Integrating Multimodal Learning and Explainable AI for Enhanced and Interpretable Prostate Lesion Classification

Claudio Giovannoni[1], Carlo Metta[2*], Andrea Berti[2], Sara Colantonio[2], Anna Monreale[1], Francesca Pratesi[2], Salvatore Rinzivillo[2]

[1]Department of Computer Science, University of Pisa, Lungarno Antonio Pacinotti, 43, Pisa, 56126, Italy.
[2]ISTI-CNR, Via Moruzzi 1, Pisa, 56127, Italy.

*Corresponding author(s). E-mail(s): carlo.metta@isti.cnr.it;
Contributing authors: claudio.giovannoni@phd.unipi.it; andrea.berti@isti.cnr.it; sara.colantonio@isti.cnr.it; anna.monreale@unipi.it; francesca.pratesi@isti.cnr.it; rinzivillo@isti.cnr.it;

Abstract

Artificial Intelligence systems could find many important applications in the medical field, holding excellent potential for improving disease diagnosis, treatment identification and selection. These opportunities are often jeopardized by the lack of interpretability of such systems, slowing down AI adoption. To overcome the issue, we first introduce an analytical framework exploiting *multimodal deep learning* for the classification of prostate lesions using Magnetic Resonance Imaging (MRI) data and clinical information on the patients. Then, we propose a *multimodal explainability* approach based on visual explanations to interpret the proposed model decision-making process and identify how the different modalities contribute to each specific prediction. Our findings, based on the PI-CAI Grand Challenge dataset, demonstrate the potential of combining multimodal data with eXplainable AI (XAI) to enhance prostate cancer diagnosis, improving model predictive performance, interpretability and understanding in treatment decision-making.

Keywords: Deep Learning, Explainable Artificial Intelligence, Prostate Cancer

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are opening new frontiers in healthcare and medical diagnostics, revolutionizing the way healthcare professionals diagnose and treat diseases. ML models have become more and more sophisticated, and their performance comes at the cost of a clear interpretation of their inner workings. The complexity and opaqueness of such models make them "black-boxes" and, when it comes to making high-stakes decisions, interpretation of the ML models becomes a critical building block for a trustworthy interaction between humans and AI systems.

In the current research panorama, clinical classification models are developed with the aim of establishing robust decision support solutions. However, beside their demonstrated predictive performance, not all of these include a complete XAI analysis accounting for the reliability, interpretability in order to guarantee trustworthiness and transparency. In this regard, XAI aims at making ML models more interpretable and reliable, providing explanations describing the rationale behind the model output and supporting the decision making process of domain experts. In the medical setting, the inherent multimodal nature of data is predominant. Nonetheless, most current clinical multimodal frameworks rely on standard early or late fusion strategies [7, 52, 61] and do not leverage XAI as an intrinsic part of the multimodal pipeline. In addition, integration of tabular metadata features into convolutional computer vision models has not yet been sufficiently explored, often lacking meaningful spatial mapping techniques that preserve semantic relationships among features. Even among works applying multimodal learning approaches, XAI modules are often shallow, composed of unimodal techniques, not fully exploiting the insights present within input data, beside for the classifier's training. We argue that the multimodal medical data sources must be further exploited through multiple XAI methods both in terms of the input data modality explained as well as the scope of the explanations produced with the aim to maximize model interpretability. To address the aforementioned research gap, we propose a strategy based on this paradigm, which we call multimodal explainability, extending multimodality across input (for model training) and output (for model explainability).

In this paper, we present a novel approach for interpretable clinically significant prostate cancer (csPCa) classification, combining multiple visual, i.e. magnetic resonance imaging (MRI), and tabular data. Unlike prior works (e.g. in reference [25]) that adopt late fusion architectures or focus on separate explainability layers, we aim at offering a coherent and interpretable end-to-end diagnostic pipeline.

Methodologically, key novelties include: (i) data preprocessing composed of volumetric-to-2D slice conversion, automated relevant slice selection and automatic region of interest (ROI) cropping and centering; (ii) hybrid multimodal fusion technique integrating input data at both feature-level and model-level through transfer learning, pixel mapping of tabular data and integration of complementary multimodal MRI along the input color channel; (iii) comprehensive post-hoc multimodal XAI, producing visual and quantitative insights at both local and global interpretability scopes to increase clinical trust and model transparency.

Our framework employs an intrinsically, as opposed to extrinsically, multimodal explanation process. The term "intrinsically multimodal" refers to the explanation

pipeline being aligned with the input modalities used for model training. It does not imply that the model is intrinsically explainable in the architectural sense. The explanations remain within the same data modalities as those used for training and do not leverage external modalities or knowledge bases, which would characterize extrinsically multimodal explanation systems. Therefore, explanations are directly derived from the input modalities without integrating additional modalities solely for explanation purposes and not for model training. An extrinsically multimodal explanation process, on the contrary, would introduce explanation modalities outside the one used in model input, to create an enriched or more accessible XAI setup.

The model is trained on a recently published prostate MRI dataset provided by the Prostate Imaging Cancer AI (PI-CAI) Grand Challenge [9], which includes medical exams from EU research centers and exams previously collected for the ProstateX challenge [28]. PI-CAI is composed of three MRI acquisition modalities, i.e. Apparent Diffusion Coefficient or ADC, T2-weighted or T2w, and Diffusion-Weighted Imaging or DWI, as well as tabular metadata for patients with prostate lesions. Among the tabular metadata, two labels of clinical significance of the tumor are provided, namely the Gleason score and the corresponding ISUP grade [55].
This framework consists of four major steps, or modules:

- Data Preprocessing: MRI acquisitions are processed to extract prostate gland volumes and reduce dimensionality. Clinical information in the form of tabular data is processed into compact image representations, preserving the original proximity feature of the neighborhood structure. This is crucial for CNN classifiers, which rely on spatial relationships of images to extract meaningful patterns.
- Input Fusion: Tabular and image are integrated using a hybrid fusion approach that combines modalities at both the feature level, by stacking them into a 4D array, and at the model level, by transferring knowledge from a model trained only on tabular data and another trained on multimodal images into the multimodal classifier.
- Multimodal Learning: A CNN is trained on multimodal data integrating information from: *i)* the three MRI modalities as image channels (T2w, ADC, and DWI), and *ii)* the embedding matrix of tabular metadata. This neural model is trained to distinguish patients with an ISUP score less than 2 (non-csPCa cases), against patients with an ISUP score equal or greater than 2 (csPCa cases).
- Explanation: XAI techniques are applied to extract saliency maps that highlight channels (modes) and regions of the most relevance for model prediction.

Results show that our multimodal explainability pipeline provides significant advantages in terms of classification performance and interpretability compared to standard unimodal approaches. We achieve a remarkable improvement in the model performance by exploiting multimodal information, reaching an AUC score of 91.2 at patient level and 0.872 at slice level under 5-fold patient-level cross-validation (the "patient level" analysis considers the whole 3D MRI to assess overall outcomes, while "slice level" focuses on evaluations of individual 2D slices). Additionally, we gain in terms of interpretability: saliency maps provide insight into how different modalities contribute to the final prediction, increasing trust and reliability of the decision model.

Our work represents a significant step forward in the development of explainable AI systems for prostate cancer diagnosis. The multimodal approach and the use of XAI techniques open new lines of research in this field to improve the accuracy, reliability, and interpretability of ML models for the diagnosis and treatment of various diseases. The findings are detailed in the following sections, providing an in-depth analysis of the data, methods, and results. In Section 2, we explore the literature on AI and ML in healthcare, specifically focusing on diagnostic applications in imaging and the importance of XAI techniques. In Section 3, we detail the methodology: the preprocessing of MRI acquisitions, the architecture of our models, and the rationale behind the multimodal strategies. Section 3.1 details the specifics of the dataset used in this study. Here, we discuss the steps useful for extracting the most informative slices from the MRI scans, ensuring our models were trained on data that accurately represent the prostate gland characteristics. Section 3.2 outlines the development of neural network models, the architectures details, as well as the integration of multimodal sources. In Section 4, we present the experimental results, the comparison with baseline models and discuss the most relevant metrics. We detail how explainability is integrated to generate insights into which features are most influential in the models decision processes. Lastly, Section 5 concludes and discusses future research directions.

2 Related Work

2.1 AI and ML in healthcare

The literature is replete with studies demonstrating the efficacy of ML models in interpreting medical data. Most works focus on the ability of such models to learn essential statistics from medical images. Litjens et al. [15] provide a review of ML applications in medical imaging, highlighting their success in various tasks, including the detection of lesion, organ segmentation, and disease classification. Similarly, Shen et al. [47] underscore the transformative power of machine learning in medical imaging, pointing to its ability to significantly improve diagnostic accuracy. Recent contributions continue to explore advanced neural architectures in biomedical imaging. Notably, adaptive migration networks have been proposed to tackle multimodal image classification tasks with domain shifts [60]. Similarly, multitask deep models have been successfully applied to real-time video diagnosis in clinical settings, leveraging explainability through Shapley values [14].

Focusing on prostate cancer, AI-driven models have shown remarkable proficiency. The study by Yoo et al. [61] stands out as a seminal work. This research is notable for being the first to effectively propose the strategy of stacking different MRI modalities to improve the accuracy of prostate cancer detection. Despite the limitations imposed by the relatively small size of the dataset, the authors demonstrate the significant potential of CNNs to extract and learn critical features from multimodal MRI. However, the above mentioned approaches typically focus on combining MRI modalities only, without integrating heterogeneous data types such as clinical or demographic tabular variables. Moreover, most of these models prioritize performance without addressing the transparency of the decision process. The role and contribution of each modality often remain opaque, limiting their usability in real-world clinical contexts.

2.2 XAI in Healthcare

The application of ML in this field presents numerous challenges and critical issues that can harness its full capabilities. One of the primary difficulties lies in the heterogeneity and complexity of medical data, which includes a wide range of formats, from structured electronic health records to unstructured clinical notes and multimodal imaging data. This diversity requires sophisticated preprocessing techniques and ML models capable of handling this complexity.

Another complication is the inherent distortion in medical datasets, often resulting from uneven distributions between populations. Moreover, the "black box" nature of modern ML models poses a significant challenge to their clinical adoption, as it obscures the rationale behind their predictions. This has led to an increasing emphasis on XAI [19], which aims to make AI decisions understandable to the users. Techniques such as LIME [43], LORE [20] and Grad-CAM [46] have been successfully applied in the healthcare domain. In video-based diagnostic workflows, multitask Shapley explanation networks have recently been proposed to enable real-time and interpretable clinical decisions [14], highlighting the synergy between high-performance ML and explainability requirements in medical practice. Research on XAI methods includes a wide range of studies aimed at clarifying predictive models in various domains, including brain cancer [63], skin lesions [32–34], lung cancer [59] and pancreatic cancer [5]. These studies leverage generative and predictive techniques to provide insight into the model internal processes. Significant work has been done to interpret data from tabular sources, such as electronic health records [31], clinical pathways [39] and medication prescriptions [38].

In prostate cancer diagnostics, the application of XAI techniques improves the interpretability of AI models used to analyze MRI acquisitions. These approaches help delineate the features that most significantly influence the model outputs, facilitating a better understanding of AI predictions in clinical decision making [22, 23, 42, 50]. The integration of AI and ML into healthcare diagnostics represents a significant step forward in the search to improve patient outcomes through more accurate and efficient disease detection. The concurrent development of XAI methodologies addresses the critical need for transparency and trust in AI applications, ensuring that healthcare professionals can understand and leverage AI tools effectively. Nonetheless, current XAI techniques are generally applied post hoc and mostly in unimodal settings, focusing on medical images or structured data alone. There is a lack of frameworks that provide modality-specific explanations in truly multimodal contexts, where fused inputs contribute differently to the prediction. This represents a key gap in the literature, especially in applications where decisions affect clinical practice.

2.3 Multimodal Explainability

The training of ML models using the combination of heterogeneous data sources has been explored since the 1980s. One of the first multimodal learning tasks belongs to Natural Language Processing (NLP) and involved both Audio and Visual modalities for Speech recognition [62]. With technological progress, the advent of neural network

architectures, Big Data and the rise in computing power, this research is witnessing a powerful comeback.

Multimodal learning particularly suits the application of XAI solutions, providing a deeper level of interpretability by exploiting multimodal inputs through multiple explanation techniques targeted to specific modalities and scopes. The work by Wang et al. [56] explores colorectal cancer detection by proposing multitask Shapley explanation networks (EMSEN) capable of real-time explainable polyp detection and classification using multimodal image colonoscopy inputs. EMSEN employs Shapley values to generate local saliency maps, highlighting relevant pixel regions influencing classifier decisions.

Another notable contribution is the work of Mini Han Wang et al. [57]: a black box model paired with post-hoc XAI and trained on multi source medical IoT data. Features extracted from multimodal image retinal scans (i.e. color fundus photography, optical coherence tomography, ultra-wide and fluorescein angiography fundus images) are combined for age-related macular degeneration diagnosis, which is challenging due to limited data availability and strong class imbalance. Visual-based XAI methods such as Grad-CAM [46] are used to generate saliency maps, enhancing interpretability and optimizing CNN layers through attention mechanisms.

In [58] the authors apply a CNN for skin lesion classification applying a multimodal explainability approach by producing local and global visual and tabular explanations. In reference [37], the authors propose multimodal explainability through XAI orchestration: an adaptive and interactive tool endowed with a multimodally extrinsic explanation process, enabled though integration of retrieval augmented generation (RAG) technique, aiding domain experts by generating summaries of combined multimodal data, AI predictions and explanations.

Despite these notable contributions and several other prominent examples, the application of multimodal techniques to the model's explanation modules remains limited. This calls for a paradigm shift towards multimodal explainability in the field, to promote trustworthiness, reliability and enhance intepretability of AI applications in medicine.

Multimodal learning involves also challenges, such as the opacity in the use of multimodal information and the increased complexity of addressing intermodal and intramodal relationships among data. As explored in [6], there are several challenges related to multimodality. These are connected to (i) the representation of multimodal data into inputs; (ii) the mapping of one modality to another one; (iii) the identification of direct relationships; (iv) the fusion of joint information and the way knowledge is transferred to different modalities. Recent work in computer vision has also addressed such challenges outside healthcare, e.g., by introducing multimodal architectures to detect and ground inconsistencies in manipulated cross-modal data streams [44], showing that interpretability remains a central issue even beyond clinical domains.

We overcome some of these challenges by exploiting the complementarity of information provided by multimodal medical data in a framework combined with XAI for medical experts.

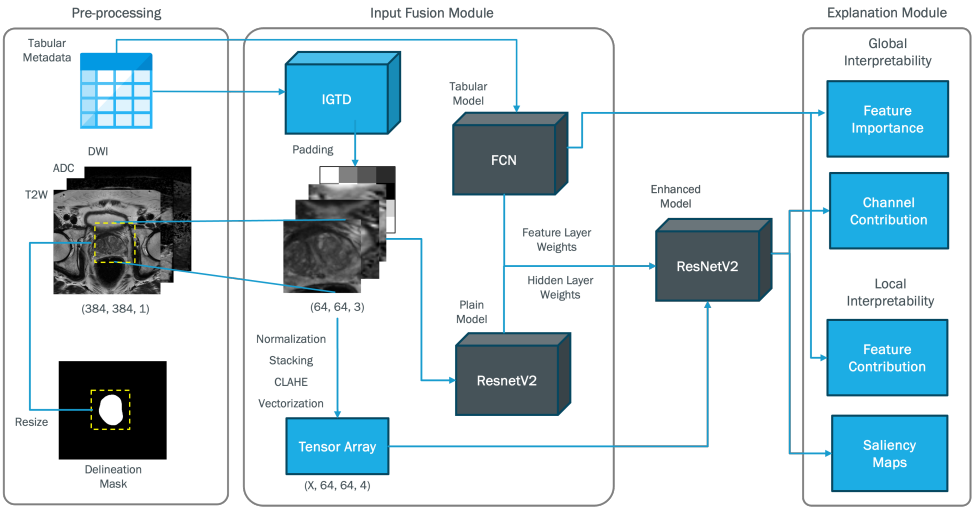

Fig. 1: The framework has three stages: pre-processing, where patient-related visual and tabular modalities are extracted and processed; integrating multimodal features for classification; and explanation module, which generates explanations through quantitative and visual-based methods.

Although recent efforts have explored multimodal fusion for diagnostic tasks, most models either treat tabular data and imaging separately or merge them without structure-aware integration. Additionally, very few works have quantitatively assessed the contribution of metadata once fused with imaging modalities. Our approach addresses these deficiencies by (i) integrating tabular data as a spatially structured image channel via IGTD, and (ii) introducing quantitative XAI metrics to assess the alignment between saliency and expert annotations, thus bridging the gap between multimodal prediction and interpretability.

3 Analytical Framework

In this section, we describe the analytical framework to address the problem of classifying the clinical significance of prostate lesion by exploiting multimodal data. A graphical representation of this framework is shown in Figure 1. The section is divided into three subsections, each dedicated to a pivotal step of our framework.

3.1 Dataset and Preprocessing

The dataset used in this research [45] aims to unify and standardize modern medical guidelines while ensuring a meaningful validation of prostate-AI technologies for clinical translation. It comprises 1,500 anonymized bi-parametric MRI scans from 1,476 patients, collected between 2012 and 2021 from different medical centers in The Netherlands and Norway. The patient exams include, in addition to medical images, clinical variables such as patient age, prostate volume, PSA level, PSA density, and fundamental acquisition parameters, including scanner manufacturer and model name, and diffusion b-value. The bi-parametric MRI imaging includes the following: axial,

sagittal, and coronal T2w; axial high b-value DWI; and axial ADC. The csPCa lesions were delineated by experienced radiologists. In prostate cancer detection, MRI is crucial for medical investigation and lesion identification [18], among which we consider three modalities: T2-weighted (T2w), High b-value Diffusion Weighted (DWI) and Apparent Diffusion Coefficient (ADC) maps, all in axial perspective.

DWI technique allows for the quantification and measuring of water molecules movement within tissues, helping to assess the cellular density, with areas of restricted water diffusion indicating potential cancerous lesions. The b-value identifies properties of the gradients used to create DWI images: the higher the b-value, the stronger the diffusion effects. HBV images emphasize the diffusion of water molecules over shorter timescales and are particularly useful for probing tissues with high cell levels. Inclusion of these imaging has been shown to improve the detection of prostate cancer by several studies [3]. ADC maps are derived from differently weighted DWI images mapping and reproduce the calculated diffusion of water molecules. ADC images complement DWI by offering quantitative information about tissue characteristics. The dataset delineates a significant effort to employ all available formats collectively to enhance classifier detection capabilities, mirroring the diagnostic processes used by experts [17]. Among 1,500 records, 1,075 patients have benign or indolent PCa, while 425 are identified with csPCa. The latter comprises 220 cases with annotations from human experts, leaving 205 positive cases unannotated.

3.1.1 Data Preparation

We apply a preprocessing step to the PI-CAI dataset to adapt it to our classification task. The dataset is curated to include only the most recent medical exam for each patient, with each 3D image converted into a number of slices ranging from 18 to 30. We note that every slice is centered on the prostate gland. The slices, resized to a 384×384 resolution, are cropped to a specific Region Of Interest (ROI) of 64×64 pixels centered on the prostate using the segmentation provided for each patient. Segmentation is generated from algorithms developed in [10, 27]. For each slide, the optimal ROI coordinates are determined by the minimum perimeter of prostate gland segmentation. For each patient, we select the most significant slices, determined by prostate's volume. This was assessed using the segmentation masks provided in the PI-CAI dataset. Specifically, for each mask's volume, we computed the number of segmented pixels per slice, ranking them by visibility of prostatic tissue. Slices with higher segmentation density were selected to ensure maximal coverage of the organ of interest.

We experimented with multiple slice selection strategies, assessing the top $15-10-5$ slices. Selecting the top 5 slices provided to be the best setup. This is likely due to reduced noise and higher signal-to-noise ratio. Therefore, the top 5 slices were chosen for each modality and each patient, which also corresponded to the 5 contingent most central and adjacent slices of the complete scan (see Figure 2).

Patient records containing unaligned prostate acquisitions, which could negatively affect the cropping quality of the prostate area, are excluded. The final version of the dataset consists of 21,222 images, derived from the original MRI of 1,427 distinct patients. The images are labelled according to the ground truth: MRI slices of patients

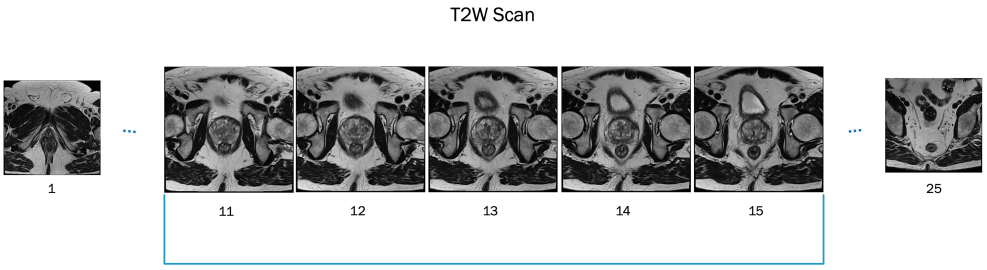

Fig. 2: The five most relevant slices are selected. Here, T2w slices 11–15 show the highest prostate density.

with ISUP scores less than 2 are labelled as non-csPCa, while those with ISUP scores equal to or greater than 2 as csPCa. Data are evaluated using 5-fold stratified group cross-validation at patient level, ensuring that all slices from the same patient appear in exactly one fold. Stratification is performed on the binary clinical significance label (csPCa vs non-csPCa) in order to preserve class proportions across folds. In each fold, models are trained on four folds and evaluated on the held-out fold. Performance metrics are reported as mean ± standard deviation across folds.

To address class imbalance ($\approx 0.3 - 0.7$) towards non-csPCa cases, we applied oversampling of the positive class (csPCa cases) samples to form more balanced mini-batches during training. Additionally, we employed a cost-sensitive weighted binary cross-entropy loss (wbce), with higher weight to the positive ($w_1 = 7$) than the negative samples ($w_0 = 3$) to emphasize errors on the csPCa cases.

3.1.2 Multimodal Fusion

In addition to image from multiple modalities, the dataset contains clinical metadata describing patient conditions. Thus, we propose a hybrid fusion strategy for jointly integrating these tabular data and the image modalities within the same classifier.

As reported in [16], the ways in which heterogeneous data sources can be combined are categorized into three categories. Early fusion is the combination of modalities at the feature level, usually through linear combination or stacking, or in a common latent space through encoding approaches. Late fusion, on the other hand, is the combination of knowledge extracted by processing each data source into a unimodal model. Hybrid fusion is a combination of the two approaches. Intermediate fusion, which is often considered a subgroup of hybrid, is the integration of modalities at an intermediate stage of the pipeline.

We achieve the optimal multimodal learning results through a hybrid fusion approach. First, we select a CNN architecture to classify MRI acquisitions. The CNN is trained using the three image modalities. To integrate clinical information, we adopt a dual strategy: (i) tabular features of metadata are processed through a unimodal fully connected neural network, and the resulting feature vectors are attached to the final multimodal classifier; (ii) tabular data are converted into an image representation via pixel mapping, ensuring that structural relationships among features are preserved during conversion.

To this end, we employ the IGTD [66] algorithm, which assigns features to pixel positions based on similarity and generates the corresponding image for each tabular row. This technique optimizes feature placement by minimizing the distance difference between their rankings and pixel positions in the generated image. As a result of the tabular pixel mapping (TPM), a bi-dimensional vector matrix with a shape of 64×64, matching the dimensions of the other image modalities, is generated for every patient. Due to the limited number of tabular features, the resulting matrix contains only 4 dense pixels, corresponding directly to these features. Given this sparsity, the metadata matrix is processed through padding by duplicating dense values across the image space to enhance feature representation and improve the network's capability to capture meaningful patterns. This allows for integration of tabular metadata into the fourth channel of the input vector, alongside the other image modalities, ensuring cohesive data alignment for classification.

The processed training data are then normalized and enhanced via the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique [41], which improves local contrast by locally adjusting pixel intensities. CLAHE is not applied to TMP matrices, as it would distort the semantic magnitude and intensity relationships inherent to the features. Finally, the input tensor of the model is created by stacking the four different modalities on the third axis according to the neural network input requirements.

3.2 Multimodal Learning

The architecture of our predictive models (Figure 1) is the key aspect to understand the innovative approach taken in this research.

3.2.1 Plain Model

The first step consists to train a classifier exclusively on three MRI modalities as three image channels: ADC, T2w, and DWI. For this purpose, we propose to use a ResNet50-v2 [24]. This choice is motivated by its remarkable ability to capture deep feature hierarchies from complex image data. In the following, we refer to this model as *Plain Model*.

The convolutional layers are designed to extract and learn features from the input images progressively. These layers are configured with varying numbers of filters (from 64 to 2048), and standard kernel sizes (3×3). The initial layers capture basic features like edges and textures, while the deeper layers learn more complex patterns relevant to prostate lesions identification. Each convolutional layer utilizes the ReLU (Rectified Linear Unit) activation function to introduce non-linearity.

Two fully connected layers of 128 units each are added to the stack, before the output layer, which are, in turn, regularized with a standard dropout at 20%. The output layer of the model uses a sigmoid activation function to output a probability, indicating the likelihood that a lesion is clinically significant.

The network, pre-trained on the ImageNet dataset, is then finetuned over the PI-CAI preprocessed dataset for 200 epochs. We opted for SGD optimizer with learning rate of 10^{-5}, which is scheduled to decay by a factor of ten every 40 epochs.

Additionally, a momentum of 0.9 accelerates the optimizer convergence and enhances stability.

3.2.2 Enhanced Model

The next phase aims to extend the capabilities of the Plain Model to accommodate a fourth channel in the width dimension, incorporating tabular metadata embeddings alongside the original three image channels. To adapt the model to this multimodal input, we propose the aforementioned two-step fusion process via transfer learning, resulting in the creation of the *Enhanced Model*.

We first extract the weights from the Plain Model, trained on the three-channel image dataset. To initialize the Enhanced Model, we transfer the pretrained weights from the Plain Model wherever applicable. Since the only architectural difference lies in the number of input channels in the first convolutional layer (3 vs. 4), we retain the pretrained weights for the first three channels and randomly initialize the weights for the fourth channel (corresponding to the TMP). Specifically, we transfer all convolutional layers from the network trained on IGTD-projected TPM. During the multimodal fine-tuning, early convolutional layers are frozen to preserve tabular-induced patterns, while deeper layers are left trainable to adapt to the image modality. This strategy allows the tabular knowledge to serve as a structural prior for the subsequent multimodal representation. This process ensures that the model benefits from the image recognition capabilities developed during the Plain Model training. The second step focuses on integrating the tabular metadata into the model. For this purpose, we restrict our analysis to four tabular features: patient age, PSA (Prostate Specific Antigen), PSAD (Prostate Specific Antigen Density) and prostate volume. A separate neural network, referred to as *Tabular Model*, is trained to classify clinical relevance based solely on tabular metadata. After training, the last feature layer of the Tabular Model is extracted and appended to the final tensor of the Enhanced Model.

Tabular data undergoes a preprocessing step before being fed into the Tabular Model. Standard normalization of numerical variables, encoding of the target variable into a boolean column were applied, along with the removal of records with missing values, to prepare the data for neural network processing. The configuration of the Tabular Model includes a series of three fully connected layers, each with 128 neurons, ReLu activation functions, and dropout with 20% rate to prevent overfitting and ensure model generalizability. Features are extracted from the final layer of the Tabular Model. These features represent distilled insights from the tabular data, capturing essential patterns and relationships. They are then integrated into the Enhanced Model, enriching its predictive capacity.

The addition of tabular data embeddings requires a reconfiguration of the input layer to process a new dimension. The model once again employs a ResNet50-v2 architecture, with the same specifications and training hyperparameters as the Plain Model, but this time with an input expanded to four channels. This includes the three channels from the Plain Model plus an embedding channel constructed from tabular data, standardized and processed as described in Section 3.1. The model was initialized, for the common neuronal structure, with weights extracted from the Plain Model and then finetuned on the 4-channel dataset for an additional 200 epochs. Enriched with

the knowledge from both the Plain and Tabular models, this Enhanced Model achieves superior performance. Detailed results and comparison are presented in Section 4.

3.2.3 Model Training and Optimization Procedure

To enhance reproducibility and clarify the learning pipeline, we provide in Algorithm 1 a high-level pseudocode outlining the multimodal model training and inference process, including tabular pretraining, tabular-to-visual feature transformation, fusion with imaging, and explanation.

Algorithm 1: Multimodal Training and Inference Pipeline

Input: MRI slices I, tabular metadata T, segmentation maps S
Output: Trained multimodal model f with interpretability outputs
`// Slice selection`
foreach *scan i in dataset* **do**
    Extract prostate segmentation density d_i from S_i;
    Select top 5 slices with highest d_i;
    Crop and resize slices to fixed ROI resolution;
`// Tabular model pretraining`
Train tabular model f_T on T;
Compute SHAP values for T and convert into saliency maps M_T;
`// IGTD transformation`
foreach *sample t in T* **do**
    Generate TPM image I_T^t from M_T^t and metadata;
    Concatenate I_T^t as 4th channel with corresponding MRI I;
`// Multimodal CNN training`
Initialize CNN encoder f_I and classification head f_{cls};
Freeze all CNN layers except final block of f_I;
Train $f = f_{cls}(f_I(I, I_T))$ using wbce loss;
`// Explainability`
Apply Grad-CAM on image channels for visual explanation;
Overlay SHAP maps from tabular input for joint interpretation;
return f

To clarify the optimization choices, Table 1 summarizes the key hyperparameters employed across the entire multimodal pipeline. These include not only training settings for the neural networks, but also key preprocessing choices, feature integration mechanisms, and explainability configurations. Hyperparameter values were selected based on preliminary grid searches, literature guidance, and empirical tuning within each training fold.

3.3 Multimodal Explanation

To address the challenge of explaining the predictions of the neural model described above, we propose a global and local approach based on visual and quantitative XAI techniques within an intrinsically multimodal approach. This method extracts explanations directly from the input data modalities, without introducing any additional

Table 1: Hyperparameters and configurations across the proposed pipeline.

Component	Hyperparameter	Value / Setting
Plain Model	Learning rate	10^{-5}
	Optimizer	SGD
	Batch size	64
	Epochs	200
	Input channels	ADC, T2w, DWI
	Pretrained weights	ImageNet
Tabular Model	Hidden layers	[128, 128, 128]
	Activation	ReLU
	Dropout	0.20
	Optimizer	Adam
	Learning rate	10^{-3}
	Batch size	64
	Epochs	50
Multimodal Model	Backbone	ResNet50-v2 (pretrained)
	Fusion method	TPM + concatenation
	Learning rate	10^{-5}
	Optimizer	SGD
	Batch size	64
	Epochs	200
	Pixel Matrix Shape	64×64 spatial tabular channel
	Freezing	Conv layers
Slice Selection Strategy	Number of slices	5
	Selection criterion	Highest density from segmentation masks (prostate density within the segmentation)
	ROI cropping size	64×64
Data Augmentation	CLAHE	$p = 0.5$
Explainability Metrics	Quantitative-based	Segmentation Density and Highest Activation Ratio
	Visual-based	Saliency maps (Grad-CAM)
	Metadata saliency	Feature importance from Tabular Model (SHAP)

modality solely for the purpose of explanation [16]. More specifically, we apply several visual-based local methods to achieve intepretability from the activations of the Enhanced Model. Grad-CAM provides explanations for decisions made by neural networks over single decision outputs, highlighting the important regions in the input image for predicting the model outcome. This method suits particularly our objective, as it allows to analyze the contribution of each channel of the multimodal input. Our focus is on extracting saliency maps from each channel, with a twofold purpose, determining the contribution of MRI channels and of the fourth channel. By analyzing the saliency maps generated for the MRI channels, we identify which of these images contribute most significantly to the model predictions and understand from which parts

of the images such contributions are derived. In addition to the three MRI channels, our model incorporates a fourth channel based on tabular metadata embeddings. Considering the contributions from each of the four channels, we intend to measure the value added by integrating diverse data modalities.

For the channel involving tabular metadata embeddings, we proceed with a quantitative approach involving the extraction of the average contribution rather than a detailed visual saliency map. This approach is justified by the nature of the tabular data channel, where spatial detail is less meaningful, and the emphasis is on understanding the overall influence of tabular information on the model predictions.

The outcomes of this analysis are presented in the next section. This in-depth exploration provides a clearer understanding of the model internal reasoning. Furthermore, in Section 4.4, we go deeper into the significance of feature extraction from the tabular data channel. This analysis is crucial as it complements the multimodal insights obtained from MRI scans by incorporating clinical markers and demographic features. The incorporation of SHapley Additive exPlanations (SHAP) [30] provides a quantitative-based approach to assess the impact of each feature, highlighting their roles in detecting clinically significant prostate cancer cases. The extraction of feature importance supports the findings discussed in this section by validating the relevance of the modalities and regions identified through XAI techniques.

To ensure complete reproducibility of the methods proposed in this work, all scripts for dataset creation, preparation, preprocessing, and model training, as well as the entire implementation, can be found in the project repository [1].

4 Experiments and Results

This section presents our evaluation framework that provides a comprehensive analysis, including accuracy, F1 score, and AUC score metrics, along with visual interpretations given by Grad-CAM saliency maps. Additionally, we compare the models performance with existing literature in this field.

4.1 Model Performance and Ablation Study

The performance of the models is summarized in Table 2, showing their accuracy, F1 score, and AUC score. All reported metrics are computed on the held-out fold and aggregated across the 5 folds as mean $\pm$ standard deviation. The results prove the effectiveness of both models to diagnose clinically significant prostate cancer, at both slice and patient level. Table 2 also shows the improvement of performance introduced by the tabular metadata. The improvement is more evident for the slice level prediction where we can observe an increase of about 3% and 4% for F1 and AUC score, respectively.

In order to dissect the incremental contribution to our model, we conduct an ablation study focusing on three major additions: 1) the inclusion of multiple MRI channels, 2) the application of transfer learning to the last layer of the network, and 3) the inclusion of multimodal embeddings as the fourth channel, stacked on top of

[1] https://github.com/cgiova/multimodal_xaI_prostate

Table 2: Performance metrics of the models with incremental enhancements.

Model	Level	Accuracy	F1 Score	AUC Score
Plain Unimodal	Slice	74.3 ± 0.4	72.7 ± 0.5	73.0 ± 0.4
	Patient	78.2 ± 0.4	77.2 ± 0.3	77.3 ± 0.3
Plain	Slice	85.0 ± 0.3	82.5 ± 0.4	83.0 ± 0.3
	Patient	90.0 ± 0.3	86.5 ± 0.3	89.1 ± 0.2
Plain + Transfer Learning	Slice	85.7 ± 0.3	82.4 ± 0.3	84.6 ± 0.4
	Patient	90.1 ± 0.2	86.6 ± 0.3	90.0 ± 0.2
Plain + Multimodal Embedding	Slice	86.9 ± 0.3	85.0 ± 0.3	86.2 ± 0.3
	Patient	90.4 ± 0.2	**87.5 ± 0.2**	91.1 ± 0.2
Enhanced	Slice	**87.2 ± 0.3**	**86.0 ± 0.2**	**87.2 ± 0.2**
	Patient	**91.0 ± 0.2**	87.2 ± 0.3	**91.2 ± 0.2**

the other three MRI channels. This study is designed to evaluate their individual and combined impacts on model performance.

The ablation study involves evaluating intermediate models that incorporate either of these additions separately. Specifically, we examine the "Plain Unimodal" model, which uses a single image modality (i.e., the most commonly used T2w), the "Plain + Transfer Learning" model, which applies transfer learning to the last layer of the Plain Model, and the "Plain + Multimodal" model, which adds only the multimodal embedding as the fourth channel. The performance of these models is compared against the baseline Plain model and the Enhanced model to understand the relative contributions. Beyond assessing performance, this ablation setup also allows us to examine the interpretability implications of our hybrid approach compared to standard modality modality integration baselines.

The results in Table 2 indicate that both transfer learning and multimodal embedding contribute positively to the model predictive performance. Among these, a significant improvement is observed with the addition of the multimodal embedding. However, using only a single imaging modality, the most commonly used T2w, yields results well below those of the multimodal models and the Plain model with three MRI image channels. Future work may explore further enhancements and the integration of additional data modalities to advance the state-of-the-art in this field.

4.2 Visual-based Explainability

A central part of our analysis involves the use of Grad-CAM to visualize the regions of the MRI scans that most influence the models predictions. A graphic representation is provided in Figure 3, with MRI channels and their corresponding saliency maps.

In our search for the best method to explain models decision, following recent literature [8], we evaluated various visual-based techniques beyond Grad-CAM. We tested

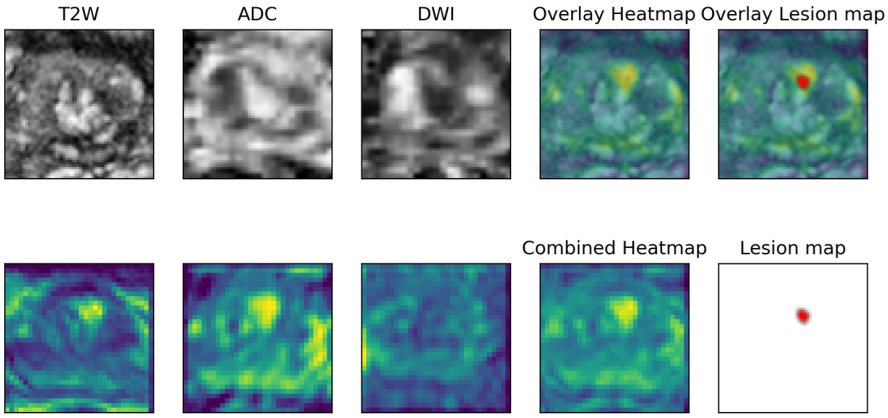

Fig. 3: The top row includes T2w, ADC and DWI images. Below are the saliency maps for each modality, a combined heatmap highlighting key regions, and a lesion segmentation map. The top-right image overlays the T2w scan with the combined heatmap and lesion segmentation.

SmoothGrad [12], a method designed to produce smoother saliency maps by averaging the gradients over multiple noise-perturbed inputs. However, it performed slightly worse in our context. We claim that the metrics we used to evaluate explanation quality (Figure 5) were sensitive to the smoothing property of SmoothGrad, which can worsen the specificity of saliency to the most critical features as it blurs the distinction between areas of low and high activation by averaging them together: we measure a drop from 5% to 10% in the Segmentation Density while the Highest Activation Ratio remains almost unchanged. We formally define these new metrics in the next section.

We also experimented with IntGrad [51], which calculates the features importance by integrating gradients along a path from a baseline image to the input. However, finding an appropriate baseline for the MRI images was difficult and perhaps not entirely feasible in our context. Using a completely black image, as often done, was not effective for our MRI data, possibly because it misrepresented the absence of signals. For the same reason, a white image was discarded as baseline.

RISE [40] is renowned for its ability to highlight important pixels without needing details about the model architecture. However, it falls short in providing channel-specific insight, crucial for our work, as it aggregates effects across the entire image without separating the contributions of different MRI scan types.

LRP [36] is a model-specific method that offers detailed backpropagation of relevance. However, its application demands extensive integration with the model architecture, rendering it impractical for use across various classifiers due to the necessity for unique configurations and calibrations for each model.

After considering these options, we choose Grad-CAM. It achieves the best balance for our needs in terms of clarity, ease of use, and relevance. For a better understanding of model behavior, we conducted an analysis on a sample of 50 cases, measuring the average activation extracted via Grad-CAM across the four channels in the Enhanced Model and the three channels in the Plain Model. Table 3 provides these values and

Table 3: Average (normalized) Grad-CAM activations over 50 image samples.

Channel	Plain Model Density		Enhanced Model Density	
	non-csPCa	csPCa	non-csPCa	csPCa
T2w	0.343	0.345	0.317	0.295
ADC	0.367	0.383	0.296	0.338
DWI	0.285	0.263	0.249	0.225
Tabular Data Embeddings	-	-	0.135	0.140

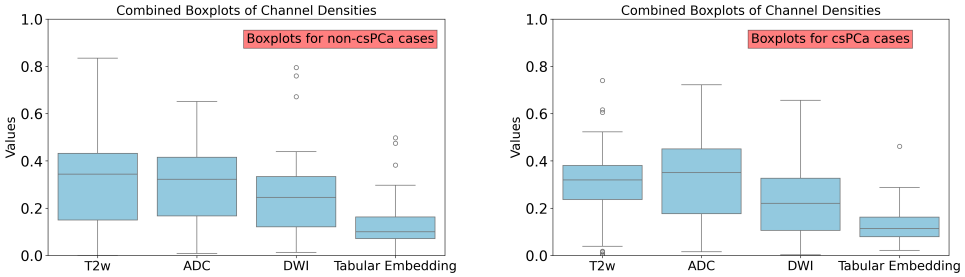

Fig. 4: Boxplots of non-csPCa (left) and csPCa (right) cases across the four distinct channels.

insight into the relative importance and impact of each channel on the model decision-making process. Furthermore, a more granular presentation is shown in Figure 4 with boxplots for each of the four contributions.

4.3 Activation Density Analysis

In order to further explore the saliency map density, we defined two new metrics to evaluate the goodness of the saliency map with respect to the human based image segmentation provided by a domain expert.

- **Segmentation Density:** We consider the ground truth of the lesion segmentation and calculate the average density of the saliency map around this segmentation.
- **Highest Activation Ratio:** Given V the volume of the segmentation, and selecting the top V points from the saliency map with the highest activation, this metric measures how many of these points lie within the saliency map segmented area.

These two metrics reflect the intuition to give high value to the saliency map activations within the area identified by the lesion, ensuring that other areas do not also exhibit high activation outside the lesion area. For instance, a saliency map, in which both the area outlined by the lesion and other areas outside the lesion exhibit strong activation, might have a high Segmentation Density but a low Highest Activation Ratio. Conversely, a saliency map whose points of highest activation reside within the lesion area, but these points are not sufficiently active and visible in the saliency map, would have a high Highest Activation Ratio but a low Segmentation Density.

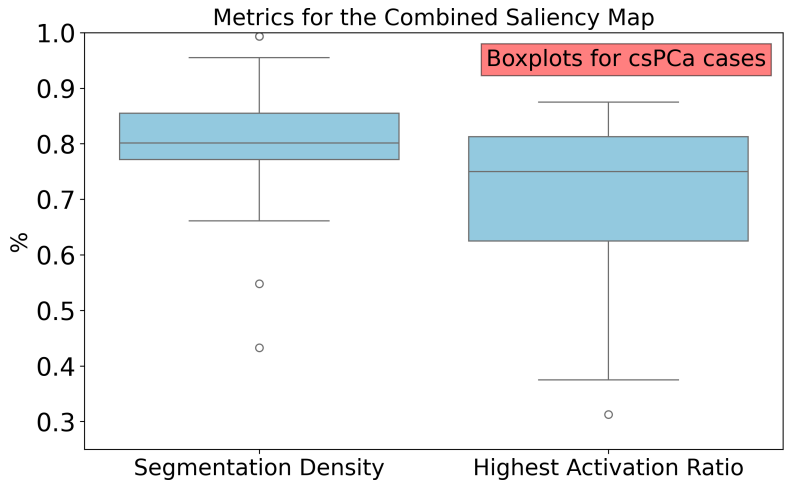

Fig. 5: Saliency maps evaluation through Segmentation Density and Highest Activation Ratio.

We measure these metrics over 40 case studies for which we have the area of the lesion identified by domain experts. Figure 5 reports the results of the saliency maps evaluations for these case studies, proving that the saliency maps exhibit good quality in identifying the lesion area. When comparing Figure 4 with Figure 5, the saliency map shows activation in the lesion area that is, on average, about 3 times that of the entire saliency map average density, further affirming the effectiveness of the methodology employed. To the best of our knowledge, both Segmentation Density and Highest Activation Ratio are novel metrics introduced in this work. We propose them as complementary quantitative tools to evaluate the alignment between saliency maps and expert-annotated lesion areas. These metrics are designed to go beyond qualitative visual assessment by offering interpretable indicators of how precisely the model focuses its attention on clinically relevant regions in medical images.

4.4 Feature Importance from Tabular Data

An important aspect to inspect is the significance of the fourth channel of our dataset, which encapsulates embeddings derived from tabular data. Unlike the first three channels that directly represent MRI scans and possess inherent spatial semantics, the fourth channel embeddings abstract patient-specific tabular information. Due to the non-linear nature of the embedding process used to integrate tabular data, directly attributing the significance of individual features becomes challenging. Consequently, our analysis is focused on evaluating the importance of tabular features as predictors within the Tabular Model, which is trained exclusively on tabular data. It is important to clarify that this SHAP-based analysis reflects only the feature importance within the unimodal Tabular Model, and does not capture the role of tabular data after its transformation into the visual representation and integration into the multimodal CNN. Therefore, this analysis should be interpreted as a complementary insight into the original semantic value of metadata, not as a full explanation of how the

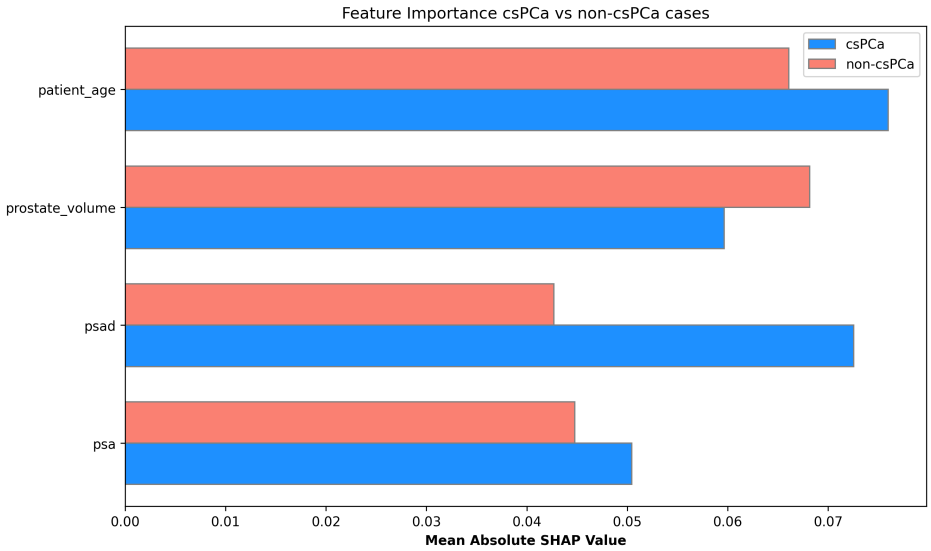

Fig. 6: Comparing mean absolute SHAP values of clinical features for csPCa (blue) and non-csPCa (red).

CNN exploits the fourth input channel during joint training. This approach remains highly pertinent, since the Enhanced Model derives significant value from the Tabular Model insights through the application of transfer learning techniques. We employed a combination of SHAP [30] values analysis and custom graphical plots to estimate the influence of clinical markers and demographic features particularly distinguishing between csPCa and non-csPCa cases.

First, we recall that we focus on four key features extracted from clinical annotations: the volume of the prostate, the level of Prostate-Specific Antigen (PSA), PSA Density (PSAD), and the age of the patient. The analysis is divided in two main parts:

Aggregated Feature Importance. We computed the mean absolute SHAP values for each tabular feature, dividing the data based on the model predictions into positive (csPCa) and negative (non-csPCa) labels. This differentiation enabled us to observe not only the general impact of features at the model level, but also how their relevance varied across distinct diagnostic outcomes. The resulting bar plot (Figure 6) provides a comparative view of feature importance.

Disaggregated Feature Importance. To complement the aggregated view, we plot the SHAP values against each feature value in Figure 7. This visualization shows the influence of individual feature on the model prediction, offering a granular perspective on how specific ranges affect the model prediction. The analysis of the aggregated feature importance (Figure 6), highlights distinct patterns in how features influence the model predictions. Notably, the PSAD shows a prominent mean absolute SHAP value in predicting csPCa, reinforcing its critical role in the detection of clinically significant cases. This observation aligns with clinical expectations, as PSA levels are a well-known marker in prostate cancer screening [2]. Conversely, the prostate volume and patient age appear to have a more pronounced impact in non-csPCa classifications.

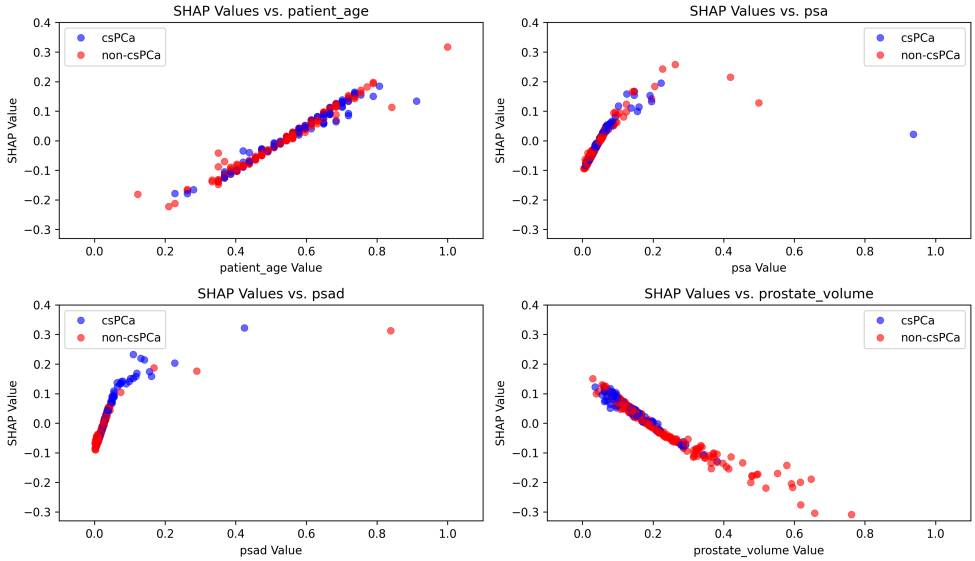

Fig. 7: Mean SHAP values of clinical feature across for csPCa (blue) vs. non-csPCa (red) predictions.

This reflects the community understanding that individual differences in prostate size and how age influences cancer must be considered.

Prostate Volume. A larger prostate volume is often associated with a lower risk of detecting clinically significant prostate cancer using PSA-based screening methods [54]. This is because a larger gland can produce more PSA naturally, which means that a higher PSA level in the context of a larger prostate may not necessarily indicate the presence of aggressive cancer. Instead, it might reflect benign prostatic hyperplasia (BPH), a common condition in older men characterized by prostate enlargement. Studies [49] have shown that when adjusting for PSA levels, men with larger prostates are less likely to have aggressive prostate cancer detected on biopsy, making prostate volume a critical factor in differentiating between BPH and csPCa.

Patient Age. Age is a known risk factor for prostate cancer. However, the relationship between age and the aggressiveness of prostate cancer is complex. Older age is associated with a higher incidence of indolent, less aggressive forms of prostate cancer [11]. Clinically significant prostate cancers, which are more likely to require intervention, have a different age distribution, with a peak incidence in slightly younger age groups. Therefore, in non-csPCa cases, including low-grade tumors that are less likely to progress, age plays a crucial role. Older patients are more likely to be monitored through active surveillance rather than aggressive treatment. The analysis of the disaggregated feature importance (Figure 7), shows how the model considers the features differently when predicting the likelihood of csPCa versus non-csPCa, with each feature having varying degrees of importance based on their SHAP values. Specifically:

- **Patient Age**. The plot shows a positive correlation for csPCa, suggesting that older age contributes to a higher SHAP value, indicating an increased likelihood of csPCa. For non-csPCa, the impact seems less pronounced.
- **PSA**. Both csPCa and non-csPCa show a positive trend, with higher PSA values contributing to higher SHAP values, but it is more pronounced for csPCa, suggesting that PSA is a stronger predictor for csPCa.
- **PSAD**. Similarly to PSA, higher PSAD values correlate with higher SHAP values for csPCa, implying that PSAD is a significant feature for predicting csPCa.
- **Prostate Volume**. There is a negative correlation for csPCa, indicating that smaller prostates have a stronger association with csPCa. For non-csPCa, as prostate volume increases, its association with non-csPCa increases.

4.5 Baseline Comparison

To contextualise the contribution of our hybrid fusion strategy against established multimodal learning paradigms, we implement three representative fusion strategies on the PI-CAI dataset using the same 5-fold stratified cross-validation, class-balancing scheme, and loss weighting described earlier.

Early Fusion: MRI only. Following the channel stacking strategy introduced by Yoo et al. [61] and further validated by Bertelli et al. [7], the three MRI modalities (T2w, ADC, DWI) are stacked as independent input channels of a single ResNet50-v2 classifier, without incorporating any clinical metadata. This configuration corresponds to our Plain model and its results are reproduced from Table 2 as the primary baseline.

Late Fusion: Independent CNN and FCN. Following Taguelmimt et al. [52], a ResNet50-v2 image encoder and our Tabular Model FCN are trained independently; their scalar outputs are then combined via a learned linear layer. Both branches are pre-trained separately and only the combination weights are fine-tuned end-to-end.

Intermediate Fusion: Feature Concatenation. Motivated by Zhang et al. [65], independent ResNet50-v2 encoders process each single-channel MRI modality; their feature vectors are concatenated together with the tabular FCN output before a shared two-layer MLP classifier.

Results are summarised in Table 4. The Enhanced Model outperforms all baselines at both evaluation levels across every metric. Late fusion improves over the Plain model in AUC, yet the decoupled optimisation of the two branches and the absence of cross-modal transfer learning limit its gains, particularly in F1 score at patient level where the improvement over Plain is marginal. Intermediate feature concatenation advances further, suggesting that joint end-to-end training of modality-specific encoders extracts more complementary representations than decision-level combination. Nevertheless, it still falls short of the Enhanced Model by approximately 0.8 AUC points at patient level, demonstrating that our proposed hybrid combination of IGTD-based tabular mapping, channel stacking, and dual transfer learning provides discriminative signal that simpler concatenation strategies cannot replicate.

It is worth noting that the replicated baselines yield slightly lower absolute performance than originally reported in the respective works, which is expected given that PI-CAI is a substantially larger and more heterogeneous dataset than those used

Table 4: Comparison of multimodal fusion strategies on PI-CAI. † Results reproduced from Table 2.

Model	Level	Accuracy	F1 Score	AUC Score
Early Fusion – MRI only [7, 61] (Plain†)	Slice	85.0 ± 0.3	82.5 ± 0.4	83.0 ± 0.3
	Patient	90.0 ± 0.3	86.5 ± 0.3	89.1 ± 0.2
Late Fusion – CNN + FCN [52]	Slice	85.7 ± 0.4	83.6 ± 0.4	84.8 ± 0.3
	Patient	90.3 ± 0.3	86.6 ± 0.4	90.0 ± 0.3
Intermediate Fusion – Feature Concat. [65]	Slice	86.2 ± 0.3	85.8 ± 0.4	86.6 ± 0.3
	Patient	90.4 ± 0.3	86.9 ± 0.3	90.4 ± 0.2
Hybrid Fusion – Enhanced (proposed)	Slice	**87.2 ± 0.3**	**86.0 ± 0.2**	**87.2 ± 0.2**
	Patient	**91.0 ± 0.2**	**87.2 ± 0.3**	**91.2 ± 0.2**

in the original works; larger cohorts tend to reduce overfitting and inflated variance typical of smaller studies, resulting in more conservative but more reliable estimates.

Regarding explainability, our hybrid fusion provides a unified, modality-level explanation process by combining channel-wise GradCAM visual-based XAI across MRI modalities with a quantitative estimate of the tabular channel's contribution within the same model. In contrast, late and intermediate fusion baselines produce branch-specific explanations by separately processing imaging and tabular modalities and integrating them at a late stage, while early fusion (MRI-only) does not include the tabular modality.

4.6 Model Results Discussion

The comparison with existing literature poses a significant challenge in the field of AI-assisted medical imaging, particularly regarding prostate cancer detection. Unlike other domains, where established datasets, benchmarking metrics, and procedures exist, this area faces complications that hinder direct comparisons across studies. **Dataset Variability**: Each study is based on a distinct dataset, varying not only in size but also in the balance between classes. This diversity stems also from the inherently unique cohorts of patients each research effort targets, influenced by geographical, demographic, and disease prevalence factors. **Diverse Data Sources**: The data used in these studies originate from varied medical equipment and institutions, introducing discrepancies in image quality, resolution, and perspectives. Such variability is compounded by differences in medical imaging protocols, scanner settings, and even the interpretation by radiologists across different hospitals or medical centers.

Our work aligns with and, in some aspects, surpasses the recent literature [7, 48, 53, 61], due to the incorporation of multimodal learning. As demonstrated in Section 4.5, the fusion strategies proposed by Taguelmimt et al. [52] and Zhang et al. [65] were directly replicated on PI-CAI and systematically outperformed by

our hybrid approach. Furthermore, advancements include fully automated segmentation models utilizing multi-modal MRI images [64] and interactive explainable deep learning models that inform decision-making [1]. Previous research explores various approaches to classify prostate cancer lesions, employing methods such as pixel-level analysis, 2D and 3D ROI classification, and automated deep learning pipelines. Techniques ranged from VGGNet-inspired CNNs to multi ResNet combinations, achieving an AUC score in the range of 0.73-0.93, with higher values typically reported on smaller and less heterogeneous cohorts.

Our study not only positions itself at the forefront of prostate cancer prediction research but also addresses the limitations of current approaches on the social issues arising from lack of complete model interpretability. The proposed multimodal approach represents a step forward in medical AI applications, offering comprehensive diagnostic insights by integrating diverse data sources. The integration of a multimodal explainability pipeline composed of visual-based, quantitative based, local and global XAI distinguishes our approach, enhancing clinical interpretability and trust in AI-assisted decision-making. We believe that this approach is a solid foundation for future research in the medical domain, advocating a shift towards multimodal approaches in both diagnostic procedures and the field of multimodal explainability.

5 Conclusion

We proposed an interpretable analytical framework leveraging *multimodal deep learning* for the classification of prostate lesions using MRI acquisitions and clinical information in tabular format. To achieve model interpretability, we applied both global and local multimodal explainability by generating visual-based and quantitative-based insights, evaluating the intra-modal relationships among the input modalities. Results show that this approach reaches remarkable performance in terms of predictive power and a high degree of model interpretability, highlighting interesting insights on the two categories of patients.

As for possible improvements and future directions, our hybrid modality fusion strategy inherently presents limitations related to model and dataset dependencies, which may constrain its immediate scalability and generalizability to similar medical scenarios and data. This is typical of multimodal explainability, where the strength of exploiting complementary knowledge among heterogeneous training data to improve performance often comes at the expense of effortless scalability and direct reproducibility. Within the medical AI domain specifically, deep learning models often face performance degradation when applied to data collected in contexts or hospitals different from those used during training, emphasizing the inherent complexity of deploying such methods across diverse medical settings.

We believe that the scalability issues in our approach are outweighed by the substantial social benefits provided through multimodal explainability. By integrating multiple XAI techniques across diverse modalities and interpretability scopes, our approach significantly contributes to opening the black-box models. This enhanced interpretability represents, in our view, one of the most important aspects in fostering trust, validation, and adoption of AI systems in real-world clinical practice.

Nonetheless, we have dedicated significant efforts to ensure thorough documentation of our approach and have made our source code publicly available, in the attempt of increase transparency and encourage reproducibility.

Another limitation of the current study is the reliance on static datasets, which might not capture the dynamic nature of clinical environments where new data continually emerges. Although a very small section patients had multiple studies, we selected only the most clinically relevant examination, as our primary objective was to introduce and validate an end-to-end multimodal explainability approach specifically tailored for prostate cancer detection. Nonetheless, this decision limits the model's ability to capture temporal variations that naturally occur in clinical practice, representing an important consideration for future extensions of this work.

As for model performance, the dependence on the quality and diversity of the data used for training can be seen as a potential limit to its effectiveness in broader, more demographically heterogeneous populations. Although we mitigate partition sensitivity through 5-fold patient-level cross-validation, external validation on independent cohorts remains an important direction for future work. In particular, generalizability to unseen patient cohorts or institutions using different acquisition protocols remains an open challenge. A viable direction involves leveraging domain adaptation or continual learning techniques to adjust the trained models when deployed in new clinical settings, ensuring robustness across heterogeneous populations.

Future research should improve the robustness and utility of our findings, with particular emphasis on adapting our multimodal deep learning frameworks for real-time clinical application. Additionally, the potential of using more recent deep learning-based predictive models should not be underestimated: examples include ConvNext [29], DenseNet [26] or Dendrite Active Connection [35]. Moreover, the use of more sophisticated and tailored explainability models that specifically address the nuances of multimodal data, as well as the exploration of more interpretable models [4, 13] should also be considered in future developments.

Regarding the XAI step of the pipeline, our explanation process preserves an intrinsically multimodal approach which, as previously described, aligns the explanation modalities with those used during training. However, we acknowledge that extending this to an extrinsically multimodal process (i.e. generating explanations though modalities not originally present in the model input data) could offer additional significant benefits in terms of accessibility to medical experts, contributing to an overall sounder interpretability process. On a similar note, a particularly promising avenue is the advancement of multimodal XAI techniques that support real-time, on-demand, personalized explanations through the enrichment of the explanation process based on the specific user. These enhancements should focus on explaining the contributions of individual modalities within a multimodal approach.

Another limitation of our current framework lies in the lack of fine-grained attribution techniques directly applied to the tabular-pixel map in the fourth channel. While we provide global feature importance from the Tabular Model and activation-based insights from Grad-CAM, we do not explicitly isolate the CNN's learned behavior over the metadata image channel. Future work will explore channel-specific attribution

methods, such as Layer-wise Relevance Propagation, on this specific input to quantify its exact contribution to the model's decision-making process.

As for the dataset employed, although the PI-CAI challenge dataset offers a large-scale and multicenter benchmark, its curated nature may not reflect the variability of real-world clinical environments. Recent commentary on PI-CAI [21] emphasizes the challenges of generalizability in prostate cancer detection using AI. We acknowledge that deep learning models often fail to transfer well across hospitals or acquisition protocols, particularly in medical imaging. To address this, future research must include rigorous hyperparameter sensitivity analysis, external validation even on smaller datasets, and fairness audits across demographic variables such as age and ethnicity. These efforts are essential to ensure robust and equitable deployment of AI models in prostate cancer diagnostics.

In conclusion, while this study marks substantial progress in applying AI to diagnose prostate cancer, pursuing these suggested directions is vital. Advancing these areas will address current limitations and significantly boost the real-world impact of AI technologies in medical diagnostics, ensuring that these systems are not only advanced but also aligned closely with the needs of real clinical practice.

Acknowledgements. Funded by the EU under G.A. 101120763-TANGO. Research also funded by PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the EU under NextGeneration EU program and under Horizon 2020 programme: G.A. 871042 *SoBigData++*, G.A. 101092749 *CREXDATA*, ERC-2018-ADG G.A. 834756 *XAI*, G.A. 952215 *TAILOR* and "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013.

References

[1] (2023) Interactive explainable deep learning model informs prostate cancer diagnosis at mri. Radiology 307(4):e222276

[2] Adhyam M, Gupta AK (2012) A review on the clinical utility of psa in cancer prostate. Indian J Surg Oncol 3(2)

[3] Agarwal HK, Mertan FV, Sankineni S, et al (2017) Optimal high b-value for diffusion weighted MRI in diagnosing high risk prostate cancers in the peripheral zone. Journal of magnetic resonance imaging : JMRI 45(1):125–131

[4] Angelov P, Kangin D, Zhang Z (2023) Towards interpretable-by-design deep learning algorithms. arXiv:2311.11396

[5] B S, Bhargavi MS (2023) An xai approach to predictive analytics of pancreatic cancer. In: 2023 ICIT, pp 343–348

[6] Baltrušaitis T, Ahuja C, Morency LP (2019) Multimodal machine learning: A survey and taxonomy. IEEE Trans on Pattern Anal and Mach Intel 41(2):423–443

[7] Bertelli E, Mercatelli L, Marzi C, et al (2022) Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric mri. Front Oncol

[8] Bodria F, Giannotti F, Guidotti R, et al (2023) Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc 37:1719–1778

[9] Bosma J, et al. (2023) The PI-CAI challenge - grand challenge

[10] Bosma J, Saha A, Hosseinzadeh M, et al (2023) Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI. Radiology: Artificial Intelligence 5(5):e230031

[11] Clark R, Vesprini D, Narod SA (2022) The effect of age on prostate cancer survival. Cancers (Basel) 14(17):409–412

[12] Daniel S, Nikhil T, Been K, et al (2017) Smoothgrad: removing noise by adding noise. In: Workshop on Visualization for Deep Learning, ICML

[13] Di Cecco A, Metta C, Fantozzi M, et al (2024) Glonets: Globally connected neural networks. In: Advances in Intelligent Data Analysis XXII. Springer, pp 53–64

[14] Dujuan WWang nad Xinwei, Sutong Wang amd Yunqiang Y (2023) Explainable multitask shapley explanation networks for real-time polyp diagnosis in videos. IEEE Transactions on Industrial Informatics 19(6):7780–7789

[15] Geert L, Thijs K, Babak EB, et al (2017) A survey on deep learning in medical image analysis. Med Image Analytics 42

[16] Giovannoni C, Metta C, Monreale A, et al (2025) A survey on multimodal explainable artificial intelligence, under submission

[17] Grant KB, Agarwal HK, Shih JH, et al (2025) Comparison of calculated and acquired high b value diffusion-weighted imaging in prostate cancer. Abdominal Imaging 40(3):578–586

[18] Grover VP, Tognarelli JM, Crossey MM, et al (2025) Magnetic resonance imaging: Principles and techniques: Lessons for clinicians. Journal of Clinical and Experimental Hepatology 5(3):246–255

[19] Guidotti R, Monreale A, Ruggieri S, et al (2018) A survey of methods for explaining black box models. ACM Computing Surveys 51(93)

[20] Guidotti R, Monreale A, Ruggieri S, et al (2019) Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems 34(6):14–23

[21] Haider MA (2025) PI-CAI: a landmark AI study for prostate cancer detection on MRI. European Radiology 35(8):4930–4931

[22] Hamm CA, Baumgärtner GL, Biessmann F, et al (2023) Interactive explainable deep learning model informs prostate cancer diagnosis at mri. Radiology 307

[23] Hassan MR, Islam MF, Uddin MZ, et al (2022) Prostate cancer classification from ultrasound and mri images using deep learning based explainable artificial intelligence. Future Generation Computer Systems 127:462–472

[24] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 IEEE CVPR, pp 770–778

[25] He L, Li H, Chen M, et al (2021) Deep multimodal learning from mri and clinical data for early prediction of neurodevelopmental deficits in very preterm infants. Frontiers in Neuroscience 15

[26] Huang G, Liu Z, Maaten LVD, et al (2017) Densely connected convolutional networks. In: 2017 IEEE CVPR. IEEE Computer Society, pp 2261–2269

[27] Joeran B, Anindo S, Matin H, et al (2023) Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. Radiology Artificial Intelligence 5(5)

[28] Litjens G, Debats O, Barentsz J, et al (2022) SPIE-AAPM PROSTATEx Challenge Data (Version 2) [dataset]

[29] Liu Z, Mao H, Wu CY, et al (2022) A convnet for the 2020s. In: Proceedings of the IEEE/CVF CVPR, pp 11976–11986

[30] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: NeurIPS, pp 4765–4774

[31] Martino FD, Delmastro F (2023) Explainable ai for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev 56

[32] Metta C, Guidotti R, Yin Y, et al (2021) Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling. In: IEEE SCC

[33] Metta C, Beretta A, Guidotti R, et al (2023) Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning. IJDA

[34] Metta C, Beretta A, Guidotti R, et al (2024) Advancing dermatological diagnostics: Interpretable ai for enhanced skin lesion classification. Diagnostic 14(7)

[35] Metta C, Fantozzi M, Papini A, et al (2024) Increasing biases can be more efficient than increasing weights. In: IEEE/CVF WACV

[36] Montavon G, Binder A, Lapuschkin S, et al (2019) Layer-Wise Relevance Propagation: An Overview, Springer International Publishing, pp 193–209

[37] Pahud de Mortanges A, Luo H, Shu SZ, et al (2024) Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. npj Digital Medicine 7(1):195

[38] Nagendran M, Festor P, Komorowski M, et al (2023) Quantifying the impact of ai recommendations with explanations on prescription decision making. Digit Med 6

[39] Panigutti C, Perotti A, Pedreschi D (2020) Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: Conference on Fairness, Accountability, and Transparency, pp 629–639

[40] Petsiuk V, Das A, Saenko K (2018) Rise: Randomized input sampling for explanation of black-box models. ArXiv

[41] Pizer SM, Amburn EP, Austin JD, et al (1987) Adaptive histogram equalization and its variations. Comp Vision, Graphics, and Image Processing 39(3):355–368

[42] Ramírez-Mena A, Andrés-León E, Alvarez-Cubero MJ, et al (2023) Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. Computer Methods and Programs in Biomedicine 240:107719

[43] Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? explaining the predictions of any classifier. In: ACM SIGKDD, pp 1135–1144

[44] Rui S, Tianxing W, Jianlong W, et al (2024) Detecting and grounding multimodal media manipulation and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(8):5556–5574

[45] Saha A, Bosma J, et al. (2024) Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. The Lancet Oncology 25(7):879–887. Publisher: Elsevier

[46] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE ICCV, pp 618–626

[47] Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. Annual Review of Biomedical Engineering 19

[48] Singhal N, Soni S, Bonthu S, et al (2022) A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. Sci Rep 12

[49] Stephan C, Lein M, Jung K, et al (1997) The influence of prostate volume on the ratio of free to total prostate specific antigen in serum of patients with prostate carcinoma and benign prostate hyperplasia. Cancer 79(1):104–109

[50] Suh J, Yoo S, Park J, et al (2020) Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. BJU Int

126(6)

[51] Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International Conference on Machine Learning, vol 70. PMLR, pp 3319–3328

[52] Taguelmimt K, Andrade-Miranda G, Harb H, et al (2025) Towards more reliable prostate cancer detection: Incorporating clinical data and uncertainty in mri deep learning. Computers in Biology and Medicine 194:110440

[53] Takeuchi T, Hattori-Kato M, Okuno Y, et al (2019) Prediction of prostate cancer by deep learning with multilayer artificial neural network. Can Urol Assoc J 13(5)

[54] Tang P, Jin XL, Uhlman M, et al (2013) Prostate volume as an independent predictor of prostate cancer in men with psa of 10-50 ng ml(-1). Asian J Androl 15(3):409–412

[55] Van Leenders GJ, Van Der Kwast TH, Grignon DJ, et al (2020) The 2019 international society of urological pathology (isup) consensus conference on grading of prostatic carcinoma. American Journal of Surgical Pathology 44(8):E87 – E99

[56] Wang D, Wang X, Wang S, et al (2023) Explainable multitask shapley explanation networks for real-time polyp diagnosis in videos. Trans on Ind Inf 19(6):7780–7789

[57] Wang MH, Chong KKl, Lin Z, et al (2023) An Explainable Artificial Intelligence-Based Robustness Optimization Approach for Age-Related Macular Degeneration Detection Based on Medical IOT Systems. Electronics 12(12):2697

[58] Wang S, Yin Y, Wang D, et al (2022) Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis. IEEE Transactions on Cybernetics 52(12):12623–12637

[59] Wani NA, Kumar R, Bedi J (2024) Deepxplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. Computer Methods and Programs in Biomedicine 243:107879

[60] Wenping M, Mengru M, Licheng Jea (2024) An adaptive migration collaborative network for multimodal image classification. IEEE Transactions on Neural Networks and Learning Systems 35(8):10935–10949

[61] Yoo S, Gujrathi I, Haider MA, et al (2019) Prostate cancer detection using deep convolutional neural networks. Sci Rep 9

[62] Yuhas B, Goldstein M, Sejnowski T (1989) Integration of acoustic and visual speech signals using neural networks. IEEE Comm Magazine 27(11):65–71

[63] Zeineldin RA, Karar ME, Elshaer Z, et al (2022) Explainability of deep neural networks for mri analysis of brain tumors. Int Journal CARS 17

[64] Zhang Y, Ma X, Li M, et al (2025) Generalist medical foundation model improves prostate cancer segmentation from multimodal mri images. npj Dig Med 8(1):372

[65] Zhang YF, Zhou C, Guo S, et al (2024) Deep learning algorithm-based multimodal mri radiomics and pathomics data improve prediction of bone metastases in primary prostate cancer. Journal of Cancer Research and Clinical Onc 150(2):78

[66] Zhu Y, Brettin T, Xia F, et al (2021) Converting tabular data into images for deep learning with convolutional neural networks. Scientific Reports 11(1):11325. Number: 1 Publisher: Nature Publishing Group