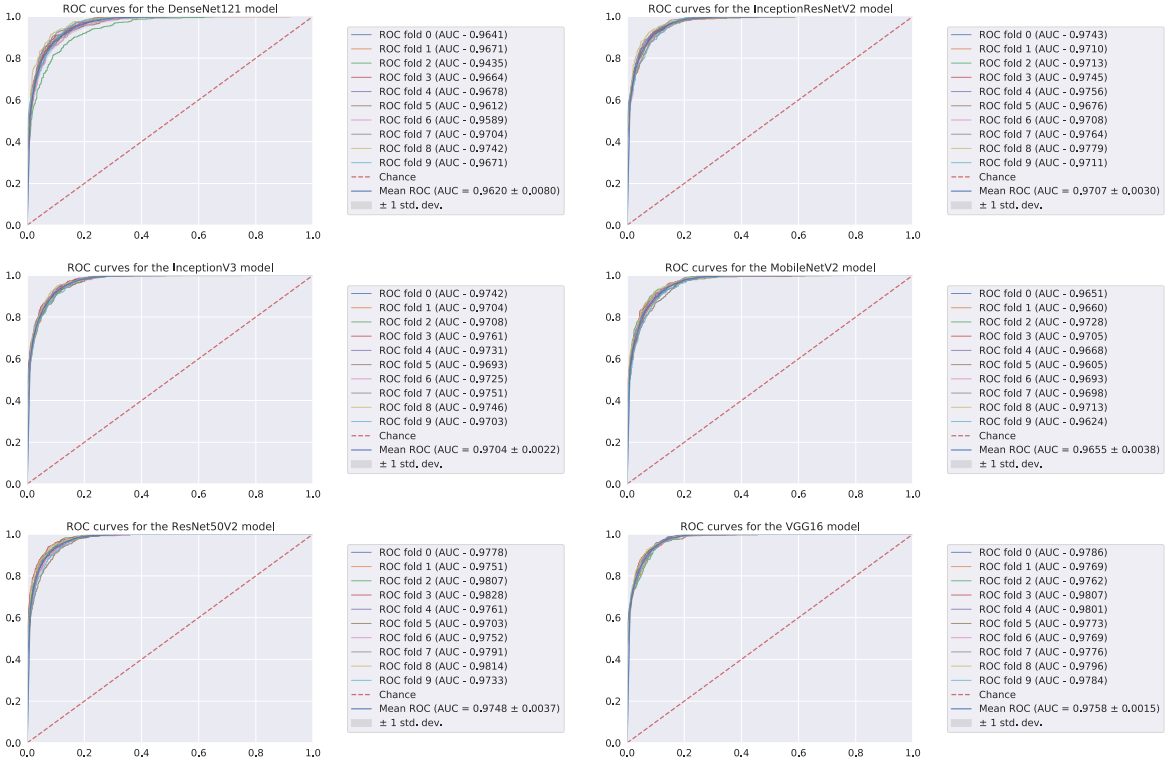
In general, the ResNet50V2 and VGG16 models showed the best results for the chest X-ray classification. This better performance can be associated with the organization of the models. For ResNet50V2, we can highlight the residual blocks that allow an adaptation of the weights to remove filters that were not useful for the final decision (HE et al., 2016). As for VGG16, the performance may indicate that classifying X-Ray features from lower levels, such as more basic forms, is better for differentiating viral pneumonia, bacterial pneumonia, COVID-19, and normal. However, the VGG16 is computationally heavier and requires more training time. Also, VGG16 has a vanishing gradient problem.

Figure 21 presents the ROC curves for each fold and model in the test fold. The AUC showed the greatest stability when comparing the accuracy, sensitivity, specificity, and F1-score metrics, with a variation of only ±0.80%.

Regarding convolutional models, we assessed the impact of using multimodal data in survival analysis with the ResNet50V2 architecture, integrating features extracted from chest X-ray images with tabular embeddings from Model B. Below, in the Tables 24 e 25 we discuss the results obtained by Model C in comparison with Models A and B.

The results for the C-index indicate that Model C presents improvements in relation to Model A. For $t = 1$, Model C achieves a C-index of 0.695 ± 0.01 in $\Delta t = 1$, compared to 0.666 ± 0.02 for Model A and 0.693 ± 0.02 for Model B. This small

Figure 21 – Classification performance for each fold and model in terms of ROC curves in the test fold.



Source: Elaborated by the author.

improvement over Model B suggests that the integration of multimodal data can contribute to discrimination slightly better. As we increase the prediction time, Model C presents a C-index slightly higher than Model A but very close to Model B. For $t = 3$, the C-index of Model C is 0.742 ± 0.012 in $\Delta t = 1$, compared to 0.584 ± 0.04 for Model A and 0.740 ± 0.01 for Model B. This pattern holds for $t = 5$ and $t = 7$, where the differences are minimal. This indicates that although adding multimodal data has some positive effect, this impact is not significantly greater than that already observed with Model B tabular embeddings.

Table 24 – Comparison of Model C performance for different C-index prediction and evaluation time points (average and ± standard deviation). The bigger, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.695 ± 0.01	0.711 ± 0.02	0.701 ± 0.01	0.703 ± 0.01
$t = 3$	0.742 ± 0.01	0.718 ± 0.01	0.730 ± 0.01	0.715 ± 0.01
$t = 5$	0.725 ± 0.01	0.729 ± 0.01	0.721 ± 0.01	0.709 ± 0.01
$t = 7$	0.692 ± 0.02	0.698 ± 0.01	0.696 ± 0.01	0.690 ± 0.01

Source: Elaborated by the author.

For the Brier score, the results also show a small improvement with Model C compared to Models A and B. For $t = 1$, the Brier score for Model C is 0.065 ± 0.00 at

$\Delta t = 1$, compared to 0.060 ± 0.00 for Model A and 0.066 ± 0.00 for Model B. Although Model C presents a slightly lower value than Model B, the difference is small. At $t = 3$, Model C's Brier score is 0.094 ± 0.00 at $\Delta t = 1$, compared to 0.097 ± 0.00 for Model A and 0.094 ± 0.00 for Model B . This trend of subtle improvements continues for $t = 5$ and $t = 7$, where the differences between Model C and Model B remain marginal.

Table 25 – Comparison of Model C performance for different prediction and evaluation time points for the Brier score (mean and ± standard deviation). The smaller, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.065 ± 0.00	0.073 ± 0.00	0.103 ± 0.00	0.146 ± 0.00
$t = 3$	0.094 ± 0.00	0.142 ± 0.00	0.171 ± 0.00	0.190 ± 0.00
$t = 5$	0.133 ± 0.00	0.193 ± 0.00	0.189 ± 0.00	0.202 ± 0.00
$t = 7$	0.164 ± 0.00	0.199 ± 0.00	0.206 ± 0.00	0.209 ± 0.00

Source: Elaborated by the author.

The results indicate that using multimodal data in Model C provides subtle improvements over Model B but without a significant increase in overall performance. The C-index and Brier score of Model C are slightly better compared to Model B, especially for shorter prediction times. However, these improvements are insufficient to conclude that adding multimodal data has a decisive impact. These findings suggest that, although integrating features extracted from X-ray images with tabular embeddings may offer some advantages, the main improvement in the performance of predictive models still comes from using embeddings for categorical and continuous variables.

6.4 MultSurv model

This section investigates the contribution of the CheXReport architecture in MultSurv to survival analysis. Firstly, in Section 6.4.1, we evaluate the performance of the CheXReport architecture for suggesting chest X-ray image reports. Then, in Section 6.4.2, we evaluate and discuss in detail the results of the MultSurv model for survival analysis.

Specifically, the CheXReport is composed of an encoder-decoder architecture. However, unlike most architectures adopted in the literature, we propose a fully transformer architecture. Adopting an improved full transformer architecture with Swin Transformer blocks allows the CheXReport architecture to extract better intrinsic visual features and relationships in X-ray images, integrating these visual features with textual elements. We adopted an encoder with SwinTransformer blocks that extract refined features and intra-report the features to capture radiological

findings from chest X-ray images. The decoder comprises a block of word embeddings connected with N_d Swin Transformer blocks, which are responsible for merging visual and textual resources.

6.4.1 CheXReport

This section investigates the contributions of the encoder and decoder in CheXReport. In Table 26, we present the results for the models based on ResNet101-V2, ResNet50-V2, and Swin Transformer as encoders and LSTM and Transformer as decoders.

Table 26 – Ablation study for the contributions of the encoder and decoder in CheXReport performance in the MIMIC-CXR dataset. Higher is better in all columns. For each column, the bold values denote the best results.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU-4	METEOR	ROUGE
ResNet101-V2 + LSTM	0.227	0.119	0.012	0.007	0.023	0.109	0.235
ResNet101-V2 + Transformer	0.243	0.126	0.084	0.019	0.035	0.123	0.237
ResNet50-V2 + LSTM	0.230	0.120	0.080	0.020	0.030	0.115	0.238
ResNet50-V2 + Transformer	0.248	0.135	0.093	0.032	0.043	0.128	0.245
CheXReport Swin-T	0.339	0.212	0.143	0.105	0.122	0.139	0.270
CheXReport Swin-S	0.344	0.215	**0.149**	0.116	0.123	**0.147**	0.280
CheXReport Swin-B	**0.354**	**0.225**	0.145	**0.127**	**0.130**	**0.147**	**0.284**

Source: Elaborated by the author.

6.4.1.1 Transformer decoder effect

Comparing the results of the models with LSTM-decoders and models with Transformers-decoders in Table 26, it is possible to see a substantial increase in model performance across all metrics. The increase may be related to the already known capacity of Transformer-based networks to capture long-term information, which may reflect a greater ability of the Transformer decoder to transform visual characteristics into more coherent and contextually accurate reports. Additionally, the Transformer decoder's ability to process input sequences of arbitrary length can generate longer captions that provide more detail about the image, leading to higher BLEU scores. Overall, the results suggest that incorporating a Transformer decoder can improve the accuracy and coherence of generated subtitles.

6.4.1.2 Swin Transformer encoder effect

The results in Table 26 for the models based on CheXReport demonstrate the impact of using a Swin Transformer as encoder and decoder in a sequence-to-sequence model for generating chest X-ray reports. Compared to the baseline model that uses a ResNet101-V2 or ResNet50-V2 backbone and an LSTM encoder-decoder architecture, replacing the encoder with a Swin Transformer (CheXReport Swin-T) improves all evaluation metrics. This suggests that the Swin Transformer effectively encodes visual information from medical images into a latent representation that can be used for captions.

Furthermore, increasing the encoder (CheXReport Swin-S and Swin-B) also leads to better performance than the base model. This indicates that Swin Transformer is suitable for generating coherent and accurate natural language descriptions of medical images. Notably, the most improvement is seen in the BLEU-4 score, which measures the fluency and coherence of generated subtitles. This suggests that Swin Transformer is particularly effective at generating longer and more detailed captions that accurately describe the content of the input image. We hypothesize that this behavior may be related to the Swin Transformer's ability to capture information hierarchically and gradually merge neighboring patches, allowing local representations to flow into deeper layers. This behavior favors detecting small changes common in chest X-ray images. These results suggest that using a Swin Transformer as an encoder and decoder in a sequence-to-sequence model can improve the quality of medical image captions.

6.4.1.3 Model Complexity

When considering the full complexity of the models, evaluating both the computational efficiency and the performance outcomes is essential. ResNet50-V2, while providing better performance than the ResNet101-V2 as indicated in Table 26, does not match the superior results of Swin Transformers. The ResNet50-V2 model, although less computationally intensive than ResNet101-V2, falls short in generating detailed and coherent reports, as seen in its lower BLEU-4 and ROUGE scores.

On the other hand, the hierarchical representation learning in Swin Transformers allows for better capture of local and global features, leading to more accurate and contextually relevant descriptions. In conclusion, while ResNet50-V2 offers a baseline performance, the Swin Transformer models, especially the Swin-B variant, provide the best balance of computational efficiency and high-quality report generation, making them the preferred choice for chest X-ray report generation.

6.4.1.4 Qualitative Analysis

In Table 27, we present randomly selected representative test samples to qualitatively compare the methods from the MIMIC-CXR test set for the ResNet101-V2 + LSTM base model and the CheXReport Swin-B model. Analyzing the results of the ResNet101-V2 + LSTM model concerning the ground truth, we can see the model's tendency to produce a smaller number of tokens. Furthermore, the ResNet101-V2 + LSTM model has difficulties in identifying small changes present in the images, such as the presence of coronary calcifications or the presence of surgical clips. Moreover, the ResNet101-V2 + LSTM model produced hallucinations compared to the ground truth, as for the fourth case in Table 27 where the model indicates the presence of a catheter not in the X-ray image of the case.

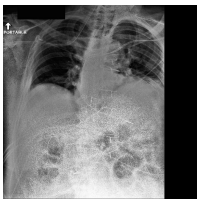
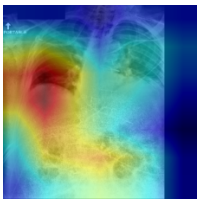
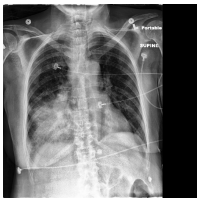
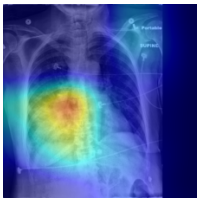
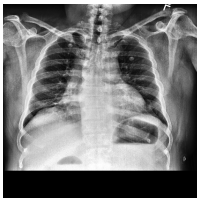
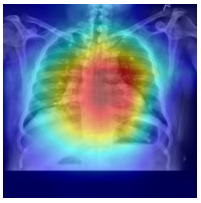
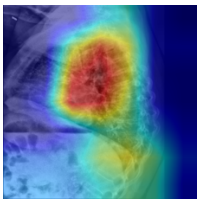
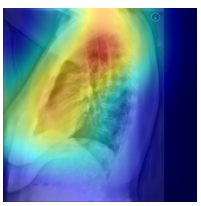
When we compare the CheXReport Swin-B model with the ground truth, we can see a greater similarity of the sentences and with greater detail. This process indicates that the CheXReport Swin-B model can extract visual information and synthesize it into text more efficiently and close to the language adopted by radiologists. Furthermore, the model could also detect elements not reported in the ground truth, as in the first case that presents the presence of surgical clips. Finally, compared to the base model ResNet101-V2 + LSTM, the suggested report produced by the proposed CheXReport Swin-B model demonstrates greater accuracy, covering essential radiological findings in the X-ray image. The observations show the ability of our proposed method to provide accurate radiological reports with less tendency to present hallucinations like the reference model.

6.4.1.5 Comparison of CheXReport with the State of the Art

We compared the CheXReport with several state-of-the-art models for generating chest X-ray reports for the MIMIC-CXR dataset: Competence-based Multimodal curriculum learning framework (CMCL) (LIU; GE; WU, 2022), Contrastive attention (CA) (MA et al., 2021), Meshed-memory transformer ($\mathcal{M}^2$ TR) (NOORALAHZADEH et al., 2021), Prior guided attention (PGA) (YAN et al., 2022), Cross-modal memory networks (CMN) (CHEN et al., 2022), Relational memory and memory-driven conditional layer normalization (RM+MCLN) (CHEN et al., 2020), Medical vision language learner MedViLL (MOON et al., 2022), and knowledge graph with dynamic structure and contrastive learning (DCL) (LI et al., 2023). Furthermore, our comparison against state-of-the-art takes into account models of different approaches to generating chest X-ray reports, such as contrastive learning (MA et al., 2021), knowledge graph (LI et al., 2023), and curriculum learning (LIU; GE; WU, 2022).

In Table 28, we present the results reported by the original studies for the

Table 27 – Comparison of ground truth reports with reports generated by the ResNet101-V2 + LSTM and CheXReport Swin-B models for chest X-ray images randomly selected from the MIMIC-CXR test set.

Image	Ground Truth	ResNet101-V2 + LSTM	CheXReport Swin-B	Attention Map †
	As compared to the previous radiograph, the patient has received a nasogastric tube. The tip projects over the middle parts of the stomach. No complications, notably no pneumothorax. The apical lung regions are not included on the image.	There are no indications of complications observed on this image. No signs of pneumothorax.	No pneumothorax is identified. Possible pleural effusion. Presence of a nasogastric tube. The tube's tip is projected over the middle part of the stomach. Heart in normal size. Surgical clips project over the patient spine.	
	There is no pneumothorax. Small-to-moderate right pleural effusion is stable. There is more consolidation in the right lung outside the entirely consolidated right middle lobe, and suggestion of new consolidation at the base of the left lung which might be due to spillover pneumonia originating in the obstructed right middle lobe. Extensive mediastinal and right hilar adenopathy unchanged. No left pleural effusion. No cardiomegaly.	Pneumothorax detected in the left lung. Small pleural effusion on the right lung. There is no left pleural effusion, and cardiomegaly is not present.	There is no pneumothorax. No left pleural effusion. No cardiomegaly. Presence of a small-to-moderate right pleural effusion. Consolidation is observable in the right and left lung, extending beyond the already consolidated right middle lobe. Extensive mediastinal and right hilar adenopathy.	
	The heart size and mediastinal contours are prominent but similar to prior studies. The lungs are clear. There is no pleural effusion or pneumothorax.	No changes in the lungs.	No significant findings in the lungs. No presence of pleural effusion or pneumothorax. Enlarged heart and mediastinum size. Surgical clips project over the patient spine.	
	Prominent interstitial markings are again seen, not significantly changed. There is no overt pulmonary edema. There is no pleural effusion. Cardiomediastinal silhouette is stable. Coronary artery calcifications and/or stents are noted. Chronic compression deformity in the lower thoracic spine.	Unchanged prominent interstitial markings. No pneumothorax or pleural effusion. There is no evidence of pulmonary edema. Catheter in normal position.	Prominent interstitial markings with no significant changes. No evident signs of pulmonary edema or pleural effusion. The cardiomediastinal silhouette is stable. There are findings of coronary artery calcifications or stents. There is a chronic compression deformity in the lower thoracic spine.	
	The cardiac, mediastinal and hilar contours appear stable. The heart is normal in size. There is no pleural effusion or pneumothorax. The lungs appear clear. The patient is status post anterior cervical fusion. Surgical clips project over the left upper quadrant. There has been no significant change.	No changes in the lungs. The heart is abnormal size	The lung volumes are low and clear. Specifically no evidence of pneumotórax or pleural effusion. Osseous structures are normal. Cardiac and mediastinal are normal. Catheter in normal position. Surgical clips project over the left upper quadrant.	

† Aggregated Attention Map for the Encoder Last Layer.

Source: Elaborated by the author.

MIMIC-CXR dataset. It is possible to see that the CheXReport achieves results similar to state-of-the-art results for generating chest X-ray reports for the MIMIC-CXR for the performance metrics analyzed. The CheXReport outperforms the state-of-the-art for the BLEU-4 and ROUGE metrics. For other metrics, the model achieves results close to state-of-the-art models. These results demonstrate the effectiveness of the CheXReport in extracting features from chest X-ray images to construct suggested findings in radiological reports.

Table 28 – Comparison with state-of-the-art methods on MIMIC-CXR dataset. All metrics for the state-of-the-art are directed cited from the original paper. Higher is better in all columns. For each column, the bold values denote the best results.

Model	Backbone	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU-4	METEOR	ROUGE
CA (MA et al., 2021)	ResNet-50	0.350	0.219	0.152	0.109	-	0.151	0.283
$\mathcal{M}^2$ TR (NOORALAHZADEH et al., 2021)	DenseNet	**0.378**	0.232	**0.154**	0.107	-	0.145	0.272
PGA (YAN et al., 2022)	ResNet101	0.356	0.222	0.151	0.111	-	0.140	0.280
CMN (CHEN et al., 2022)	ResNet101	0.353	0.218	0.148	0.106	-	0.142	0.278
CMCL (LIU; GE; WU, 2022)	ResNet-50	0.344	0.217	0.140	0.097	-	0.145	0.272
RM+MCLN (CHEN et al., 2020)	ResNet101	0.353	0.218	0.145	0.103	-	0.142	0.277
MedViLL (MOON et al., 2022)	ResNet-50	-	-	-	0.126	-	-	-
DCL (LI et al., 2023)	ViT	-	-	-	0.109	-	0.150	0.284
ASGMD (XUE et al., 2024)	ResNet-101/Resnet152	0.372	**0.233**	**0.154**	0.112	-	**0.152**	**0.286**
CheXReport (Ours)	Swin-B	0.354	0.225	0.145	**0.127**	**0.130**	0.147	**0.286**

Source: Elaborated by the author.

The CheXReport performances in terms of BLEU-1 (0.378 vs. 0.354), BLEU-2 (0.233 vs. 0.225), BLEU-3 (0.154 vs. 0.145), and METEOR (0.152 vs. 0.147) are mainly inferior to $\mathcal{M}^2$TR. This performance may be related to the two-phase method adopted to generate the report suggestions. In the $\mathcal{M}^2$ TR, high-level context information is extracted and refined by a language model to generate the final Chest X-ray report suggestions (NOORALAHZADEH et al., 2021). This process can favor the model's performance for shorter text sequences due to the use of base sentences similar to those found in radiological reports to signal the presence or absence of a finding in high-level context information. This is not replicated in longer sentences and corroborated for the BLEU-4, METEOR, and ROUGE metrics, which do not present such a variation in relation to state-of-the-art metrics.

Meanwhile, we can see from the performance metrics of Table 28 that the CheXReport presents a superior result, especially when evaluating anagrams with long sentences. This performance may be related to the capabilities of networks based on Transformer architectures to capture long-term features and produce long sentence sizes (VASWANI et al., 2017). In terms of semantic characteristics of the report produced by CheXReport, we captured the fluency and semantics of the report generated in relation to the ground truth according to ROUGE's performance. This characteristic can be directly related to the embedding model used, which favors understanding standard linguistic characteristics and can incorporate the visual

characteristics extracted by the Swin Transformer encoder throughout the inference process.

The CheXReport's slightly lower scores in shorter anagram metrics (BLEU-1 to BLEU-3) do not necessarily diminish its efficacy or applicability in clinical settings. In fact, the superior performance in BLEU-4 and ROUGE suggests that the CheXReport generates more extensive, coherent, and contextually rich reports. This aspect is crucial in medical report generation, where detail and accuracy are paramount. The ability to construct longer sentences that accurately encapsulate the nuances of a medical image can be more beneficial than mere brevity. It is also important to mention that despite small advances in current literature, the suggestion of findings on chest X-rays is complex and presents relative stability in related studies. Furthermore, this complexity is also reflected in models with more significant numbers of parameters, directly influencing the explainability of these models.

Moreover, the Transformer architecture's prowess in capturing long-term dependencies and producing coherent, longer sentences aligns well with the typical requirements of medical report writing. This capability is reflected in CheXReport's performance in the ROUGE metric, which assesses the fluency and semantic alignment of the generated reports with the ground truth. The effective integration of visual characteristics extracted by the Swin Transformer encoder throughout the report generation process further enhances the model's ability to produce contextually rich and medically relevant reports.

6.4.2 Survival Analysis Results

Incorporating the contextual vectors of the CheXReport architecture in MultSurv model can better extract the intrinsic visual characteristics and relationships of the X-ray images and combine them with the textual resources. In this sense, combining visual embeddings with clinical and laboratory data embeddings aims to explore the complex interactions of variables describing the patient's current health status. In Tables 29 and 30, we present the results for the MultSurv model in the test set.

The results presented in Table 29 indicate that the MultSurv model exhibits robust performance in survival analysis, particularly compared to previous models (Model A, Model B, and Model C). The MultSurv model demonstrated superior performance, especially in the early prediction days. For $t = 1$, MultSurv model achieved a C-index of 0.723 ± 0.08 at $\Delta t = 1$, compared to 0.666 ± 0.02 for Model A, 0.693 ± 0.02 for Model B and 0.695 ± 0.01 of Model C. As the prediction time increases, the C-index performance of MultSurv model remains consistently higher. This suggests that integrating CheXReport's visual and textual features with clinical and laboratory data

Table 29 – Comparison of MultSurv model performance for different C-index prediction and evaluation time points (average and ± standard deviation). The bigger, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.723 ± 0.08	0.735 ± 0.01	0.711 ± 0.02	0.706 ± 0.01
$t = 3$	0.742 ± 0.00	0.729 ± 0.03	0.735 ± 0.00	0.726 ± 0.00
$t = 5$	0.726 ± 0.06	0.731 ± 0.01	0.722 ± 0.03	0.714 ± 0.04
$t = 7$	0.725 ± 0.01	0.715 ± 0.02	0.702 ± 0.03	0.695 ± 0.03

Source: Elaborated by the author.

provides a more comprehensive view of a patient's health status, enabling better survival predictions.

Table 30 – Comparison of MultSurv model performance for different prediction and evaluation time points for the Brier score (average and ± standard deviation). The smaller, the better.

Prediction Time	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
$t = 1$	0.071 ± 0.00	0.024 ± 0.01	0.097 ± 0.00	0.147 ± 0.00
$t = 3$	0.088 ± 0.01	0.112 ± 0.00	0.152 ± 0.00	0.180 ± 0.00
$t = 5$	0.117 ± 0.00	0.161 ± 0.01	0.182 ± 0.01	0.199 ± 0.01
$t = 7$	0.150 ± 0.00	0.168 ± 0.00	0.186 ± 0.00	0.193 ± 0.01

Source: Elaborated by the author.

The results in Table 30 demonstrate that the MultSurv model maintains a relatively low Brier score throughout the different prediction time points. For $t = 1$, the Brier score is 0.071 ± 0.00 at $\Delta t = 1$, higher than 0.060 ± 0.00 for Model A, 0.066 ± 0.00 for Model B and 0.065 ± 0.00 of Model C. Although the MultSurv model presents a slightly higher value than Model A for immediate predictions, this difference is compensated by the superiority in longer prediction times.

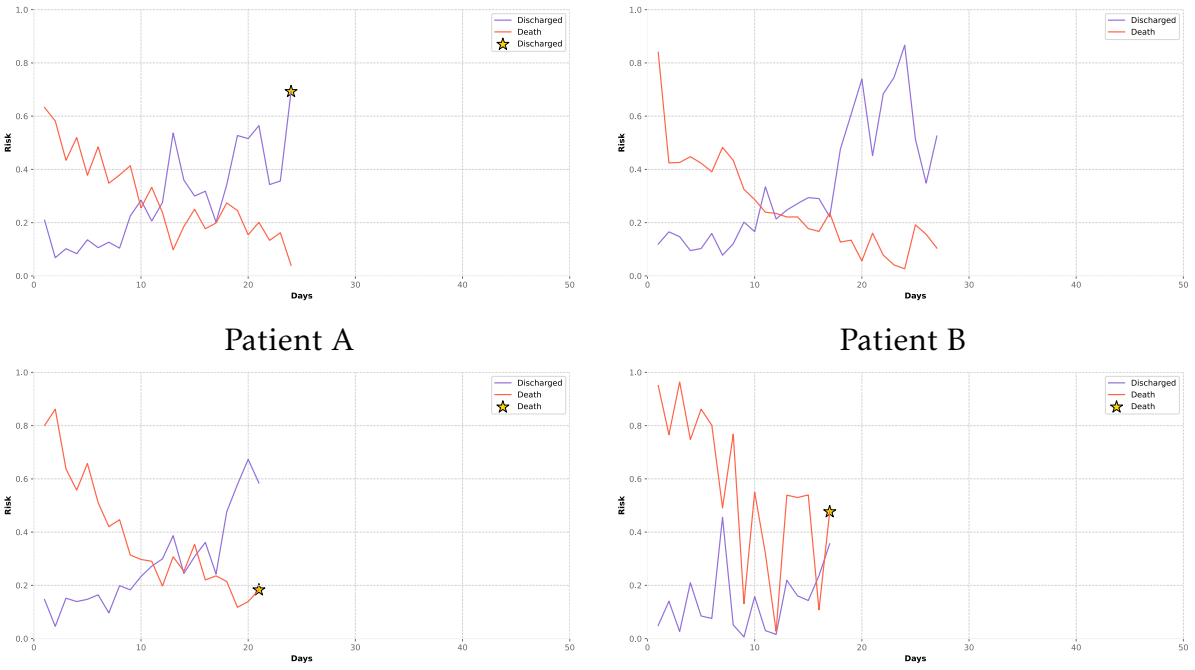
Therefore, compared to Models A, B, and C, the MultSurv model showed improvements in the discrimination and calibration of survival predictions, especially for short and medium-term prediction periods. These advances are particularly relevant to the clinical context, where data-driven decisions can improve patient care and optimize the allocation of healthcare resources.

6.4.3 Qualitative Analysis

In this section, we present a qualitative analysis of the MultSurv model results. First, we evaluated the ability to predict independent risks for death and discharge for patients in the test set (Figure 22). Patients were randomly selected from the subgroups

of censored, death, discharge, and one patient with prediction error. Next, we assessed the global SHAP values to verify the importance of the variables for predicting patient survival. Finally, we evaluate which regions of the X-ray images the MultSurv model is capturing with greater importance for making the prediction.

Figure 22 – Independent risks for discharge and death for patients in the test set over the 50 days. In purple is the risk for discharge, and in red is the risk for death. The star indicates when and what event happened.



Source: Elaborated by the author.

Patient A demonstrates the risk over time of a hospitalized patient who was discharged at the end of hospitalization, indicating that the MultSurv model was correct. It is also possible to observe that as the discharge event approached, the model increased the risk associated with discharge, indicating that the MultSurv model can possibly identify an improvement in the patient's health condition. It is also important to highlight a standard behavior of the model of maintaining a higher risk of death for all patients in the first days of hospitalization, which may be a direct reflection of the health condition of hospitalized patients, who were mainly in severe cases.

Meanwhile, patient B demonstrates the risks over time until the patient's right censoring. In other words, the patient was transferred or gave up treatment at the hospital, and it was impossible to observe what happened after this period. If we analyze the graph (Figure 22), we can see that the risk of discharge was considerably higher during the period in which the patient was censored, which may indicate a possible transfer to a less complex hospital to free up places for patients with severe

conditions.

When we analyze the risk prediction behavior for patient C, we can see that the model does not associate the highest risk with the observed event. Initially, the model identifies a high risk of death, which decreases over time, while the risk of discharge increases inconsistently in relation to the observed event. We observed intersections between these curves over the days, indicating changes in the model's risk perception. It is worth noting that despite the global error, the model still predicted a relative risk of death of around 20% and with a slight upward trend.

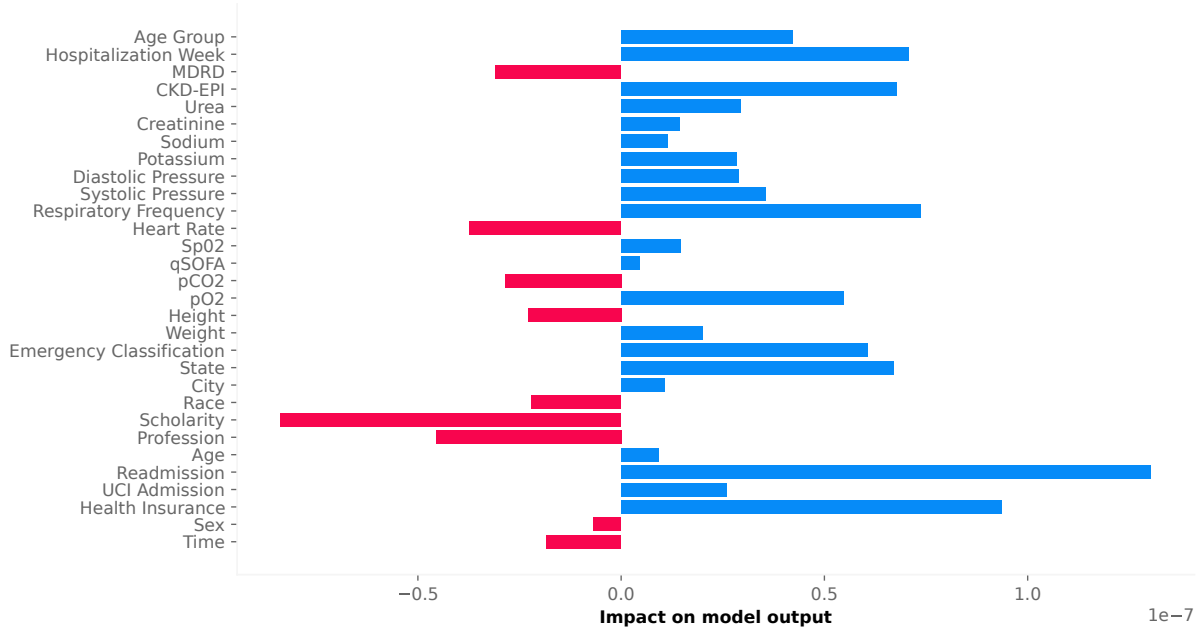
Finally, analyzing the risks associated with patient D, we can see that there is a fluctuation in relation to the patient's risk of death. The fluctuations in the risk curves suggest that patient D's health status may have fluctuated during the observed period, which may have influenced the model's predictions. These fluctuations indicate the importance of continuously monitoring the patient's health status to adjust treatment and ensure a more effective intervention.

In Figure 23, we present the global SHAP values for the model's tabular variables. Positive values contribute to an increase in the patient's risk of death, and negative values decrease the patient's risk. Analyzing the impact of variables on model results reveals critical insights into the determinants of patient outcomes during the COVID-19 pandemic. Variables such as ICU Admission and CKDEPI have positive impacts, highlighting the importance of intensive therapy and renal function in patient survival. Sp02 and potassium levels underline the relevance of vital signs in determining the prognosis of patients with COVID-19. These results suggest that continuous monitoring of these clinical parameters may be critical in making timely and informed decisions about patient care.

On the other hand, the model indicates that sociodemographic factors such as educational level, profession, and race have a negative impact on patient outcomes, reflecting broader social inequalities in access to and quality of healthcare. The negative influence of these variables suggests that disparities in socioeconomic status and racial origin affect health outcomes, potentially due to underlying systemic issues such as access to health care, pre-existing health conditions, and social determinants of health.

SHAP also allows you to evaluate the impact of each variable for each model prediction. In Figure 24, we present an example of SHAP values for a specific observation from the test set. The positive impact of respiratory and heart rates suggests that these vital signs are associated with a worse prognosis, possibly reflecting the severity of the patient's clinical condition. On the other hand, socioeconomic variables such as Race and Profession have negative impacts, indicating that, for this patient, these factors are associated with a lower chance of death. Diastolic pressure and scholarity also show negative impacts, suggesting that

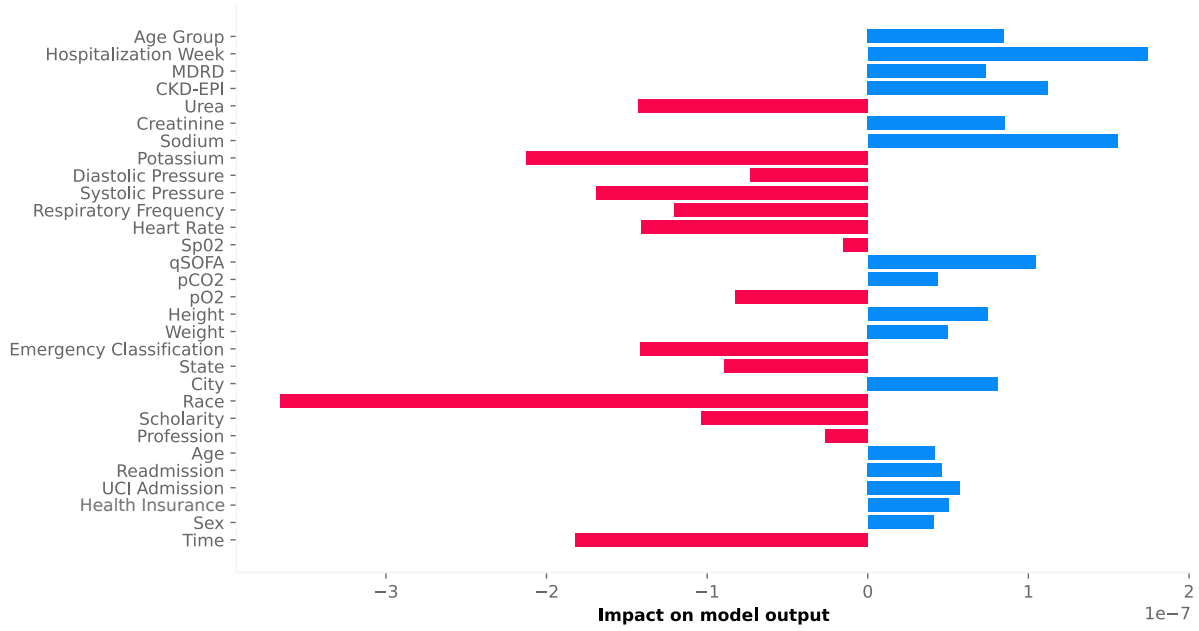
Figure 23 – Global SHAP of each tabular feature on the MultSurv model.



Source: Elaborated by the author.

better socioeconomic conditions and lower diastolic pressure can positively influence the patient's health, reducing the risk of death.

Figure 24 – Local SHAP for a specific observation.

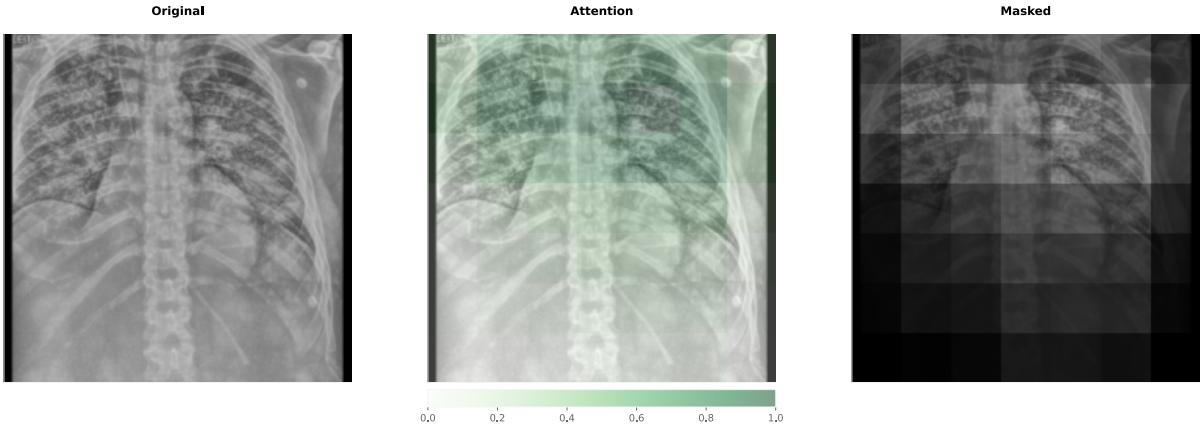


Source: Elaborated by the author.

When we analyze the attention of the CheXReport architecture on the specific patient's X-ray image (Figure 25), we can see that the model gives greater attention to the upper areas of the lungs, as indicated by the darker regions in the attention scale.

This distribution suggests that the model identifies these areas as critical for assessing the patient's health status. Focusing attention on these regions may be associated with the detection of common anomalies in cases of COVID-19, such as lung opacities or infiltrations, which are often observed in the upper part of the lungs.

Figure 25 – Attention over the image for the last layer of the CheXReport.



Source: Elaborated by the author.

6.5 Discussion

In this section, we discuss and compare the MultSurv model with the state-of-the-art methods for survival analysis. A detailed discussion of the state-of-the-art in multimodal survival analysis for hospitalized patients is presented in Chapter 3. One of the main challenges of current survival analysis methods is capturing information from multimodal data. In this sense, for methods based purely on data, we only use sociodemographic, clinical, and laboratory data for training. All methods were trained with the same data set. In Tables 31 e 32, we compare the results obtained by Models A, B, C, and MultSurv in terms of C-index and Brier Score.

Cox-based methods such as CoxTime, CoxCC, and DeepSurv are widely used in survival analysis due to their simplicity and interpretability. However, we observed that the Cox methods perform worse than the MultSurv model in terms of C-index (Table 32). For example, at forecast time $t = 1$, CoxCC achieves a C-index of 0.463 ± 0.07 for $\Delta t = 1$, while DeepSurv and CoxTime have 0.361 ± 0.01 and 0.377 ± 0.08, respectively. These methods are limited by the assumption of proportionality of risks, which may not hold in clinical contexts, where risk factors can interact in complex ways and change over time. In terms of Brier score, for $t = 1$, CoxCC, DeepSurv, and CoxTime present Brier scores of 0.010 ± 0.00, 0.010 ± 0.00, and 0.10 ± 0.00, respectively, for $\Delta t = 1$, while MultSurv model has a slightly higher value of 0.071 ± 0.00. The advantage of Cox methods in immediate predictions can be attributed to their simplicity and the fact that they are designed for scenarios with

Table 31 – Comparison of MultSurv model in relation to various methods for the C-index (average and ± standard deviation). The bigger, the better.

Algorithms	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
Prediction Time $t = 1$				
CoxTime†	0.377 ± 0.08	0.312 ± 0.01	0.380 ± 0.01	0.406 ± 0.04
CoxCC†	0.463 ± 0.07	0.340 ± 0.01	0.400 ± 0.01	0.436 ± 0.02
DeepSurv†	0.361 ± 0.01	0.326 ± 0.07	0.378 ± 0.09	0.410 ± 0.02
PCHazard†	0.555 ± 0.08	0.552 ± 0.01	0.506 ± 0.03	0.524 ± 0.03
DeepHit†	0.533 ± 0.01	0.462 ± 0.01	0.480 ± 0.06	0.456 ± 0.07
N-MTLR†	0.453 ± 0.01	0.452 ± 0.06	0.465 ± 0.03	0.453 ± 0.05
Model A†	0.666 ± 0.02	0.622 ± 0.02	0.629 ± 0.01	0.648 ± 0.01
Model B†	0.693 ± 0.02	0.708 ± 0.02	0.702 ± 0.02	0.701 ± 0.01
Model C	0.695 ± 0.01	0.711 ± 0.02	0.701 ± 0.01	0.703 ± 0.01
MultSurv model	0.723 ± 0.08	0.735 ± 0.01	0.711 ± 0.02	0.706 ± 0.01
Prediction Time $t = 3$				
CoxTime†	0.282 ± 0.07	0.388 ± 0.09	0.418 ± 0.07	0.420 ± 0.03
CoxCC†	0.362 ± 0.01	0.435 ± 0.01	0.461 ± 0.02	0.463 ± 0.05
DeepSurv†	0.340 ± 0.01	0.404 ± 0.01	0.430 ± 0.02	0.439 ± 0.04
PCHazard†	0.541 ± 0.04	0.492 ± 0.03	0.503 ± 0.03	0.476 ± 0.04
DeepHit†	0.472 ± 0.01	0.534 ± 0.06	0.510 ± 0.06	0.495 ± 0.06
N-MTLR†	0.410 ± 0.07	0.452 ± 0.05	0.447 ± 0.05	0.460 ± 0.04
Model A†	0.584 ± 0.04	0.635 ± 0.03	0.639 ± 0.02	0.633 ± 0.02
Model B†	0.740 ± 0.01	0.720 ± 0.02	0.728 ± 0.02	0.714 ± 0.02
Model C	0.742 ± 0.01	0.718 ± 0.01	0.730 ± 0.01	0.715 ± 0.01
MultSurv model	0.742 ± 0.00	0.729 ± 0.03	0.735 ± 0.00	0.726 ± 0.00
Prediction Time $t = 5$				
CoxTime	0.387 ± 0.09	0.420 ± 0.01	0.423 ± 0.02	0.396 ± 0.02
CoxCC†	0.359 ± 0.01	0.438 ± 0.04	0.451 ± 0.06	0.441 ± 0.04
DeepSurv†	0.362 ± 0.08	0.425 ± 0.05	0.437 ± 0.03	0.438 ± 0.03
PCHazard†	0.454 ± 0.04	0.487 ± 0.05	0.480 ± 0.04	0.486 ± 0.04
DeepHit†	0.547 ± 0.04	0.526 ± 0.03	0.512 ± 0.06	0.523 ± 0.03
N-MTLR†	0.457 ± 0.01	0.451 ± 0.06	0.467 ± 0.04	0.475 ± 0.04
Model A†	0.605 ± 0.02	0.617 ± 0.03	0.614 ± 0.02	0.611 ± 0.02
Model B†	0.722 ± 0.02	0.728 ± 0.01	0.720 ± 0.01	0.706 ± 0.02
Model C	0.725 ± 0.01	0.729 ± 0.01	0.721 ± 0.01	0.709 ± 0.01
MultSurv model	0.726 ± 0.06	0.731 ± 0.01	0.722 ± 0.03	0.714 ± 0.04
Prediction Time $t = 7$				
CoxTime	0.371 ± 0.08	0.404 ± 0.05	0.382 ± 0.03	0.358 ± 0.02
CoxCC†	0.420 ± 0.06	0.443 ± 0.07	0.437 ± 0.04	0.430 ± 0.03
DeepSurv†	0.396 ± 0.06	0.432 ± 0.06	0.440 ± 0.04	0.422 ± 0.03
PCHazard†	0.494 ± 0.09	0.478 ± 0.09	0.496 ± 0.08	0.479 ± 0.08
DeepHit†	0.563 ± 0.01	0.510 ± 0.07	0.520 ± 0.04	0.533 ± 0.07
N-MTLR†	0.420 ± 0.01	0.467 ± 0.05	0.476 ± 0.04	0.452 ± 0.03
Model A†	0.605 ± 0.02	0.617 ± 0.03	0.614 ± 0.02	0.611 ± 0.02
Model B†	0.694 ± 0.03	0.696 ± 0.02	0.695 ± 0.01	0.688 ± 0.01
Model C	0.692 ± 0.02	0.698 ± 0.01	0.696 ± 0.01	0.690 ± 0.01
MultSurv model	0.725 ± 0.01	0.715 ± 0.02	0.702 ± 0.03	0.695 ± 0.03

† Trained only with tabular data.

Source: Elaborated by the author.

less temporal variability. However, its effectiveness decreases in long-term predictions, where data complexity and temporal variability play a more significant role.

Adopting mechanisms based on neural networks can reduce the dependence on the assumption of risk proportionality in Cox models. For example, the N-MTLR presents a competitive C-index at different time points. For $t = 1$, the N-MTLR C-index is 0.453 ± 0.01 for $\Delta t = 1$, which already demonstrates an improvement over traditional methods such as CoxCC and CoxTime. Meanwhile, PCHazard demonstrated intermediate performance with better results for short-term predictions for the C-index and Brier score. However, PCHazard's performance deteriorated faster than MultSurv model in long-term predictions.

A considerable improvement in performance metrics is obtained with the DeepHit model, which can model time-dependent covariates and deal with competing risks, which traditional models struggle with. However, DeepHit could not incorporate the patient's health history into the model (LEE; YOON; SCHAAR, 2019). In this sense, it is possible to observe that Model A achieves substantially better performance by incorporating contextual information from health history through an RNN.

However, based on the results obtained for Model B, it is possible that incorporating a temporal context of the patient's health status may not capture all the relationships in the data. In this sense, Model B incorporates embeddings for tabular data, capturing non-linear relationships and improving the discrimination and calibration of predictions compared to Model A. For example, if we have a categorical variable "disease type" with values such as "diabetes", "hypertension", and "asthma", each of these categories would be represented by a vector of real values that can capture the similarity between "diabetes" and "hypertension" more effectively than a binary representation.

At first, Model C incorporation of visual characteristics did not present a considerable performance improvement. The stabilization in performance indices may be related to the complexity of the radiological findings present in the images, which may not be captured with simple classification models. In this sense, adopting architectures, such as CheXReport, that can correlate reports with visual features can represent a more detailed description of radiological findings.

Although MultSurv model's main performance gain is related to using embeddings for tabular data, incorporating the CheXReport architecture allows for a more comprehensive analysis of patient health, capturing nuances that unimodal approaches might miss. The performance of MultSurv model on the C-index metric highlights its superiority over existing methods, especially in the context of short-term and medium-term predictions. As shown in Table 31, MultSurv model achieves a C-index of 0.723 ± 0.08 for $t = 1$ and $\Delta t = 1$, which is higher than the scores

Table 32 – Comparison of MultSurv model in relation to various methods for the Brier (mean and ± standard deviation). The smaller, the better.

Algorithms	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 7$
Prediction Time $t = 1$				
CoxTime†	0.010 ± 0.00	0.023 ± 0.00	0.037 ± 0.00	0.056 ± 0.00
CoxCC†	0.010 ± 0.00	0.023 ± 0.00	0.038 ± 0.00	0.057 ± 0.00
DeepSurv†	0.010 ± 0.00	0.023 ± 0.00	0.037 ± 0.00	0.056 ± 0.00
PCHazard†	0.028 ± 0.07	0.054 ± 0.05	0.066 ± 0.04	0.071 ± 0.03
DeepHit†	0.012 ± 0.00	0.026 ± 0.00	0.041 ± 0.00	0.061 ± 0.00
N-MTLR†	0.011 ± 0.00	0.026 ± 0.00	0.043 ± 0.00	0.066 ± 0.00
Model A†	0.060 ± 0.00	0.092 ± 0.00	0.130 ± 0.00	0.163 ± 0.00
Model B†	0.066 ± 0.00	0.074 ± 0.00	0.102 ± 0.00	0.147 ± 0.00
Model C	0.065 ± 0.00	0.073 ± 0.00	0.103 ± 0.00	0.146 ± 0.00
MultSurv model	0.071 ± 0.00	0.024 ± 0.01	0.097 ± 0.00	0.147 ± 0.00
Prediction Time $t = 3$				
CoxTime†	0.036 ± 0.00	0.051 ± 0.00	0.072 ± 0.00	0.098 ± 0.00
CoxCC†	0.037 ± 0.00	0.052 ± 0.00	0.073 ± 0.00	0.073 ± 0.00
DeepSurv†	0.036 ± 0.00	0.051 ± 0.00	0.072 ± 0.00	0.097 ± 0.00
PCHazard†	0.080 ± 0.02	0.082 ± 0.01	0.081 ± 0.01	0.107 ± 0.00
DeepHit†	0.040 ± 0.00	0.056 ± 0.00	0.077 ± 0.00	0.104 ± 0.01
N-MTLR†	0.042 ± 0.00	0.060 ± 0.00	0.085 ± 0.00	0.117 ± 0.01
Model A†	0.097 ± 0.00	0.136 ± 0.00	0.168 ± 0.00	0.186 ± 0.00
Model B†	0.094 ± 0.00	0.141 ± 0.00	0.172 ± 0.00	0.191 ± 0.00
Model C	0.094 ± 0.00	0.142 ± 0.00	0.171 ± 0.00	0.190 ± 0.00
MultSurv model	0.088 ± 0.01	0.112 ± 0.00	0.152 ± 0.00	0.180 ± 0.00
Prediction Time $t = 5$				
CoxTime†	0.068 ± 0.01	0.092 ± 0.00	0.119 ± 0.00	0.139 ± 0.01
CoxCC†	0.069 ± 0.01	0.094 ± 0.00	0.120 ± 0.00	0.140 ± 0.00
DeepSurv†	0.068 ± 0.01	0.092 ± 0.00	0.118 ± 0.00	0.137 ± 0.00
PCHazard†	0.087 ± 0.01	0.099 ± 0.01	0.136 ± 0.01	0.151 ± 0.01
DeepHit†	0.073 ± 0.01	0.098 ± 0.01	0.126 ± 0.01	0.149 ± 0.01
N-MTLR†	0.080 ± 0.00	0.109 ± 0.01	0.144 ± 0.01	0.174 ± 0.01
Model A†	0.143 ± 0.00	0.174 ± 0.00	0.192 ± 0.00	0.195 ± 0.00
Model B†	0.134 ± 0.00	0.194 ± 0.00	0.190 ± 0.00	0.203 ± 0.00
Model C	0.133 ± 0.00	0.193 ± 0.00	0.189 ± 0.00	0.202 ± 0.00
MultSurv model	0.117 ± 0.00	0.161 ± 0.01	0.182 ± 0.01	0.199 ± 0.01
Prediction Time $t = 7$				
CoxTime†	0.117 ± 0.01	0.146 ± 0.01	0.163 ± 0.01	0.177 ± 0.01
CoxCC†	0.119 ± 0.01	0.147 ± 0.01	0.164 ± 0.01	0.178 ± 0.00
DeepSurv†	0.117 ± 0.01	0.144 ± 0.00	0.161 ± 0.00	0.175 ± 0.00
PCHazard†	0.133 ± 0.01	0.161 ± 0.01	0.180 ± 0.02	0.203 ± 0.01
DeepHit†	0.124 ± 0.01	0.154 ± 0.01	0.175 ± 0.01	0.191 ± 0.01
N-MTLR†	0.140 ± 0.01	0.178 ± 0.02	0.206 ± 0.02	0.231 ± 0.02
Model A†	0.185 ± 0.00	0.202 ± 0.00	0.202 ± 0.00	0.212 ± 0.00
Model B†	0.165 ± 0.00	0.200 ± 0.00	0.208 ± 0.00	0.211 ± 0.00
Model C	0.164 ± 0.00	0.199 ± 0.00	0.206 ± 0.00	0.209 ± 0.00
MultSurv model	0.150 ± 0.00	0.168 ± 0.00	0.186 ± 0.00	0.193 ± 0.01

† Trained only with tabular data.

Source: Elaborated by the author.

of other models trained only with tabular data, such as DeepHit (0.533 ± 0.01) and PMF (0.490 ± 0.01). Even state-of-the-art models like DynamicDeepHit (Model A) achieve lower C-index values (0.666 ± 0.02).

Finally, the results demonstrate that embeddings and incorporating multimodal data are strategies that can improve the performance of survival analysis models. Although still relevant, traditional techniques show limitations in capturing the complexities inherent in digital health data, especially when compared to models that use DL. By capturing a more comprehensive view of a patient's health status, the MultSurv model allows healthcare professionals to adjust their treatment strategies more accurately and effectively, potentially improving patient outcomes and optimizing healthcare resources.

6.6 Final Remarks

In this chapter, we present and discuss the results for MultSurv model. We started with the evaluation of the dataset collected in the MDH Consortium. We then performed an ablation study to evaluate the impact of different modules on the survival analysis of patients hospitalized for COVID-19. Our data were limited to patients admitted to the HCPA. Finally, we compared the results obtained with state-of-the-art methods for survival analysis.

The main scientific contribution of this dissertation was the proposal of a survival analysis model with the capacity to process multimodal and longitudinal data from patients hospitalized with COVID-19. The model can serve as a second opinion in the context of scarcity of hospital resources, prioritization of care, or improvements in patient treatment for healthcare professionals. Furthermore, this model can be adapted and serve as a basis for survival analysis in other contexts involving multimodal and longitudinal data.

Finally, the results demonstrate that the MultSurv model surpasses current methods in the literature in survival analysis in terms of discriminative performance. Furthermore, we support decision-making with local and global interpretable analyses of the MultSurv model. Although the results are promising, there are several directions for future research. Generalization of the model to different populations and geographic contexts should be explored, as well as the incorporation of new types of multimodal data, such as genomic data. Furthermore, integration with real-time healthcare systems and assessment of the clinical impact of the MultSurv model in daily practice are important steps towards validating and improving the model's utility.

7 CONCLUSION

This dissertation was guided by the research question: "How to develop a deep learning architecture that leverages dynamic multimodal data to enhance survival analysis predictions while ensuring the explainability of the model's outputs?". In this sense, in Chapter 3, we investigate the different aspects involving survival analysis for multimodal data from hospitalized patients. The RLR aimed to evaluate key methods and identify current research gaps. Current literature still has gaps in incorporating multimodal data and treating competing risks effectively.

Therefore, to answer the research question and the gaps in the literature, in Chapter 4, we present a proposal for the MultSurv model, a multimodal model for survival analysis with the capability to deal with competing risks. The model uses an architecture that can extract temporal information and concatenate representations of different modalities in multitask networks. Furthermore, we adopted a strategy for generating embeddings for categorical and continuous data in vector representations, allowing the model to capture non-linear relationships between variables. Finally, MultSurv model extracts the visual characteristics of X-ray images using the CheXReport architecture, which seeks to improve the representation of radiological findings for predicting patient risks.

When analyzing the results obtained by MultSurv model, we can highlight that using tabular and visual embeddings, combined with the CheXReport architecture, allowed the capture of more complex relationships between clinical, laboratory, and imaging data. This integrated approach resulted in better survival predictions, especially in short and medium-term periods. The results demonstrated that the MultSurv model outperformed traditional models (Model A) and those that use only tabular data (Model B) or multimodal data without the CheXReport architecture (Model C). The superior C-index and Brier score suggests the effectiveness of MultSurv model in discriminating events and maintaining a stable calibration over time.

Despite the complexity of the MultSurv model architecture, it is still possible to identify the influences of variables on survival predictions. The global analysis of SHAP values and regions of attention in X-ray images can provide transparency in model decisions and improve confidence in the practical application of predictions. Furthermore, the possibility of visualizing the independent risks of death and discharge over time allows us to understand the evolution of patients better. This possibility enables supporting clinical decisions in hospital environments to manage critical patients.

We envision that using a survival analysis model, such as the MultSurv model for pandemics, can assist in efficiently allocating resources essential during periods of

high demand. With the ability to predict clinical outcomes based on patient-specific data, such as clinical, laboratory, and imaging information, the model assists in risk stratification and prioritization of treatments. This prediction enables more accurate clinical decisions and can improve patient triage and optimize the use of ICU beds and ventilators, reducing the burden on healthcare systems and potentially increasing survival rates.

In conclusion, this dissertation presented the MultSurv model for survival analysis in COVID-19 patients, combining multimodal data, explainability mechanisms, and a competing risks architecture. MultSurv model was designed to be integrated into hospital routines, helping to triage and monitor patients. Its practical application goes beyond COVID-19, offering a scalable and adaptable solution for various clinical scenarios. The model's ability to predict independent risks of discharge and death and its detailed interpretability provide actionable insights to improve patient management and outcomes. The contributions from this dissertation and the multidisciplinary collaboration throughout the doctorate resulted in several publications detailed in Section 7.1. Finally, in Section 7.2, the limitations of the study and directions for future work are presented.

7.1 Publications

In this section, we present the publications made throughout the research of this dissertation. The following publications are directly related to the theme of this work:

- **ZEISER, F. A.** ; COSTA, C. A. ; RAMOS, G. O. ; MAIER, A. ; RIGHI, R. R. . CheXReport: A transformer-based architecture to generate chest X-ray reports suggestions. EXPERT SYSTEMS WITH APPLICATIONS, 2024.

- **ZEISER, F. A.** ; SANTOS, I. ; BOHN, H. ; COSTA, C. A. ; RAMOS, G. O. ; MAIER, A. ; ANDRADE, J. R. M. ; BACELAR, A. . Pleural Effusion Classification on Chest X-ray Images with Contrastive Learning. In: 19th International Conference on Web Information Systems and Technologies, 2023, Rome, Italy. 19th International Conference on Web Information Systems and Technologies, 2023.

- **ZEISER, F. A.**; DE MATOS, H. V. ; SCHMITT, A. B. ; COSTA, C. A. ; RAMOS, G. O. . Integration of Epidemiologic, Socioeconomic, and Sociodemographic Indicators to Predict Early COVID-19 In-Hospital Outcomes. In: 20th Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2023), 2023, Belo Horizonte. 20th Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2023), 2023.

- **ZEISER, F. A.** ; COSTA, C. A. ; RAMOS, GABRIEL ; BOHN, HENRIQUE ; SANTOS, ISMAEL ; DONIDA, B. ; BRUN, A.P.O. ; ZARICHTA, N . Generating X-ray Reports Using Global Attention. In: XIX Encontro Nacional de Inteligência. Artificial e Computacional, 2022, Campinas. XIX Encontro Nacional de Inteligência. Artificial e Computacional, 2022.

- **ZEISER, F. A.**; DONIDA, B.; DA COSTA, C. A.; RAMOS, G. O.; SCHERER, J. N.; BARCELLOS, N. T.; ALEGRETTI, A. P.; IKEDA, M. L. R.; MULLER, A. P. W. C.; BOHN, H.; SANTOS, I.; BONI, L.; ANTUNES, R. S.; RIGHI, R. R.; RIGO, S. J. First and second COVID-19 waves in Brazil: A cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality. The Lancet Regional Health - Americas, v. 6, p. 100107, 2022.

- **ZEISER, F. A.**; COSTA, C. A.; RAMOS, G. O.; BOHN, H.; SANTOS, I.; RIGHI, R. Evaluation of Convolutional Neural Networks for COVID-19 Classification on Chest X-Rays. In: 10th Brazilian Conference on Intelligent System (BRACIS 2021), São Paulo, 2021.

- **ZEISER, F. A.**; COSTA, C. A. ; RAMOS, G. O. Convolutional Neural Networks Evaluation for COVID-19 Classification on Chest Radiographs. In: LatinX in AI (LXAI) Research at ICML 2021, 2021.

In addition, several publications were developed in the context of the project and in collaboration with other researchers. The following list presents the other publications:

- Published articles:
  - GUBERT, L. C. ; **ZEISER, F. A.** ; COSTA, C. A. ; KUNST, R. . Classification and Prediction of Hypoglycemia in Patients with Type 2 Diabetes Mellitus Using Data from the EHR and Patient Context. In: 9th International Conference on Internet of Things, Big Data and Security, 2024, Angers, France. 9th International Conference on Internet of Things, Big Data and Security, 2024.
  - BERTONI, A. P. S. ; VALANDRO, C. ; BRASIL, R. A. ; **ZEISER, F. A.** ; WINK, M. R. ; FURLANETTO, T. W. ; DA COSTA, C. A. . NT5E DNA methylation in papillary thyroid cancer: Novel opportunities for precision oncology. MOLECULAR AND CELLULAR ENDOCRINOLOGY, v. 570, p. 111915, 2023.
  - RODRIGUES, V. F. ; RIGHI, R. ; DA COSTA, C. A. ; **ZEISER, F. A.** ; ESKOFIER, B. ; MAIER, A. ; KIM, D. . Digital health in smart cities:

Rethinking the remote health monitoring architecture on combining edge, fog, and cloud. HEALTH AND TECHNOLOGY, v. 1, p. 1, 2023.

- LIMA, G. ; **ZEISER, F. A.** ; DA SILVEIRA, A. ; RIGO, S. ; RAMOS, G. O. . An encoder-decoder deep neural network for binary segmentation of seismic facies. COMPUTERS & GEOSCIENCES, v. 1, p. 105507, 2023.
- KUHN, G. ; **ZEISER, F. A.** ; ROEHE, ADRIANA ; COSTA, C. A. ; RAMOS, G. O. . Aprendizado profundo para assistência histopatológica: um estudo de classificação de micrometástases em câncer de mama. In: Simpósio Brasileiro de Computação Aplicada à Saúde, 2023, São Paulo. XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2023), 2023.
- GOLDSCHMIDT, G. ; **ZEISER, F. A.** ; RIGHI, R. ; COSTA, C. A. . ARTERIAL: A Natural Language Processing Model for Prevention of Information Leakage from Electronic Health Records. In: XIII Brazilian Symposium on Computing Systems Engineering, 2023, Porto Alegre. XIII Brazilian Symposium on Computing Systems Engineering, 2023.
- ZONTA, T. ; COSTA, C. A. ; **ZEISER, F. A.** ; RAMOS, G. O. ; RIGHI, R. R. ; KUNST, R. . A Predictive Maintenance model for Optimizing Production Schedule using Deep Neural Networks. JOURNAL OF MANUFACTURING SYSTEMS, v. 626, p. 450-462, 2022.
- FREITAS, S. A. ; **ZEISER, F. A.** ; RAMOS, G. O. ; COSTA, C. A. . DeepCADD: a Deep Learning Architecture for Automatic Detection of Coronary Artery Disease. In: International Joint Conference on Neural Networks (IJCNN), 2022, Padua. 2022 International Joint Conference on Neural Networks, 2022.
- **ZEISER, F. A.**; COSTA, C. A.; RAMOS, G. O.; BOHN, H.; SANTOS, I.; ROEHE, A. V. DeepBatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. Expert Systems With Applications, v. 185, p. 115586, 2021.
- **ZEISER, F. A.**; COSTA, C. A.; ROEHE, A. V.; RIGHI, R.; MARQUES, N. M. C. Breast cancer intelligent analysis of histopathological data: A systematic review. Applied Soft Computing, v. 113, p. 107886, 2021.
- LIMA, G. ; RAMOS, G. O. ; RIGO, S. J. ; **ZEISER, F. A.**; SILVEIRA, A. . Binary Segmentation of Seismic Facies Using Encoder-Decoder Neural Networks. In: LatinX in AI Workshop @ NeurIPS 2020, 2020, Online. Proc. of LatinX in AI Workshop @ NeurIPS 2020, 2020.

- Articles under review::

  - RODRIGUES NETO, J. ; KUHN, G. ; **ZEISER, F. A.** ; ROEHE, A. V. ; COSTA, C. A. ; RAMOS, G. O. Deep learning for histopathological assistance: a classification-segmentation model to detect micrometastases in breast cancer. 2024.
  - ZONTA, T.; COSTA, C. A.; **ZEISER, F. A.**; RAMOS, G. O.; RIGHI, R. R.; KUNST, R. A Degradation Index Model for Maintenance Prediction Run-to-failure in Production Systems. Computers & Industrial Engineering, 2024.

- Published book chapters:

  - COSTA, C. A. ; **ZEISER, F. A.** ; RIGHI, R. R. ; ANTUNES, R. S. ; ALEGRETTI, A. P. ; BERTONI, A. P. ; RAMOS, G. O. ; MELLO, B. H. ; VANIN, F. ; BERTOLETTI, O. A. ; RIGO, S. J. . Internet of Things and Machine Learning for Smart Healthcare. 2024.
  - BERTONI, A. P. S. ; RODRIGUES, V. F. ; **ZEISER, F. A.** ; MELLO, B. H. ; COSTA, C. A. ; DONIDA, B. ; RIGO, S. J. ; RIGHI, R. R. . Internet das Coisas de Saúde: aplicando IoT, interoperabilidade e aprendizado de máquina com foco no paciente. Minicursos do XXII Simpósio Brasileiro de Computação Aplicada à Saúde. 1ed.: , 2022, v. , p. 1-.

- Published abstracts:

  - BOHN, HENRIQUE ; **ZEISER, F. A.** ; COSTA, C. A. . Classificação de Derrame Pleural em Radiografias do Tórax com Contrastive. In: XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023, São Leopoldo. XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023.
  - SANTOS, I. ; **ZEISER, F. A.** ; COSTA, C. A. ; RAMOS, G. O. . Atenção Global para Sugestão de Achados em Radiografias de Tórax. In: XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023, São Leopoldo. XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023.
  - SCHMITT, A. B. ; DE MATOS, H. V. ; **ZEISER, F. A.** ; COSTA, C. A. . MODELO PARA PREDIÇÃO DE ÍNDICE DE MORTALIDADE HOSPITALAR DE COVID-19. In: XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023, São Leopoldo. XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023.
  - DE MATOS, H. V. ; SCHMITT, A. B. ; **ZEISER, F. A.** ; COSTA, C. A. . MODELO DE MACHINE LEARNING PARA PREVISÃO DE SOBREVIVÊNCIA DE PACIENTES. In: XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023, São Leopoldo. XXX Mostra Unisinos de Iniciação Científica e Tecnológica, 2023.

- HENTZ, R. ; **ZEISER, F. A.** . DETECÇÃO DE MINERAÇÃO ILEGAL EM FLORESTAS COM RESNET50: UMA ABORDAGEM COM REDE NEURAL CONVOLUCIONAL.. In: XVI Seminário De Iniciação Científica E Seminário Integrado De Ensino, Pesquisa E Extensão, 2023, Chapecó. XVI Seminário De Iniciação Científica E Seminário Integrado De Ensino, Pesquisa E Extensão, 2023.
- BOHN, H. C. ; **ZEISER, F. A.** ; COSTA, C. A. . Primeira e Segunda Onda da COVID-19 no Brasil: Um Estudo Restrospectivo de Pacientes Hospitalizados. In: XXIX Mostra Unisinos de Iniciação Científica e Tecnológica, 2022, São Leopoldo. XXIX Mostra Unisinos de Iniciação Científica e Tecnológica, 2022.
- BONI, L. ; SANTOS, ISMAEL ; **ZEISER, F. A.** . Estratégias dos serviços de atenção primária no Brasil para manejo e controle da infecção COVID-19. In: VII Congresso Sul-Brasileiro de Medicina de Família e Comunidade, 2022, Porto Alegre. VII Congresso Sul-Brasileiro de Medicina de Família e Comunidade, 2022.
- BOHN, H. C. ; **ZEISER, F. A.** ; COSTA, C. A. ; ROEHE, A. V. . Sistema de Apoio à Decisão baseado em Aprendizado Profundo para Interpretação e Diagnóstico de Câncer de Mama em Imagens Histológicas. In: XXVVIII Mostra Unisinos de Iniciação Científica e Tecnológica, 2021, São Leopoldo. Anais da XXVVIII Mostra Unisinos de Iniciação Científica e Tecnológica. São Leopoldo: Casa Leiria, 2021. p. 332-333.

7.2 Limitations and Future Work

The MultSurv model has some limitations, as described below. First, the data used in this study were derived from a single healthcare institution, the HCPA, limiting the generalizability of the findings. The homogeneity of the dataset, particularly concerning the demographic and regional characteristics of the patient population, might restrict the applicability of the MultSurv model to other settings with different patient demographics and healthcare practices. Second, the study relies on the quality and completeness of EHRs and imaging data. Missing or inconsistent data could adversely impact the model's performance. Although techniques to handle missing data were employed, the inherent uncertainties associated with such data could lead to biases in the model's predictions.

Third, the MultSurv model's reliance on chest X-ray images and associated clinical data may not fully capture the complex and multifactorial nature of COVID-19 and its progression. While integrating multimodal data enhances the model's predictive

capabilities, the absence of additional modalities, such as computed tomography scans not included in the dataset and comprehensive patient histories, could limit the model's effectiveness in real-world clinical scenarios. Moreover, the interpretability of DL models remains a challenge. Although the MultSurv model incorporates explainability mechanisms, the complexity of the underlying algorithms can make it difficult for clinicians to fully trust and understand the model's predictions. This black-box nature of DL models can be a barrier to their adoption in clinical practice.

The evaluation of the MultSurv model was conducted retrospectively. Prospective validation in diverse clinical settings is essential to confirm the model's utility and reliability in real-time decision-making processes. Additionally, the impact of the MultSurv model on clinical outcomes, resource allocation, and patient management was not assessed and warrants future investigation. Lastly, while the model demonstrated robustness in short to medium-term predictions, its performance for long-term predictions was not thoroughly evaluated. The dynamic nature of patient health status and the evolution of COVID-19 treatments over time necessitate continuous model updates and validations to maintain accuracy and relevance.

Future work should focus on addressing these limitations. To improve the generalizability of the MultSurv model, future research should incorporate data from multiple healthcare institutions across diverse geographic regions. Expanding the dataset to include other types of medical imaging, such as computed tomography scans, patient histories, and genomic data, can provide a more comprehensive view of a patient's health status. Also, conducting prospective validation studies in diverse clinical environments is crucial for assessing the real-world performance of the MultSurv model. Such studies will help determine the model's impact on clinical decision-making, patient outcomes, and healthcare resource management.

Enhancing the transparency of the MultSurv model will help build trust among clinicians, facilitating its integration into routine clinical practice. Given the evolving nature of pandemics, incorporating mechanisms for adaptive and continual learning into the MultSurv model is important. This will enable the model to update itself with new data and emerging trends, helping to maintain its accuracy and relevance over time. Techniques such as transfer learning and online learning could be employed to achieve this goal. Future work should also address the ethical and privacy concerns of using AI in healthcare. Ensuring the protection of patient data and addressing potential biases in the model is critical for its responsible deployment. Developing guidelines and frameworks for the ethical use of AI in clinical settings will support the broader acceptance of such technologies.

REFERENCES

ACHARYA, Krishna Prasad; GHIMIRE, Tirth Raj; SUBRAMANYA, Supram Hosuru. Access to and equitable distribution of covid-19 vaccine in low-income countries. **npj Vaccines**, Nature Publishing Group, v. 6, n. 1, p. 1–3, 2021.

ACR, ACR; RADIOLOGY, American College of et al. **Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection**. [S.l.]: American College of Radiology Richmond, VA, 2020.

ÅHLIN, Philip; ALMSTRÖM, Peter; WÄNSTRÖM, Carl. When patients get stuck: a systematic literature review on throughput barriers in hospital-wide patient processes. **Health Policy**, Elsevier, v. 126, n. 2, p. 87–98, 2022.

ALBAWI, Saad; MOHAMMED, Tareq Abed; AL-ZAWI, Saad. Understanding of a convolutional neural network. In: IEEE. **2017 international conference on engineering and technology (ICET)**. [S.l.], 2017. p. 1–6.

ANSELL, Christopher; SØRENSEN, Eva; TORFING, Jacob. The covid-19 pandemic as a game changer for public administration and leadership? the need for robust governance responses to turbulent problems. **Public Management Review**, Taylor & Francis, p. 1–12, 2020.

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

BANERJEE, Satanjeev; LAVIE, Alon. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Available at: <https://www.aclweb.org/anthology/W05-0909>.

BARBEY, Aron K. Network neuroscience theory of human intelligence. **Trends in cognitive sciences**, Elsevier, v. 22, n. 1, p. 8–20, 2018.

BEDFORD, Juliet et al. A new twenty-first century science for effective epidemic response. **Nature**, Nature Publishing Group, v. 575, n. 7781, p. 130–136, 2019.

BENGIO, Yoshua; SIMARD, Patrice; FRASCONI, Paolo. Learning long-term dependencies with gradient descent is difficult. **IEEE transactions on neural networks**, IEEE, v. 5, n. 2, p. 157–166, 1994.

BENGIS, RG et al. The role of wildlife in emerging and re-emerging zoonoses. **Revue scientifique et technique-office international des epizooties**, Citeseer, v. 23, n. 2, p. 497–512, 2004.

BERGSTRA, James; BENGIO, Yoshua. Random search for hyper-parameter optimization. **Journal of machine learning research**, v. 13, n. 2, 2012.

BETTHÄUSER, Bastian A; BACH-MORTENSEN, Anders M; ENGZELL, Per. A systematic review and meta-analysis of the evidence on learning during the covid-19 pandemic. **Nature Human Behaviour**, Nature Publishing Group, v. 7, n. 3, p. 375–385, 2023.

BRADBURN, Mike J et al. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. **British journal of cancer**, Nature Publishing Group, v. 89, n. 3, p. 431–436, 2003.

BRANDAL, Lin T et al. Outbreak caused by the sars-cov-2 omicron variant in norway, november to december 2021. **Eurosurveillance**, European Centre for Disease Prevention and Control, v. 26, n. 50, p. 2101147, 2021.

BURKI, Talha. Covid-19 in latin america. **The Lancet Infectious Diseases**, Elsevier, v. 20, n. 5, p. 547–548, 2020.

BURSZTYN, Leonardo et al. **Misinformation during a pandemic**. [S.l.], 2020.

CAMINADE, Cyril; MCINTYRE, K Marie; JONES, Anne E. Impact of recent and future climate change on vector-borne diseases. **Annals of the New York Academy of Sciences**, Wiley Online Library, v. 1436, n. 1, p. 157–173, 2019.

CARTER, Jane V. et al. Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves. **Surgery**, v. 159, n. 6, p. 1638 – 1645, 2016.

CHAUDHARI, Sneha et al. An attentive survey of attention models. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, v. 12, n. 5, p. 1–32, 2021.

CHEN, Zhihong et al. Cross-modal memory networks for radiology report generation. **arXiv preprint arXiv:2204.13258**, 2022.

____. Generating radiology reports via memory-driven transformer. **arXiv preprint arXiv:2010.16056**, 2020.

CHO, Kyunghyun et al. On the properties of neural machine translation: Encoder-decoder approaches. **arXiv preprint arXiv:1409.1259**, 2014.

CHOLLET, François. Xception: Deep learning with depthwise separable convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 1251–1258.

CHOLLET, Francois. **Deep learning with Python**. [S.l.]: Simon and Schuster, 2021.

CLARK, Taane G et al. Survival analysis part i: basic concepts and first analyses. **British journal of cancer**, Nature Publishing Group, v. 89, n. 2, p. 232–238, 2003.

COHEN, Joseph Paul et al. **COVID-19 Image Data Collection: Prospective Predictions Are the Future**. 2020.

COPPIN, Ben. **Inteligência artificial**. 1. ed. Rio de Janeiro: LTC, 2010. 636 p.

COX, David R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

DEVLIN, Jacob. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DICK, Stephanie. Artificial intelligence. **Harvard Data Science Review**, PubPub, v. 1, n. 1, 2019.

DOGO, EM et al. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In: IEEE. **2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)**. [S.l.], 2018. p. 92–99.

DONIDA, Bruna; COSTA, Cristiano André da; SCHERER, Juliana Nichterwitz. Making the covid-19 pandemic a driver for digital health: Brazilian strategies. **JMIR Public Health Surveill**, v. 7, n. 6, p. e28643, Jun 2021.

DOSOVITSKIY, Alexey et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2021.

DOZAT, Timothy. Incorporating nesterov momentum into adam. 2016.

DUCHI, John; HAZAN, Elad; SINGER, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. **Journal of machine learning research**, v. 12, n. 7, 2011.

ENGELEN, Jesper E Van; HOOS, Holger H. A survey on semi-supervised learning. **Machine Learning**, Springer, v. 109, n. 2, p. 373–440, 2020.

EZHILAN, Madeshwari; SURESH, Indhu; NESAKUMAR, Noel. Sars-cov, mers-cov and sars-cov-2: a diagnostic challenge. **Measurement**, Elsevier, v. 168, p. 108335, 2021.

FOTSO, Stephane. Deep neural networks for survival analysis based on a multi-task framework. **arXiv preprint arXiv:1801.05512**, 2018.

FU, Yu et al. Severity-onset prediction of covid-19 via artificial-intelligence analysis of multivariate factors. **Heliyon**, Elsevier, v. 9, n. 8, 2023.

FUKUSHIMA, Kunihiko. Visual feature extraction by a multilayered network of analog threshold elements. **IEEE Transactions on Systems Science and Cybernetics**, IEEE, v. 5, n. 4, p. 322–333, 1969.

GE, Xing-Yi et al. Isolation and characterization of a bat sars-like coronavirus that uses the ace2 receptor. **Nature**, Nature Publishing Group, v. 503, n. 7477, p. 535–538, 2013.

GEIRHOS, Robert et al. Comparing deep neural networks against humans: object recognition when the signal gets weaker. **arXiv preprint arXiv:1706.06969**, 2017.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. 1. ed. Cambridge: MIT Press, 2017.

GORISHNIY, Yury et al. Revisiting deep learning models for tabular data. **Advances in Neural Information Processing Systems**, v. 34, p. 18932–18943, 2021.

GRAHAM, Barney S; SULLIVAN, Nancy J. Emerging viral diseases from a vaccinology perspective: preparing for the next pandemic. **Nature immunology**, Nature Publishing Group, v. 19, n. 1, p. 20–28, 2018.

GRENNAN, Dara. What is a pandemic? **Jama**, American Medical Association, v. 321, n. 9, p. 910–910, 2019.

GRIFFIN, Kelly M et al. Hospital preparedness for covid-19: a practical guide from a critical care perspective. **American journal of respiratory and critical care medicine**, American Thoracic Society, v. 201, n. 11, p. 1337–1344, 2020.

GRONVALL, Gigi Kwik. The scientific response to covid-19 and lessons for security. **Survival**, Taylor & Francis, v. 62, n. 3, p. 77–92, 2020.

HARKER, Julie; KLEIJNEN, Jos. What is a rapid review? a methodological exploration of rapid reviews in health technology assessments. **International Journal of Evidence-Based Healthcare**, Wiley Online Library, v. 10, n. 4, p. 397–410, 2012.

HAYKIN, Simon S. **Neural networks and learning machines**. Third. Upper Saddle River, NJ: Pearson Education, 2009.

HE, Kaiming et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HINTON, Geoffrey; SRIVASTAVA, Nitish; SWERSKY, Kevin. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. **Cited on**, v. 14, n. 8, p. 2, 2012.

HO, Matthew et al. Longitudinal dynamic clinical phenotypes of in-hospital covid-19 patients across three dominant virus variants in new york. **International Journal of Medical Informatics**, Elsevier, v. 181, p. 105286, 2024.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HOO, Zhe Hui; CANDLISH, Jane; TEARE, Dawn. **What is an ROC curve?** [S.l.]: BMJ Publishing Group Ltd and the British Association for Accident . . . , 2017. 357–359 p.

HUANG, Chaolin et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. **The lancet**, Elsevier, v. 395, n. 10223, p. 497–506, 2020.

HUANG, Sandy et al. Adversarial attacks on neural network policies. **arXiv preprint arXiv:1702.02284**, 2017.

HUANG, Xin et al. Tabtransformer: Tabular data modeling using contextual embeddings. **arXiv preprint arXiv:2012.06678**, 2020.

HUGHES, James M et al. The origin and prevention of pandemics. **Clinical Infectious Diseases**, The University of Chicago Press, v. 50, n. 12, p. 1636–1640, 2010.

IOFFE, Sergey; SZEGEDY, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR. **International conference on machine learning**. [S.l.], 2015. p. 448–456.

JAMES, Nick; MENZIES, Max; RADCHENKO, Peter. Covid-19 second wave mortality in europe and the united states. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, v. 31, n. 3, p. 031105, 2021.

JANIESCH, Christian; ZSCHECH, Patrick; HEINRICH, Kai. Machine learning and deep learning. **Electronic Markets**, Springer, v. 31, n. 3, p. 685–695, 2021.

JEONG, Yeojin; SUNG, Joohon. An automated deep learning method and novel cardiac index to detect canine cardiomegaly from simple radiography. **Scientific Reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 14494, 2022.

JIANG, Fei; GUTERMAN, Elan. Survival analysis. **Statistical Methods in Epilepsy**, Chapman and Hall/CRC, p. 124–142, 2024.

JING, Longlong; TIAN, Yingli. Self-supervised visual feature learning with deep neural networks: A survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 43, n. 11, p. 4037–4058, 2020.

JOHNSON, Alistair EW et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. **Scientific data**, Nature Publishing Group UK London, v. 6, n. 1, p. 317, 2019.

KAPLAN, Edward L; MEIER, Paul. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KATZMAN, Jared L et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. **BMC medical research methodology**, Springer, v. 18, p. 1–12, 2018.

KHANGURA, Sara et al. Evidence summaries: the evolution of a rapid review approach. **Systematic reviews**, BioMed Central, v. 1, n. 1, p. 1–9, 2012.

KINGMA, Diederik P; BA, Jimmy. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KLEIN, John P. **Survival analysis: Techniques for censored and truncated data**. [S.l.]: Springer Science Business Media, Inc, 2003.

KLEINBAUM, David G et al. Kaplan-meier survival curves and the log-rank test. **Survival analysis: a self-learning text**, Springer, p. 55–96, 2012.

KRAUSE, Ben et al. Multiplicative lstm for sequence modelling. **arXiv preprint arXiv:1609.07959**, 2016.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KVAMME, Håvard; BORGAN, Ørnulf. Continuous and discrete-time survival prediction with neural networks. **Lifetime data analysis**, Springer, v. 27, n. 4, p. 710–736, 2021.

KVAMME, Håvard; BORGAN, Ørnulf; SCHEEL, Ida. Time-to-event prediction with neural networks and cox regression. **Journal of machine learning research**, v. 20, n. 129, p. 1–30, 2019.

LABATIE, Antoine et al. Proxy-normalizing activations to match batch normalization while removing batch dependence. **Advances in Neural Information Processing Systems**, v. 34, 2021.

LEACH, Melissa et al. Post-pandemic transformations: How and why covid-19 requires us to rethink development. **World development**, Elsevier, v. 138, p. 105233, 2021.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Yann A et al. Efficient backprop. In: **Neural networks: Tricks of the trade**. [S.l.]: Springer, 2012. p. 9–48.

LEE, Changhee; YOON, Jinsung; SCHAAR, Mihaela Van Der. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 67, n. 1, p. 122–133, 2019.

LEE, Changhee et al. Deephit: A deep learning approach to survival analysis with competing risks. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2018. v. 32, n. 1.

LEGG, Shane; HUTTER, Marcus et al. A collection of definitions of intelligence. **Frontiers in Artificial Intelligence and applications**, IOS press, v. 157, p. 17, 2007.

LEISMAN, Daniel E et al. Cytokine elevation in severe and critical covid-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. **The Lancet Respiratory Medicine**, Elsevier, v. 8, n. 12, p. 1233–1244, 2020.

LI, Long-quan et al. Covid-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. **Journal of medical virology**, Wiley Online Library, v. 92, n. 6, p. 577–583, 2020.

LI, Mingjie et al. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 3334–3343.

LIN, Chin-Yew. Rouge: A package for automatic evaluation of summaries. In: **Text summarization branches out**. [S.l.: s.n.], 2004. p. 74–81.

LITEWKA, Sergio G; HEITMAN, Elizabeth. Latin american healthcare systems in times of pandemic. **Developing world bioethics**, Wiley Online Library, v. 20, n. 2, p. 69–73, 2020.

LIU, Fenglin; GE, Shen; WU, Xian. Competence-based multimodal curriculum learning for medical report generation. **arXiv preprint arXiv:2206.14579**, 2022.

LIU, Ze et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 10012–10022.

LIU, Zhuang et al. A convnet for the 2020s. **arXiv preprint arXiv:2201.03545**, 2022.

LIU, Ze et al. Video swin transformer. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 3202–3211.

MA, Xuewei et al. Contrastive attention for automatic chest x-ray report generation. **arXiv preprint arXiv:2106.06965**, 2021.

MACEDO, Ana; GONCALVES, Nilza; FEBRA, Claudia. Covid-19 fatality rates in hospitalized patients: systematic review and meta-analysis. **Annals of epidemiology**, Elsevier, v. 57, p. 14–21, 2021.

MATTHAY, Michael A et al. Acute respiratory distress syndrome. **Nature reviews Disease primers**, Nature Publishing Group, v. 5, n. 1, p. 1–22, 2019.

MCCABE, Ruth et al. Adapting hospital capacity to meet changing demands during the covid-19 pandemic. **BMC medicine**, Springer, v. 18, p. 1–12, 2020.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943.

MEGANCK, Rita M; BARIC, Ralph S. Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases. **Nature medicine**, Nature Publishing Group, v. 27, n. 3, p. 401–410, 2021.

MEHTAR, Shaheen et al. Limiting the spread of covid-19 in africa: one size mitigation strategies do not fit all countries. **The Lancet Global Health**, Elsevier, v. 8, n. 7, p. e881–e883, 2020.

MINSKY, Marvin; PAPERT, Seymour. Perceptron: An introduction to computational geometry. **Cambridge tiass., HIT**, v. 479, p. 480, 1969.

MISRA, Ishan; MAATEN, Laurens van der. Self-supervised learning of pretext-invariant representations. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2020. p. 6707–6717.

MITCHELL, Tom M et al. Machine learning. **Burr Ridge, IL: McGraw Hill**, v. 45, n. 37, p. 870–877, 1997.

MOON, Jong Hak et al. Multi-modal understanding and generation for medical images and text via vision-language pre-training. **IEEE Journal of Biomedical and Health Informatics**, IEEE, v. 26, n. 12, p. 6070–6080, 2022.

MURTAUGH, Paul A et al. Primary biliary cirrhosis: prediction of short–term survival based on repeated patient visits. **Hepatology**, LWW, v. 20, n. 1, p. 126–134, 1994.

NASSIF, Ali Bou et al. Speech recognition using deep neural networks: A systematic review. **IEEE access**, IEEE, v. 7, p. 19143–19165, 2019.

NOORALAHZADEH, Farhad et al. Progressive transformer-based generation of radiology reports. **arXiv preprint arXiv:2102.09777**, 2021.

ORAN, Daniel P; TOPOL, Eric J. The proportion of sars-cov-2 infections that are asymptomatic: a systematic review. **Annals of internal medicine**, American College of Physicians, v. 174, n. 5, p. 655–662, 2021.

OSTERHOLM, Michael. Preparing for the next pandemic. **New England Journal of Medicine**, v. 352, n. 18, p. 1839–1842, 2005.

OSTERHOLM, Michael T. **Preparing for the next pandemic**. [S.l.]: Routledge, 2017.

OTTER, Daniel W; MEDINA, Julian R; KALITA, Jugal K. A survey of the usages of deep learning for natural language processing. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 32, n. 2, p. 604–624, 2020.

PANDEY, Navin et al. Transforming a general hospital to an infectious disease hospital for covid-19 over 2 weeks. **Frontiers in public health**, Frontiers Media SA, v. 8, p. 382, 2020.

PANEL, One Health High-Level Expert et al. Developing one health surveillance systems. **One Health**, Elsevier, p. 100617, 2023.

PANWAR, Harsh et al. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. **Chaos, Solitons & Fractals**, Elsevier, v. 140, p. 110190, 2020.

PAPINENI, Kishore et al. Bleu: a method for automatic evaluation of machine translation. In: . [S.l.: s.n.], 2002. p. 311–318.

PARK, Seo Young et al. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). **Korean Journal of Radiology**, Korean Society of Radiology, v. 22, n. 10, p. 1697, 2021.

PATEL, Anita; JERNIGAN, Daniel B et al. Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak—united states, december 31, 2019–february 4, 2020. **Morbidity and mortality weekly report**, Centers for Disease Control and Prevention, v. 69, n. 5, p. 140, 2020.

PENCINA, Michael J; D'AGOSTINO, Ralph B. Evaluating discrimination of risk prediction models: the c statistic. **Jama**, American Medical Association, v. 314, n. 10, p. 1063–1064, 2015.

PHILIPP, WS Kopper et al. Deeppamm: Deep piecewise exponential additive mixed models for complex hazard structures in survival analysis. **arXiv preprint arXiv:2202.07423**, 2022.

PIRET, Jocelyne; BOIVIN, Guy. Pandemics throughout history. **Frontiers in microbiology**, Frontiers, p. 3594, 2021.

RAHMAN, Md Tanvir et al. Zoonotic diseases: etiology, impact, and control. **Microorganisms**, MDPI, v. 8, n. 9, p. 1405, 2020.

RAWAT, Waseem; WANG, Zenghui. Deep convolutional neural networks for image classification: A comprehensive review. **Neural Computation**, v. 29, n. 9, p. 2352–2449, 2017. Available at: <https://doi.org/10.1162/neco_a_00990>.

ROBBINS, Herbert; MONRO, Sutton. A stochastic approximation method. **The annals of mathematical statistics**, JSTOR, p. 400–407, 1951.

ROSENBLATT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUSSELL, Stuart; NORVIG, Peter. **Artificial intelligence: a modern approach**. [S.l.: s.n.], 2020.

RUUSKA, Salla et al. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. **Behavioural Processes**, v. 148, p. 56 – 62, 2018. ISSN 0376-6357.

SAEED, Numan et al. Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. **arXiv preprint arXiv:2403.10603**, 2024.

SAIT, Unais et al. **Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-Rays).** 2020.

SANDLER, Mark et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 4510–4520.

SANTURKAR, Shibani et al. How does batch normalization help optimization? **Advances in neural information processing systems**, v. 31, 2018.

SHANG, Yunfeng; LI, Haiwei; ZHANG, Ren. Effects of pandemic outbreak on economies: evidence from business history context. **Frontiers in Public Health**, Frontiers, v. 9, p. 146, 2021.

SHI, Heshui et al. Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study. **The Lancet infectious diseases**, Elsevier, v. 20, n. 4, p. 425–434, 2020.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SIMPSON, Scott et al. Radiological society of north america expert consensus statement on reporting chest ct findings related to covid-19. endorsed by the society of thoracic radiology, the american college of radiology, and rsna. **Journal of thoracic imaging**, Wolters Kluwer Health, 2020.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information processing & management**, Elsevier, v. 45, n. 4, p. 427–437, 2009.

SRIVASTAVA, Nitish et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

STRUYF, Thomas et al. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19. **Cochrane Database of Systematic Reviews**, John Wiley & Sons, Ltd, n. 2, 2021.

SUTTON, Richard S; BARTO, Andrew G. **Reinforcement learning: An introduction**. [S.l.]: MIT press, 2018.

SZEGEDY, Christian et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2017.

____. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2818–2826.

TAN, Mingxing; LE, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International conference on machine learning**. [S.l.], 2019. p. 6105–6114.

TONG, Rui; ZHU, Zhongsheng; LING, Jia. Comparison of linear and non-linear machine learning models for time-dependent readmission or mortality prediction among hospitalized heart failure patients. **Heliyon**, Elsevier, v. 9, n. 5, 2023.

TRICCO, Andrea C et al. A scoping review of rapid review methods. **BMC medicine**, BioMed Central, v. 13, n. 1, p. 1–15, 2015.

VASWANI, Ashish et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VERGANO, Marco et al. Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the italian perspective during the covid-19 epidemic. **Critical Care**, Springer, v. 24, n. 1, p. 165, 2020.

WAN, Yuchai; ZHOU, Hongen; ZHANG, Xun. An interpretation architecture for deep learning models with the application of covid-19 diagnosis. **Entropy**, Multidisciplinary Digital Publishing Institute, v. 23, n. 2, p. 204, 2021.

WANG, Dawei et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. **Jama**, American Medical Association, v. 323, n. 11, p. 1061–1069, 2020.

WANG, Jun et al. icovid: interpretable deep learning framework for early recovery-time prediction of covid-19 patients. **NPJ digital medicine**, Nature Publishing Group, v. 4, n. 1, p. 1–13, 2021.

WANG, Ping; LI, Yan; REDDY, Chandan K. Machine learning for survival analysis: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 51, n. 6, p. 1–36, 2019.

WANG, Quan Qiu et al. Covid-19 risk and outcomes in patients with substance use disorders: analyses from electronic health records in the united states. **Molecular psychiatry**, Nature Publishing Group, v. 26, n. 1, p. 30–39, 2021.

WENHAM, Clare et al. Preparing for the next pandemic. **BMJ**, BMJ Publishing Group Ltd, v. 373, 2021. Available at: <https://www.bmj.com/content/373/bmj.n1295>.

WIEGREBE, Simon et al. Deep learning for survival analysis: a review. **Artificial Intelligence Review**, Springer, v. 57, n. 3, p. 65, 2024.

WIERSINGA, W Joost et al. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19): a review. **Jama**, American Medical Association, v. 324, n. 8, p. 782–793, 2020.

WONG, Ho Yuen Frank et al. Frequency and distribution of chest radiographic findings in patients positive for covid-19. **Radiology**, Radiological Society of North America, v. 296, n. 2, p. E72–E78, 2020.

WU, Yonghui et al. **Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation**. 2016.

WU, Zhiqiang et al. Sars-cov-2's origin should be investigated worldwide for pandemic prevention. **The Lancet**, Elsevier, v. 398, n. 10308, p. 1299–1303, 2021.

XU, Kelvin et al. Show, attend and tell: Neural image caption generation with visual attention. In: PMLR. **International conference on machine learning**. [S.l.], 2015. p. 2048–2057.

XU, Ming et al. Accurately differentiating between patients with covid-19, patients with other viral infections, and healthy individuals: multimodal late fusion learning approach. **Journal of Medical Internet Research**, JMIR Publications Inc., Toronto, Canada, v. 23, n. 1, p. e25535, 2021.

XUE, Youyuan et al. Generating radiology reports via auxiliary signal guidance and a memory-driven network. **Expert Systems with Applications**, Elsevier, v. 237, p. 121260, 2024.

YAMGA, Eric et al. Interpretable clinical phenotypes among patients hospitalized with covid-19 using cluster analysis. **Frontiers in Digital Health**, Frontiers, v. 5, p. 1142822, 2023.

YAN, Bin et al. Prior guided transformer for accurate radiology reports generation. **IEEE Journal of Biomedical and Health Informatics**, IEEE, v. 26, n. 11, p. 5631–5640, 2022.

YU, Chao et al. Reinforcement learning in healthcare: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, v. 55, n. 1, p. 1–36, 2021.

ZEISER, Felipe André et al. Evaluation of convolutional neural networks for covid-19 classification on chest x-rays. In: BRITTO, André; DELGADO, Karina Valdivia (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2021. p. 121–132. ISBN 978-3-030-91699-2.

____. Breast cancer intelligent analysis of histopathological data: A systematic review. **Applied Soft Computing**, Elsevier, p. 107886, 2021.

____. Segmentation of masses on mammograms using data augmentation and deep learning. **Journal of Digital Imaging**, Springer, p. 1–11, 2020.

____. First and second covid-19 waves in brazil: A cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality. **The Lancet Regional Health-Americas**, Elsevier, v. 6, p. 100107, 2022.

ZHANG, Dingwen et al. Weakly supervised object localization and detection: a survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, 2021.

ZHANG, Haotian; ZHANG, Lin; JIANG, Yuan. Overfitting and underfitting analysis for deep learning based end-to-end communication systems. In: IEEE. **2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)**. [S.l.], 2019. p. 1–6.

ZHANG, Man et al. A survey of semi-and weakly supervised semantic segmentation of images. **Artificial Intelligence Review**, Springer, v. 53, n. 6, p. 4259–4288, 2020.

ZHONG, Zhusi et al. De-biased disentanglement learning for pulmonary embolism survival prediction on multimodal data. **IEEE Journal of Biomedical and Health Informatics**, IEEE, 2024.

ZHOU, Zhi-Hua. A brief introduction to weakly supervised learning. **National science review**, Oxford University Press, v. 5, n. 1, p. 44–53, 2018.

ZHU, Na et al. A novel coronavirus from patients with pneumonia in china, 2019. **New England journal of medicine**, Mass Medical Soc, 2020.

ZUIDERVELD, Karel. Graphics gems iv. In: HECKBERT, Paul S. (Ed.). **Graphics Gems**. San Diego, CA, USA: Academic Press Professional, Inc., 1994. chap. Contrast Limited Adaptive Histogram Equalization, p. 474–485.