UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

AI-DRIVEN PREDICTION OF LUNG CANCER PATIENT RESPONSES TO IMMUNOTHERAPY

NOEMI CICALA

Thesis supervisor
MIQUEL FERRIOL GALMÉS

Thesis co-supervisor
DAMIANO PIOVESAN

Tutor: LUIS ANTONIO BELANCHE MUÑOZ (Department of Computer Science)

Degree
Master's Degree in Data Science

Master's thesis

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

22/10/2025

Acknowledgements

I would like to sincerely thank Dr. Miquel Ferriol-Galmés for his continuous guidance, patience, and support throughout this project. His careful attention to every question and uncertainty, combined with his expertise and professionalism, has been invaluable in shaping this thesis and in helping me grow as a researcher. I feel truly fortunate to have had the opportunity to work under his mentorship.

I am also deeply grateful to the doctors and staff at Hospital Clínic de Barcelona for granting me access to their clinical datasets, an opportunity that was both rare and essential for this work. Beyond providing the data, they allowed me to present my results and offered insightful feedback, making the collaboration between clinicians and data scientists, which is at the heart of this thesis, a tangible and enriching experience.

Abstract

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, and the advent of immunotherapy has significantly improved survival outcomes for a subset of patients. However, accurately predicting treatment response remains a major clinical challenge due to the complexity and heterogeneity of patient profiles.

In this thesis, we explore an AI-driven predictive framework for modeling and interpreting patient responses to immunotherapy, based on a retrospective cohort including demographic, clinical, and molecular features. Multiple clinically relevant endpoints are investigated, encompassing both survival and response indicators: **Overall Survival (OS)** status and duration, **Progression-Free Survival (PFS)**, **Time to Progression (TTP)**, **Best Overall Response (BOR)** and its modified variants, as well as **Responder Quality Groups** derived from both raw and log-transformed TTP.

Three main modeling paradigms, **classical machine learning**, **conventional neural networks**, and **advanced neural architectures**, are systematically compared. Through targeted hyperparameter tuning and regularization, we mitigate overfitting and show that model complexity should be adapted to each endpoint rather than uniformly increased.

Model interpretability is ensured through **SHAP-based feature** attributions, which identify biologically plausible predictors of treatment outcome. Finally, ethical and sustainability aspects are discussed, highlighting the social and economic value of AI-assisted decision support in clinical oncology, despite the environmental costs of model training.

Overall, this work demonstrates that interpretable and endpoint-specific AI models can effectively complement clinical expertise, improving patient stratification and promoting a more personalized and sustainable approach to lung cancer immunotherapy.

Contents

List of Abbreviations

Clinical and Biological Abbreviations

Abbreviation	Definition
ALK	Anaplastic lymphoma kinase (molecular mutation)
ASR	Age-standardized incidence rate
BDS	Baseline sum of diameters
BMI	Body mass index (anthropometric measure of weight/height)
BOR	Best Overall Response
CR	Complete response
DC	Dendritic cells
dNLR	Derived neutrophil-to-lymphocyte ratio (biomarker)
EGFR	Epidermal growth factor receptor (molecular mutation)
ICI	Immune checkpoint inhibitors (immune-modulating agents)
KRAS	Kirsten rat sarcoma virus (molecular mutation)
LDCT	Low-dose computed tomography
LDH	Lactate dehydrogenase (biomarker)
LIPI	Lung immune prognostic index (biomarker)
MHC	Major histocompatibility complex
NK	Natural killer cells
NSCLC	Non-small cell lung cancer
OS	Overall Survival
PD	Progressive disease
PD-1 / PD-L1	Programmed cell death protein 1 / its ligand
PFS	Progression-Free Survival
PR	Partial response
SCLC	Small cell lung cancer
SD	Stable disease
SOD	Sum of diameters
TAA	Tumor-associated antigens
TMB	Tumor mutation burden
TME	Tumor microenvironment
TNM	Tumor–Node–Metastasis staging system
TSA	T cell-mediated cytotoxicity targeting tumor-specific antigens
TTP	Time to Progression

Computational and Machine Learning Abbreviations

Abbreviation	Definition
AI	Artificial intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic curve
CDSS	Clinical decision support system
CNN	Convolutional neural network
DL	Deep learning
FD	Federated learning
HL	Hidden layer
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine learning
MSE	Mean Squared Error
SVC / SVR	Support Vector Classifier / Support Vector Regressor
sMAPE	Symmetric Mean Absolute Percentage Error
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic minority over-sampling technique
XAI	Explainable artificial intelligence

List of Figures

List of Tables

1 Introduction

1.1 Context

Lung cancer remains the leading cause of cancer-related mortality globally, posing a significant challenge to public health. The Global Cancer Statistics 2022 indicate that lung cancer is the most frequently diagnosed cancer worldwide, with 2.48 million new cases, and the deadliest, resulting in approximately 1.82 million deaths in the same year [1]. The corresponding age-standardized incidence rate (ASR, World[1]) was 23.6 per 100,000 and the mortality rate was 16.8 per 100,000, underscoring its status as the most lethal malignancy across diverse populations.

These statistics highlight the critical necessity for more efficacious and enduring therapeutic strategies. Conventional treatments, including surgery, chemotherapy, and radiation therapy, frequently exhibit restricted effectiveness in advanced disease stages. Consequently, there is an urgent demand for novel therapeutic modalities capable of eliciting sustained and durable responses.

In recent years, research [2] has significantly advanced our understanding of the immune system's role in cancer, leading to major breakthroughs in cancer **immunotherapy**. Immunotherapy aims to re-engage the immune system, particularly T cell–mediated cytotoxicity targeting tumor-specific (TSA) and tumor-associated antigens (TAA), to selectively eliminate cancer cells. Immune-modulating agents may also increase the presence of tumor-specific antibodies, natural killer cells (NKs), dendritic cells (DCs(, macrophages and cytokines, further enhancing anti-tumor activity. One of the most impactful developments [3] has been the introduction of immune checkpoint inhibitors (ICIs), which have reshaped the treatment paradigm for lung cancer. These drugs function by blocking inhibitory pathways, such as the PD-1/PD-L1 axis, that tumors exploit to escape immune attack. By removing these "brakes" on the immune system, ICIs restore T-cell function and promote sustained immune activity against cancer cells. ICIs have demonstrated significant clinical benefits in non-small cell lung cancer (NSCLC), offering durable responses in a subset of patients, often regardless of tumor histology or mutational status. Their emergence has shifted immunotherapy from an experimental option to a mainstream treatment strategy in lung cancer. However, not all patients respond, and the variability in treatment outcomes remains a major clinical challenge.

Understanding which patient-specific variables influence immunotherapy effectiveness is

[1] An ASR is a weighted mean of the age-specific rates; the weighting is based on the population distribution of a standard population. The most frequently used standard population is the World (W) Standard Population. The calculated incidence or mortality rate is then called the age-standardized incidence or mortality rate (W), and is expressed per 100 000 person-years.

therefore a pressing priority. Identifying these factors would enable more precise patient selection and the development of personalized treatment strategies, maximizing benefit while minimizing unnecessary exposure to ineffective therapies.

In this context, the central objective of this thesis is to develop artificial intelligence (AI)-based models capable of predicting individual patient responses to immunotherapy in lung cancer. By leveraging multi-dimensional clinical and molecular data, this work aims not only to build accurate predictive frameworks, but also to identify the most informative variables associated with treatment outcomes. Through this dual approach, prediction and interpretability, this research contributes to the growing field of precision oncology, with the ultimate goal of supporting clinicians in making better-informed, personalized treatment decisions for lung cancer patients.

1.2 Motivation

The healthcare industry is currently undergoing a significant transformation, driven by increasing costs, resource constraints, and the growing complexity of patient care. Healthcare systems worldwide are grappling with numerous challenges, including unequal access to medical services, rising demand due to aging populations, clinician shortages, and fragmented information. These systemic issues have been further exacerbated by global health crises, such as the COVID-19 pandemic, which revealed critical weaknesses in resource allocation, diagnostic capacity, and data sharing mechanisms [4].

Concurrently, traditional treatment approaches for lung cancer, such as surgery, chemotherapy and radiation therapy, continue to demonstrate limited success, particularly in advanced-stage cases. NSCLC and SCLC often present at later stages when curative interventions are no longer feasible. Although low-dose computed tomography (LDCT) is currently used for screening, its adoption remains low and is hindered by a high rate of false positives, radiation risks, and a lack of infrastructure for large-scale implementation. Chemotherapy and targeted therapies have improved progression-free survival in some cases, but they are still associated with toxicity, limited efficacy, and a lack of sustained responses. In the case of SCLC, despite initial responsiveness to treatment, rapid relapse and poor survival outcomes remain the norm, underscoring the urgent need for more effective, personalized treatment modalities [2].

Simultaneously, the biomedical field is experiencing a revolution catalyzed by advances in AI, big data, and machine learning. AI is increasingly being integrated into domains such as diagnostic imaging, genomics, drug discovery, and robotic surgery. One of AI's most promising roles is in clinical decision support, where it can process vast amounts of heterogeneous data—including clinical records, genomic profiles, and imaging data—far

beyond the capacity of any human clinician. Moreover, AI systems are not subject to the same cognitive and cultural biases as human decision-makers, offering more consistent and scalable insights across diverse patient populations.

Recent developments further demonstrate AI's growing impact across all stages of cancer therapy [5]. For example, studies have shown that large language models such as GPT-4 can identify accurate diagnoses in complex medical records in nearly 40% of cases, and propose relevant differential diagnoses in over two-thirds of cases, according to a 2023 study published in JAMA [6]. In parallel, biomedical vision-language models like Microsoft's LLaVA-Med have demonstrated the capability to infer pathological conditions from imaging data such as CT and X-rays. In the field of cervical cancer screening, AI-assisted diagnostic tools have already achieved interpretation accuracy exceeding 99% on negative slides and detection rates above 99.9% on positive abnormalities, according to real-world applications by Huawei in collaboration with KingMed Diagnostics. Beyond diagnostics, AI is beginning to transform oncology drug development. It plays a growing role in the design of small molecule inhibitors, optimizing manufacturing pipelines, and predicting therapeutic responses. The fusion of oncology big data and deep learning is setting the stage for a new paradigm in personalized medicine—one that is not only reactive, but also anticipatory and highly individualized.

In the context of lung cancer, and particularly in the era of immunotherapy, AI offers a compelling opportunity. While immune checkpoint inhibitors have significantly improved outcomes for a subset of patients, response rates vary widely, and we still lack reliable tools to predict which patients will benefit. Given the complexity and heterogeneity of tumor biology, traditional statistical methods fall short in capturing non-linear, multi-dimensional relationships within the data. This is where AI becomes indispensable. With the ability to model intricate patterns across thousands of patient variables, AI can help uncover novel predictors of immunotherapy response, stratify patients more effectively, and ultimately guide precision oncology. Recent studies [7] have demonstrated the utility of AI-driven models that integrate radiology, pathology, genomics, and proteomics data to predict key biomarkers such as PD-L1 expression, tumor mutation burden (TMB), and characteristics of the tumor microenvironment (TME). These biomarkers are critical in determining eligibility and likely response to immunotherapy. As such, machine learning is paving the way toward a "digital biopsy" approach—capable of providing a comprehensive, non-invasive assessment that may eventually replace traditional single-modality evaluations. The convergence of these technologies enables clinicians not only to identify patients most likely to benefit from immune checkpoint inhibitors but also to anticipate potential side effects, enhancing patient safety and treatment personalization. In this vision, AI is not simply a tool for automation, but a catalyst for a new era of data-driven, adaptive, and patient-specific cancer care.

1.3 Objectives

The overarching objective of this thesis is to explore the potential of AI–based models for predicting patient responses to immunotherapy in lung cancer, with a particular focus on understanding the impact of different patient-specific variables. Rather than aiming exclusively at achieving optimal predictive accuracy, this work emphasizes the interpretability of results and the identification of patterns that may shed light on which factors are more (or less) influential in shaping treatment outcomes.
To achieve this general aim, the thesis pursues the following specific objectives:

- **Data modeling**: to construct a structured representation of patient data by integrating available clinical and demographic information relevant to immunotherapy outcomes.

- **Comparative modeling** : to design, train, and evaluate a range of machine learning (ML) and deep learning (DL) models, with the dual purpose of comparing their predictive performance and assessing their suitability with respect to interpretability and clinical applicability.

- **Feature importance** and **variable impact**: to analyze and interpret the role of patient-specific variables in determining immunotherapy responses, thereby providing greater awareness of which factors appear more (or less) influential within the available dataset.

- **Interpretability** and **explainability**: to apply interpretable AI techniques that not only clarify model decisions but also highlight clinically meaningful patterns, ensuring that results remain understandable and actionable.

- **Critical assessment** and **clinical perspective**: to discuss the limitations of the current dataset and identify gaps where additional variables could substantially improve predictive power. This thesis therefore aims to establish a methodological foundation for future studies that may leverage richer and more comprehensive data to build more robust predictive frameworks for precision oncology.

Through this approach, the thesis contributes both to the understanding of variable importance in immunotherapy outcomes and to the longer-term goal of enabling more precise and personalized treatment strategies for lung cancer patients.

2 Background

2.1 Lung Cancer and Current Treatment Landscape

2.1.1 Epidemiology and Risk Factors

Lung cancer remains the leading cause of cancer-related mortality worldwide, with an estimated 1.8 million deaths in 2020, accounting for 18% of all cancer deaths. The primary risk factor is tobacco smoking, but non-smokers can also develop lung cancer due to factors such as exposure to secondhand smoke, occupational hazards (e.g., asbestos, radon), air pollution, hereditary cancer syndromes, and previous chronic lung diseases [8].

2.1.2 Histological Subtypes

Lung cancer is classified into two main histological subtypes [8]:

- NSCLC: Comprising approximately 85% of cases, NSCLC includes subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. It generally exhibits slower progression compared to SCLC.
- SCLC: Representing about 15% of cases, SCLC is characterized by rapid growth and early metastasis. It is strongly associated with smoking and has a poorer prognosis.

2.1.3 Diagnostic Approaches

Diagnosis involves a combination of clinical evaluation, imaging techniques, bronchoscopy, biopsy for histopathological examination, and molecular testing to identify specific genetic mutations or biomarkers [8].

2.1.4 Treatment Modalities

Treatment strategies are tailored based on cancer type, stage and patient health status[8]:

- **Surgery**: Indicated for early-stage NSCLC when the tumor is localized and resectable.
- **Chemotherapy**: Utilized in both NSCLC and SCLC, often as adjuvant or neoadjuvant therapy, and for advanced stages.
- **Radiotherapy**: Applied in localized disease or as palliative care in advanced stages.
- **Targeted Therapy**: Involves drugs that target specific genetic mutations found in some NSCLC tumors.
- **Immunotherapy**: Emerging as a significant treatment option, particularly for advanced NSCLC.

2.1.5 Immunotherapy in Lung Cancer

Recent advances in immunology have significantly improved our understanding of how the immune system interacts with cancer, paving the way for immunotherapy as a transformative treatment for lung cancer. The main goal of cancer immunotherapy is to activate or re-activate the immune system (particularly TSA and TAA) to selectively target and eliminate cancer cells. Immune-modulatory agents may also enhance anti-tumor responses by increasing tumor-specific antibodies, NK cells, DCs, macrophages, and cytokines in the blood plasma [2].

Historically, lung cancer was considered less responsive to immunotherapy due to mechanisms of immune evasion, including reduced immunosurveillance, secretion of immunosuppressive cytokines, and downregulation of major histocompatibility complex (MHC) expression. However, recent technical advances have uncovered the molecular basis of lung cancer immunogenicity, allowing the development of multiple immunotherapeutic strategies.

Key types of lung cancer immunotherapy include:

- **ICIs**: Monoclonal antibodies targeting inhibitory pathways such as PD-1/PD-L1 and CTLA-4, restoring T-cell function and promoting sustained anti-tumor activity.
- **Therapeutic vaccines and autologous cellular therapies**: Designed to elicit tumor-specific immune responses.
- **Immune modulators**: Agents that enhance the activity of NK cells, DCs, and cytokines to reinforce anti-tumor immunity.

Each therapeutic approach has distinct advantages and limitations, and current research increasingly focuses on combination strategies that integrate ICIs with chemotherapy, targeted therapy, or other immunomodulatory agents to improve efficacy and durability of responses.

2.2 Clinical endpoints in oncology

2.2.1 Standard Outcome Variables

In order to evaluate and predict patient outcomes after the initiation of first-line immunotherapy in advanced lung cancer, several clinical endpoints, used in clinical trials and real-world studies, each capturing different aspects of patient benefit and treatment efficacy, have been considered.
The main outcomes are briefly introduced below.

Overall Survival (OS) — Status and Duration

OS is the time from a prespecified origin (typically randomization or treatment start) to death from any cause. Patients alive at analysis are censored at the last known alive date [9].
It can be expressed both as :

- a binary variable **OS Status**, coded as 1 if the patient is dead, 0 if alive at censoring
- and as a time-to-event variable **OS Duration**, the time from origin date to death or censoring (expressed in days or months).

Recognized as the most reliable endpoint in oncology, it is typically selected as the preferred endpoint when survival can be adequately evaluated[10].

Progression-Free Survival (PFS)

PFS is defined as the time from randomization until objective tumor progression or death, whichever occurs first, reflecting both tumor control and patient survival. Tumor progression is determined according to the response assessment criteria specified in the clinical trial (most frequently RECIST 1.1) [11].
As such, it provides an intermediate endpoint that can be evaluated earlier than OS.

Best Overall Response (BOR)

BOR is defined as the best categorical treatment response achieved by a patient during study participation, assessed according to RECIST 1.1 [2].
Responses are categorized as shown in the Table 3 [12]:

[2]Response Evaluation Criteria in Solid Tumors (RECIST) are a set of published rules that provide an objective measurement of tumor burden in response to conventional systemic therapy, characterizing lesions as measurable vs non-measurable and target vs non-target

Response Category	Criteria
Complete Response (CR)	Requires all of: • Disappearance of all target and non-target lesions • Pathological lymph nodes reduced to <10 mm in short axis • No new lesions
Partial Response (PR)	Requires all of: • At least 30% decrease in sum of diameters (SOD) of target lesions compared to baseline sum diameters (BSD) • Non-progressive disease of non-target lesions • No new lesions
Stable Disease (SD)	Not meeting criteria for PD or PR
Progressive Disease (PD)	Either one of: • Any new lesions • At least 20% relative and 5 mm absolute increase of SOD of target lesions compared to the smallest SOD ever recorded

Table 3: RECIST 1.1 response categories and their criteria

Time to Progression (TTP)

TTP is the duration between treatment initiation and the first documented progression, but unlike PFS it does not count death without progression as an event [9]. It isolates the event of progression and can provide insights into tumor dynamics independently of mortality.

2.2.2 Derived and Adapted Outcome Variables

While the endpoints described above (OS, PFS, BOR, TTP) represent the standard clinical measures of efficacy in oncology, several methodological challenges emerged when applying them directly in the machine learning setting of this thesis. These included

class imbalance, small sample sizes for certain categories, and the need for more practically informative outcome groupings. To address these issues, additional derived variables were constructed, either by merging categories, redefining outcome classes, or applying data-driven grouping methods. Examples include:

Modified BOR categories

The CR category was extremely rare in the dataset. To mitigate sparsity and improve model robustness, CR was either removed or merged with PR, given their conceptual similarity as indicators of tumor shrinkage.

Responder quality groups

Based on TTP, patients were stratified into groups of good, medium, and poor responders using clustering techniques. This transformation provided a more interpretable stratification of treatment benefit, closer to clinical decision-making needs.

These adaptations were not intended to replace standard clinical definitions but to enable a more meaningful and feasible modeling process in the context of the available data.

2.2.3 Comparative Analysis of Endpoints

The selection of a specific endpoint has important methodological implications, influencing both the clinical interpretation and the appropriate machine learning formulation. The Table 4 summarizes the main strengths and weaknesses of each outcome.

Endpoint	**Strengths**	**Limitations**	**ML task implications**
OS	Gold-standard endpoint	Requires long follow-up	Regression (**OS duration**) or Binary classification (**OS status**)
PFS	Shorter follow-up than OS	Affected by criteria choice	Regression
BOR	Intuitive categories	Ignores duration of response	Multiclass classification
TTP	Focused on progression dynamics	Risk of underestimating poor outcomes	Regression
Modified BOR	Reduces sparsity and improves class balance; CR/PR merging retains conceptual meaning	Loss of granularity; CR-specific information diluted or lost	Multiclass classification
Responder groups	Clinically intuitive; balances classes; captures heterogeneity	Non-standard endpoint; sensitive to clustering choices; loss of temporal resolution	Multiclass classification

Table 4: Strengths, limitations and ML task implications of both standard and derived oncology endpoints.

3 State of the art

3.1 Artificial intelligence in oncology

AI has rapidly emerged as a transformative tool in modern medicine, offering innovative solutions to complex problems in clinical decision-making, diagnostics, and personalized care. In the field of oncology, where the management of cancer requires the integration of vast amounts of clinical, pathological, and imaging data, AI holds particular promise [13]. At its core, AI encompasses a broad range of computational techniques, including ML and DL, that enable systems to recognize patterns, learn from large datasets, and generate predictions that support physicians in patient care [14].

3.1.1 History of AI in Oncology

The evolution of AI in oncology has followed a trajectory shaped by advances in computing power, data availability, and algorithmic innovation. In its early stages, during the 1960s–1980s, AI in medicine largely revolved around expert systems. These systems, such as MYCIN and INTERNIST-1, used rule-based logic to support clinical decision-making in infectious diseases and internal medicine. While they demonstrated the potential of computer-assisted diagnosis, their rigid reliance on handcrafted rules limited their applicability to oncology, where the heterogeneity of tumor biology exceeded the representational power of such systems [15].

A significant shift occurred in the early 2000s with the emergence of machine learning, allowing models to learn patterns directly from data rather than relying on predefined rules. The availability of digitized medical imaging and electronic health records provided fertile ground for data-driven methods. The introduction of DL and convolutional neural networks (CNNs) around 2012 marked a turning point, enabling automated image recognition at a level comparable to human experts. These advances quickly permeated oncology, where imaging plays a central role in diagnosis, staging, and treatment response monitoring [16].

Recent reviews emphasize that the convergence of big data, omics technologies, and AI has led to a new paradigm—computational oncology, where AI algorithms integrate imaging, molecular, and clinical data to predict outcomes and guide personalized therapies [17].

3.1.2 Opportunities and Promises of AI in Oncology

AI offers unprecedented opportunities across the entire cancer care continuum, from early detection and diagnosis to prognosis, treatment optimization, and follow-up.

- **Early Detection and Diagnosis**
  AI algorithms, particularly DL-based models, have demonstrated human-level performance in detecting malignancies in medical imaging. In breast, lung, and skin cancer, AI systems can identify suspicious lesions and classify tumor subtypes with remarkable accuracy, potentially reducing diagnostic delays and human error [18]. Furthermore, pathology AI can analyze digitized whole-slide images to detect mitotic figures, grade tumors, and predict molecular alterations directly from histopathological images.

- **Predictive and Prognostic Modeling**
  By integrating multimodal data—including genomic, transcriptomic, radiomic, and clinical variables—AI enables the development of predictive models for survival, recurrence, and treatment response. This supports personalized medicine by identifying which patients are most likely to benefit from a specific therapy [19].

- **Clinical Decision Support Systems (CDSS)**
  AI-driven CDSS assist oncologists in treatment selection and drug dosing by synthesizing complex datasets into actionable insights. For instance, recent studies have evaluated the "data readiness" of oncology datasets for AI-driven decision-making in melanoma and skin cancer treatment [20].

- **Federated and Collaborative Learning (FL)**
  To overcome privacy and data fragmentation issues, FL has emerged as a paradigm allowing AI models to be trained across multiple institutions without data sharing. FL in oncology has shown potential for improving model generalizability across breast, lung, and prostate cancer datasets [21].

- **Explainable AI (XAI)**
  As AI systems become increasingly complex, explainability becomes crucial to ensure clinician trust and regulatory compliance. XAI methods, such as SHAP, LIME, and Grad-CAM, are increasingly used to visualize and interpret model decisions, particularly in breast cancer diagnosis and radiological applications [22].

3.1.3 Limitations and Challenges of AI in Oncology

While AI holds immense promise in oncology, its clinical implementation faces several critical challenges. These barriers span technical, regulatory, and ethical domains, highlighting that AI is not yet a fully mature or universally deployable technology.

- **Data Quality, Quantity, and Representativeness**
  High-quality labeled datasets remain scarce in many cancer domains. Small, unbalanced, or monocentric datasets lead to overfitting and poor model generalization

[18]. Furthermore, clinical data are often heterogeneous—spanning structured EHR entries, imaging archives, and unstructured physician notes.

- **Bias and Fairness**
  AI models may inadvertently learn and propagate existing biases in healthcare systems, underperforming in underrepresented populations due to imbalanced training data [23, 24]. Ensuring fairness across demographic and socioeconomic groups remains an ongoing challenge in oncology.

- **Explainability and Transparency**
  Most high-performing DL models are "black boxes," providing little insight into the reasoning behind predictions. Lack of interpretability poses barriers to clinical trust and adoption [17].

- **Clinical Validation and Generalizability**
  Many studies demonstrating AI success in oncology are retrospective or based on limited datasets. When deployed in real-world clinical settings, model performance often declines. Prospective, multicenter clinical trials are urgently needed to validate these tools [18].

- **Ethical, Legal, and Regulatory Challenges**
  AI raises questions regarding accountability, patient consent and data governance. The legal and ethical framework for AI-assisted clinical decisions remains under development [25].

- **Infrastructure and Workforce Limitations**
  Integrating AI systems into clinical workflows requires digital infrastructure, data harmonization and clinician training. Many healthcare systems, particularly in low-resource settings, lack these prerequisites [26].

3.2 A focus on AI in Immunotherapy

Immunotherapy, particularly ICIs targeting the PD-1/PD-L1 axis, has transformed the therapeutic landscape of NSCLC. However, only a subset of patients derive durable benefit and immune-related toxicities can be substantial; therefore, accurate patient selection remains a critical unmet need. AI has been increasingly investigated as a tool to improve prediction of response, toxicity, and optimal treatment sequencing in lung cancer immunotherapy by integrating imaging, pathology, genomic, and clinical data modalities [27, 28].

Biomarkers and data modalities

Current FDA-approved or clinically used biomarkers for ICI selection in NSCLC include tumor PD-L1 immunohistochemistry and, in some contexts, TMB; nevertheless, both biomarkers have limited sensitivity and specificity. AI approaches aim to either enhance existing biomarker assessment (for instance, automated scoring of PD-L1 or digital pathology signatures) or derive novel, composite biomarkers from multi-modal data (radiomics, pathomics, genomics, transcriptomics, and clinical variables) that better stratify benefit and risk [28].

Radiomics, quantitative feature extraction from CT, PET/CT or MRI, has been widely studied in NSCLC, showing associations with tumor immune microenvironment, PD-L1 expression, and outcomes after ICIs. Several studies and reviews report that CT-based radiomic signatures can predict response and progression-free survival in ICI-treated cohorts, although heterogeneity in study design and feature sets remains an issue [29, 27, 30]. Digital pathology (H&E whole-slide image analysis and immunohistochemistry image analysis) paired with deep learning has also demonstrated promise in predicting ICI response and extracting surrogate molecular features without explicit genomic testing [31].

Multi-omics integration, combining radiomics, pathomics and molecular profiles, is an active area: machine learning models trained across modalities can capture complementary signals (e.g., radiologic heterogeneity + gene expression signatures) that improve predictive performance over single-modality models [28, 32].

AI methods and predictive performance

A broad range of AI/ML methods has been applied, from traditional machine learning classifiers (random forests, gradient boosting, SVM) on engineered radiomic features to end-to-end deep learning and ensemble models applied to images and multi-modal inputs. Recent multicenter deep learning studies and ensemble CT-based models reported encouraging discrimination for ICI benefit (AUCs and concordance indices often in the acceptable-to-good range), though reported metrics vary widely due to cohort differences and retrospective designs [32, 31, 33].

XAI techniques have been used to increase clinician interpretability of predictions and to highlight biologically plausible regions or features associated with response (e.g., peritumoral heterogeneity, necrosis patterns) [34].

Clinical evidence and validation

While many retrospective and single-center studies show promise, high-quality prospective evidence is still limited. Multicenter cohorts and external validation are increasingly reported: notable recent multicenter efforts have used pathology-based deep learning and CT radiomics applied to international datasets, demonstrating improved robustness and transportability compared to earlier single-center models [31, 27]. Systematic reviews and meta-analyses confirm the potential of imaging-based AI biomarkers but highlight methodological heterogeneity and frequent lack of full external validation [35, 34].

In summary, AI offers multiple avenues to improve the selection of NSCLC patients for immunotherapy, from enhanced biomarker quantification to integrated multi-modal predictors of benefit and harm. Translational success will depend on rigorous external validation, prospective evaluation, transparent reporting, and multidisciplinary collaboration between oncologists, radiologists, pathologists, data scientists, and regulatory stakeholders [27, 31, 28].

4 Methodology

4.1 Datasets Overview

This section introduces the datasets employed in this thesis work. All datasets were provided by *Hospital Clínic* of Barcelona and contain sensitive clinical information of patients diagnosed with lung cancer and treated with immunotherapy. Initially, three datasets, that we called Dataset A, Dataset B and Dataset C, were delivered and have been previously described in related works [36, 37, 38]. After the thesis work had already started, a fourth dataset was made available, representing a unified version of the initial three. In the following, each dataset is briefly described.

4.1.1 Dataset A, Dataset B and Dataset C

Table 5 summarizes the main characteristics of the three initial datasets provided by *Hospital Clínic*.

Dataset A [36]

It includes a large cohort (1485 patients, 89 variables) with a broad spectrum of variables: demographic and lifestyle factors, tumor characteristics and biomarkers (e.g., PD-L1, TMB), pre-therapy and immunotherapy data, detailed information on metastasis locations, treatment response, survival outcomes, and longitudinal blood tests (both baseline and cycle 2). Despite its richness, this dataset suffers from considerable missingness, especially in molecular variables.

Dataset B [37]

This dataset is more compact (431 patients and only 23 variables) and focuses on a core set of predictors: demographics, immunotherapy line, therapy response, survival data (OS, PFS), and basic biomarkers (LDH, dNLR, LIPI). Its structure is clean and almost free of missing values, which makes it highly suitable for predictive modeling despite its smaller sample size.

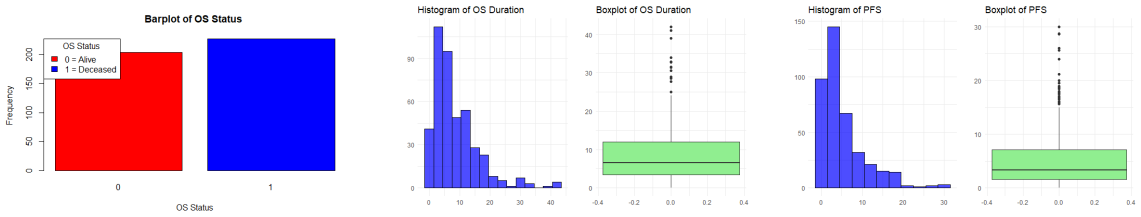
Dataset C [38]

It represents an intermediate case (930 patients, 148 variables), covering the entire clinical pathway from diagnosis to treatment lines, molecular profiling (EGFR, KRAS, ALK, etc.), metastasis details, therapy response, survival outcomes, and a wide range of laboratory measures. However, its high dimensionality, redundancy, and substantial missing data make preprocessing more challenging.

Dataset	Rows	Columns	Variable Groups	Missingness	Strengths	Weaknesses
A	1485	89	Demographics Tumor characteristics Pre-therapy Immunotherapy Metastasis Blood tests Survival	Moderate ≈ 33.5 %	Large sample size Rich set of variables	Considerable missing data
B	431	23	Demographics Immunotherapy Survival Therapy response Biomarkers	Low ≈ 0.9 %	Compact Minimal missing values	Smaller sample size Fewer variables
C	930	148	Demographics Tumor biomarkers Molecular profile Metastasis Treatments Blood tests Survival	High ≈ 40.6 %	Broad coverage of patient clinical journey	Redundancy Many missing values

Table 5: Comparative overview of the initial datasets provided by Hospital Clínic

Despite their differences in scope and level of detail, the three datasets share a common backbone of variables. All of them include basic demographic and lifestyle information (age, sex, smoking history), tumor histology, and essential clinical outcomes such as OS, in Fig. 1a and Fig. 1b, and PFS in Fig. 1c.



(a) OS Status (b) OS Duration (c) PFS

Figure 1: Visual representation of the three response variables contained in Dataset B.

Overall, while Dataset A and Dataset C provide broader coverage, Dataset B offers a more compact and consistent structure with minimal missingness and its variables are directly linked to survival and treatment response, justifying its selection for the subsequent analyses.

4.1.2 Dataset D

In addition to the three initial datasets, Hospital Clínic later provided a fourth dataset, resulting from the integration and harmonization of the previous sources. This unified dataset contains 1117 patients and 231 variables, thus representing the most comprehensive collection available. Compared to the initial datasets, it offers a broader coverage of the patient's clinical pathway, from diagnosis to multiple lines of therapy.
The dataset includes the following major groups of variables:

- **Personal information**: hospital of treatment, patient identifier, demographic data (sex, age, date of birth), anthropometric measures (weight, height, BMI), and smoking history (smoker, pack-years).
- **Diagnosis**: dates of diagnosis (both initial and advanced disease), TNM staging (8th edition), and performance status at diagnosis.
- **Metastasis**: detailed information on the presence and localization of metastases at diagnosis (lung, pleura, liver, brain, bone, etc.) and the number of affected sites.
- **Biopsy and molecular profile**: histology and histological patterns, biopsy site and timepoint, PD-L1 expression (percentage and categories), molecular testing and major oncogenic drivers (EGFR, KRAS, ALK, ROS1, MET).
- **Immunotherapy progression (iPD)**: sites and number of lesions at progression, presence of new lesions, and organ-specific progression.
- **Immune-related toxicity (iTOX)**: type, grade, timing, and treatment of immune-related adverse events, including corticosteroid use.
- **Laboratory tests**: blood counts and biochemical parameters (Hb, LDH, CRP, albumin, etc.) measured at different timepoints (pre-treatment, baseline, cycle 2, first CT scan).
- **Therapy lines and response**: information on up to six lines of treatment, including start and end dates, type of therapy (monotherapy vs. combination), best response, progression details, and assessment criteria.
- **Survival and follow-up**: OS, cause of death, date of last contact, and follow-up status.

This dataset represents the richest and most heterogeneous source of information, combining clinical, molecular, laboratory, and therapeutic data. However, its size and complexity also introduce challenges in terms of preprocessing, redundancy and missingness.

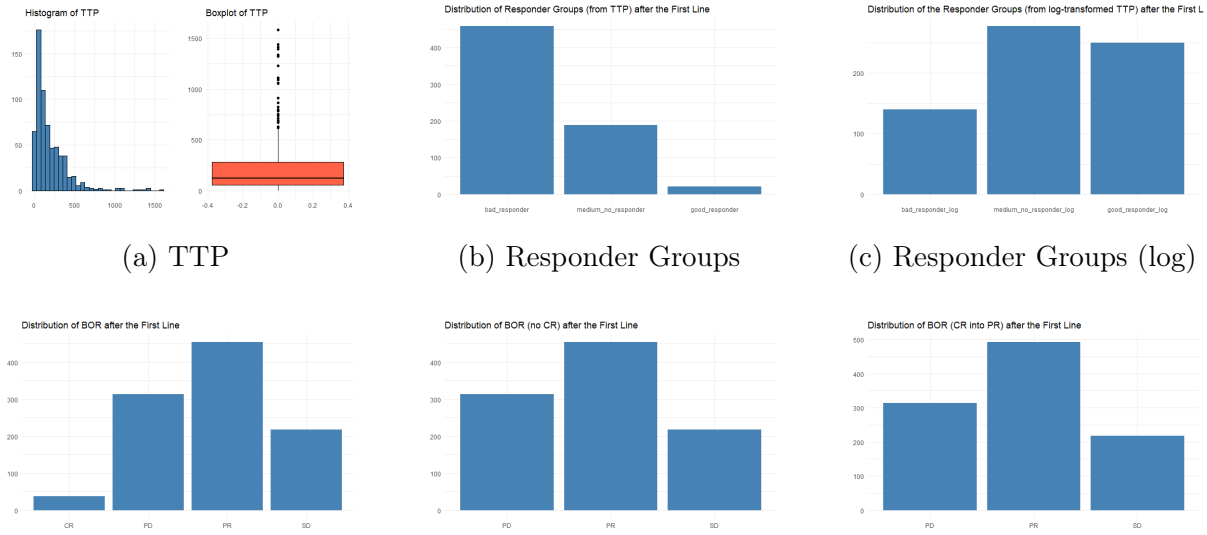
Exploratory Analysis

Following an extensive statistical, qualitative and quantitative analysis of the unified dataset, and after a preprocessing phase (including date standardization, data cleaning, handling of missing values and derivation of new features), a reduced set of variables, that appeared most relevant with respect to the objectives of the thesis, was identified.
The selected features cover four main dimensions:

- **Demographics and baseline status**: age, sex, smoking history, and performance status at advanced disease diagnosis.
- **Tumor burden and metastasis**: number of metastatic sites and organ-specific involvement.
- **Clinical and biological markers**: histology, baseline hematological values.
- **Treatment and outcomes**: progression during first-line immunotherapy, best response and categorical variables derived from progression.

As shown by the histogram in Fig. 2a of the TTP variable, its distribution is strongly right-skewed. This skewness leads to an unbalanced proportion when deriving categorical groups (via K-means clustering) as shown in Fig. 2b meant to represent different patient conditions. Although the resulting categories may be unevenly sized, the log-transformation was applied to better normalize the distribution, seen in Fig. 2c, and facilitate subsequent modeling, while still reflecting a realistic clinical scenario.

A similar imbalance is observed in the best response variable, as shown in Fig. 2d. In particular, the category CR is almost entirely absent compared to the other three categories (PR, SD, PD).
This class imbalance influenced the modeling strategies employed in later chapters: either CR was merged with the *Partial Response* category due to conceptual similarity, shown in Fig. 2f , or it was excluded from certain analyses to ensure more robust model performance, seen in Fig. 2e



(a) TTP (b) Responder Groups (c) Responder Groups (log)

(d) BOR (original) (e) BOR (no CR) (f) BOR (CR into PR)

Figure 2: Visual representation of the six response variables obtained from Dataset D.

4.1.3 Methodological Considerations from a Data Science Perspective

From a machine learning standpoint, the choice of endpoint has direct implications on model design, evaluation, and interpretability. Each type of predictive task introduces specific challenges and limitations.

Regression (OS duration, PFS and TTP)

Predicting survival times or progression durations is inherently more complex than classification. These variables are continuous, often right-skewed, and subject to censoring (patients who are alive or progression-free at last follow-up). Standard regression models cannot easily accommodate censored observations, which motivates the use of specialized survival models. Handling censoring correctly is essential to avoid biased estimates, but it increases model complexity and evaluation requires survival-specific metrics (C-index, time-dependent AUC), which are less intuitive than accuracy or F1-score.

Binary classification (OS status)

Reducing complex outcomes to a binary label can simplify modeling and interpretation, but this comes at a cost. In advanced lung cancer, death or disease progression unfortunately occurs for the vast majority of patients, leading to strong class imbalance and limiting the clinical usefulness of a "yes/no" prediction. Moreover, a binary endpoint provides little information on when the event will happen or what the expected clinical trajectory is, making it less valuable for treatment planning.

Multiclass classification (BOR, Modified BOR and Responder groups)

Multiclass problems are conceptually richer than binary tasks. Predicting whether a patient will achieve a complete or partial response, stable disease, or progression provides

clinically actionable insights and better reflects heterogeneity in treatment benefit. Compared to binary classification, multiclass predictions can serve as proxies for treatment durability and quality of life: for example, distinguishing between "stable disease" and "progression" conveys important prognostic differences. However, multiclass problems require more data per class, and rare outcomes (e.g., complete response) can be especially difficult to predict reliably.

4.1.4 Class Balancing Techniques

Given the inherent imbalance in several of the response variables, we considered two commonly adopted strategies to mitigate this issue.

Class weighting

Instead of altering the sample distribution, this approach adjusts the learning objective by assigning higher loss weights to the minority class. In practice, the loss function is reweighted so that errors on minority-class instances incur a larger penalty. Class weights are typically computed as the inverse of the class frequencies, which ensures that the optimization process pays proportionally more attention to rare outcomes.

SMOTE oversampling

Synthetic Minority Over-sampling Technique (SMOTE) generates new synthetic samples of the minority class by interpolating between existing instances in the feature space. In this way, the training set becomes more balanced, potentially improving the classifier's ability to recognize underrepresented outcomes. However, SMOTE may introduce noise or unrealistic samples, especially when applied to small and heterogeneous datasets.

4.2 Model architecture

4.2.1 Classical Machine Learning Models

To establish reference benchmarks, a broad set of classical machine learning models, covering both classification and regression tasks, was implemented. These algorithms are well suited for tabular clinical data, providing interpretable results and robust comparisons against more complex methods.

Linear Models

Linear models, including *Logistic Regression* and *Ridge Classifier* for classification, as well as *Linear Regression*, *Ridge*, *Lasso*, and *ElasticNet* for regression, are simple and interpretable. They allow a direct assessment of the relationship between predictors and outcomes, making them valuable for understanding feature effects.

Tree-Based Models

Tree-based models, such as *Decision Tree*, *Random Forest*, *Gradient Boosting*, *XGBoost*, and *LightGBM*, are highly flexible and can naturally capture non-linear relationships and interactions. They are robust to outliers, work well with mixed-type features, and provide feature importance measures, making them suitable for both classification and regression tasks.

Instance-Based Models

k-Nearest Neighbors (kNN) methods can predict outcomes by comparing a new sample to similar instances in the training set. This applies both to *kNN Classifier* for classification and *kNN Regressor* for regression, offering an intuitive, similarity-based approach.

Kernel-Based Models

Support Vector methods, including *Support Vector Classifier (SVC)* and *Support Vector Regressor (SVR)*, are effective in high-dimensional spaces and capable of learning complex non-linear relationships via kernel functions. They provide a balance between flexibility and regularization, useful for both types of tasks.

Probabilistic Models

Gaussian Naïve Bayes is a probabilistic classifier that relies on conditional independence assumptions among predictors, offering a simple and computationally efficient baseline for classification problems.

4.2.2 Conventional Neural Network Models

As the core contribution of this thesis, we designed and implemented feed-forward neural networks (multi-layer perceptrons, MLPs) tailored to the clinical tabular data. Starting from a baseline architecture, we systematically explored several design choices and regularization techniques in order to mitigate overfitting and improve generalization:

- **Depth and width:** variation in the number of hidden layers and neurons per layer, to evaluate the trade-off between model capacity and overfitting.
- **Regularization strategies:** L_1, L_2 and their combination (ElasticNet) to constrain the weight magnitudes and improve stability.
- **Dropout:** random inactivation of neurons during training to reduce co-adaptation and enhance robustness.
- **Batch Normalization:** normalization of intermediate activations to stabilize training and accelerate convergence.

The motivation to explore neural network models stemmed from their ability to capture highly non-linear relationships and complex feature interactions, which are often present in clinical and molecular data. Unlike classical models multi-layer perceptrons provide a flexible architecture capable of jointly handling numerical, categorical, and binary variables, as well as derived temporal features.
Moreover, despite the limited sample size available after preprocessing, neural networks represent a promising framework for future research in larger multicentric cohorts. Their adaptability to heterogeneous data types and their potential to learn hierarchical feature representations made them a natural candidate to investigate.

4.2.3 Advanced Neural Network Architectures

In order to address the limitations of standard MLPs, particularly their tendency to overfit and their limited ability to generalize in the presence of scarce data, several more sophisticated neural architectures were investigated. The rationale was to assess whether modifications in depth, width, regularization strategies, or the integration of attention mechanisms could lead to more robust representations of the clinical and molecular features.

Architectures emphasizing depth or width

Funnel MLP [shown in Fig. 3a]: after several experiments with shallow feed-forward architectures, which suffered from severe overfitting and suboptimal generalization, I adopted a progressively compressed multilayer design. By enforcing a structural bottleneck across layers, this architecture encourages the network to distill redundant information into a compact latent representation rather than memorizing noise. This rationale is supported by the well-established effectiveness of autoencoder-like architectures for dimensionality reduction, where high-dimensional data are mapped into low-dimensional codes that preserve the essential structure of the input while discarding spurious variability [39]. Such compression not only facilitates downstream tasks such as classification and visualization, but also acts as a powerful form of regularization, aligning with prior findings that multilayer bottleneck networks can outperform traditional linear techniques like PCA in extracting meaningful representations.
Wideres MLP [shown in Fig. 3b]: after preliminary experiments, it became evident that simply increasing depth was not a reliable solution, as it often led to optimization difficulties such as vanishing gradients [3] and degradation of training accuracy. To overcome these issues, an architecture that increases the width of hidden layers while incorporating residual connections, was adopted. This design choice was motivated by two main

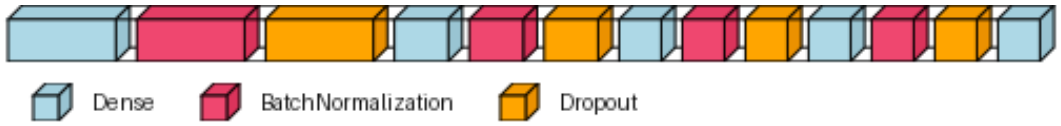
[3]The vanishing gradient problem arises when, during backpropagation, gradients shrink exponentially as they propagate through many layers. This is especially common with particular activation functions, whose derivatives saturate near zero, causing weight updates in earlier layers to become negligibly small. As a result, training deep networks becomes slow, unstable, or may fail to converge.

factors: first, residual pathways provide stable gradient flow and effectively address the degradation problem that hampers the training of deep plain networks [40]; second, wider hidden representations have been shown to capture richer feature interactions and, in many cases, improve performance more efficiently than blindly stacking additional layers [41]. In this sense, widening residual architectures preserves the stabilizing benefits of skip connections while enhancing representational capacity, leading to faster convergence and more effective generalization.

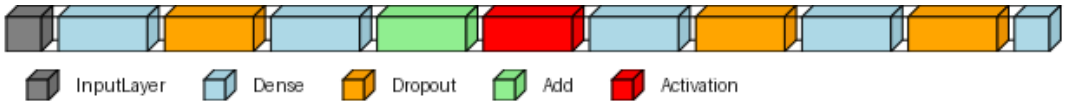
Stacked Narrow MLP [shown in Fig. 3c]: this variant takes the opposite approach, relying on deeper but very narrow layers. The motivation was that a "slim but deep" architecture could construct hierarchical abstractions with fewer parameters overall, thereby reducing overfitting risk while still enabling multiple levels of representation. This rationale aligns with theoretical insights on deep feedforward networks with piecewise-linear activations, which demonstrate that depth—rather than width alone—can exponentially increase the number of linear regions in the input space, effectively re-using intermediate computations across layers [42]. In this sense, even narrow deep networks are capable of approximating highly complex functions with relatively few parameters, leveraging compositional structure to achieve a favorable trade-off between expressivity and model size.

Parallel Branch MLP [shown in Fig. 3d]: motivated by the heterogeneity of the data modalities, this model processes different subsets of features in parallel branches before merging them. The goal was to allow each branch to specialize in its own feature space, promoting a more structured integration of clinical and molecular information than a single shared pathway could achieve.[4]

[4]The idea of parallel specialized branches is inspired by the Inception architecture [43], a convolutional neural network originally developed for image classification. Inception introduced the use of *inception modules*, blocks combining convolutions of different sizes and pooling layers, which enabled the network to capture both local and global patterns efficiently, reducing redundancy while maintaining high representational power.

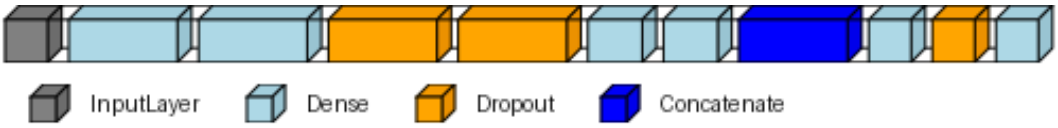


(a) Funnel MLP (Layered view)



(b) Wideres MLP (Layered view)

(c) Stacked Narrow MLP (Layered view)



(d) Parallel Branch MLP (Layered view)

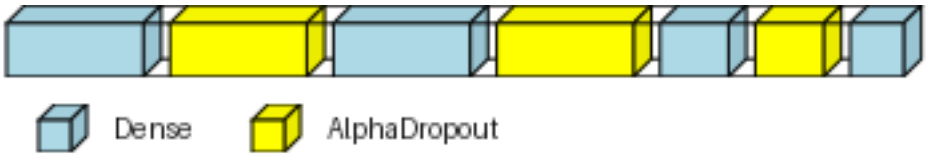
Figure 3: Architectures emphasizing depth or width: layered view representations of the models.

Architectures for stability and regularization

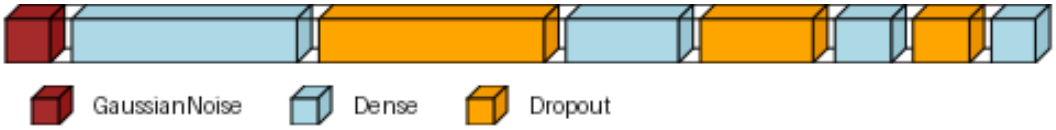
The **Self-Norm MLP** [shown in Fig. 4a] employs self-normalizing activation functions ($SELU$) that maintain normalized activations across layers, reducing the need for explicit normalization. This choice was motivated by the difficulties typically encountered when training very deep feedforward networks, which often suffer from high variance in training error and instability due to stochastic gradient descent, dropout, or perturbations in normalization parameters [44]. Self-normalizing neural networks (SNNs) are designed to push activations toward zero mean and unit variance, effectively stabilizing learning and avoiding vanishing or exploding gradients. By leveraging SELU activations, the Self-Norm MLP enables robust training of deeper architectures without the need for batch or layer normalization, while still allowing multiple levels of abstract feature representation.

The **NoisyWide MLP** [shown in Fig. 4b] injects random noise into the activations or weights during training, encouraging robustness and preventing over-reliance on specific features. This design choice is motivated by the well-established regularization effect of noise injection, which can improve generalization by discouraging the network from memorizing specific patterns in the training data [45]. In modern deep networks, especially

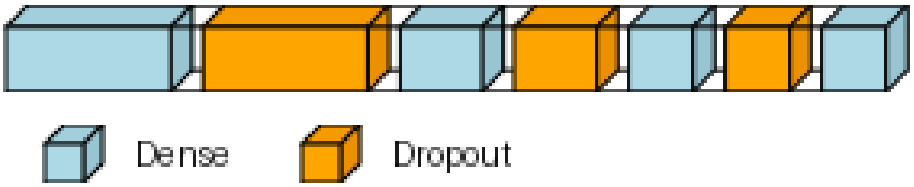
fully connected architectures with many layers or complex latent structures, noise added to gradients or activations helps exploration of the parameter space, allowing the optimizer to escape poor local minima and reducing sensitivity to initialization [46]. By introducing controlled stochastic perturbations, the NoisyWide MLP leverages these principles to train deeper and wider networks more effectively, improving stability and performance without relying solely on architectural constraints or explicit regularization techniques.
The **L2Light MLP** [shown in Fig. 4c] applies a strong L2 penalty to the weights, acting as a lightweight but effective regularization strategy to control model complexity. Weight decay is known to improve generalization by constraining the growth of parameters: it suppresses irrelevant components of the weight vector and mitigates the influence of noise in the training targets [47]. Unlike approaches that explicitly reduce the number of free parameters, L2 regularization encourages the network to learn the smallest set of weights necessary to solve the task, thereby reducing overfitting while preserving representational capacity. In this way, the L2Light MLP offers a computationally efficient form of regularization, particularly well-suited for tabular datasets where overly complex models tend to memorize noise rather than extract meaningful patterns.



(a) Self-Norm MLP (Layered view)



(b) NoisyWide MLP (Layered view)



(c) L2Light MLP (Layered view)

Figure 4: Architectures for stability and regularization: layered view representations of the models.

Attention-based variants

To further investigate the potential of attention mechanisms in clinical data, several MLP variants were designed by integrating an attention block at different stages of the architecture. The rationale is that attention can automatically highlight the most informative variables by assigning different weights to features depending on the learned context, thus addressing challenges such as noisy signals, missing values, and complex feature interactions. In healthcare applications, attention has emerged as a powerful alternative to recurrent approaches, offering a more efficient way to capture long-range dependencies in multivariate clinical data [48]. Beyond sequential domains, recent work on tabular data has also shown that attention-based architectures can bridge the performance gap between standard MLPs and tree-based ensembles by generating contextual embeddings that are both robust and interpretable [49]. These insights motivated the exploration of attention-augmented MLPs in our study, briefly described below:

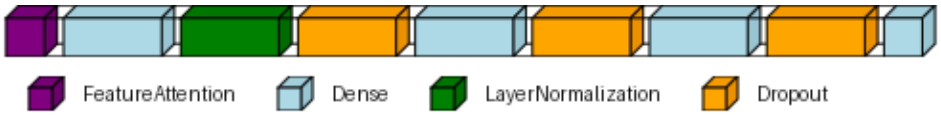
- **Input-level attention**: inspired by TabNet [50], this variant applies attention directly to raw input features (see Fig. 5a). The motivation was to allow the model to highlight the most relevant variables from the outset, effectively acting as an adaptive feature selector.

- **Intermediate attention**: motivated by works like TabTransformer [49], here attention operates after initial hidden layers, so that it acts on transformed representations rather than raw data (see Fig. 5b). This design tests whether attention can better capture interactions once low-level noise has been filtered.

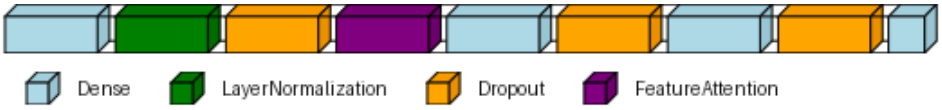
- **Pre-output attention**: positioned immediately before the final layer (see Fig. 5c), with the aim of directly influencing the classification decision by selectively weighting the latent features most relevant to the outcome.

- **Multiple attention blocks**: drawing inspiration from Transformer architectures [51], this approach inserts attention at different depths of the network (see Fig. 5d). The rationale is to capture complementary patterns at multiple levels of abstraction, from local interactions to global dependencies.

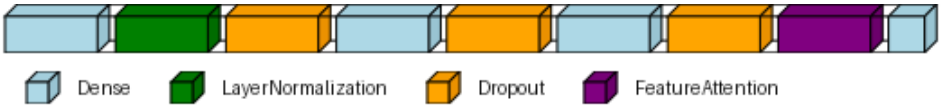
- **Residual attention**: building on Residual Attention Networks [52], this variant integrates attention with skip connections (see Fig. 5e), aiming to stabilize training while maintaining the benefits of dynamic feature weighting.



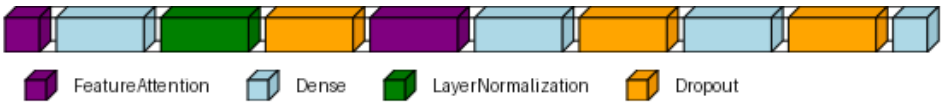
(a) Attention block on the input level



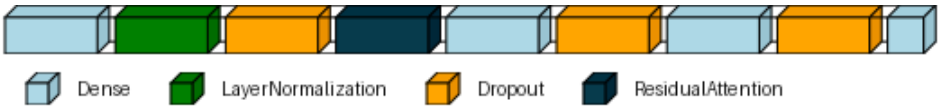
(b) Intermediate Attention block



(c) Attention block on the output level



(d) Multiple Attention blocks



(e) Residual Attention block

Figure 5: Schematic representation of the different attention-based neural network architectures.

4.3 Model Interpretability using SHAP Values

To ensure comparability and interpretability across different model families (classical machine learning algorithms, shallow neural networks and deeper architectures) we employed *SHAP* (SHapley Additive exPlanations) values [53].

SHAP is a unified framework for feature attribution grounded in cooperative game theory. It computes the contribution of each input variable to the model's output as a Shapley value, i.e., the expected marginal contribution of a feature across all possible feature subsets. Formally, for a model $f(x)$, the SHAP value ϕ_i of feature i is given by:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right],$$

where F is the full set of input features and S denotes any subset excluding feature i.

Since the exact computation of SHAP values is intractable for complex models, various approximation methods have been developed. Following the taxonomy proposed by Lundberg and Lee [53], we used different explainers according to the model type:

- **TreeExplainer** for ensemble-based models (e.g., Random Forest, XGBoost), which leverages the structure of decision trees for an exact and efficient computation of SHAP values.
- **LinearExplainer** for linear and logistic regression models, where SHAP values coincide with standardized regression coefficients.
- **DeepExplainer** for standard feed-forward neural networks, which builds on DeepLIFT to propagate attributions layer by layer.
- **KernelExplainer** for model-agnostic estimation, used for MLPs and more complex architectures (e.g., residual or attention-based models). This explainer approximates the SHAP values through a locally weighted linear regression around each sample.

To visualize and interpret the global feature contributions derived from SHAP values, the `shap.summary_plot()` function was employed, which provides a compact yet informative representation of both feature importance and feature effects [54]. Each point in the summary plot corresponds to a single Shapley value associated with a specific feature and sample: the x-axis represents the magnitude and direction of the feature's impact on the model output, while the y-axis lists the features ordered by their overall importance. The color of each point encodes the original feature value (from low to high), thus enabling a direct visualization of whether higher or lower values of a feature tend to increase or decrease the predicted probability.

This visualization combines global and local interpretability: the spread of points along the x-axis illustrates the variability of a feature's contribution across the dataset, while the color gradient reveals potential non-linear relationships between feature values and model outputs. In addition, the vertical jittering of points prevents overlap, making it possible to perceive the overall distribution of SHAP values per feature. In this work, SHAP summary plots were generated for each output class to facilitate class-specific interpretation of the predictive patterns.

This approach allowed a unified, quantitative comparison of feature relevance across diverse architectures, facilitating both the validation of model behavior and the interpretation of clinically meaningful patterns.

5 Experimentation and Evaluation

5.1 Experimental Setup

To ensure a fair and efficient evaluation, a subset of features was selected based on their relevance as predictors of the target variable, thus reducing both dimensionality and computational costs. The benchmark was carried out by comparing classical machine learning algorithms with more advanced neural network architectures, in order to assess the added value of deep learning approaches.

Prior to training, the unified dataset underwent a preprocessing phase that included one-hot encoding of categorical variables to ensure proper model convergence. Subsequently, the data were partitioned into training, validation, and test sets, following a stratified split where appropriate, to guarantee robust and unbiased performance assessment.

All implementation notebooks and source code are available at `https://github.com/noemicic/immunotherapy-outcome-prediction`.

5.1.1 Data Balancing and Scaling Strategies

Given the presence of class imbalance in several response variables, and considering the potential sensitivity of certain models to feature scaling, multiple configurations were systematically evaluated. In particular, each model was trained under the conditions shown in Table 6:

Table 6: Preprocessing and imbalance-handling strategies tested across models.

Strategy	**Description**
No scaling	Raw features used directly as input.
Scaling	Standardization of features without any balancing technique.
No scaling + Class weights	Class imbalance mitigation through weighted loss during training, without feature scaling.
Scaling + Class weights	Combination of class weighting and feature scaling.
Scaling + SMOTE (automatic)	Synthetic oversampling of the minority class with default SMOTE parameters, previewed by feature scaling.
Scaling + SMOTE (custom)	Oversampling with SMOTE under a custom configuration, previewed by scaling.

5.2 Evaluation Metrics

5.2.1 Classification Metrics

Binary Classification

For the binary classification tasks, model performance was assessed using a set of complementary metrics in order to capture different aspects of predictive capability. In this case, the main evaluation metric is **Accuracy**, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. However, due to the presence of class imbalance, greater emphasis was placed on the F1-score and the Area Under the ROC Curve (AUC), which provide more robust performance indicators in such settings.
The **F1-score** is given by:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2}$$

with

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}. \tag{3}$$

The **AUC** is computed as the area under the Receiver Operating Characteristic (ROC) curve, which characterizes the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{4}$$

where

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}. \tag{5}$$

In addition to these scalar metrics, the **Confusion Matrix** played a crucial role in model evaluation. It provides a comprehensive view of the classifier's predictions by displaying the distribution of true and predicted labels.

Multi-class Classification

For multi-class classification tasks, the *weighted* versions of standard metrics were used to account for class imbalance.

The **Accuracy** is defined as the proportion of correct predictions over all samples:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6}$$

Instead, **Weighted Precision**, **Recall** and **F1-score** are computed as:

$$\text{Precision}_{\text{weighted}} = \sum_{i=1}^{C} w_i \cdot \frac{TP_i}{TP_i + FP_i}, \tag{7}$$

$$\text{Recall}_{\text{weighted}} = \sum_{i=1}^{C} w_i \cdot \frac{TP_i}{TP_i + FN_i}, \tag{8}$$

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{C} w_i \cdot \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{9}$$

where C is the number of classes, TP_i, FP_i, FN_i denote the true positives, false positives, and false negatives for class i, and w_i is the support proportion of class i.
Macro-averaged AUC-ROC is computed if probabilistic outputs are available:

$$\text{AUC}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \text{AUC}_i \tag{10}$$

Finally, the **Confusion Matrix** again served as an essential diagnostic tool to evaluate multi-class predictions. By highlighting which classes were most frequently confused, it enabled a deeper understanding of model weaknesses and guided subsequent tuning and model selection, often proving more informative than aggregated metrics alone.

5.2.2 Regression Metrics

For regression tasks, we evaluated the model performance using the following metrics.
The **Mean Absolute Error (MAE)** measures the average absolute difference between predicted and actual values.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{11}$$

Instead, the **Mean Squared Error (MSE)** measures the average squared difference between predicted and actual values.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12}$$

The **Mean Absolute Percentage Error (MAPE)** measures the average absolute percentage difference between predicted and actual values.

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\max(y_i, \epsilon)} \tag{13}$$

where y_i and $\hat{y}_i$ are the actual and predicted values for sample i, n is the number of samples, and ϵ is a small constant added to avoid division by zero.

However, during the experimental phase, it was observed that the **MAPE** metric often produced extremely large or unstable values due to the presence of actual values close to zero in the denominator, which caused numerical instability. Therefore, in the results section, the **Symmetric Mean Absolute Percentage Error (sMAPE)** was reported instead, as a more robust alternative.
The **sMAPE** normalizes the error with respect to the average of the actual and predicted values, thus preventing the divergence observed with MAPE:

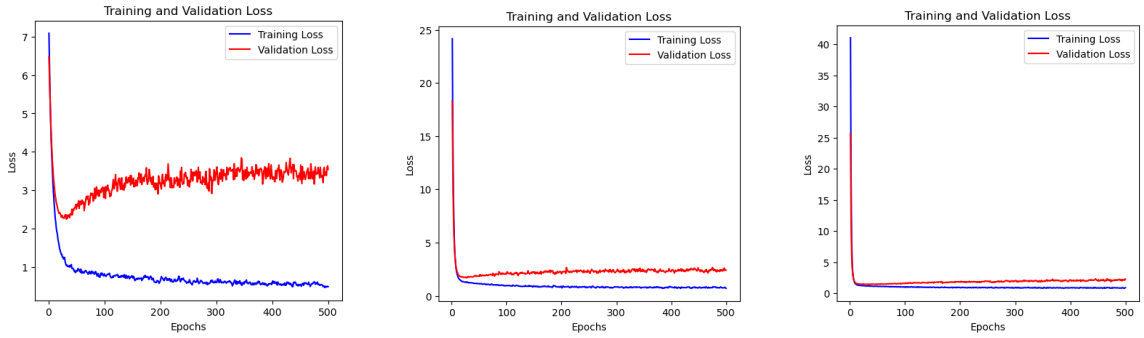
$$\text{sMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{14}$$

This metric provides a more stable and interpretable percentage-based measure of error, especially in the presence of actual values approaching zero.

5.3 Impact of Hyperparameters

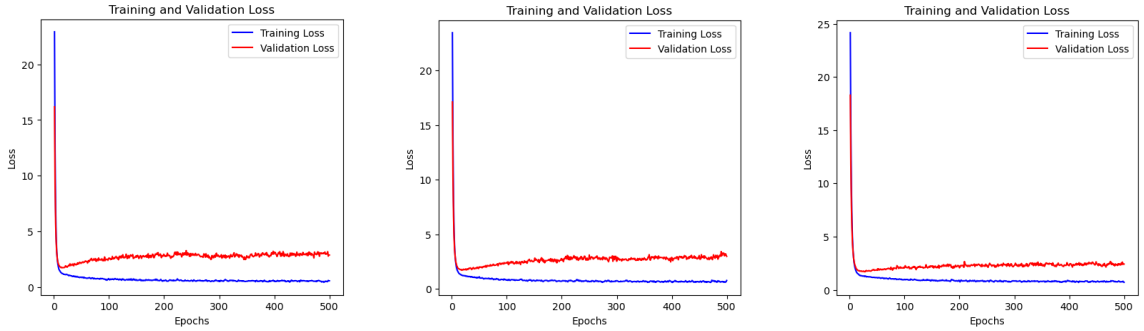
While all three model families were evaluated, the hyperparameter optimization process was primarily conducted for the Conventional Neural Networks. For classical ML and Advanced NN models, existing architectures or well-established implementations were employed, with only minor parameter tuning (mostly through grid search) to select the most competitive configurations within each family. In contrast, the Conventional NN models required a more systematic fine-tuning of both architectural and regularization parameters to achieve satisfactory generalization.

Among all the tested hyperparameters, the network depth, number of neurons per layer, **L2 regularization strength** and **dropout rate** had the most significant influence on model performance and overfitting behavior. Increasing any of these parameters initially improved generalization, as reflected by a reduced gap between training and validation losses, but excessively large values led to underfitting and a decline in predictive metrics.



(a) L2 = 0.01 (b) L2 = 0.05 (c) L2 = 0.1

Figure 6: Training and validation loss curves for the BOR prediction task under different L2 regularization strengths. All models share the same architecture (2 hidden layers, 512 neurons each, dropout 0.4, batch normalization).



(a) Dropout = 0.2 (b) Dropout = 0.3 (c) Dropout = 0.4

Figure 7: Training and validation loss curves for the BOR prediction task under different dropout rates. All models share the same architecture (2 hidden layers, 512 neurons each, L2 regularization 0.05, batch normalization).

Overall, the optimal trade-off between model stability and predictive accuracy was achieved with two hidden layers, 512 neurons per layer, a dropout rate around 0.4, and L2 regularization in the range of 0.05–0.1. These intermediate configurations consistently yielded robust generalization across endpoints, confirming that moderate regularization and controlled architectural complexity are key to balancing expressiveness and overfitting in conventional neural networks.

5.4 Results

As discussed in Section 5.1.1, several data preprocessing strategies were evaluated prior to model training, including feature scaling, class balancing through SMOTE oversampling, and the use of class weights. These techniques were tested both individually and in combination, in order to assess their effect on model stability and predictive performance across different endpoints. In the context of binary and multi-class classification tasks, specifically for the prediction of OS Status, BOR (both original and modified definitions), and Responder Quality Groups (under the two derived stratifications), the application of feature scaling to numerical variables proved to be the most consistent and effective preprocessing strategy. Models trained on scaled datasets generally achieved higher performance and more stable convergence across the majority of architectures and prediction tasks, while the additional use of SMOTE or class weighting did not systematically improve results and, in some cases, led to overfitting or unstable behavior. For this reason, only the results obtained with scaled features (without further balancing techniques) are reported and discussed in the following sections. To empirically support this choice, Table 7 presents a representative comparison of the different preprocessing strategies.

Table 7: Comparison of preprocessing strategies for the conventional NN model on the BOR (original) classification task.

Strategy	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix				
							CR	PR	SD	PD
						CR	0	0	5	0
No scaling	**0.517**	0.273	**0.517**	0.357	0.500	PR	0	0	42	0
						SD	0	1	94	0
						PD	0	1	39	0
							CR	PR	SD	PD
						CR	0	1	2	2
Scaling	0.478	**0.472**	0.478	**0.473**	0.584	PR	1	17	17	7
						SD	0	23	59	13
						PD	0	12	17	11
							CR	PR	SD	PD
						CR	5	0	0	0
No Scaling + Class weights	0.028	0.001	0.028	0.002	0.554	PR	41	0	1	0
						SD	94	1	0	0
						PD	39	1	0	0
							CR	PR	SD	PD
						CR	0	1	2	2
Scaling + Class weights	0.429	0.434	0.429	0.428	0.567	PR	1	18	16	7
						SD	1	28	51	15
						PD	1	11	19	9
							CR	PR	SD	PD
						CR	0	2	1	2
Scaling + SMOTE (automatic)	0.429	0.446	0.429	0.431	**0.589**	PR	1	20	16	5
						SD	2	29	50	14
						PD	2	14	16	8
							CR	PR	SD	PD
						CR	0	1	0	4
Scaling + SMOTE (custom)	0.401	0.395	0.401	0.397	0.586	PR	0	20	26	6
						SD	3	23	55	14
						PD	3	10	19	8

It should be noted that feature scaling was applied whenever required by the specific model type, both for classification and regression tasks. In particular, algorithms based on distance metrics or gradient optimization (e.g., linear models, support vector machines, neural networks) were trained on scaled features, whereas tree-based models (e.g., Random Forest, Gradient Boosting) were used with raw input data, since they are inherently insensitive to feature scaling.

In the following sections, the predictive performance of the developed models is presented for each endpoint considered; the results are summarized in tables [5], reporting the main performance metrics across different model families, grouped as follows:

- **Classical Machine Learning Models**
  A grid search over the main hyperparameters was performed to identify the best configuration in terms of predictive performance, with model evaluation based on k-fold cross-validation to ensure robust and unbiased estimates.

- **Conventional Neural Network Models**
  A systematic exploration of architectural and regularization parameters was carried out to identify optimal configurations and reduce overfitting. The number of hidden layers varied between 1 and 2, neurons per layer ranged from 64 to 1024, dropout rates from 0.2 to 0.5, and L2 regularization from 0 to 0.1. Additionally, ElasticNet penalties (combining L1 and L2 regularization) and batch normalization layers were tested. Given the large search space, only representative configurations are reported in the result tables.

- **Advanced Neural Network models**
  Finally, the advanced neural network models, including attention-based architectures, underwent extensive hyperparameter tuning to mitigate overfitting, identified as the main limiting factor across all tasks. Interestingly, the best-performing configurations were often close to the default parameter settings, suggesting a high sensitivity of these models to overparameterization.

[5] For each metric, the best result is highlighted in **bold**, while the second-best result is underlined.

5.4.1 Dataset B

OS Status

Table 8: Performance of classical ML models on the OS Status classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix
Logistic Regression	0.691	0.718	0.667	0.691	**0.739**	0: 28, 11 / 1: 14, 28
Random Forest	0.679	0.674	0.738	0.705	**0.739**	0: 24, 15 / 1: 11, 31
Gradient Boosting	0.679	0.691	0.691	0.691	0.736	0: 26, 13 / 1: 13, 29
XGBoost	0.667	0.660	0.738	0.697	0.735	0: 23, 16 / 1: 11, 31
LGBM	0.654	0.646	0.738	0.689	0.733	0: 22, 17 / 1: 11, 31
SVC	0.519	0.519	**1.000**	0.683	0.523	0: 0, 39 / 1: 0, 42
KNN	0.469	0.493	0.857	0.626	0.461	0: 2, 37 / 1: 6, 36
Ridge Classifier	**0.704**	**0.725**	0.691	**0.707**	–	0: 28, 11 / 1: 13, 29
Gaussian Naive Bayes	0.494	0.571	0.095	0.163	0.712	0: 36, 3 / 1: 38, 4
Decision Tree	0.580	0.587	0.643	0.614	0.652	0: 20, 19 / 1: 15, 27

Among all the tested models, the *Logistic Regression* emerged as the most balanced choice for predicting the OS status of lung cancer patients treated with immunotherapy. Despite not achieving the highest overall accuracy (0.691 vs. 0.704 for *Ridge Classifier*), it reached the best AUC-ROC (0.739) and exhibited a well-balanced confusion matrix, indicating good generalization across both classes. In addition, Logistic Regression offers high interpretability, which is crucial in clinical applications where understanding the contribution of each variable is as important as the predictive performance itself. Therefore, this model represents the most appropriate trade-off between accuracy, robustness, and explainability for this specific clinical prediction task.

Table 9: Performance of conventional NN models on the OS Status classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix		
							0	1
1 HL + Dropout	0.593	0.622	0.548	0.582	0.665	0	25	14
						1	19	22
							0	1
1 HL + L2 + Dropout	0.667	0.674	**0.691**	**0.682**	0.691	0	25	14
						1	13	29
							0	1
1 HL + Norm Layer + L2 + Dropout	0.642	0.697	0.548	0.613	0.659	0	29	10
						1	19	23
							0	1
1 HL + BatchNorm + L2 + Dropout	0.642	0.651	0.667	0.659	0.677	0	24	15
						1	14	28
							0	1
2 HL + L2 + Dropout	**0.679**	**0.711**	0.643	0.675	0.692	0	28	11
						1	15	27
							0	1
2 HL + Norm Layer + L2 + Dropout	0.654	0.694	0.595	0.641	**0.695**	0	28	11
						1	17	25
							0	1
2 HL + Norm Layer + ElasticNet + Dropout	0.642	0.697	0.548	0.613	0.677	0	29	10
						1	19	23

Among all the tested conventional neural network architectures, the *2 HL + L2 + Dropout* model achieved the best overall balance between performance and generalization. It reached the highest accuracy (0.679) and precision (0.711), together with a competitive AUC-ROC (0.692), while maintaining a well-balanced confusion matrix. The inclusion of L2 regularization and dropout across two hidden layers effectively mitigated overfitting, as also confirmed by the training curves, allowing the network to learn meaningful patterns without memorizing the training data.

Table 10: Performance of advanced NN models on the OS Status classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix
						0 1
Funnel MLP	0.642	0.638	0.714	0.674	0.690	0 22 17
						1 12 30
						0 1
WideRes MLP	0.617	0.649	0.571	0.608	0.683	0 26 13
						1 18 24
						0 1
Stacked Narrow MLP	0.605	0.625	0.595	0.610	0.662	0 24 15
						1 17 25
						0 1
Parallel Branch MLP	0.704	0.750	0.643	0.692	0.706	0 30 9
						1 15 27
						0 1
Self-Norm MLP	0.667	0.667	0.714	0.690	0.667	0 24 15
						1 12 30
						0 1
NoisyWide MLP	0.667	0.703	0.619	0.658	0.686	0 28 11
						1 16 26
						0 1
L2Light MLP	0.617	0.622	0.667	0.644	0.687	0 22 17
						1 14 28
						0 1
Input-level attention block	**0.716**	0.721	**0.738**	**0.729**	0.741	0 27 12
						1 11 31
						0 1
Intermediate attention block	**0.716**	0.721	**0.738**	**0.729**	0.721	0 27 12
						1 11 31
						0 1
Pre-output attention block	0.691	0.730	0.643	0.684	0.736	0 29 10
						1 15 27
						0 1
Multiple attention blocks	**0.716**	**0.788**	0.619	0.693	**0.810**	0 32 7
						1 16 26
						0 1
Residual attention block	0.679	0.711	0.643	0.675	0.700	0 28 11
						1 15 27

Among the advanced neural architectures, the *Input-level Attention Block* model achieved the best overall performance, with the highest accuracy (0.716), recall (0.738), and F1-score (0.729), along with a competitive AUC-ROC (0.741). Its confusion matrix reveals a well-balanced trade-off between classes, minimizing false negatives, a key aspect in clinical applications where identifying high-risk patients is critical. Moreover, the use of an attention mechanism at the input stage allows the network to learn feature relevance from the very beginning, enhancing both predictive performance and interpretability. The training curves also indicate a good control of overfitting, confirming the robustness and reliability of this model for predicting OS status in patients undergoing immunotherapy.

OS Duration

Table 11: Performance of classical ML models on the OS Duration regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
Linear Regression	47.675	5.395	75.951
Ridge	47.421	5.299	75.537
Lasso	44.377	5.204	74.333
ElasticNet	44.377	5.204	74.333
Decision Tree Regressor	46.174	5.167	69.780
Random Forest Regressor	<u>42.882</u>	5.068	70.123
Gradient Boosting Regressor	43.180	5.052	<u>68.819</u>
XGBoost Regressor	43.471	<u>5.008</u>	**68.340**
LightGBM Regressor	46.585	5.186	69.020
Support Vector Regressor (SVR)	**42.503**	**4.899**	71.371
K-Nearest Neighbors Regressor (kNN)	51.336	5.531	73.239

Among the tested models, the *XGBoost Regressor* achieved the best overall trade-off between predictive accuracy and robustness, with competitive MSE and MAE values and the lowest sMAPE (68.34%). While the *Support Vector Regressor* obtained slightly lower absolute errors (MSE = 42.50, MAE = 4.90), its kernel-based structure may be more sensitive to noise and outliers, and provides limited interpretability. In contrast, *XGBoost* offers a more stable performance across different survival time ranges, benefiting from built-in regularization and feature importance analysis.

Table 12: Performance of conventional NN models on the OS Duration regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
1 HL + Dropout	51.017	5.313	73.090
1 HL + L2 + Dropout	46.521	4.886	70.130
1 HL + Norm Layer + L2 + Dropout	51.437	4.973	74.140
1 HL + BatchNorm + L2 + Dropout	52.054	5.409	78.150
2 HL + L2 + Dropout	48.124	4.850	68.920
2 HL + Norm Layer + L2 + Dropout	**44.470**	<u>4.609</u>	<u>67.140</u>
2 HL + Norm Layer + ElasticNet + Dropout	<u>44.809</u>	**4.499**	**66.790**

Among all conventional neural network architectures, the *2 HL + Norm Layer + ElasticNet + Dropout* model achieved the best overall performance for OS duration prediction, with the lowest MAE (4.499) and sMAPE (66.79%), and a competitive MSE (44.81). The combination of normalization and ElasticNet regularization (L1 + L2) effectively balanced bias and variance, reducing overfitting while preserving model expressiveness. Furthermore, the training curves indicated a more controlled gap between training and validation losses compared to other architectures, confirming the model's superior generalization capability.

Table 13: Performance of advanced NN models on the OS Duration regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
Funnel MLP	**40.797**	4.795	69.580
WideRes MLP	43.784	4.764	68.840
Stacked Narrow MLP	43.475	4.556	66.540
Parallel Branch MLP	<u>40.932</u>	4.503	<u>66.180</u>
Self-Norm MLP	79.376	6.731	79.550
NoisyWide MLP	42.970	4.639	67.960
L2Light MLP	44.008	4.757	68.150
Input-level attention block	44.670	<u>4.312</u>	67.490
Intermediate attention block	45.929	4.424	68.990
Pre-output attention block	50.713	4.520	70.940
Multiple attention blocks	43.282	**4.199**	**63.280**
Residual attention block	48.144	4.766	68.490

Among advanced neural architectures, the *Multiple Attention Blocks MLP* achieved the best overall performance for OS duration prediction, with the lowest MAE (4.199) and sMAPE (63.28%). Although the *Parallel Branch MLP* yielded slightly lower MSE, the multiple attention configuration demonstrated a more stable training behavior and significantly reduced overfitting, as indicated by the smaller gap between training and validation losses. This design allows the network to hierarchically capture and reweight relevant clinical features across layers, improving both predictive accuracy and model robustness.

PFS

Table 14: Performance of classical ML models on the PFS regression task.

Model	MSE	MAE	sMAPE (%)
Linear Regression	<u>17.996</u>	<u>2.785</u>	<u>64.118</u>
Ridge	18.041	2.801	65.036
Lasso	18.041	2.788	66.060
ElasticNet	18.041	2.788	66.060
Decision Tree Regressor	19.516	2.986	67.177
Random Forest Regressor	18.991	3.098	70.794
Gradient Boosting Regressor	24.053	3.232	67.746
XGBoost Regressor	20.501	3.198	69.350
LightGBM Regressor	19.446	3.024	67.218
Support Vector Regressor (SVR)	**17.783**	**2.582**	**60.795**
K-Nearest Neighbors Regressor (kNN)	19.415	3.101	67.938

The *Support Vector Regressor* achieved the best overall performance for PFS prediction, outperforming both linear and ensemble-based models: this improvement likely stems from the SVR's ability to capture mild non-linear interactions among clinical variables and its robustness to outliers due to the ε-insensitive loss. These characteristics make it particularly suitable for modeling the heterogeneous progression patterns observed in patient data.

Table 15: Performance of conventional NN models on the PFS regression task.

Model	MSE	MAE	sMAPE (%)
1 HL + Dropout	22.596	3.312	78.050
1 HL + L2 + Dropout	22.185	3.173	71.370
1 HL + Norm Layer + L2 + Dropout	24.214	3.302	70.040
1 HL + BatchNorm + L2 + Dropout	26.474	3.359	70.920
2 HL + L2 + Dropout	21.230	<u>2.952</u>	<u>61.720</u>
2 HL + Norm Layer + L2 + Dropout	<u>21.185</u>	3.002	64.600
2 HL + Norm Layer + ElasticNet + Dropout	**21.013**	**2.741**	**60.720**

Among conventional neural architectures, the *2 HL + Norm Layer + ElasticNet + Dropout* model achieved the best performance in predicting PFS: this outcome can be attributed to the ElasticNet's combined ability to enforce sparsity (via the L1 term) while maintaining weight stability (through the L2 component). Together with dropout and normalization, this configuration effectively controlled overfitting, as also reflected by the smooth and well-aligned training and validation loss curves.

Table 16: Performance of advanced NN models on the PFS regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
Funnel MLP	19.943	2.743	<u>59.740</u>
WideRes MLP	19.918	2.868	61.700
Stacked Narrow MLP	22.177	2.900	61.960
Parallel Branch MLP	19.180	2.749	60.960
Self-Norm MLP	60.146	5.269	80.930
NoisyWide MLP	**17.223**	<u>2.518</u>	**59.500**
L2Light MLP	19.260	2.777	60.390
Input-level attention block	24.377	2.946	63.990
Intermediate attention block	<u>18.691</u>	2.696	61.420
Pre-output attention block	21.103	**2.497**	61.860
Multiple attention blocks	19.813	2.794	64.250
Residual attention block	22.863	3.061	65.010

Among the advanced architectures, the *NoisyWide MLP* achieved the best overall performance across all regression metrics, showing both low prediction error and stable convergence. The combination of wide layers and noise injection likely enhanced its capacity to model complex nonlinear relationships while mitigating overfitting; so this model provides a more balanced and theoretically consistent trade-off between expressivity and generalization.

5.4.2 Dataset D

BOR (original)

Table 17: Performance of classical ML models on the BOR (original) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix				
							CR	PR	SD	PD
						CR	0	4	0	1
Logistic Regression	0.445	0.424	0.445	0.431	0.602	PR	3	56	14	22
						SD	0	26	6	8
						PD	1	17	5	19
							CR	PR	SD	PD
						CR	0	2	2	1
Random Forest	0.440	0.427	0.440	0.428	**0.609**	PR	4	55	12	24
						SD	2	25	6	7
						PD	1	20	2	19
							CR	PR	SD	PD
						CR	0	2	2	1
Gradient Boosting	0.429	0.411	0.429	0.414	0.567	PR	2	52	13	28
						SD	0	26	3	11
						PD	3	11	5	23
							CR	PR	SD	PD
						CR	0	2	1	2
XGBoost	0.462	0.425	0.462	**0.434**	0.593	PR	2	59	9	25
						SD	0	29	4	7
						PD	0	18	3	21
							CR	PR	SD	PD
						CR	0	5	0	0
LGBM	0.489	0.383	0.489	0.428	0.534	PR	1	70	5	19
						SD	0	35	0	5
						PD	2	20	1	19
							CR	PR	SD	PD
						CR	0	5	0	0
SVC	**0.522**	0.273	**0.522**	0.358	0.500	PR	0	95	0	0
						SD	0	40	0	0
						PD	0	42	0	0
							CR	PR	SD	PD
						CR	0	2	0	3
KNN	0.319	0.380	0.319	0.276	0.483	PR	0	25	0	70
						SD	0	4	0	36
						PD	0	9	0	33
							CR	PR	SD	PD
						CR	3	0	1	1
Ridge Classifier	0.324	**0.571**	0.324	0.346	–	PR	24	20	22	29
						SD	12	3	12	13
						PD	13	2	3	24
							CR	PR	SD	PD
						CR	0	0	5	0
Gaussian Naive Bayes	0.242	0.139	0.242	0.132	0.552	PR	4	0	82	9
						SD	1	0	38	1
						PD	2	0	34	6
							CR	PR	SD	PD
						CR	0	2	1	2
Decision Tree	0.407	0.449	0.407	0.422	0.534	PR	6	47	15	27
						SD	3	14	10	13
						PD	1	13	11	17

Although the *SVC* achieved the highest accuracy, its confusion matrix revealed a strong bias toward the predominant class, indicating poor generalization under severe class imbalance. Conversely, the *Logistic Regression* model provided the most balanced and clinically meaningful predictions. Despite not attaining the top performance across all metrics, it achieved the second-best F1-score and AUC-ROC, while maintaining a reasonable distribution of predictions across classes. Importantly, its misclassifications mostly occurred

between clinically related categories (e.g., PR and SD), suggesting that the model effectively captured the underlying structure of the data.

Table 18: Performance of conventional NN models on the BOR (original) classification task.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**				
							CR	PR	SD	PD
						CR	0	2	2	1
1 HL + Dropout	0.407	0.402	0.407	0.401	0.588	PR	2	54	11	28
						SD	0	21	5	14
						PD	2	17	8	15
							CR	PR	SD	PD
						CR	0	2	2	1
1 HL + L2 + Dropout	0.440	0.427	0.440	0.426	0.573	PR	2	57	9	27
						SD	0	29	4	17
						PD	1	16	6	19
							CR	PR	SD	PD
						CR	0	2	2	1
1 HL + Norm Layer + L2 + Dropout	0.423	0.425	0.423	0.423	0.568	PR	3	55	17	20
						SD	0	18	8	14
						PD	2	18	8	14
							CR	PR	SD	PD
						CR	0	4	0	1
1 HL + BatchNorm + L2 + Dropout	0.445	0.438	0.445	0.441	0.588	PR	2	56	17	20
						SD	1	17	12	10
						PD	1	23	5	13
							CR	PR	SD	PD
						CR	0	3	1	1
2 HL + L2 + Dropout	<u>0.456</u>	<u>0.471</u>	<u>0.456</u>	<u>0.450</u>	<u>0.613</u>	PR	1	53	8	33
						SD	1	17	8	14
						PD	1	16	3	22
							CR	PR	SD	PD
						CR	0	2	2	1
2 HL + Norm Layer + L2 + Dropout	**0.484**	**0.475**	**0.484**	**0.476**	**0.624**	PR	1	60	10	24
						SD	1	18	10	11
						PD	0	18	6	18
							CR	PR	SD	PD
						CR	0	3	1	1
2 HL + Norm Layer + ElasticNet + Dropout	0.434	0.442	0.434	0.433	0.624	PR	1	51	13	30
						SD	1	16	10	13
						PD	0	19	5	18

Among the conventional neural network models, the *2HL + Normalization Layer + L2 + Dropout* architecture achieved the best overall performance across all metrics, with the highest accuracy, F1-score (0.476) and AUC-ROC (0.624). The confusion matrix confirms its superior balance in classifying the major response categories (PR and PD), with fewer extreme misclassifications compared to simpler architectures. However, its predictions for the SD class remain suboptimal, as the model tends to confuse SD cases with PR or PD, indicating that the feature space separating SD responses is not well captured. As for the CR class, no samples were correctly predicted, a limitation shared across all models, primarily due to the extremely low number of CR examples in the dataset.

Table 19: Performance of advanced NN models on the BOR (original) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix				
							CR	PR	SD	PD
						CR	0	2	0	3
Funnel MLP	0.368	0.407	0.368	0.374	0.538	PR	2	42	18	33
						SD	0	15	5	20
						PD	1	11	10	20
							CR	PR	SD	PD
						CR	0	2	2	1
WideRes MLP	0.445	**0.466**	0.445	0.438	0.576	PR	4	55	3	33
						SD	1	20	6	13
						PD	2	16	4	20
							CR	PR	SD	PD
						CR	0	1	3	1
Stacked Narrow MLP	0.451	0.446	0.451	0.448	0.583	PR	2	58	16	19
						SD	1	19	11	9
						PD	2	20	7	13
							CR	PR	SD	PD
						CR	0	2	2	1
Parallel Branch MLP	0.462	0.453	0.462	<u>0.455</u>	0.568	PR	2	58	10	25
						SD	0	23	9	8
						PD	1	14	10	17
							CR	PR	SD	PD
						CR	0	4	0	1
Self-Norm MLP	0.456	<u>0.463</u>	0.456	0.449	0.568	PR	1	59	5	30
						SD	0	18	8	14
						PD	2	17	7	16
							CR	PR	SD	PD
						CR	0	2	1	2
NoisyWide MLP	0.440	0.432	0.440	0.434	0.563	PR	3	58	10	24
						SD	2	21	6	11
						PD	0	15	11	16
							CR	PR	SD	PD
						CR	0	1	3	1
L2Light MLP	0.478	0.462	0.478	**0.461**	**0.623**	PR	1	60	9	25
						SD	1	19	6	14
						PD	0	19	2	21
							CR	PR	SD	PD
						CR	0	4	0	1
Input-level attention block	0.423	0.430	0.423	0.415	<u>0.596</u>	PR	0	47	13	35
						SD	0	17	7	16
						PD	0	15	4	23
							CR	PR	SD	PD
						CR	0	4	0	1
Intermediate attention block	<u>0.517</u>	0.373	<u>0.517</u>	0.409	0.594	PR	0	86	0	9
						SD	0	39	0	1
						PD	0	34	0	8
							CR	PR	SD	PD
						CR	0	5	0	0
Pre-output attention block	**0.522**	0.273	**0.522**	0.358	0.562	PR	0	95	0	0
						SD	0	40	0	0
						PD	0	42	0	0
							CR	PR	SD	PD
						CR	0	4	0	1
Multiple attention blocks	0.451	0.329	0.451	0.379	0.537	PR	0	70	0	25
						SD	0	32	0	8
						PD	0	30	0	12
							CR	PR	SD	PD
						CR	0	0	4	1
Residual attention block	0.363	0.404	0.363	0.370	0.584	PR	2	38	18	37
						SD	2	19	7	12
						PD	1	10	10	21

Among the advanced neural network architectures, the *L2Light MLP* achieved the most balanced and reliable performance on the BOR classification task. It reached the highest F1-score (0.461) and AUC-ROC (0.623), metrics that are particularly meaningful in this context of class imbalance. The confusion matrix further supports this conclusion: the model correctly classifies the majority of PR and PD cases, showing improved discrimination compared to the other advanced configurations. However, the SD class remains the model's main weakness, as predictions for this category are still frequently misassigned to adjacent response levels (PR or PD). The CR class, once again, is not correctly predicted by any model due to the severe underrepresentation of this outcome in the dataset.

TTP

Table 20: Performance of classical ML models on the TTP regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
Linear Regression	**35155.467**	132.353	75.714
Ridge	36091.202	127.139	<u>74.749</u>
Lasso	<u>35922.264</u>	128.319	**74.541**
ElasticNet	39326.668	<u>125.334</u>	74.855
Decision Tree Regressor	39575.999	144.132	78.304
Random Forest Regressor	36169.648	140.036	78.052
Gradient Boosting Regressor	37004.075	140.460	78.577
XGBoost Regressor	36149.613	138.789	78.181
LightGBM Regressor	39827.604	146.716	80.585
Support Vector Regressor (SVR)	39368.327	**124.939**	75.073
K-Nearest Neighbors Regressor (kNN)	40116.481	140.640	78.326

For TTP prediction, which proved to be the most challenging regression task, the *Lasso* regression emerged as the most suitable model. Despite a slightly higher MSE compared to plain Linear Regression, it achieved better MAE and sMAPE, while providing a sparse and interpretable set of coefficients.

Table 21: Performance of conventional NN models on the TTP regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
1 HL + Dropout	41716.475	136.632	84.440
1 HL + L2 + Dropout	41652.859	133.970	81.240
1 HL + Norm Layer + L2 + Dropout	38218.637	**133.326**	83.890
1 HL + BatchNorm + L2 + Dropout	38590.560	136.496	81.160
2 HL + L2 + Dropout	40324.516	138.643	83.520
2 HL + Norm Layer + L2 + Dropout	38585.394	134.763	79.600
2 HL + Norm Layer + ElasticNet + Dropout	**37737.977**	134.426	**79.340**

Also for TTP prediction, the *2 HL + Norm Layer + ElasticNet + Dropout* model achieved the best overall performance. This architecture not only provided the lowest MSE and sMAPE, but also effectively controlled overfitting, as reflected in the stable training and validation loss curves. Its combination of depth and regularization allows it to capture complex non-linear relationships while maintaining generalization, making it the most suitable model for this challenging and highly variable task.

Table 22: Performance of advanced NN models on the TTP regression task.

Model	**MSE**	**MAE**	**sMAPE (%)**
Funnel MLP	43116.379	135.053	78.350
WideRes MLP	39797.237	139.178	80.930
Stacked Narrow MLP	40677.398	132.831	79.720
Parallel Branch MLP	**37974.128**	132.181	79.960
Self-Norm MLP	52163.729	164.963	82.900
NoisyWide MLP	41646.106	134.947	79.410
L2Light MLP	39279.156	135.039	**77.620**
Input-level attention block	45527.178	152.222	85.070
Intermediate attention block	41006.528	**129.187**	78.880
Pre-output attention block	43669.865	129.287	78.980
Multiple attention blocks	41928.502	130.769	77.680
Residual attention block	41528.337	133.740	82.000

For the TTP regression task, the *L2Light MLP* demonstrated the best overall perfor-

mance, particularly in terms of sMAPE, while maintaining competitive MSE and MAE. Its light L2 regularization, combined with a moderately deep and simple architecture, allowed the network to capture key non-linear relationships while effectively controlling overfitting.

BOR (modified: no CR class)

Table 23: Performance of classical ML models on the BOR (modified: no CR class) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							PR	SD	PD
Logistic Regression	0.360	0.447	0.360	0.375	0.567	PR	30	30	34
						SD	7	11	10
						PD	14	17	22
							PR	SD	PD
Random Forest	0.480	0.483	0.480	0.479	0.597	PR	57	19	18
						SD	14	9	5
						PD	28	7	18
							PR	SD	PD
Gradient Boosting	0.491	**0.502**	0.491	0.490	**0.606**	PR	59	22	13
						SD	13	10	5
						PD	29	7	17
							PR	SD	PD
XGBoost	0.446	0.427	0.456	0.433	0.574	PR	59	17	18
						SD	13	6	9
						PD	36	4	13
							PR	SD	PD
LGBM	0.480	0.480	0.480	0.479	0.600	PR	56	18	20
						SD	14	8	6
						PD	28	5	20
							PR	SD	PD
SVC	**0.509**	0.283	**0.509**	0.364	0.484	PR	89	5	0
						SD	28	0	0
						PD	52	1	0
							PR	SD	PD
KNN	0.297	0.091	0.297	0.139	0.509	PR	0	0	94
						SD	0	0	28
						PD	1	0	52
							PR	SD	PD
Ridge Classifier	0.337	0.421	0.337	0.353	–	PR	28	33	33
						SD	8	10	10
						PD	15	17	21
							PR	SD	PD
Gaussian Naive Bayes	0.200	0.448	0.200	0.115	0.489	PR	1	87	6
						SD	0	28	0
						PD	1	46	6
							PR	SD	PD
Decision Tree	0.400	0.455	0.400	0.418	0.521	PR	40	29	25
						SD	10	9	9
						PD	19	13	21

Among the classical machine learning models, the *Gradient Boosting* algorithm achieved the most consistent results on the BOR (modified: no CR class) task. It obtained the highest Precision (0.502) and F1-score (0.490), together with the top AUC-ROC value (0.606), confirming its ability to maintain a good balance between sensitivity and specificity across classes. The confusion matrix further supports this outcome, showing that most PR and PD instances were correctly classified, while misclassifications were primarily distributed among clinically similar categories. This suggests that, after removing the highly underrepresented CR class, Gradient Boosting was able to better capture the remaining inter-class relationships without overfitting.

Table 24: Performance of conventional NN models on the BOR (modified: no CR class) classification task.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**			
							PR	SD	PD
1 HL + Dropout	0.406	0.416	0.406	0.407	<u>0.562</u>	PR	50	22	22
						SD	14	9	5
						PD	29	12	12
							PR	SD	PD
1 HL + L2 + Dropout	0.417	0.416	0.417	0.416	0.554	PR	52	18	24
						SD	16	6	6
						PD	30	8	15
							PR	SD	PD
1 HL + Norm Layer + L2 + Dropout	0.417	0.434	0.417	<u>0.424</u>	0.561	PR	47	21	26
						SD	15	8	5
						PD	25	10	18
							PR	SD	PD
1 HL + BatchNorm + L2 + Dropout	0.406	<u>0.455</u>	0.406	0.421	0.558	PR	45	29	20
						SD	15	10	3
						PD	20	17	16
							PR	SD	PD
2 HL + L2 + Dropout	<u>0.429</u>	**0.479**	<u>0.429</u>	**0.441**	**0.580**	PR	41	26	27
						SD	10	14	4
						PD	18	15	20
							PR	SD	PD
2 HL + Norm Layer + L2 + Dropout	0.360	0.364	0.360	0.360	0.531	PR	46	22	26
						SD	17	7	4
						PD	32	11	10
							PR	SD	PD
2 HL + Norm Layer + ElasticNet + Dropout	**0.440**	0.448	**0.440**	**0.441**	0.569	PR	51	20	23
						SD	9	12	7
						PD	30	9	14

Among conventional neural architectures, the *2 HL + L2 + Dropout* model achieved the most robust and balanced performance on the BOR (modified: no CR class) classification task. It obtained the highest Precision (0.479), F1-score (0.441), and AUC-ROC (0.580), confirming its ability to generalize across the remaining classes. The confusion matrix further highlights this improvement, with a notably better prediction of the SD class

compared to previous configurations. This suggests that the combination of two hidden layers and L2 regularization, supported by dropout, provided an effective trade-off between model complexity and regularization strength, improving class-level discrimination without overfitting.

Table 25: Performance of advanced NN models on the BOR (modified: no CR class) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							PR	SD	PD
Funnel MLP	0.423	0.453	0.423	0.434	0.527	PR	44	29	21
						SD	13	9	6
						PD	25	7	21
							PR	SD	PD
WideRes MLP	0.434	0.450	0.434	0.440	0.568	PR	51	17	26
						SD	14	9	5
						PD	23	14	16
							PR	SD	PD
Stacked Narrow MLP	0.394	0.442	0.394	0.407	0.562	PR	37	27	32
						SD	10	9	9
						PD	19	11	23
							PR	SD	PD
Parallel Branch MLP	0.440	0.456	0.440	<u>0.445</u>	0.573	PR	49	22	23
						SD	14	11	3
						PD	26	10	17
							PR	SD	PD
Self-Norm MLP	0.457	<u>0.465</u>	0.457	**0.461**	0.583	PR	52	18	24
						SD	13	7	8
						PD	25	7	21
							PR	SD	PD
NoisyWide MLP	0.423	0.442	0.423	0.431	**0.594**	PR	46	22	26
						SD	17	7	4
						PD	23	9	21
							PR	SD	PD
L2Light MLP	0.394	0.446	0.394	0.403	0.584	PR	38	30	25
						SD	8	16	4
						PD	21	17	15
							PR	SD	PD
Input-level attention block	0.411	**0.491**	0.411	0.431	<u>0.585</u>	PR	38	29	27
						SD	5	11	12
						PD	16	14	23
							PR	SD	PD
Intermediate attention block	<u>0.491</u>	0.386	<u>0.491</u>	0.423	0.512	PR	74	0	20
						SD	23	0	5
						PD	41	0	12
							PR	SD	PD
Pre-output attention block	0.434	0.339	0.434	0.376	0.536	PR	66	0	28
						SD	24	0	4
						PD	43	0	10
							PR	SD	PD
Multiple attention blocks	**0.537**	0.289	**0.537**	0.375	0.499	PR	94	0	0
						SD	28	0	0
						PD	53	0	0
							PR	SD	PD
Residual attention block	0.411	0.460	0.411	0.426	0.548	PR	43	24	27
						SD	10	12	6
						PD	19	17	17

Among advanced neural architectures on the BOR (modified: no CR class) task, the *Self-Norm MLP* model shows the most balanced and clinically meaningful performance. It achieves the highest F1-score (0.461) and competitive Precision (0.465) and AUC-ROC (0.583), indicating a robust generalization across classes. The confusion matrix highlights that the majority of PR and PD samples are correctly classified, while the SD class

remains challenging. Overall, this configuration effectively balances the need for class-level accuracy with the prevention of overfitting in this highly imbalanced dataset.

BOR (modified: merging CR class with PR one)

Table 26: Performance of classical ML models on the BOR (modified: CR class merged with PR one) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							PR	SD	PD
Logistic Regression	0.500	0.468	0.500	**0.477**	0.593	PR	67	12	21
						SD	29	5	6
						PD	20	3	19
							PR	SD	PD
Random Forest	0.456	0.398	0.456	0.423	0.594	PR	63	13	24
						SD	34	1	5
						PD	21	2	19
							PR	SD	PD
Gradient Boosting	0.462	0.449	0.462	0.447	**0.605**	PR	56	13	31
						SD	27	5	8
						PD	16	3	23
							PR	SD	PD
XGBoost	0.495	0.469	0.495	0.473	0.584	PR	62	11	27
						SD	27	4	9
						PD	15	3	24
							PR	SD	PD
LGBM	0.484	0.454	0.484	0.462	0.580	PR	61	14	25
						SD	28	4	8
						PD	18	1	23
							PR	SD	PD
SVC	**0.550**	0.302	**0.550**	0.390	0.505	PR	100	0	0
						SD	40	0	0
						PD	42	0	0
							PR	SD	PD
KNN	0.291	0.422	0.291	0.216	0.517	PR	14	0	86
						SD	4	0	36
						PD	3	0	39
							PR	SD	PD
Ridge Classifier	0.401	0.477	0.401	0.408	–	PR	34	34	32
						SD	9	16	15
						PD	12	7	23
							PR	SD	PD
Gaussian Naive Bayes	0.264	**0.695**	0.264	0.156	0.542	PR	1	88	11
						SD	0	39	1
						PD	0	34	8
							PR	SD	PD
Decision Tree	0.429	0.423	0.429	0.426	0.520	PR	56	29	22
						SD	29	6	5
						PD	17	9	16

For the BOR task with the CR class merged into PR, *Logistic Regression* achieves the best overall balance in terms of F1-score (0.477) and Recall (0.500), making it the most reliable classical model for this modified setup. While the confusion matrix reveals that SD

remains challenging and PD predictions are slightly less accurate compared to *Gradient Boosting* or *XGBoost*, *Logistic Regression* provides the most consistent and clinically meaningful predictions overall, especially for the combined PR category.

Table 27: Performance of conventional NN models on the BOR (modified: CR class merged with PR one) classification task.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**			
							PR	SD	PD
1 HL + Dropout	0.418	0.418	0.418	0.414	0.574	PR	56	16	28
						SD	20	4	16
						PD	19	7	16
							PR	SD	PD
1 HL + L2 + Dropout	<u>0.478</u>	**0.484**	<u>0.478</u>	<u>0.478</u>	0.586	PR	60	15	25
						SD	20	8	12
						PD	14	9	19
							PR	SD	PD
1 HL + Norm Layer + L2 + Dropout	0.445	0.452	0.445	0.443	0.570	PR	59	12	29
						SD	17	6	17
						PD	20	6	16
							PR	SD	PD
1 HL + BatchNorm + L2 + Dropout	0.440	<u>0.479</u>	0.440	0.451	<u>0.591</u>	PR	51	14	35
						SD	15	11	14
						PD	15	9	18
							PR	SD	PD
2 HL + L2 + Dropout	**0.506**	**0.484**	**0.506**	**0.486**	**0.593**	PR	67	12	21
						SD	23	6	11
						PD	22	1	19
							PR	SD	PD
2 HL + Norm Layer + L2 + Dropout	0.396	0.424	0.396	0.403	0.571	PR	49	17	34
						SD	19	6	15
						PD	15	10	17
							PR	SD	PD
2 HL + Norm Layer + ElasticNet + Dropout	0.473	0.477	0.473	0.468	0.586	PR	59	10	31
						SD	23	9	8
						PD	20	4	18

For the BOR task with CR merged into PR, *2 HL + L2 + Dropout* clearly stands out as the best conventional NN model. It achieves the highest Accuracy (0.506), F1-score (0.486), and Recall (0.506), and its confusion matrix shows a good balance across all three classes, especially for PR and PD. It's worth noting that *1 HL + L2 + Dropout* is competitive in terms of SD predictions: the model captures this class slightly better, but at the cost of lower performance on PR, highlighting a trade-off between overall metrics and class-specific recall.

Table 28: Performance of advanced NN models on the BOR (modified: CR class merged with PR one) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							PR	SD	PD
Funnel MLP	0.396	0.409	0.396	0.400	0.544	PR	54	18	28
						SD	17	6	17
						PD	22	8	12
							PR	SD	PD
WideRes MLP	0.434	0.443	0.434	0.437	0.550	PR	56	16	28
						SD	22	8	10
						PD	16	11	15
							PR	SD	PD
Stacked Narrow MLP	0.495	<u>0.489</u>	0.495	<u>0.489</u>	<u>0.607</u>	PR	64	10	26
						SD	23	10	7
						PD	18	8	16
							PR	SD	PD
Parallel Branch MLP	0.451	0.444	0.451	0.447	0.566	PR	63	17	20
						SD	20	10	10
						PD	20	13	9
							PR	SD	PD
Self-Norm MLP	0.467	0.456	0.467	0.460	0.556	PR	62	15	23
						SD	25	11	4
						PD	21	9	12
							PR	SD	PD
NoisyWide MLP	0.495	0.481	0.495	0.481	0.587	PR	68	7	25
						SD	22	7	11
						PD	20	7	15
							PR	SD	PD
L2Light MLP	0.517	**0.511**	0.517	**0.511**	0.600	PR	67	12	21
						SD	18	11	11
						PD	20	6	16
							PR	SD	PD
Input-level attention block	0.434	0.445	0.434	0.437	0.525	PR	54	29	27
						SD	19	8	13
						PD	19	6	17
							PR	SD	PD
Intermediate attention block	**0.571**	0.442	**0.571**	0.479	0.606	PR	91	0	9
						SD	38	0	2
						PD	29	0	13
							PR	SD	PD
Pre-output attention block	<u>0.560</u>	0.427	<u>0.560</u>	0.452	**0.626**	PR	94	0	6
						SD	38	0	2
						PD	34	0	8
							PR	SD	PD
Multiple attention blocks	0.533	0.397	0.533	0.452	0.538	PR	85	0	15
						SD	29	0	11
						PD	30	0	12
							PR	SD	PD
Residual attention block	0.462	0.453	0.462	0.448	0.604	PR	59	9	32
						SD	23	4	13
						PD	16	5	21

For the BOR task with CR merged into PR, among advanced NN models, *L2Light MLP* emerges as the preferred choice. While *Stacked Narrow MLP* is competitive in terms of F1-score and overall metrics, *L2Light* shows a slightly better balance in the confusion

matrix: more PR instances are correctly classified, and SD predictions, although just one more correct instance, improve the model's practical reliability.

Responder Groups (stratification from TTP)

Table 29: Performance of classical ML models on the Responder Groups (stratification by k-means from TTP) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
Logistic Regression	0.474	**0.570**	0.474	0.501	0.585	bad	32	25	14
						medium	12	22	4
						good	2	4	1
							bad	medium	good
Random Forest	0.586	0.486	0.586	0.500	**0.644**	bad	64	7	0
						medium	34	4	0
						good	6	1	0
							bad	medium	good
Gradient Boosting	0.578	0.518	0.578	0.540	0.604	bad	56	15	10
						medium	27	11	0
						good	5	2	0
							bad	medium	good
XGBoost	0.578	0.527	0.578	0.546	0.609	bad	55	15	1
						medium	26	12	0
						good	5	2	0
							bad	medium	good
LGBM	0.586	0.530	0.586	**0.550**	0.592	bad	57	13	1
						medium	27	11	0
						good	4	3	0
							bad	medium	good
SVC	0.328	0.107	0.327	0.162	0.500	bad	0	71	0
						medium	0	38	0
						good	0	7	0
							bad	medium	good
KNN	**0.612**	0.375	**0.612**	0.465	0.514	bad	71	0	0
						medium	38	0	0
						good	7	0	0
							bad	medium	good
Ridge Classifier	0.448	0.553	0.448	0.480	–	bad	31	25	15
						medium	12	19	7
						good	3	2	2
							bad	medium	good
Gaussian Naive Bayes	0.164	0.489	0.164	0.153	0.495	bad	4	32	35
						medium	1	10	27
						good	1	1	5
							bad	medium	good
Decision Tree	0.509	0.477	0.509	0.492	0.557	bad	46	25	0
						medium	25	13	0
						good	5	2	0

Among classical ML models, *XGBoost* achieved the best overall balance between performance metrics (notably F1-score and AUC-ROC), which are the most meaningful indicators in this imbalanced setting. However, it failed to correctly identify any good responders. In contrast, it's notable to underline that *Logistic Regression*, though less performant overall, was one of the very few models capable of detecting this minority class.

Table 30: Performance of conventional NN models on the classification on the Responder Groups (stratification by k-means from TTP) task.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**			
							bad	medium	good
1 HL + Dropout	0.534	0.502	0.535	0.518	0.627	bad	48	22	1
						medium	24	14	0
						good	5	2	0
							bad	medium	good
1 HL + L2 + Dropout	0.578	**0.537**	0.578	**0.555**	0.616	bad	53	18	0
						medium	23	14	1
						good	4	3	0
							bad	medium	good
1 HL + Norm Layer + L2 + Dropout	0.552	0.516	0.552	0.531	0.599	bad	52	17	2
						medium	25	12	1
						good	4	3	0
							bad	medium	good
1 HL + BatchNorm + L2 + Dropout	**0.586**	0.530	**0.586**	0.551	0.525	bad	56	15	0
						medium	26	12	0
						good	5	2	0
							bad	medium	good
2 HL + L2 + Dropout	0.560	0.511	0.560	0.533	0.629	bad	53	18	0
						medium	26	12	0
						good	3	4	0
							bad	medium	good
2 HL + Norm Layer + L2 + Dropout	0.569	0.531	0.569	0.548	0.628	bad	53	17	1
						medium	24	13	1
						good	3	4	0
							bad	medium	good
2 HL + Norm Layer + ElasticNet + Dropout	0.569	0.534	0.569	0.539	**0.660**	bad	56	15	0
						medium	28	9	1
						good	2	4	1

Here, the best-performing model is the *2 HL + Norm Layer + ElasticNet + Dropout*: this model not only achieved the highest AUC-ROC among all conventional NN architectures, but it also showed a more balanced performance across all metrics. Importantly, in the confusion matrix, it is the only model able to correctly predict at least one instance of the good responder class, which is clinically crucial given the extreme class imbalance.

Table 31: Performance of advanced NN models on the Responder Groups (stratification by k-means from TTP) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
Funnel MLP	0.517	0.453	0.517	0.477	0.515	bad	53	17	1
						medium	31	7	0
						good	5	2	0
							bad	medium	good
WideRes MLP	0.526	0.477	0.526	0.499	0.589	bad	51	20	0
						medium	28	10	0
						good	3	4	0
							bad	medium	good
Stacked Narrow MLP	0.552	0.520	0.552	0.536	0.572	bad	50	21	0
						medium	23	14	1
						good	3	4	0
							bad	medium	good
Parallel Branch MLP	0.560	0.564	0.560	0.539	0.573	bad	53	18	0
						medium	27	11	0
						good	3	3	1
							bad	medium	good
Self-Norm MLP	**0.586**	**0.574**	**0.586**	**0.577**	**0.635**	bad	51	19	1
						medium	21	16	1
						good	3	3	1
							bad	medium	good
NoisyWide MLP	0.526	0.464	0.523	0.487	0.556	bad	54	16	1
						medium	30	7	0
						good	4	3	0
							bad	medium	good
L2Light MLP	0.500	0.460	0.500	0.479	0.569	bad	48	23	0
						medium	28	10	0
						good	3	4	0
							bad	medium	good
Input-level attention block	0.552	0.518	0.552	0.534	0.586	bad	49	22	0
						medium	23	15	0
						good	4	3	0
							bad	medium	good
Intermediate attention block	0.526	0.495	0.526	0.510	0.583	bad	46	25	0
						medium	23	15	0
						good	6	1	0
							bad	medium	good
Pre-output attention block	0.535	0.510	0.535	0.523	0.547	bad	46	25	0
						medium	22	16	0
						good	4	3	0
							bad	medium	good
Multiple attention blocks	0.517	0.494	0.517	0.505	0.557	bad	45	26	0
						medium	23	15	0
						good	4	3	0
							bad	medium	good
Residual attention block	0.491	0.475	0.491	0.483	0.539	bad	44	25	2
						medium	24	13	1
						good	5	2	0

In this case, the *Self-Norm MLP* clearly stands out as the best-performing model among the advanced neural network architectures. It achieves the highest scores across all key metrics demonstrating both robust overall performance and stability; importantly, the

confusion matrix shows that, unlike most other models, it is able to correctly predict at least one instance of the good responder class while maintaining strong predictions for the bad and medium groups.

Responder Groups (stratification from log-transformed TTP)

Table 32: Performance of classical ML models on the Responder Groups (stratification by k-means from log-transformed TTP) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
Logistic Regression	0.513	0.513	0.513	**0.510**	0.653	bad	15	6	10
						medium	11	19	14
						good	5	11	26
							bad	medium	good
Random Forest	**0.521**	**0.528**	**0.521**	0.505	**0.675**	bad	8	13	10
						medium	5	24	15
						good	1	12	29
							bad	medium	good
Gradient Boosting	0.436	0.441	0.436	0.428	0.609	bad	9	13	9
						medium	8	17	19
						good	1	16	25
							bad	medium	good
XGBoost	0.504	0.506	0.504	0.477	0.659	bad	5	19	7
						medium	4	26	14
						good	1	13	28
							bad	medium	good
LGBM	0.453	0.438	0.453	0.433	0.630	bad	6	12	13
						medium	9	19	16
						good	1	13	28
							bad	medium	good
SVC	0.359	0.349	0.359	0.348	0.484	bad	12	9	10
						medium	13	9	22
						good	9	12	21
							bad	medium	good
KNN	0.385	0.502	0.385	0.224	0.518	bad	0	31	0
						medium	0	44	0
						good	0	41	1
							bad	medium	good
Ridge Classifier	0.444	0.432	0.444	0.421	–	bad	16	5	10
						medium	14	9	21
						good	6	9	27
							bad	medium	good
Gaussian Naive Bayes	0.444	0.440	0.444	0.423	0.588	bad	4	11	15
						medium	6	23	15
						good	1	17	24
							bad	medium	good
Decision Tree	0.402	0.400	0.402	0.397	0.541	bad	8	11	12
						medium	12	20	12
						good	2	21	19

In this setting, the *Logistic Regression* appears to be the most balanced and reliable model, despite *Random Forest* achieving slightly higher global metrics such as accuracy, precision, and AUC-ROC. What really supports this choice is the confusion matrix: *Logistic*

Regression correctly predicts 15 instances of the bad class, 19 of the medium, and 26 of the good, showing a more even distribution across classes. In contrast, *Random Forest* underperforms on the bad class, correctly predicting only 8 instances, which compromises the model's reliability in distinguishing non-responders.

Table 33: Performance of conventional NN models on the classification on the Responder Groups (stratification by k-means from log-transformed TTP) task.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**			
							bad	medium	good
1 HL + Dropout	0.393	0.405	0.393	0.381	0.569	bad	6	10	15
						medium	6	19	19
						good	1	20	21
							bad	medium	good
1 HL + L2 + Dropout	0.385	0.405	0.385	0.376	0.578	bad	8	9	14
						medium	6	14	24
						good	1	18	23
							bad	medium	good
1 HL + Norm Layer + L2 + Dropout	0.376	0.385	0.376	0.375	<u>0.605</u>	bad	9	11	11
						medium	9	19	16
						good	2	24	16
							bad	medium	good
1 HL + BatchNorm + L2 + Dropout	**0.444**	**0.439**	**0.444**	**0.426**	0.597	bad	7	11	13
						medium	8	17	19
						good	1	13	28
							bad	medium	good
2 HL + L2 + Dropout	0.393	0.417	0.393	0.391	0.583	bad	9	9	13
						medium	6	16	22
						good	1	20	21
							bad	medium	good
2 HL + Norm Layer + L2 + Dropout	<u>0.419</u>	<u>0.419</u>	<u>0.419</u>	<u>0.415</u>	**0.611**	bad	10	7	14
						medium	11	17	16
						good	2	18	22
							bad	medium	good
2 HL + Norm Layer + ElasticNet + Dropout	0.393	0.390	0.393	0.386	0.580	bad	8	9	14
						medium	11	16	17
						good	2	18	22

For this task, the two most competitive models appear to be *1 HL + BatchNorm + L2 + Dropout* and *2 HL + Norm Layer + L2 + Dropout*. While their overall metrics are fairly consistent and comparable, the key difference lies in the confusion matrices. The 2 HL model predicts slightly more instances of the bad class, but at the expense of the good class, which is underpredicted. In contrast, the 1 HL model correctly predicts a higher number of good cases, which are the rare and most clinically relevant responders, without a significant drop in performance on the bad class. Therefore, considering the importance of capturing the good class in this highly imbalanced setting, the *1 HL + BatchNorm + L2 + Dropout* model is preferred.

Table 34: Performance of advanced NN models on the Responder Groups (stratification by k-means from log-transformed TTP) classification task.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
Funnel MLP	0.385	0.377	0.385	0.370	0.566	bad	6	11	14
						medium	8	15	21
						good	3	15	24
							bad	medium	good
WideRes MLP	0.419	0.410	0.419	0.401	**0.637**	bad	6	11	14
						medium	8	17	19
						good	2	14	26
							bad	medium	good
Stacked Narrow MLP	0.444	0.428	0.444	0.425	0.614	bad	5	9	17
						medium	8	23	13
						good	2	16	24
							bad	medium	good
Parallel Branch MLP	0.393	0.390	0.393	0.380	0.626	bad	6	10	15
						medium	8	17	19
						good	2	17	23
							bad	medium	good
Self-Norm MLP	0.359	0.361	0.359	0.351	0.581	bad	6	13	12
						medium	8	17	19
						good	2	21	19
							bad	medium	good
NoisyWide MLP	0.393	0.395	0.393	0.385	0.605	bad	9	7	15
						medium	10	14	20
						good	2	17	23
							bad	medium	good
L2Light MLP	0.402	0.409	0.402	0.393	0.588	bad	9	5	17
						medium	9	14	21
						good	2	16	24
							bad	medium	good
Input-level attention block	**0.496**	**0.493**	**0.496**	**0.484**	<u>0.636</u>	bad	9	9	13
						medium	8	21	15
						good	2	12	28
							bad	medium	good
Intermediate attention block	0.393	0.391	0.393	0.386	0.552	bad	7	16	8
						medium	10	17	17
						good	2	18	22
							bad	medium	good
Pre-output attention block	<u>0.462</u>	<u>0.485</u>	<u>0.462</u>	<u>0.468</u>	0.628	bad	15	14	2
						medium	16	19	9
						good	4	18	20
							bad	medium	good
Multiple attention blocks	0.385	0.393	0.385	0.379	0.563	bad	9	9	13
						medium	8	14	22
						good	2	18	22
							bad	medium	good
Residual attention block	0.368	0.371	0.368	0.364	0.597	bad	10	8	13
						medium	9	13	22
						good	4	18	20

For this task, the most competitive advanced NN models are *Input-level attention block* and *Pre-output attention block*. While both models achieve similar overall metrics, the key difference lies in the confusion matrices and the interpretability of the attention mech-

anism. The *Pre-output* model predicts more instances of the bad class, but this comes at the expense of underpredicting the good class, which is the rare and most clinically relevant category. In contrast, the *Input-level attention block* not only maintains competitive metrics but also correctly identifies a higher number of good cases. From a conceptual perspective, applying attention at the input level allows the model to assess the importance of features directly at the beginning of the network, providing more interpretable insights into which variables drive the predictions, rather than only capturing their impact at the final layers. Therefore, considering both performance on the minority class and interpretability, the *Input-level attention block* is preferred.

5.5 Model Performance Comparison

To consolidate the findings obtained from the extensive experimental analysis, the best-performing model from each architectural category, classical Machine Learning, conventional Neural Networks and advanced Neural Network architectures, was selected for further comparison. This summarizing step enables a clearer evaluation of how model complexity and architectural design influence predictive performance on the given clinical endpoint. For each clinical endpoint, a dedicated summary table will illustrate the best model identified within each architectural family. The choice of these models is grounded in the performance analysis and reasoning discussed in the previous sections.

5.5.1 OS Status

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix		
							0	1
Logistic Regression	0.691	0.718	0.667	0.691	0.739	0	28	11
						1	14	28
							0	1
2 HL + L2 + Dropout	0.679	0.711	0.643	0.675	0.692	0	28	11
						1	15	27
							0	1
Input-level attention block	**0.716**	**0.721**	**0.738**	**0.729**	**0.741**	0	27	12
						1	11	31

Table 35: Performance of the best model from each architectural family on the OS Status classification task.

Among the evaluated architectures, the *input-level attention block*, representing the advanced neural network models, achieved the best performance across all evaluated metrics; this suggests that the direct incorporation of an attention mechanism at the feature-input stage effectively enhanced the model's ability to discern relevant patterns associated with survival outcomes. Given that the OS status represents a relatively straightforward binary endpoint, it is plausible that such a focused architecture can capture the underlying structure of the data more efficiently than deeper or more regularized configurations, without overfitting.

5.5.2 OS Duration

Model	MSE	MAE	sMAPE (%)
XGBoost Regressor	43.471	5.008	68.340
2 HL + Norm Layer + ElasticNet + Dropout	44.809	4.499	66.790
Multiple attention blocks	**43.282**	**4.199**	**63.280**

Table 36: Performance of the best model from each architectural family on the OS Duration regression task.

For the OS duration regression task, the *Multiple attention blocks* architecture achieved the best overall performance, with notably lower MSE, MAE, and sMAPE values compared to both classical and conventional models. This result suggests that attention-based mechanisms can effectively capture intricate dependencies among clinical and biological predictors, offering an adaptive weighting of feature contributions that enhances regression accuracy. Unlike the OS status task, where simplicity may have favored direct *Input-level attention block* architecture, the continuous nature of OS duration appears to benefit from a more expressive architecture capable of learning hierarchical feature interactions.

5.5.3 PFS

Model	MSE	MAE	sMAPE (%)
Support Vector Regressor (SVR)	17.783	2.582	60.795
2 HL + Norm Layer + ElasticNet + Dropout	21.013	2.741	60.720
NoisyWide MLP	**17.223**	**2.518**	**59.500**

Table 37: Performance of the best model from each architectural family on the PFS regression task.

For the PFS regression task, the *NoisyWide MLP* outperformed both classical and conventional architectures, achieving the lowest MSE, MAE, and sMAPE values. This result is particularly relevant given the intrinsic complexity of the PFS variable, which is affected by heterogeneous and often noisy clinical dynamics. The superior performance of this advanced neural architecture suggests that incorporating noise during training and widening the network improves generalization and robustness, enabling the model to better capture subtle nonlinear patterns underlying progression dynamics. Overall, these findings reinforce the idea that attention or noise-enhanced deep architectures can be particularly advantageous for survival regression tasks involving high variability and nonlinearity.

5.5.4 BOR (original)

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix				
							CR	PR	SD	PD
						CR	0	4	0	1
Logistic Regression	0.445	0.424	0.445	0.431	0.602	PR	3	56	14	22
						SD	0	26	6	8
						PD	1	17	5	19
							CR	PR	SD	PD
						CR	0	2	2	1
2 HL + Norm Layer + L2 + Dropout	**0.484**	**0.475**	**0.484**	**0.476**	**0.624**	PR	1	60	10	24
						SD	1	18	10	11
						PD	0	18	6	18
							CR	PR	SD	PD
						CR	0	1	3	1
L2Light MLP	0.478	0.462	0.478	0.461	0.623	PR	1	60	9	25
						SD	1	19	6	14
						PD	0	19	2	21

Table 38: Performance of the best model from each architectural family on the BOR (original) classification task.

Among the tested models, the conventional neural network (*2 HL + Norm Layer + L2 + Dropout*) achieved the best overall performance on the BOR classification task. This result suggests that, for moderately complex endpoints such as BOR, where response categories are interrelated and not strictly separable in a linear fashion, architectures of intermediate complexity can offer the most favorable trade-off between expressiveness and generalization. While classical models may be too rigid to capture subtle nonlinear dependencies, advanced networks might overfit due to limited data availability and noise. The conventional NN, by integrating regularization and normalization layers, appears to provide sufficient flexibility to model clinically meaningful distinctions while maintaining stability in the predictions.

5.5.5 TTP

Model	MSE	MAE	sMAPE (%)
Lasso	**35922.264**	**128.319**	**74.541**
2 HL + Norm Layer + ElasticNet + Dropout	37737.977	134.426	79.340
L2Light MLP	39279.156	135.039	77.620

Table 39: Performance of the best model from each architectural family on the TTP regression task.

Among the evaluated models, the *Lasso Regressor* achieved the best overall performance on the TTP regression task, outperforming both the conventional and advanced neural architectures. This finding highlights an important aspect of modeling complex clinical endpoints: when the target variable is highly heterogeneous and potentially noisy, simpler models with strong regularization can provide more stable and reliable predictions. In this context, the Lasso's sparsity-inducing property likely contributed to filtering out irrelevant or weakly informative predictors, reducing variance and improving generalization. Conversely, deeper or wider networks may have struggled to converge toward meaningful representations due to the limited data scale and intrinsic variability of the TTP endpoint.

5.5.6 BOR (modified)

No CR class

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**	**Confusion Matrix**			
							PR	SD	PD
Gradient Boosting	**0.491**	**0.502**	**0.491**	**0.490**	**0.606**	PR	59	22	13
						SD	13	10	5
						PD	29	7	17
							PR	SD	PD
2 HL + L2 + Dropout	0.429	0.479	0.429	0.441	0.580	PR	41	26	27
						SD	10	14	4
						PD	18	15	20
							PR	SD	PD
Self-Norm MLP	0.457	0.465	0.457	0.461	0.583	PR	52	18	24
						SD	13	7	8
						PD	25	7	21

Table 40: Performance of the best model from each architectural family on the BOR (modified: no CR class) classification task.

Among the tested models, *Gradient Boosting* achieved the best overall performance on the BOR task after the exclusion of the CR class: however, the confusion matrices reveal that its advantage mainly arises from accurate predictions on the predominant PR category, while its sensitivity toward SD and PD remains limited. Conversely, the conventional and advanced neural networks, despite slightly lower global metrics, produced a more balanced classification among the less represented response groups. This suggests that while ensemble-based classical models remain strong general-purpose learners, neural approaches might offer a finer differentiation when class-level interpretability and balance are prioritized.

CR merged into PR

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							PR	SD	PD
Logistic Regression	0.500	0.468	0.500	0.477	0.593	PR	67	12	21
						SD	29	5	6
						PD	20	3	19
							PR	SD	PD
2 HL + L2 + Dropout	0.506	0.484	0.506	0.486	0.593	PR	67	12	21
						SD	23	6	11
						PD	22	1	19
							PR	SD	PD
L2Light MLP	**0.517**	**0.511**	**0.517**	**0.511**	**0.600**	PR	67	12	21
						SD	18	11	11
						PD	20	6	16

Table 41: Performance of the best model from each architectural family on the BOR (modified: CR class merged with PR one) classification task.

Among the compared models, the *L2Light MLP* emerged as the most effective approach for the modified BOR task (with CR merged into PR), achieving the best overall performance across all reported metrics. Notably, this advanced neural network exhibits superior predictive capability for both the PR and SD categories, suggesting an enhanced ability to capture intermediate response patterns that are typically harder to model with linear or shallow architectures. However, its predictions for the PD group remain slightly less accurate compared to classical or conventional models, possibly due to the lower sample representation and the higher variability characterizing progressive cases. Overall, this result indicates that advanced neural models can leverage their architectural flexibility to better model nuanced response dynamics, while still facing challenges in rare or heterogeneous subgroups.

5.5.7 Responder Groups

From TTP

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
XGBoost	0.578	0.527	0.578	0.546	0.609	bad	55	15	1
						medium	26	12	0
						good	5	2	0
							bad	medium	good
2 HL + Norm Layer + ElasticNet + Dropout	0.569	0.534	0.569	0.539	**0.660**	bad	56	15	0
						medium	28	9	1
						good	2	4	1
							bad	medium	good
Self-Norm MLP	**0.586**	**0.574**	**0.586**	**0.577**	0.635	bad	51	19	1
						medium	21	16	1
						good	3	3	1

Table 42: Performance of the best model from each architectural family on the Responder Groups (stratification by k-means from TTP) classification task.

Among the tested models, the *Self-Norm MLP* stands out as the most effective architecture for predicting the Responder Quality Groups derived from TTP. It achieves the highest overall performance across all major metrics and displays a more balanced confusion matrix, correctly identifying a small number of good responders, a class typically underrepresented and hard to detect. This result highlights the ability of advanced neural networks to capture subtle, nonlinear relationships even in the presence of severe class imbalance, outperforming both classical and conventional architectures.

From log-transformed TTP

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Confusion Matrix			
							bad	medium	good
Logistic Regression	**0.513**	**0.513**	**0.513**	**0.510**	**0.653**	bad	15	6	10
						medium	11	19	14
						good	5	11	26
							bad	medium	good
1 HL + BatchNorm + L2 + Dropout	0.444	0.439	0.444	0.426	0.597	bad	7	11	13
						medium	8	17	19
						good	1	13	28
							bad	medium	good
Input-level attention block	0.496	0.493	0.496	0.484	0.636	bad	9	9	13
						medium	8	21	15
						good	2	12	28

Table 43: Performance of the best model from each architectural family on the Responder Groups (stratification by k-means from log-transformed TTP) classification task.

Interestingly, when the responder groups were defined from the log-transformed TTP, the classical *Logistic Regression* model outperformed both neural network architectures. This

outcome suggests that the logarithmic transformation effectively linearized the relationship between predictors and the outcome, allowing a simpler model to capture the main discriminative patterns. Moreover, the reduced complexity of the transformed task may have limited the advantages of deep models, which could instead overfit minor variations in a relatively small and imbalanced dataset. This highlights how, in certain clinically grounded settings, an appropriate data transformation can make a simpler model both more stable and more interpretable.

5.6 Interpretability and Insights

To gain insight into the decision mechanisms of the best-performing models identified in the previous section, an interpretability analysis was conducted using SHAP values, as discussed in Section 4.3. Given the clinical relevance and diversity of predictive tasks, three key response variables were selected for this analysis: the original BOR classes, and the responder quality groups derived from both the raw and the log-transformed TTP variables. These endpoints capture distinct yet complementary aspects of treatment outcome, categorical tumor response, temporal dynamics of disease progression and their normalized representation, making them particularly suitable for interpretability assessment.

BOR (original)

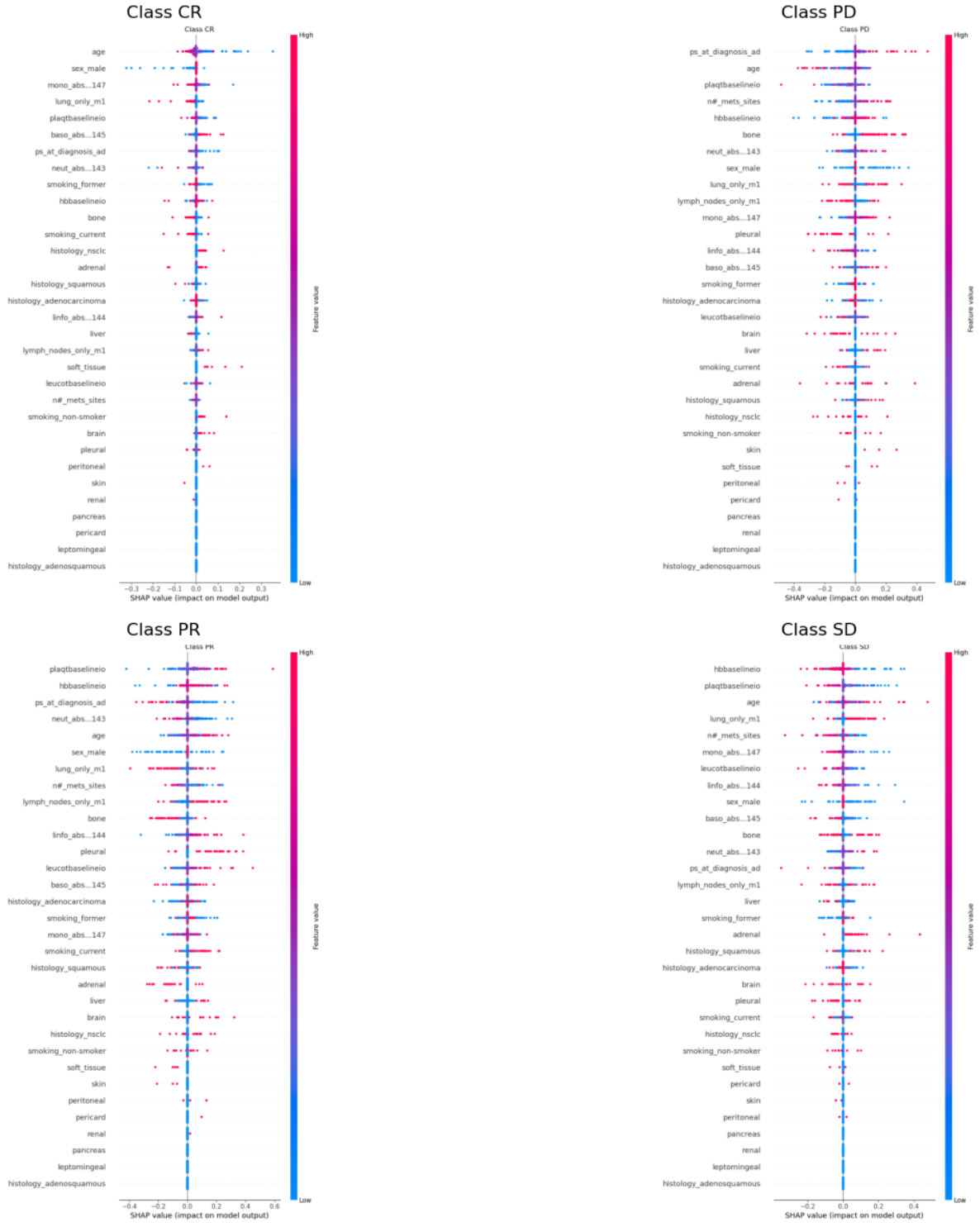


Figure 8: SHAP summary plots for the BOR (original) classification task obtained with the best-performing conventional neural network.

Across the four response categories (CR, PR, SD, PD), several variables consistently emerge as influential, while others show class-specific effects. In details, the image shows:

- **Class CR**

  Patients predicted as complete responders are mainly characterized by lower monocyte, basophil, and platelet counts at diagnosis, all positively influencing the model's confidence toward CR. In contrast, higher platelet levels at baseline, a squamous histology, and a history of former smoking are negatively associated with this class. Overall, the model identifies complete responders as patients with lower systemic inflammation and non-squamous tumor types.

- **Class PR**

  For partial responders, higher hemoglobin and platelet levels, older age, and the presence of pleural metastases have a positive impact on the predicted probability of PR. Conversely, bone, brain, and liver metastases contribute negatively, suggesting that patients with a more limited or localized disease are more likely to experience partial tumor shrinkage. This indicates that PR is associated with a favorable hematologic profile and less aggressive metastatic spread.

- **Class SD**

  Predictions of stable disease are driven by higher hemoglobin, platelet, and neutrophil counts, which positively affect the likelihood of SD. On the contrary, the presence of pleural involvement and an adenocarcinoma histology are linked to negative SHAP values, pushing predictions away from SD. This pattern suggests an intermediate biological profile, in which systemic inflammation is present but without strong markers of response or progression.

- **Class PD**

  Patients predicted as progressing are mainly associated with a poor performance status at diagnosis, higher platelet and hemoglobin levels, and male sex, all exerting positive effects on PD predictions. Conversely, lower neutrophil and monocyte counts and isolated lymph node involvement have negative impacts, reducing the model's confidence toward progression. Overall, PD patients are identified as having a more aggressive clinical phenotype, marked by elevated hematologic parameters and reduced performance status.

In summary, the SHAP analysis highlights that:
1. **Hematological and inflammatory markers** (such as neutrophil, platelet, and hemoglobin levels) are consistently influential across all response categories.
2. **Tumor histology** and the **pattern of metastatic spread** act as key discriminants between responders and non-responders.
3. The model captures **clinically meaningful trends**, linking lower systemic inflammation and non-squamous histotypes to favorable outcomes, while poor performance status and high inflammatory burden correspond to progression.

Responder Groups (from TTP)

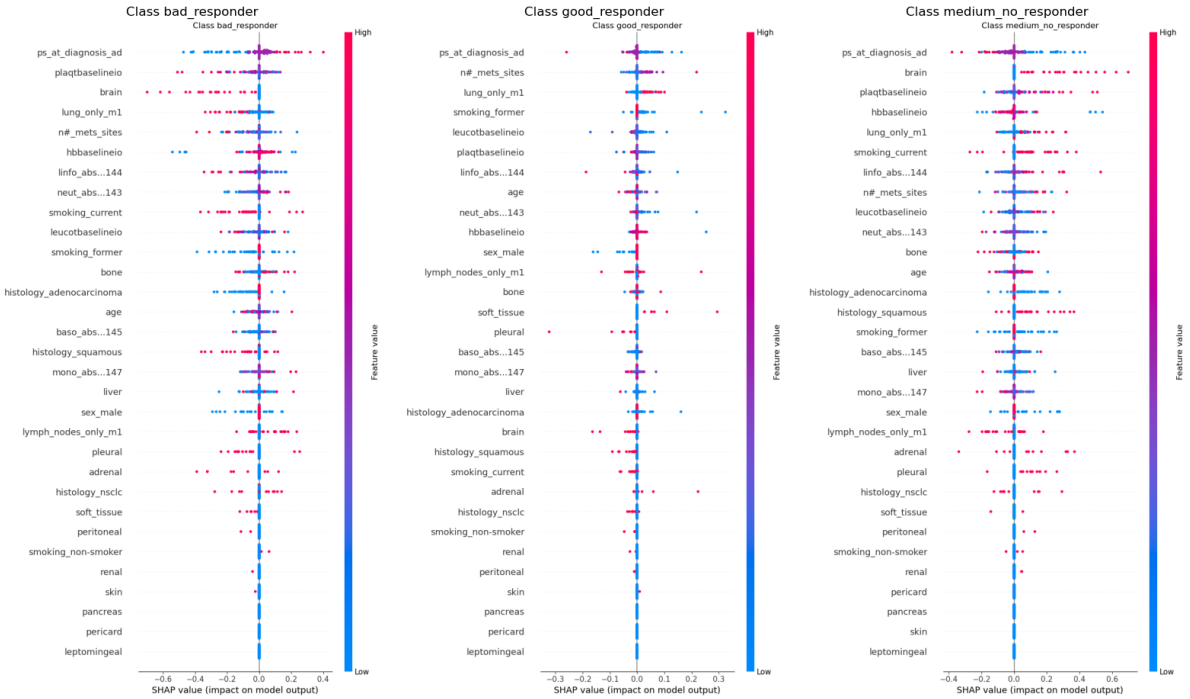


Figure 9: SHAP summary plots for the Responder Groups (from TTP) classification task obtained with the best-performing advanced neural architecture.

Across the three response categories several variables consistently emerge as influential, while others show class-specific effects.
In details:

- **Bad (responder) Class**
  The model identifies performance status at diagnosis and platelet baseline levels as key variables with a strong positive SHAP impact, indicating that worse performance and higher platelet counts are associated with poorer treatment outcomes. Additionally, the presence of brain or lung-only metastases and a higher number of metastatic sites also contribute positively to this class, consistently describing patients with a more aggressive disease profile.

- **Medium (responder) Class**
  In this group, similar variables remain relevant but with attenuated effects. Performance status and hemoglobin baseline values tend to balance between protective and adverse influences, while smoking habits (especially current smoking) and metastatic patterns play a moderate role. This intermediate class reflects a mixed clinical picture, capturing patients whose response to treatment is neither strongly positive nor clearly negative.

- **Good (responder) Class**
  The SHAP values show that lower performance status, fewer metastatic sites, and lung-only metastases are the main drivers of a favorable response, together with a younger age and lower platelet counts. These features delineate patients with a limited disease burden and overall better health, aligning well with expected clinical correlations.

Overall, the model effectively distinguishes responder groups by leveraging clinically coherent patterns, where **disease extent, baseline functional status, and systemic inflammatory markers** emerge as the most influential determinants of treatment response.

Responder Groups (from log-transformed TTP)

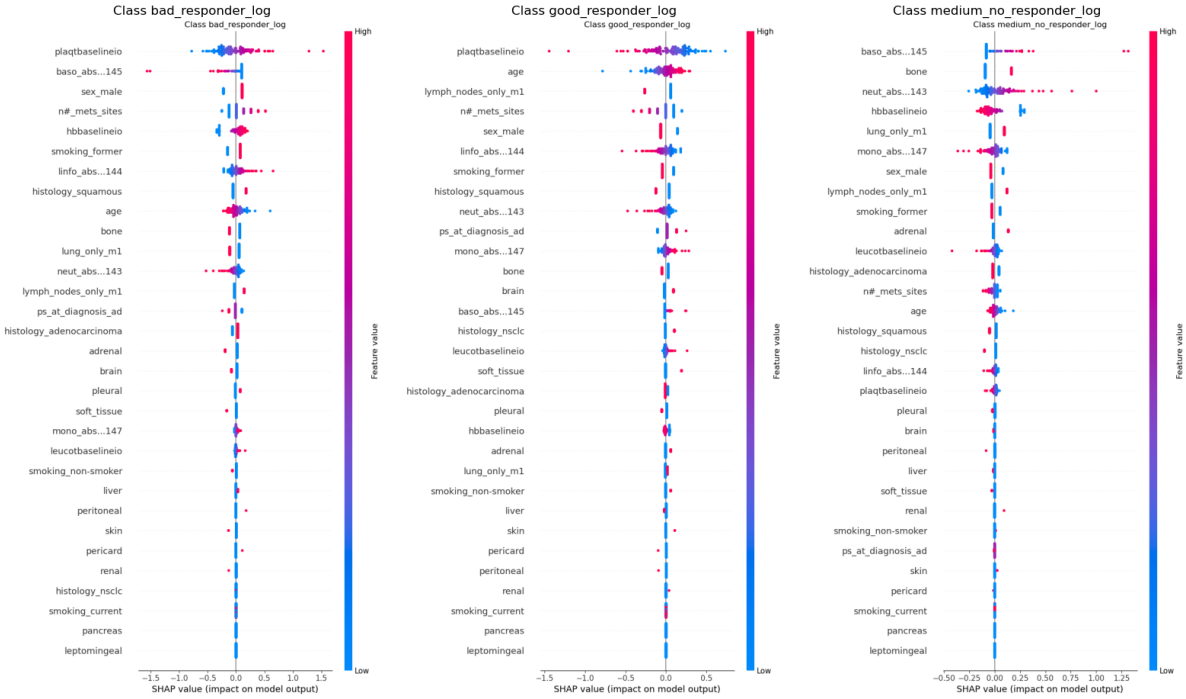


Figure 10: SHAP summary plots for the Responder Groups (from log-transformed TTP) classification task obtained with the best-performing classical ML model.

Across the three response categories derived from the log-transformed progression variable, similar clinical drivers emerge as key determinants of treatment outcome, though with subtle shifts in their relative influence and directionality. In detail:

- **Bad (responder) Class**
  The model again highlights platelet baseline levels and performance status at diagnosis as major contributors to poor outcomes, both showing a strong positive SHAP impact. High platelet counts remain associated with a higher likelihood of belonging to the bad responder group. Additionally, multiple metastatic sites and

the presence of brain or adrenal involvement are consistent positive predictors of poor response. Notably, hematological markers such as neutrophil and lymphocyte counts gain importance, suggesting that systemic inflammation becomes more predictive when progression is expressed on a logarithmic scale.

- **Medium (responder) Class**
  This intermediate group retains a composite pattern, but the SHAP distribution indicates a more balanced influence among hematological variables (basophils, neutrophils, monocytes) and disease extent indicators. The lung-only metastases variable shows a mild negative SHAP contribution, consistent with more favorable disease patterns within this group. The log transformation appears to reduce the dominance of clinical performance and strengthen the relative impact of immune cell metrics, hinting at a subtler interplay between systemic biology and disease stability.

- **Good (responder) Class**
  Favorable outcomes continue to be associated with lower platelet counts, younger age, fewer metastatic sites, and lymph-node–only metastases. The contribution of histological subtype (especially squamous vs. adenocarcinoma) becomes more evident, whereas the effect of performance status appears less pronounced. Overall, the model delineates this group as patients with limited disease burden, lower systemic inflammation, and better biological reserve, in line with clinically coherent expectations.

Overall, after the log-transformation, the model maintains a strong ability to distinguish responder groups through clinically consistent variables, but **hematologic** and **inflammatory parameters** gain relative weight, suggesting that expressing progression in logarithmic form enhances sensitivity to systemic biological signals rather than purely disease-burden indicators.

6 Ethical and Sustainability Considerations

The integration of AI models into the clinical decision-making process inevitably raises ethical and sustainability considerations. In this work, the application of machine learning and deep learning methods to predict patient outcomes and treatment responses has been guided by principles of social, economic, and environmental responsibility.

From a social sustainability perspective, the proposed models aim to support clinicians in providing more accurate and personalized prognostic information to patients. The ability to estimate, with a certain degree of confidence, treatment response or survival duration can help guide therapeutic choices, improve communication with patients, and ultimately enhance the quality of care. Importantly, AI is not intended to replace the clinician's judgment, but rather to augment it, offering interpretable insights into how different clinical and biological variables contribute to outcomes. This interpretability fosters transparency and trust, ensuring that predictive models remain a tool for empowerment rather than automation.

From an economic perspective, optimizing treatment strategies through predictive modeling can significantly reduce costs associated with immunotherapy: given the high price of each treatment dose and the considerable financial burden on healthcare systems, being able to identify patients most likely to benefit from specific therapies can promote a more efficient and equitable allocation of medical resources. Such data-driven stratification aligns with the broader goals of sustainable healthcare, balancing clinical efficacy with economic feasibility.

Finally, regarding environmental sustainability, it must be acknowledged that training complex AI models is computationally intensive and energy demanding. Although this represents a non-negligible environmental cost, the long-term benefits, both social and economic, may justify the trade-off. By improving patient management and reducing overtreatment or ineffective therapeutic cycles, these models may indirectly contribute to lowering the overall environmental footprint of medical practice.

In summary, the adoption of AI in healthcare should be pursued responsibly, considering not only its predictive performance but also its ethical, social, and ecological implications. Striking a balance among these dimensions is essential to ensure that technological progress truly serves patients, clinicians, and society as a whole.

7 Conclusions

This work explored the development and comparison of different modeling strategies, spanning from classical machine learning algorithms to conventional and advanced neural architectures, for predicting multiple clinical endpoints derived from patients treated with immunotherapy. Through a systematic evaluation, it emerged that model performance does not increase monotonically with architectural complexity; rather, it is highly dependent on the intrinsic nature of the endpoint and the data representation adopted.

For regression tasks such as the TTP, simpler models like Lasso regression achieved the best performance, likely due to the relatively limited sample size and the dominance of linear components within the input–output relationships. In contrast, classification tasks involving higher non-linearity, such as the responder quality groups derived through unsupervised clustering, benefited from more expressive neural architectures, particularly self-normalizing networks, which demonstrated better calibration and generalization capabilities even in imbalanced settings. Interestingly, when the TTP variable was log-transformed, the problem structure changed substantially, and simpler models once again outperformed deeper networks. This result emphasizes that data transformations can sometimes reveal latent linear relationships that complex architectures might otherwise obscure or overfit.

Beyond predictive accuracy, model interpretability through SHAP analysis revealed that a consistent set of clinically meaningful variables drives treatment response across different model configurations. Features related to disease extent, baseline functional status, and systemic inflammation (such as platelet and neutrophil counts) emerged as the most influential predictors, while their relative contribution varied depending on the endpoint transformation. Notably, after log-transforming TTP, hematologic and immune-related markers gained greater explanatory power, suggesting that subtle biological effects become more evident when non-linear scaling is applied to clinical outcomes.

Overall, these findings suggest that model complexity should be chosen adaptively with respect to the statistical and biological properties of the endpoint under investigation, rather than being imposed a priori. The results also highlight the importance of data preprocessing and representation, as transformations such as normalization or log-scaling can significantly affect the expressivity required from the predictive model.

From a clinical standpoint, the interpretability achieved through SHAP analysis provides tangible value for oncologists and multidisciplinary care teams. By quantifying the contribution of each variable to individual predictions, these models can assist clinicians in identifying the patient-specific factors most strongly associated with either favorable or

poor responses to immunotherapy. For instance, visualizations of feature impact allow clinicians to discern how inflammatory markers or metastatic patterns influence predicted outcomes, thereby offering an intuitive understanding of the biological mechanisms underlying the model's decision process.

Such interpretability has several practical implications: first, it supports personalized treatment planning, by highlighting which clinical or laboratory factors might suggest a higher likelihood of benefit from immunotherapy versus alternative therapies; second, it provides diagnostic transparency, enabling experts to verify whether the model's reasoning aligns with established clinical knowledge or to uncover non-obvious associations that might warrant further clinical investigation. Finally, at a population level, aggregating SHAP-based insights across patients can help prioritize prognostic biomarkers for future trials or refine inclusion criteria for ongoing studies, thereby bridging data-driven insights with clinical research and decision support.

From a broader perspective, this work demonstrates the feasibility and flexibility of combining classical and modern modeling paradigms within the same analytical framework, paving the way for future research aimed at improving interpretability, robustness, and clinical relevance in the prediction of machine learning-based immunotherapy outcomes. In particular, future developments should consider the integration of molecular and genomic alterations, such as KRAS, EGFR, ALK, or ROS1 mutations, since recent studies have shown that these molecular profiles substantially influence both disease progression and response to immune checkpoint inhibitors. Incorporating such features could enhance model generalizability and provide deeper biological insights into the mechanisms underlying treatment response in lung cancer patients.

References

[1] International Agency for Research on Cancer. Global cancer observatory: World fact sheet. `https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf`, 2022.

[2] Aritraa Lahiri, Avik Maji, {Pravin D.} Potdar, Navneet Singh, Purvish Parikh, Bharti Bisht, Anubhab Mukherjee, and {Manash K.} Paul. Lung cancer immunotherapy: progress, pitfalls, and promises. *Molecular Cancer*, 22(1), December 2023. Publisher Copyright: © 2023, The Author(s).

[3] Keisuke Onoi, Yusuke Chihara, Junji Uchino, Takayuki Shimamoto, Yoshie Morimoto, Masahiro Iwasaku, Yoshiko Kaneko, Tadaaki Yamada, and Koichi Takayama. Immune checkpoint inhibitors for lung cancer treatment: A review. *Journal of Clinical Medicine*, 9(5), 2020.

[4] Mohammed Yousef Shaheen. Applications of artificial intelligence (ai) in healthcare: A review, 09 2021.

[5] Beilei Liu, Hongyu Zhou, Li-Cheng Tan, Kin Siu, and Xin-Yuan Guan. Exploring treatment options in cancer: Tumor treatment strategies. *Signal transduction and targeted therapy*, 9:175, 07 2024.

[6] Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80, 07 2023.

[7] Qing Gao, Luyu Yang, Mingjun Lu, Renjing Jin, Huan Ye, and Teng Ma. The artificial intelligence and machine learning in lung cancer immunotherapy. *Journal of Hematology Oncology*, 16, 05 2023.

[8] World Health Organization. Lung cancer, June 2023.

[9] Amanda Delgado and Achuta Guddati. Clinical endpoints in oncology - a primer. *American journal of cancer research*, 11:1121–1131, 04 2021.

[10] U.S. Food and Drug Administration. Clinical trial endpoints for the approval of cancer drugs and biologics (guidance for industry). `https://www.fda.gov/media/71195/download`, December 2018.

[11] Kathleen Ruchalski, Marta Braschi-Amirfarzan, Michael Douek, Victor Sai, Antonio Gutierrez, Rohit Dewan, and Jonathan Goldin. A primer on recist 1.1 for oncologic imaging in clinical drug trials. *Radiology: Imaging Cancer*, 3(3):e210008, 2021. PMID: 33988475.

[12] Radiopaedia. Response evaluation criteria in solid tumours (recist). https://radiopaedia.org/articles/response-evaluation-criteria-in-solid-tumours.

[13] Erjia Zhu, Amgad Muneer, Jianjun Zhang, Yang Xia, Xiaomeng Li, Caicun Zhou, John Heymach, Jia Wu, and Xiuning Le. Progress and challenges of artificial intelligence in lung cancer clinical translation. *npj Precision Oncology*, (2025) 9:210, 07 2025.

[14] Prem Ramkumar, Kyle Kunze, Heather Haeberle, Jaret Karnuta, Bryan Luu, and Benedict Nwachukwu. Clinical and research medical applications of artificial intelligence: Fundamentals for the orthopaedic surgeon. *Arthroscopy The Journal of Arthroscopic and Related Surgery*, 37, 08 2020.

[15] Elena Fountzilas, Tillman Pearce, Mehmet Baysal, Abhijit Chakraborty, and Apostolia Tsimberidou. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *npj Digital Medicine*, 8, 01 2025.

[16] Hideyuki Shimizu and Keiichi Nakayama. Artificial intelligence in oncology. *Cancer Science*, 111, 03 2020.

[17] Likhitha Kolla and Ravi Parikh. Uses and limitations of artificial intelligence for oncology. *Cancer*, 130, 03 2024.

[18] C. Zhang, J. Xu, R. Tang, Y. Liang, J. Liu, W. Wang, Q. Zhao, X. Li, Y. Chen, and Z. Yang. Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *Journal of Hematology & Oncology*, 16(1):114, 2023.

[19] Fatma Zahra Abdeldjouad, Menaouer Brahami, and Mohammed Sabri. Evaluating the effectiveness of artificial intelligence in predicting adverse drug reactions among cancer patients: A systematic review and meta-analysis. *arXiv preprint arXiv:2404.05762*, 2024.

[20] Joscha Grüger, Tobias Geyer, Tobias Brix, Michael Storck, Sonja Leson, Laura Bley, Carsten Weishaupt, Ralph Bergmann, and Stephan A. Braun. Ai-driven decision support in oncology: Evaluating data readiness for skin cancer treatment. *arXiv preprint arXiv:2503.09164*, 2025.

[21] Anshu Ankolekar, Sebastian Boie, Maryam Abdollahyan, Emanuela Gadaleta, Seyed Alireza Hasheminasab, Guang Yang, Charles Beauville, Nikolaos Dikaios, George Anthony Kastis, Michael Bussmann, Sara Khalid, Hagen Kruger, Philippe Lambin, and Giorgos Papanastasiou. Advancing oncology with federated learning: transcending boundaries in breast, lung, and prostate cancer. a systematic review. *medRxiv*, 2024.

[22] Samita Bai, Sidra Nasir, Rizwan Ahmed Khan, Alexandre Meyer, and Hubert Konik. Breast cancer diagnosis: A comprehensive exploration of explainable artificial intelligence (xai) techniques. *arXiv preprint arXiv:2406.00532*, 2024.

[23] D. S. Char, M. D. Abràmoff, and C. Feudtner. Identifying bias and ensuring fairness in artificial intelligence for oncology. *JCO Clinical Cancer Informatics*, 6:e2200042, 2022.

[24] R. B. Parikh, Z. Obermeyer, and A. S. Navathe. Addressing bias in artificial intelligence in health care. *Nature Medicine*, 29:109–118, 2023.

[25] Eliza Froicu, Ioana Creanga, Vlad-Adrian Afrăsânie, Bogdan Gafton, Teodora Alexa-Stratulat, Lucian Miron, Diana Pușcașu, Vladimir Poroch, Gema Bacoanu, I. Radu, and Mihai-Vasile Marinca. Artificial intelligence and decision-making in oncology: A review of ethical, legal, and informed consent challenges. *Current Oncology Reports*, 27:1002–1012, 06 2025.

[26] European Society for Medical Oncology (ESMO). Ai is transforming cancer diagnosis—but systemic barriers still hold it back. `https://dailyreporter.esmo.org/homepage/digital-oncology/ai-is-transforming-cancer-diagnosis-but-systemic-barriers-still-hold-it-back`.

[27] Laila C. Roisman, Waleed Kian, Alaa Anoze, Vered Fuchs, Maria Spector, Roee Steiner, Levi Kassel, Gilad Rechnitzer, Iris Fried, Nir Peled, and Naama R. Bogot. Radiological artificial intelligence - predicting personalized immunotherapy outcomes in lung cancer. *NPJ Precision Oncology*, 7:99, 2023.

[28] Ting Mei, Ting Wang, and Qinghua Zhou. Multi-omics and artificial intelligence predict clinical outcomes of immunotherapy in non-small cell lung cancer patients. *Clinical and Experimental Medicine*, 24, 03 2024.

[29] Jie Zheng, Shuang Xu, Guoyu Wang, and Yiming Shi. Applications of ct-based radiomics for the prediction of immune checkpoint markers and immunotherapeutic outcomes in non-small cell lung cancer. *Frontiers in Immunology*, 15, 08 2024.

[30] Ian Janzen, Cheryl Ho, Barbara Melosky, Qian Ye, Jessica Li, Gang Wang, Stephen Lam, Calum MacAulay, and Ren Yuan. Machine learning and computed tomography radiomics to predict disease progression to upfront pembrolizumab monotherapy in advanced non-small-cell lung cancer: A pilot study. *Cancers*, 17(1):58, 2024.

[31] Jhimli Mitra, Soumya Ghose, and Rajat Thawani. Clinically explainable prediction of immunotherapy response integrating radiomics and clinico-pathological information in non-small cell lung cancer. *Cancers*, 17(16), 2025.

[32] Maliazurina Saad, Lingzhi Hong, MD Aminu, Natalie Vokes, Pingjun Chen, Morteza Salehjahromi, Kang Qin, Sheeba Sujit, Xuetao Lu, Elliana Young, Qasem Al-Tashi, Rizwan Qureshi, Carol wu, Brett Carter, Steven Lin, Percy Lee, Saumil Gandhi, Joe Chang, Ruijiang Li, and Jia Wu. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by ct-based ensemble deep learning: a retrospective study. *The Lancet Digital Health*, 5, 05 2023.

[33] Siyun Lin, Zhuangxuan Ma, Yuanshan Yao, Hou Huang, Wufei Chen, Dongfang Tang, and Wen Gao. Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model. *Diagnostic and Interventional Radiology*, 31, 01 2025.

[34] Marta Ligero, Bente Gielen, Victor Navarro, Pablo Cresta Morgado, Olivia Prior, Rodrigo Dienstmann, Paolo Nuciforo, Stefano Trebeschi, Regina Beets-Tan, Evis Sala, Elena Garralda, and Raquel Perez-Lopez. A whirl of radiomics-based biomarkers in cancer immunotherapy, why is large scale validation still lacking? *npj Precision Oncology*, 8(1):42, 2024. Open access perspective / review.

[35] Xinyu Yuan, Heli Xu, Junkai Zhu, Zixuan Yang, Boyue Pan, Lin Wu, and Huanhuan Chen. Systematic review and meta-analysis of artificial intelligence for image-based lung cancer classification and prognostic evaluation. *NPJ precision oncology*, 9:300, 08 2025.

[36] Laura Mezquita, Isabel Preeshagul, Edouard Auclin, Diana Saravia, Lizza Hendriks, Hira Rizvi, Wungki Park, Ernest Nadal, Patricia Martin-Romano, Jose C. Ruffinelli, Santiago Ponce, Clarisse Audigier-Valette, Simona Carnio, Felix Blanc-Durand, Paolo Bironzo, Fabrizio Tabbò, Maria Lucia Reale, Silvia Novello, Matthew D. Hellmann, Peter Sawan, Jeffrey Girshman, Andrew J. Plodkowski, Gerard Zalcman, Margarita Majem, Melinda Charrier, Marie Naigeon, Caroline Rossoni, AnnaPaola Mariniello, Luis Paz-Ares, Anne Marie Dingemans, David Planchard, Nathalie Cozic, Lydie Cassard, Gilberto Lopes, Nathalie Chaput, Kathryn Arbour, and Benjamin Besse. Predicting immunotherapy outcomes under therapy in patients with advanced nsclc using dnlr and its early dynamics. *European Journal of Cancer*, 151:211–220, 2021.

[37] Monica Pierro, Capucine Baldini, Edouard Auclin, Hélène Vincent, Andreea Varga, Patricia Martin Romano, Perrine Vuagnat, Benjamin Besse, David Planchard, Antoine Hollebecque, Stéphane Champiat, Aurélien Marabelle, Jean-Marie Michot, Christophe Massard, and Laura Mezquita. Predicting immunotherapy outcomes in older patients with solid tumors using the lipi score. *Cancers*, 14(20), 2022.

[38] Mihaela Aldea, Jose Carlos Benitez, and Laura Mezquita. The lung immune prognostic index (lipi) stratifies prognostic groups in advanced non-small cell lung cancer (nsclc) patients. *Translational Lung Cancer Research*, 9(4), 2020.

[39] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[41] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[42] Guido Montufar, Razvan Pascanu, Kyunghyun Cho, and Y. Bengio. On the number of linear regions of deep neural networks. *NIPS 2014*, 02 2014.

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[44] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017.

[45] Chris Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7:108–116, 01 1995.

[46] Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Lukasz Kaiser, Karol Kurach, Ilya Sutskever, and James Martens. Adding gradient noise improves learning for very deep networks, 2017.

[47] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Proceedings of the 5th International Conference on Neural Information Processing Systems*, NIPS'91, page 950–957, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[48] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and diagnose: clinical time series analysis using attention models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[49] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020.

[50] Sercan Ömer Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *CoRR*, abs/1908.07442, 2019.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[52] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *CoRR*, abs/1704.06904, 2017.

[53] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[54] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 2022.