Università degli Studi di Palermo

AREA RICERCA E TRASFERIMENTO TECNOLOGICO
SETTORE DOTTORATI E CONTRATTI PER LA RICERCA
U. O. DOTTORATI DI RICERCA

D079 - PhD School of Biomedicine, Neuroscience and Advanced Diagnostics
Department of Biomedicine, Neurosciences and Advanced Diagnostics (BiND)
Disciplinary Scientific Sector ING-INF/05

Innovations in Medical Image Analysis and Explainable AI for Transparent Clinical Decision Support Systems

AUTHOR
FRANCESCO PRINZI

Ph.D. COORDINATOR
PROF. FABIO BUCCHIERI

TUTOR
PROF. SALVATORE VITABILE

CYCLE XXXVI
YEAR 2023

“Do, or do not… There is no try!”

Summary

This thesis explores innovative methods designed to assist clinicians in their everyday practice, with a particular emphasis on Medical Image Analysis and Explainability issues.

The main challenge lies in interpreting the knowledge gained from machine learning algorithms, also called black-boxes, to provide transparent clinical decision support systems for real integration into clinical practice. For this reason, all work aims to exploit Explainable AI techniques to study and interpret the trained models. Given the countless open problems for the development of clinical decision support systems, the project includes the analysis of various data and pathologies.

The main works are focused on the most threatening disease afflicting the female population: Breast Cancer. The works aim to diagnose and classify breast cancer through medical images by taking advantage of a first-level examination such as Mammography screening, Ultrasound images, and a more advanced examination such as MRI. Papers on Breast Cancer and Microcalcification Classification demonstrated the potential of shallow learning algorithms in terms of explainability and accuracy when intelligible radiomic features are used. Conversely, the union of deep learning and Explainable AI methods showed impressive results for Breast Cancer Detection. The local explanations provided via saliency maps were critical for model introspection, as well as increasing performance.

To increase trust in these systems and aspire to their real use, a multi-level explanation was proposed. Three main stakeholders who need transparent models have been identified: developers, physicians, and patients. For this reason, guided by the enormous impact of COVID-19 in the world population, a fully Explainable machine learning model was proposed for COVID-19 Prognosis prediction exploiting the proposed multi-level explanation. It is assumed that such a system primarily requires two components: 1) inherently explainable inputs such as clinical, laboratory, and radiomic features; 2) Explainable methods capable of explaining globally and locally the trained model.

The union of these two requirements allows the developer to detect any model bias, the doctor to verify the model findings with clinical evidence, and justify decisions to patients.

These results were also confirmed for the study of coronary artery disease. In particular machine learning algorithms are trained using intelligible clinical and radiomic features extracted from pericoronaric adipose tissue to assess the condition of coronary arteries.

Eventually, some important national and international collaborations led to the analysis of data for the development of predictive models for some neurological disorders. In particular, the predictivity of handwriting features for the prediction of depressed patients was explored. Using the training of neural networks constrained by first-order logic, it was possible to provide high-performance and explainable models, going beyond the trade-off between explainability and accuracy.

Acknowledgements

Expressing gratitude is a profound obligation when I reflect on the extraordinary journey of my three-year doctoral program, enriched by satisfaction, invaluable experiences, and genuine connections. I extend my heartfelt thanks to all the people who played a key role in turning the arduous journey of a doctoral student into an extraordinary one, thanks to their unwavering commitment to continuous dialogue and support. At every stage of my doctoral journey, these outstanding people stood by me, encouraging and supporting me in the most difficult moments. Their unflinching support has been instrumental in my academic success, and for that, I am deeply grateful.

Aware that all these words cannot express the gratitude I feel toward these people, I want to say thank you:

I would like to express my heartfelt gratitude to my supervisor, ***Professor Salvatore Vitabile***, for his unwavering support, guidance, and mentorship throughout the duration of my PhD program. He has been an outstanding mentor and I am delighted to have had the opportunity to work under his guidance. This project would not have been possible without his kind and constant support. Time and time, he has proven to be an inspiring mentor and person to talk to and draw from. I am deeply grateful for his contributions to my academic and personal growth. I also want to express my gratitude for Professor Vitabile's unwavering patience and faith in my abilities. His dedication to my development as a student and researcher has been truly inspiring.

I must express my deepest gratitude to ***Professor Pietro Lio'***, who enriched my personal and academic experience as a doctoral student, throughout my visit at the University of Cambridge. He allowed me to meet and form life lasting relationships with incredible colleagues, researchers and friends. Through his continuous encouragement, he has shown me the beauty of research and the importance of authentic relationships. Professor Lio's consistent dedication to the well-being and success of his students is as an example for my future career as a paramount accomplishment for any supervisor.

I am grateful to ***Professor Salvatore Gaglio*** for his undoubted faith in my abilities from the very beginning and for instilling in me a deep enthusiasm for the field of artificial intelligence. His vast knowledge has continuously fueled my inspiration since my master's years. I clearly remember his willingness to engage in dialogue in an attempt to provide me with guidance for my decisions after my master's degree. Professor Gaglio will forever remain an enduring source of inspiration in my academic journey.

I would like to express my gratitude to my friend ***Eng. Carmelo Militello***, whose consistent engagement and challenging discussions have played a significant role in the achievements of my doctoral journey. He has been a faithful companion in countless jobs, and our shared experiences have facilitated the growth and improvement of my skills, fostered by our mutual exchange of ideas.

I'd also like to extend my appreciation to my friend Eng. ***Marco Insalaco***, who initially introduced me to and guided me through laboratory activities. He instilled in me the practicality and attention to detail that are crucial elements in the successful completion of research projects.

I want to thank my labmates, Tiziana, Alberto, and Giovanni, for their continuous support in the countless lab activities.

Finally, I must thank all the people I met in Cambridge, who helped make my visiting experience one of the most meaningful experiences of my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The development of Clinical Decision Support Systems (CDSS) using machine learning (ML) techniques has proven remarkable progress in the area of healthcare. These systems have emerged as invaluable tools, empowering healthcare professionals to make well-informed decisions, enhance diagnostic accuracy, and deliver personalized treatment recommendations. However, the increasing complexity and adoption of shallow learning (SL) and deep learning (DL) models have raised concerns regarding their "black-box" nature, wherein the decision-making process becomes opaque and challenging to interpret. This lack of transparency has prompted the need for Explainable Artificial Intelligence (XAI) to shed light on the inner workings of these complicated models, ensuring high accuracy and comprehensible behavior to clinicians, patients, and other stakeholders.

The central focus of modern CDSS lies in achieving a balance between model explainability and predictive accuracy. Shallow learning approaches (e.g., logistic regression, decision trees, etc.) offer relatively straightforward interpretations, making them attractive for many applications in healthcare. These models provide insights into the features driving their decisions, allowing clinicians to understand the reasoning behind the system's recommendations. However, they may lack the complexity needed to capture the hidden patterns and insights present in the data, potentially limiting their overall accuracy and the range of medical conditions they can effectively treat.

In contrast, deep learning methods, particularly neural networks (NNs), have demonstrated unprecedented success in numerous domains, including computer vision, natural language processing, and speech recognition. Their ability to learn hierarchical representations from data enables them to extract highly informative features and find

patterns that may not be evident in shallow models. Consequently, deep learning models have shown extremely promising in CDSS for tasks such as medical imaging analysis. Nonetheless, their inherent complexity often leads to the aforementioned "black-box" issue, hindering their explainability. For this reason, concerns have been raised about the potential risks of blindly relying on their decisions in critical medical scenarios. In an attempt to seek a trade-off between explainability and accuracy in CDSS, researchers and developers have explored a range of innovative approaches to enhance model transparency and explainability.

An option falls in the explainable-by-design methods generating significant interest in recent research. These methods design and modify conventional deep learning architectures to inherently promote interpretability. For instance, training of neural networks constrained by logical rules has been proposed, which allows for explanations using the formalism of first-order logic. Explainable-by-design models strike a balance between performance and interpretability, offering a promising pathway to alleviate the black-box challenge in CDSS without compromising predictive accuracy. However, despite these efforts, a significant challenge persists in training CDSS with small datasets. Acquiring large, well-annotated datasets in the medical domain is often hindered by privacy concerns, ethical considerations, and the complexities involved in obtaining annotations for diverse medical conditions. As a result, deep learning models may struggle to generalize effectively, and there is a risk of overfitting when limited data are used for training.

Researchers have also explored the incorporation of radiomic features to perform interpretable handcrafted feature extraction from medical images. Radiomics analysis involves extracting a wide array of biomarkers aiming to characterize the texture, shape, and intensity patterns within an image. These handcrafted features can provide valuable insights into disease characteristics, effectively complementing the deep learning models' predictions. For this reason, radiomic features offer a transparent and interpretable representation of medical images, empowering clinicians to validate and understand the model's decision process through human-interpretable image-derived biomarkers. In addition, radiomics enables the use of shallow learning methods that appear attractive when modest data are available. Nevertheless, researchers must tread carefully, ensuring that the simplicity of shallow models does not sacrifice essential insights or compromise patient care.

The development of CDSS using shallow learning and deep learning techniques holds immense potential to revolutionize healthcare practices and improve patient outcomes. Nevertheless, the dichotomy between explainability and accuracy remains a critical

challenge that requires innovative solutions. While striving to bridge this gap, understanding the trade-offs between shallow learning and deep learning approaches becomes paramount. XAI methods can facilitate the integration of systems into real clinical practice, as they enable the proper debugging of model training. They also empower clinicians to clinically validate the models by comparing their findings with existing clinical evidence. Most importantly, these methods allow patients to understand the reasons behind a specific decision, providing them with transparency and clarity about their medical care. Moreover, addressing the complexities of training CDSS with small datasets is essential to ensure robust and reliable models in the face of data scarcity. By striking a harmonious balance between model transparency and predictive power, it is possible to forge a path towards more trustworthy, effective, and ethically sound CDSS, empowering clinicians and ultimately benefiting patients worldwide.

1.1 Objectives and Open Questions

The objective of my Ph.D. project is to comprehensively investigate the potential of machine learning methods in supporting clinical practice, aiming to unlock their full capabilities while addressing the associated challenges. These cutting-edge techniques have exhibited remarkable performance in various domains, but their integration into the complex and sensitive healthcare environment demands careful consideration of several critical factors.

The availability of well-annotated datasets remains a pivotal concern, and this study explores innovative approaches to achieve successful model training even with limited data.

The imperative for interpretability and explainability in system outputs is paramount, as clinicians and medical professionals must have a clear understanding of the reasoning behind the model's decisions to foster trust and acceptance in real-world applications. To address the challenge of interpretability explicitly, this research will explore the potential of radiomic feature extraction as an alternative to deep feature extraction. By comparing deep learning methods with radiomic feature extraction techniques, this thesis aims to provide valuable insights into the applicability of interpretable shallow learning techniques in medical imaging and diagnostics.

What is the trade-off to be made between explainability and accuracy? This balancing is essential for the development of Explainable Clinical Decision Support Systems (X-CDSS).

By exploring these multifaceted aspects, this thesis focuses on enhancing the efficacy of machine learning models in clinical practice and contributing to the ongoing discourse

on responsible AI adoption in the medical domain.

1.2 My Contributions

Objectives and Research Questions have been analyzed and addressed in numerous clinical settings, leading to the development of several publications in the following domains:

Breast Cancer classification in DCE-MRI: The work concerns the classification of breast cancer in DCE-MRI using radiomic features. The first preliminary results were presented at the *"International Conference on Applied Intelligence and Informatics"*, and the article received the best paper award. Subsequently, the work was expanded and deepened and is currently under review in an international journal.

- **Prinzi, F.**, Orlando, A., Gaglio, S., & Vitabile, S. Breast Cancer Classification through Multivariate Radiomic Time Series Analysis in DCE-MRI Sequences. *Expert Systems with Applications, Under Review*
- **Prinzi, F.**, Orlando, A., Gaglio, S., Midiri, M., & Vitabile, S. (2022, September). ML-Based Radiomics Analysis for Breast Cancer Classification in DCE-MRI. In *International Conference on Applied Intelligence and Informatics* (pp. 144-158). Cham: Springer Nature Switzerland. **Best Paper Award**. https://doi.org/10.1007/978-3-031-24801-6_11

Breast Cancer Detection in Mammograms: The works deal with the early detection of breast cancer in mammograms using Yolo. The first preliminary results were presented at the *"Italian Workshop on Neural Networks"*. Subsequently, the article was expanded and published in the Cognitive Computation journal.

- **Prinzi, F.**, Insalaco, M., Orlando, A., Gaglio, S., & Vitabile, S. (2023). A Yolo-Based Model for Breast Cancer Detection in Mammograms. *Cognitive Computation*, 1-14. https://doi.org/10.1007/s12559-023-10189-6
- **Prinzi, F.**, Insalaco, M., Gaglio, S., & Vitabile, S. (2023). Breast cancer localization and classification in mammograms using YoloV5. In *Applications of Artificial Intelligence and Neural Systems to Data Science* (pp. 73-82). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3592-5_7

Breast Cancer Classification in Ultrasounds: The works deal with the classification in Ultrasound images using Radiomics. The first preliminary results were presented at the *"18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics"*. Subsequently, the article was expanded and submitted to La Radiologia

Medica journal.

- Bartolotta, T.V., Militello, C., **Prinzi, F.**, Ferraro, F., Rundo, L., Zarcaro, C., Di Marco, C., Olrando, A., Matranfa, D., & Vitabile, S. (2023). Artificial intelligence-based, semi-automated segmentation for the extraction of ultrasound-derived radiomics features in breast cancer: a prospective multicenter study. *La Radiologia Medica, Under Review*

- **Prinzi, F.**, Militello, M., Bartolotta, T.V. & Vitabile, S. (2023). Breast Cancer Malignancy Prediction by means of an Explainable Model based on a Multimodal Signature. Proceedings of *"18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics"*, *In press.*

Breast Microcalcification Classification in Mammogram: Radiomic features were used to develop an interpretable signature for the detection and classification of breast microcalcification. The paper is currently under review by the Journal of Digital Imaging

- **Prinzi, F.**, Orlando, A., Gaglio, S., & Vitabile, S. Interpretable Radiomic Signature for Breast Microcalcification Detection and Classification. *Journal of Digital Imaging, Under Review*

COVID-19 Prognosis Prediction: the work is derived from participation in an international competition (Hackathon https://ai4covid-hackathon.ing.unimore.it/) to predict the prognosis of COVID-19 patients. The proposed solution was awarded among the three best solutions in terms of explainability and led to the development of an under-review work. A detailed analysis of the extraction of wavelet-derived radiomic features has already been published.

- **Prinzi, F.**, Militello, M., Scichilone, N., Gaglio, S. & Vitabile, S. Explainable Machine-Learning Models for COVID-19 Prognosis Prediction using Clinical, Laboratory and Radiomic Features. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3327808

- **Prinzi, F.**, Militello, C., Conti, V., & Vitabile, S. (2023). Impact of Wavelet Kernels on Predictive Capability of Radiomic Features: A Case Study on COVID-19 Chest X-ray Images. Journal of Imaging, 9(2), 32. *Journal of Imaging*, 23(12), 5677. https://doi.org/10.3390/jimaging9020032

Explainable Depression Detection: this collaboration formed between *University of Palermo*, *University of Campania Luigi Vanvitelli* and *University of Cambridge*, presented preliminary results at the *"Italian Workshop on Neural Networks"* proposing

an explainable-by-design model for predicting depressed patients through writing and drawing tasks.

- **Prinzi, F.**, Barbiero, P., Cordasco, C., Pietro, L., Vitabile, S, & Esposito, A. Explainable Depression Detection Using Handwriting Features. *"The Italian Workshop on Neural Networks"*, *In press.*

Coronary Artery Disease prediction: the work deal with the use of CT radiomic features and clinical biomarkers for predicting coronary artery disease:

- Militello, C., **Prinzi, F.**, Sollami, G., Rundo, L., La Grutta, L., & Vitabile, S. (2023). CT radiomic features and clinical biomarkers for predicting coronary artery disease. *Cognitive Computation*, 15(1), 238-253. https://doi.org/10.1007/s12559-023-10118-7

Diffuser-based Data Augmentation and Vision Transform: The works deal with the training of vision Tranfrormer in small dataset scenario. The problem of melanoma classification was presented at *"18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics"*[8] and published in Sensors[9]. The same strategies were applied in Mammogram for breast cancer Classification, and presented at the *"Italian Workshop on Neural Networks"*

- Cannata, S., Cicceri, G., Cirrincione, G., Currieri, T., Lovino, M., Militello, C., **Prinzi, F.** & Vitabile, S. (2023). Diffuser Data Augmentation for ViT-based Classification of Dermatoscopic Melanoma Images. Proceeding of *"18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics"*, *In press.*

- Cirrincione, G., Cannata, S., Cicceri, G., **Prinzi, F.**, Currieri, T., Lovino, M., Militello, C., Pasero, E. & Vitabile, S. (2023). Transformer-Based Approach to Melanoma Detection. *Sensors*, 23(12), 5677. https://doi.org/10.3390/s23125677

- Cannata, S., Cicceri, G., Cirrincione, G., Currieri, T., Lovino, M., Militello, M., **Prinzi, F.**, Pasero, E. & Vitabile, S. (2023). ViT-based Classification of Mammogram Images: Impact of Data Augmentation Techniques. *"The Italian Workshop on Neural Networks"*, *In press.*

Shallow and Deep Learning Classifiers in Medical Image Analysis: This paper encapsulated the main findings from the study of the state-of-the-art application of machine learning methods for medical image analysis. This study has been collected and is currently under review at an International Journal.

- **Prinzi, F.**, Currieri, T., Gaglio, T., & Vitabile, S. Shallow and Deep Learning Classifiers in Medical Image Analysis. *European Radiology Experimental, Under Review.*

1.3 Structure of the Thesis

The preliminary phase of my research involved a thorough examination and analysis of the current state of machine learning and Explainable AI methods. Chapter 2 elucidates these concepts to equip the reader with the essential knowledge needed to grasp the subsequent chapters. Subsequent research activities have led to the implementation of machine learning approaches applied to distinct clinical domains. Consequently, in Chapter 3 all papers related to the analysis of breast cancer were discussed, considering Radiomics and deep learning approaches. Then the works related to COVID-19 prognosis prediction were exposed in Chapter 4, emphasizing the Explainability issues. In Chapter 5 the radiomic workflow is applied for Coronary Artery Disease prediction, showing the importance of an interpretable feature extraction to clinically validate the results. Chapter 6 exposes a new explainable-by-design model to overcome the concept of trade-off between model explainability and accuracy, applied to the prediction of depressed patients. Finally, Chapter 7 Conclusion explains and summarizes the main findings of the thesis.

Chapter 2

Background

Medical Image Analysis represents a large slice of Artificial Intelligence applications in medicine. It involves the interpretation and evaluation of medical images to aid several clinical tasks, including diagnosis, prognosis, and treatment planning. Medical imaging technologies, such as MRI, CT scans, and X-rays, generate complex visual data, and extracting meaningful information becomes pivotal. Introducing CDSS into this arena has been transformative. These systems use advanced machine learning algorithms and computer vision techniques to assist clinicians in making more informed and accurate decisions. Through these methods, the objective is to reduce diagnostic errors and support the clinical process.

The first step to analyzing medical images through machine learning methods is to find a salient representation (e.g., biomarkers, features, embeddings) of the regions of interest. this step is called *feature extraction* and is one of the crucial steps in medical image analysis. A great amount of information is derived from the images, which goes beyond what is visually perceptible. Radiomic feature extraction provides an efficient method for quantifying tumor heterogeneity and potentially improving disease prognosis, classification, and in general to support the physician diagnostics process. Contrarily, deep feature extraction employs machine learning algorithms (e.g., using deep neural networks) to identify patterns within high-dimensional data, thereby enhancing the accuracy of diagnostic systems. However, one of the main limitations of neural networks lies in their opacity, which often makes them "black boxes" and hides the correlation between input data and the results obtained.

To resolve the opacity of machine learning models, XAI has come into the picture. XAI algorithms elucidate the internal workings of complex AI models, fostering user trust by providing intelligible and justified decision-making. From this perspective, Radiomics

could have a significant advantage. Since radiomic feature extraction is grounded in established imaging and medical principles, can be interpreted and compared with the medical literature, thereby offering more transparency and interpretability. Moreover, Radiomics enables the use of shallow learning algorithms. Despite its relatively simplistic nature, shallow learning presents a strong alternative to deep learning methods in certain situations, striking a balance between computational demand and performance. While deep learning excels in analyzing large-scale, high-dimensional data, shallow learning can efficiently process smaller datasets, retaining considerable accuracy without the requirement of intensive resources.

Altogether, the confluence of Radiomics, deep learning, shallow learning, and XAI forms a comprehensive landscape, each contributing unique strengths and complementing each other to obtain accurate and explainable medical image analysis.

2.1 Clinical Decision Support Systems

In recent years we have seen a significant surge in the use of computer-assisted tools employing AI methodologies. These innovative tools leverage the capabilities of machine learning frameworks in a wide range of applications, from gaming and commercial or financial pattern analysis to a plethora of Decision Support Systems (DSS). Specifically, within this broad spectrum, CDSSs can strengthen critical healthcare processes where informed decision-making and system reliability are crucial.

A conventional CDSS consists of software specifically crafted to serve as a direct assistance tool for clinical decision-making. Within this framework, the attributes of a patient are systematically compared with a computerized clinical knowledge repository, subsequently yielding patient-specific evaluations or suggestions. Subsequently, the provided computer-based knowledge is made available to the healthcare practitioner for decision-making purposes [1]. Contemporary CDSSs are predominantly deployed at the point-of-care, allowing clinicians to integrate their expertise with the information or recommendations supplied by the CDSS [2].

Focusing on the main field of my research activity, in radiology, CDSS are tools designed to enhance diagnostic accuracy and streamline patient care in diagnostic examinations. These systems employ advanced algorithms and machine learning techniques to assist radiologists in their interpretation of medical images, such as X-rays, CT scans, and MRIs. Images now play a progressively significant role in the context of medical data, as they contain crucial information for tasks such as disease classification, diagnosis, prognosis, etc. Nevertheless, various imaging modalities come with inherent drawbacks and potential challenges. For instance, interpreting mammography findings

can pose substantial difficulties when confronted with cases of elevated breast tissue density. Similarly, ultrasound scans, which rely on ultrasound signals, often yield image streams characterized by noise, thereby complicating the analytical process. Hence, CDSSs have been integrated into radiology practice through the application of machine learning techniques. These systems serve the purpose of offering an impartial perspective that complements the clinical judgment of physicians. Consequently, CDSS finds utility in several applications, including lesion segmentation, automated classification, and detection, among others. In fact, several examples of CDSSs based on deep learning were proposed by IBM Watson Health, DeepMind, Google for a broad spectrum of applications [2]. Additionally, Radiomics has been identified as tool for enhanced imaging and precision radiology [3, 4].

Although research is moving toward integrating CDSSs into real clinical practice, their integration requires addressing some challenges. Training high-performance deep learning architectures is a major issue when small datasets are available, that is a very common scenario in medical images. In addition, machine learning models are typically black-boxes, that is, they are able to learn highly informative and predictive patterns and features, but it is not possible for a human to interpret these findings [5]. This greatly complicates the integration of CDSSs, because an explanation of model decisions and subsequent clinical validation are necessary steps for human-machine trust and overcoming skepticism toward new technologies.

For these reasons, the next subsection discusses what is XAI, to address the explainability issue of machine learning models and overcome the ethical and legal concerns advanced by the regulatory bodies. In addition, several methods are discussed to obtain both high-performance and explainable models in medical imaging, distinguishing the advantages and disadvantages of using deep learning and shallow learning for feature extraction and model development.

2.2 Explainable AI

Despite the vital role of data-driven AI in CDSS, its deployment in the medical field presents several challenges. The advent of novel AI techniques and burgeoning data accessibility have given rise to high-performance yet opaque CDSSs. This obscurity has garnered the attention of regulatory bodies [6, 7]. For instance, the US Federal Trade Commission emphasizes the need for AI applications to exhibit transparency, explainability, fairness, and empirical soundness while promoting accountability [8]. Similarly, the European Parliament's General Data Protection Regulation (GDPR) mandates the provision of comprehensible explanations when automated decision-making occurs [9]. This opacity often elicits skepticism from healthcare professionals and patients alike,

potentially undermining the physician-patient relationship and trust [10].

Consequently, the academic community is actively pursuing strategies to enhance the transparency and explainability of AI systems [11]. Within the healthcare milieu, several factors such as the availability of large datasets [12], the presence of imbalanced or inaccurate datasets in high-dimensional spaces, and various other issues [13] can influence the reliability of AI models. Thus, promoting transparency and explainability can play a significant role in model validation, knowledge domain enhancement, and the actual use of these systems.

It is becoming increasingly apparent that it's insufficient to regard AI-based tools as mere 'black boxes'. Despite certain reservations [14, 15, 16, 17], explainability is fast becoming a mandatory criterion for these systems [18]. Especially within healthcare, understanding the comprehensive effect of each feature (global explanation) and providing an explanation of the decision-making process for each patient (local explanation) can foster trust in data-driven models and facilitate their incorporation into clinical practice.

This progression naturally aligns with the burgeoning field of XAI [19], which is garnering considerable interest [13] and playing a pivotal role in developing and deploying eXplainable Clinical Decision Support Systems (X-CDSSs) that can be efficiently and consciously employed in clinical practice.

2.2.1 The Imperative for Explainability

The necessity for explainability is driven by the extensive advantages it provides [9]. By deploying explainable methods, a broader understanding of the entire inference process, from raw data to valuable insights, can be achieved. Particularly within health informatics, it is crucial to provide results that are not just correct and valid, but also interpretable. Among the main advantages of using these methods, the following can be found:

User Acceptance and Control

User satisfaction and the likelihood of accepting algorithmic decisions are elevated when explanations are provided [20]. Facilitating user acceptance and control is particularly crucial in sectors like healthcare, defense, finance, and law, where understanding decisions and fostering trust in algorithms are paramount [21]. For instance, in a CDSS, when information about a patient's condition is delivered, the patient naturally seeks an explanation [22]. As AI-based systems become increasingly prevalent, personalized services can be ensured by learning users' preferences from their actions [23].

Insightfulness

Advancements in AI have resulted in tools capable of automating tasks traditionally performed by humans. However, this is not the only application of AI. In certain scenarios, machine learning is employed to extract insights from large datasets [23]. In the field of medicine, such techniques can enhance human knowledge in contexts beyond the reach of human capabilities alone [24].

Ethical and Legal Aspects

GDPR asserts every individual's right to comprehend the logic involved when machine intelligence supports human decision-making [25]. Similarly, the US Federal Trade Commission emphasizes the use of transparent, explainable, and fair AI [8]. Legal liability assessment is another burgeoning interest area. An AI-based model may introduce new liability areas by causing unanticipated and undesirable situations (e.g., an accident caused by a self-driving car, classifications based on skin color, etc.). Thoroughly understanding the model's decisions is essential to evaluate liability in such scenarios [23].

Explanatory Debugging

Employing explanation methods facilitates the validation and enhancement of trained models by studying errors and detecting anomalies. This paves the way for interactive machine learning tools [26] for explanatory debugging [27]. Such tools analyze system outcomes to augment training examples, rectify incorrect labels, and designate new input features. Additionally, a human operator can audit rules in an explainable AI system to study the system's generalization capability on unknown real-world data. This is especially significant for AI-based tools developed for medical applications, as model performance may deteriorate with data from different sources (e.g., different medical image acquisition protocols/systems, hospitals, etc.). Explainable AI could enhance algorithm robustness and boost clinicians' confidence in CDSS [28].

2.2.2 Distinguishing Between Transparent and Explainable Models

This section elucidates the distinctions between intrinsically transparent algorithms and algorithms that necessitate specific 'explainers' to make their functionality comprehensible. The model design phase involves deciding between implementing transparent models or black-box models that require XAI methods for the explanation. Models interpretability primarily hinges on two characteristics: the input features and the machine learning algorithms employed.

2.2.2.1 Intrinsically Transparent Models

An intrinsically transparent model has a transparent structure that lends itself to an understandable decision process, thereby eliminating the need for additional explanation methods. Such fully interpretable models are achievable by incorporating comprehensible features and inherently interpretable algorithms. A feature is deemed comprehensible if it can be associated with a concept that is readily understood by humans (intelligible).

Examples of such intelligible features include clinical, radiomic, genomics, laboratory, and many other attributes. Conversely, deep features extracted *via* neural network are less readily understood. From an algorithmic perspective, Logistic Regression, Decision Trees, and Naive Bayes are classified as transparent. The convergence of these two aspects — comprehensible features and transparent algorithms — claims an intrinsically interpretable system for two reasons: the impact of features on the model's decision-making can be quantified, and the model's conclusions can be validated against the existing clinical literature.

However, designing a transparent model introduces a layer of complexity [29]. Typically, interpretable features are less informative than deep features, which are extracted through deep neural network's abstraction mechanism. Furthermore, transparent algorithms may falter when dealing with complex and nonlinear data relationships. Consequently, in certain contexts, models designed to be transparent may exhibit suboptimal performance. Despite this, for certain applications, the use of transparent models may suffice [30].

2.2.2.2 Black-Box Models

Recognized black-box algorithms are unparalleled in finding hidden patterns in the data. For this reason, explanation methods have to be employed to study the decision process and the learned features [31, 32]. In the case one of the two characteristics (interpretable input features and transparent by-design algorithms) is missing, explainable methods have to be exploited. For example, Artificial Neural Networks, Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Tree Ensemble (TE), are defined as black-boxes and require explanation methods for their introspection. When black-box algorithms are used and the features are intelligible, explanation methods allow estimating the most important features for prediction. Typically, these methods are called *post-hoc* algorithms because applied after model training. *Post-hoc* algorithms can be applied to calculate the contribution of tabular features and images (discussed in the next section). Considering the great generalization capabilities of these models, it is nevertheless worth making an effort to add — in the classic development and

implementation pipeline — a step for model explanation, to increase human-machine confidence. Model explainability becomes much more complicated when unintelligible features are used and becomes difficult to validate the results, also with the physician's support. Using unintelligible features makes it hard to compare the findings with the medical literature because the meaning of the features is difficult to define or even unknown (as happens with the learned features in deep architectures). Methods for calculating saliency maps, discussed in the following sections, allow the explanation of the features extracted *via* neural networks for image analysis. However, their explanation is only local and makes global medical validation difficult. In addition, the saliency maps represent a qualitative explanation tool, which may still be subject to inter-operator variability.

2.2.3 Achieving Explainability

As previously highlighted, models that are inherently opaque necessitate *post-hoc* processing to ensure their explainability and clinical validation. The integration of an explanation layer within these "black-box" models promotes both high performance and introspection. Consequently, growing interest and expansion have been observed in AI-based tools and XAI. The development of frameworks aiding the scientific community in interpreting models at both global and local levels, for understanding models trained on text, tabular data, and imaging analysis, further justifies this trend.

2.2.3.1 Explaining Models for Image Analysis

In the medicine scenario, where the notion of big data is often absent due to the challenges in data collection and annotation, deciphering trained models becomes critical. Convolutional-based architectures have emerged as a prominent choice for image analysis due to their capacity to automatically learn hierarchical representations from data. However, a noteworthy drawback of CNNs is their insatiable appetite for vast amounts of data during the training process. Such hunger for data can pose a significant challenge in practical scenarios where obtaining large, high-quality datasets may be laborious or resource-intensive. Deep architectures trained on very small datasets are highly susceptible to bias, prone to overfitting, and require an examination of the learned features to validate the trained model. However, with the introduction of numerous medical benchmarks, novel architectures, and the application of Transfer Learning techniques [33], the use of deep architectures has been invigorated, making them a common choice for medical image analysis.

Explainability for deep architectures in medical image analysis is often centered around the notion of saliency maps. Saliency maps aim to highlight image regions (or pixels)

that hold significant informational value for prediction [34]. Several techniques were proposed for saliency map computation. The study [35] employs the computation of class score gradient related to the input image to visualize the class concept and compute a specific activation map for a given image. A more robust explanation *via* activation maps is proposed in [36] through integrated gradients, which adhere to the principles/axioms of 'sensitivity' and 'implementation invariance'. A network using deconvolution to visualize the convolutional network is proposed in [37] and [38], with a variant of the deconvolution approach with guided back-propagation proposed in [39]. Very popular algorithms for saliency maps computation belongs to the 'class activation maps' (CAM) category, introduced in [40]. To overcome the CAM limitations of visualizing only the last layer and specific CNN architectures, GradCAM is introduced [41], followed by GradCAM++ [42] for enhanced visual explanations and multi-object scenarios. However, several methods are constantly proposed for saliency map computation [43].

Regrettably, some empirical studies have shown underwhelming results within the clinical field. For instance, a few cases have reported imprecisely localized and spatially blurred visualization [44, 45]. Moreover, it has been evidenced that the saliency map remains unchanged even when adversarial attacks result in erroneous model predictions [46]. Besides, saliency maps only offer local explanations; in fact, the methods mentioned above generate an explanation for a specific instance (e.g., a patient).

An alternative method to elucidate models for image analysis involves extracting deterministic quantitative features with clearly defined significance. The aim is to ascertain the importance of these comprehensible features and compare the model findings with existing clinical literature. Radiomics, which will be elaborated upon in subsequent chapters, serves as an exemplary tool for such applications.

2.2.3.2 Explaining Models for Tabular Data

In the context of clinical practice, imaging data constitutes only a small fraction of the information that can influence the healthcare process, both in terms of volume and type. In fact, models can be trained using a variety of features including clinical, radiomic, genomic, phenotypic, and so on. When dealing with tabular data, algorithms such as Random Forest, XGBoost, and Support Vector Machine often become the primary choice due to their advantage over deep architectures in not requiring vast amounts of data for training. However, since most of these algorithms are inherently black-box, a plethora of methods for their introspection have been proposed [47].

LIME [48] stands as one of the most recognized algorithms employed by XAI researchers for local explanations, while DLIME [49] seeks to address LIME's limitations. The

SHapley Additive exPlanations (SHAP) method [50], leveraging Shapley values to calculate feature importance, enabling both global and local explanations, is perhaps the most used XAI algorithm within the scientific community. The Anchors method [51] constructs and employs 'if-then' rules to denote local conditions sufficient for prediction. Permutation Importance [52] is employed to study feature importance by assigning p-values based on their permuted significance. In [53], a comprehensive description of the main methods of XAI is provided.

While some of these XAI algorithms (e.g., SHAP and LIME) can also be employed to explain models for medical image analysis, the methods discussed in the previous subsection (e.g., GradCAM) were found to be more common for saliency map computation.

2.2.4 Identifying the Stakeholders for an Explainable CDSS

Developing an explainable CDSS demands a multi-tiered explanation strategy, catering to the perspectives of the *developer*, the *clinician*, and the *patient*. Figure 2.1 illustrates how these three crucial perspectives integrate into the process and decision-making flow of a CDSS, thereby leading to an X-CDSS. To the conventional CDSS implementation, an additional step of explainability has to be added to provide a fully X-CDSS.

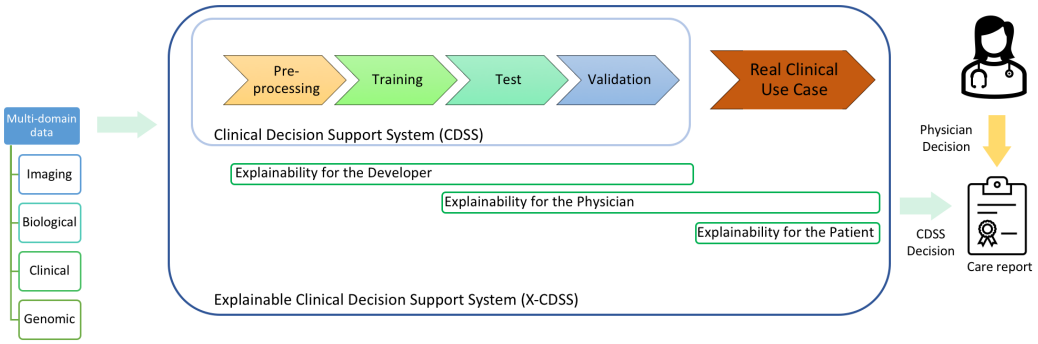


Figure 2.1: General workflow of an Explainable Clinical Decision Support System.

2.2.4.1 The Developer's Perspective

The developer must validate the model after the conventional training pipeline. For instance, in medical imaging, it is critical to ensure that the functionalities learned remain constant even when the training distribution undergoes minor changes (distributional drift) [23]. Typically, a model trained on images acquired by a specific machine/hospital with its own protocol/settings will likely not perform well on images taken with a different setup [54]. Therefore, it's anticipated that feature importance should remain unaltered with the variation of the hospital. To verify this, a model can

be trained by dividing data by the hospital (using a 'leave one center out' modality), then evaluating the importance of the features on a global scale for each model, and ultimately considering only the most important common features obtained from each source (i.e., hospitals). This approach aids in discarding features that are less significant and robust. In addition, feature harmonization can impact heavily on model training, avoiding the distributional drift phenomenon.

2.2.4.2 The Clinician's Perspective

The significance of XAI for clinicians cannot be overstated, as it offers a transformative approach to harnessing the power of AI in the medical field while ensuring transparency and trust. XAI provides clinicians with the unique capability to obtain both global and local explanations for AI-driven decisions, enabling them to delve into the underlying reasoning of the models. By gaining insights into how the AI arrived at specific conclusions, clinicians can verify and confirm the clinical evidence with greater confidence, aligning the findings with established medical literature and best practices. This empowers clinicians to make well-informed decisions, validate AI-driven diagnoses, and understand the reasoning behind the models, ultimately enhancing patient care and safety. Moreover, the ability to compare AI-generated results with existing medical knowledge not only enhances medical research and practice but also fosters a seamless integration of AI into clinical workflows, creating a mutually beneficial partnership between AI technology and healthcare professionals.

2.2.4.3 The Patient's Perspective

Lastly, it is vital to offer an explanation to the patient. A local approach yields a model explanation for each distinct case, mirroring how a clinician justifies a decision to a patient (e.g., a therapy choice, a diagnosis, etc.). To ensure the clinical validation and justification of the global and local explanation, collaboration with a medical expert in the domain is crucial. It is the domain expert—in this case, the physician—who is capable of discerning whether a local explanation is rational and can thus consider it valid. Additionally, with the implementation of the GDPR, it becomes a legal prerequisite for using CDSS in actual clinical practice, as medical data are personal and sensitive: any system using it for automated decision-making support must be capable of explaining its decision-making process (GDPR - Art.15), and a person has the right to request human intervention to check/review the decision of the AI-based CDSS (GDPR - Art.22). The regulations remain somewhat undefined, and it is uncertain whether this type of global and local explanation meets the legal requirement. However, these rules certainly signify a stride toward integrating CDSS into actual clinical practice.

2.2.5 Interpretability and Explainability

Interpretability and explainability are two closely related concepts in the field of machine learning. The concept of interpretability was widely used, but no formal definition had been proposed [55]. In fact, the two terms are often used interchangeably in the literature [53], although some differences have been established. In [56], interpretability is defined as the science of comprehending what a model did (or might have done). The same authors state that interpretability alone is insufficient and explainable models are interpretable by default, but the reverse is not always true. More recently, was stated that interpretability and explainability have escaped a clear universal definition [57]. In [58] was said that interpretability is mostly connected with the intuition behind the outputs of a model, while explainability is associated with the internal logic and mechanics that are inside a machine learning system. Sometimes 'intelligibility' is used as a synonym for interpretability [23]. In addition, explainability emphasizes presenting insights and justifications for a model's output in a manner that is accessible and meaningful to stakeholders, such as end-users or regulatory bodies. In practice, these concepts often overlap, as methods designed to enhance interpretability often contribute to improved explainability and *vice versa*, fostering a synergy that advances the trustworthy deployment of machine learning systems.

2.3 Biomarkers Extraction in Medical Imaging

2.3.1 Deep Feature Extraction

Deep Neural Networks, particularly Convolutional Neural Networks, have revolutionized the process of feature extraction from medical images. The architectural composition of a CNN, which includes convolutional, pooling, and fully connected layers, enables an automatic and hierarchical representation of complex data patterns. The abstraction mechanism that is engaged for this purpose is crucial and operates in a tiered fashion. In the initial stages of this network, low-level features such as edges or colors are identified. Deeper into the network, these elementary features are amalgamated and processed to discern more complex, higher-level features, such as shapes or specific objects. This layered abstraction of patterns, often referred to as 'features of features', means each layer uses the outputs of the preceding layer as its inputs. Convolution operations are crucial in this process. They enable the network to extract spatial features from images while preserving their hierarchical nature. Convolution works by moving a filter (or kernel) over the input data and computing the dot product at each position. The result is a feature map, highlighting the locations of a particular feature in the image. Convolutional layers capture local patterns and spatial dependencies

in the data, making them critical for tasks such as image and video processing. Furthermore, the spatial invariance property of CNNs allows them to recognize patterns irrespective of their location in the image [59]. Combined with pooling operations that reduce dimensionality while maintaining the most salient features, CNNs can efficiently handle large inputs and extract meaningful features. The unique architectural elements of CNNs, such as convolution operations and their mechanism of abstraction, pave the way for effective feature extraction from complex datasets. The extracted features can then be used to boost the performance of machine learning models in a plethora of applications, making CNNs an invaluable tool in the domain of artificial intelligence.

Self-attention-based architectural paradigms, with Transformers being a notable example [60], have garnered preeminence within the field of natural language processing (NLP). Conversely, in the domain of computer vision, convolutional architectures continue to maintain their ascendancy [61]. Motivated by the accomplishments observed in NLP, various attempts have been made to amalgamate CNN-based architectures with self-attention mechanisms. The one most widely used appears as the Vision Transformer (ViT) [62]. ViT has emerged as a prominent model for image classification and feature extraction tasks in the field of computer vision. Unlike CNNs that process an image locally using convolutions, ViTs treat an image as a sequence of patches and process it globally, similar to how transformers handle text sequences in natural language processing tasks. In image classification tasks, ViTs have demonstrated promising performance by capitalizing on the transformer's self-attention mechanism, which allows for capturing long-range dependencies between patches in an image. This enables them to learn more global and complex patterns, which can be beneficial for tasks requiring a broader understanding of the image context. Features extracted *via* ViT can be used in various downstream tasks, enhancing the performance of machine learning models. However, ViTs also have certain disadvantages compared to CNNs. One of the major limitations is their computational demand. Training Transformers require significantly more data and computational resources compared to their convolutional counterparts. Additionally, while Transformers can capture long-range dependencies, they may overlook local spatial hierarchies and correlations that are effectively captured by convolutional layers in CNNs. For these reasons, while ViTs have demonstrated remarkable success in image classification and feature extraction tasks, their usage requires careful consideration of the trade-offs between their advantages and the costs associated with training and implementation. In fact, the authors [62] recognize that Transformers do not possess certain innate characteristics, such as translation equivariance and locality, commonly found in CNNs. As a result, when trained on limited datasets, Transformers may not yield robust generalizations. This observation is beginning to be echoed in other research studies [63].

2.3.2 Radiomic Feature Extraction

2.3.2.1 Introduction to Radiomics

Radiomics is an innovative, multidisciplinary method aiming to convert images or part of them (regions of interest, ROIs) into highly informative biomarkers (or features) [64, 65]. The extracted radiomic features can offer a quantitative perspective to complement the qualitative assessment performed by a radiologist, showing significant potential to enhance the diagnostic process. More specifically, the development of predictive models enables the correlation of radiomic biomarkers with clinical outcomes, improving diagnostic and prognostic accuracy. Radiomic feature extraction is a detailed and carefully coordinated process that involves a multitude of steps. The importance of developing reproducible and explainable studies can be addressed using the radiomic workflow. In fact, it is crucial to define a clear and well-structured processing pipeline, where every step - including image acquisition, segmentation, feature definition and extraction, feature selection, and model setup - must be carefully considered to ensure the repeatability of the radiomic analysis [66, 67].

Radiomic features are calculated using mathematical formulas applied to the ROI shape and the gray level histograms or texture-defining matrices, earning them the name *hand-crafted* features. The ROI typically corresponds to the anatomical structure or pathology under study. This mask, which highlights the ROI, serves as the principal guide during the extraction process. Radiomic feature extraction within a specific ROI holds the promise of focusing and extracting valuable information exclusively from the targeted area. This approach has the potential to unlock crucial insights for medical analysis. However, one critical challenge lies in ensuring the accurate delineation of the ROI. The process of segmenting the ROIs is highly operator-dependent, leading to significant inter-variability among different operators. These discrepancies in ROI delineation can result in inconsistent and unreliable radiomic features, undermining the consistency and validity of the extracted information [68, 69, 70]

The radiomic workflow has been applied in several medical context: to predict involvement of lungs in COVID-19 and pneumonia using CT [71]; to predict myocardial function improvement in cardiac MR images in patients after coronary artery bypass grafting [72]; for molecular subtype classification of low-grade gliomas in MR imaging [73]; in breast cancer for predicting prognostic biomarkers and molecular subtypes in MRI[74], to predict axillary lymph node status [75], to predict the nodal status in ultrasound considering clinically negative breast cancer patients [76]; and for many other applications [77, 78, 79, 80, 81, 82, 83].

In general, this technique has found prominent use in the medical imaging field due to

its ability to convert routine clinical images into meaningful data and information with high productivity, selectivity, and sensitivity [64]. Moreover, this accurate and precise feature extraction method doesn't necessitate a large amount of data, as in the case of deep learning architectures, making it particularly relevant for medical applications where data availability is a major concern. However, the radiomic features extraction involves the setting of numerous parameters and there is a high risk of model overfitting due to the high dimensionality achieved through extraction.

2.3.2.2 Standardized radiomic features

To guarantee task reproducibility, it is essential to standardize the process of feature extraction. As a result, an effort was made to promote the standardization of radiomic features, leading to the introduction of the Image Biomarker Standardization Initiative (IBSI)[84]. In the following studies, PyRadiomics [85], a software for extracting radiomic features, was employed, ensuring compliance with the IBSI standards. It is possible to divide the radiomic features into the following categories:

- First-order features: The first-order features provide essential information about the overall distribution of pixel intensities within the ROI, including statistical measures such as mean, median, standard deviation, minimum, maximum, skewness, and kurtosis.
- Gray Level Co-occurrence Matrix features: The Gray Level Co-occurrence Matrix (GLCM) was definite by Haralikc *et al.* [86] to capture the distribution of co-occurrence pixel values at a given distance d and angle θ. The features of this class require setting the distances d between the center voxel and the neighbor. The value of GLCM features was calculated for each angle θ separately and then the mean of these was returned. This latter approach should reduce the risk of overfitting.
- Shape features: These features are in no way related to the intensities of the voxels but to the size and shape of the mask. It is possible to extract 2D and 3D features.
- Gray Level Run Length Matrix features: The grey-level run length matrix (GLRLM) gives the size of homogeneous runs for each grey level, where the runs are consecutive pixels with the same gray level value. GLRLM features were introduced by Galloway *et al.* [87], later expanded by [88] and tested to the 3D case by Xu *et al.* [89], proving a great discriminatory for textures lung. The same considerations of GLCM could be made for the choice of angles.
- Neighboring Gray Tone Difference Matrix features: The Neighboring Gray Tone

Difference Matrix (NGTDM) was introduced by Amadasun *et al.* [90] and has been used to extract features capable of highlighting changes in the intensity of the voxels. This feature class requires setting the distance between a gray value and the average gray value of its neighbors.

- Gray Level Size Zone Matrix features: The Gray Level Size Zone Matrix (GLSZM) was introduced by Thibault *et al.* [91] to take into account that homogeneous texture is composed of large areas of the same intensity and not small groups of pixels or segments in a certain direction.

- Gray Level Dependence Matrix features: As described in [85] a Gray Level Dependence Matrix (GLDM) quantifies gray level dependencies in an image, where a gray level dependency is defined as the number of connected voxels within some distance that is dependent on the center voxel [92]. It requires the setting of the distance and the α direction.

2.3.2.3 Higher-level Radiomic features

The same features mentioned in the previous paragraph can be extracted from various transformed images, including Wavelet transforms, Fourier transforms, Laplacian-of-Gaussian filtering, exponential, logarithmic filtering, and others. Among the countless higher-level features, those derived from Wavelets appear to hold the most promise and widespread usage.

Wavelet-derived features have proven to be highly predictive in a range of scenarios: identifying tumor types in early-stage lung nodules *via* CT scans [93], predicting responses to neoadjuvant chemotherapy treatments for breast cancer through MRI [94], forecasting responses to low-dose rate radiotherapy treatment for gastric carcinoma using CT [95], detecting liver cirrhosis [96], differentiating glioblastoma multiforme from brain metastases through MRI [97], and grading pulmonary lesions in COVID-19 cases using CT [98].

The Discrete Wavelet Transform (DWT) is formulated through the high-pass h_ψ and low-pass h_ϕ filtering operation, representing a dilated and translated version of a particular signal. Wavelet-derived features are computed based on image four decompositions for 2D images such as X-rays, mammograms, and ultrasounds. Conversely, wavelet transform generates eight decompositions for 3D volumes such as CT and MRI scans. This results in multi-resolution images. In addition, there are also multiple families of Wavelet transform, some optimized for noise reduction and others for image compression. Nevertheless, they all generally provide a multi-resolution representation of the original image. Given these considerations, it's evident that the predictive capability

of wavelet-derived features surpasses that of original features. Thus, it's beneficial to examine and compare the behavior of wavelet kernels, to provide sound recommendations for their use. However, this also leads to a significant increase in the number of features which, if not handled correctly, can cause predictive systems to fall into the curse of dimensionality [99].

2.3.3 Radiomic *vs.* Deep Feature Extraction

The Radiomics approach presents numerous benefits when compared to deep extraction methods. While the latter demands substantial datasets for effective model training, radiomic feature extraction can be accomplished using smaller datasets [100]. Furthermore, shallow learning algorithms are suitable for categorizing radiomic data and for establishing predictive models based on Radiomics. A key advantage of radiomic features is their inherent interpretability, each feature has a clear meaning. The application of shallow learning algorithms and inherently intelligible features allows for both local and global explanations of the predictive models. Moreover, it provides physicians with the capacity to *i)* clinically justify research findings, *ii)* build trust in computerized systems, and *iii)* promote their incorporation into clinical procedures. Furthermore, radiomic features ensure a deterministic and reproducible tool for feature extraction.

The effectiveness of Radiomics-based methods has been validated in a multitude of disease settings and throughout various stages of healthcare delivery, including diagnosis [101], prognosis [102], assessment of treatment responses [103], and disease progression monitoring [104]. This has yielded promising outcomes [105]. However, deep learning architectures have the capability to derive abstract and higher-level features, rendering deep features more informative than their radiomic counterparts. Indeed, several studies have demonstrated that deep features outperform radiomic features in terms of overall performance [106, 107, 100, 108]. Consequently, the specific context and design prerequisites necessitate a balanced approach between the advantages and disadvantages of machine learning and deep learning methods.

A major drawback of radiomic features pertains to their reliance on ROIs segmentation. Segmentations can be performed automatically or manually, both with their unique challenges. Automated segmentations require validation by medical professionals, adding an extra step to the process. This validation is necessary to ensure accuracy and proper representation of the medical scenario, which may not always be fully captured by the algorithm. Manual segmentations, on the other hand, introduce the potential for inter-operator variability as they are directly dependent on the individual clinician's interpretation and judgment. Such variability can affect the consistency and

replicability of the radiomic analysis. Thus, conducting robustness studies of these features becomes a mandatory requirement. It helps in the identification and elimination of features that are excessively susceptible to such variations, thereby enhancing the reliability and accuracy of radiomic studies. These studies can also provide insights into standardizing the segmentation process to minimize the inconsistencies introduced by operator-dependent variables [68, 69, 70].

2.4 Machine Learning Methods in Medical Applications

2.4.1 Shallow Learning

Shallow learning techniques, also known as traditional or classic machine learning methods, refer to a class of simple algorithms that learn from data to make predictions or decisions. These techniques are called "shallow" because they typically have a simpler architecture compared to deep learning algorithms. Shallow learning methods rely on tabular data in the form of feature vectors for training. These algorithms expect structured input where each data point is represented as a fixed-length vector with explicit features. Therefore, when working with image data, it becomes necessary to convert the images into feature vectors using techniques such as radiomic feature extraction. In addition, in order to reduce the overfitting risk, preprocessing and feature selection turns out to be key step preceding training. Several well-known algorithms belong to the shallow learning paradigm, including Linear Regression, Logistic Regression, Random Forest, Decision Tree, XGBoost, Support Vector Machines, etc.

2.4.1.1 Feature preprocessing

Feature Scaling

Feature normalization or standardization is a crucial step in the preparation of data for machine learning models. It refers to the process of rescaling the features to follow a standard scale. Without normalization, features with larger scales can inappropriately influence the model, leading to longer training times and reducing the importance of the other features. By normalizing features, each feature contributes approximately proportionately to the final decision, ensuring better convergence and often leading to more accurate models. This is especially vital for algorithms that rely on gradient descent for optimization or those that compute distances, like k-means clustering and support vector machines. For other algorithms such as Decision Tree-based, feature scaling is not mandatory.

MinMax Normalization: This technique rescales the features between a specified range (typically 0 to 1). The minimum value of the feature becomes 0, and the maximum

value becomes 1. Given a value x, its normalized value x' is computed as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization (Z-score normalization): Standardization aims to rescale features to have a mean 0 and a standard deviation 1. Given a value x, its standardized value z is computed as:

$$z = \frac{x - \mu}{\sigma}$$

Where μ is the mean of the feature values and σ is their standard deviation.

Data Harmonization

Feature harmonization is an essential process, especially in multi-center studies where data is acquired from various centers using potentially different protocols or equipment. When data originates from multiple sources, inherent variations may arise, leading to inconsistencies in the dataset's distribution. Such inconsistencies can be attributed to the differences in acquisition procedures, hardware specifications, or other external conditions unique to each center. One significant challenge posed by these variations is the phenomenon of distributional drift. This refers to the changes in data distributions over time or across centers, where the statistical properties of the collected data might differ considerably. If unaddressed, distributional drift can lead to decreased model performance, as a machine learning model trained on data from one center might not generalize well to data from another center. Radiomic features are extremely susceptible to the distributional drift phenomenon. Feature harmonization aims to mitigate these issues by adjusting and standardizing the features across different datasets, ensuring consistency and improving the robustness of subsequent analyses or predictive modeling.

One example of data harmonization method used in several research works is ComBat [109, 110]. A salient advantage of the ComBat method is its capacity to directly engage with features that have previously been extracted from images. This obviates the necessity to access the original images during the harmonization process. Its primary objective is to align the feature distributions across various imaging protocols, ensuring consistency and reducing potential disparities introduced by different data acquisition methods.

2.4.1.2 Feature selection

Feature selection is a crucial step in the machine learning pipeline to identify and retain the most informative features from a dataset. The primary goal is to simplify the model, reduce training time, and counteract the curse of dimensionality, potentially leading to enhanced model performance. The methods for feature selection can be broadly categorized into three main types: filter methods, wrapper methods, and embedded methods.

Wrapper methods

Encapsulate the machine learning model within the selection process. These methods assess subsets of variables to maximize model performance, iteratively adding or removing features to determine the best feature combination. Examples include forward selection, backward elimination, and recursive feature elimination. In this thesis, the Sequential Forward Floating Selection (SFFS) algorithm was mainly used [111]. SFFS is a greedy search algorithm used to select a subset of features that is most relevant to the problem. Using the floating variant, a larger number of feature subset combinations can be sampled because an additional exclusion step is computed to remove features once they are included. While they can provide optimized feature subsets tailored to the model, they can also be computationally expensive, especially with high-dimensional data.

Filter methods

Filter methods evaluate the relevance of features based on their intrinsic properties, independent of any machine learning model. Being model-agnostic, they are generally faster than wrapper methods but might not capture feature interactions specific to a given model. They often use statistical measures, such as correlation coefficients, or mutual information, to rank and select features.

- **Near Zero Variance Analysis:** This step was designed to eliminate features that lack information content. Any feature with a variance lower than a specified threshold is considered uninformative and subsequently discarded.
- **Correlation Analysis:** The goal here is to identify and remove highly correlated features, thus minimizing redundancy. The Spearman correlation coefficient is typically employed for pairwise comparisons of features. A value larger than 0.80 are commonly used to consider two feature correlated [112, 113, 114, 115].

- **Statistical Test:** When working with very small datasets, it is common to use statistical tests to identify and select features that have significant relationships with the output. The rationale is to focus on those features that demonstrate strong, statistically evident associations, thereby potentially improving the model's accuracy with limited data. However, this approach can be risky. Machine learning models are adept at uncovering non-linear relationships between inputs and outputs, nuances that traditional statistical tests might overlook. Moreover, while a feature may not exhibit statistical significance alone, it can become predictive when combined or interacted with other features. By prematurely discarding such features based solely on univariate statistical tests, complex patterns and relationships in the data can be compromised. Thus, while statistical tests can guide feature selection, relying exclusively on them can limit the potential of machine learning models to capture the full complexity of the data. For continuous variables, one common non-parametric test is the Mann-Whitney U test. This test is used to determine if there are significant differences between two independent groups on a continuous or ordinal dependent variable. For binary or categorical variables, there are several tests, but the most common one is the Chi-squared test. Another option is Fisher's exact test, which is especially useful when the sample sizes are small [116].

- **Mutual Information:** Uses a measure of entropy, termed "mutual information," to determine the features to be incorporated in the reduced data set [117]. To elaborate, mutual information assesses the dependence between two random variables. Specifically, it quantifies the extent to which knowledge about one variable informs us about the other.

Embedded methods

Embedded methods incorporate feature selection as part of the model training process. For instance, regularization methods like Lasso or decision tree-based algorithms inherently perform feature selection by assigning lower weights or importance to less relevant features. Two methods used in this thesis:

- **L1-based:** This approach employs a linear model with an L1 penalty, which inherently reduces the number of features by setting certain coefficients to zero. The underlying premise is that a linear model subjected to an L1 norm penalty yields sparse solutions. For example, a Linear Support Vector Classifier algorithm is trained using this method, and only features associated with non-zero coefficients are retained for subsequent modeling.

- **Tree-based:** This method exploits decision-tree algorithms for feature selection. Specifically, tree-based estimators compute feature importances based on impurity metrics [118]. These importances aid in differentiating between pertinent and irrelevant features, allowing for the exclusion of the latter.

2.4.1.3 Class Imbalance Management

An important and common issue of classifier training lies in the use of imbalanced datasets. In an imbalanced data scenario, one class significantly outnumbers the others, which poses a significant challenge in machine learning. Such an imbalance can lead models to become biased towards the majority class, often resulting in poor predictive performance for the minority class. To address this issue, a range of techniques have been developed to either oversample the minority class, undersample the majority class, or both.

Among these techniques, SMOTE (Synthetic Minority Over-sampling Technique) stands out as one of the most popular and effective methods for handling imbalanced datasets. Developed by Chawla *et al.* [119], SMOTE works by generating synthetic samples in the feature space. Instead of simply replicating minority class instances, SMOTE selects two or more similar instances and perturbs an instance's feature values, creating a "synthetic" instance that, while not an exact duplicate, is consistent with existing instances. This oversampling strategy not only boosts the number of minority class samples but also introduces variability, making models less likely to overfit compared to simple oversampling.

2.4.1.4 Shallow Learning Methods

Tree Ensemble algorithms have demonstrated exceptional performance, particularly in classifying small datasets [120, 121, 122]. Random Forest (RF) and XGBoost (XGB) are two widely used Tree Ensemble algorithms, both renowned for their effectiveness in various classification tasks. In XGB, the primary objective is to minimize the model's loss function by incorporating weak learners using gradient descent, making it a type of Boosting Ensemble Method. On the other hand, RF uses the bagging technique to construct multiple weak learners by considering random subsets of features and bootstrap samples of the data. The decisions of each learner are then aggregated, forming an ensemble model through a process known as Bagging Ensemble Method. Support Vector Machines (SVM) is a powerful machine learning algorithm primarily used for classification tasks. The primary objective of SVM is to identify the optimal hyperplane among this infinite set. The SVM algorithm considers some data more important than others for finding the best hyperplane: the support vectors. They are

the samples (data points) most important to define the position and orientation of the best decision boundary (i.e., the separating hyperplane). The distance between the separating hyperplane and the support vectors is called the margin. These support vectors are crucial in defining the best decision boundary. SVM aims to find the hyperplane with the largest margin, making it more robust to new, unseen data. In cases where a linear hyperplane cannot effectively separate the data, SVM can use kernel functions to transform the original feature space into a separable space, allowing for nonlinear classification.

Alongside SVM, Tree Ensemble algorithms rank among the most frequently used techniques in various machine learning applications [120, 121, 122, 123].

2.4.2 Deep Learning

When data is unstructured and complex, such as images or time series, deep learning models generally become the first choice. These models are inherently designed to handle high-dimensionality data, variance, and complexity, making them adept at uncovering hidden patterns within such datasets. Ad discussed in the previous Section 2.3.1, deep learning methods are an invaluable tool for feature extraction. Leveraging layered architectures, deep learning models like CNNs and Recurrent Neural Networks (RNNs) transform the raw, unstructured data through multiple processing layers. For instance, CNNs, have the ability to learn hierarchical patterns from pixel level to complex object details, which a shallow learning algorithm may overlook. Similarly, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are extensively used for time-series data, exploiting their ability to remember previous inputs in their hidden layers, thereby creating an internal context. Thus, deep learning models have become an essential tool for dealing with unstructured data, offering remarkable accuracy and predictive power where traditional machine learning methods may fail.

Deep learning architectures, particularly CNNs, are widely employed to solve classification problems but also object detection tasks. Each of these tasks, while interconnected, has distinct objectives. Classification is a task in which the model is trained to assign input data to one of several predefined categories. For instance, in image classification, a model might be trained to identify whether a given image contains a malignant or benign tumor. The model makes its prediction based on the entirety of the input and doesn't concern itself with the location or number of instances of the class within the image. On the other hand, object detection goes a step further by not only identifying what objects are present in an image but also determining where they are located. It is a more complex problem, as it involves both classification (what) and localization (where). An object detection model would, for example, identify and localize multiple

instances of classes within an image. So, if an image contains one benign and one malignant cancer, the model would classify the objects and provide bounding boxes around each cancer. In both these tasks, deep learning models like CNNs-based have proven to be immensely effective. They can automatically learn and extract features from raw data through multiple layers of processing, which greatly enhances their predictive performance for both classification and object detection tasks.

In this thesis, CNNs and Transformers are the main deep learning methods, for this reason introduced in the following subsections.

2.4.2.1 Convolutional Network Foundamentals

Shallow learning training for medical image analysis requires that the images or ROIs are described through feature vectors. These features can be manually designed using Radiomics, capturing the relationships between gray scales, texture, and shape of a ROI. In contrast, Convolutional Neural Networks intrinsically incorporate feature extraction into their processes [124]. When presented with an image or ROI, CNNs independently identify the relevant features for the task (e.g., classification, object detection). This built-in capability has propelled CNNs to the forefront of medical image analysis [125, 126, 127]. Their foundational concept is influenced by studies centered on the brain's visual cortex [128, 129]. Due to their significance, CNNs have spurred vast research interest, giving rise to a plethora of complex architectures. This segment delves into the mechanics of renowned architectures, including VGG, ResNet, Inception, and more [130].

At its essence, a CNN consists of multiple sequential layers. It initiates with the input layer, representing the image. This image is fundamentally a matrix of pixels, denoted by width, height, and channels. Colored images are composed of three channels (RGB), while grayscale presents one single channel. Subsequently, Convolutional and Pooling Layers alternate.

Convolutional Layers employ convolution operations to convert input images into more expressive spaces, converting the input image dimensions into a feature vector. This convolutional segment, dedicated to feature extraction, applies a convolution matrix — commonly referred to as a kernel or filter — to the input image. Conventionally in image processing, kernel configurations are tailored based on targeted features, including elements like edge detection and noise mitigation [131]. However, within CNNs, kernel values adjust during training, with the network itself determining the function of each filter. The convolutional processes yield feature maps. For optimization and performance enhancement, these maps may undergo dimension reduction *via* Pooling Layers. These layers can leverage techniques such as average pooling (taking the mean

of feature map regions) or max pooling (selecting the paramount value).

The alternation of Convolutional and Pooling Layers equips the CNN to identify both foundational and advanced features. The initial layers discern elementary features, but as progression occurs, these features mature into abstract, critical elements for classification. At the end of this process, the emergent feature vector feeds into a Multilayer Perceptron, responsible for the final classification. This terminal phase boasts densely interconnected layers, often termed dense layers. It is possible to replace the dense layer with the shallow learning methods discussed in the previous sections.

2.4.2.2 Vision Transformer Foundamentals

The Vision Transformer (ViT) is a novel neural network architecture for computer vision tasks. It leverages the self-attention mechanism typically found in transformer models for NLP analysis (e.g., BERT, GPT). Instead of processing images using traditional convolutional layers, the ViT approach begins by dividing an input image into fixed-size non-overlapping patches. Each patch is then linearly embedded into a flat vector, and a positional encoding is added to retain spatial information. These vectors serve as the input sequence to a series of transformer encoder layers. Each transformer encoder consists of multi-head self-attention mechanisms and feed-forward neural networks, interspersed with layer normalization and residual connections. Following the transformer encoders, a classification token is appended to the sequence's start, and after being processed through the model, it's used for the final classification. A feed-forward head on top of the processed classification token yields the final prediction. In particular:

1. *Input Image*: The raw image input.
2. *Patching*: The input image is divided into fixed-size patches, often non-overlapping. For example, for an image of size 384×384 and patch size 16×16.
3. *Patch Embedding*: Each patch is linearly embedded into a flat vector. This is typically done using a simple feed-forward neural network that transforms the patch from a 16×16×3 tensor (assuming RGB image) to a 1D vector.
4. *Positional Encoding*: Since the transformer does not have an inherent knowledge of the spatial position of the patches, positional encodings are added to the patch embeddings. These encodings ensure that the transformer recognizes the relative or absolute position of patches within the image.

   Transformer Encoder Layers: This is where the core computations happen:

   (a) Multi-Head Self-Attention: Each input vector is processed through multiple

'heads' of self-attention layers. These heads allow the model to focus on different parts of the image and capture various types of relations and patterns. The outputs from the heads are concatenated and linearly transformed.

(b) Feed-forward Neural Networks: After the attention mechanism, each position (patch embedding) is passed through a feed-forward neural network (independently). This network is the same for each position.

(c) Residual Connections: Both the attention and feed-forward mechanisms have skip (or residual) connections around them. This promotes easier training and better gradient flow.

(d) Layer Normalization: Normalization is applied before each sub-layer, and the result is passed through a non-linear activation function (usually ReLU or GELU).

5. *Final Layer (Head)*: The processed classification token is passed through a feed-forward head (often just a linear layer) to produce the final classification outputs.

2.4.2.3 Transfer Learning

The remarkable success of neural Networks is closely linked to the rise of extensive databases containing hundreds of thousands of samples. In medical scenarios, obtaining such vast databases is quite challenging. Given that NNs for classification tasks operate on supervised learning principles, each sample data requires a corresponding label. The act of matching every sample with its appropriate label is termed annotation. The label can be a manual delineation of a ROI (segmentation mask), a histological examination results, or a combination of both. Concisely, annotating can be costly, complex, and sometimes invasive. Consequently, this often leads to datasets limited to mere hundreds or even dozens of samples, insufficient for comprehensive NN training.

Transfer Learning (TL) emerges as a solution in situations where training on a source big database aids in enhancing the generalization to a target dataset. In the context of neural networks, it's common practice to initially train a NN using a larger dataset for a primary task, called *source* dataset. Subsequently, the acquired weights are adjusted and repurposed for a secondary dataset, called *target* dataset [132]. One approach is to employ the model trained on the source dataset for feature extraction and fine-tuning only the dense layers (those associated with classification) on the target dataset. However, this requires a high degree of similarity between the target and source datasets, particularly in the features they encompass. An alternative strategy involves transferring all the weights from the source dataset and optimizing them for the target dataset. Employing TL on a target database typically leads to enhanced

classification in the initial epochs, accelerated learning, and superior final classification performance [133]. Furthermore, this method isn't exclusive to NNs; it's applicable to general machine learning models [134], Bayesian Networks (BNs) [135], and within Reinforcement Learning contexts [136].

2.4.2.4 Data Augmentation

Data augmentation (DA) plays a pivotal role in the field of machine learning. At its core, data augmentation involves introducing variety into a dataset by applying various modifications and transformations to the original data samples. Data augmentation increases the amount of data available for training and aims to extend and introduce noise into the original data distribution, improving the generalization in real cases. In addition, it is widely used in for class balancing. A balanced dataset ensures that the model doesn't become biased towards the more prevalent classes and provides a more generalized performance. In essence, data augmentation enhances the robustness of a model by exposing it to varied data instances, preventing overfitting, and promoting better generalization to unseen data. Data augmentation techniques are based on the knowledge that, in nature, data rarely follow an orderly distribution; Thus, by training a model on a noisier version of the original data, it is possible to train the model on real-world scenarios.

There are several data augmentation techniques. The most common involve geometric and intensity transformations, better known as traditional techniques. Recently, with the spread of generative models, data augmentation with generated images also appears to be a promising technique.

Traditional Data Augmentation

Traditional or geometric data augmentation includes a wide plethora of techniques designed to introduce spatial and visual variations into a dataset. One of the most common methods is rotation, where images or data points are turned by a certain angle, simulating the different orientations an object might be viewed from in real-world scenarios. Flipping is another widely-used technique, which involves reflecting data samples across their vertical or horizontal axis. This can be especially useful for recognizing objects or patterns regardless of their spatial orientation. Adding noise to data serves as an effective way to simulate real-world imperfections and disturbances, ensuring models are trained to be more robust and not overly sensitive to minor changes. Meanwhile, contrast enhancement alters the visual disparity between the light and dark areas in an image, aiding in the recognition of objects in varying lighting conditions. All these geometric augmentations, by introducing a diversified set of transformations,

ensure that machine learning models are better equipped to handle the numerous conditions they might encounter when deployed in real-world environments.

It is possible to apply data augmentation techniques in two ways:

- *As preprocessing step*: in this case, the training images are augmented before training through some established transformations. Then the original dataset plus the augmented dataset represents the input for training.
- *During the training*: the purpose is to transform the input directly during training. Specifically, random transformations are applied to each batch, which can vary from batch to batch. For example, it is possible to apply a flip upper-down with a certain probability (0.5 probability) or apply a transformation in a certain range for example (0, 30 degrees). For each image, these transformations will then be applied with the defined probabilities and ranges, which will differ in each batch.

Diffuser-based Data Augmentation

In recent years, Diffusion models have gained prominence in the domain of synthetic image generation. They surpassed other established methods such as Variational Autoencoders (VAE) [137] and Generative Adversarial Networks (GAN) [138], as highlighted by Dhariwal *et al.* [139]. These models have the capability to produce data samples that can fill gaps in the existing dataset, especially in areas where traditional augmentation techniques might not suffice. For example, GANs can generate entirely new images that capture the inherent distribution of a dataset, providing a richer set of samples for training. On the other hand, diffusion models work by introducing noise into a data sample and then gradually denoising it. This process effectively simulates the natural variability and imperfections present in real-world data. By using diffusion processes, it is possible to generate a continuum of data variations, making it possible to train models with additional synthetic images. In essence, through the use of generative and diffusion models, data augmentation has transcended basic geometric transformations, paving the way for more sophisticated and enhanced datasets. Considering the cost of training and inference of generative models, typically the dataset is augmented before training with synthetic data. Then training is performed considering the comprehensive dataset.

2.4.3 Shallow Learning and Deep Learning

Shallow learning and deep learning represent two subsets of machine learning that have been applied to a wide array of tasks in various fields. Deep learning architectures, such

as Vision Transformers, Convolutional Neural Networks, and others, offer a revolutionary approach to image processing and feature extraction. These models are capable of directly learning from raw image data without the need for explicit feature engineering or vectorization. By employing multiple layers of hierarchical representations, deep learning models can automatically extract meaningful patterns and features from the images, allowing them to grasp complex spatial structures and relationships. This end-to-end learning process enables deep learning models to achieve remarkable performance on various image-related tasks. These models have the added advantage of being able to use transfer learning, a technique where a pre-trained model on one task can be fine-tuned for a related task. This can significantly reduce training time and data requirements, allowing for the leveraging of existing knowledge across domains. Deep learning models can capture complex patterns and nonlinear relationships within the data but are often computationally expensive, require large amounts of data, and may lead to overfitting if not properly regulated. The primary difference between the two approaches lies in their architectural depth, leading to varied abilities in feature extraction and representation.

Neural Networks are high-performance models, but their training involves setting a plethora of hyperparameters [140]. These include the selection of activation functions [141], which can be part of the training process, the choice of optimizer and loss function, and determining the learning rate for weight updates. Other parameters involve deciding the number of epochs, the batch size for weight updates, as well as the architecture details like the number of layers and neurons in each layer. The challenge lies in the fact that there's no universal rule for setting these hyperparameters, which can vary depending on the specifics of the case. Factors such as the dataset size, the number of features under consideration, and the type of task being performed, whether classification or regression, can influence these settings. A noteworthy point to consider is the depth of the network architecture. While it may seem advantageous to implement deep architectures for enhanced extraction of feature hierarchies, the Universal Approximation Theorem states that even NNs with a single hidden layer can serve as universal approximators [142, 143]. Hence, there is a balance to be struck when designing the depth of the model.

Shallow learning models are often more interpretable, computationally efficient, and can perform well with less data. Shallow learning generally excels in tasks where simplicity and transparency are essential, and the underlying patterns are not overly complex. Deep learning is often favored for complex tasks involving unstructured data like images, speech, or text, where higher-level abstraction is required. As guidelines, shallow

learning can be more suitable when computational resources are limited, interpretability is crucial, or data is scarce. In contrast, deep learning could be the better choice for tackling more complex problems where large labeled datasets are available, transfer learning is a viable option, and the model's complexity and computational cost are justified by the task's demands. In each case, research moves to achieve a trade-off between explainability and accuracy. Figure 2.2 shows this trade-off focused mainly on medical image analysis. The radiomic feature extraction and the use of shallow learning methods bring the solution towards a more explainable model. In fact, the use of intrinsically intelligible data and XAI methods allows a comprehensive insight into the trained shallow learning models.

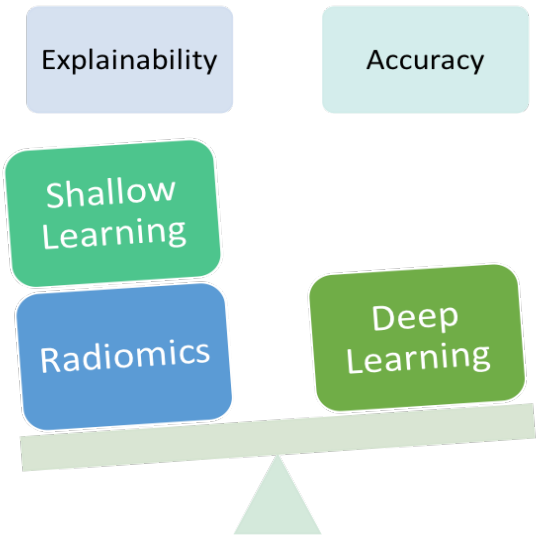


Figure 2.2: Trade-off between Explainability and Accuracy.

Moreover, a significant advantage of deep learning models lies in their ability to train with less reliance on additional information or preprocessing. In fact, in traditional Radiomics approaches, an annotated mask identifying the region of interest is crucial, along with the label indicating the class of the lesion. This mask is used for extracting features from the highlighted region, a process that requires significant manual intervention and domain expertise.

In stark contrast, deep learning models are inherently designed to identify and prioritize salient regions or features within the data during the training process. As a result, they do not necessitate a predefined mask for feature extraction. Leveraging their multi-layered architecture, deep learning models learn to assign importance to regions in the data that are most relevant to the task automating the feature extraction process. This autonomous capability of deep learning models is particularly beneficial in tasks like object detection. Traditional shallow learning methods are not suitable for such

tasks, as they often require handcrafted features and lack the capacity to learn spatial hierarchies, crucial for object detection. On the other hand, deep learning models can detect and localize objects in an image without requiring explicit annotations for each object's features.

2.4.4 How to Choose a Classifier

Selecting an appropriate training strategy and classifier is a notable challenging process, demanding careful consideration of numerous facets [144]. The renowned "No free lunch" theorem [145], fundamentally asserts that, on average, no pair of algorithms can outperform all others across the entirety of conceivable problems. This suggests that some algorithms may perform as well as extremely simplistic approaches, like random search, making it challenging to define one algorithm as superior to another. Nevertheless, depending on the specific task at hand, there are instances where particular algorithms are more advisable than others [146, 147].

In the course of this thesis research, four key characteristics have been identified that should serve as guiding principles when making decisions regarding the choice of a classifier: i) task analysis, ii) dataset size, iii)explainability requirements, and iv)available computing resource.

2.4.4.1 Task Analysis

The initial choice of classification approach depends on the task and the characteristics of the input. In particular, the algorithms introduced in the preceding section are not universally suited for both binary and multiclass classification scenarios. For instance, SVM and Logistic Regression (LR) are primarily designed for binary classification and necessitate the implementation of specialized techniques, like One-versus-Rest or One-versus-One strategies [148, 149], to adapt them to multiclass classification tasks.

The previously discussed algorithms possess versatility, capable of handling both continuous and discrete features. Nevertheless, in certain situations, specific configurations might become essential. For instance, when employing KNN, it might be necessary to use the Hamming distance metric for binary variables.

2.4.4.2 Dataset Size

The exponential expansion of accessible data is a driving force behind the evolution of machine learning techniques. While there is no strict threshold defining the minimum number of instances required to train an algorithm, it is generally regarded as questionable to work with fewer than 50 instances [150]. Some statistical analyses have focused

on establishing a relationship between the number of features and training samples. For instance, in the case logistic regression, it has been observed that a minimum of 10 to 15 samples per feature is necessary to yield reasonably stable estimates [151]. In the context of small datasets, it is advisable to opt for simpler algorithms such as logistic regression or linear SVM and to avoid deep learning algorithms. Tree Ensemble, on the other hand, has demonstrated its efficacy for classification tasks with limited data [121, 120, 152], and it is frequently used in conjunction with SVM. Deep learning solutions are more suitable when dealing with abundant data, and they can be justified for small datasets only if a larger dataset is leveraged for transfer learning purposes.

2.4.4.3 Explainability Requirements

As discussed in Section 2.2 the significant insufficient transparency of machine learning algorithms represents a pivotal challenge for the integration of these systems into clinical practice. For this reason, XAI has emerged to address the problem of poor interpretability, to make the learned logic accessible and the process understandable by humans [153, 23, 31, 154].

Shallow learning algorithms such as DTs, LR inherently offer interpretability, meaning their decision-making processes can be understood without the need for XAI techniques. This makes them preferred choices when dealing with limited data, and when simple, linear models suffice. In contrast, other shallow learning algorithms such as SVM and Tree Ensembles lack inherent explicability, but various XAI methods can be applied to provide both global and local explanations [9].

When interpretable inputs like clinical or laboratory data are used, explanations are easily accessible. Conversely, learned features, such as those extracted *via* CNNs, are often unintelligible. In such cases, explanations frequently involve the computation of saliency maps, which highlight the most influential regions within images during the prediction process, providing a kind of local explanation. However, it's worth noting that saliency maps can sometimes yield inconsistent explanations, as demonstrated by Gu *et al.* [46] and Zhang *et al.* [155]. As a result, shallow learning solutions are often preferred over deep learning approaches when explainability is a paramount concern.

2.4.4.4 Available Computing Resources

Contemporary mid-range computers offer ample computing resources for training shallow learning algorithms. However, when it comes to deep learning models, the demands are considerably higher, necessitating a high-performance graphics processing unit (GPU). Moreover, GPUs with substantial memory become essential, especially when working with architectures containing millions of parameters. In this context,

cloud computing services such as Google Colaboratory (https://colab.research.google.com/) can serve as an excellent solution, particularly for deep learning training in small to medium-scale applications. They provide the necessary computational power and memory resources, facilitating the development and training of complex deep learning models without the need for extensive local hardware investments.

Chapter 3

Machine Learning Applications in Breast Cancer

This chapter will focus on breast cancer analysis. Several medical image modalities were exploited, including MRI [156, 157], Mammogram [158, 159, 160, 161] and Ultrasounds [162, 163].

3.1 Introduction

Breast cancer stands as the most prevalent tumor among women [164]. Evidence from previous randomized trials and incidence-based mortality studies suggest that participation in breast screening programs significantly lowers breast cancer mortality rates [165]. However, there are persistent concerns related to false positives and negatives. Many of these inaccuracies can be attributed to factors such as the masking effect of dense breasts, and human errors including radiologist perception and decision-making mistakes. Moreover, the inherent imaging characteristics of tumors also exacerbate this issue, as benign masses often appear similar to malignant ones, and conversely, malignant masses sometimes mimic benign ones [166].

Radiologist's diagnostic process aims to describe the regions of interest using the Breast Imaging-Reporting and Data System (BI-RADS) Atlas [167]. The BI-RADS lexicon includes standardized terms and descriptors that help radiologists communicate their findings accurately and consistently. It categorizes breast imaging findings into different levels of suspicion for malignancy, ranging from 0 to 6:

0. Incomplete: Additional imaging or evaluation is required.
1. Negative: The breast tissue appears to be entirely normal.

2. Benign: Findings indicate a benign (non-cancerous) condition.

3. Probably Benign: Findings are likely benign, but short-term follow-up is recommended.

4. Suspicious: Findings are suspicious for malignancy, with increasing levels of suspicion (4A, 4B, 4C) indicating higher likelihood.

5. Highly Suggestive of Malignancy: Findings are highly suspicious for cancer.

6. Known Biopsy-Proven Malignancy: A biopsy has confirmed the presence of cancer.

Each category includes specific descriptors for different types of abnormalities, such as masses, calcifications, asymmetries, architectural distortions, and other features. The BI-RADS lexicon also provides guidance on the appropriate follow-up and management based on the assigned category. Using machine learning methods, the goal is to develop models that automatically assess the severity of injuries, providing support to the physician's diagnostic process. In addition, explainability is a prerequisite for the validation and justification of model decisions.

There are several diagnostic tests for the prevention and characterization of breast cancer. In this thesis, three different image acquisitions were considered: Mammography, Ultrasound, and Magnetic Resonance. These three modalities are both important imaging techniques used in the detection and diagnosis of breast cancer. The choice of which imaging modality to use depends on various factors, including the purpose of the examination, the patient's age, breast density, and specific clinical indications. In many cases, both mammography and ultrasound may be used together to provide a comprehensive evaluation of breast health. Typically, MRI is considered a more advanced examination that is performed in the presence of abnormalities found by previous modalities. For this reason, each proposed work uses a different dataset and methods, which will be discussed separately in each subsection below.

3.2 Breast Cancer Classification in DCE-MRI using Radiomic Features

Dynamic Contrast-Enhanced magnetic resonance imaging (DCE-MRI) plays a central role in the diagnosis of breast lesions by providing both morphological and hemodynamic information. This imaging technique evaluates the vasculature at multiple time points after intravenous contrast injection, allowing quantitative analysis of signal changes through enhanced dynamic properties [168]. Examining variations in the uptake of contrast agent, including factors such as initial peak enhancement and the presence of a delayed washout period, it is possible to improve the malignity specificity prediction. It is established that malignant lesions show an immediate increase in signal intensity, whereas benign lesions show a slower increase in signal intensity [169]. Therefore, DCE-MRI is an MRI sequence whose signal intensity changes due to the contrast agent. The physician's aim is to evaluate the signal changes obtained by the entire sequence to diagnose the disease. However, although DCE-MRI has been widely used to improve upon MRI in characterizing breast lesions [170], its specificity remains suboptimal [171, 172, 173].

Deep learning models have shown impressive results in medical image analysis. Convolutional-based architectures were employed in several tasks related to breast cancer, including, classification, segmentation, detection, etc. [174, 175, 176]. Also for Time Series Analysis [177] several deep learning methods were proposed. These methods focus on extracting informative features and exploiting the abstraction mechanism of deep neural networks, as discussed in Section 2.3.1. However, training deep architectures places a significant need for large amounts of well-annotated data. More importantly, the extracted deep features are unintelligible, making it difficult to give medical interpretation to the trained models. In practice, deep features are often interpreted using a saliency map, which shows the most important pixels for prediction. Saliency maps only generate a local explanation (i.e. for each instance of the dataset), while they do not allow a global explanation of the model. However, to clinically validate the systems and compare them with the medical literature, a comprehensive explanation is required.

For this reason, the radiomic workflow (discussed in detail in Section 2.3.2) was implemented for the DCE-MRI analysis in several works. Four separate heuristic parameter maps were used by Gibbs *et al.* [178] to train support vector machines. They focused on BI-RADS 4 or 5 classification of breast lesions less than 1 cm across a data set of 165 lesions. In Chu *et al.* [179] parameter maps were mined and the 133 court was used to train the logistic regression model. In their study, Parekh *et al.* [180]

collected the radiomic characteristics of multiparametric breast MRI images, including DCE-MRI. They then grouped and ranked these features using the isoSVM algorithm, with a dataset consisting of 124 patients. Zhang *et al.* [115] leveraged five imaging modalities for feature extraction, including DCE-MRI and a support vector machine trained for the classification task. Nagarajan *et al.* [181] deals with extracting specific features from five post-contrast images. They then evaluated the performance of a support vector regressor and a fuzzy k nearest neighbor classifier for classifying small lesions. Militello *et al.* [101] leveraged several feature selection algorithms on a court of 111 patients and a trained support vector machine for classification.

In this work, two separate analyses were performed for breast cancer classification in DCE-MRI, using a proprietary multi-protocol dataset acquired at the University Hospital "Paolo Giaccone" (Palermo, Italy).

- The first one follows the previous works, in which radiomic features extracted from the time instants of the DCE-MRI sequences are used as input for shallow learning classifiers [156]. In particular, extraction of radiomic features was performed at seven instants of the DCE-MRI sequence (one pre-contrast and six post-contrast images), considering the original images, and filtered with a Laplacian of Gaussian filter and wavelet transform. The features of the seven instants are used to select the most descriptive features for breast cancer classification. Then the goal was to identify the best instant of the DCE-MRI sequence for classification. Random Forest, XGBoost, and Support Vector Machines are compared for classification and tested on independent test dataset. An explanation of the best results obtained is also provided through the Shapley values [50] to show the overall importance of each feature in the final prediction.

- The second one proposes a novel approach for the DCE-MRI analysis [157]. In fact, to the best of our knowledge, no work has been proposed to classify breast cancer in DCE-MRI using time series analysis algorithms. Specifically, all seven instants of the DCE-MRI sequence were analyzed simultaneously. A comparative analysis of several time series analysis algorithms was performed. Figure 3.1 shows the general workflow. In particular, after feature extraction and harmonization, the best predictive features were selected through the implemented multi-instant feature selection and univariate time series classification. Then, the most accurate univariate models are clustered through a voting mechanism to implement a multivariate time series classifier. The goal of this model is to mimic a physician's diagnostic assessment of a DCE-MRI sequence by evaluating the changes and trends in radiomic features produced by contrast agent administration. Additionally, through the use of intelligible radiomic features, it was

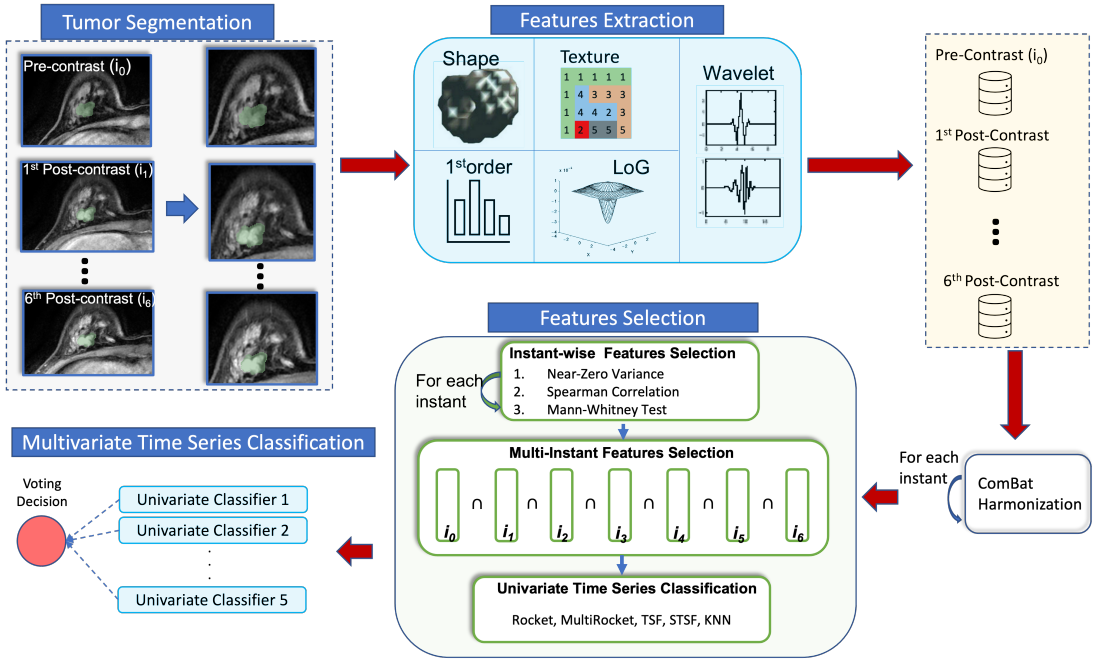


Figure 3.1: General workflow for breast cancer classification using Time Series Analysis methods.

possible to interpret model results and validate results clinically.

3.2.1 Materials and Methods

This section introduces two distinct approaches for breast cancer classification in DCE-MRI. The first approach uses traditional shallow learning methods to identify the most predictive instant within the DCE-MRI sequence. The second approach leverages the entire sequence through sequence analysis methods for classification. Both approaches share the same dataset, feature extraction, and preprocessing procedures. However, there are important variations in feature selection and classification methods employed between the two approaches.

3.2.1.1 Dataset

Patient Population

One hundred sixty-six breast mass enhancements were included in the DCE-MRI studies. The masses presented a mean size of 15.3 ±10.5 (Range: 3-75;), acquired in 103 female and 1 male patients with a mean age of 51 ± 11 (Range: 31-79 years). Examinations were performed at University Hospital "Paolo Giaccone" (Palermo, Italy), from April 2018 to March 2020. Two expert radiologists assessed the benign and malignant class in consensus. Seventy-three samples were recognized as 2-3 BI-RADS and 93 as

DCE-MRI sequence for each sample were transformed into time series of length 7 (one pre-contrast and 6 post-contrast). An initial univariate analysis was performed to select the best features and classifier for time-series analysis. Then, multivariate time series classification is performed through a voting mechanism, considering the best algorithms and features found in the univariate phase. Model and feature evaluation were performed taking into account the accuracy calculated in a stratified 10-fold cross-validation procedure repeated 20 times. Finally, the multivariate time series classification model is evaluated on an independent test set. No shape features were considered for time series analysis, considering the mask was fixed for all time instants.

Tuning univariate time-series feature Selection

Considering the large number of features selected after the previous selection and preprocessing step, an initial univariate time series analysis was performed. In fact, the classifiers deployed for time series classification are initially used in a univariate manner for two main purposes: 1) feature selection to determine the most discriminating features for time series classification and 2) evaluation to determine the optimal time series classifier. The univariate feature selection step represents a wrapper approach because the feature selection process is based on a specific machine learning algorithm.

Rocket and MultiRocket

The RandOm Convolutional KErnel Transform (ROCKET)[193] algorithm is a kernel-based classifier that applies random convolutions kernel on the time series to produce two main features:

- maximum (max): the maximum value (equivalent to global max pooling)
- portion of positive values (ppv): $ppv(Z) = \frac{1}{n}\sum_{i=0}^{n}[z_i > 0]$, where Z is the output of the convolution operation.

Then, a RidgeClassifierCV algorithm is used for classification, As example, using 500 kernels Rocket produces 1000 features for each time series. Despite the high-dimensionality of the extracted features and the dataset's small size, Rocket has been shown to provide high classification accuracy when used as input to a linear classifier (e.g., ridge regression) [193].

Rocket classifier is used as the reference algorithm for time series classification. However, the MultiRocket algorithm has also been exploited as it outperforms its predecessors Rocket and MiniRocket[194] in terms of accuracy[195]. The MultiRocket uses fixed kernels such as MiniRocket, with a fixed length and weights and dilatations. Additionally, it uses 4 pooling operators on the convolution output:

- Proportion of Positive Values (PPV), the same as described in Rocket.
- Mean of Positive Values (MPV), to capture the magnitude of the positive values. It is defined as $MPV(Z) = \frac{1}{m}\sum_{i=1}^{m} z_i^+$ where z^+ represents a vector of positive value of length m.
- Mean of Indices of Positive Values (MIPV), to capture information about the relative location of positive values. It is defined as:

$$MIPV = \begin{cases} \frac{1}{m}\sum_{j=1}^{m} i_j^+ & if \;\; m > 0 \\ -1 & otherwise \end{cases} \tag{3.1}$$

- Longest Stretch of Positive Values (LSPV) returns the maximum length of any subsequence of positive values, calculated as $LSPV(Z) = max[j-i|\forall_{i \leq k \leq j} z_k > 0]$

The main novelty of the algorithms draws inspiration from DrCIF algorithm [196]: the original time series is converted into its first-order difference and used also for feature extraction. For this reason, MultiRocket enhances the number and the meaning of the extracted features [195].

Time Series Forest and Supervised Time Series Forest

The Time Series Forest (TSF) algorithm [197] is a time interval-based classification algorithm. TSF trains a random forest model with features extracted by the interval split into $\sqrt{m}$ intervals. Features include mean, standard deviation, and slope, and m represent the length of the series. Additionally, the Supervised TSF (STSF)[198] was implemented. STFS improves TSF in terms of efficiency by selecting only discriminatory intervals *via* a supervised step. Median, interquartile range, minimum, and maximum were introduced features. There is evidence that for some datasets, STSF achieves comparable accuracy to state-of-the-art time series classification methods while being significantly more efficient.

K-Nearest Neighbors

The K-Nearest Neighbors classifier for time series is a distance-based algorithm where specific metrics are used to compute the distances between samples. It represents a reference for time series classification because it is simple and does not require tuning of numerous hyperparameters. In this work comprehensive comparison of different metrics was made. In particular, distances based on Dynamic Time Warping (DTW)[199], derivative DTW (DDTW)[200], weighted DTW (WDTW) and weighted

Table 3.6: MultiRocket validation accuracy of the five most accurate features. The last line represents the average accuracy value.

Category	Feature	Class	Accuracy
FO	Energy	Original	0.694
FO	TotalEnergy	Original	0.687
NGTDM	Busyness	LoG $\sigma = 2$	0.679
GLCM	Imc2	Wavelet HLH	0.676
GLDM	DependenceEntropy	Original	0.673
			0.681

Table 3.7: TSF validation accuracy of the five most accurate features. The last line represents the average accuracy value.

Category	Feature	Class	Accuracy
FO	Energy	Original	0.632
FO	TotalEnergy	Original	0.636
GLDM	LDHGLE	Wavelet LLH	0.635
GLCM	Imc2	Wavelet HLH	0.630
GLSZM	SmallAreaEmphasis	Wavelet HHH	0.641
			0.635

Table 3.8: STSF validation accuracy of the five most accurate features. The last line represents the average accuracy value.

Category	Feature	Class	Accuracy
FO	TotalEnergy	Original	0.651
NGTDM	Strength	Original	0.648
GLCM	Imc2	Wavelet HLH	0.677
GLSZM	SmallAreaEmphasis	Wavelet HHH	0.652
GLDM	LDE	Wavelet HHH	0.648
			0.654

trend becomes even more pronounced when only the top five features are considered, as evidenced in Tables 3.7 and 3.8. Here, STSF achieves an accuracy of 0.654, while TSF achieves an accuracy of 0.635. In contrast, the performance obtained for KNN, when considering various distance metrics, falls significantly behind Rocket-based and TSF-based algorithms (refer to Supplemental Material, Table A.2).

Among the best-performing models, which include Rocket-based and TSF-based algorithms, features such as Total Energy and Icm2 stand out as highly predictive for each time series classifier. Energy features, in particular, demonstrate strong predictive power in three out of the four leading models.

Multivariate Analysis

The five top features discussed for the Rocket models were used for multivariate analysis *via* a voting mechanism. The meaning five features express:

- *Original First Order Energy*: is a metric that quantifies the magnitude of voxel values within an image. A higher value indicates a larger sum of the squares of these voxel values. Visually, the lesion with high Energy should exhibit a very bright appearance characterized by very high intensities.
- *Original First Order TotalEnergy*: is a metric derived by scaling the Energy feature with respect to the volume of the voxel in cubic millimeters.
- *Original ngtdm Strength*: exhibits a high value when an image demonstrates a slow transition in intensity, accompanied by more pronounced variations in gray-level intensities.
- *Wavelet - HLH glcm Imc2*: measures the complexity of the texture.
- *LoG - First Order 90Percentile*

The validation performance of the five best features resulting from the univariate analysis were reported in Table 3.9. Metrics were computed during the 20-repeated 10-fold CV and reported considering the mean and standard deviation.

Table 3.10 shows the validation performance for the multivariate time series analysis compared with the best instant for the instant-wise analysis. Performance on the test set were instead reported in Table 3.11 for both the multivariate time series analysis and the instant-wise analysis.

During the testing phase, the Multivariate Rocket model demonstrated superior performance in terms of accuracy, sensitivity, and NPV compared to the instant-wise model. Conversely, the instant-wise model exhibited a higher AUC-ROC, specificity,

Table 3.9: The validation performance of the top five features using the Rocket algorithm. (O: original, W: Wavelet, LoG2: LoG with $\sigma = 2$.)

Model	Accuracy	AUC-ROC	Specificity	Sensitivity
O FO Energy	0.717 ± 0.133	0.718 ± 0.133	0.719 ± 0.184	0.717 ± 0.196
O FO TotalEnergy	0.736 ± 0.109	0.736 ± 0.110	0.727 ± 0.160	0.745 ± 0.173
O NGTDM Strength	0.696 ± 0.111	0.696 ± 0.110	0.610 ± 0.169	0.782 ± 0.150
W HLH GLCM Imc2	0.728 ± 0.119	0.727 ± 0.120	0.666 ± 0.192	0.789 ± 0.160
LoG2 FO 90Percentile	0.671 ± 0.120	0.671 ± 0.121	0.681 ± 0.182	0.661 ± 0.172

Table 3.10: The validation performance of the multivariate time series classification using a voting mechanism and its comparison against the previous instant-wise analysis. (MR: Multivariate Rocket; I-W: Instant-wise)

Model	Accuracy	AUC-ROC	Specificity	Sensitivity	PPV	NPV
MR	0.742 ± 0.117	0.743 ± 0.118	0.710 ± 0.171	0.775 ± 0.156	0.742 ± 0.131	0.767 ± 0.138
I-W	0.710 ± 0.130	0.741 ± 0.135	0.738 ± 0.177	0.683 ± 0.178	0.743 ± 0.153	0.703 ± 0.145

and PPV. However, it's worth noting that the difference between specificity and sensitivity is significantly smaller for the Multivariate Rocket model, indicating a more balanced performance when compared to the instant-wise model. It is noteworthy that the time series approach yielded substantial improvements over the instant-wise approach, particularly evident in the results of the 20-repeated 10-fold cross-validation (as shown in Table 3.10). Given the relatively small dataset size, the cross-validation procedure provides a more precise evaluation of performance. In this context, the Multivariate Rocket model demonstrated higher accuracy and AUC-ROC, with slightly lower specificity but substantially higher sensitivity. The PPV values were comparable between both models, with Multivariate Rocket having a slightly higher PPV. Furthermore, the standard deviation for each metric was lower, indicating greater stability in performance metrics.

3.2.3 Discussion

In these works, the time series structure of the DCE-MRI acquisition was analyzed in two different ways: (1) the Instant-wise analysis to evaluate the best time instant of the series for classification, and (2) the Time Series Analysis to assume that the whole series in its entirety is more informative than individual time instants.

For the first, separate analysis for each instant within the DCE-MRI sequence was conducted. The Random Forest model displayed encouraging performance when trained on features extracted from the third post-contrast instant. This result can be explained

Table 3.11: The overall model performance achieved using the Rocket algorithm and its comparison against the previous instant-wise analysis.

Model	Accuracy	AUC-ROC	Specificity	Sensitivity	PPV	NPV
Multivariate Rocket	0.852	0.852	0.823	0.882	0.833	0.875
Instant-wise	0.823	0.877	0.882	0.764	0.866	0.789

by considering that, during the third post-contrast instant, the contrast agent is absorbed effectively by both malignant and benign lesions. This emphasizes their unique characteristics, which enhances the model's ability to distinguish between them. However, it's important to note that the instant-wise analysis doesn't fully harness the potential of the DCE-MRI sequence because it handles classification independently for each instant.

In this view, the second proposed method introduces several novelties and advantages. Firstly, an analysis of the DCE-MRI acquisition series using time series classification algorithms was conducted, considering all sequence instants simultaneously. This approach operates on the assumption that the entire series contains more valuable information than examining individual time instants separately. Moreover, this approach aligns closely with the diagnostic process employed by radiologists, who evaluate the complete sequence to make judgments regarding the benign or malignant nature of the lesion. The multi-protocol dataset closely replicates the complexities of the real clinical setting, thereby introducing challenges that enhance the validity of this research. To maintain data consistency, data harmonization was performed to align the distributions from the two distinct protocols. Furthermore, the multi-instant feature selection enabled the selection of informative features for time series classification, irrespective of the particular time instant within the sequence.

A comprehensive comparison of various time series classification algorithms was conducted, and the results revealed that Rocket and MultiRocket performed well in scenarios involving small datasets and short time series. Specifically, when considering the top five radiomic features for Rocket, the most favorable results were achieved by aggregating them using a voting mechanism. When evaluating the performance through a 20-repeated 10-fold cross-validation, the Multivariate Rocket model consistently outperformed the instant-wise model across all metrics, except for specificity and, to a slight extent, PPV, while exhibiting significantly lower standard deviation. This trend persisted during the test phase, where the Multivariate Rocket model demonstrated higher sensitivity but lower specificity when compared to the instant-wise model. However, when considering the balance between specificity and sensitivity, the Multivariate Rocket model emerged as a more balanced option than the instant-wise model. It's

worth noting that the features derived from the Laplacian of Gaussian (LoG) were not as influential as the Original and Wavelet-derived features in achieving these results.

Literature Comparison

Although using a two-protocol dataset certainly increases the complexity of the classification task, the performance obtained are still superior or in line with the state-of-the-art.

A high specificity (97-100%) and a low sensitivity (56-67%) were obtained by Gibbs *et al.* [178], extracting radiomic features from three parameter maps and using a support vector machine as a classifier. Similarly, results were obtained by Zhang *et al.* [115] using a support vector machine. In particular, a specificity of 0.800 and sensitivity of 0.714 were obtained considering only the pharmacokinetic parameters maps. Also in Militello *et al.* [101] a higher specificity (0.741 ± 0.114) with respect to sensitivity (0.709 ± 0.176) was achieved using a support vector machine. In Zhou *et al.* [179] an opposite trend was computed: focusing on radiomic analysis, a sensitivity of 85% with respect to a specificity of 65%. Furthermore, in Parekh *et al.* [180], several image sequences were involved in the radiomic analysis. In particular, a higher sensitivity (0.93) was observed in comparison to specificity (0.85).

Compared with the validation performance obtained in the instant-wise analysis, only the specificity resulted in slightly lower (0.710 ± 00171 vs. 0.738 ± 0.177). The sensitivity resulted significantly higher (0.775 ± 0.156 vs. 0.683 ± 0.178). It is possible to extend the same consideration to the test set. Except compared with the Parekh *et al.* [180], in which features were extracted from several sequences (DCE-MRI, DCE High Spatial Resolution, DWI, ADC map, T1, and T2), the achieved performance results in line or higher. In fact, a more balanced specificity and sensitivity were calculated, meaning fair benign and malignant classification rates.

Furthermore, when compared to clinical diagnostic performance, the Multivariate Rocket classifier demonstrates a similar trend in sensitivity and specificity. In particular, was proved that MRI provided an overall sensitivity and specificity of 94.6% and 74.2%, respectively, while for the contrast-enhanced MRI, overall sensitivity and specificity were 91.5% and 64.7% [206]. Focusing on DCE-MRI, a sensitivity of 93.2% and a specificity of 71.1% was computed by Zhang *et al.* [207], while 0.74 of specificity and 0.87 of sensitivity by Dong *et al.* [208]. In this view, the trained model is coherent with the clinical diagnostic performance, showing a slightly lower sensitivity and higher specificity.

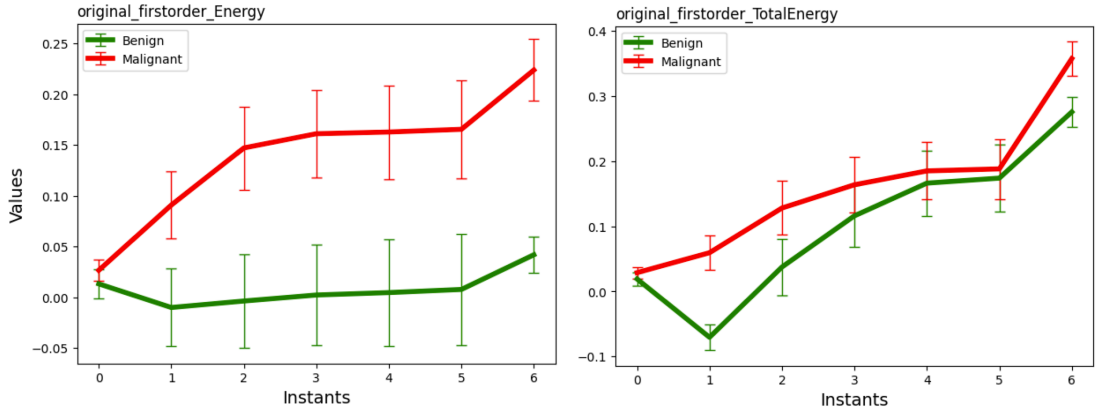


Figure 3.7: Average trend of Feature Energy and Total Energy for the 81 benign lesions (green) and 85 malignant lesions (red).

Model Interpretation and Clinical Validation

The primary benefit of extracting radiomic features is the comprehensibility of these features. Radiomic feature extraction enables model introspection while preserving a high classification accuracy. In this specific context, model introspection enables us to compare the model's results with the actual physician diagnosis process.

The key discoveries revolve around features that are closely associated with alterations in the intensity of gray levels in images, namely Energy and Total Energy features. In Figure 3.2, it is visually apparent that the initial consequence of contrast agent administration is the elevation of gray level (signal) intensities. This increase aligns precisely with what the Energy and Total Energy features describe. Figure 3.7 illustrates the average trend of these two features and clearly demonstrates this phenomenon. Specifically, the malignant lesion displays a swift rise in energy and, consequently, signal intensity, which persists until the final moment of the DCE-MRI sequence. The quantitative explanation of these model findings paves the way for clinical validation. In particular, the much more rapid initial growth observed in malignant lesions as opposed to benign ones suggests that the contrast agent is absorbed more swiftly in malignant lesions, as previously documented in Padhani *et al.*'s study [169]. Furthermore, Figure 3.7 highlights that the most significant disparities between the two trends are concentrated in the initial moments. It is noteworthy that peak enhancement usually occurs within the first 2 minutes following the injection of the contrast agent [209]. Additionally, the elevated NGTDM Strength values suggest that the ROIs undergo a gradual shift in intensity, characterized by a prevalence of pronounced and coarse fluctuations in gray-level intensities. In this specific case, as depicted in Figure 3.8, the most substantial difference is observed in the final time frame of the sequence, where benign lesions exhibit higher Strength values. This implies that when the contrast agent is absorbed

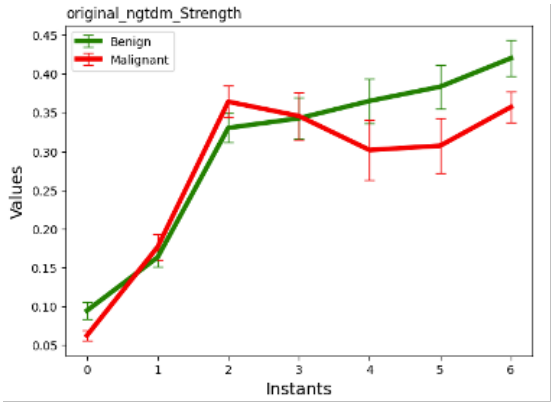


Figure 3.8: Average trend of the NGTDM Strength feature for benign lesions (green) and malignant lesions (red).

by both types of lesions, benign lesions display a more consistent pattern with fewer rapid intensity fluctuations. In fact, previous research has demonstrated that malignant lesions typically exhibit heterogeneous internal enhancement during the delayed phase[210, 211].

The explanation of features becomes more complex when we delve into high-level features such as those derived from wavelet transformations. This complexity arises because the clinical interpretation of results is typically conducted by physicians using the original images, which significantly differ from the transformed images (e.g., those subjected to Wavelet Transformation or LoG filtering). Consequently, establishing a direct association between radiomic features and clinical findings can be challenging and may not be equitable. Nonetheless, from a quantitative perspective, the other series of features display distinct patterns between benign and malignant tumors, offering valuable insights for distinguishing between the two (See Figure A.1 of Supplemental Materials for the average trend of the other selected radiomic sequences).

3.3 Breast Microcalcification Detection and Classification in Mammogram using an Interpretable Radiomic Signature

Breast calcifications refer to small deposits of calcium salts, typically measuring less than 1 mm in diameter [212], which appear radiopaque on mammograms. While they are generally common and often non-cancerous, they can serve as an early indication of breast cancer when observed on mammograms. Approximately one-third of all malignant lesions detected during screening mammography include breast calcifications [213, 214]. Mammography exclusively identifies around 50% of non-palpable breast cancers and nearly 95% of cases of ductal carcinoma-in-situ (DCIS) through microcalcification patterns [215, 216]. Furthermore, a comprehensive meta-analysis conducted by Brennan *et al.* [217] revealed that although other mammographic abnormalities such as masses, architectural distortion, asymmetry, palpability of the lesion, and lesion size strongly correlate with the upstaging of DCIS, cases of DCIS presenting solely as calcifications can also hide invasive disease. The classification of breast microcalcifications can vary based on their size, shape, extent, density, and distribution pattern as observed on mammograms [218]. In clinical practice, calcification diagnosis is primarily based on radiologists' evaluation of their morphology and distribution, following the guidelines outlined in the BI-RADS Atlas [167]. It's important to note that the rates of false-positive biopsy recommendations for calcifications can be quite high, ranging from 30% to 87% [219, 220]. Additionally, localizing calcifications can be more challenging in mammographic images with low contrast and in dense breast tissues [221]. Consequently, the sensitivity of screening for detecting malignant calcifications remains relatively low. Many detectable calcifications are not immediately identified for further investigation and are only noticed during subsequent screening rounds, often when the disease has already advanced to an invasive stage [222]. To address this situation, it is possible to improve the diagnostic process for physicians by incorporating a quantitative approach.

As discussed in Section 2.3.2, Radiomics provides a quantitative viewpoint complementary to the radiologist. Among all the benefits already discussed, Radiomics has additional strengths in this work. It is possible to extract radiomic features from ROIs at the original spatial resolution, avoiding any image resizing as is the case of deep feature extraction (e.g., *via* neural networks). Especially in the case of microcalcifications, in which the ROI sizes are about 1mm [212] (e.g., a few pixels), the resizing can greatly reduce the information content. Furthermore, as previously highlighted, since it is well known the meaning each radiomic feature expresses, it becomes feasible to interpret the findings of the trained models and derive crucial clinical insights. This

interpretability stands as a fundamental prerequisite for instilling trust and validating the performance of these trained systems [223, 11].

Radiomics workflow have also been applied to the analysis of microcalcifications in breast imaging. Lei *et al.*[224] conducted research on predicting benign BI-RADS 4 calcifications using radiomic features, and they developed a nomogram that incorporated these features along with the menopausal state as predictive factors. Similarly, Stelzer *et al.* [225] focused on the classification of BI-RADS 4 microcalcifications, exploring radiomic approaches. Marathe *et al.* [226] presented a quantitative approach for classifying amorphous calcifications into benign and actionable (high-risk and malignant) categories using radiomic techniques. Loizidou *et al.* [227] used a proprietary dataset comprising two sequential screening mammogram rounds. They employed temporal subtraction between recent and prior mammograms to distinguish between healthy tissue and microcalcifications, as well as to differentiate benign from suspicious microcalcifications. In Fanizzi *et al.* [228], both radiomic and wavelet features were employed for classification tasks, including distinguishing between normal and abnormal cases and classifying calcifications as benign or malignant.

As it emerged, it is a common practice to divide the analysis of microcalcifications into two distinct tasks: detection and classification. The detection task is focused on distinguishing microcalcifications from healthy breast tissue. In contrast, the classification task assumes that microcalcifications have already been detected and involves differentiating between malignant and benign ones. Due to the small size of microcalcifications, the detection process is highly sensitive and can be influenced by factors such as human perception, breast density, and the characteristics of the cancer itself [166]. Radiomics offers a quantitative perspective in addition to the visual assessment conducted by physicians. This quantitative approach can effectively support and enhance the diagnostic process, providing valuable assistance in both microcalcification detection and classification tasks.

In this study, a radiomic signature was developed for training machine learning models to detect and classify breast microcalcifications [160]. In this research, a proprietary dataset was collected at the Radiology section of the University Hospital "Paolo Giaccone" in Palermo, Italy. Figure 3.9 illustrates the general workflow of the study. The dataset was segmented into three categories: healthy tissue, benign microcalcifications, and malignant microcalcifications. The same training pipeline was applied to address two specific tasks: Task 1, which involved distinguishing between benign and malignant microcalcifications, and Task 2, which focused on discriminating healthy tissue from microcalcifications. Following the feature extraction process, the SMOTE method was applied to balance the data by oversampling the benign microcalcification samples.

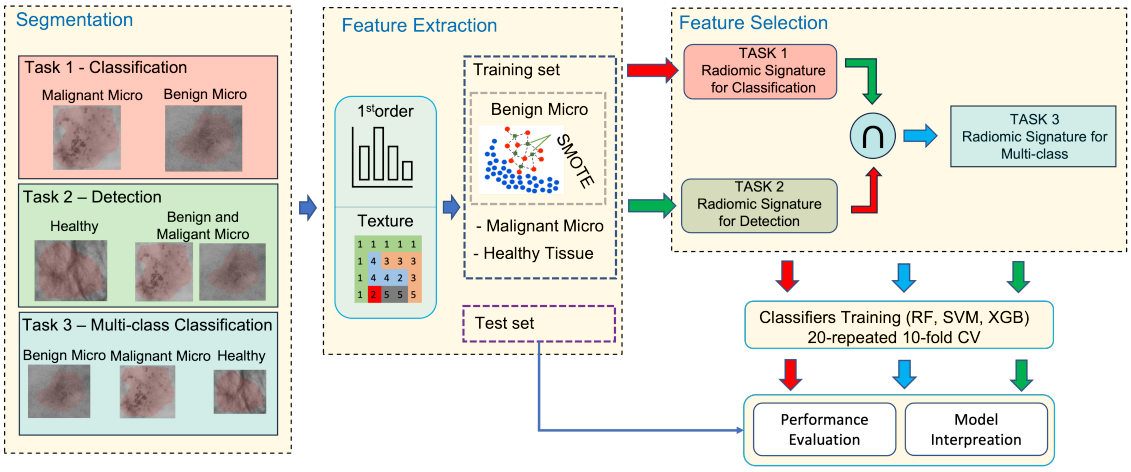


Figure 3.9: Overall architecture for breast microcalcification classification and detection.

Several feature selection steps were employed to identify the most informative features for both tasks. The intersection of the selected features from Task 1 and Task 2 was used to train a multi-class model, enabling simultaneous differentiation between healthy tissue, benign microcalcifications, and malignant microcalcifications (Task 3). Three different shallow learning algorithms — SVM, RF, and XGB — were compared for both detection and classification tasks. To evaluate the models, a 20-repeated 10-fold cross-validation strategy was employed. Finally, the performance of the trained models was assessed on the test set, and their interpretability was analyzed. In summary, the study proposed a radiomic signature capable of distinguishing between healthy breast tissue and both benign and malignant microcalcifications, demonstrating its potential for improving the detection and classification of these crucial indicators in breast imaging.

The study presents the following key components:

- A well-structured processing pipeline [150], to establish a meaningful radiomic signature for breast calcifications.
- Development of a multi-class model that has the capability to differentiate between healthy breast tissue, benign microcalcifications, and malignant microcalcifications.
- an interpretation of the more informative radiomic features to provide a trusted system supporting the decision-making processes.

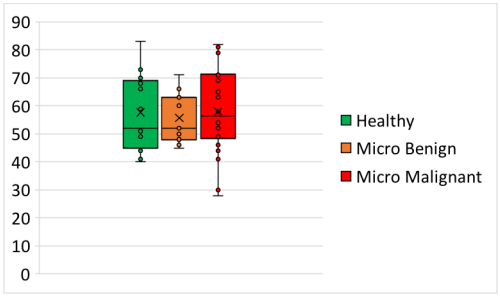


Figure 3.10: Patients age comparison among the three groups.

3.3.1 Materials and Methods

3.3.1.1 Dataset Description and Segmentation

The dataset comprises a total of 161 images acquired using a Fujifilm Full Field Digital Mammography system at the Radiology section of the University Hospital "Paolo Giaccone" in Palermo, Italy. These images have a spatial resolution of 4728×5928 pixels and a pixel size of 50 µm.

The dataset was categorized into three groups:

- Healthy patients: 76 images. Age: 57.6 ± 12.7 years, with an age range of $40 - 83$ years
- Patients with benign microcalcifications: 26 images. Age: 55.7 ± 8.6 years, with an age range of $45 - 71$ years.
- Patients with malignant microcalcifications: 59 images. Age: 58.0 ± 14.4 years, with an age range of $28 - 82$ years.

Figure 3.10 presents box plots comparing the age distributions among these groups.

The ITK-SNAP toolkit was used for ROIs segmention. Healthy ROIs were chosen randomly and manually segmented, while for microcalcification images, manual segmentation was employed to identify clusters of neighboring microcalcifications. In total, 380 segmentations of healthy tissue, 136 benign microcalcifications, and 242 malignant microcalcifications were obtained. An expert radiologist conducted the annotations to identify abnormal regions.

Three distinct tasks were undertaken.

- First Task: involved detecting benign and malignant microcalcifications versus healthy tissue, with 378 samples in one group and 380 in the other.
- Second Task: focused on classifying benign versus malignant microcalcifications,

with 136 samples in the benign category and 242 in the malignant category.

- Third Task: regards multi-class classification, considering benign, malignant microcalcifications, and healthy tissue, with 136, 242, and 380 samples, respectively.

3.3.1.2 Radiomic Feature Extraction

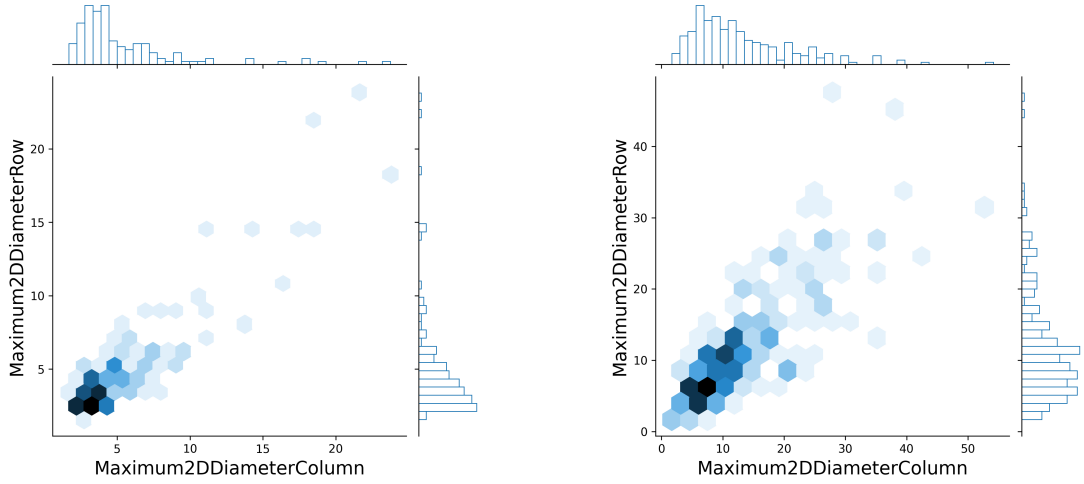
In this work, 93 radiomic features were extracted, following the IBSI [84] and using the PyRadiomics toolkit [85].

An image gray level discretization bin-width of 25 was chosen. This bin-width was determined by considering the average gray level range of 5419, which is calculated as the difference between the maximum and minimum gray levels. With this bin-width, approximately 216 bins can be created in the histogram using the formula $\frac{mean-range}{bin-width}$. Values of around 256 bins are frequently used [229].

The extracted features belong to intensity (or first-order (FO)) and textural features, as discussed in Section 2.3.2. In particular, first-order features were extracted, as well as texture features computed from the following matrices: GLCM, GLRLM, NGTDM, GLSZM and GLDM.

There are several reasons why 2D shape features were not included in the analysis:

- The primary goal was to create a signature independent of the specific segmentation generated. Instead, the focus was on features related to texture and gray level intensity, ensuring the signature's robustness across different segmentation methods.
- As shown in Figures 3.11a and 3.11b, malignant microcalcifications tend to have larger segmentations compared to benign ones on average. Including shape features could introduce a significant bias in the models, potentially leading to discrimination based solely on shape characteristics rather than considering the more relevant texture and gray level intensity information.
- The segmentations were intentionally kept coarse since the primary objective was to detect and classify clusters of microcalcifications, not individual microcalcifications. In such cases, fine-grained shape features may not provide additional discriminatory power and might even introduce noise into the analysis.



(a) Sizes of benign microcalcifications.

(b) Sizes of malignant microcalcifications.

Figure 3.11: Microcalcifications size representation. Maximum 2D diameter Row (Column) is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the column-slice (row-slice). These magnitudes represent the size width and height of lesions.

3.3.1.3 Feature Selection

Two distinct feature subsets were selected for the detection and classification tasks. This selection process involved various analytical techniques, including variance analysis, correlation analysis, and the assessment of statistical significance. The primary objective was to identify a subset of radiomic features that are both informative and non-redundant, ensuring that the selected features contribute meaningfully to the tasks and do not introduce unnecessary complexity or collinearity into the models [150] (discussed in detail in Section 2.4.1.1)

A threshold of 0.01 was considered for variance analysis. A $|\rho| < 0.85$ was considered to discard correlate features (Spearman's rank correlation coefficient) [114, 113]. The Mann-Whitney U test was used to test the class differences (healthy tissue vs. microcalcifications and benign vs. malignant microcalcifications). A $p < 0.05$ was considered statistically significant.

Ultimately, the SFFS algorithm [111], was implemented for the purpose of selecting the most suitable subset of features for each model (RF, SVM, and XGB). The application of SFFS was carried out separately for the tasks of detection and classification. Specifically, the remained features after the initial analyses, including variance assessment, correlation examination, and statistical significance testing, were used as input

for SFFS. The models used for the SFFS procedure were trained in a 10-fold cross-validation strategy. Accuracy was the metric to maximize. For training the multi-class model, which regards both detection and classification tasks simultaneously, the approach involved the common subset of features that had been separately identified for the two tasks.

3.3.1.4 Model Training and Test

Accurate extraction of radiomic features proves to be effective in situations where data is limited, which stands in contrast to the data-intensive nature of deep learning methods [100]. Furthermore, radiomic features offer a viable way to take advantage of shallow training methods when working with tabular data and small dataset [120, 121, 122]. In this particular study, three distinct classifiers were employed: SVM, RF, and XGB.

The process of feature selection and model training was conducted independently for both the detection and classification tasks. Consequently, it was feasible to treat both tasks as binary classifications. Prior to initiating the feature selection and training phases for all three tasks, the dataset was split into two parts: 80% of the data was reserved for feature selection and training purposes, while the remaining 20% was exclusively allocated for testing. Furthermore, taking into account the class imbalance observed between benign and malignant microcalcifications, the SMOTE method was applied [119] to the training set to balance the two classes: the minority class (benign) was oversampled by adding synthetic data, thereby balancing it with the majority class (malignant). Importantly, SMOTE was not applied to the test set. Given the limited dataset size, a robust evaluation approach was adopted, involving 20 repetitions of stratified 10-fold cross-validation. This approach was employed to assess the validation performance effectively. As a result, the validation performance metrics were reported by considering both the mean and standard deviation for each metric. The model that demonstrated the highest accuracy during the validation phase was selected for further testing.

Moreover, the features that were common between those selected for the detection and classification tasks were used to train the multi-class model. The same training and testing procedure was applied in this case to ensure consistency and comparability with the binary tasks.

Table 3.12: Selected features for detection and classification tasks, before applying SFFS.

Feature	Class	Det	Clas
10Percentile	FO	X	X
90Percentile	FO	X	X
Energy	FO		X
Entropy	FO	X	X
Kurtosis	FO		X
Maximum	FO	X	X
Minimum	FO	X	X
Skewness	FO	X	X
Autocorrelation	GLCM		X
Contrast	GLCM	X	X
DependenceVariance	GLDM		X
LargeAreaLowGrayLevelEmphasis	GLSZM	X	X
Busyness	NGTDM		X
Contrast	NGTDM	X	

3.3.2 Results

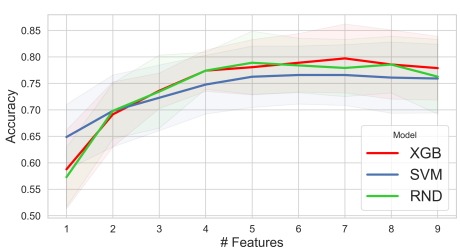
To comprehensively assess model performance, a range of evaluation metrics was employed, including Accuracy, Area Under the Receiver Operating Characteristic (AUC-ROC), Specificity, Sensitivity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). To ensure a fair and consistent comparison among the trained models, the same random seed was set for all probabilistic terms within the algorithms and for generating the splits during the stratified cross-validation.

The experiments were conducted within a Python 3.7 environment. For RF, the training involved the bootstrap technique, 100 estimators, and the Gini criterion. XGB was set with 100 estimators, a maximum depth of 6, 'gain' as the importance type, a binary logistic loss function, and a learning rate of 0.3. SVM employed the Radial basis function as the kernel, a regularization parameter of $C = 1.0$, and a kernel coefficient calculated as $1/(n_{features} * variance)$. Prior to SVM training, feature standardization was applied.

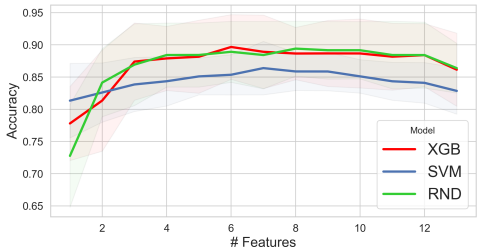
Furthermore, for multi-class training, the one-vs-rest strategy was employed for SVM, and the softmax loss function was used for XGB.

3.3.2.1 Features Selected

Table 3.12 displays the features selected for both tasks following the initial steps of variance analysis, correlation analysis, and the statistical test. In particular, there was a significant overlap between the two feature subsets. For both the detection and classification tasks, each model (e.g., SVM, XGB, RF) was trained using the same



(a) SFFS in detection task. On average, 7 is the features number that maximizes the accuracy of the three models.



(b) SFFS in classification task. On average, 9 is the features number that maximizes the accuracy of the three models.

Figure 3.12: The graph generated *via* SFFS shows the accuracy value for each model (XGB, SVM, and RND) considering several features subset. The x-axis is represented the $n-th$ step of the algorithm; the y-axis is instead shown the accuracy value.

number of features. These feature counts were determined through SFFS, selecting the smallest radiomic signature that yielded the highest accuracy. Specifically, Figures 3.12a and 3.12b present accuracy results considering the different subsets chosen *via* SFFS for the detection and classification tasks, respectively. As shown in Figure 3.12a, on average, a signature size of seven features maximizes accuracy across all three models in the detection task. Conversely, Figure 3.12b illustrates that a set of nine features optimizes accuracy for the classification task. As a result, seven and nine features were chosen for training in the detection and classification tasks, respectively. In the detection task, the NGTDM Contrast feature was the first one selected *via* SFFS for each considered model. However, NGTDM Contrast did not prove to be statistically significant for the classification task. In contrast, for the classification task, the FO Entropy feature was the first feature selected *via* SFFS for each model. Additionally, FO Entropy, GLCM Contrast, and GLSZM LargeAreaLowGrayLevelEmphasis were consistently among the most frequently selected features *via* SFFS (selected in at least 5 out of the 6 models considered)

Taking into account the features that were statistically significant ($p < 0.05$) for both the detection and classification tasks, a common subset of 8 features was identified, as presented in Table 3.12. This common subset of features was used to tackle the two tasks simultaneously, addressing a multi-class problem with three classes: healthy tissue, benign microcalcifications, and malignant microcalcifications. For this reason, the results section is structured to present the outcomes of these three tasks separately, reflecting the performance and findings associated with each specific task.

3.3.2.2 Performance of the three Tasks

The performance evaluation during feature selection *via* SFFS was conducted using a 10-fold stratified cross-validation approach (refer to Figures 3.12a, 3.12b). The cross-validation process was repeated only once due to the computational complexity of the SFFS algorithm. Conversely, for the model training phase, a 10-fold cross-validation was repeated 20 times to ensure a more precise assessment of the models, as shown in Figures 3.13 and 3.14. This repeated cross-validation approach provides a robust estimate of model performance. Ultimately, the most accurate model identified during the validation phase was chosen for testing on an independent test dataset, as presented in Tables 3.13, 3.14, and 3.15. This procedure helps verify the model's generalization and performance in real scenarios.

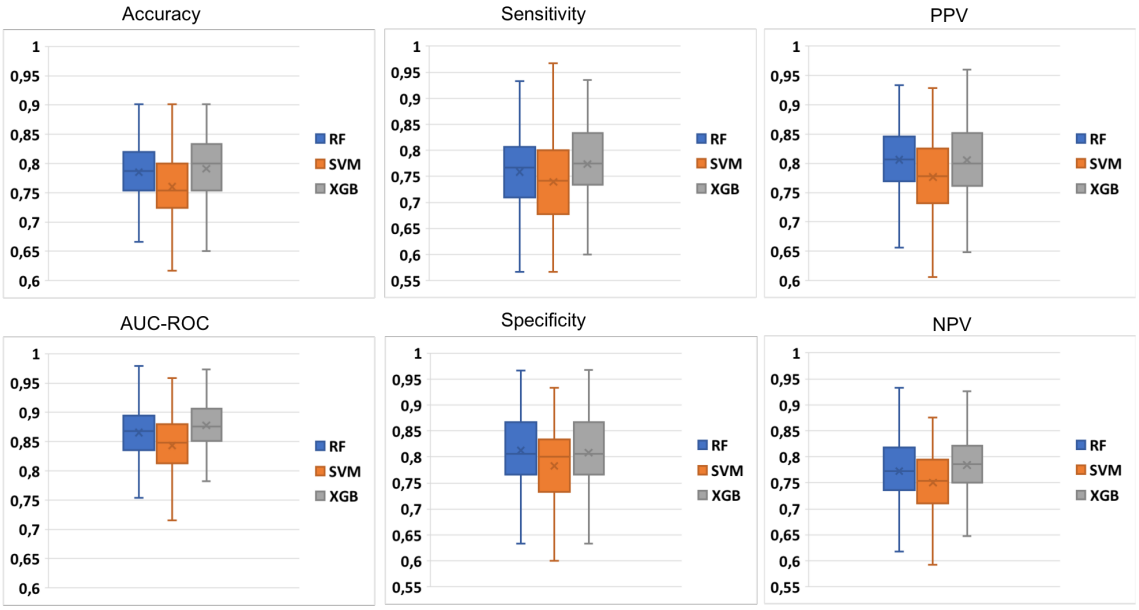


Figure 3.13: Validation performance for the detection task computed during the 20-repeated 10-fold cross-validation procedure.

Detection Performance

The primary objective of this task was to classify healthy tissue from microcalcifications. In the training set, there were 306 samples of healthy tissue and 302 samples of microcalcifications, while the test set comprised 78 microcalcifications and 74 healthy tissue samples. Figure 3.13 provides an overview of the validation performance calculated during the 20 repetitions of the 10-fold cross-validation process. It's evident that XGB achieved a higher performance during the validation phase, nearly comparable to RF. Across all models, a notable trend was observed wherein specificity was higher than sensitivity. This indicates that the models exhibited a stronger ability to correctly

identify healthy tissue samples as compared to their ability to detect microcalcifications.

Table 3.13 presents the metrics computed during the test phase. While SVM demonstrated lower performance compared to XGB and RF in the validation phase, it exhibited superior generalization capabilities when applied to unseen data. Specifically, SVM achieved an AUC-ROC of 0.865 in the test phase. RF and XGB also achieved promising AUC-ROC performance, with values of 0.859 and 0.854, respectively. However, it's important to highlight that there was a significant imbalance between sensitivity and specificity, with a higher specificity than sensitivity.

Table 3.13: Test performance for the detection task.

Metric	RF	SVM	XGB
Accuracy	0.756	0.789	0.750
AUC-ROC	0.859	0.865	0.854
Sensitivity	0.729	0.783	0.702
Specificity	0.782	0.794	0.794
PPV	0.760	0.783	0.764
NPV	0.753	0.794	0.738

Table 3.14: Test performance for the classification task.

Metric	RF	SVM	XGB
Accuracy	0.868	0.868	0.842
AUC-ROC	0.921	0.927	0.933
Sensitivity	0.931	0.863	0.909
Specificity	0.781	0.875	0.750
PPV	0.854	0.904	0.833
NPV	0.892	0.823	0.857

Classification Performance

The objective of this task was to classify benign and malignant microcalcifications. In the training set, there were 198 malignant microcalcifications and 198 benign microcalcifications, considering both real samples and synthetic samples generated *via* SMOTE for the benign class. The test set comprised 44 malignant microcalcifications and 32 benign microcalcifications.

Figure 3.14 illustrates the validation performance, while the achieved performance in the test phase is presented in Table 3.14. Similar to the detection task, SVM exhibited lower performance compared to XGB and RF during the validation phase. However, in the test phase, decision tree-based models (RF and XGB) performed less effectively than SVM. Once again, there was a noticeable imbalance between sensitivity and specificity. Nevertheless, it's essential to note that the models achieved very high overall performance, with AUC-ROC values of 0.921, 0.927, and 0.933 for RF, SVM, and XGB, respectively. For decision tree-based models (RF and XGB), sensitivity was higher than specificity, indicating their greater ability to correctly identify malignant microcalcifications as opposed to benign ones.

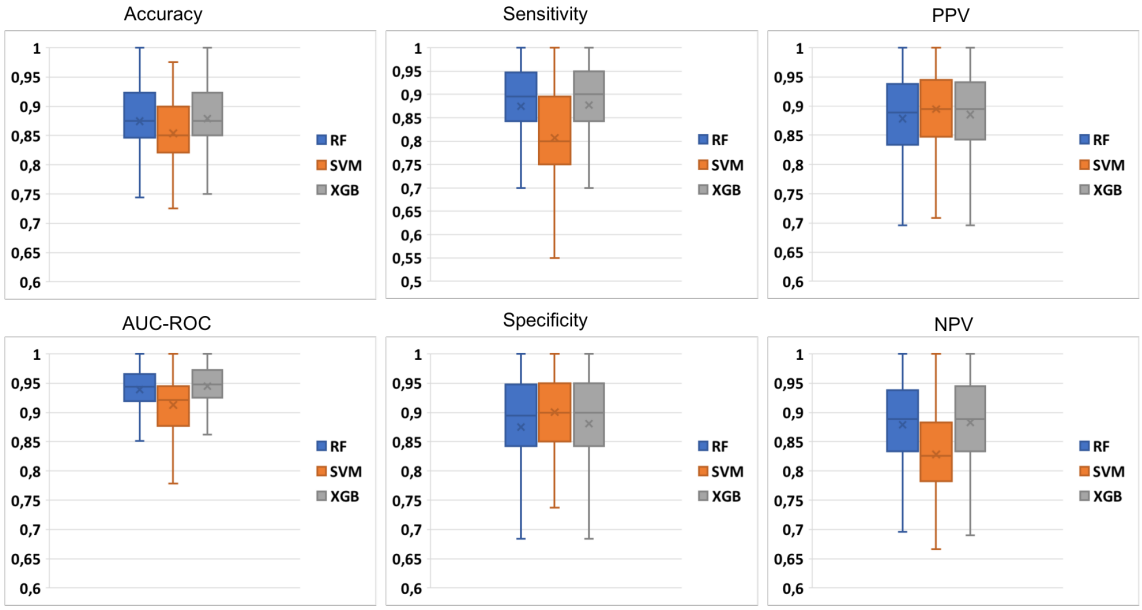


Figure 3.14: Validation performance for the classification task computed during the 20-repeated 10-fold cross-validation procedure.

Multi-class Model Performance

Leveraging the common set of discriminating features identified for both the detection and classification tasks (as shown in Table 3.12), SVM, RF, and XGB were trained for the multi-class classification task. SVM employed the one-vs-rest strategy, while XGB used the softmax loss function. In this multi-class classification scenario, the training set consisted of 198 malignant microcalcifications, 198 benign microcalcifications (comprising 104 real samples and 97 generated *via* SMOTE), and 198 healthy tissue samples. The 198 healthy samples were randomly selected from the original pool of 380 to ensure class balance during training. For the test set, 78 healthy tissue samples, 44 benign microcalcifications, and 32 malignant microcalcifications were used.

In Table 3.15, the test set performance is presented. Healthy tissue classification exhibits a high specificity but a low sensitivity, indicating that the model excels at detecting microcalcifications. This trend is also observed in the case of benign microcalcifications, where the model's performance is better in detecting malignant microcalcifications and healthy tissue. Consequently, detecting malignant microcalcifications is relatively straightforward in all scenarios. In particular, the decision tree-based models outperform the SVM classifiers across the board. They achieve higher AUC-ROC and accuracy scores when classifying the three different classes. This suggests that the tree-based models are better suited for multi-class classification tasks.

Table 3.15: Multi-class classification test performance, for simultaneous detection and classification task.

Model	Class	Acc	AUC	Sens	Spec	PPV	NPV
RF	Healthy	0.746	0.810	0.679	0.815	0.791	0.712
	B Micro	0.818	0.860	0.593	0.877	0.558	0.891
	M Micro	0.811	0.890	0.772	0.827	0.641	0.900
SVM	Healthy	0.694	0.783	0.538	0.855	0.792	0.643
	B Micro	0.792	0.849	0.687	0.819	0.500	0.909
	M Micro	0.824	0.840	0.818	0.649	0.927	0.882
XGB	Healthy	0.740	0.830	0.679	0.802	0.779	0.709
	B Micro	0.811	0.856	0.625	0.860	0.540	0.897
	M Micro	0.824	0.876	0.750	0.854	0.673	0.895

3.3.3 Discussion

The study aimed to address the challenge of diagnosing breast microcalcifications by developing a data-driven system to aid physicians in their diagnostic process. This system leveraged the radiomic workflow to convert medical images into highly informative features, providing a quantitative perspective that complements the visual assessments made by doctors. Given the difficulty of diagnosing microcalcifications due to their minute size, data-driven systems can play a pivotal role in improving accuracy. Microcalcifications often progress into invasive lesions, underscoring the importance of early detection to prevent advanced disease stages and facilitate appropriate treatment. In this context, the combination of the radiomic workflow with shallow learning techniques can support physicians in their diagnostic efforts and also enable the interpretation of features and the creation of explainable models. The development of explainable models is essential for model validation and for comparing research findings with existing medical literature [123]. Furthermore, explainability enhances the usability and acceptance of AI models [223]. In various complex decision-based tasks, the interpretability of AI-based systems can become a crucial feature [11]. The work presented in this study yielded significant results, both in terms of predictive performance and in the insights gained through the interpretability of radiomic features.

Model performance and findings

In terms of performance evaluation, the detection results are indeed promising. The models achieved AUC-ROC scores of 0.859, 0.856, and 0.854 for Random Forest, Support Vector Machine, and XGBoost, respectively. In particular, the performance improves when focusing solely on the classification of malignant versus benign microcalcifications. In this specific scenario, the AUC-ROC scores rise to 0.921, 0.927, and

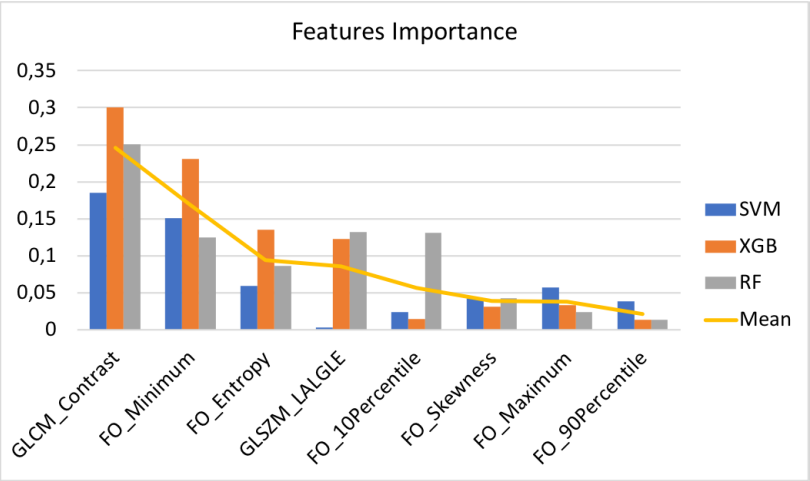


Figure 3.15: Features importance computed *via* the Mean Score Decrease method.

0.933 for RF, SVM, and XGB, respectively. This result is particularly significant because it demonstrates the system's capability to detect microcalcifications that have the potential to develop into invasive cancers. The disparity in performance between the two tasks reaffirms that the primary challenge in microcalcification analysis lies in the detection phase. This underscores the critical role of accurate detection in early diagnosis during screening, as it is the pivotal task in identifying potentially harmful lesions.

One of the key findings of this study is the identification of a shared radiomic signature, and its relate importance, between the detection and classification tasks, as illustrated in Figure 3.15. This signature was determined using the Mean Score Decrease method available in the ELI5 framework [230]. In particular, three features stand out as the most important: GLCM Contrast, FO Entropy, and FO Minimum. The GLCM Contrast is a measure of the local intensity variation, so a larger value correlates with a greater disparity in intensity values among neighboring pixels. A higher Contrast was foundin healthy tissue with respect to microcalcification. A higher Minimum was found for the healthy tissue with respect to microcalcification: this is intuitive because the microcalcification intensity is much lower compared with healthy tissue. Finally, a higher Entropy was found in microcalcifications compared with the healthy tissue. With the Entropy is possible to measure the uncertainty/randomness in the image values.

Unlike deep learning architectures that produce latent spaces with limited comparability and reproducibility across studies, the radiomic approach allows for the meaningful integration of significant features from different research works. This is because each radiomic feature has a well-defined meaning and interpretation, in contrast to deep features. As a result, significant overlap with other studies was found. For example,

Entropy and Minimum have been identified as important features in studies related to PET and MRI for breast cancer phenotypes and prognosis [231]. Entropy has also been found to be significant in multiparametric MRI for breast cancer tissue characterization [180, 232], and the GLCM Contrast has been highlighted in similar contexts [180]. Furthermore, the Minimum feature has been recognized as relevant in DCE-MRI for predicting Sentinel Lymph Node Metastasis [233].

Comparison

Several papers addressed the microcalcification analysis through Radiomics. Although the following works use different datasets, a qualitative comparison can be made. In particular, Stelzer *et al.* [225] have focused only on BI-RADS 4 microcalcification, analyzing a dataset consisting of 150 benign and 76 malignant microcalcifications. They exploited the radiomic workflow for classification, in an attempt to avoid unnecessary benign biopsies. To the extracted features, the principal component analysis (PCA) was applied and a multilayer perceptron was trained. They obtained an AUC-ROC of 0.82-0.83, and found also the GLCM Contrast the most important feature contributing to PCA. Lei *et al.* [224] focused also on BI-RADS 4 calcifications to discriminate benign from malignant calcifications. They selected 6 radiomic features and uses also the menopausal state to train an SVM model, reaching an AUC-ROC of 0.80, a PPV of 73.53 and NPV of 84.21. Marathe *et al.* [226] analyzed 276 amorphous calcifications (200 benign and 76 malignant). They extracted the radiomic features from the foreground and background masks, and global features from dilated foreground masks. Using the LightGBM classifier they obtained an AUC-ROC of 0.73, a sensitivity of 1.0 and a specificity of 0.35. In addition, they proved that in small dataset scenario, local and global radiomic features allows higher performance with respect to VGG-16 and ResNet-50 deep architecture. In Fanizzi *et al.* [228] the healthy ROIs were considered to train two different classifiers normal vs. abnormal and benign vs. malignant. From the Breast Cancer Digital Repository [234] 130 microcalcifications (75 benign and 55 malignant) and 130 healthy ROIs were selected. They used the wavelet Haar transform before the feature extraction process. The selected features were used to train the random forest model, obtaining a median AUC-ROC value of 98.16% and 92.08% for the detection and classification tasks, respectively. As discussed, an opposite trend was found: the classification model performed better than the detection model. Loizidou *et al.* [227] acquired a proprietary dataset considering two sequential screening mammogram rounds, to distinguish between normal tissue vs. microcalcifications, and benign vs. suspicious microcalcifications. For the two tasks, radiomic features from the recent mammogram (RM) and from the temporal subtracted (TS) mammograms were extracted. Then, several machine learning classifiers were compared, considering the

3.5 YOLO-based Model for Breast Cancer Detection enhanced by Explainable methods.

In breast bancer diagnosis process, the primary physician objective is to identify all ROIs within the entire mammogram, including masses, calcifications, distortions, and more. Detecting these abnormalities at an early stage is of utmost importance for planning subsequent examinations, treatment strategies, or interventions. Failure to detect these abnormalities can lead to irreversible damage to the patient. Consequently, breast cancer detection represents one of the most complex and crucial tasks in breast cancer. Regrettably, many solutions proposed in the literature do not aim to comprehensively analyze the entire mammogram. Instead, they limit the detection process to patch classification: manually selecting and cropping ROIs and then training classifiers to distinguish these cropped regions. However, to truly support and replicate the diagnostic process followed by physicians, an architecture capable of autonomously detecting all ROIs across the entire mammogram is required.

The advent of RetinaNet, Faster R-CNN, and YOLO has spurred the advancement of systems for breast cancer detection [238, 239, 240, 241]. These frameworks have undeniably contributed significantly to the field. However, they bring about two primary challenges:

1. **Learning Whole Mammogram Features**: These models need to learn the features of the whole mammogram. During training, resizing the images to fit the model's input requirements can lead to the loss of critical details, potentially impacting their ability to detect abnormalities accurately.

2. **Increased Error Rates**: As the model is trained to detect all ROIs within all healthy tissue patches (non-ROIs), it inevitably faces an increase in the error rate due to the challenging nature of this task.

Nevertheless, it's worth noting that YOLO has demonstrated its effectiveness across various scenarios, often outperforming its competitors in terms of accuracy and inference speed [242].

In this study, a breast cancer detection model based on YOLOv5 was introduced, designed to assist physicians in their diagnostic procedures. A comprehensive comparison, evaluating various feature extractors, including different versions of YOLOv5 (nano, small, medium, and large), Darknet53 as proposed in YOLOv3 [243], and the Vision Transformer [62] was conducted. Additionally, to evaluate the performance of these models, a proprietary dataset was acquired and meticulously annotated. This dataset was collected at the Radiology section of the University Hospital "Paolo Giaccone" in

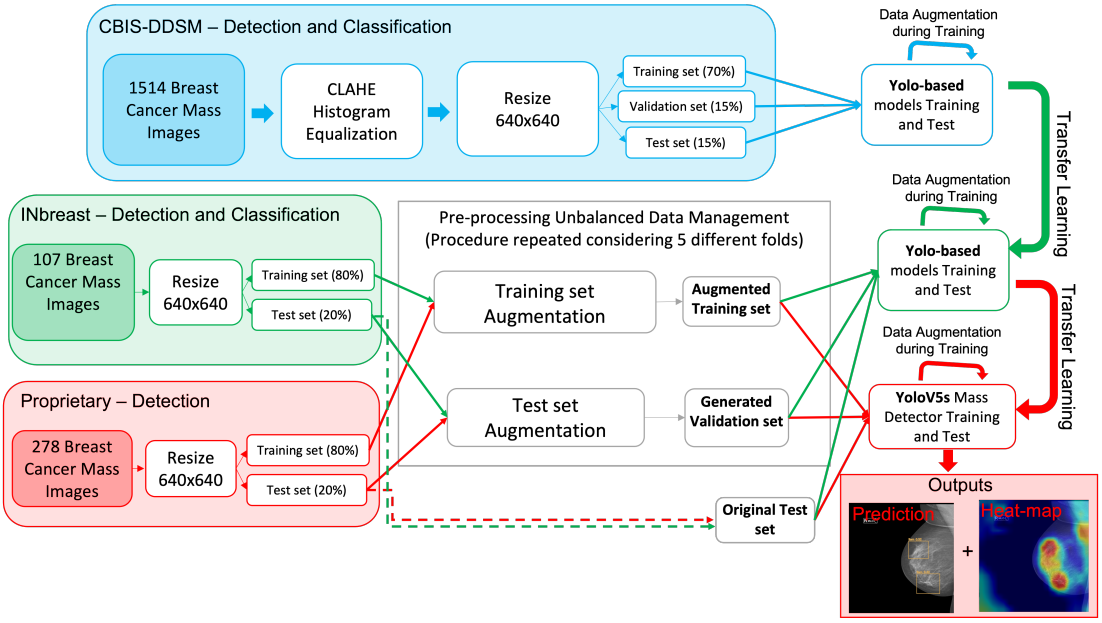


Figure 3.18: The overall architecture for breast cancer detection using Yolo-based architecture.

Palermo, Italy. Recognizing the importance of extensive data for effective deep learning architecture training, the transfer learning (TL) technique was employed. Recent research has demonstrated that training with small datasets, leveraging pre-trained models, represents a promising direction in developing reliable systems in the medical field [244]. To achieve this, the CBIS-DDSM dataset [245] and the INbreast dataset [246] were employed as source datasets [159], while the proprietary dataset served as the target dataset. In particular, the proprietary dataset contains lesions that present a greater challenge for recognition, including asymmetries and distortions, which hold significant clinical relevance [247]. The workflow of the experiments is depicted in Figure 3.18.

Despite the impressive performance of deep learning models, their widespread adoption is hindered by their inherent black-box nature, where the internal workings are not readily understandable to users [9]. This issue has raised critical concerns related to legal considerations, user acceptance, and trust, as extensively discussed in Section 2.2.

To facilitate the integration of these systems into real clinical practice, it is imperative to address the challenge of model explainability. In this study, the gradient-free method known as Eigen-CAM [248] was employed for generating saliency maps, and it was compared with the Occlusion Sensitivity method. These saliency maps served a crucial role in verifying the learning model's decision-making process and highlighting the most important pixels involved in the prediction.

Presenting the detected ROIs in the form of heatmaps can significantly assist physicians in their evaluation process. In fact, typically, ROIs are predicted and displayed only when they surpass a certain confidence threshold, which means that the most challenging-to-identify regions may not meet this threshold. By using heatmaps, it is possible to guide the physician's attention to these important areas, thereby supporting them in the complex, labor-intensive, and demanding task of mammogram evaluation.

To summarize, this work presents the following contributions. YoloV5-based architectures were compared with the previous YoloV3 model and considering the Vision Transformer block. In addition, to propose a transparent decision support system, Eigen-CAM, and Occlusion Sensitivity were used as Explainable AI algorithms [249, 7] to compute the saliency maps. The generated saliency maps were used for two main reasons: 1) as an explanatory debugging tool and to prevent inadequate outputs [250, 28], and 2) to guide the physician's attention even on predicted ROIs with low confidence [158].

3.5.1 Materials and Methods

3.5.1.1 Datasets

Open-source Datasets

The CBIS-DDSM dataset [245] represents a refined edition of the Digital Database for Screening Mammography (DDSM) dataset. It is composed of scanned film mammograms. A total of 1514 masses including 1618 lesions (850 benign and 768 malignant) were considered.

Conversely, the INbreast [246] dataset consists of Full-Field Digital mammograms (FFDM). Only the 107 benign and malignant images were selected. Lesions with a BI-RADS score greater than 3 were classified as malignant, while the other scores were categorized as benign. Certain images contained multiple lesions, resulting in the identification of a total of 40 benign and 75 malignant ROIs.

Proprietary

The dataset comprises 278 FFDMs that collectively contain 307 lesions. These lesions were meticulously annotated by expert radiologists who specialize in identifying abnormal regions. The images were acquired using a Fujifilm Full Field Digital system located in the Radiology section of the University Hospital "Paolo Giaccone" in Palermo, Italy. The images have a spatial resolution of 5928 $\times$ 4728 pixels and a pixel size of 50 µm. The benign lesions represent 17.6% of the dataset and 82.4% are malignant.

The acquired case series include *distortions* and *asymmetries* lesions which are particularly difficult cases to study. Detecting and diagnosing distortions in medical imaging can be especially demanding, given their distinctive features such as the presence of spicules extending from a central point, focal retractions, or straightening at the edges of the parenchyma [251]. Consequently, distortions often rank among the most frequently missed abnormalities in clinical practice [252]. Asymmetries are characterized by unilateral accumulations of fibroglandular tissue that do not fulfill the criteria for being categorized as masses. It has been estimated that approximately 20% of cases involving asymmetries are linked to malignancy, underscoring the significance of this area in research and clinical investigation [247]. The dataset is composed of masses (62%), asymmetries (15%), and distortions (23%). Given the large class imbalance, the proprietary dataset was used only for detection. The two open-source datasets were used also for classification.

3.5.1.2 Data Preprocessing

Yolo training requires the coordinates of the bounding-boxes containing the lesion and the class of each lesion (if present). For the CBIS-DDSM and INbreast datasets, the coordinates were computed using the smallest rectangle containing the segmented lesion. Instead, for the proprietary dataset the coordinates were computed from the square region that inscribes the circle containing the lesion.

Despite the CBIS-DDSM dataset has an acceptable size for deep learning, it is composed of scanned film mammograms. This results in noisy and poorly detailed images. For this reason, contrast-limited adaptive histogram equalization (CLAHE) was applied in the CBIS-DDSM images [253]. For all datasets, images were resized using the Lanczos Filter to 640x640 size [254, 255]. In addition 0-255 was considered as the interval for graylevels rescale. The CBIS-DDSM dataset was divided randomly considering 70% training, 15% validation, and 15% test set.

Considering the smaller size of INbreast and proprietary datasets compared with CBIS-DDSM, they were divided into training (80%) and test set (20%). In addition, to address the imbalanced classes issue the following section delves into two key aspects: data augmentation for achieving class balance and generating the validation set, as well as the procedure aimed at enhancing the training process.

Data Augmentation

Given the significant class imbalance within the INbreast and proprietary datasets, augmentation techniques were employed to increase the number of images in the minority class (benign) within the training set. Figure 3.19 provides a summary of the

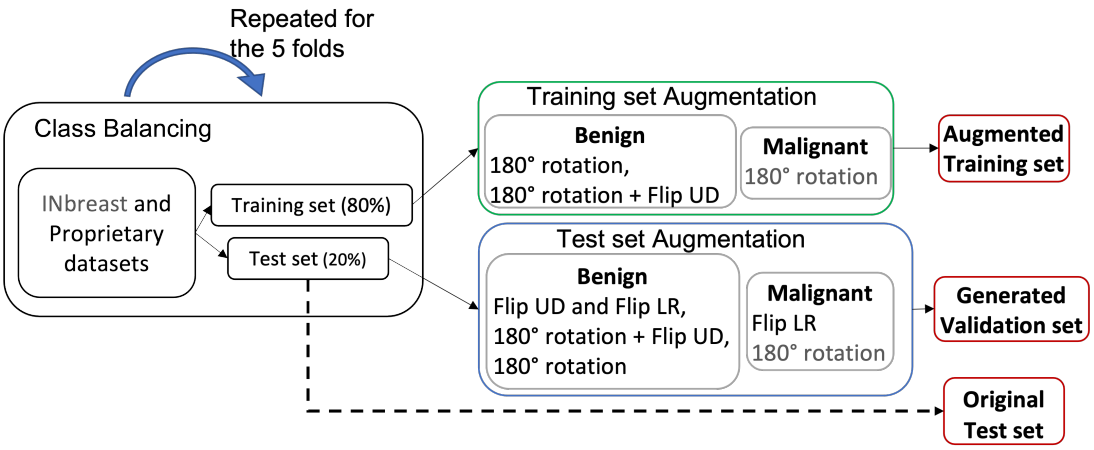


Figure 3.19: Transformations for class balancing and validation set creation. The procedure was repeated implementing the 5-fold cross-validation.

transformations used in the data augmentation process. Specifically, for benign images, transformations included a 180° rotation and a 180° rotation followed by a flip UD.

Additionally, as recommended in [239], the remaining portion of the test dataset was augmented to create the validation set. This augmentation involved the following transformations:

- For benign images: Flip UD, 180° rotation + Flip UD, Flip Left-Right (LR), and 180° rotation.
- For malignant images: Flip LR.

Furthermore, to address the smaller class difference within the INbreast dataset, 180° rotations were also considered for malignant masses.

This process effectively yielded a balanced validation set. Furthermore, it's worth noting that the procedure outlined for both the INbreast and proprietary datasets was repeated across five distinct partitions of training and test sets, thus employing a 5-fold cross-validation approach.

The other transformations were performed during the training. Image translation, rotation, scale, shear, flip UD, flip LR, and HSV augmentation were considered. Furthermore, all three datasets included a limited number of multi-lesion images. To enhance the model's ability to identify multiple lesions within a single image, the mosaic technique was employed. This augmentation method involves creating a 2x2 grid image that incorporates the target image and three randomly selected images from the dataset. The mosaic technique enhances training for two primary reasons: firstly, the

merging of these four images results in multiple ROIs within a single image, thereby enhancing the model's capacity to simultaneously recognize multiple ROIs. Secondly, to maintain the same input size, the four merged images and their respective ROIs are resized, thereby improving the detection of smaller lesions.

Three different data augmentation configurations were set: low, medium, and high. Table 3.20 shows the parameter set for each configuration. For HSV, translation, rotation, scale, and shear the value indicates the random range for the transformation. In the context of flip and mosaic transformations, the assigned value represents the probability of executing the transformation. Thus, a value of 0.5 is regarded as a substantial level of augmentation, as it implies that both augmented and non-augmented images are taken into account during the training process.

Table 3.20: Setting for data augmentation during the training phase.

Level	H,S,V	Translation	Rotation	Scale	Shear	Flip(UD,LR)	Mosaic
low	0.0, 0.0, 0.0	0.1	5.0	0.1	5.0	(0.5, 0.5)	0.0
med	0.007, 0.35, 0.2	0.3	10.0	0.3	5.0	(0.5, 0.5)	1.0
high	0.015, 0.7, 0.4	0.3	20.0	0.3	10.0	(0.5, 0.5)	0.5

3.5.1.3 Yolo-based Architectures and Training

Similar to other single-stage object detectors, Yolo comprises three key components: the Backbone, the Neck, and the Head. The Backbone is a CNN responsible for extracting and consolidating image features. The Neck facilitates feature extraction optimized for detecting small, medium, and large objects effectively. The three feature maps optimized for small, medium, and large objects, are then fed into the Head, which consists of convolutional layers used for making the final predictions. Yolo's methodology requires the image to be divided into a grid, and for each grid cell, a prediction is made. This prediction is represented as a 6-tuple denoted as $y = (p_c, b_x, b_y, b_h, b_w, c)$, where (b_x, b_y, b_h, b_w) specify the center coordinates (x, y) and dimensions (height, width) of the predicted bounding box, p_c signifies the probability of an object's presence within the cell, and c indicates the predicted class. To allow for the detection of multiple objects within the same grid cell, the concept of anchors is employed. Consequently, the 6-tuple prediction is generated for each predefined anchor. Each version of Yolo exhibits its own distinctive characteristics, primarily pertaining to the structure of the feature extractor, namely, the backbone.

YoloV3

YoloV3 represents a significant advancement over its predecessors, offering increased depth and improved accuracy. However, it comes at the cost of requiring more training

time and data. In YoloV3, the backbone architecture employed is Darknet53, as documented in [243]. Darknet53 is a hybrid design that combines elements from Darknet19 (used in YoloV2 [256]) and includes residual network components like the BottleNeck [61]. This hybrid approach was designed to enhance the capabilities of Darknet19 while also maintaining the efficiency exhibited by larger networks such as ResNet-101/152. Darknet53 considers the inclusion of shortcut connections, which enable the model to capture finer-grained information. This enhancement is particularly advantageous for improving the detection performance of small objects. For feature extraction suitable for various sizes of objects, YoloV3 incorporates the Feature Pyramids Network (FPN) [257]. FPN specializes in detecting both large and small objects, making the model versatile in handling a wide range of object scales. Additionally, YoloV3 employs a non-maximum suppression technique to select the most relevant bounding box when multiple overlapping bounding boxes are generated, further refining the model's object detection capabilities.

YoloV5

In YoloV5, the Backbone architecture employed is CSPDarknet53. This backbone builds upon the Darknet53 architecture proposed in [243]. It incorporates a CSPNet strategy [258], which involves partitioning the feature map of the base layer into two segments and subsequently merging them through a cross-stage hierarchy. This design enhances the feature extraction process, leading to improved performance. For feature extraction across multiple scales, YoloV5 uses the Neck component PAnet [259] to create a Feature Pyramids Network (FPN). This FPN enables the extraction of multi-scale feature maps, optimizing the model's ability to detect objects of varying sizes. YoloV5 is implemented in different versions, including Nano, Small, Medium, Large, and Extra-Large, each with variations in the number of convolutional kernels used, thereby affecting the total number of parameters. In this work, a comparative analysis was conducted between the Nano, Small, Medium, and Large versions of YoloV5.

YoloV5-Transformer

Unlike convolutional networks, Transformers have the capability to model complex relationships among small patches within an image. The fundamental concept behind a Transformer block is the assumption that the image can be divided into a sequence of patches, with each patch being converted into a vector representation. These vectorized image patches are then used to create lower-dimensional linear embeddings, which are subsequently inputted into a Transformer Encoder. This Transformer Encoder employs Multi-Head Attention mechanisms to identify both local and global dependencies within the image. Studies have demonstrated that incorporating a Transformer block

into convolutional networks can enhance efficiency and overall accuracy [260]. In the context of YoloV5, the Transformer block is integrated into the penultimate layer of the backbone. Specifically, it is positioned among the three convolutional layers that precede the spatial pyramid pooling layer.

Models Training and Evaluation Protocol

Due to the small size of both the INbreast and proprietary datasets, training a deep architecture like Yolo could potentially compromise the reliability of the resulting models. As a solution, although the CBIS-DDSM dataset consists mainly of scanned film mammograms, it was used as the source dataset for initial model training. This setup allows for the application of the TL technique on the INbreast and proprietary target datasets. Both the source and target datasets are labeled, and the TL technique employed is known as Inductive Transfer Learning [33].

Considering that Yolo simultaneously addresses a regression task to predict bounding box coordinates and two classification tasks to predict object detection confidence and class scores, two distinct loss functions were employed. For the regression task, the Complete IoU Loss was employed, while for the classification tasks, Binary Cross-Entropy with Logits loss functions were used in both cases.

The results were evaluated using common metrics for object detection tasks, including Precision, Recall, and Average Precision (AP). The AP is defined as the area under the precision-recall curve. In this evaluation, the Intersection over Union (IoU) threshold was set to 0.5. For CBIS-DDSM and INbreast datasets, AP was calculated separately for detecting malignant lesions (M AP) and benign lesions (B AP), as well as the mean of the two classes (mAP).

3.5.1.4 Models Explanation

It is imperative to verify the trained models before their deployment in clinical settings. For this reason, the developed model generates prediction explanations as a secondary output. Saliency maps are a valuable tool in this regard, as they have the ability to unveil the pixels or regions that have played a crucial role in the system's decision-making process. This functionality effectively highlights all potential ROIs for the benefit of the physician, aiding in the interpretability and trustworthiness of the model's predictions.

Numerous gradient-based techniques, such as CAM (Class Activation Map) [40], Grad-CAM (Gradient-weighted Class Activation Mapping) [41], and GradCAM++ [42], have

been developed to enhance the interpretability and transparency of deep learning models. These methods, in particular, are designed for class-discriminative visualization and rely on class probability scores for gradient calculations. However, gradient-based methods have a notable drawback: they entail additional computational overhead when backpropagating any quantity. Furthermore, they assume that the model's classifiers make correct decisions. When incorrect decisions are made, these methods tend to produce inaccurate or distorted visualizations [248]. Consequently, the localization accuracy of these techniques tends to be weak, especially in cases where incorrect predictions are involved.

Furthermore, traditional CNNs typically provide class distributions for each sample, whereas YOLO's output comprises bounding box coordinates, object presence probabilities within each cell, and class distributions. These unique characteristics often render the YOLO output non-differentiable, making it impractical for the implementation of gradient-based algorithms. Consequently, many object detection studies use gradient-free methods as a means of architecture interpretation [261, 262, 263]. To address these challenges, this study introduces Eigen-CAM for the computation of saliency maps and conducts a comparison with the Occlusion Sensitivity method.

Eigen-CAM is a gradient-free approach used to calculate and display the principal components of the learned features from the convolutional layers. This method offers an intuitive and versatile solution that can be applied to various deep learning models. In Eigen-CAM, the underlying assumption is that all the spatial features considered relevant and learned throughout the hierarchy of the CNN model will be retained during the optimization process, while non-relevant features will be subjected to regularization or smoothing to enhance interpretability and visualization.

Eigen-CAM is computed considering the input image I of size $i \times j$ projected onto the last convolutional layer $L = K$ and is given by: $O_{L=K} = W_{L=K}^{T} I$. The matrix $O_{L=K} = U\Sigma V^{T}$ is factorized using the singular value decomposition to obtain the principal components. The activation map is given by the projection on the first eigenvector $L_{Eigen-CAM} = O_{L=K} V_1$, where V_1 is the first eigenvector in the V matrix.

Like Eigen-CAM, Occlusion Sensitivity is a technique relevant to image detection tasks. It shares the characteristics of being gradient-free and not dependent on the particular architecture employed. Occlusion Sensitivity evaluates alterations in activations that arise from the occlusion of various regions within an image [264].

Saliency maps were proposed as a valuable tool for enhancing the predictions generated by YoloV5, and they play a crucial role in aiding physicians during the diagnostic process, particularly when the model's predictions are not entirely accurate. YoloV5

typically offers predictions only when it surpasses a specific confidence threshold. The purpose of using saliency maps is to identify all potential ROIs and address potential false negatives. In fact, many cancer types progress to an invasive stage precisely because early predictions often fail, even in the presence of preliminary signs. In contrast to YoloV5's predictions, saliency maps provide all potential ROIs, even those with low confidence levels. This expanded approach could result in an increase in false positives. Given this context, physicians receive two outputs:

- The conventional YoloV5 output, balances precision and recall and provides only those ROIs that exceed a certain confidence level. This output focuses on higher-confidence predictions.
- Saliency maps, which propose all potential ROIs, including those with lower confidence levels. These potential ROIs may serve as early indicators of cancer, even if their probability of being actual lesions does not surpass the threshold.

This dual-output approach provides a comprehensive perspective, allowing physicians to make informed decisions and potentially detect cancer at an earlier stage.

3.5.2 Results

The experiments were conducted in Python 3 environment on Google Colaboratory Pro. The PyTorch implementation provided by Ultralytics [265] was used for the experiments, and the training process was monitored using the Weights & Biases platform [266].

One hundred epochs with a batch size of 16 were set for the training process. Model selection was based on the validation mAP, and the best model was determined through a weighted combination of mAP@0.5 and mAP@0.5:0.95, with weights assigned as 0.9 and 0.1, respectively.

3.5.2.1 Performance on CBIS-DDSM and INbreast

The CBIS-DDSM dataset served as source dataset for evaluating the optimal YoloV5 architecture and for hyperparameter optimization, considering the nano, small, medium, and large versions. Additionally, it was used as a source dataset to implement inductive TL and enhance the model's generalization capabilities on INbreast and proprietary FFDM images.

Due to the multitude of hyperparameters involved, an initial analysis was conducted using the default values proposed for each YoloV5-based model. The results achieved for each version of YoloV5 are summarized in Table 3.21. In particular, the Nano and

Table 3.21: Comparison of the Nano, Small, Medium, and Large architectures of YoloV5 on the CBIS-DDSM dataset, considering all default hyperparameters.

Model	B AP	M AP	Precision	Recall	mAP
n	0.257	0.479	0.473	0.408	0.368
s	0.257	0.518	0.447	0.427	0.387
m	0.280	0.514	0.489	0.403	0.397
l	0.239	0.488	0.491	0.377	0.364

Large versions exhibited lower mean mAP compared to the Small and Medium versions. In contrast, the Small model, when compared to the Medium model, displayed a more balanced precision and recall pair while maintaining a significantly smaller parameter count, approximately one-third of the Medium model. Consequently, all subsequent experiments were exclusively performed using the Small model.

Table 3.22 indicates that the histogram equalization method has a positive impact on model performance. Furthermore, it shows that the Adam optimizer, with a learning rate of 0.001, outperforms the default Stochastic Gradient Descent (SGD) optimizer, which uses a learning rate of 0.01. As a result, experiments were conducted to assess the influence of data augmentation using the equalized dataset and the Adam optimizer. Table 3.22 also demonstrates how the results improve as data augmentation is increased. This extensive data augmentation underscores the importance of having substantial amounts of data when training a deep architecture such as Yolo. It validates the decision to use the CBIS-DDSM dataset as the source dataset for TL on the INbreast and proprietary datasets, given the improvement in performance observed with data augmentation.

Table 3.22: Performance of YoloV5 Small version, considering the equalized CBIS-DDSM dataset, Adam optimizer, and the three data augmentation configurations.

Hyps	B AP	M AP	Precision	Recall	mAP
Equal	0.300	0.501	0.487	0.408	0.400
Adam+Equal	0.321	0.555	0.487	0.464	0.438
aug-low	0.241	0.49	0.46	0.394	0.366
aug-med	0.337	0.549	0.497	0.487	0.433
aug-high	0.361	0.634	0.566	0.482	0.498

Exploiting the optimized hyperparameters established on the CBIS-DDSM dataset, both YoloV3 and YoloV5-Transformer models were trained on the CBIS-DDSM dataset to implement TL technique on the INbreast target dataset. The results are summarized in Table 3.23. Due to the dataset's size, performance was assessed using 5-fold Cross Validation, and the mean and standard deviation were reported for each metric.

For all experiments, the best training protocol determined for the CBIS-DDSM dataset,

Table 3.23: 5-Fold results for the three used architectures on INbreast dataset (Tr is for Transformer; NoTL is the training without transfer learning).

Model	B AP	M AP	Precision	Recall	mAP
YoloV3	0.585 ± 0.093	0.890 ± 0.036	0.785 ± 0.012	0.695 ± 0.104	0.738 ± 0.061
YoloV5s-Tr	0.642 ± 0.060	0.894 ± 0.054	0.799 ± 0.118	0.742 ± 0.146	0.771 ± 0.048
YoloV5s-NoTL	0.652 ± 0.051	0.890 ± 0.047	0.835 ± 0.059	0.713 ± 0.770	0.771 ± 0.038
YoloV5s	0.771 ± 0.131	0.898 ± 0.069	0.854 ± 0.097	0.729 ± 0.100	0.835 ± 0.098

which includes using the Adam optimizer, applying high levels of data augmentation, and employing a batch size of 16, was applied.

Furthermore, INbreast was also trained from scratch to showcase the difference in accuracy between models with and without TL. The YoloV5s model outperforms its previous version, YoloV3, and also the YoloV5-Transformer. It's worth noting that YoloV3 has a feature extractor with significantly more parameters than YoloV5s and Transformer (approximately 61 million vs. 7 million), necessitating a larger amount of data for effective training. Despite having a comparable number of parameters to YoloV5s, the YoloV5-Transformer version exhibited lower performance. When comparing YoloV5s trained from scratch with YoloV5s trained using TL on INbreast, there is an improvement of 0.061 in mAP and 0.119 in B AP. This improvement highlights the effectiveness of TL in leveraging knowledge from the source dataset (CBIS-DDSM) to enhance performance on the target dataset (INbreast).

The observed performance imbalance among classes underscores the dataset's characteristics, where the detection rate for benign lesions, representing the minority class, is lower than that for malignant lesions across all considered models.

3.5.2.2 Performance on Proprietary dataset

The YoloV5s model emerged as the most accurate choice for the two open-source datasets and was subsequently employed for lesion detection on the proprietary dataset. The model trained on the CBIS-DDSM as the source dataset and INbreast as the target dataset served as the starting checkpoint for training on the proprietary dataset. This approach allowed the model trained on the proprietary dataset to inherit the knowledge acquired from both the CBIS-DDSM and INbreast datasets.

Figure 3.20 illustrates the difference in validation mAP observed during training with and without transfer learning. In particular, transfer learning resulted in a higher initial mAP, faster mAP growth in the early training epochs, and a higher mAP asymptote, in accordance with the principles of transfer learning [133]. This outcome was further validated on the test set, where an mAP of 0.561 was achieved without transfer learning,

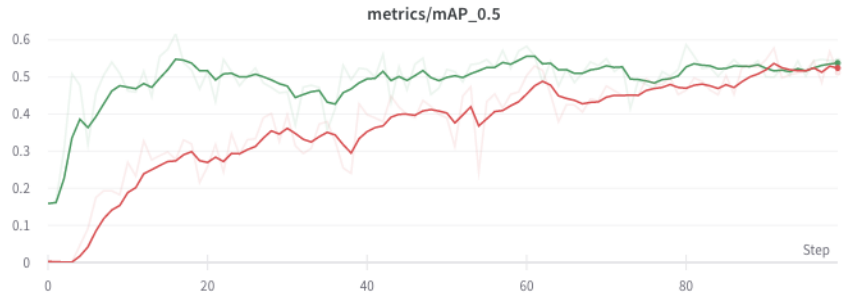


Figure 3.20: Training performance with (green) and without (red) transfer learning on the proprietary dataset.

compared to an mAP of 0.61 with transfer learning.

Table 3.24 presents the results obtained within the context of the 5-Fold cross-validation strategy for the proprietary dataset.

Table 3.24: 5-Fold results on the proprietary dataset, considering the training with and without transfer learning.

Model	Precision	Recall	mAP
YoloV5s no-TL	0.665 ± 0.054	0.541 ± 0.043	0.561 ± 0.053
YoloV5s TL	0.726 ± 0.110	0.591 ± 0.063	0.621 ± 0.035

3.5.2.3 Model Explanation and Improvements

To assess the performance using XAI methods, a manual analysis was conducted on a subset of the proprietary dataset, which included 50 (51 lesions). No healthy images were considered in this evaluation. The primary objective was to evaluate the differences in false positives and false negatives using two XAI techniques: Eigen-CAM and Occlusion Sensitivity.

In this qualitative analysis, the generated saliency maps did not exhibit complete overlap, as demonstrated in Figures 3.21 and 3.22. This phenomenon is consistent with existing literature, which has also highlighted the limited overlap between saliency maps generated by different methods [155, 15, 267]. Specifically, when using Occlusion Sensitivity, the regions associated with lesions appeared to be slightly illuminated compared to Eigen-CAM, where they were more prominently highlighted.

In addition to the qualitative observations, the quantitative analysis indicated the superiority of Eigen-CAM for this specific object detection task in mammography.

Table 3.25 provides a summary of the results obtained from the analysis. In the selected subset, the Yolo model correctly detected 41 lesions but missed 15 lesions (false

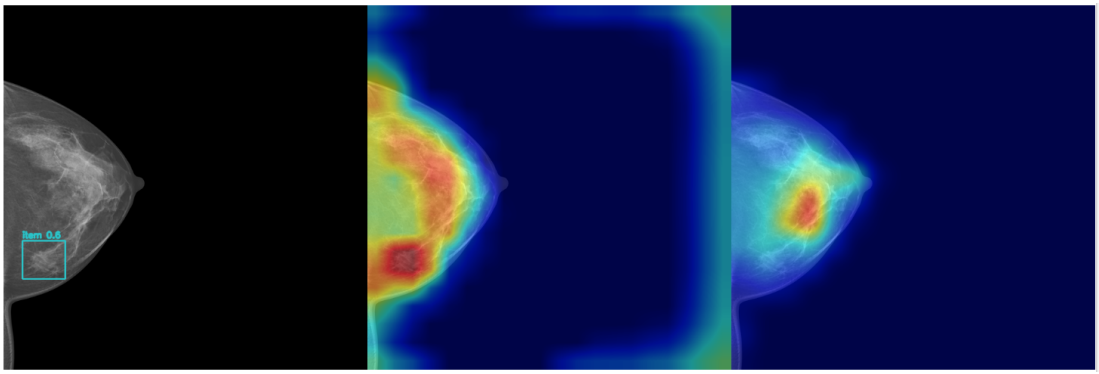


Figure 3.21: Example of a bounding-box prediction on the left and the respective saliency map on the center (Eigen-CAM) and on the right (Occlusion sensitivity). The ROI is correctly predicted with a confidence index of 0.6. However, also other suspicious areas are highlighted on the saliency map.

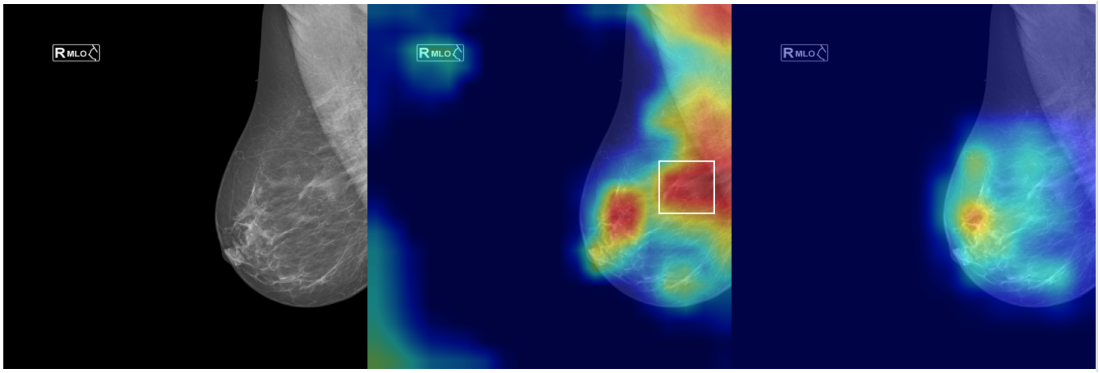


Figure 3.22: Example of wrong prediction on the left and the respective saliency map on the center (Eigen-CAM) and on the right (Occlusion sensitivity). Despite the error, the saliency map calculated *via* Eigen-Cam provides several suspicious ROIs, as well as the miss-detected lesion (marked with the white bounding-box).

negatives) and incorrectly identified 19 non-existent lesions (false positives). However, improved results were observed when Eigen-CAM was employed. Out of the 56 lesions, 52 were correctly detected, reducing the false negatives to only 4. However, the use of Eigen-CAM led to an increase in false positives, totaling 34. Conversely, the Occlusion Sensitivity method did not perform as well as Eigen-CAM. It exhibited an increase in false negatives to 20 and a higher number of false positives, totaling 55.

3.5.3 Discussion

Yolo Performance for Breast Cancer Detection

The proposed work for breast bancer Detection introduces several novelty and advantages. In particular, it involves the use of three distinct datasets, each contributing

Table 3.25: Performance variation through the use of saliency maps.

Model	Lesions #	TP	FP	FN
Yolo-based	56	41	19	15
Eigen-CAM	56	52	34	4
OS	56	36	55	20

its unique characteristics to the research. The CBIS-DDSM dataset, being the largest among the three, serves as an excellent choice for deep training, allowing for the development of robust models. However, it is important to note that the CBIS-DDSM dataset primarily consists of scanned film mammograms, resulting in images that are distinct from the FFDM images commonly encountered in clinical practice. On the other hand, the INbreast and the proprietary FFDM datasets represent valuable benchmarks for testing Yolo on real clinical practice images. These datasets provide a more representative sample of the images encountered in actual clinical scenarios. For this reason, the CBIS-DDSM dataset was suitable as source to find an optimal pre-training, as opposed to the commonly used COCO dataset. This decision reflects the aim of aligning the pre-training data with the domain of interest, which is crucial for achieving better performance on the target datasets, especially when dealing with medical images. Indeed, the COCO dataset is primarily used for object recognition tasks, composed of objects such as cars, people, and various other real-world items. These objects exhibit significantly different distributions when compared to breast cancer in mammograms. Therefore, for all the experiments, the transfer learning technique was used employing the CBIS-DDSM dataset as the source dataset.

Considering the evolutionary nature of Yolo architectures, which aim to enhance both accuracy and inference speeds, it was not immediately evident that YoloV5 would outperform YoloV3. Furthermore, among the various architectures, the smaller version (YoloV5s) exhibited the highest accuracy, even when compared to YoloV5s-Transformer. The performance achieved on the proprietary dataset was slightly inferior to that on the INbreast dataset. However, it's important to acknowledge that the proprietary dataset contains three times as many lesions, which allows for a more precise evaluation of the model's capabilities. Additionally, despite both datasets focusing on breast cancer analysis, their distributions, and consequently their training requirements, differ. INbreast was acquired with a pixel size of 70µm while the proprietary dataset with pixel size of 50 µm. Regarding the spatial resolution INbreast 3328×4084 or 2560×3328 and the proprietary dataset 5928×4728.

Furthermore, a significant distinction lies in the heterogeneity of the datasets. Specifically, in the case of INbreast, the dataset comprises 107 abnormalities, which are exclusively masses, with just two cases of asymmetries. Conversely, the proprietary

dataset primarily consists of masses (62%), but it also includes asymmetries (15%) and distortions (23%). The inclusion of these diverse lesion types, accounting for 38% of the dataset, introduces an additional challenge for precise detection. In fact, according to BI-RADS [167], an architectural distortion (AD) refers to a non-definite visible mass. AD is not always indicative of cancer and may instead represent various benign processes and high-risk lesions [268]. AD accounts for a significant portion of breast cancers missed during screening, ranging from 12% to 45% [269]. Asymmetries refer to areas of fibroglandular tissue that are visible on just one mammographic projection, often arising due to the superimposition of normal breast tissue. Various types of asymmetries exist. For instance, the developing asymmetry carries a 15% risk of malignancy [270]. Conversely, global symmetry is typically considered a normal variant.

Despite the increased complexity introduced by considering various lesion types, this approach aligns the system more closely with real-world clinical scenarios. Consequently, the obtained results are promising and underscore the feasibility of addressing breast cancer detection without simplifying the task to mere patch classification.

Literature Comparison

Conducting a precise comparison with other studies poses challenges due to variations in datasets, preprocessing methods, and training protocols. Nevertheless, Table 3.26 provides an overview of some related works that share similarities with this work study.

In [271] OPTIMAM dataset (OMI-H), composed of about 5300 mammograms, was used as source dataset to perform TL on INbreast dataset. Using the Faster R-CNN architecture, they obtained an AUC-ROC of 0.79 and 0.95 for benign and malignant lesion detection. YoloV1 was used in [238], resulting in 99,5 and 99,9 for benign and malignant lesion detection in the DDSM dataset. Yolo9000 (e.g. YoloV2) is used in [272]: in contrast to this work, localization, and classification performance were evaluated separately on the INbreast dataset. In particular, first, the lesions are localized, and then only the localized ones are classified, resulting in a detection accuracy of 97.2 and a classification accuracy of 95.3. The most similar work to ours in terms of evaluation protocol and workflow was proposed by Aly *et al.* [239]. Using YoloV3, they obtained an AP of 94.2 and 84.6 for benign and malignant detection, respectively. However the reported best results are computed using a higher image spatial resolution (832×832 vs. our 640×640), and the results were reported in 5-fold cross-validation only for 448x448 spatial resolution. In fact, comparing the achieved results on the best fold with their result on 608×608 images, an AP of 88.5 (vs. their 87.5), and 92.2 (vs. their 80.8) for benign and malignant detection was abotained, respectively.

Table 3.26: Comparison between the proposed and other breast cancer detection works, considering the INbreast dataset. (Det: Detection; Cls: Classification; Acc: Accuracy; AP: Average Precision; → is for TL from dataset1 to dataset2.)

Paper	Architecture	Dataset	Performance
[271]	Faster R-CNN	Optimam → INbreast	AUC B: 0.79; M: 0.95
[238]	YoloV1	DDSM	AUC B: 99.5; M: 99.9
[272]	YoloV2	DDSM & INbreast	Det. Acc: 97.2; Cls Acc (AUC): 95.3
[239]	YoloV3	INbreast	AP B: 94.2; M: 84.6
Proposed	YoloV5s	CBIS-DDSM → INbreast	AP B:0.771 ± 0.131; M:0.898 ± 0.069

Importance of Explainability

Despite achieving encouraging performance, it's crucial for the system to be both accurate and trusted by physicians to integrate it into real clinical practice. To address this need, introspection and explanation of the trained model were performed using the Eigen-CAM method. Figure 3.21 and Figure 3.22 display two saliency maps generated through Eigen-CAM and the Occlusion Sensitivity methods. Specifically, the former image represents a correct prediction, while the latter demonstrates an incorrect prediction. In Figure 3.21, the Eigen-CAM heatmap predominantly highlights the area around the predicted lesion. However, it's also recommended for the physician to pay attention to other regions of the image. Conversely, in Figure 3.22, the model makes an error in its prediction (misses detection). Here, the advantage of using a gradient-free method becomes apparent. The Eigen-CAM heatmap identifies several salient areas that warrant the physician's attention, aiding in error identification and potentially improving overall diagnosis.

Furthermore, the saliency maps shown in Figure 3.21 and 3.22 primarily emphasize activations within the breast region. Any minimal activations observed outside this area, as seen in the Eigen-CAM maps, can be attributed to artifacts and are not considered confounding factors for the physician's interpretation. It is possible to speculate that the slight activations at the black edges of the images may assist in aligning the coordinates of the bounding boxes predicted on the opposite side of the image, where only the background is present. In addition, the obtained saliency maps are class-independent, as supported by clinical literature findings. Mammography is typically employed as a screening examination aimed at identifying various abnormalities. In contrast, other examination modalities like MRI are more informative for characterization purposes and are considered secondary examinations [101].

Based on these findings, Eigen-CAM emerges as the more suitable method compared to Occlusion Sensitivity for generating saliency maps in object detection tasks. Although it inevitably led to an increase in false positives, the substantial reduction in false

negatives is of paramount clinical significance. This reduction is particularly crucial from a clinical perspective as it facilitates early diagnosis and aids in scheduling further examinations, thereby ruling out the progression of invasive lesions.

Given these considerations, saliency maps should complement, rather than replace, the outputs of the Yolo model. Yolo's predictions tend to be stringent with a minimal number of false positives, whereas Eigen-CAM's predictions are more conservative with a minimal number of false negatives. It's important to emphasize that these outputs should be regarded as qualitative tools that always require clinical radiologic evaluation. Thus, it remains the responsibility of the physician to determine which areas merit additional examination and consideration.

Chapter 4

Explainable Machine-Learning Models for COVID-19 Prognosis Prediction using Clinical, Laboratory and Radiomic Features

4.1 Introduction

The global spread of the SARS-CoV-2 virus has had profound and destructive impacts on several aspects, including the economy, society, and public health. Although the proliferation of less severe variants and the accessibility of vaccines have led to a decline in mortality rates, accurately predicting health-threatening symptoms at an early stage remains a key challenge. [291, 292, 293].

Chest CT scans have shown high sensitivity in detecting COVID-19 [294, 295]. However, CXRs have emerged as a more sustainable and efficient approach for managing the substantial daily caseload [296]. Additionally, when CXR images are integrated with clinical and laboratory data, their prognostic accuracy improves [297, 298, 299]. In particular, several studies proposed the use of machine learning models to enhance the prediction of COVID-19 prognosis. However, optimizing model accuracy is not the sole consideration. In crucial settings, such as clinical environments, it is imperative to guarantee the interpretability of the trained models. These models must undergo technical validation by engineers to strengthen their robustness and reliability, as well

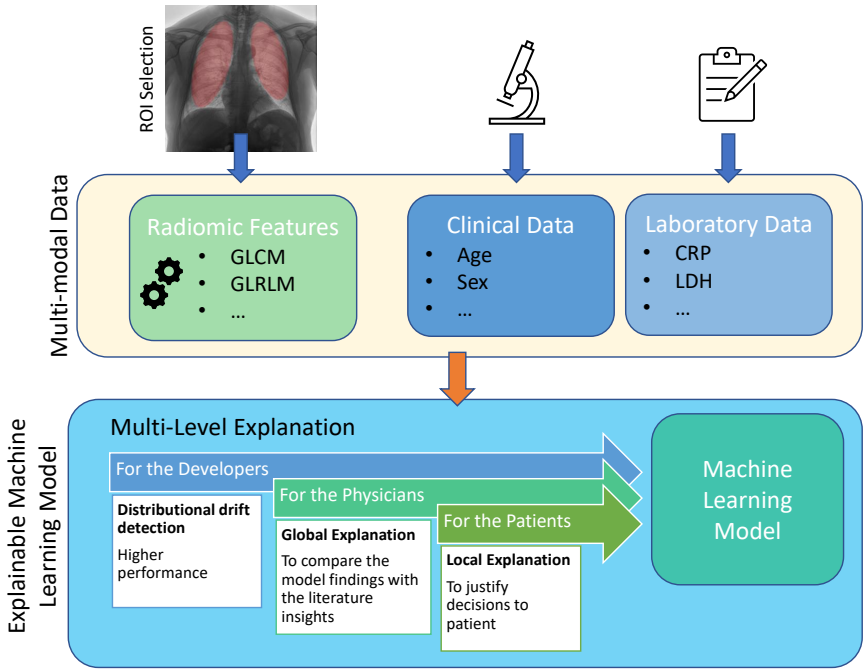


Figure 4.1: The proposed multilevel explainability makes it possible to focus on the needs of key stakeholders involved in the healthcare process.

as clinical validation by physicians to verify their efficacy and align them with existing clinical evidence. One primary approach to achieving model interpretability is by using inherently interpretable inputs. While clinical and laboratory features are inherently comprehensible to humans, imaging features may lack interpretability depending on the extraction process.

For instance, despite the introduction of various techniques aimed at elucidating the features extracted through deep neural networks, their inherent nature often lacks comprehensibility. These approaches primarily focus on computing saliency maps to highlight the areas most influence the model's decision-making process [300, 301]. To illustrate this, recent studies have highlighted a limitation in the Grad-CAM method's ability to distinguish multifocal lesions [44, 45]. Moreover, it has been observed that different methods can yield conflicting outcomes [155]. Additionally, these techniques provide only a local explanation for a specific instance (i.e., a patient), thereby preventing a comprehensive evaluation of the systems on a global scale. Consequently, saliency maps have yet to establish as an objective tool for validating clinical findings.

In recent years, particularly within the field of radiology, Radiomics has emerged as a powerful tool for extracting features. The advantages of using radiomic features have been extensively discussed in Section 2.3.2. The primary strength of radiomic features

resides in their inherent interpretability, as the significance of each feature is well-known. Nonetheless, radiomic studies frequently encounter challenges related to reproducibility. Moreover, many studies within the literature tend to present informative radiomic signatures without delving into a comprehensive clinical explanation or interpretation. Intelligible inputs represent the initial phase towards achieving an interpretable model. However, machine learning algorithms, such as SVM and Tree Ensemble, which are often considered as black boxes, have seen the development of numerous techniques [9, 47] for their *post-hoc* explanation [55]. These methods offer both global and local explanations, facilitating insights for various stakeholders in the healthcare process, including clinicians, technicians, nurses, general practitioners, healthcare policymakers, and patients [11].

Despite the deep features may be more informative, the extraction of higher-level radiomic features has shown great benefits. In particular, wavelet-derived features have shown exceptional predictivities in numerous scenarios [93, 94, 95, 96, 97, 98]. However, wavelets are frequently employed without careful consideration of the specific kernel type involved. In many cases, the commonly adopted practice is to use the default kernel for feature extraction, without prior evaluation to determine which kernel is better suited for the particular clinical scenario.

This study aimed to develop predictive models for COVID-19 prognosis prediction. Clinical, laboratory, and radiomic features were used as inputs and SVM and RF were implemented as classifiers. Additionally, various feature selection strategies were employed. Initially, unimodal models relied solely on clinical and laboratory data were evaluated, followed by models using only CXR radiomic features. Subsequently, multimodal models that combined both clinical and CXR were considered. The use of the machine learning algorithms mentioned above, in conjunction with intrinsically interpretable features, enabled the implementation of the multi-level explanation approach, as depicted in Figure 4.1. This approach involves both global and local explanations. The global explanation is employed for model introspection to assess the contribution of individual features, identify phenomena such as distributional drift, and validate any previously established clinical evidence. The local explanation, on the other hand, is employed to clarify predictions for each patient. The combination of intrinsically explainable inputs with global and local explanations serves as the foundation for developing an Explainable Clinical Decision Support System (X-CDSS) [302].

This work presents several significant contributions, including:

- The study introduces a multi-level explainability framework that considers perspectives from developers, physicians, and patients. This approach allows for a

comprehensive understanding of the models by assessing the role of each feature and quantifying their contribution to the final prognosis decision.

- A detailed examination of two shallow learning classifiers, namely SVM and RF, with the goal of defining predictive models for COVID-19 patient prognosis (i.e., distinguishing between MILD and SEVERE cases).
- The implementation of various feature selection strategies to identify the optimal feature set, which comprises both radiomic and clinical/laboratory features.
- An in-depth analysis by comparing different wavelet kernels and by evaluating their impact on the predictive capabilities of radiomic models.

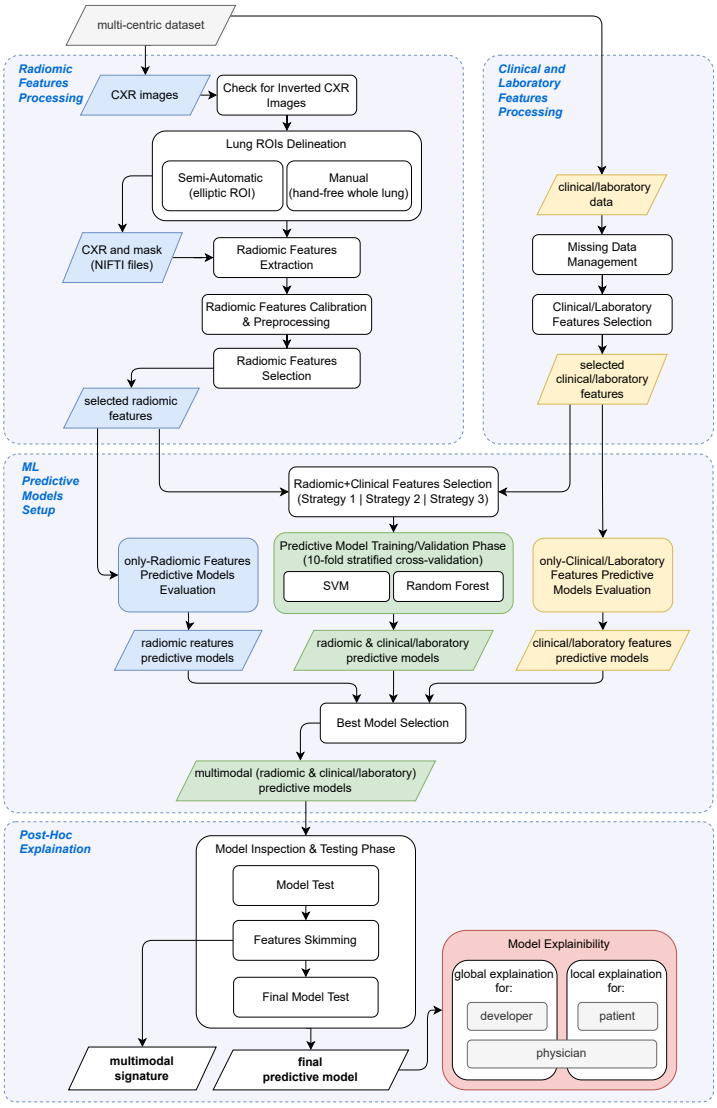


Figure 4.2: Overall flow diagram for Explainable COVID-19 prognosis prediction.

Figure 4.2 shows the overall workflow. Four high-level blocks were implemented: *i)* radiomic features processing, *ii)* clinical and laboratory features processing, *iii)* setup of the machine learning predictive models, and, finally, *iv)* implementation of the *post-hoc* explanation using the proposed multi-level explanation.

4.2 Related Works

With the global spread of the COVID-19 pandemic, there has been a substantial increase in interest in radiomic analysis. This approach has proven valuable in deriving additional insights related to the diagnosis, severity [303, 102, 299], and prognosis [304, 260, 297, 305, 45, 298] of the disease. These studies have explored both unimodal data, such as CT or CXR imaging [260, 305, 303], and multimodal data, which combines imaging with clinical information [297, 102, 299, 298]. While CT scans offer high-quality images with complex details, CXRs are a more rapid and efficient option. CXRs are particularly valuable in healthcare systems where the need for numerous daily examinations must be balanced with resource sustainability, as is often the case in public healthcare settings such as the National Health Service (NHS).

Angeli *et al.* [297] assessed the prognostic value of integrating CT scans with clinical and laboratory data for COVID-19 patients. They extracted Pulmonary Involvement (PI) and Pulmonary Consolidation (PC) scores from 301 CT images. Feature selection and model training using Logistic Regression were performed to predict improvement/recovery *vs.* ICU admission or death. In particular, the PC score showed no significant association (AUC = 0.722), but integrating PI and PC scores with demographic, comorbidities, and laboratory data improved AUC to 0.841. In Shiri *et al.* [305], 14339 CT images were used to predict overall survival outcomes. Radiomic features (texture, intensity, and shape) were extracted from lung segmentations generated by COLI-Net [306]. Various classifiers were employed, with the ANOVA feature selector and Random Forest yielding the best performance (AUC = 0.83, sensitivity = 0.81, specificity = 0.72). Wang *et al.* [298] employed 188 patient CT scans to predict disease progression (aggravation or improvement). Radiomic features from lesion ROIs, along with demographic and laboratory data, were integrated and selected using ICC and F-test methods. Different classifiers were tested, with AUC values of 0.843 for clinical features, 0.813 for radiomic features, and 0.865 for the combined set. In the work of Xu *et al.* [303], 284 CT images were classified into four COVID-19 progression groups. Radiomic features were extracted and selected using K-best and ElasticNet algorithms. SVM achieved microaverage and macroaverage AUCs of 0.89 and 0.90, respectively, on the test dataset. Finally, Shi *et al.* [299] developed a radiomic nomogram for COVID-19 severity classification using clinical, laboratory, and radiomic features. A Multi-Task

U-Net 2D was used for segmentation, and LASSO regression for radiomic features selection. The resulting model achieved an impressive AUC of 0.978 in the validation cohort.

In Soda *et al.* [102], the authors conducted prognosis prediction for COVID-19 patients using a dataset of 820 Chest X-Ray (CXR) images, classifying cases as MILD or SEVERE. They investigated the predictive capabilities of clinical/laboratory features, radiomic features, and their combination. For clinical features, they assessed both shallow learning and deep learning methods, using SVM and MLP. Regarding CXR images, three distinct approaches were explored: handcrafted, hybrid, and end-to-end deep learning. In the handcrafted approach, radiomic features were extracted using a pixel-based method [307] and lung segmentation *via* U-Net with manual refinement. In the hybrid approach, a range of CNNs including AlexNet, VGG, ResNet, DensNet, SqueezeNet, MobileNet, and their variants were trained to extract deep features from CXR images. These deep features were then combined with clinical and laboratory data and selected through Mutual Information and Recursive feature elimination. Classifiers such as SVM, Logistic Regression, and Random Forest were applied. For the end-to-end deep learning approach, deep features extracted using CNNs, with ResNet50 yielding the best results, were concatenated with clinical features. These deep features underwent processing through a dense structure, and similarly, clinical features were processed through a dense structure before being combined. The model was trained using Stochastic Gradient Descent. Results demonstrated that considering only CXR imaging, deep features outperformed radiomic features, achieving an accuracy of 0.705 compared to 0.65. The combination of clinical and imaging features significantly improved performance across all three approaches. In particular, the hybrid approach, using GoogleNet and Logistic Regression, yielded the highest accuracy. Additionally, the authors used Grad-CAM to provide explanations for their results, highlighting the regions within CXR images that contributed to the predictions. In Barbano *et al.* [308], several deep architectures based on ResNet-18 and DenseNet-121 were proposed. The most effective model was DenseNet-121, pre-trained on the CheXpert dataset comprising 224000 CXR images and tested on the CORDA-SLG dataset containing 451 CXR images. This model successfully classified COVID-19 patients as positive or negative, achieving sensitivity of 0.79, specificity of 0.82, and an AUC of 0.84. The authors also employed Grad-CAM to enhance the interpretability of their results.

In Guarrasi *et al.* [309], the authors explored various convolutional architectures and dense networks for prognosis prediction, specifically distinguishing between MILD and SEVERE cases using a dataset identical to that used in a previous study [102]. They

also included an additional 283 CXRs for external validation. Among the trained networks, an ensemble of three CNNs (GoogleNet-based, VGG-based, and ResNet-based) along with one MLP for clinical data was employed. The ensemble achieved an accuracy of 77.90 ± 1.27 when considering both imaging and clinical features. Additionally, they applied Grad-CAM to generate saliency maps of the three CNNs and Integrated Gradient for the MLP. In Borghesi *et al.* [304], the authors introduced the Brixia score as a means to assess COVID-19 infection using CXRs. They divided each image into six zones representing different parts of the lungs and assigned a score from 0 to 3 to indicate the level of impairment in each zone. The sum of these scores resulted in a total score ranging from 0 to 18. This Brixia score was manually assigned to 100 CXRs by an experienced thoracic radiologist and used to distinguish between recovery and death outcomes. Weighted Kappa (k_w) and the Mann-Whitney U-test were employed to compare the CXR scores with the final outcomes in selected patients, yielding a k_w of 0.82 with a 95% confidence interval of 0.79–0.86. In Signoroni *et al.* [45], the authors proposed the BS-Net for predicting the Brixia score using a dataset of 5000 CXRs. The BS-Net used a semi-quantitative approach to leverage the sensitivity of CXRs and radiologists' ability to identify COVID-19 pneumonia. It employed an end-to-end scheme that include segmentation, alignment, and score prediction. Key components included ResNet-18 for feature extraction, Nested U-Net for segmentation, alignment with synthetic transformations, optional hard self-attention, ROI pooling for Brixia score prediction, and Feature Pyramid Network for combining multi-scale feature maps. The model employed sparse categorical cross-entropy (SCCE) with a Mean Absolute Error contribution for joint multi-class classification and regression to predict the Brixia score accurately. Furthermore, to enhance the explainability of the Grad-CAM algorithm, the authors introduced a method inspired by LIME using the concept of super-pixels. This approach computed the difference between probability maps generated by different model replicas with a single super-pixel masked to zero, aiding in understanding network activity in lung areas and improving localization capability.

The existing literature suggests that the issue of interpretability in various studies has not received sufficient attention. Moreover, in cases where attempts have been made to provide interpretability, the resulting saliency maps have often been found to be unsatisfactory and inconsistent in their interpretations [301, 44]. This limitation is further highlighted in the work by Signoroni *et al.* [45], which introduced a custom approach to address the critical shortcomings of Grad-CAM. In this research, the primary focus was to underscore the critical importance of explainability, by presenting a methodology that effectively meets the needs of developers, physicians, and patients. The work aims to bridge the gap in interpretability and provide clear and reliable insights for all stakeholders involved.

Table 4.1: Multi-centric dataset characteristics used for the predictive models training/validation and the testing phases.

Hospital	**Phase**	**Images**	**SEVERE (%)**	**MILD (%)**	**LIEVE (%)**
A	train/validation	120	85 (70.83)	35 (29.17)	n.a.
B	train/validation	104	45 (43.27)	59 (56.73)	n.a.
C	train/validation	151	81 (46.36)	70 (53.64)	n.a.
D	train/validation	139	63 (45.33)	76 (54.67)	n.a.
E	train/validation	101	46 (45.55)	55 (54.45)	n.a.
F	train/validation	488	248 (50.81)	151 (30.94)	89 (18.25)
All	train/validation	1103	568 (50.36)	446 (46.6)	89 (3.04)
F	test	486	180 (37.04)	306 (62.96)	n.a.

4.3 Materials and Methods

Figure 4.2 shows the overall workflow. The next subsections describe each block of the processing pipeline in detail.

4.3.1 Multi-Centric Dataset Description

The dataset consists of information from 1589 COVID-19 patients, composed of clinical, laboratory, and CXR data. These patients were categorized into three prognosis groups: 'SEVERE,' 'MILD,' and 'LIEVE'. The classification was based on the level of hospital support they received. The 'SEVERE' category includes patients who needed non-invasive ventilation support, intensive care unit (ICU) care, or unfortunately, those who did not survive. Patients falling outside this category were considered 'MILD' [102]. Importantly, this dataset was collected from a total of six different hospitals, ensuring diversity and broad representation.

The dataset was split into two sets: one containing 1103 patients used for training and validation of the predictive models, and the other with 486 patients designated for the testing phase. This partitioning, with 1103 training/validation cases and 486 testing cases, was defined by the organizing committee of the Covid CXR Hackathon competition, which provided access to the dataset [310]. Table 4.1 presents the class distribution across the different hospitals.

CXR Images Details

The CXR images were provided in .PNG format with a 16-bit depth. From a preliminary qualitative assessment, it was evident that the dataset displayed significant heterogeneity in terms of both image size and overall quality. Table 4.2 provides an overview of the size distribution across the different centers, denoted as A, B, C, D,

Table 4.2: Variability in CXR image size across the different hospitals. Only the top 3 most frequent sizes (along with the number of images and percentage) are reported for each center.

Hospital	1^{st} **most frequent size** $(r \times c) : \#imgs - \%imgs$	2^{nd} **most frequent size** $(r \times c) : \#imgs - \%imgs$	3^{rd} **most frequent size** $(r \times c) : \#imgs - \%imgs$
A	$(4280 \times 3520) : 75 - 62.5\%$	$(2500 \times 2048) : 20 - 16.6\%$	$(2772 \times 2771) : 10 - 8.3\%$
B	$(4240 \times 3480) : 90 - 86.5\%$	$(2846 \times 2330) : 2 - 1.9\%$	$(2836 \times 2336) : 2 - 1.9\%$
C	$(2866 \times 2350) : 66 - 43.7\%$	$(3000 \times 3000) : 6 - 3.9\%$	$(2917 \times 2402) : 6 - 3.9\%$
D	$(2648 \times 2208) : 33 - 23.7\%$	$(2140 \times 1760) : 21 - 15.1\%$	$(2648 \times 2176) : 21 - 15.1\%$
E	$(4280 \times 3520) : 33 - 32.6\%$	$(2880 \times 2880) : 24 - 23.7\%$	$(2936 \times 3080) : 8 - 7.9\%$
F	$(2836 \times 2336) : 392 - 80.3\%$	$(2336 \times 2836) : 17 - 3.4\%$	$(2012 \times 2012) : 7 - 1.4\%$

E, and F. As a result, these images exhibit significant variability in terms of inherent image quality and acquisition conditions. Regarding image quality, the dataset contains not only natively digital images but also images obtained by scanning X-Ray plates. These scanned images typically exhibit lower quality due to the conversion process. Furthermore, some images within the dataset display an inverted pattern compared to the typical representation of X-Ray images. In the conventional representation, bones appear as hyperintense regions (indicating high density), while lung areas appear hypointense (indicating low density). However, in certain cases, the dataset includes images with an opposite pattern, where bones are represented as hyperintense regions. This inversion necessitated adjustments to align with the conventional representation. Additionally, in terms of the clinical context, the dataset comprises images of patients with a range of medical devices, including permanent life-support devices such as pacemakers and temporary ones such as tubes for forced ventilation, thoracic electrodes, and monitoring wires. This diversity in patient conditions and equipment further contributes to the dataset's complexity and heterogeneity.

Clinical and Laboratory Features Selection and Data Imputation

For each patient, clinical and laboratory data were associated with the CXR image (the complete features list is provided in Section A.2.1). The *prognosis* feature served as the label for supervised training. However, for input into the predictive models, only 23 features were considered. Here's a breakdown of the prior feature selection process:

- *Excluded Features (3)*: Three features, namely "Hospital," "Position," and "Death," were excluded a priori and not used as input features;
- *Features with High Missing Data (5)*: Five features, including "Fibrinogen," "PCT," "dDimer," "SaO2," and "Obesity," were omitted because they had a missing data percentage exceeding 50%;
- *Features Not Present in the Test Set (6)*: An additional six features, consisting of

"OxPercentage," "CardiovascularDisease," "IschemicHeartDisease," "AtrialFibrillation," "HeartFailure," and "Ictus," were excluded from consideration because they were not present in the test set.

To address missing values in the remaining 23 clinical features, both univariate and multivariate data imputation techniques were employed. In the univariate imputation method, missing values for each feature were replaced using either the mean or median value of the available data for that specific feature. In the multivariate imputation method, at each step of the imputation process, one of the feature columns with missing values was designated as the output, and the remaining feature columns were treated as inputs for a regressor. A regressor was then used to predict the missing values of the feature under consideration.

4.3.2 Lung ROIs Delineation Assesment

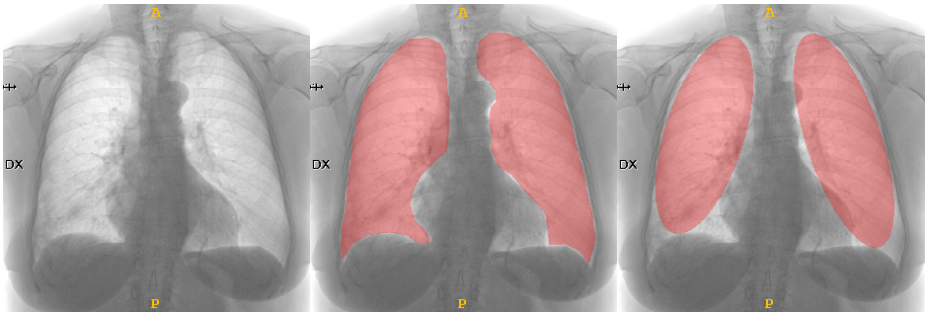
A custom tool was developed using MATLAB to define ROIs within the lung for the purpose of extracting radiomic features. This tool included two distinct segmentation methods:

1. *Manual Whole Lung Delineation*: In this approach, a radiologist with over three years of experience in X-Ray annotation manually delineated the boundaries of both the left and right lungs. The delineations were performed in collaboration with a senior radiologist.

2. *Semi-Automated Elliptical ROI Delineation*: This method involved a semi-automatic process to identify the largest elliptical region fully enclosed within the lung boundaries. The operator only needed to position the bounding box over the lung, and the implemented software automatically located the ellipse. This delineation method was designed to concentrate on the central area of the lungs while excluding peripheral zones.

The GUI allows 1) interactive selection of the two selection modes; 2) execution of segmentation; and 3) final saving. Specifically, the image and its mask were saved in NIFTI format.

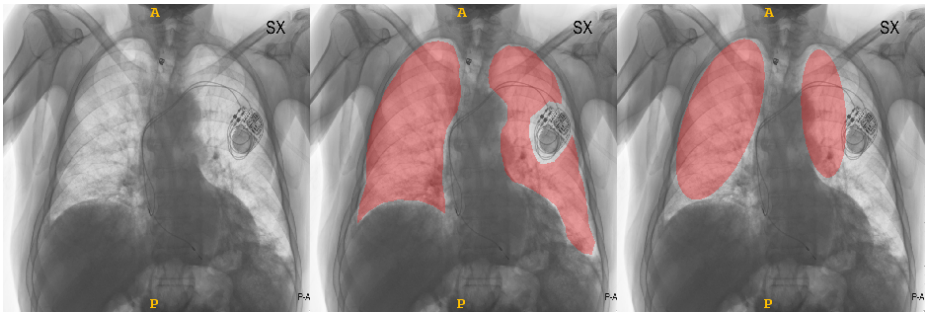
Figure 4.3 shows two examples related to the implemented annotation modalities: in (a) and (d) original CXR images of MILD and SEVERE patients, respectively; in (b) and (e) the *hand-free whole lung* delineations; in (c) and (f) the *semi-automatic elliptic ROI* delineations. Radiomic features serve to quantify the distribution and texture of lung tissues. The presence of external health-supporting devices, such as pacemakers, monitoring wires, respirator pipes, and others (as summarized in Table 4.3), can significantly affect the extracted feature values. Value in parenthesis in Table 4.3 represents



(a) Example of MILD COVID-19 patient

(b) hand-free whole lung delineation

(c) semi-automatic elliptic ROI delineation



(d) example of SEVERE COVID-19 patient

(e) hand-free whole lung delineation

(f) semi-automatic elliptic ROI delineation

Figure 4.3: Two examples of MILD and SEVERE patients with the related annotation modalities.

the percentage of the samples calculated with respect to the number of images in the corresponding prognosis class. To address this issue, ROIs that contained these external health-supporting devices were excluded in both delineation methods, as illustrated in the severe case depicted in Figure 4.3.

4.3.3 Radiomic Features Extraction

A total of 1023 features were extracted by means of the PyRadiomics [85, 84] toolkit. In particular, 93 original features were extracted, considering: first-order intensity histogram statistics, GLCM, GLRLM, GLSZM, GLDM and NGTDM. Then the same

Table 4.3: Number of patients with health-supporting on the CXR image.

Subset	**Prognosis**	**Health-supporting devices patients**
training/validation (1103)	MILD (535) SEVERE (568)	23 (4.30%) 51 (8.98%)
testing (486)	MILD (306) SEVERE (180)	16 (5.23%) 37 (20.55%)

features were extracted considering Laplacian of Gaussian (LoG) and Wavelets filtered images. For LoG filtering three different values of σ were considered ($\sigma \in \{1, 3, 5\}$), collecting 279 features ($279 = 93 \times 3$); Moreover, to determine the optimal quantization level, the features were extracted considering different *binWidth* values ($binWidth \in \{8, 16, 32, 64, 128, 256\}$). For the wavelet-derived features, an in-depth analysis was performed and described in the next subsection. After this analysis, for Wavelets transform, the Haar kernel and two decomposition levels ($levels \in \{1, 2\}$) were considered, obtaining 651 features ($651 = 93 \times 7$). Finally, 930 features were extracted from the filtered images.

4.3.4 Wavelet-derived Features Extraction

The widespread adoption of wavelet transforms in various signal and image processing applications arises from their unique ability to capture information across both the frequency and time domains. In the context of this research, the discrete wavelet transform (DWT) to CXR images was applied. The computation of DWT involves the use of two essential functions: the *scaling function* and the *wavelet function* [311]. These functions play a crucial role in the transformation process, allowing DWT to effectively extract valuable information from the input images while maintaining the ability to perform multi-resolution analysis for various practical applications. This involved subjecting the images to high-pass h_ψ and low-pass h_ϕ filtering operations, effectively decomposing the images into two distinct components: high-frequency (details) and low-frequency (approximation) components. This decomposition process results in the generation of subimages at different resolutions, facilitating multi-resolution analysis [182, 183]. Consequently, DWT has found extensive applications in the field of image processing, with a specific focus on tasks such as denoising [184] and compression [185, 186].

In this study *Biorthogonal* (Bior1.5), *Coiflets* (Coif1), *Daubechies* (Db3), *Discrete Meyer* (Dmey), *Haar*, *Reverse Biorthogonal* (Rbio1.5), and *Symlets* (Sym2) wavelet families [312] were considered [313]. The following are the main applications of wavelet families:

- **Biorthogonal**: commonly used for denoising, in particular when white Gaussian noise is present [314];
- **Reverse Biorthogonal**: for compression [315] and denoising [316];
- **Coiflet**: for compression [317] and denoising [318];
- **Daubechies**: provides excellent performance in compression and are popular choice in medical imaging applications [319];

Table 4.4: Number of coefficients that define the kernel length.

Wavelet Kernel	Coefficients Number
Bior1.5	10
Coif1	6
Db3	6
Dmey	62
Haar	2
Rbio1.5	10
Sym2	4

- **Discrete Meyer**: in general used for multi-resolution analysis [320] and some variants for edge and blocking artifact reduction [321];
- **Haar**: the first introduced for wavelet transforms and several generalizations and modifications were proposed [322]. It is one of the most widely used and has many medical imaging applications, including image fusion [323], and compression in radiography [324], CT, and MRI [325];
- **Symlets**: a modified version of *Daubechies* wavelets with increased symmetry [326], used for signal decomposition including characterization of fabric texture [327].

To obtain the desired decomposition results, specific kernels of each wavelet family had to be experimentally selected. These kernel selections were made with the intention of preserving the visual and qualitative similarity between the decomposed images and the original image. The details of these chosen kernels and their corresponding number of coefficients are presented in Table 4.4. In particular, for all wavelet families except for *Dmey*, kernels with a number of coefficients equal to or less than 10 were the preferred choices.

4.3.5 Radiomic Features Calibration and Preprocessing

Features calibration and preprocessing were performed by following the steps [150]:

1. **Quantization level analysis**: The quantization level was determined based on the highest number of radiomic features, taking into account the Intraclass Correlation Coefficient (ICC) [85]. This analysis was instrumental in establishing the optimal bin width, which aimed to maximize the number of features exhibiting robustness in terms of ICC. In this research, the two-way random-effects model was adopted denoted as ICC(3,1) [328, 68].
2. **Near-zero variance analysis**: features exhibiting variance less than or equal

to 0.01 were discarded;

3. **Redundant features analysis**: highly correlated features (values greater than 0.85) were removed, using the Spearman correlation for pairwise feature comparison [112, 113, 114, 115].

4. **Statistical analysis**: to assess the difference between the distributions of MILD and SEVERE cases, the Mann-Whitney U test was employed for each feature that had been selected in the preceding steps. A significance threshold of 0.05 was applied. The p-values obtained were subjected to adjustment using the Bonferroni–Holm method [329].

4.3.6 Features Selection and Predictive Model Setup

Elliptic *vs.* Handcrafted Segmentation Evaluation

The process started with assessing the effectiveness of different segmentation techniques, namely, *hand-free whole lung* and *automated elliptic ROI*, in terms of their predictivity. This evaluation was carried out using the SFS method [111]. The SFS was configured to work in both forward (SFS) and floating mode (SFFS).

To evaluate performance, a 10-fold cross-validation with stratification was employed, and SVM and RF algorithms were used as classifiers. For all experiments involving SVM, data normalization was applied.

Radiomic and Clinical/Laboratory Feature Selection

An initial feature selection was performed to assess the individual performance of each unimodal model, specifically the clinical/laboratory and radiomic models. SFFS was employed to investigate how performance changes as the number of features increases. The purpose was to determine the optimal number of features to maximize accuracy.

For the clinical/laboratory features, SFFS was considered to analyze the accuracy trend considering all 23 features. In contrast, for the radiomic features, SFFS was set to select the best 30 features. Throughout these feature selection experiments using SFFS, a stratified 10-fold cross-validation approach was employed to ensure robust evaluation.

Subsequently, three strategies were applied:

- *Selection Strategy 1*: SFFS was applied considering the clinical/laboratory and radiomic features selected in the preprocessing selection step.

- *Selection Strategy 2*: SFFS was applied to all the clinical/laboratory and all radiomic features.

- *Selection Strategy 3*: SFFS was applied to the optimal number of radiomic features selected in the preprocessing selection step and all the clinical/laboratory features. This strategy was implemented to balance the ratio between clinical/laboratory and radiomic features.

Model Training and Test

Before moving to the training and validation stage, several preparatory steps were completed. These included imputation for missing data in the clinical and laboratory dataset, calibration, and preprocessing of the radiomic features. Following these steps, a stratified 10-fold cross-validation was conducted, and this process was repeated 20 times to fine-tune hyperparameters. During this cross-validation process, the model with the highest accuracy was selected as the best-performing model. Subsequently, this chosen model was evaluated for its performance on the test dataset.

4.3.7 Multi-Level Explainability

To successfully develop and incorporate a CDSS into actual clinical practice, it's essential that the system is not only effective but also transparent and comprehensible to its users. This is why the work proposes a multi-level explanation framework that considers the viewpoints of both the developer and the various stakeholders participating in the healthcare process, such as physicians and patients.

Developer Perspective

The developer's primary objective is to train models that can effectively make predictions on new, unseen data. While stringent validation protocols can help identify overfitting issues within the training dataset, there is a risk that the model may not perform well when faced with a distribution that is even slightly different from the one it was trained on [23]. This situation, often referred to as *distribution drift*, can be mitigated by incorporating explainable AI algorithms into the model.

To identify and mitigate the problem of distributional drift, the Mean Decrease Accuracy (MDA) method was implemented, which is part of the ELI5 framework [230]. The importance of features were calculated using a Leave One Center Out (LOCO) procedure, with each center representing one of the six hospitals (A, B, C, D, E, and F). The LOCO evaluation process involves iterative partitioning of the dataset samples for each center. During each iteration, five out of the six centers are designated for training, while the remaining one is reserved for testing. The following methodology was used to drop center-dependent features and select only the descriptive features of the COVID-19 prognosis. In particular:

- the MDA method was used to calculate the features' importance of each center. For example, to compute feature importance for hospital A, all hospitals were used for the training, and A for test and MDA computation. This procedure is repeated for each hospital.
- according with MDA method, a positive weight is representative of significant features, *vice versa* a negative weight is representative of unstable features.
- to define a feature as stable, 3 different criteria were established, selecting features that in at least 3, 4, or 5 centers (out of 6) obtained positive weights; features with less than 3 positive weights (across all hospitals) were considered dependent on the acquisition center and then discarded.

Through this procedure, three distinct subsets of features were derived based on their weights across the six centers. These subsets include features with at least 3, 4, or 5 positive weights. To assess the effectiveness of this approach, model performance were evaluated with and without this debugging step, taking into account only the features present in these three subsets. The reasoning behind this evaluation is that by excluding features that lack stability across multiple centers (those features dependent on the specific hospital rather than being informative of the underlying phenomenon), can potentially improve the generalization capabilities of the trained models.

Physician Perspective

Physicians play a crucial role in ensuring that the patterns learned by the model align with clinical evidence. This is achieved by using inherently interpretable features, such as clinical, laboratory, and radiomic data. By comparing the model's results with established clinical practices and detecting any inconsistencies with medical literature, physicians can validate the model's performance.

To provide a global explanation, the SHapley Additive exPlanations analysis method [50] was employed. Specifically, SHAP was used to identify the features that influence the model's output, guiding it toward either a SEVERE or MILD prediction. This step involved collaboration with a medical team, enabling them to verify the results obtained from the model with relevant medical literature. This collaborative effort ensures that the model's predictions align with established medical knowledge and clinical expertise.

Patient Perspective

In compliance with the GDPR [25], which requires explanations for users receiving system decisions, particularly patients, a local explanation is provided for each specific instance. To fulfill this requirement, a local explanation is generated using the SHAP

Table 4.5: AUROC values obtained in training (*mean* ± *standard deviation*) and in testing phases for each wavelet kernel.

Wavelet Kernel	**Machine Learning Model**			
	SVM		**RF**	
	Train	**Test**	**Train**	**Test**
Bior1.5	0.725 ± 0.044	**0.689**	0.711 ± 0.047	**0.706**
Coif1	0.710 ± 0.046	0.670	0.708 ± 0.044	0.679
Db3	0.708 ± 0.051	0.676	0.690 ± 0.044	0.653
Dmey	0.700 ± 0.049	0.650	0.678 ± 0.047	0.662
Haar	$\mathbf{0.734} \pm 0.047$	0.677	$\mathbf{0.726} \pm 0.046$	**0**.686
Rbio1.5	0.700 ± 0.050	0.649	0.697 ± 0.047	0.649
Sym2	0.718 ± 0.044	0.671	0.704 ± 0.047	0.689

analysis method. The SHAP analysis was also used to obtain a local explanation and to evaluate the features pushing the model toward a SEVERE or MILD decision.

4.4 Experimental Results

4.4.1 Wavelet-derived Feature Evaluation

For the evaluation of the optimal wavelet kernel, a comparison of different kernel families was performed. In particular, extracting the radiomic features from each transformed image considering the various kernels, the discussed preprocessing pipeline was applied. On the selected features, the SFFS method was applied, and then models were trained and tested.

Table 4.5 displays the AUROC results from the experimental trials in both the training and testing phases for each wavelet kernel used. The metrics reported for the training phase are presented as the *mean* ± *standard deviation*, as these values were computed by averaging results across 20 repetitions of the 10-fold stratified cross-validation.

Among the wavelet kernels, *Db3*, *Dmey*, and *Rbio1.5* consistently demonstrated the poorest performance across all machine learning models. AUROC is widely regarded as a key metric for assessing overall diagnostic accuracy, with higher values indicating better discrimination ability of the biomarkers [330]. In particular, there were overlapping AUROC values observed during testing for the *Bior1.5*, *Coif1*, *Haar*, and *Sym2* kernels. Eventually, the Haar kernel was selected for use as the primary kernel.

Figure 4.4, shows the confusion matrices obtained by the machine learning classifiers considering *Haar* as the wavelet kernel.

Table 4.6: Quantization level analysis results.

Bin-Width	Robust Features
8	319
16	407
32	**573**
64	488
128	416
256	381

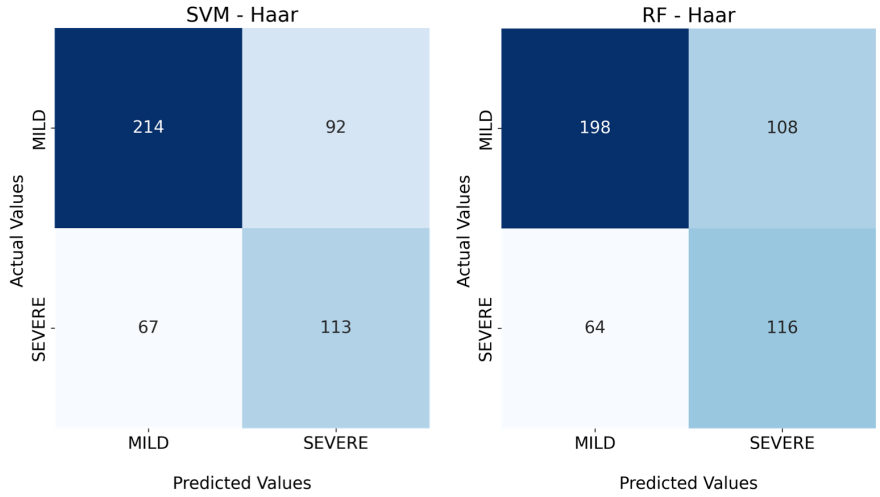


Figure 4.4: The confusion matrices on the test set obtained with the *Haar* kernel features.

4.4.2 Radiomic Features Preprocessing and Lung Delineation Selection

Several calibration and preprocessing steps were undertaken to identify features that were both robust and informative while eliminating redundancy. The process involved employing ICC analysis to determine the optimal quantization level from the set $binWidth \in \{8, 16, 32, 64, 128, 256\}$. Table 4.6 displays the count of robust features for each bin width. A bin width of 32 was chosen, adhering to the criterion of $ICC \geq 0.85$, and this choice was subsequently employed in all subsequent stages of the processing pipeline.

Then, the number of radiomic features was progressively reduced within each preprocessing step, and a final set of 40 features was obtained (see Table 4.7).

In conclusion, Table 4.8 presents the accuracy values, computed using the SFFS method during the 10-fold cross-validation procedure, for evaluating the optimal lung delineation approach. This evaluation considered two approaches: *hand-free whole lung*

Table 4.7: Calibration and preprocessing of radiomic features.

Preprocessing Step	**Analysis Method**	**Remaining Features**
initial features	n.a.	1023
near-zero variance analysis	$variance \leq 0.01$	354
redundant features analysis	Spearman correlation ($cutoff = 0.85$)	57
statistical analysis	Mann-Whitney U rank test ($p < 0.05$)	40

delineation and *automated elliptic ROI* delineation. In particular, both SVM and RF models achieved higher accuracy when features extracted from elliptic ROIs were used.

Clinicians provided justification for this result, noting that the elliptical modality specifically focuses on the central region of the lung, which is considered the most representative when compared to peripheral areas. As a result, features extracted from elliptic ROIs were chosen for use in subsequent experiments.

Table 4.8: Evaluation and choice of the best lung delineation approach. With both classifiers (i.e., SVM and RF), the automated elliptic ROI modality shows slightly better behaviour than the hand-free whole lung modality. Radiomic features considered here belong to both types (original and filtered).

Classifier	**Delineation Approach**	**Selected Features**	**Accuracy**
SVM	whole lung	19	0.673 ± 0.050
	elliptic ROI	22	0.710 ± 0.036
RF	whole lung	12	0.687 ± 0.071
	elliptic ROI	16	0.703 ± 0.057

4.4.3 Imputation of Missing Values in Clinical Data

The SFFS method was employed to determine the optimal imputation approach for clinical data. Table 4.9 displays the results obtained with SVM and RF using the three different imputation methods. No statistically significant differences were observed among the various approaches. However, the mean imputation method yielded a smaller standard deviation in the results. Consequently, in line with [102], the mean imputation method was selected for data imputation.

4.4.4 Feature Selection and Model Training

As previously mentioned, the feature selection process uses SFFS [111] to identify the optimal subset of features that maximized accuracy within a 10-fold cross-validation procedure.

Figure A.2 presents the initial results of feature selection for evaluating unimodal models. Based on the number of features that maximized accuracy, Table 4.10 presents the

Table 4.9: Comparison of imputation approaches.

Classifier	**Imputation Approach**	**Selected Features**	**Accuracy**
SVM	Mean	11	0.750 ± 0.031
	Median	11	0.753 ± 0.046
	LR	10	0.748 ± 0.042
RF	Mean	15	0.728 ± 0.031
	Median	15	0.746 ± 0.038
	LR	14	0.737 ± 0.042

Table 4.10: Preliminary feature selection result for unimodal models obtained by SVM and RF.

	SVM		**RF**	
Metrics	**Radiomic Features**	**Clinical / Laboratory**	**Radiomic Features**	**Clinical / Laboratory**
Accuracy	0.694 ± 0.039 [0.686, 0.701]	0.750 ± 0.041 [0.742, 0.758]	0.672 ± 0.044 [0.664, 0.680]	0.721 ± 0.038 [0.714, 0.728]
Sensitivity	0.668 ± 0.056 [0.658, 0.678]	0.772 ± 0.050 [0.763, 0.781]	0.659 ± 0.065 [0.647, 0.671]	0.736 ± 0.054 [0.726, 0.746]
Specificity	0.720 ± 0.062 [0.708, 0.731]	0.724 ± 0.064 [0.712, 0.736]	0.685 ± 0.062 [0.673, 0.697]	0.703 ± 0.061 [0.692, 0.714]
AUC	0.741 ± 0.044 [0.732, 0.749]	0.804 ± 0.041 [0.796, 0.812]	0.719 ± 0.049 [0.710, 0.728]	0.794 ± 0.039 [0.787, 0.801]
# Features	**22**	**11**	**16**	**15**

performance results achieved by SVM and RF models when considering either clinical/laboratory features or only radiomic features.

Following the preliminary selection, three feature selection strategies that combine both clinical/laboratory and radiomic features were implemented to assess the multimodal model.

- *Selection Strategy 1*: in this case, 22 radiomic and 11 clinical/laboratory features were considered for SVM and 16 radiomic and 15 clinical/laboratory for RF.
- *Selection Strategy 2*: SFFS was applied on all the clinical/laboratory (23) and all radiomic (40) features.
- *Selection Strategy 3*: In this case, 22 radiomic and 23 clinical/laboratory for SVM, 16 radiomic and 23 clinical/laboratory for RF.

Table 4.11 provides a summary of the multimodal training and validation performance results obtained for the three feature selection strategies, using a 20-repeated stratified 10-fold cross-validation approach. As anticipated, the combined use of clinical laboratory and radiomic features leads to improved model performance when compared to unimodal models, as observed in previous studies [297, 298].

Table 4.11: Performance obtained in the training/validation phase by the SVM and RF classifiers, with the 10-fold stratified CV procedure (20 repetitions were performed). For each metric, the *mean value* ± *standard deviation* and the confidence interval are reported. (C/L is for clinical/laboratory)

	SVM			**RF**		
Evaluation	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**
Accuracy	0.748 ± 0.040 [0.741, 0.755]	0.755 ± 0.039 [0.748, 0.762]	0.760 ± 0.036 [0.753, 0.768]	0.741 ± 0.040 [0.733, 0.748]	0.746 ± 0.041 [0.738, 0.754]	0.746 ± 0.042 [0.738, 0.754]
Sensitivity	0.741 ± 0.060 [0.730, 0.752]	0.768 ± 0.051 [0.758, 0.778]	0.781 ± 0.052 [0.771, 0.790]	0.743 ± 0.054 [0.733, 0.753]	0.753 ± 0.059 [0.742, 0.764]	0.747 ± 0.057 [0.736, 0.758]
Specificity	0.755 ± 0.053 [0.745, 0.765]	0.742 ± 0.060 [0.731, 0.753]	0.738 ± 0.056 [0.728, 0.748]	0.738 ± 0.064 [0.726, 0.750]	0.740 ± 0.061 [0.729, 0.751]	0.745 ± 0.061 [0.734, 0.756]
AUC	0.803 ± 0.041 [0.795, 0.811]	0.816 ± 0.041 [0.808, 0.824]	0.827 ± 0.035 [0.820, 0.834]	0.812 ± 0.040 [0.805, 0.819]	0.813 ± 0.042 [0.805, 0.821]	0.815 ± 0.039 [0.808, 0.822]
Features (C/L, radiomic)	**18** (5, 13)	**38** (16, 22)	**21** (14, 7)	**21** (12, 9)	**38** (17, 21)	**21** (14, 7)

Table 4.12: Performance obtained in the testing phase by the SVM and RF classifiers.

	SVM			**RF**		
Metrics	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**
Accuracy	0.707	0.720	0.709	0.705	0.697	0.706
Sensitivity	0.700	0.794	0.794	0.727	0.688	0.755
Specificity	0.712	0.647	0.624	0.683	0.702	0.656
AUC	0.775	0.783	**0.778**	**0.800**	0.795	0.796

4.4.5 Predictive Models Test

Given their superior performance, the multimodal models were employed for the testing phase. In this context, AUC was used to identify the best predictive model.

Remarkably, the AUC values obtained during the testing phase exhibited only minimal decreases compared to the training and validation phases, demonstrating promising generalization capabilities. Table 4.12 presents a summary of the results achieved in the testing phase for both SVM and RF models. RF outperformed SVM, and the *Selection Strategy 1* yielded the highest performance, attaining an AUROC of 0.800. In conclusion, it can be inferred that the Random Forest model, coupled with *Selection Strategy 1*, ensures the highest level of performance.

4.4.6 Model Inspection and Final Test Performance

To enhance the model's generalization capabilities, additional feature selection was conducted. Specifically, MDA calculated in LOCO mode, was employed to identify and remove features that exhibited distributional drift. Beginning with the top-performing

model (Random Forest + *Selection Strategy 1*), and using the MDA weights computed across the six centers (hospitals), the set of input features was additionally reduced.

Table A.3 provides a list of features with positive weight in more than 3, 4, and 5 centers concurrently. Subsequently, three feature subsets were created, each comprising 17, 11, and 6 features, respectively. These subsets were then used to retrain the models and recompute the test performance. Table 4.13 illustrates the improvements achieved by eliminating features susceptible to distributional drift. Specifically, when using the features with positive weight in 4 centers simultaneously, an accuracy of 0.733 and an AUROC of 0.819 were obtained, compared to an accuracy of 0.705 and an AUROC of 0.800 when not managing distributional drift.

Table 4.13: Performance obtained in the testing phase by the RF classifier after MDA features skimming.

RF	**Positive Weights (th=3+)**	**Positive Weights (th=4+)**	**Positive Weights (th=5+)**
Accuracy	0.710	**0.733**	0.681
Sensitivity	0.711	**0.761**	0.722
Specificity	**0.709**	0.705	0.640
AUC	0.800	**0.819**	0.765

4.5 Discussion and Analysis

In this research, a machine learning model was designed to offer explainable predictions for COVID-19 prognosis. The primary objective of this model is to assist healthcare professionals in distinguishing between various disease progressions. Given the importance of providing explanations for the decision-making process in clinical settings, a multi-level explanation framework was proposed, considering the diverse stakeholders involved in model development and clinical decision-making. These stakeholders include developers, physicians, and patients. The approach incorporates inherently interpretable clinical, laboratory, and radiomic features, enabling introspection into the model's decision-making process. Both global and local explanation methods were proposed to ensure a comprehensive understanding of the model's predictions.

4.5.1 Clinical Validation

The use of the SHAP Tree Explainer [50] facilitated the interpretation and clinical validation of the model's outcomes. Figure 4.5 highlights the selected features with the greatest influence on the trained model. The beeswarm plot generated by SHAP was employed to assess the importance of each feature in the trained model. For each test sample, Shapley values were computed and aggregated in the graph. The features'

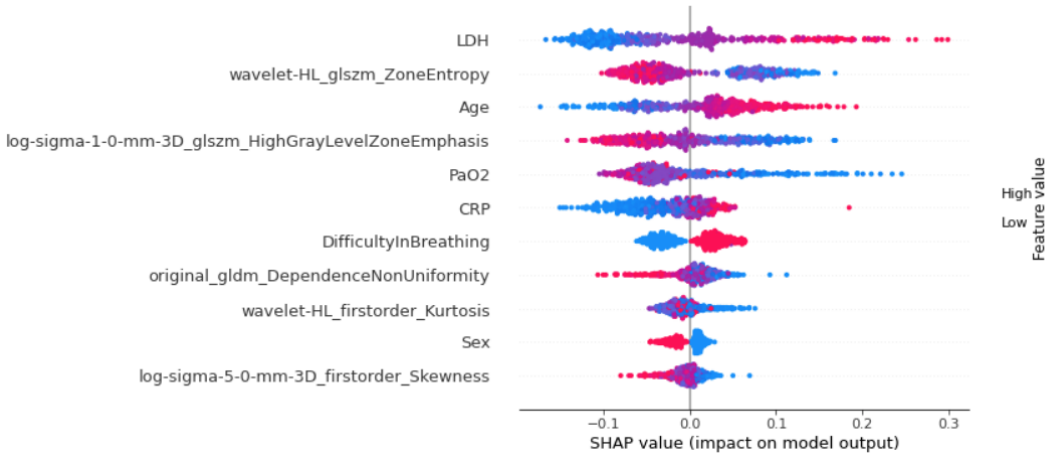


Figure 4.5: SHAP beeswarm plot.

importance is presented in descending order, with LDH emerging as the most influential feature in classification, followed by ZoneEntropy (Wavelet HL, GLSZM), and so forth. The color of the dots on the plot signifies whether feature values are low (blue) or high (red). Dots positioned to the left or right of the vertical line (Shapley value equal to 0.0) indicate that a specific feature tends to guide the model toward a prediction of MILD or SEVERE, respectively. This visualization aids in comprehending the model's decision-making process and the impact of individual features on predictions. In particular, the clinical and laboratory features exhibited significant correspondence with findings commonly applied in clinical practice:

- patients with high values of *Lactate DeHydrogenase Concentration* (LDH) in the blood are generally predisposed to SEVERE diseases, while low values seem to have greater resistance and are limited to MILD diseases [331];
- low values of *Partial Pressure of oxygen* (PaO2) in arterial blood are indicative of SEVERE disease, while high values indicate a MILD level disease [332];
- the clinical evidence confirmed that the high values of *C-Reactive Protein* (CRP) are an indicator of SEVERE disease, while parameter low values indicate a MILD level disease [333];
- male subjects [334], and, naturally, older subjects are more exposed to severe disease (*Sex, Age*).

The prominence of laboratory parameters such as LDH and CRP for the trained model aligns with their well-established association with the most severe forms of COVID-19 disease. These parameters serve as indicators of the inflammatory cascade, making their significance unsurprising. Previous studies have already confirmed their value as

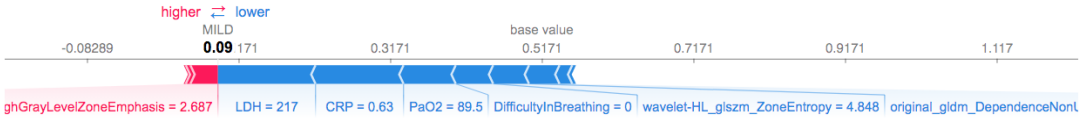
powerful predictors of clinical deterioration. For instance, high levels of serum LDH have been consistently identified as one of the most reliable indicators of clinical worsening in COVID-19 patients. Similarly, elevated CRP levels in the bloodstream have been shown to predict the most unfavorable clinical outcomes in individuals diagnosed with COVID-19. In a recent study [335], both LDH and CRP were identified as contributors to improved diagnostic accuracy for COVID-19 in suspected patients with respiratory symptoms. Another study [336] demonstrated that these laboratory parameters, when combined with radiological features, can effectively predict the need for invasive ventilation in patients with COVID-19 pneumonia. These findings underscore the clinical relevance of the model on LDH and CRP features in assessing COVID-19 prognosis. It is indeed logical to anticipate that low levels of PaO2 are closely associated with severe COVID-19 disease, given that they reflect a state of lung failure. This observation aligns with the early stages of the pandemic, when it became evident that hypoxemia at the time of diagnosis was indicative of the most severe COVID-19 cases, which carried the highest risk of severe respiratory distress and mortality. PaO2, as a lung functional parameter, plays a pivotal role in widely-used predictive scores for identifying COVID-19 patients at risk of acute respiratory failure and mortality [337].

The current study's findings underscore the significance of incorporating lung functional and laboratory parameters into the discriminative approach, as these parameters serve as valuable indicators of hyper-inflammatory processes and lung involvement.

Regarding radiomic features, wavelet-derived and LoG-derived features exhibited strong discriminatory properties. However, the most crucial features were found within the GLSZM (Grey-Level Size Zone Matrix) category, which quantifies the presence of grey-level zones in an image. Here, a grey-level zone is defined as a group of connected pixels that share the same grey-level intensity.

- for *HighGrayLevelZoneEmphasis*, a higher value indicates a greater proportion of higher grey-level values and size zones in the image. In this case, high values mean that the lung is more uniform (with large uniform regions) and no lesions are present;
- for *ZoneEntropy*, SEVERE patients show a more heterogeneous texture. Hence, the behavior of ZoneEntropy is analogous to HighGrayLevelZoneEmphasis in the classification process.

The use of Shapley values has enabled the generation of local explanations to evaluate predictions for individual patients. Figure 4.6 provides an illustrative example of two patients predicted as MILD and SEVERE, respectively. In the graph, features leading the model's prediction toward a MILD outcome are represented in blue, while



(a) Local explanation of a MILD case

(b) Local explanation of a SEVERE case

Figure 4.6: SHAP local explanation.

features leading the model to predict SEVERE are represented in red. In the first case (Figure 4.6a), the patient exhibits normal LDH (217), CRP (0.63), PaO2 (89.5), and no signs of respiratory distress. These factors collectively contribute to the model predicting a MILD prognosis. Conversely, in the second case (Figure 4.6b), the patient presents a high LDH value (680) and a medium/low value of ZoneEntropy, which drives the prediction towards a SEVERE prognosis.

These examples clearly illustrate how explainability transforms a predictive model into a CDSS for physicians, enabling them to comprehensively understand and justify the model's predictions. Such interpretability is invaluable in clinical practice, as it aids in informed decision-making and enhances trust in the model's recommendations.

4.5.2 Performance Discussion and Literature Comparison

In the testing phase, RF trained using the skimmed signature with 4+ positive weights in the LOCO modality achieved an accuracy=0.733 and AUC=0.819. It demonstrates promising generalization capabilities and minimal performance degradation with respect to training/validation performance.

This work can be compared fairly with [102, 309, 308], the only literature works using a subset of the dataset used in this work. In particular, considering only CXR images, in [102] the accuracy obtained (on a subset of 820 cases) with Radiomics was 65.8 ± 1.50 against 74.2 ± 1.0 with deep features; in [309] the best model yielded an accuracy of 73.36 ± 1.95 using deep features. These results improved when clinical features were also considered: [102] obtained an accuracy of 76.9 ± 5.4, while [309] got 77.90 ± 1.27. The accuracy values reported by [102] are those obtained in the training/validation phase. Also in [308] deep architectures were proposed, with the best one giving sensitivity=0.79, specificity=0.82, and AUC=0.84. In summary, the obtained results are promising and in line with the literature on the same dataset or

on a subset[102, 309, 308].

Nevertheless, it is important to highlight the explainability and accuracy trade-off of the solution. In fact, in [102] and [309] deep learning approaches slightly improve performances compared with the trained RF model. However, deep features extracted by CNN do not guarantee a high level of explainability. From a clinical point of view is difficult to correlate the deep features learned with morpho-functional characteristics of a disease found by physicians. Through the use of intrinsically interpretable clinical and radiomic features, the proposed multi-level explanation improves the model's clinical validation.

In addition, as shown in Table A.4, few works focus on explainable solutions. In [102, 308] and [309], saliency maps are used to realize explanation. The decrease in performance (compared with deep approaches proposed in [102, 308, 309]) obtained in this study is reasonable and justifies the choice of improving the explainability for a clinically compliant solution.

Chapter 5

CT radiomic features and clinical biomarkers for predicting coronary artery disease

5.1 Introduction

Epicardial adipose tissue (EAT) is a metabolically active reserve of visceral fat located between the pericardium's cardiac serosa and the myocardium. It includes more than 80% of the cardiac surface, including the right ventricle's free wall, and the atrioventricular and interventricular grooves, and it surrounds the initial segments of the coronary arteries [338]. Numerous studies have demonstrated that EAT serves various roles, such as providing mechanical support to coronary vessels, storing energy due to its high free fatty acid content, and contributing to thermoregulation [339].

Coronary artery disease (CAD) is a leading cause of death and morbidity worldwide [340]. Coronary CT angiography (CCTA) has gained clinical acceptance and currently plays a pivotal role in evaluating CAD. As a non-invasive and cost-effective imaging tool, CCTA holds great promise for reducing the global socioeconomic burden of CAD [341]. The identification of high-risk atherosclerotic plaque markers in CCTA, including low attenuation, positive remodeling, spotty calcification, and the napkin-ring sign, enables highly specific identification of patients at an elevated risk for major adverse cardiac events. These markers are associated with adverse outcomes and can predict ischemia even in non-obstructive lesions [342]. Recent research has focused on analyzing CT attenuation in epicardial and pericoronary adipose tissue as an indirect indicator of coronary atherosclerosis and plaque inflammation [114]. Inflammation is

a critical component of atherosclerosis and a consistent pathological feature of unstable atherosclerotic plaques. Increased CT attenuation in adipose tissue adjacent to an atherosclerotic plaque is considered a marker of inflammation [343].

Various imaging methods have been developed to measure epicardial and pericoronary adipose tissue, including echocardiography, CT, and MRI. Among these, CT stands out due to its higher spatial resolution, enabling a more precise assessment of EAT [344]. Traditionally, quantifying EAT has been a complex task requiring manual measurements performed by highly skilled personnel, using only a fraction of the available data. However, recent advancements have led to the development of accurate and reliable semi-automated software for EAT quantification, such as quartile attenuation analysis. This software has the potential to establish more significant associations between fat characteristics and various clinical scenarios [345].

Quantification methods that rely solely on visible characteristics discernible to the naked eye capture only a fraction of the available information. This limitation often results in a rather simplistic parameter that exhibits significant overlap between patients and healthy controls. However, the field of Radiomics offers a solution by significantly expanding the quantitative information that can be extracted from CT images (Section 2.3.2). These features are used to identify imaging patterns associated with clinical features or outcomes [341]. Leveraging radiomic data has the potential to enhance the diagnostic and predictive capabilities of CCTA, leading to improved risk stratification for future cardiac events [346].

The need for explainable models and interpretable features brings the project design again toward shallow learning approaches. As discussed in Section 2.2, model explainability has become a fundamental requirement in clinical contexts. In fact, some mandatory aspects for clinicians and patients must be considered:

- *clinician's needs*: once the developer deems the model as valid, clinicians can then verify its accuracy against clinical evidence. This validation process helps build trust among clinicians in these computerized systems and promotes their adoption in clinical practice;
- *patient's needs*: A local explanation of the model's result for an individual patient represents how a doctor explains their decisions for a specific clinical case. In this context, a domain expert, typically a clinician, assesses whether the local explanation aligns with their clinical knowledge, making it sensible and valid in the specific patient's context.

To achieve this goal, after establishing and configuring predictive models, the inclusion

of specific features within the identified signatures was analyzed with the guidance of a physician's team.

The objective of this research is to construct predictive models for CAD prediction by leveraging a combination of clinical and radiomic features. Through the use of shallow learning algorithms, it is possible to generate patterns that are interpretable. Additionally, feature selection techniques were applied to identify the most robust predictive signature [347].

The key contributions of this study are as follows:

- a well-structured processing pipeline, according to the literature indications [150], enabling the definition of robust biomarkers;
- the implementation of multimodal predictive models, based on both clinical and radiomic features, able to predict CAD;
- to provide a trusted system supporting cognitive and decision-making processes [348] in the medical domain by means of machine learning algorithms and interpretable clinical and radiomic features.

5.2 Materials and Methods

For a comprehensive overview of the processing pipeline, refer to Figure 5.1. Furthermore, Figure 5.2 provides a breakdown of the 'alternatives' implemented for each of the processing steps, such as feature selection methods and machine learning classifiers.

5.2.1 Dataset Description

The dataset used in this study comprises 118 CCTA series that were collected between October 2019 and January 2020 at the Policlinico University Hospital 'Paolo Giaccone' of Palermo, Italy. Initially, the dataset consisted of 135 cases, which were subject to an initial assessment by two radiologists with over 10 years of experience. These radiologists evaluated the cases based on criteria such as image quality. Subsequently, 17 CCTA series were excluded from the study due to poor quality, characterized by low opacification of coronary arteries and motion artifacts.

The final dataset consists of 84 male and 34 female individuals, with an average age of 60.33 ± 13.2 years. These individuals were categorized into two groups: 'without CAD' (40 cases) and 'with CAD' (78 cases).

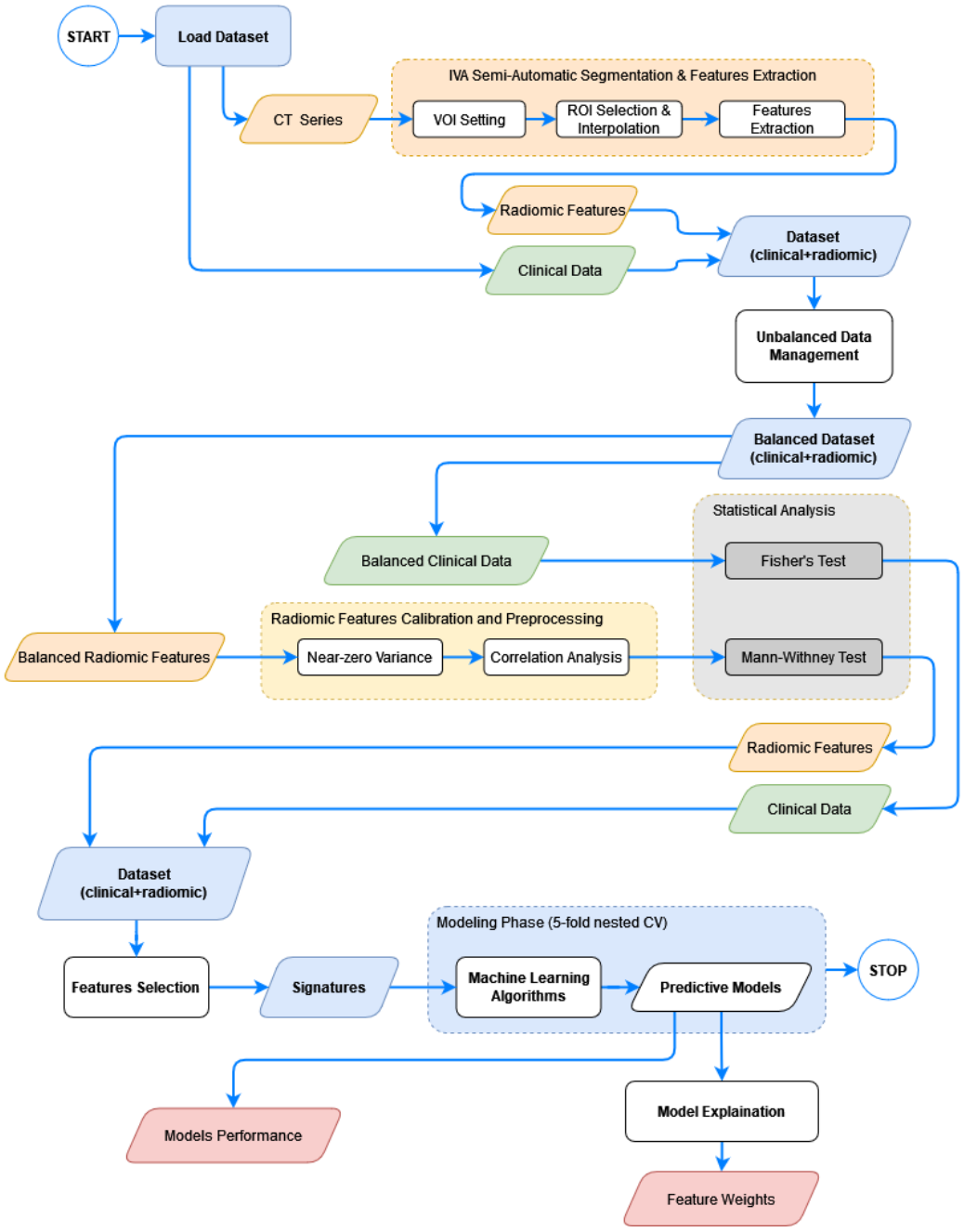


Figure 5.1: Overall flow diagram depicting the whole processing pipeline implemented in this study.

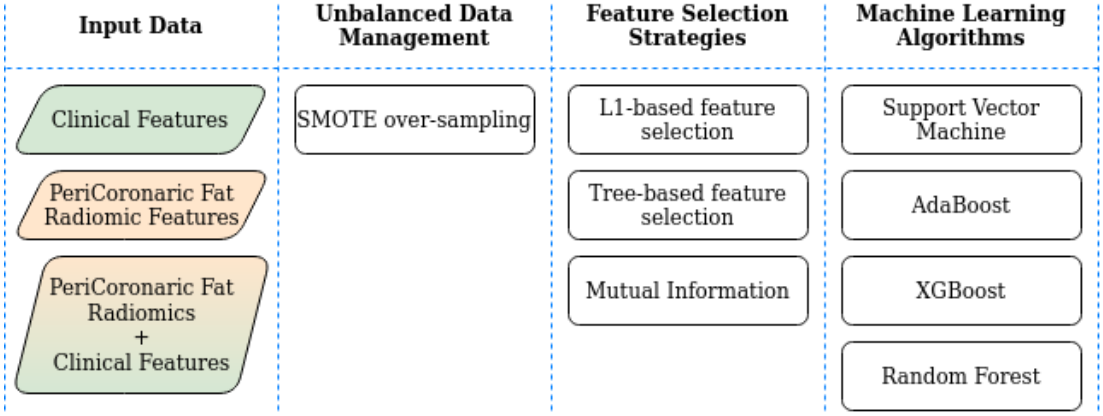


Figure 5.2: Processing alternatives of the crucial pipeline steps.

5.2.2 Clinical Features

The following clinical features were considered in this study: *age*, *sex*, *body-mass index (BMI)*, *family history*, *smoking*, *diabetes*, *hypertension*, *cholesterol*, *obesity*, *current hypertension*, *statin treatment*, *peripheral vasculopathy*, *prior acute myocardial infarction (AMI)*.

5.2.3 Pericoronaric Adipose Tissue Segmentation

The pericoronary fat around the anterior interventricular artery (IVA) VOI was considered for feature extraction. This precise choice of a specific area, the IVA, is also justified by the need to ensure the reproducibility of the study.

To facilitate this, a semi-automatic computer-assisted tool was implemented using Matlab. This tool simplifies the detection of a cylindrical region around the IVA through a few straightforward steps. Here's the process:

1. It starts by selecting the volume of interest (VOI) containing VAT. This is done by drawing a rectangle on the slice where the IVA is most clearly visible along its axis. This rectangle defines an area with dimensions (x, y). Along the z-axis, the greater of these two dimensions, either x or y, is chosen as the dimension for the VOI. As a result, a parallelepiped with dimensions $(x, y, \max(x, y))$ is established in space, forming the VOI around the IVA.

2. After identifying the VOI, at intervals of every *stepROI* slice, the operator inserts a circular ROI centered on the IVA.

3. Once the initial ROIs are manually drawn, the system takes over and automatically interpolates these ROIs onto the remaining slices within the defined range of interest.

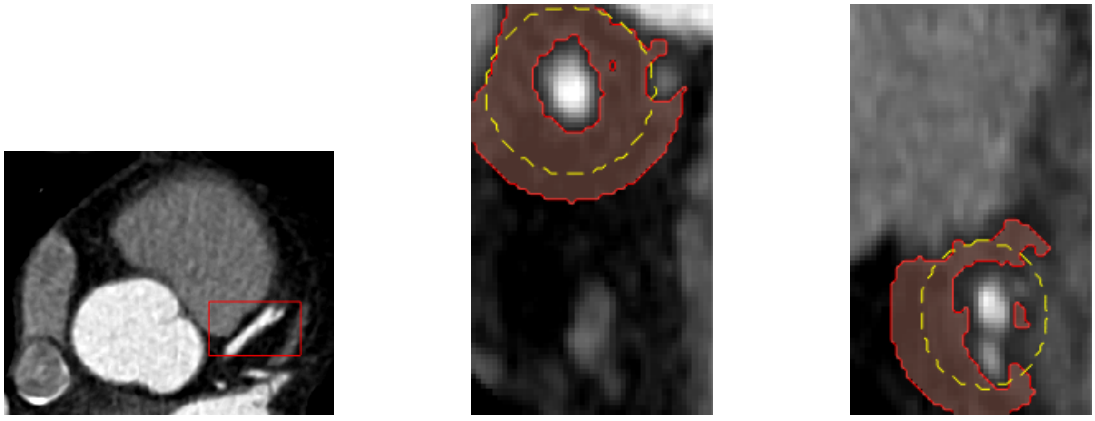
This interpolation process significantly reduces the number of slices that need manual ROI drawing, and it is inspired by the method proposed in [345].

Algorithm 1 describes the semi-automatic pseudocode. Moreover, Figure 5.3 shows the initial step concerning VOI and ROI setting. The three views (a, b, c) and the 3D volume-rendering model (d) of the segmented pericoronaric adipose tissue are depicted in Figure 5.4.

Algorithm 1 Pericoronaric Adipose Tissue Segmentation

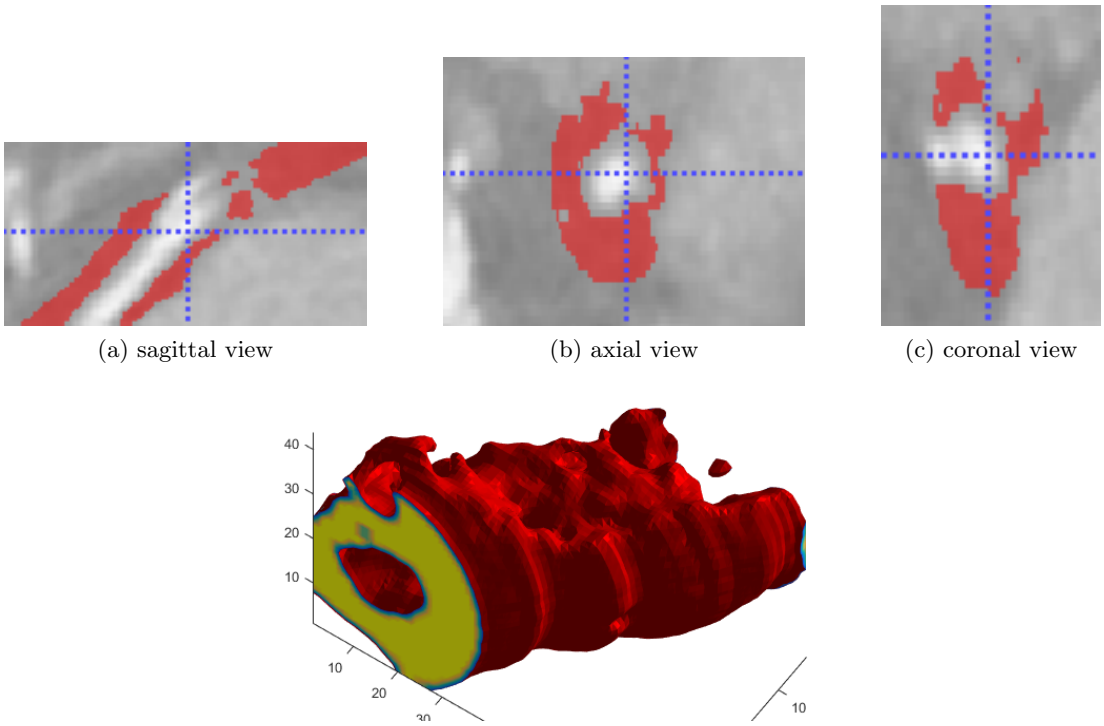
Input: Volume of Interest (VOI) containing the IVA
Output: segmentation mask of the pericoronaric adipose issue within the VOI

1: selection of the CT slice where the IVA is most visible along its axis.
2: selection of the VOI containing the IVA. The user selects the VOI by inserting an interactive bounding-box (see Figure 5.3a) in the slice displayed. (The size of the rectangle will be the size of the VOI on the (x, y) plane. With respect to the $z - axis$, the size of the VOI will be the maximum between x and y.)
3: extraction of the VOI (i.e., a parallelepiped) with dimensions $(x, y, \max(x, y))$
4: from the first (Figure 5.3b) the last (Figure 5.3c) slice of the VOI, enter ROIs every $stepROI$ slices. Jointly with the radiologists, was chosen $stepROI = 5$.
5: from the first to the last slice of the VOI, where the ROIs were not manually placed, they are automatically determined by interpolation.
6: segmentation of adipose tissue around IVA (and contained within ROIs) belonging to the range [-175,-15] (values expressed in Hounsfield Units).
7: saving the VOI and segmentation mask in NIfTI format.



(a) VOI selection (b) initial ROI (c) final ROI

Figure 5.3: In **(a)** selection of the VOI in the slice where the IVA is most visible. In **(b)** and **(c)** the ROIs are inserted around the IVA in the initial and the final slices, respectively.



(a) sagittal view (b) axial view (c) coronal view

(d) 3D volume-rendering

Figure 5.4: In **a)**, **b)** and **c)** the three views of the segmented adipose tissue around the IVA. In **d)** the corresponding 3D volume-rendering reconstruction of the pericoronaric adipose tissue around the IVA.

5.2.4 Radiomic Features Extraction

The extraction of the radiomic features was done by means of PyRadiomics [85]: a total of 93 features were extracted. The extraction was performed without any resampling to avoid interpolation artifacts. Radiomic features were extracted from the 3D ROIs delineated in the previous step. The following five feature categories were extracted and considered (detailed discussed in 2.3.2): GLCM, GLRLM, GLSZM, GLDM, NGTDM.

Given that the ROI extracted around the IVA has a cylindrical shape, shape-based features were excluded. This decision was based on the fact that shape-based features are not pertinent to the clinical problem.

5.2.5 Imbalanced Data Management

Dealing with imbalanced datasets in classification tasks can lead to poor performance on the minority class. To mitigate this issue, one approach is to oversample the minority class, creating new instances from existing ones. In this work the SMOTE method was implemented [119].

5.2.6 Radiomic Features Preprocessing and Statistical Analysis

Feature preprocessing is mandatory in order to define robust imaging biomarkers, as discussed in [150] and Section 2.4.1.1. In particular, features with a variance less than 0.01 and a correlation coefficient higher than 0.9 were discarded.

The Mann-Whitney U test and Fisher's exact test were used to test the difference between the variable distributions for continuous and categorical variables respectively. A p-value lower than 0.05 was considered as the threshold for statistical significance.

5.2.7 Features Selection Methods

Incorporating numerous features into a model can increase its complexity and elevate the risk of overfitting during classification. This is because some features may introduce noise that can potentially harm the model's performance.

Several feature selection methods were implemented to define multimodal signatures (e.g., sets of both radiomic and clinical features). Furthermore, assessments by considering only clinical or radiomic features separately were conducted, resulting in unimodal signatures. This enables the evaluation of the improvements achieved by using multimodal signatures. In this work. L1-based, tree-based and mutual information feature selections were implemented (described in Section 2.4.1.1).

5.2.8 Modeling Phase

The predictive modeling was performed by exploiting different machine learning algorithms (namely, SVM, Random Forest, AdaBoost and XGBoost). These classifiers were trained and tested within a nested 5-fold cross-validation framework, as described in [349]. The chosen classifiers were trained using the features selected in the previous step to produce binary classification results (i.e., distinguishing between 'with CAD' and 'without CAD' cases). The use of the nested CV allowed to train a classification model where the hyperparameters also need to be optimized. In fact, nested CV estimates the generalization error of the underlying model and its hyperparameter search. Figure 5.5 depicts the nested 5-fold cross-validation approach adopted in this study.

To evaluate model performance accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and area under the curve (AUROC) were considered.

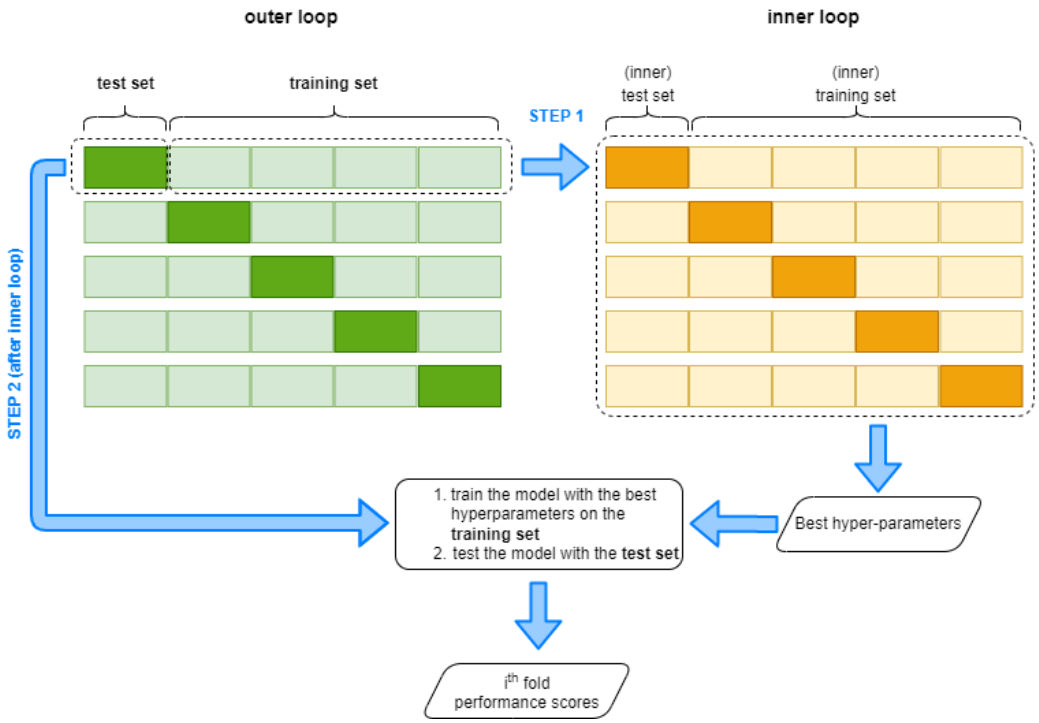


Figure 5.5: Diagram depicting the nested 5-fold cross-validation approach used in this study.

5.3 Experimental Results

The conducted experiments had the goal of assessing the effectiveness of the constructed predictive models for characterizing coronary artery disease.

5.3.1 Features selection and Modeling

Table 5.1 shows details concerning steps implemented for calibration and preprocessing of radiomic and clinical features.

Table 5.1: Radiomic and clinical features preprocessing.

Step	**Analysis Method**	**Remaining Features**
initial radiomic features	n.a.	93
near-zero variance analysis	$variance \leq 0.01$	80
redundant features analysis	Spearman ($cutoff = 0.9$)	42
statistical analysis (radiomic features)	Mann-Whitney U rank test ($p < 0.05$)	30
initial clinical features	n.a.	12
statistical analysis (clinical features)	Fischer's test ($p < 0.05$)	11

The feature selection phase resulted in the identification of unimodal signatures, which consist of only one type of features (either clinical or radiomic), as well as multimodal signatures that incorporate both radiomic and clinical features. These signatures were selected to serve as inputs for the machine learning algorithms, as presented in Table A.5. To start, a 'discovery' phase aimed at identifying the optimal machine learning algorithm was conducted. During this discovery phase, 10 repetitions of nested 5-fold cross-validation were performed, assessing the accuracy of the predictive models exclusively. The results of this phase can be found in Table 5.2.

Table 5.2: Accuracy values obtained in the modeling phase considering the machine learning algorithms and the feature selection methods used.

Feature Selection	**Accuracy**			
Method	**SVM**	**AdaBoost**	**RF**	**XGBoost**
L1-based	0.687 ± 0.069	0.683 ± 0.072	0.730 ± 0.070	0.690 ± 0.073
Tree-based	0.693 ± 0.072	0.692 ± 0.082	0.744 ± 0.075	0.710 ± 0.072
Mutual Information	0.717 ± 0.067	0.678 ± 0.086	0.724 ± 0.084	0.700 ± 0.074

Subsequently, 100 repetitions were computed to calculate all other relevant metrics for the best predictive model identified during the discovery phase, as summarized in Table 5.3.

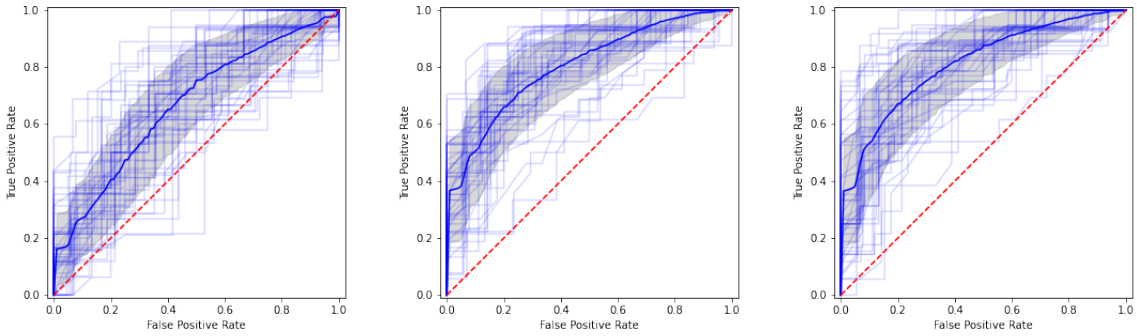
Figure 5.6 illustrates the ROC curves obtained when considering both unimodal and multimodal signatures, with more specific details for *a)* exclusively clinical features, *b)* solely radiomic features, and *c)* a combination of clinical and radiomic features. In the figure, the thicker blue curve represents the ROC curve averaged across the 100 repetitions of cross-validation. The lighter blue, thinner curves represent the individual ROC curves for some of the CV repetitions. The transparent gray band surrounding the ROC curve depicts the standard deviation.

5.3.2 Model Explainability

The notion of explainability extends beyond the mere identification of a signature, which is essentially a set of biomarkers capable of predicting clinical outcomes such as diagnosis, prognosis, or treatment response. In the subsequent discussion, the connection between trends in features is elucidated. The approach involves: *i)* Quantifying the significance of each feature's contribution to the final model's decision-making process; *ii)* Providing clinical justifications for why the identified features hold discriminative value for the task. This analysis was focused on the best machine learning classifier (i.e., Random Forest). In particular, the feature importance was computed by the accumulation of the Mean Decrease in Impurity (MDI) within each Decision Tree composing the

Table 5.3: Performance obtained by the Random Forest model considering the 3 features selection methods.

Feature Type	**Metric** ($mean \pm stdDev$)	**L1-based**	**Tree-based**	**Mutual Information**
clinical	Accuracy	0.628 ± 0.084	0.628 ± 0.084	0.626 ± 0.076
	Sensitivity	0.642 ± 0.133	0.642 ± 0.133	0.646 ± 0.121
	Specificity	0.621 ± 0.128	0.621 ± 0.128	0.613 ± 0.118
	PPV	0.630 ± 0.115	0.630 ± 0.115	0.626 ± 0.109
	NPV	0.637 ± 0.124	0.637 ± 0.124	0.636 ± 0.116
	AUROC	0.684 ± 0.088	0.684 ± 0.088	0.666 ± 0.081
radiomic	Accuracy	0.659 ± 0.077	0.720 ± 0.078	0.713 ± 0.078
	Sensitivity	0.691 ± 0.112	0.767 ± 0.112	0.762 ± 0.112
	Specificity	0.635 ± 0.126	0.681 ± 0.130	0.672 ± 0.124
	PPV	0.656 ± 0.115	0.709 ± 0.112	0.700 ± 0.112
	NPV	0.673 ± 0.113	0.747 ± 0.118	0.740 ± 0.118
	AUROC	0.741 ± 0.081	0.819 ± 0.074	0.803 ± 0.076
clinical + radiomic	Accuracy	0.719 ± 0.080	0.735 ± 0.072	0.739 ± 0.079
	Sensitivity	0.740 ± 0.125	0.766 ± 0.113	0.770 ± 0.120
	Specificity	0.708 ± 0.128	0.713 ± 0.122	0.716 ± 0.124
	PPV	0.719 ± 0.117	0.730 ± 0.110	0.733 ± 0.111
	NPV	0.734 ± 0.121	0.755 ± 0.115	0.759 ± 0.121
	AUROC	0.793 ± 0.077	0.819 ± 0.070	0.820 ± 0.076



(a) only clinical features (AUROC=0.666 ± 0.081)

(b) only radiomic features (AUROC=0.803 ± 0.076)

(c) clinical radiomic features (AUROC=0.820 ± 0.076)

Figure 5.6: ROC curves obtained by the best predictive model (Random Forest + Mutual Information) considering unimodal (**a)** and **b)**) and multimodal (**c)**) signatures.

Forest. The mean and standard deviation of the accumulation were calculated by considering the best model (in terms of accuracy) obtained in each of the 100 repetitions of the nested CV.

Based on the MDI analysis, both clinical and radiomic features make contributions to the prediction. As illustrated in Figure 5.7, the feature *"age"*, alongside *"Total Energy"*, *"Gray Level Variance"* (GLV), and *"Gray Level Non-Uniformity Normalized"* (GLNN), emerged as among the most influential discriminative features. It is worth noting that *"age"* can sometimes be considered a confounding factor for classifiers, which is why it is occasionally treated separately from the other features. However, in this study, the "age" feature was included in the same manner as the other features because it holds clinical significance, even though it alone is not sufficient for accurate diagnoses. Indeed, as reported in the literature, epicardial fat characteristics can provide valuable support for predicting coronary artery disease [350]. Furthermore, as indicated by the weights obtained from the MDI analysis, the other clinical features do not carry as much relevance, underscoring the necessity for radiomic features.

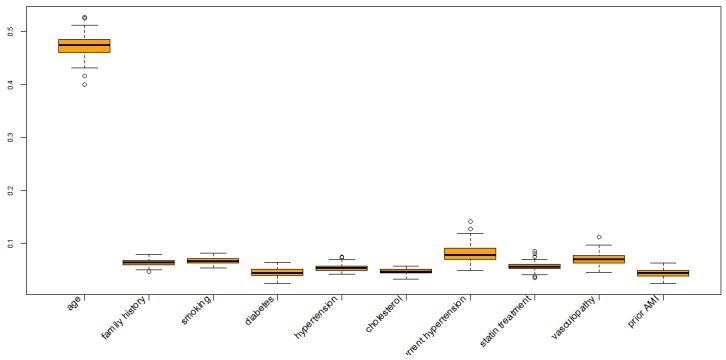
5.4 Discussion

This research has successfully showcased that combinations of radiomic features serve as valuable biomarkers for evaluating the diagnosis of CAD patients. This study was meticulously designed to implement various feature selection methods, including L1-based, tree-based, and mutual information approaches. Additionally, a range of machine learning classifiers, specifically SVM, Random Forest, AdaBoost, and XGBoost were considered, all of which are well-suited for handling small-size datasets.

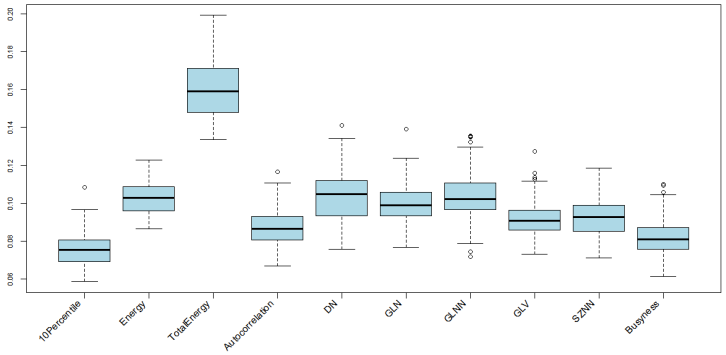
The experimental results demonstrate a substantial enhancement in performance when the multimodal signatures is used, which combine both clinical and radiomic features. Specifically, the best predictive model, employing mutual information and Random Forest, achieved an AUROC of 0.820 ± 0.076. In contrast, the weakest unimodal model, relying solely on clinical risk factors, achieved only an AUROC of 0.666 ± 0.081. Remarkably, the multimodal model exhibited a significant improvement of approximately 23% ($\Delta_{AUROC} = 0.154$). Furthermore, it's worth noting that even using radiomic features alone resulted in improved performance, with an AUROC of 0.803 ± 0.076 compared to relying solely on clinical features.

To provide clinical justifications for the results, the MDI analysis was employed to assess the importance of features in the prediction process. This approach elucidates the experimental findings. According to the MDI weights, the most crucial features are as follows:

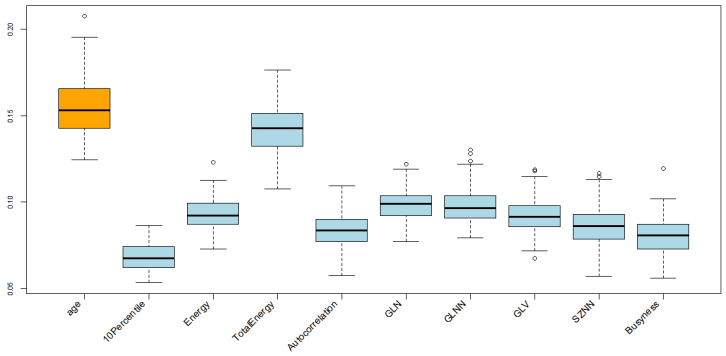
- *Age* (weight=0.131). High values of "age" are associated with a higher likelihood of belonging to the "with CAD" class, while low values of "age" are associated with the "without CAD" class. This observation aligns with clinical intuition, as



(a) only clinical features



(b) only radiomic features



(c) clinical and radiomic features

Figure 5.7: Feature weights (importance) of the signatures composed of **a)** only clinical features, **b)** only radiomic features, and **c)** clinical and radiomic features.

older patients generally have a greater probability of developing CAD over time.

- *Total Energy* (weight=0.176) is derived from "Energy" which measures the magnitude of voxel values within an image. Specifically, "Total Energy" represents

the value of the "Energy" feature scaled by the voxel volume. Larger values of "Total Energy" indicate a greater sum of the squares of these voxel values. In particular, high "Total Energy" values are associated with the "without CAD" class, while low "Total Energy" values are associated with the "with CAD" class. This observation can be explained in the context of extracted pericoronary fat, which typically falls within the range of [-175, -15] in terms of Hounsfield Units (HU). More negative values of HU correspond to more stable clinical conditions of coronary arteries, as indicated in the literature [351]. Therefore, the higher "Total Energy" values, which imply a greater sum of squared voxel values (potentially corresponding to more negative HU values), are indicative of a more stable clinical condition and are associated with the "without CAD" class.

- *Gray Level Variance (GLV)* (weight=0.102) and *Gray Level Non-Uniformity Normalized (GLNN)* (weight=0.098) both are correlated with the variability of gray-level intensity values within the image. A lower value for these features indicates greater homogeneity in intensity values, implying that the image regions have more consistent or uniform gray-level patterns. Interestingly, high "GLV" and "GLNN" values are associated with the "with CAD" class, whereas low values are associated with the "without CAD" class. This behavior appears to align with findings in the literature [352, 353]. Greater inhomogeneity or variability in gray-level intensity values can be indicative of increased loco-regional pericoronary inflammation, which is consistent with the association of these features with the "with CAD" class.

To provide readers with a comprehensive overview of the results achieved by related studies addressing a similar problem, a comparison was conducted and included Table A.6.

In all of the referenced studies [354, 355, 356, 357], an extensive set of radiomic features was used, including both original features and those derived from convolutions with 'LoG' (Laplacian of Gaussian) and 'Wavelets' kernels. This comprehensive approach results in a substantial number of features, potentially exceeding a thousand.

However, it's important to note that a practical guideline in machine learning suggests having at least 5-10 samples (i.e., patients) for each feature in a binary classification model [358, 64]. Consequently, when dealing with a limited number of samples, increasing the number of features, particularly to a large extent, may introduce redundancy among features and exacerbate the curse of dimensionality problem.

Chapter 7

Conclusion

The research conducted in this thesis has aimed to create, develop, and assess innovative computer-assisted tool, designed to assist clinicians in their everyday practice, with a particular emphasis on medical image analysis and explainability issues.

Significant progress in machine learning has primarily been observed within contexts characterized by the availability of large amounts of data. However, within the domain of medicine, data collection and annotation are prohibitively expensive, rendering the architectures suggested in the literature for large datasets unsuitable for direct application in small datasets.

Consequently, the solutions presented here addressed the challenge of building high-performance architectures while preemptively mitigating the risk of overfitting due to the use of limited data. In fact, the availability of well-annotated datasets remains a pivotal concern, and this study explored innovative approaches to face the issue.

The radiomic workflow facilitates the extraction of interpretable biomarkers and paves the way for the development of shallow learning techniques for the analysis of tabular data. It is well-established that shallow architectures demand a smaller volume of training data compared to deep architectures. In fact, applications in breast cancer classification using MRI, Mammography, and Ultrasounds as well as COVID-19 prognosis and Coronary artery disease predictions show that the use of radiomic features and shallow learning techniques allows highly accurate and explainable models. The incorporation of Explainable AI techniques and interpretable biomarkers enabled both global and local explanations while preserving the significance of the model inputs, thereby allowing the model findings study and a comparison with the established clinical literature. Clinical validation of these systems represents a crucial aspect in promoting their acceptance and use, a process that can solely be achieved through the

use of interpretable features and explainable models. Indeed, providing comprehensive explanations plays a pivotal role for developers in the detection of potential undesired outcomes during the inference process. Moreover, it fulfills the essential function of validating and cross-referencing the model's outputs with pre-existing clinical literature. Local explanations, specifically, empower model decisions by enabling those directly impacted, such as patients, to comprehend them, while physicians provide justifications for these decisions.

It has been widely shown that radiomic biomarkers and shallow learning methods maintain an advantage by providing high-performance and highly interpretable models. However, deep models excel at extracting highly informative features, thereby enabling the implementation of more accurate models than those based on shallow learning approaches. Nonetheless, the process of deep feature extraction remains opaque, rendering the meaning of these features unknown. Consequently, the association of these features with clinical findings and the subsequent clinical validation of results become unfeasible. In practice, the explanation of deep features typically relies on the generation of saliency maps, but this approach gives rise to three primary concerns: *i)* saliency maps provide only a local explanation, precluding a global perspective; *ii)* different methods for calculating saliency maps often yield conflicting results, complicating the determination of the most reliable approach; *iii)* saliency maps are qualitative and are subject to inter-operator subjectivity in their evaluation. These findings have been widely discussed in breast cancer detection in mammograms using Yolo and for COVID-19 prognosis prediction. In the first case, it was revealed that the use of Explainable AI methods can enhance the diagnostic performance of the trained model, although several methods yield different saliency maps and offer only local explanations. Furthermore, it was demonstrated that leveraging deep architectures is advisable when there's an opportunity to leverage source databases and apply transfer learning to smaller target datasets. The same results were confirmed when ViT architectures were trained for skin cancer classification and breast cancer classification. In fact, the use of geometric and diffures-data augmentation resulted in significant accuracy improvements. In the second case, for COVID-19 prognosis prediction, it was established that incorporating inherently interpretable features leads to significantly improved trust in these systems and provides valuable clinical insights.

For this reason, the dilemma arises: what is the trade-off between accuracy and explainability? Is it right to optimize one aspect rather than another?

The most recent research on detecting depressed patients has introduced an explainable-by-design model that aims to overcome the explainability/accuracy trade-off. This approach combines the First-Order Logic employed in the old highly-explainable expert

systems, with the capacity of neural networks to uncover relationships within data. By doing so, it enables the use of neural networks without compromising the system's explainability.

Setting aside the critique of XAI methods [375, 14, 16, 17], this thesis empathize the critical significance of architectural decisions, such as features and classifiers, in the development of CDSS. It is essential to consider both explainability and accuracy of equal importance, enabling the creation of Explainable or Transparent CDSS that can be effectively used in real clinical practice.

Appendix A

General Appendix

A.1 Breast Cancer Classification in DCE-MRI

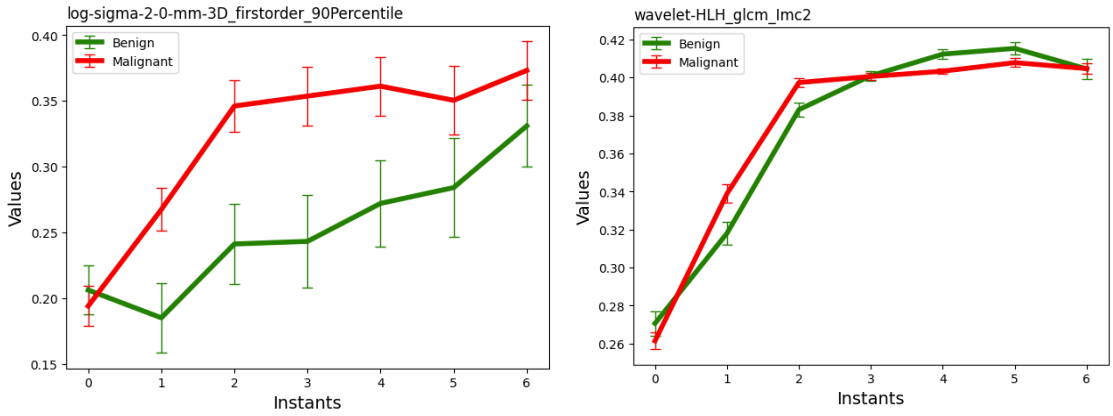


Figure A.1: LoG first-order 90-percentile and wavelet-HLH glcm Imc2 average trends.

Table A.1: Univariate Time series methods results for each radiomic feature (O: original; W: wavelet, L-[2,3]: LoG with σ [2,3]); LDHGLE_LargeDependenceHighGrayLevelEmphasis.

Features	**multirocket**	**rocket**	**TSF**	**STSF**	**Average**
O_firstorder_Energy	0,694	0,717	0,632	0,629	0,668
O_firstorder_TotalEnergy	0,687	0,737	0,636	0,651	0,678
O_glcm_Correlation	0,590	0,547	0,581	0,568	0,571
O_gldm_DependenceEntropy	0,673	0,657	0,596	0,597	0,631
O_ngtdm_Strength	0,664	0,696	0,599	0,648	0,652
L-2_firstorder_90Percentile	0,626	0,671	0,628	0,622	0,637
L-2_gldm_LDHGLE	0,621	0,583	0,580	0,542	0,582
L-2_ngtdm_Busyness	0,679	0,648	0,624	0,603	0,639
L-3_gldm_LDHGLE	0,642	0,641	0,587	0,607	0,619
W-LLH_glcm_Correlation	0,647	0,637	0,523	0,549	0,589
W-LLH_gldm_LDHGLE	0,620	0,637	0,635	0,605	0,624
W-LHL_gldm_LDHGLE	0,638	0,606	0,623	0,637	0,626
W-LHH_gldm_LDHGLE	0,527	0,497	0,571	0,541	0,534
W-LHH_ngtdm_Strength	0,623	0,569	0,624	0,647	0,616
W-HLL_firstorder_Kurtosis	0,590	0,627	0,572	0,566	0,589
W-HLL_gldm_LDHGLE	0,610	0,625	0,554	0,570	0,590
W-HLH_glcm_Imc2	0,676	0,729	0,630	0,677	0,678
W-HLH_gldm_LDHGLE	0,601	0,543	0,578	0,583	0,576
W-HLH_ngtdm_Strength	0,673	0,585	0,609	0,610	0,619
W-HHL_gldm_LDHGLE	0,592	0,596	0,571	0,568	0,582
W-HHH_glcm_Autocorrelation	0,588	0,617	0,546	0,560	0,578
W-HHH_gldm_LargeDependenceEmphasis	0,594	0,578	0,571	0,648	0,598
W-HHH_glszm_SmallAreaEmphasis	0,663	0,507	0,641	0,652	0,616
W-LLL_firstorder_Minimum	0,564	0,569	0,550	0,574	0,564
Average	**0,628**	**0,617**	**0,594**	**0,602**	

Table A.2: Univariate Time series KNN results for each radiomic feature and each distance (O: original; W: wavelet, L-[2,3]: LoG with σ [2,3], LDHGLE: LargeDependenceHighGrayLevelEmphasis, LDE: LargeDependenceEmphasis, SAE: SmallAreaEmphasis).

Features	**DDTW**	**DTW**	**EDR**	**ERP**	**EUC**	**LCSS**	**TWE**	**WDTW**	Average
O_firstorder_Energy	0,616	0,575	0,522	0,592	0,586	0,508	0,565	0,575	**0,583**
O_firstorder_TotalEnergy	0,621	0,583	0,577	0,539	0,552	0,508	0,565	0,584	**0,565**
O_glcm_Correlation	0,526	0,493	0,554	0,509	0,496	0,508	0,498	0,491	**0,592**
O_gldm_DependenceEntropy	0,604	0,578	0,575	0,516	0,499	0,555	0,521	0,578	**0,568**
O_ngtdm_Strength	0,590	0,603	0,639	0,610	0,625	0,635	0,601	0,603	**0,565**
L-2_firstorder_90Percentile	0,571	0,562	0,565	0,555	0,576	0,505	0,551	0,564	**0,509**
L-2_gldm_LDHGLE	0,536	0,532	0,579	0,537	0,563	0,509	0,560	0,537	**0,549**
L-2_ngtdm_Busyness	0,583	0,596	0,628	0,596	0,580	0,604	0,585	0,596	**0,613**
L-3_gldm_LDHGLE	0,638	0,630	0,628	0,602	0,611	0,539	0,554	0,624	**0,558**
W-LLH_glcm_Correlation	0,523	0,507	0,526	0,569	0,522	0,508	0,538	0,505	**0,543**
W-LLH_gldm_LDHGLE	0,608	0,543	0,552	0,510	0,579	0,463	0,545	0,538	**0,594**
W-LHL_gldm_LDHGLE	0,565	0,663	0,591	0,641	0,627	0,581	0,605	0,664	**0,605**
W-LHH_gldm_LDHGLE	0,573	0,456	0,542	0,515	0,460	0,568	0,475	0,461	**0,524**
W-LHH_ngtdm_Strength	0,568	0,567	0,547	0,584	0,578	0,550	0,571	0,585	**0,546**
W-HLL_firstorder_Kurtosis	0,506	0,566	0,552	0,534	0,573	0,551	0,557	0,565	**0,619**
W-LHH_gldm_LDHGLE	0,508	0,552	0,517	0,504	0,506	0,524	0,573	0,558	**0,503**
W-HLH_glcm_Imc2	0,605	0,580	0,677	0,615	0,657	0,508	0,636	0,582	**0,569**
W-HLH_gldm_LDHGLE	0,487	0,580	0,532	0,558	0,558	0,521	0,570	0,574	**0,556**
W-HLH_ngtdm_Strength	0,597	0,606	0,621	0,665	0,650	0,581	0,620	0,611	**0,528**
W-HHL_gldm_LDHGLE	0,524	0,531	0,477	0,594	0,572	0,535	0,554	0,533	**0,612**
W-HHH_glcm_Autocorrelation	0,574	0,549	0,504	0,517	0,515	0,506	0,531	0,553	**0,548**
W-HHH_gldm_LDE	0,559	0,559	0,556	0,598	0,611	0,529	0,576	0,561	**0,621**
W-HHH_glszm_SAE	0,545	0,643	0,568	0,689	0,686	0,507	0,652	0,634	**0,545**
W-LLL_firstorder_Minimum	0,544	0,544	0,476	0,489	0,473	0,488	0,580	0,541	**0,531**
Average	**0,566**	**0,567**	**0,563**	**0,568**	**0,569**	**0,533**	**0,566**	**0,567**	

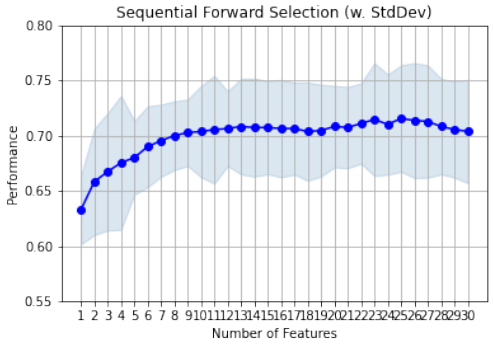
A.2 COVID-19 Prognosis Prediction

A.2.1 Provided Clinical and Laboratory Data

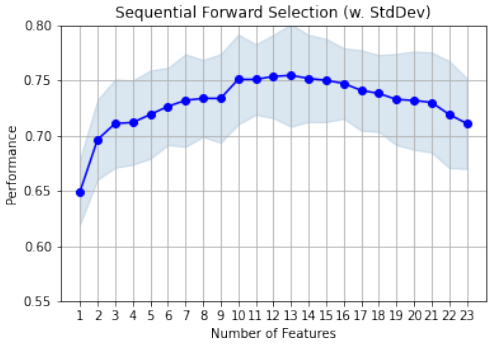
This appendix provides the complete list of clinical and laboratory features associated with CXR images.

In particular, clinical data are: *Hospital*, *Age*, *Sex*, *Positivity at Admission*, *Temperature*, *Days of Fever*, *Cough*, *Difficulty in Breathing*, *Cardiovascular Disease*, *Ischemic Heart Disease*, *Atrial Fibrillation*, *Heart Failure*, *Ictus*, *High Blood Pressure*, *Diabetes*, *Dementia*, *Chronic Obstructive Bronchopneumopathy (BPCO)*, *Cancer*, *Chronic Kidney Disease*, *Respiratory Failure*, *Obesity*, *Position*, *Prognosis*, *Death*.

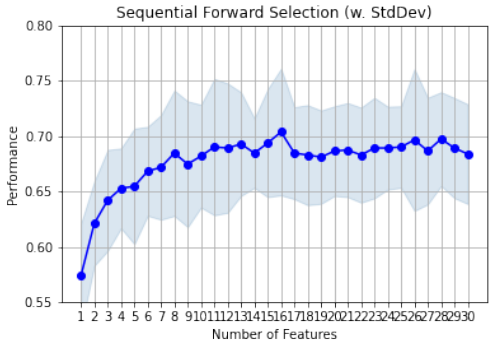
Instead, laboratory data are: *White Blood Cell (WBC)*, *Red Blood Cell (RBC)*, *C-Reactive Protein (CRP)* , *Fibrinogen*, *Glucose* , *Procalcitonin (PCT)*, *Lactate Dehydrogenase (LDH)*, *International Normalized Ratio (INR)*, *D-Dimer*, *Oxigen Percentage*, *Partial Pressure of Oxygen (PaO2)*, *Arterial Oxygen Saturation (SaO2)*, *Partial Pressure of Carbon Dioxide (PaCO2)*, *pH*.



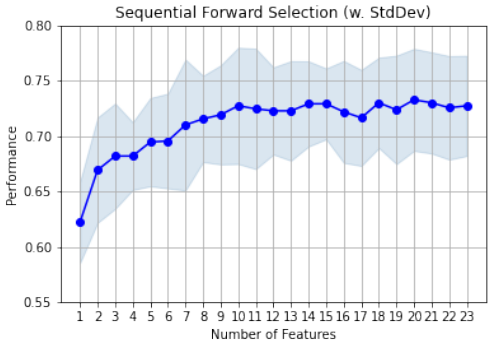
(a) SVM classifier with radiomic features



(b) SVM classifier with clinical features



(c) RF classifier with radiomic features



(d) RF classifier with clinical features

Figure A.2: Accuracy trend obtained in the *preliminary selection* by SVM (in (a) and (b)) and RF (in (c) and (d)) classifiers, during the features selection, performed using SFFS algorithm: in (a) and (c) results on radiomic features; in (b) and (d) results on clinical features.

Table A.3: Starting from the 21 features selected *via* the *Selection Strategy 1* (see Subsection 4.3.6) with the RF classifier, several selection criteria were applied (i.e., th=3+, 4+, and 5+, respectively) to remove distributional drift-affected features. The last three columns refer to the number of positive weights.

Feature Name	**Category**	**Image Type**	**th=3+**	**th=4+**	**th=5+**
Age	n.a.	n.a.	X	X	X
Sex	n.a.	n.a.	X	X	-
DaysFever	n.a.	n.a.	-	-	-
DifficultyInBreathing	n.a.	n.a.	X	X	X
WBC	n.a.	n.a.	X	-	-
RBC	n.a.	n.a.	X	-	-
CRP	n.a.	n.a.	X	X	-
LDH	n.a.	n.a.	X	X	X
PaO2	n.a.	n.a.	X	X	X
Diabetes	n.a.	n.a.	-	-	-
Cancer	n.a.	n.a.	X	-	-
RespiratoryFailure	n.a.	n.a.	-	-	-
Kurtosis	first order	original	X	-	-
DependenceNonUniformity	gldm	original	X	X	-
HighGrayLevelZoneEmphasis	glszm	LoG ($\sigma = 1.0mm$)	X	X	-
Maximum	first order	LoG ($\sigma = 3.0mm$)	X	-	-
Skewness	first order	LoG ($\sigma = 5.0mm$)	X	X	-
Kurtosis	first order	wavelet HL	X	X	X
ZoneEntropy	glszm	wavelet HL	X	X	X
HighGrayLevelZoneEmphasis	glszm	wavelet HH	-	-	-
ZoneEntropy	glszm	wavelet HH	X	-	-
Remaining Features			**17**	**11**	**6**

Table A.4: Literature approaches and comparison. (*) In the *Dataset / Modality / Centers* column, the values between round parenthesis represent the number of images used for training, validation, and testing phases, respectively. (Exp: explainability)

Reference	**Task**	**Dataset (*) / Modality / Centers**	**Features**	**Method**	**Results**	**Exp**
Angeli *et al.* [297]	recovery *vs.* ICU or Death	301 / CT / 1	imaging, demographic, laboratory	LR	AUC=0.841	n.a.
Shiri *et al.* [305]	survival prediction	14339 / CT / 19	radiomic	LR, LASSO, LDA, RF, AdaBoost, Naïve Bayes, MLP	AUC=0.83, sens=0.81, spec=0.72	n.a.
Wang *et al.* [298]	aggravation *vs.* improvement	188 / CT / 1	radiomic, clinical	LR, SVM, DT, RF, XGBoost	AUC=0.843 (radiomic), AUC=0.813 (clinical), AUC=0.865 (combined)	n.a.
Xu *et al.* [303]	early, progressive, severe, or absorption stages	284 / CT / 1	radiomic	SVM	AUC=0.90	n.a.
Shi *et al.* [299]	infection severity	260 / CT / 3	clinical, laboratory, radiomic	LR multivariate	AUC=0.978	n.a.
Borghesi *et al.* [304]	recovery *vs.* death	100 / CXR / 1	Brixia score	weighted Kappa, Mann-Whitney U-test	kw=0.82	n.a.
Signoroni *et al.* [45]	Brixia score prediction	5000 / CXR / n.a.	deep	BS-Net	MAE=0.441	super-pixel maps
Soda *et al.* [102]	mild *vs.* severe	820 / CXR / 6	clinical, radiomic, deep	SVM, LR, RF, MLP, CNNs	acc=0.769±0.054	Grad-CAM
Barbano *et al.* [308]	COVID-positive *vs.* COVID-negative	451 (129+322) / CXR / 1	clinical, deep	fully-connected ANN	AUC=0.84	Grad-CAM
Guarrasi *et al.* [309]	mild *vs.* severe	1103 (820+283) / CXR / 6	clinical, deep	CNNs	acc=73.36±1.95 (only CXR), acc=77.61±1.10 (CXR+clinical)	feature importance, Grad-CAM
Proposed approach	mild *vs.* severe	1589 (820+283+486) / CXR / 6	clinical, radiomic	RF, SVM	AUC=0.819, acc=0.733	multi-level (SHAP analysis)

A.3 Coronary Artery Disease Prediction

Table A.5: Multimodal signatures obtained by the features selection methods considering clinical and radiomic features. Cells containing clinical and radiomic features are highlighted with orange and blue colors, respectively.

Feature Selection Method	Signature	Feature Category
L1-based (1 clinical + 9 radiomic)	age	clinical
	10Percentile	FO
	Mean	FO
	Minimum	FO
	DependenceNonUniformity	GLDM
	LargeDependenceHighGrayLevelEmphasis	GLDM
	LargeDependenceLowGrayLevelEmphasis	GLDM
	GrayLevelNonUniformity	GLSZM
	SizeZoneNonUniformity	GLSZM
	Busyness	NGTDM
Tree-based (4 clinical + 16 radiomic)	age	clinical
	current hypertension	clinical
	statin treatment	clinical
	vasculopathy	clinical
	10Percentile	FO
	90Percentile	FO
	Energy	FO
	Mean	FO
	TotalEnergy	FO
	ClusterShade	GLCM
	LargeDependenceHighGrayLevelEmphasis	GLDM
	LargeDependenceLowGrayLevelEmphasis	GLDM
	GrayLevelNonUniformity	GLSZM
	GrayLevelNonUniformityNormalized	GLSZM
	GrayLevelVariance	GLSZM
	LowGrayLevelZoneEmphasis	GLSZM
	SizeZoneNonUniformity	GLSZM
	SmallAreaHighGrayLevelEmphasis	GLSZM
	SmallAreaLowGrayLevelEmphasis	GLSZM
	ZoneEntropy	GLSZM
Mutual Information (1 clinical + 9 radiomic)	age	clinical
	10Percentile	FO
	Energy	FO
	TotalEnergy	FO
	Autocorrelation	GLCM
	GrayLevelNonUniformity	GLSZM
	GrayLevelNonUniformityNormalized	GLSZM
	GrayLevelVariance	GLSZM
	SizeZoneNonUniformityNormalized	GLSZM
	Busyness	NGTDM

Table A.6: Literature comparison

Literature Work	**Input ROI for Radiomic Features**	**Assessed Outcome**	**Dataset Type/Patients**	**Number of Radiomic Features**	**Machine Learning Classifier**	**Findings**
Kolossvary *et al.* (2017) [354]	coronary artery plaques	napkin-ring sign	CCTA / 30 plaques with napkin-ring sign vs. 30 matched plaques without napkin-ring sign	4440	linear regression	AUC=0.92 (radiomic); AUC=0.77 (conventional parameter)
Lin *et al.* (2020) [355]	PCAT of proximal RCA	myocardial infarction vs. stable CAD vs. No CAD	CCTA / 60 patients with acute myocardial infarction were matched with 60 controls	1103	extreme gradient boosting	AUC=0.76 (clinical); AUC=0.77 (clinical + PCAT atten.); AUC=0.87(clinical + PCAT atten. + radiomic)
Hu *et al.* (2022) [356]	PCAT proximal to the coronary artery	coronary plaques prediction	105 patients with calcified plaques and 95 patients with non-calcified plaques	828	GLN, KNN, SVM, RF, NN	AUC=0.60 (clinical); AUC=0.97 (radiomic); AUC=0.97 (clinical + radiomic)
Kalykakis *et al.* (2022) [357]	circular ROI proximal to the coronary artery	characterizing of functionally significant coronary lesions	CTCA, PET, SPECT / 292 coronary vessels (140 with corresponding PET-MPI data and 152 with SPECT MPI data)	1765	logistic regression	$0.624 \leq AUC \leq 0.816$
Proposed approach	PCAT of proximal IVA	CAD patients vs. no CAD patients	CCTA / 40 CAD patients were matched with 78 controls	93	RF + mutual information for features selection	AUC=0.666 (clinical); AUC=0.803 (radiomic); AUC=0.820 (clinical + radiomic)

Bibliography

[1] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical decision support systems for the practice of evidence-based medicine," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 527–534, 2001.

[2] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[3] A. Giardino, S. Gupta, E. Olson, K. Sepulveda, L. Lenchik, J. Ivanidze, R. Rakow-Penner, M. J. Patel, R. M. Subramaniam, and D. Ganeshan, "Role of imaging in the era of precision medicine," *Academic radiology*, vol. 24, no. 5, pp. 639–649, 2017.

[4] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, "Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework," *Scientific reports*, vol. 7, no. 1, p. 1648, 2017.

[5] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[6] S. Kundu, "Ai in medicine must be explainable," *Nature medicine*, vol. 27, no. 8, pp. 1328–1328, 2021.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Inf. Fusion*, p. 82–115, jun 2020.

[8] A. Smith and F. Director, "Using artificial intelligence and algorithms," *FTC, Apr*, 2020.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[10] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC medical informatics and decision making*, vol. 20, pp. 1–9, 2020.

[11] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artificial Intelligence in Medicine*, vol. 133, p. 102423, 2022.

[12] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[13] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[14] A. M. Bornstein, "Is artificial intelligence permanently inscrutable," *Nautilus*, vol. 40, 2016.

[15] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[16] L. G. McCoy, C. T. Brenna, S. S. Chen, K. Vold, and S. Das, "Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based," *Journal of clinical epidemiology*, vol. 142, pp. 252–257, 2022.

[17] A. J. London, "Artificial intelligence and black-box medical decisions: accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.

[18] M. Jovanović and M. Schmitz, "Explainability as a user requirement for artificial intelligence systems," *Computer*, vol. 55, no. 2, pp. 90–94, 2022.

[19] N. Stogiannos, H. Bougias, E. Georgiadou, S. Leandrou, and P. Papavasileiou, "Analysis of radiomic features derived from post-contrast t1-weighted images and apparent diffusion coefficient (adc) maps for breast lesion evaluation: A retrospective study," *Radiography*, vol. 29, no. 2, pp. 355–361, 2023.

[20] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke, T. K. Gilbert, T. Hagendorff, S. Holm, M. Livne, *et al.*, "To explain or not to explain?—artificial intelligence explainability in clinical decision support systems," *PLOS Digital Health*, vol. 1, no. 2, p. e0000016, 2022.

[21] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[22] D. J. Koehler, "Explanation, imagination, and confidence in judgment," *Psychological bulletin*, vol. 110, no. 3, p. 499, 1991.

[23] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Communications of the ACM*, vol. 62, no. 6, pp. 70–79, 2019.

[24] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

[25] E. Parliament, D.-G. for Parliamentary Research Services, F. Lagioia, and G. Sartor, *The impact of the general data protection regulation on artificial intelligence*. Publications Office, 2021.

[26] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *Ai Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[27] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini, "A workflow for visual diagnostics of binary classifiers using instance-level explanations," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172, IEEE, 2017.

[28] M. Pocevičiūtė, G. Eilertsen, and C. Lundström, "Survey of xai in digital pathology," *Artificial intelligence and machine learning for digital pathology: state-of-the-art and future challenges*, pp. 56–88, 2020.

[29] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review," *npj Digital Medicine*, vol. 5, no. 1, p. 156, 2022.

[30] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[31] G. Alicioglu and B. Sun, "A survey of visual analytics for explainable artificial intelligence methods," *Computers & Graphics*, vol. 102, pp. 502–520, 2022.

[32] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

[33] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[36] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[37] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 international conference on computer vision*, pp. 2018–2025, IEEE, 2011.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014.

[39] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[42] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847, IEEE, 2018.

[43] X.-H. Li, Y. Shi, H. Li, W. Bai, C. C. Cao, and L. Chen, "An experimental study of quantitative evaluations on saliency methods," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, (New York, NY, USA), p. 3200–3208, Association for Computing Machinery, 2021.

[44] Y. Oh, S. Park, and J. C. Ye, "Deep learning covid-19 features on cxr using limited training data sets," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2688–2700, 2020.

[45] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, and D. Farina, "Bs-net: Learning covid-19 pneumonia severity on a large chest x-ray dataset," *Medical image analysis*, vol. 71, p. 102046, 2021.

[46] J. Gu and V. Tresp, "Saliency methods for explaining adversarial attacks," *arXiv preprint arXiv:1908.08413*, 2019.

[47] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[49] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[51] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[52] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[53] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, and S. Melacci, "Logic explained networks," *Artificial Intelligence*, vol. 314, p. 103822, 2023.

[54] F. Sardanelli and F. Podo, eds., *TitleBreast MRI for High-risk Screening*. Springer Cham, 2020.

[55] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[56] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[57] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv preprint arXiv:2012.01805*, 2020.

[58] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[59] Z.-Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, and A. G. Hauptmann, "Rethinking spatial invariance of convolutional networks for object counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19638–19648, 2022.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[63] H. Zhu, B. Chen, and C. Yang, "Understanding why vit trains badly on small datasets: An intuitive perspective," *arXiv preprint arXiv:2302.03751*, 2023.

[64] R. Gillies, P. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563—-577, 2016.

[65] P. Lambin, R. Leijenaar, and T. e. a. Deist, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749—-762, 2017.

[66] S. Ruano, G. Gallego, A. Yezzi, C. Cuevas, and N. García, "Robust image registration with global intensity transformation," in *2015 International Symposium on Consumer Electronics (ISCE)*, pp. 1–2, IEEE, 2015.

[67] S. Bignardi, R. Sandhu, and A. Yezzi, "Radar-based shape and reflectivity reconstruction using active surfaces and the level set method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[68] C. Militello, L. Rundo, M. Dimarco, A. Orlando, I. D'Angelo, V. Conti, and T. V. Bartolotta, "Robustness analysis of dce-mri-derived radiomic features in breast masses: Assessing quantization levels and segmentation agreement," *Applied Sciences*, vol. 12, no. 11, 2022.

[69] C. Xue, J. Yuan, G. G. Lo, A. T. Y. Chang, D. M. C. Poon, O. L. Wong, Y. Zhou, and W. C. W. Chu, "Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 10, 2021.

[70] R. Cattell, S. Chen, and C. Huang, "Robustness of radiomic features in magnetic resonance imaging: review and a phantom study," *Visual computing for industry, biomedicine, and art*, vol. 2, pp. 1–16, 2019.

[71] P. M. Khaniabadi, Y. Bouchareb, H. Al-Dhuhli, I. Shiri, F. Al-Kindi, B. M. Khaniabadi, H. Zaidi, and A. Rahmim, "Two-step machine learning to diagnose and predict involvement of lungs in covid-19 and pneumonia using ct radiomics," *Computers in biology and medicine*, vol. 150, p. 106165, 2022.

[72] F. Arian, M. Amini, S. Mostafaei, K. Rezaei Kalantari, A. Haddadi Avval, Z. Shahbazi, K. Kasani, A. Bitarafan Rajabi, S. Chatterjee, M. Oveisi, I. Shiri, and H. Zaidi, "Myocardial function prediction after coronary artery bypass grafting using mri radiomic features and machine learning algorithms," *Journal of digital imaging*, vol. 35, no. 6, pp. 1708–1718, 2022.

[73] L. H. T. Lam, D. T. Do, D. T. N. Diep, D. L. N. Nguyet, Q. D. Truong, T. T. Tri, H. N. Thanh, and N. Q. K. Le, "Molecular subtype classification of low-grade gliomas using magnetic resonance imaging-based radiomics and machine learning," *NMR in Biomedicine*, vol. 35, no. 11, p. e4792, 2022.

[74] J. Y. Lee, K.-s. Lee, B. K. Seo, K. R. Cho, O. H. Woo, S. E. Song, E.-K. Kim, H. Y. Lee, J. S. Kim, and J. Cha, "Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on mri," *European Radiology*, vol. 32, no. 1, pp. 650–660, 2022.

[75] J. Cheng, C. Ren, G. Liu, R. Shui, Y. Zhang, J. Li, and Z. Shao, "Development of high-resolution dedicated pet-based radiomics machine learning model to predict axillary lymph node status in early-stage breast cancer," *Cancers*, vol. 14, no. 4, p. 950, 2022.

[76] S. Bove, M. C. Comes, V. Lorusso, C. Cristofaro, V. Didonna, G. Gatta, F. Giotta, D. La Forgia, A. Latorre, M. I. Pastena, N. Petruzzellis, D. Pomarico, L. Rinaldi, P. Tamborra, A. Zito, A. Fanizzi, and R. Massafra, "A ultrasound-based radiomic approach to predict the nodal status in clinically negative breast cancer patients," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[77] S. Vicini, C. Bortolotto, M. Rengo, D. Ballerini, D. Bellini, I. Carbone, L. Preda, A. Laghi, F. Coppola, and L. Faggioni, "A narrative review on current imaging applications of artificial intelligence and radiomics in oncology: focus on the three most common cancers," *La radiologia medica*, pp. 1–18, 2022.

[78] G. Carlini, C. Gaudiano, R. Golfieri, N. Curti, R. Biondi, L. Bianchi, R. Schiavina, F. Giunchi, L. Faggioni, E. Giampieri, A. Merlotti, D. Dall'Olio, C. Sala, S. Pandolfi, D. Remondini, A. Rustici, L. V. Pastore, L. Scarpetti, B. Bortolani, L. Cercenelli, E. Brunocilla, E. Marcelli, F. Coppola, and G. Castellani, "Effectiveness of radiomic zot features in the automated discrimination

of oncocytoma from clear cell renal cancer," *Journal of Personalized Medicine*, vol. 13, no. 3, 2023.

[79] F. Giudice, S. Salerno, G. Badalamenti, G. Muto, A. Pinto, M. Galia, F. Prinzi, S. Vitabile, and G. L. Re, "Gastrointestinal stromal tumors: diagnosis, follow-up and role of radiomics in a single center experience," in *Seminars in Ultrasound, CT and MRI*, Elsevier, 2023.

[80] M. Ferro, O. de Cobelli, G. Musi, F. del Giudice, G. Carrieri, G. M. Busetto, U. G. Falagario, A. Sciarra, M. Maggi, F. Crocetto, B. Barone, V. F. Caputo, M. Marchioni, G. Lucarelli, C. Imbimbo, F. A. Mistretta, S. Luzzago, M. D. Vartolomei, L. Cormio, R. Autorino, and O. S. Tătaru, "Radiomics in prostate cancer: An up-to-date review," *Therapeutic Advances in Urology*, vol. 14, 2022.

[81] K. Aftab, F. B. Aamir, S. Mallick, F. Mubarak, W. B. Pope, T. Mikkelsen, J. P. Rock, and S. A. Enam, "Radiomics for precision medicine in glioblastoma," *Journal of neuro-oncology*, pp. 1–15, 2022.

[82] G. Spadarella, T. Perillo, L. Ugga, and R. Cuocolo, "Radiomics in cardiovascular disease imaging: from pixels to the heart of the problem," *Current Cardiovascular Imaging Reports*, pp. 1–11, 2022.

[83] R. Biondi, M. Renzulli, R. Golfieri, N. Curti, G. Carlini, C. Sala, E. Giampieri, D. Remondini, G. Vara, A. Cattabriga, M. A. Cocozza, L. V. Pastore, N. Brandi, A. Palmeri, L. Scarpetti, G. Tanzarella, M. Cescon, M. Ravaioli, G. Castellani, and F. Coppola, "Machine learning pipeline for the automated prediction of microvascularinvasion in hepatocellularcarcinomas," *Applied Sciences*, vol. 13, no. 3, 2023.

[84] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, *et al.*, "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[85] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[86] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[87] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer graphics and image processing*, vol. 4, no. 2, pp. 172–179, 1975.

[88] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognition Letters*, vol. 11, no. 6, pp. 415–419, 1990.

[89] D.-H. Xu, A. S. Kurani, J. D. Furst, and D. S. Raicu, "Run-length encoding for volumetric texture," *Heart*, vol. 27, no. 25, pp. 452–458, 2004.

[90] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Transactions on systems, man, and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, 1989.

[91] G. Thibault, B. FERTIL, C. Navarro, S. Pereira, N. Lévy, J. Sequeira, and J.-L. MARI, "Texture indexes and gray level size zone matrix application to cell nuclei classification," *10th International Conference on Pattern Recognition and Information Processing*, 11 2009.

[92] C. Sun and W. G. Wee, "Neighboring gray level dependence matrix for texture classification," *Computer Vision, Graphics, and Image Processing*, vol. 23, no. 3, pp. 341–352, 1983.

[93] R. Jing, J. Wang, J. Li, X. Wang, B. Li, F. Xue, G. Shao, and H. Xue, "A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[94] J. Zhou, J. Lu, C. Gao, J. Zeng, C. Zhou, X. Lai, W. Cai, and M. Xu, "Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in mri," *BMC cancer*, vol. 20, no. 1, pp. 1–10, 2020.

[95] Z. Hou, Y. Yang, S. Li, J. Yan, W. Ren, J. Liu, K. Wang, B. Liu, and S. Wan, "Radiomic analysis using contrast-enhanced ct: predict treatment response to pulsed low dose rate radiotherapy in gastric carcinoma with abdominal cavity metastasis," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 4, 2018.

[96] K. Kotowski, D. Kucharski, B. Machura, S. Adamski, B. G. Becker, A. Krason, L. Zarudzki, J. Tessier, and J. Nalepa, "Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features," *Computers in Biology and Medicine*, p. 106378, 2022.

[97] S. Bijari, A. Jahanbakhshi, P. Hajishafiezahramini, P. Abdolmaleki, *et al.*, "Differentiating glioblastoma multiforme from brain metastases using multidimensional radiomics features derived from mri and multiple machine learning models," *BioMed Research International*, vol. 2022, 2022.

[98] Z. Jiang, J. Yin, P. Han, N. Chen, Q. Kang, Y. Qiu, Y. Li, Q. Lao, M. Sun, D. Yang, *et al.*, "Wavelet transformation can enhance computed tomography texture features: a multicenter radiomics study for grade assessment of covid-19 pulmonary lesions.," *Quantitative imaging in medicine and surgery*, vol. 12, no. 10, pp. 4758–4770, 2022.

[99] E. Keogh and A. Mueen, *Curse of Dimensionality*. Springer US, 2017.

[100] P. Wei, "Radiomics, deep learning and early diagnosis in oncology," *Emerging topics in life sciences*, vol. 5, no. 6, pp. 829–835, 2021.

[101] C. Militello, L. Rundo, M. Dimarco, A. Orlando, R. Woitek, I. D'Angelo, G. Russo, and T. V. Bartolotta, "3d dce-mri radiomic analysis for malignant lesion prediction in breast cancer patients," *Academic Radiology*, vol. 29, no. 6, pp. 830–840, 2022.

[102] P. Soda, N. C. D'Amico, J. Tessadori, G. Valbusa, V. Guarrasi, C. Bortolotto, M. U. Akbar, R. Sicilia, E. Cordelli, D. Fazzini, *et al.*, "Aiforcovid: Predicting the clinical outcomes in patients with covid-19 applying ai to chest-x-rays. an italian multicentre study," *Medical image analysis*, vol. 74, p. 102216, 2021.

[103] M. R. Chetan and F. V. Gleeson, "Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives," *European radiology*, vol. 31, pp. 1049–1058, 2021.

[104] N. Sushentsev, L. Rundo, O. Blyuss, T. Nazarenko, A. Suvorov, V. J. Gnanapragasam, E. Sala, and T. Barrett, "Comparative performance of mri-derived precise scores and delta-radiomics models for the prediction of prostate cancer progression in patients on active surveillance," *European radiology*, vol. 32, pp. 680–689, 2022.

[105] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, "Overview of radiomics in breast cancer diagnosis and prognostication," *The Breast*, vol. 49, pp. 74–80, 2020.

[106] Q. Sun, X. Lin, Y. Zhao, L. Li, K. Yan, D. Liang, D. Sun, and Z.-C. Li, "Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region," *Frontiers in oncology*, vol. 10, p. 53, 2020.

[107] D. Truhn, S. Schrading, C. Haarburger, H. Schneider, D. Merhof, and C. Kuhl, "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri," *Radiology*, vol. 290, no. 2, pp. 290–297, 2019.

[108] C. S. Lisson, C. G. Lisson, M. F. Mezger, D. Wolf, S. A. Schmidt, W. M. Thaiss, E. Tausch, A. J. Beer, S. Stilgenbauer, M. Beer, *et al.*, "Deep neural networks and machine learning radiomics modelling for prediction of relapse in mantle cell lymphoma," *Cancers*, vol. 14, no. 8, p. 2008, 2022.

[109] J.-P. Fortin, N. Cullen, Y. I. Sheline, and et al., "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, 2018.

[110] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, pp. 118–127, 04 2006.

[111] S. Raschka, "Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack," *The Journal of Open Source Software*, vol. 3, Apr. 2018.

[112] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, *et al.*, "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[113] Q. Niu, X. Jiang, Q. Li, Z. Zheng, H. Du, S. Wu, and X. Zhang, "Texture features and pharmacokinetic parameters in differentiating benign and malignant breast lesions by dynamic contrast enhanced magnetic resonance imaging," *Oncology Letters*, vol. 16, no. 4, pp. 4607–4613, 2018.

[114] E. K. Oikonomou, M. C. Williams, C. P. Kotanidis, M. Y. Desai, M. Marwan, A. S. Antonopoulos, K. E. Thomas, S. Thomas, I. Akoumianakis, L. M. Fan, *et al.*, "A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary ct angiography," *European heart journal*, vol. 40, no. 43, pp. 3529–3543, 2019.

[115] Q. Zhang, Y. Peng, W. Liu, J. Bai, J. Zheng, X. Yang, and L. Zhou, "Radiomics based on multimodal mri for the differential diagnosis of benign and malignant breast lesions," *Journal of Magnetic Resonance Imaging*, vol. 52, no. 2, pp. 596–607, 2020.

[116] S. Parab and S. Bhalerao, "Choosing statistical test," *International journal of Ayurveda research*, vol. 1, no. 3, p. 187, 2010.

[117] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, "Feature selection via mutual information: New theoretical insights," in *2019 International Joint Conference on Neural Networks (IJCNN)*, (Budapest, Hungary, July 14–19, 2019), pp. 1–9, IEEE, 2019.

[118] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–16, 2009.

[119] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, p. 321–357, jun 2002.

[120] S. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder, "Breast cancer risk prediction using xgboost and random forest algorithm," in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pp. 1–4, IEEE, 2020.

[121] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Computers in Biology and Medicine*, vol. 128, p. 104089, 2021.

[122] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[123] V. Di Stefano, F. Prinzi, M. Luigetti, M. Russo, S. Tozza, P. Alonge, A. Romano, M. A. Sciarrone, F. Vitali, A. Mazzeo, L. Gentile, G. Palumbo, F. Manganelli, S. Vitabile, and F. Brighina,

"Machine learning for early diagnosis of attrv amyloidosis in non-endemic areas: A multicenter study from italy," *Brain Sciences*, vol. 13, no. 5, 2023.

[124] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[125] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, pp. 1–13, 2018.

[126] J. Premaladha and K. Ravichandran, "Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms," *Journal of medical systems*, vol. 40, pp. 1–12, 2016.

[127] P. Kharazmi, J. Zheng, H. Lui, Z. Jane Wang, and T. K. Lee, "A computer-aided decision support system for detection and localization of cutaneous vasculature in dermoscopy images via deep feature learning," *Journal of medical systems*, vol. 42, pp. 1–11, 2018.

[128] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[129] D. H. Hubel, "Single unit activity in striate cortex of unrestrained cats," *The Journal of physiology*, vol. 147, no. 2, p. 226, 1959.

[130] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021.

[131] J. Coady, A. O'Riordan, G. Dooly, T. Newe, and D. Toal, "An overview of popular digital image processing filtering operations," in *2019 13th International conference on sensing technology (ICST)*, pp. 1–5, IEEE, 2019.

[132] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 3320–3328, MIT Press, 2014.

[133] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264, IGI global, 2010.

[134] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[135] A. Niculescu-Mizil and R. Caruana, "Inductive transfer for bayesian network structure learning," in *Artificial intelligence and statistics*, pp. 339–346, PMLR, 2007.

[136] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey.," *Journal of Machine Learning Research*, vol. 10, no. 7, 2009.

[137] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[138] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[139] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021.

[140] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022.

[141] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete, "A survey on modern trainable activation functions," *Neural Networks*, vol. 138, pp. 14–32, 2021.

[142] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.

[143] G. V. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[144] F. Prinzi, T. Currieri, S. Gaglio, and S. Vitabile, "Shallow and deep learning classifiers in medical image analysis," *European Radiology Experimental*, Under Review.

[145] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[146] M. Kamal, A. R. Pratap, M. Naved, A. S. Zamani, P. Nancy, M. Ritonga, S. K. Shukla, F. Sammy, *et al.*, "Machine learning and image processing enabled evolutionary framework for brain mri analysis for alzheimer's disease detection," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[147] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis c virus infection," *Intelligent Medicine*, vol. 2, no. 04, pp. 193–198, 2022.

[148] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[149] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[150] N. Papanikolaou, C. Matos, and D. M. Koh, "How to develop a meaningful radiomic signature for clinical use in oncologic patients," *Cancer Imaging*, vol. 20, no. 1, pp. 1–10, 2020.

[151] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False discovery rates in pet and ct studies with texture features: a systematic review," *PloS one*, vol. 10, no. 5, p. e0124165, 2015.

[152] X. Xie, M. Yang, S. Xie, X. Wu, Y. Jiang, Z. Liu, H. Zhao, Y. Chen, Y. Zhang, and J. Wang, "Early prediction of left ventricular reverse remodeling in first-diagnosed idiopathic dilated cardiomyopathy: a comparison of linear model, random forest, and extreme gradient boosting," *Frontiers in cardiovascular medicine*, vol. 8, p. 684004, 2021.

[153] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philosophy & Technology*, vol. 31, pp. 611–627, 2018.

[154] M. Theunissen and J. Browning, "Putting explainable ai in context: institutional explanations for medical ai," *Ethics and Information Technology*, vol. 24, no. 2, p. 23, 2022.

[155] J. Zhang, H. Chao, M. K. Kalra, G. Wang, and P. Yan, "Overlooked trustworthiness of explainability in medical ai," *medRxiv*, pp. 2021–12, 2021.

[156] F. Prinzi, A. Orlando, S. Gaglio, M. Midiri, and S. Vitabile, "Ml-based radiomics analysis for breast cancer classification in dce-mri," in *International Conference on Applied Intelligence and Informatics*, pp. 144–158, Springer, 2022.

[157] F. Prinzi, A. Orlando, S. Gaglio, and S. Vitabile, "Breast cancer classification through multivariate radiomic time series analysis in dce-mri sequences," *Expert Systems with Applications*, Under Review.

[158] F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A yolo-based model for breast cancer detection in mammograms," *Cognitive Computation*, pp. 1–14, 2023.

[159] F. Prinzi, M. Insalaco, S. Gaglio, and S. Vitabile, "Breast cancer localization and classification in mammograms using yolov5," in *Applications of Artificial Intelligence and Neural Systems to Data Science*, pp. 73–82, Springer, 2023.

[160] F. Prinzi, A. Orlando, S. Gaglio, and S. Vitabile, "Interpretable radiomic signature for breast microcalcification detection and classification," *Journal of Digital Imaging*, Under Review.

[161] S. Cannata, G. Cicceri, G. Cirrincione, T. Currieri, M. Lovino, C. Militello, F. Prinzi, E. Pasero, and S. Vitabile, "Vit-based classification of mammogram images: Impact of data augmentation techniques," in *the Italian Workshop on Neural Networks*, In press, 2023.

[162] F. Prinzi, C. Militello, T. V. Bartolotta, and S. Vitabile, "Breast cancer malignancy prediction by means of an explainable model based on a multimodal signature," in *Proceedings of 18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics*, In press, 2023.

[163] T. V. Bartolotta, C. Militello, F. Prinzi, F. Ferraro, L. Rundo, C. Zarcaro, M. Di Marco, A. Orlando, D. Matranga, and S. Vitabile, "Artificial intelligence-based, semi-automated segmentation for the extraction of ultrasound-derived radiomics features in breast cancer: a prospective multicenter study," *La Radiologia Medica*, Under Review.

[164] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[165] S. W. Duffy, L. Tabár, A. M.-F. Yen, P. B. Dean, R. A. Smith, H. Jonsson, S. Törnberg, S. L.-S. Chen, S. Y.-H. Chiu, J. C.-Y. Fann, M. M.-S. Ku, W. Y.-Y. Wu, C.-Y. Hsu, Y.-C. Chen, G. Svane, E. Azavedo, H. Grundström, P. Sundén, K. Leifland, E. Frodis, J. Ramos, B. Epstein, A. Åkerlund, A. Sundbom, P. Bordás, H. Wallin, L. Starck, A. Björkgren, S. Carlson, I. Fredriksson, J. Ahlgren, D. Öhman, L. Holmberg, and T. H.-H. Chen, "Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women," *Cancer*, vol. 126, p. 2971—2979, July 2020.

[166] E. U. Ekpo, M. Alakhras, and P. Brennan, "Errors in mammography cannot be solved through technology alone," *Asian Pacific journal of cancer prevention: APJCP*, vol. 19, no. 2, p. 291, 2018.

[167] American College of Radiology BI-RADS Committee, *Acr bi-rads atlas: breast imaging reporting and data system*. American College of Radiology, 2013.

[168] J. Xiao, H. Rahbar, D. S. Hippe, M. H. Rendi, E. U. Parker, N. Shekar, M. Hirano, K. J. Cheung, and S. C. Partridge, "Dynamic contrast-enhanced breast mri features correlate with invasive breast cancer angiogenesis," *NPJ breast cancer*, vol. 7, no. 1, p. 42, 2021.

[169] A. R. Padhani, "c," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 16, no. 4, pp. 407–422, 2002.

[170] R. H. El Khouli, K. J. Macura, M. A. Jacobs, T. H. Khalil, I. R. Kamel, A. Dwyer, and D. A. Bluemke, "Dynamic contrast-enhanced mri of the breast: quantitative method for kinetic curve type assessment," *AJR. American journal of roentgenology*, vol. 193, no. 4, p. W295, 2009.

[171] A. Orlando, M. Dimarco, R. Cannella, and T. V. Bartolotta, "Breast dynamic contrast-enhanced-magnetic resonance imaging and radiomics: state of art," *Artificial Intelligence in Medical Imaging*, 2020.

[172] C. K. Kuhl, S. Schrading, C. C. Leutner, N. Morakkabati-Spitz, E. Wardelmann, R. Fimmers, W. Kuhn, and H. H. Schild, "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," *Journal of clinical oncology*, vol. 23, no. 33, pp. 8469–8476, 2005.

[173] Y. Zhang, H. Ren, *et al.*, "Meta-analysis of diagnostic accuracy of magnetic resonance imaging and mammography for breast cancer," *Journal of cancer research and therapeutics*, vol. 13, no. 5, p. 862, 2017.

[174] K. Rautela, D. Kumar, and V. Kumar, "A systematic review on breast cancer detection using deep learning techniques," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 4599–4629, 2022.

[175] M. F. Mridha, M. Hamid, M. M. Monowar, A. J. Keya, A. Q. Ohi, M. Islam, J.-M. Kim, *et al.*, "A comprehensive survey on deep-learning-based breast cancer diagnosis," *Cancers*, vol. 13, no. 23, p. 6116, 2021.

[176] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[177] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[178] P. Gibbs, N. Onishi, M. Sadinski, K. M. Gallagher, M. Hughes, D. F. Martinez, E. A. Morris, and E. J. Sutton, "Characterization of sub-1 cm breast lesions using radiomics analysis," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 5, pp. 1468–1477, 2019.

[179] J. Zhou, Y. Zhang, K.-T. Chang, K. E. Lee, O. Wang, J. Li, Y. Lin, Z. Pan, P. Chang, D. Chow, M. Wang, and M.-Y. Su, "Diagnosis of benign and malignant breast lesions on dce-mri by using radiomics and deep learning with consideration of peritumor tissue," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 3, pp. 798–809, 2020.

[180] V. S. Parekh and M. A. Jacobs, "Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric mri," *NPJ breast cancer*, vol. 3, no. 1, pp. 1–9, 2017.

[181] M. B. Nagarajan, M. B. Huber, T. Schlossbauer, G. Leinsinger, A. Krol, and A. Wismüller, "Classification of small lesions in breast mri: evaluating the role of dynamically extracted texture features through feature selection," *Journal of medical and biological engineering*, vol. 33, no. 1, 2013.

[182] D. Ravichandran, R. Nimmatoori, and M. Gulam Ahamad, "Mathematical representations of 1d, 2d and 3d wavelet transform for image coding," *Int. J. Adv. Comput. Theory Eng*, vol. 5, pp. 1–8, 2016.

[183] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[184] A. Al Jumah, "Denoising of an image using discrete stationary wavelet transform and various thresholding techniques," *Journal of Signal and Information Processing*, 2013.

[185] Ç. P. Dautov and M. S. Özerdem, "Wavelet transform and signal denoising using wavelet method," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2018.

[186] M. Boix and B. Canto, "Wavelet transform application to the compression of images," *Mathematical and computer modelling*, vol. 52, no. 7-8, pp. 1265–1270, 2010.

[187] R. D. Chitalia and D. Kontos, "Role of texture analysis in breast mri as a cancer biomarker: a review," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 927–938, 2019.

[188] S. Peng, L. Chen, J. Tao, J. Liu, W. Zhu, H. Liu, and F. Yang, "Radiomics analysis of multiphase dce-mri in predicting tumor response to neoadjuvant therapy in breast cancer," *Diagnostics*, vol. 11, no. 11, p. 2086, 2021.

[189] M. Mahrooghy, A. B. Ashraf, D. Daye, C. Mies, M. Feldman, M. Rosen, and D. Kontos, "Heterogeneity wavelet kinetics from dce-mri for classifying gene expression based breast cancer recurrence risk," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pp. 295–302, Springer, 2013.

[190] Y. Li, S. Ammari, L. Lawrance, A. Quillent, T. Assi, N. Lassau, and E. Chouzenoux, "Radiomics-based method for predicting the glioma subtype as defined by tumor grade, idh mutation, and 1p/19q codeletion," *Cancers*, vol. 14, no. 7, p. 1778, 2022.

[191] H. M. Pereira, M. E. Leite Duarte, I. Ribeiro Damasceno, L. A. de Oliveira Moura Santos, and M. H. Nogueira-Barbosa, "Machine learning-based ct radiomics features for the prediction of pulmonary metastasis in osteosarcoma," *The British Journal of Radiology*, vol. 94, no. 1124, p. 20201391, 2021.

[192] F. Orlhac, J. J. Eertink, A.-S. Cottereau, J. M. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, and I. Buvat, "A guide to combat harmonization of imaging biomarkers in multicenter studies," *Journal of Nuclear Medicine*, vol. 63, no. 2, pp. 172–179, 2022.

[193] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.

[194] A. Dempster, D. F. Schmidt, and G. I. Webb, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, (New York, NY, USA), p. 248–257, Association for Computing Machinery, 2021.

[195] C. W. Tan, A. Dempster, C. Bergmeir, and G. I. Webb, "Multirocket: multiple pooling operators and transformations for fast and effective time series classification," *Data Mining and Knowledge Discovery*, pp. 1–24, 2022.

[196] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "Hive-cote 2.0: a new meta ensemble for time series classification," *Machine Learning*, vol. 110, no. 11, pp. 3211–3243, 2021.

[197] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A time series forest for classification and feature extraction," *Information Sciences*, vol. 239, pp. 142–153, 2013.

[198] N. Cabello, E. Naghizade, J. Qi, and L. Kulik, "Fast and accurate time series classification through supervised interval search," in *2020 IEEE International Conference on Data Mining (ICDM)*, (Los Alamitos, CA, USA), pp. 948–953, IEEE Computer Society, nov 2020.

[199] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[200] E. J. Keogh and M. J. Pazzani, *Derivative Dynamic Time Warping*, pp. 1–11. SIAM, 2011.

[201] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern recognition*, vol. 44, no. 9, pp. 2231–2240, 2011.

[202] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proceedings 18th international conference on data engineering*, pp. 673–684, IEEE, 2002.

[203] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, p. 792–803, VLDB Endowment, 2004.

[204] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, (New York, NY, USA), p. 491–502, Association for Computing Machinery, 2005.

[205] P.-F. Marteau, "Time warp edit distance with stiffness adjustment for time series matching," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 306–318, 2008.

[206] N. Aristokli, I. Polycarpou, S. Themistocleous, D. Sophocleous, and I. Mamais, "Comparison of the diagnostic performance of magnetic resonance imaging (mri), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review," *Radiography*, 2022.

[207] L. Zhang, M. Tang, Z. Min, J. Lu, X. Lei, and X. Zhang, "Accuracy of combined dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging for breast cancer detection: a meta-analysis," *Acta radiologica*, vol. 57, no. 6, pp. 651–660, 2016.

[208] H. Dong, L. Kang, S. Cheng, and R. Zhang, "Diagnostic performance of dynamic contrast-enhanced magnetic resonance imaging for breast cancer detection: an update meta-analysis," *Thoracic Cancer*, vol. 12, no. 23, pp. 3201–3207, 2021.

[209] R. M. Mann, C. K. Kuhl, K. Kinkel, and C. Boetes, "Breast mri: guidelines from the european society of breast imaging," *European radiology*, vol. 18, pp. 1307–1318, 2008.

[210] G. Agrawal, M.-Y. Su, O. Nalcioglu, S. A. Feig, and J.-H. Chen, "Significance of breast lesion descriptors in the acr bi-rads mri lexicon," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 115, no. 7, pp. 1363–1380, 2009.

[211] M. Tozaki, T. Igarashi, and K. Fukuda, "Positive and negative predictive values of bi-rads®-mri descriptors for focal breast masses," *Magnetic Resonance in Medical Sciences*, vol. 5, no. 1, pp. 7–15, 2006.

[212] S. Azam, M. Eriksson, A. Sjölander, M. Gabrielson, R. Hellgren, K. Czene, and P. Hall, "Predictors of mammographic microcalcifications," *International journal of cancer*, vol. 148, no. 5, pp. 1132–1143, 2021.

[213] S. Azam, M. Eriksson, A. Sjölander, M. Gabrielson, R. Hellgren, K. Czene, and P. Hall, "Mammographic microcalcifications and risk of breast cancer," *British journal of cancer*, vol. 125, no. 5, pp. 759–765, 2021.

[214] M. Muttarak, P. Kongmebhol, and N. Sukhamwang, "Breast calcifications: which are malignant," *Singapore Med J*, vol. 50, no. 9, pp. 907–914, 2009.

[215] M. Scimeca, R. Bonfiglio, E. Menichini, L. Albonici, N. Urbano, M. T. De Caro, A. Mauriello, O. Schillaci, A. Gambacurta, and E. Bonanno, "Microcalcifications drive breast cancer occurrence and development by macrophage-mediated epithelial to mesenchymal transition," *International journal of molecular sciences*, vol. 20, no. 22, p. 5633, 2019.

[216] S. A. Narod, "Age of diagnosis, tumor size, and survival after breast cancer: implications for mammographic screening," *Breast cancer research and treatment*, vol. 128, pp. 259–266, 2011.

[217] M. E. Brennan, R. M. Turner, S. Ciatto, M. L. Marinovich, J. R. French, P. Macaskill, and N. Houssami, "Ductal carcinoma in situ at core-needle biopsy: meta-analysis of underestimation and predictors of invasive breast cancer," *Radiology*, vol. 260, no. 1, pp. 119–128, 2011.

[218] T. Tot, M. Gere, S. Hofmeyer, A. Bauer, and U. Pellas, "The clinical value of detecting microcalcifications on a mammogram," in *Seminars in cancer biology*, vol. 72, pp. 165–174, Elsevier, 2021.

[219] C. K. Bent, L. W. Bassett, C. J. D'Orsi, and J. W. Sayre, "The positive predictive value of bi-rads microcalcification descriptors and final assessment categories," *American Journal of Roentgenology*, vol. 194, no. 5, pp. 1378–1383, 2010.

[220] L. J. Grimm, M. M. Miller, S. M. Thomas, Y. Liu, J. Y. Lo, E. S. Hwang, T. Hyslop, and M. D. Ryser, "Growth dynamics of mammographic calcifications: differentiating ductal carcinoma in situ from benign breast disease," *Radiology*, vol. 292, no. 1, pp. 77–83, 2019.

[221] J. Salvado and B. Roque, "Detection of calcifications in digital mammograms using wavelet analysis and contrast enhancement," in *IEEE International Workshop on Intelligent Signal Processing, 2005.*, pp. 200–205, IEEE, 2005.

[222] J. Mordang, A. Gubern-Mérida, A. Bria, F. Tortorella, R. Mann, M. Broeders, G. den Heeten, and N. Karssemeijer, "The importance of early detection of calcifications associated with breast cancer in screening," *Breast cancer research and treatment*, vol. 167, no. 2, pp. 451–458, 2018.

[223] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, 2022.

[224] C. Lei, W. Wei, Z. Liu, Q. Xiong, C. Yang, M. Yang, L. Zhang, T. Zhu, X. Zhuang, C. Liu, Z. Liu, J. Tian, and K. Wang, "Mammography-based radiomic analysis for predicting benign bi-rads category 4 calcifications," *European journal of radiology*, vol. 121, p. 108711, 2019.

[225] P. Stelzer, O. Steding, M. Raudner, G. Euller, P. Clauser, and P. Baltzer, "Combined texture analysis and machine learning in suspicious calcifications detected by mammography: Potential to avoid unnecessary stereotactical biopsies," *European Journal of Radiology*, vol. 132, p. 109309, 2020.

[226] K. Marathe, C. Marasinou, B. Li, N. Nakhaei, B. Li, J. G. Elmore, L. Shapiro, and W. Hsu, "Automated quantitative assessment of amorphous calcifications: Towards improved malignancy risk stratification," *Computers in Biology and Medicine*, vol. 146, p. 105504, 2022.

[227] K. Loizidou, G. Skouroumouni, C. Nikolaou, and C. Pitris, "An automated breast microcalcification detection and classification technique using temporal subtraction of mammograms," *IEEE Access*, vol. 8, pp. 52785–52795, 2020.

[228] A. Fanizzi, T. Basile, L. Losurdo, R. Bellotti, U. Bottigli, R. Dentamaro, V. Didonna, A. Fausto, R. Massafra, M. Moschetta, O. Popescu, P. Tamborra, S. Tangaro, and D. La Forgia, "A machine learning approach on multiscale texture analysis for breast microcalcification diagnosis," *BMC bioinformatics*, vol. 21, no. 2, pp. 1–11, 2020.

[229] S.-H. Lee, H. Park, and E. S. Ko, "Radiomics in breast imaging from techniques to clinical applications: a review," *Korean Journal of Radiology*, vol. 21, no. 7, p. 779, 2020.

[230] ELI5 Website, "Eli5 Documentation," 2022. (Last accessed 31-Mar-2022).

[231] S.-y. Huang, B. L. Franc, R. J. Harnish, G. Liu, D. Mitra, T. P. Copeland, V. A. Arasu, J. Kornak, E. F. Jones, S. C. Behr, N. M. Hylton, E. R. Price, L. Esserman, and S. Youngho, "Exploration of pet and mri radiomic features for decoding breast cancer phenotypes and prognosis," *NPJ breast cancer*, vol. 4, no. 1, pp. 1–13, 2018.

[232] V. S. Parekh and M. A. Jacobs, "Multiparametric radiomics methods for breast cancer tissue characterization using radiological imaging," *Breast cancer research and treatment*, vol. 180, no. 2, pp. 407–421, 2020.

[233] J. Liu, D. Sun, L. Chen, Z. Fang, W. Song, D. Guo, T. Ni, C. Liu, L. Feng, Y. Xia, X. Zhang, and C. Li, "Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer," *Frontiers in Oncology*, vol. 9, p. 980, 2019.

[234] R. Ramos-Pollán, M. A. Guevara-López, C. Suárez-Ortega, G. Díaz-Herrero, J. M. Franco-Valiente, M. Rubio-del Solar, N. González-de Posada, M. A. P. Vaz, J. Loureiro, and I. Ramos, "Discovering mammography-based machine learning classifiers for breast cancer diagnosis," *Journal of medical systems*, vol. 36, no. 4, pp. 2259–2269, 2012.

[235] R. Guo, G. Lu, B. Qin, and B. Fei, "Ultrasound imaging technologies for breast cancer detection and management: a review," *Ultrasound in medicine & biology*, vol. 44, no. 1, pp. 37–70, 2018.

[236] S. E. Lee, J. H. Yoon, N.-H. Son, K. Han, and H. J. Moon, "Screening in patients with dense breasts: Comparison of mammography, artificial intelligence, and supplementary ultrasound," *American Journal of Roentgenology*, 2023.

[237] E. A. Sickles and C. J. D'Orsi, "How should screening breast us be audited? the bi-rads perspective," *Radiology*, vol. 272, no. 2, pp. 316–320, 2014.

[238] M. A. Al-Masni, M. A. Al-Antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system," *Computer methods and programs in biomedicine*, vol. 157, pp. 85–94, 2018.

[239] G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, "Yolo based breast masses detection and classification in full-field digital mammograms," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105823, 2021.

[240] A. Baccouche, B. Garcia-Zapirain, C. C. Olea, and A. S. Elmaghraby, "Breast lesions detection and classification via yolo-based fusion models," *Comput. Mater. Contin*, vol. 69, pp. 1407–1425, 2021.

[241] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, and J. Kang, "Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network," *PloS one*, vol. 13, no. 9, p. e0203355, 2018.

[242] I. W. A. S. Darma, N. Suciati, and D. Siahaan, "A performance comparison of balinese carving motif detection and recognition using yolov5 and mask r-cnn," in *2021 5th International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 52–57, 2021.

[243] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[244] G. Chugh, S. Kumar, and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," *Cognitive Computation*, pp. 1–20, 2021.

[245] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[246] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[247] L. Abdelrahman, M. Al Ghamdi, F. Collado-Mesa, and M. Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Computers in biology and medicine*, vol. 131, p. 104248, 2021.

[248] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020.

[249] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, vol. 73, pp. 1–15, 2018.

[250] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137, 2015.

[251] M. A. Durand, S. Wang, R. J. Hooley, M. Raghu, and L. E. Philpotts, "Tomosynthesis-detected architectural distortion: management algorithm with radiologic-pathologic correlation," *Radiographics*, vol. 36, no. 2, pp. 311–321, 2016.

[252] O. N. Oyelade and A. E.-S. Ezugwu, "A state-of-the-art survey on deep learning methods for detection of architectural distortion from digital mammography," *IEEE Access*, vol. 8, pp. 148644–148676, 2020.

[253] T. Mahmood, J. Li, Y. Pei, F. Akhtar, M. U. Rehman, and S. H. Wasti, "Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach," *Plos one*, vol. 17, no. 1, p. e0263126, 2022.

[254] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 5, pp. 1–11, 2019.

[255] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis," *arXiv preprint arXiv:1811.02661*, 2018.

[256] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[257] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[258] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, 2020.

[259] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, 2018.

[260] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[261] Q. Tan, W. Xie, H. Tang, and Y. Li, "Multi-scale attention adaptive network for object detection in remote sensing images," in *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 218–223, IEEE, 2022.

[262] W. Li and L. Huang, "Yolosa: Object detection based on 2d local feature superimposed self-attention," *Pattern Recognition Letters*, vol. 168, pp. 86–92, 2023.

[263] M. Qiu, L. A. Christopher, S. Chien, and Y. Chen, "Attention mechanism improves yolov5x for detecting vehicles on surveillance videos," in *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–8, IEEE, 2022.

[264] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 818–833, Springer International Publishing, 2014.

[265] Ultralytics, "YoloV5 Ultralytics Github," 2022. (Last accessed 24-Jan-2023).

[266] wandb, "Weights & Biases," 2022. (Last accessed 24-Jan-2023).

[267] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *arXiv preprint arXiv:2102.13076*, 2021.

[268] T. M. Babkina, A. V. Gurando, T. M. Kozarenko, V. R. Gurando, V. V. Telniy, and D. V. Pominchuk, "Detection of breast cancers represented as architectural distortion: A comparison of full-field digital mammography and digital breast tomosynthesis," *Wiad Lek*, vol. 74, no. 7, pp. 1674–9, 2021.

[269] R. M. Rangayyan, S. Banik, and J. Desautels, "Computer-aided detection of architectural distortion in prior mammograms of interval cancer," *Journal of Digital Imaging*, vol. 23, no. 5, pp. 611–631, 2010.

[270] A. Arian, K. Dinas, G. C. Pratilas, and S. Alipour, "The breast imaging-reporting and data system (bi-rads) made easy," *Iranian Journal of Radiology*, vol. 19, no. 1, 2022.

[271] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in full field digital mammograms," *Computers in biology and medicine*, vol. 121, p. 103774, 2020.

[272] M. A. Al-Antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms," *Computer methods and programs in biomedicine*, vol. 196, p. 105584, 2020.

[273] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[274] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106126, 2023.

[275] S. Cannata, G. Cicceri, G. Cirrincione, T. Currieri, M. Lovino, C. Militello, F. Prinzi, and S. Vitabile, "Diffuser data augmentation for vit-based classification of dermatoscopic melanoma images," in *Sensors*, In press, 2023.

[276] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[277] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[278] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International conference on medical image computing and computer-assisted intervention*, pp. 603–611, Springer, 2017.

[279] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 2246–2255, 2020.

[280] G. Cirrincione, S. Cannata, G. Cicceri, F. Prinzi, T. Currieri, M. Lovino, C. Militello, E. Pasero, and S. Vitabile, "Transformer-based approach to melanoma detection," *Sensors*, vol. 23, no. 12, p. 5677, 2023.

[281] W. Hu, L. Fang, R. Ni, H. Zhang, and G. Pan, "Changing trends in the disease burden of non-melanoma skin cancer globally from 1990 to 2019 and its predicted level in 25 years," *BMC cancer*, vol. 22, no. 1, p. 836, 2022.

[282] P. P. Naik, "Cutaneous malignant melanoma: A review of early diagnosis and management," *World Journal of Oncology*, vol. 12, no. 1, p. 7, 2021.

[283] N. V. Kumar, P. V. Kumar, K. Pramodh, and Y. Karuna, "Classification of skin diseases using image processing and svm," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, pp. 1–5, IEEE, 2019.

[284] A. K. Verma, S. Pal, and S. Kumar, "Classification of skin disease using ensemble data mining techniques," *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 6, p. 1887, 2019.

[285] J. Xie, Z. Wu, R. Zhu, and H. Zhu, "Melanoma detection based on swin transformer and simam," in *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 5, pp. 1517–1521, IEEE, 2021.

[286] D. C. Malo, M. M. Rahman, J. Mahbub, and M. M. Khan, "Skin cancer detection using convolutional neural network," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0169–0176, IEEE, 2022.

[287] G. S. Krishna, K. Supriya, M. Sorgile, *et al.*, "Lesionaid: Vision transformers-based skin lesion generation and classification," *arXiv preprint arXiv:2302.01104*, 2023.

[288] The International Skin Imaging Collaboration, "The International Skin Imaging Collaboration." https://www.isic-archive.com// (visited on 11 May 2023).

[289] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018.

[290] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

[291] M. Khan, M. Mehran, Z. Haq, Z. Ullah, S. Naqvi, M. Ihsan, and H. Abbass, "Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review," *Expert Systems with Applications*, vol. 185, p. 115695, 2021.

[292] C. Combi and G. Pozzi, "Health informatics: clinical information systems and artificial intelligence to support medicine in the COVID-19 pandemic," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pp. 480–488, IEEE, 2021.

[293] A. Benfante, S. Principe, M. Cicero, M. Incandela, G. Seminara, C. Durante, and N. Scichilone, "Management of severe asthma during the first lockdown phase of SARS-CoV-2 pandemic: Tips for facing the second wave," *Pulmonary Pharmacology & Therapeutics*, vol. 73, p. 102083, 2022.

[294] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in china: a report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[295] Z. Li, S. Zhao, Y. Chen, F. Luo, Z. Kang, S. Cai, W. Zhao, J. Liu, D. Zhao, and Y. Li, "A deep-learning-based framework for severity assessment of COVID-19 with CT images," *Expert Systems with Applications*, vol. 185, p. 115616, 2021.

[296] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review," *Clinical imaging*, vol. 64, pp. 35–42, 2020.

[297] E. Angeli, S. Dalto, S. Marchese, L. Setti, M. Bonacina, F. Galli, E. Rulli, V. Torri, C. Monti, R. Meroni, G. Beretta, M. Castoldi, and E. Bombardieri, "Prognostic value of CT integrated

with clinical and laboratory data during the first peak of the COVID-19 pandemic in northern italy: A nomogram to predict unfavorable outcome," *European Journal of Radiology*, vol. 137, p. 109612, 2021.

[298] D. Wang, C. Huang, S. Bao, T. Fan, Z. Sun, Y. Wang, H. Jiang, and S. Wang, "Study on the prognosis predictive model of COVID-19 patients based on CT radiomics," *Scientific Report*, vol. 11, no. 1, p. 11591, 2021.

[299] H. Shi, Z. Xu, G. Cheng, H. Ji, L. He, J. Zhu, H. Hu, Z. Xie, W. Ao, and J. Wang, "CT-based radiomic nomogram for predicting the severity of patients with COVID-19," *Eur J Med Res*, vol. 27, no. 13, 2022.

[300] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Saliency driven image manipulation," *Machine Vision and Applications*, vol. 30, no. 2, pp. 189–202, 2019.

[301] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *The Visual Computer*, vol. 29, no. 5, pp. 381–392, 2013.

[302] F. Prinzi, C. Militello, N. Scichilone, S. Gaglio, and S. Vitabile, "Explainable machine-learning models for covid-19 prognosis prediction using clinical, laboratory and radiomic features," *Access*, Under Review.

[303] Z. Xu, L. Zhao, G. Yang, Y. Ren, J. Wu, Y. Xia, X. Yang, M. Cao, G. Zhang, T. Peng, J. Zhao, H. Yang, J. Hu, and J. Du, "Severity assessment of COVID-19 using a CT-based radiomics model," *Stem Cells International*, vol. 2021, no. 2263469, 2021.

[304] A. Borghesi and R. Maroldi, "COVID-19 outbreak in italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression," *La Radiologia medica*, vol. 125, no. 5, p. 509–513, 2020.

[305] I. Shiri, Y. Salimi, M. Pakbin, G. Hajianfar, A. Avval, A. Sanaat, S. Mostafaei, A. Akhavanallaf, A. Saberi, Z. Mansouri, D. Askari, and et al., "COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients," *Computers in Biology and Medicine*, vol. 145, p. 105467, 2022.

[306] I. Shiri, H. Arabi, Y. Salimi, A. Sanaat, A. Akhavanallaf, G. Hajianfar, D. Askari, S. Moradi, Z. Mansouri, M. Pakbin, S. Sandoughdaran, and et al., "COLI-Net: Deep learning-assisted fully automated COVID-19 lung and infection pneumonia lesion detection and segmentation from chest computed tomography images," *International Journal of Imaging Systems and Technology*, vol. 32, no. 1, pp. 12–25, 2022.

[307] W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[308] C. Barbano, E. Tartaglione, C. Berzovini, M. Calandri, and M. Grangetto, "A two-step explainable approach for COVID-19 computer-aided diagnosis from chest X-ray images," *arXiv preprint*, 2021.

[309] V. Guarrasi and P. Soda, "Multi-objective optimization determines when, which and how to fuse deep networks: an application to predict COVID-19 outcomes," 2022.

[310] Hackathon Website, "Covid CXR Hackathon Competition," 2022. (Last accessed 31-Mar-2022).

[311] P. Prasad, D. Prasad, and G. S. Rao, "Performance analysis of orthogonal and biorthogonal wavelets for edge detection of x-ray images," *Procedia Computer Science*, vol. 87, pp. 116–121, 2016.

[312] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "Pywavelets: A python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, no. 36, p. 1237, 2019.

[313] F. Prinzi, C. Militello, V. Conti, and S. Vitabile, "Impact of wavelet kernels on predictive capability of radiomic features: A case study on covid-19 chest x-ray images," *Journal of Imaging*, vol. 9, no. 2, p. 32, 2023.

[314] S. Pragada and J. Sivaswamy, "Image denoising using matched biorthogonal wavelets," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 25–32, IEEE, 2008.

[315] Z. Z. Abidin, M. Manaf, and A. S. Shibhgatullah, "Experimental approach on thresholding using reverse biorthogonal wavelet decomposition for eye image," in *2013 IEEE International Conference on Signal and Image Processing Applications*, pp. 349–353, 2013.

[316] S. K. T.N. Tilak, "Reverse biorthogonal spline wavelets in undecimated transform for image denoising," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 66–72, 2018.

[317] Rohima and M. Barkah Akbar, "Wavelet analysis and comparison from coiflet family on image compression," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–5, 2020.

[318] S. A. A. Karim, M. K. Hasan, J. Sulaiman, J. B. Janier, M. T. Ismail, and M. S. Muthuvalu, "Denoising solar radiation data using coiflet wavelets," in *AIP Conference Proceedings*, vol. 1621, pp. 394–401, American Institute of Physics, 2014.

[319] K. Wahid, "Low complexity implementation of daubechies wavelets for medical imaging applications," in *Discrete Wavelet Transforms-Algorithms and Applications*, IntechOpen, 2011.

[320] Y. Meyer, *Wavelets and Operators: Volume 1.* No. 37 in Cambridge studies in advance mathematics, Cambridge university press, 1992.

[321] M.-T. Wu, "Wavelet transform based on meyer algorithm for image edge and blocking artifact reduction," *Information Sciences*, vol. 474, pp. 125–135, 2019.

[322] P. Porwik and A. Lisowska, "The haar-wavelet transform in digital image processing: its status and achievements," *Machine graphics and vision*, vol. 13, no. 1/2, pp. 79–98, 2004.

[323] J. Bhardwaj and A. Nayak, "Haar wavelet transform–based optimal bayesian method for medical image fusion," *Medical & Biological Engineering & Computing*, vol. 58, no. 10, pp. 2397–2411, 2020.

[324] S. Narula and S. Gupta, "Image compression radiography using haar wavelet transform," *International Journal of Computer Applications*, vol. 975, p. 8887, 2015.

[325] J. Wang and K. Huang, "Medical image compression by using three-dimensional wavelet transformation," *IEEE Transactions on Medical Imaging*, vol. 15, no. 4, pp. 547–554, 1996.

[326] S. Arfaoui, A. B. Mabrouk, and C. Cattani, *Wavelet Analysis: Basic Concepts and Applications.* Chapman and Hall/CRC, 2021.

[327] R. X. Gao and R. Yan, *Wavelets: Theory and applications for manufacturing.* Springer Science & Business Media, 2010.

[328] P. Shrout and J. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological bulletin*, vol. 86, no. 2, pp. 420—-428, 1979.

[329] P. Westfall, J. Troendle, and G. Pennello, "Multiple McNemar tests," *Biometrics*, vol. 66, no. 4, pp. 1185–1191, 2010.

[330] J. Su and J. Liu, "Linear combinations of multiple diagnostic markers," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1350–1355, 1993.

[331] B. M. Henry, G. Aggarwal, J. Wong, S. Benoit, J. Vikse, M. Plebani, and G. Lippi, "Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis," *Am J Emerg Med*, vol. 38, no. 9, pp. 1722–1726, 2020.

[332] M. Tobin, A. Jubran, and L. F., "PaO2/FIO2 ratio: the mismeasure of oxygenation in COVID-19," *European Respiratory Journal*, vol. 57, p. 2100274, 2021.

[333] N. Ali, "Elevated level of c-reactive protein may be an early marker to predict risk for severity of COVID-19," *Journal of medical virology*, 2020.

[334] N. Nguyen, J. Chinn, M. De Ferrante, K. Kirby, S. Hohmann, and A. Amin, "Male gender is a predictor of higher mortality in hospitalized adults with COVID-19," *PLoS One*, vol. 16, no. 7, p. e0254066, 2021.

[335] S. Principe, A. Grosso, A. Benfante, F. Albicini, S. Battaglia, E. Gini, M. Amata, I. Piccionello, A. Corsico, and N. Scichilone, "Comparison between suspected and confirmed COVID-19 respiratory patients: What is beyond the PCR test," *Journal of Clinical Medicine*, vol. 11, no. 11, p. 2993, 2022.

[336] T. Do, S. Skornitzke, U. Merle, M. Kittel, S. Hofbaur, C. Melzig, H. Kauczor, M. Wielpütz, and O. Weinheimer, "COVID-19 pneumonia: Prediction of patient outcome by CT-based quantitative lung parenchyma analysis combined with laboratory parameters," *PloS one*, vol. 17, no. 7, p. e0271787, 2022.

[337] F. Innocenti, A. De Paris, A. Lagomarsini, L. Pelagatti, L. Casalini, A. Gianno, M. Montuori, P. Bernardini, F. Caldi, I. Tassinari, and R. Pini, "Stratification of patients admitted for SARS-CoV2 infection: prognostic scores in the first and second wave of the pandemic," *Internal and Emergency Medicine*, pp. 1–9, 2022.

[338] G. Iacobellis, A. E. Malavazos, and M. M. Corsi, "Epicardial fat: from the biomolecular aspects to the clinical practice," *The international journal of biochemistry & cell biology*, vol. 43, no. 12, pp. 1651–1654, 2011.

[339] H. S. Sacks and J. N. Fain, "Human epicardial adipose tissue: a review," *American heart journal*, vol. 153, no. 6, pp. 907–917, 2007.

[340] C. Weber and H. Noels, "Atherosclerosis: current pathogenesis and therapeutic options," *Nature medicine*, vol. 17, no. 11, pp. 1410–1422, 2011.

[341] M. Kolossváry, C. N. De Cecco, G. Feuchtner, and P. Maurovich-Horvat, "Advanced atherosclerosis imaging by CT: radiomics, machine learning and deep learning," *Journal of cardiovascular computed tomography*, vol. 13, no. 5, pp. 274–280, 2019.

[342] P. Maurovich-Horvat, M. Ferencik, S. Voros, B. Merkely, and U. Hoffmann, "Comprehensive plaque assessment by coronary CT angiography," *Nature Reviews Cardiology*, vol. 11, no. 7, pp. 390–402, 2014.

[343] D. R. Obaid, P. A. Calvert, D. Gopalan, R. A. Parker, S. P. Hoole, N. E. West, M. Goddard, J. H. Rudd, and M. R. Bennett, "Atherosclerotic plaque composition and classification identified by coronary computed tomography: assessment of computed tomography–generated plaque maps compared with virtual histology intravascular ultrasound and histology," *Circulation: Cardiovascular Imaging*, vol. 6, no. 5, pp. 655–664, 2013.

[344] L. La Grutta, P. Toia, A. Farruggia, D. Albano, E. Grassedonio, A. Palmeri, E. Maffei, M. Galia, S. Vitabile, F. Cademartiri, *et al.*, "Quantification of epicardial adipose tissue in coronary calcium score and CT coronary angiography image data sets: comparison of attenuation values, thickness and volumes," *The British Journal of Radiology*, vol. 89, no. 1062, p. 20150773, 2016.

[345] C. Militello, L. Rundo, P. Toia, V. Conti, G. Russo, C. Filorizzo, E. Maffei, F. Cademartiri, L. La Grutta, M. Midiri, *et al.*, "A semi-automatic approach for epicardial adipose tissue segmentation and quantification on cardiac CT scans," *Computers in biology and medicine*, vol. 114, p. 103424, 2019.

[346] K. Cheng, A. Lin, J. Yuvaraj, S. J. Nicholls, and D. T. Wong, "Cardiac computed tomography radiomics for the non-invasive assessment of coronary inflammation," *Cells*, vol. 10, no. 4, p. 879, 2021.

[347] C. Militello, F. Prinzi, G. Sollami, L. Rundo, L. La Grutta, and S. Vitabile, "Ct radiomic features and clinical biomarkers for predicting coronary artery disease," *Cognitive Computation*, vol. 15, no. 1, pp. 238–253, 2023.

[348] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, and O. Gambino, "Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine," *Journal of Biomedical Informatics*, vol. 108, p. 103479, 2020.

[349] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics*, vol. 6, no. 1, pp. 1–15, 2014.

[350] J. Zhou, Y. Chen, Y. Zhang, H. Wang, Y. Tan, Y. Liu, L. Huang, H. Zhang, Y. Ma, and H. Cong, "Epicardial fat volume improves the prediction of obstructive coronary artery disease above traditional risk factors and coronary calcium score: development and validation of new pretest probability models in chinese populations," *Circulation: Cardiovascular Imaging*, vol. 12, no. 1, p. e008002, 2019.

[351] M. Goeller, S. Achenbach, M. Marwan, M. K. Doris, S. Cadet, F. Commandeur, X. Chen, P. J. Slomka, H. Gransar, J. J. Cao, *et al.*, "Epicardial adipose tissue density and volume are related to subclinical atherosclerosis, inflammation and major adverse cardiac events in asymptomatic subjects," *Journal of cardiovascular computed tomography*, vol. 12, no. 1, pp. 67–73, 2018.

[352] M. Goeller, S. Achenbach, S. Cadet, A. C. Kwan, F. Commandeur, P. J. Slomka, H. Gransar, M. H. Albrecht, B. K. Tamarappoo, D. S. Berman, *et al.*, "Pericoronary adipose tissue computed tomography attenuation and high-risk plaque characteristics in acute coronary syndrome compared with stable coronary artery disease," *JAMA cardiology*, vol. 3, no. 9, pp. 858–863, 2018.

[353] S. Hedgire, V. Baliyan, E. J. Zucker, D. O. Bittner, P. V. Staziaki, R. A. Takx, J.-E. Scholtz, N. Meyersohn, U. Hoffmann, and B. Ghoshhajra, "Perivascular epicardial fat stranding at coronary CT angiography: a marker of acute plaque rupture and spontaneous coronary artery dissection," *Radiology*, vol. 287, no. 3, p. 808, 2018.

[354] M. Kolossváry, J. Karády, B. Szilveszter, P. Kitslaar, U. Hoffmann, B. Merkely, and P. Maurovich-Horvat, "Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign," *Circulation: Cardiovascular Imaging*, vol. 10, no. 12, p. e006843, 2017.

[355] A. Lin, M. Kolossváry, J. Yuvaraj, S. Cadet, P. A. McElhinney, C. Jiang, N. Nerlekar, S. J. Nicholls, P. J. Slomka, P. Maurovich-Horvat, *et al.*, "Myocardial infarction associates with a distinct pericoronary adipose tissue radiomic phenotype: a prospective case-control study," *Cardiovascular Imaging*, vol. 13, no. 11, pp. 2371–2383, 2020.

[356] G.-q. Hu, Y.-q. Ge, X.-k. Hu, and W. Wei, "Predicting coronary artery calcified plaques using perivascular fat CT radiomics features and clinical risk factors," *BMC Medical Imaging*, vol. 22, no. 1, pp. 1–10, 2022.

[357] G. Kalykakis, F. Driest, D. Terentes, A. Broersen, P. Kafouris, T. Pitsariotis, N. Anousakis Vlachochristou, A. Antonopoulos, G. Benetos, R. Liga, *et al.*, "Radiomics-based analysis by machine learning techniques improves characterization of functionally significant coronary lesions," *European Heart Journal*, vol. 43, no. Supplement_2, pp. ehac544–216, 2022.

[358] K. Koutroumbas and S. Theodoridis, *Pattern recognition*. London, United Kingdom: Academic Press, 4th ed., 2009.

[359] W. Depression, "Other common mental disorders: global health estimates," *Geneva: World Health Organization*, vol. 24, 2017.

[360] C. Woody, A. Ferrari, D. Siskind, H. Whiteford, and M. Harris, "A systematic review and meta-regression of the prevalence and incidence of perinatal depression," *Journal of affective disorders*, vol. 219, pp. 86–92, 2017.

[361] M. Semrau, A. Alem, J. L. Ayuso-Mateos, D. Chisholm, O. Gureje, C. Hanlon, M. Jordans, F. Kigozi, C. Lund, I. Petersen, *et al.*, "Strengthening mental health systems in low-and middle-income countries: recommendations from the emerald programme," *BJPsych Open*, vol. 5, no. 5, p. e73, 2019.

[362] M. Large, "Study on suicide risk assessment in mental illness underestimates inpatient suicide risk," *Bmj*, vol. 352, 2016.

[363] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, pp. 591–643, 2013.

[364] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, M. Simes, R. Berman, S. H. Koenigsberg, and R. C. Kessler, "The economic burden of adults with major depressive disorder in the united states (2010 and 2018)," *Pharmacoeconomics*, vol. 39, no. 6, pp. 653–665, 2021.

[365] A. Esposito, G. Raimo, M. Maldonato, C. Vogel, M. Conson, and G. Cordasco, "Behavioral sentiment analysis of depressive states," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 000209–000214, IEEE, 2020.

[366] C. Taleb, M. Khachab, C. Mokbel, and L. Likforman-Sulem, "Feature selection for an improved parkinson's disease identification based on handwriting," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pp. 52–56, IEEE, 2017.

[367] G. Cordasco, F. Scibelli, M. Faundez-Zanuy, L. Likforman-Sulem, and A. Esposito, "Handwriting and drawing features for detecting negative moods," in *Italian Workshop on Neural Nets*, pp. 73–86, Springer, 2017.

[368] G. Raimo, M. Buonanno, M. Conson, G. Cordasco, M. Faundez-Zanuy, G. McConvey, S. Marrone, F. Marulli, A. Vinciarelli, and A. Esposito, "Handwriting and drawing for depression detection: A preliminary study," in *Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings*, pp. 320–332, Springer, 2023.

[369] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, and S. Melacci, "Entropy-based logic explanations of neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 6046–6054, 2022.

[370] F. Prinzi, P. Barbiero, G. Cordasco, P. Lio, S. Vitabile, and A. Esposito, "Explainable depression detection using handwriting features," in *the Italian Workshop on Neural Networks*, In press, 2023.

[371] G. Bottesi, M. Ghisi, G. Altoè, E. Conforti, G. Melli, and C. Sica, "The italian version of the depression anxiety stress scales-21: Factor structure and psychometric properties on community and clinical samples," *Comprehensive psychiatry*, vol. 60, pp. 170–181, 2015.

[372] G. Cordasco, F. Scibelli, M. Faundez-Zanuy, L. Likforman-Sulem, and A. Esposito, "Handwriting and drawing features for detecting negative moods," *Quantifying and Processing Biomedical and Behavioral Signals 27*, pp. 73–86, 2019.

[373] G. Cordasco, M. Buonanno, M. Faundez-Zanuy, M. T. Riviello, L. Likforman-Sulem, and A. Esposito, "Gender identification through handwriting: an online approach," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 000197–000202, IEEE, 2020.

[374] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, and E. Haffen, "Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications," *BioMed research international*, vol. 2013, 2013.

[375] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 248–266, 2022.