This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3530156

Automated Ensemble Multimodal Machine Learning for Healthcare

Fergus Imrie*, Stefan Denner*, Lucas S. Brunschwig, Klaus Maier-Hein and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*— The application of machine learning in medicine and healthcare has led to the creation of numerous diagnostic and prognostic models. However, despite their success, current approaches generally issue predictions using data from a single modality. This stands in stark contrast with clinician decision-making which employs diverse information from multiple sources. While several multimodal machine learning approaches exist, significant challenges in developing multimodal systems remain that are hindering clinical adoption. In this paper, we introduce a multimodal framework, AutoPrognosis-M, that enables the integration of structured clinical (tabular) data and medical imaging using automated machine learning. AutoPrognosis-M incorporates 17 imaging models, including convolutional neural networks and vision transformers, and three distinct multimodal fusion strategies. In an illustrative application using a multimodal skin lesion dataset, we highlight the importance of multimodal machine learning and the power of combining multiple fusion strategies using ensemble learning. We have open-sourced our framework as a tool for the community and hope it will accelerate the uptake of multimodal machine learning in healthcare and spur further innovation.**

***Index Terms*— Multimodal Machine Learning, Medical Imaging, Machine Learning, Deep Learning, Automated Machine Learning, Cancer, Biomedicine, Healthcare Informatics**

I. Introduction

Healthcare data is increasingly diverse in origin and nature, encompassing patient records and imaging to genetic information and real-time biometrics. Machine learning (ML) can learn complex relationships from data and has shown promise for diagnostic and prognostic modeling [1], [2]. However, such approaches typically only use one type or modality of data [3], limiting their ability to consider the broader clinical context.

In contrast, clinicians make decisions based on the synthesis of information from multiple sources [4], with numerous studies showing the absence of such information can result in lower performance and decreased clinical utility [5], [6]. This is perhaps particularly the case in medical imaging. For example, almost 90% of radiologists reported that additional clinical information was important and could change diagnoses compared to imaging alone [7]. Numerous other examples exist across many specialties such as ophthalmology [8], pathology [9], and dermatology [10].

Multimodal machine learning integrates multiple types and sources of data, offering a more holistic approach that mirrors clinical decision-making processes. While multimodal ML remains in its infancy, models that incorporate multiple data modalities have been developed in areas such as cardiology [11], dermatology [12], oncology [13], [14], and radiology [15]. However, technical challenges in developing, understanding, and deploying multimodal ML systems currently prevent broad adoption in medicine beyond bespoke examples.

One possible solution to these challenges is automated machine learning (AutoML) [16]. AutoML can help design powerful ML pipelines by determining the most appropriate modeling and hyperparameter choices, while requiring minimal technical expertise from the user. To bridge the gap between clinicians and cutting-edge ML, we previously proposed an AutoML approach, AutoPrognosis [17], for constructing ensemble ML-based diagnostic and prognostic models using structured data. While AutoPrognosis has been used to develop clinical models for a number of outcomes [18]–[22], it is constrained to tabular features. Several other frameworks for automated pipeline optimization, such as Auto-sklearn [23], Auto-Weka [24], and TPOT [25], suffer the same limitation.

In this paper, we propose a general-purpose framework, AutoPrognosis-Multimodal (AutoPrognosis-M), that uses AutoML and ensemble learning to construct models that incorporate data from multiple modalities, namely imaging and tabular data. AutoPrognosis-M contains state-of-the-art imaging and tabular models and, to the best of our knowledge, is the first approach to consider multiple multimodal fusion strategies. Additionally, AutoPrognosis-M enables such models to be interrogated with explainable AI (XAI) and provides uncertainty estimates using conformal prediction, aiding understanding and helping build model trust [26].

*Fergus Imrie and Stefan Denner contributed equally.

Fergus Imrie is with the Department of Statistics, University of Oxford, Oxford, UK (e-mail: fergus.imrie@stats.ox.ac.uk).

Stefan Denner is with the Division of Medical Image Computing, German Cancer Research Center (DKFZ), Germany, and Faculty of Mathematics and Computer Science, Heidelberg University, Germany (e-mail: stefan.denner@dkfz-heidelberg.de).

Lucas S. Brunschwig is with École Polytechnique Fédérale de Lausanne, Switzerland (e-mail: lucas.brunschwig@epfl.ch).

Klaus Maier-Hein is with the Division of Medical Image Computing, German Cancer Research Center (DKFZ), Germany, Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Germany, and National Center for Tumor Diseases (NCT) Heidelberg, Germany (e-mail: k.maier-hein@dkfz-heidelberg.de).

Mihaela van der Schaar is with Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK (e-mail: mv472@cam.ac.uk).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3530156

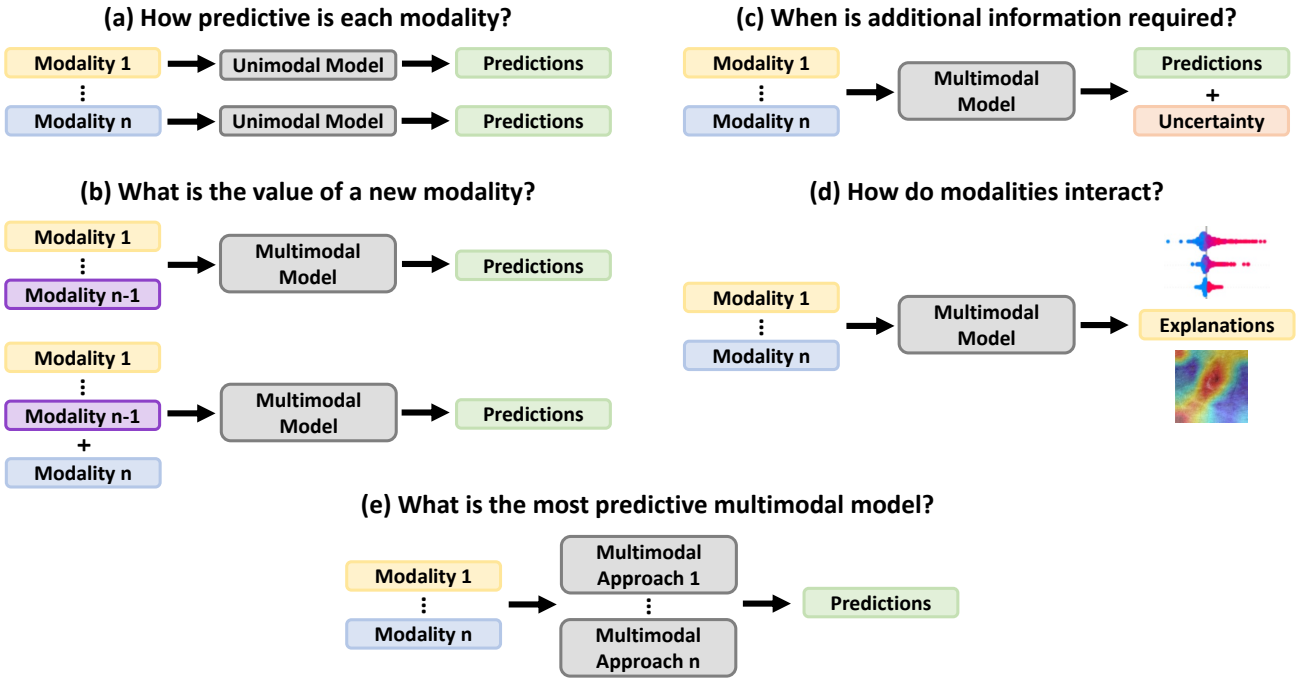


Fig. 1: **Overview of the types of questions that can be asked with multimodal machine learning.** In addition to developing powerful multimodal models (e), multimodal ML can help understand the value of each modality (a), the impact of adding a new modality (b), when an additional modality is required (c), and how the information in different modalities interacts (d).

We applied our approach in an illustrative clinical scenario: skin lesion diagnosis using both images and clinical features. Our experiments demonstrate the benefit of incorporating information from multiple modalities and highlight the impact of the multimodal learning strategy on model performance. We show different multimodal fusion strategies can be effectively combined to form ensemble models that substantially outperform any individual approach. Additionally, we quantify the value of information from each modality and show how our framework can be used to determine whether additional data is necessary on an individual patient basis.

While our experiments focus on skin lesion diagnosis, we emphasize that AutoPrognosis-M is a general-purpose approach that can be applied to any disease or clinical outcome without requiring substantial ML expertise, and can help answer a range of clinical questions (Fig. 1). We have open-sourced AutoPrognosis-M to aid the clinical adoption of multimodal ML models.

II. Methods: AutoPrognosis-M

AutoPrognosis-M enables clinicians and other users to develop diagnostic and prognostic models using state-of-the-art multimodal ML (Fig. 2). Perhaps the most significant challenge is the complex design space of model architectures and associated hyperparameters, which must be set appropriately for the specific task and data being considered. Failure to do so can significantly degrade performance; however, this often requires significant ML knowledge and expertise. This is further compounded in the multimodal setting by the different possible ways of integrating data from multiple sources.

To address this, our framework employs AutoML [24] to efficiently and effectively search the model and hyperparameter space, considering multiple fusion strategies, to construct powerful ensemble multimodal ML models. While incorporating multiple modalities can improve predictive power, this may not always be the case or a modality might be particularly expensive to acquire. AutoPrognosis-M can optimize both unimodal models and multimodal models, allowing the value of each modality to be assessed individually and the added value of an additional modality to be understood (Fig. 1a-b). Additionally, AutoPrognosis-M can identify when more information could be required using uncertainty estimation (Fig. 1c) and enables models to be debugged and understood using explainable AI (Fig. 1d).

By automating the optimization of ML pipelines across multiple modalities and determining the most suitable way of combining the information from distinct sources, we reduce the technical barrier for non-ML experts, such as clinicians and healthcare professionals, to develop multimodal ML models for problems in healthcare. We believe that AutoPrognosis-M significantly simplifies the process of training and validating multimodal ML models without compromising on the expressiveness or quality of the ML models considered.

A. Automated Machine Learning

AutoML aims to simplify and automate the process of designing and training ML models, thereby reducing the technical capabilities required to develop effective models. Human practitioners have biases about what model architectures or hyperparameters will provide the best results for a specific

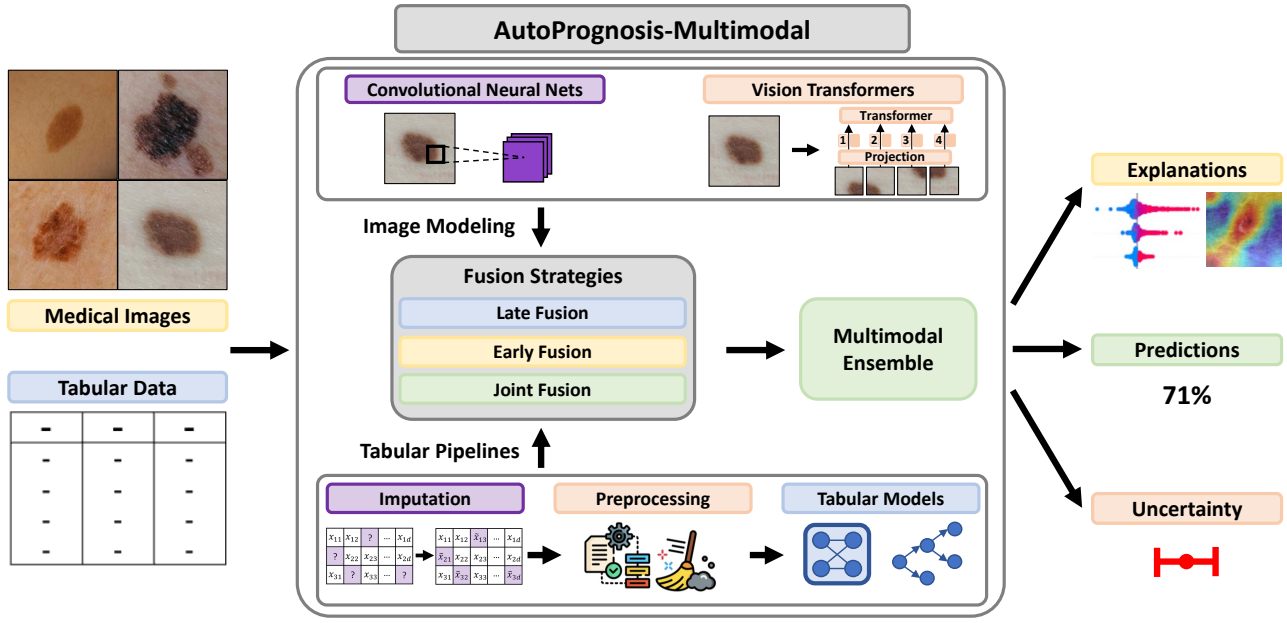


Fig. 2: **Overview of AutoPrognosis-M.** AutoPrognosis-M leverages automated machine learning to produce multimodal ensembles by optimizing state-of-the-art image and tabular modeling approaches across three fusion strategies. AutoPrognosis-M also enables such models to be interrogated with explainable AI and provides uncertainty estimates using conformal prediction.

task. While this might be helpful in some cases, often it will not and can cause inconsistency in the quality of the final predictive system [16]. AutoML helps minimize these biases by automatically searching over a more general set of models, hyperparameters, and other design choices to optimize a given objective function, returning the best configurations found. Beyond simply minimizing human biases, AutoML reduces the demand for human experts and has been shown to typically match or exceed the skilled human performance [27].

B. Unimodal approaches

1) Tabular: We implemented the tabular component of our multimodal framework using AutoPrognosis 2.0 [17]. In contrast to many other approaches for learning from tabular data, we consider full ML *pipelines*, rather than just predictive models, consisting of missing data imputation, feature processing, model selection, and hyperparameter optimization. AutoPrognosis includes 24 classification models, nine imputation methods, five dimensionality reduction, and six feature scaling algorithms. The classification approaches include linear models, such as logistic regression [28], tree-based approaches, such as random forests [29] and XGBoost [30], and neural networks [31]. Additionally, we extended the original implementation of AutoPrognosis to incorporate two recent deep learning approaches for tabular data, TANGOS [32] and FT-Transformer [33]. TANGOS [32] introduces a novel regularization for multi-layer perceptrons (MLPs) that encourages orthogonalization and specialization of latent attributions. FT-Transformer [33] is a transformer adapted for tabular data. A self-attention-based transformer is used to process embeddings of categorical and continuous features together with a classification token, which is then passed to a prediction head. The best-performing pipelines are combined in an ensemble via either a learned weighting or stacking, where a meta-model is trained on the output of the underlying pipelines. For further details about AutoPrognosis, we refer the reader to Imrie et al. [17]; detailed descriptions of the classification algorithms can be found in the original publications or James et al. [34].

2) Imaging: For imaging tasks, we employ several distinct model architectures to cater to a wide range of diagnostic and prognostic applications. Two main classes of deep learning models exist for processing images: convolutional neural networks (CNNs) [35] and vision transformers (ViTs) [36]. CNNs use convolutional layers that exploit local connectivity to extract spatial features. Due to the use of shared features and small receptive fields, convolutional layers are relatively parameter efficient. ViTs divide images into patches, which are flattened and passed to a sequence of transformer layers. The self-attention mechanism allows ViTs to learn long-range relationships between patches more readily than CNNs. These methods provide complementary strengths for image analysis, and thus we included several models from each class. Specifically, we utilized ResNet [37], EfficientNet [38], and MobileNetV2 [39] CNN architectures, as well as the standard ViT architecture [36]. Each model architecture is available in several sizes to be able to handle different task complexities. A full list of the 17 imaging models provided, together with additional details, can be found in Table III.

While general-purpose models for tabular data do not exist, the transferability of imaging models has been shown in a diverse range of disciplines [40], [41], and can be particularly effective when there is limited available data for a particular problem. Pretrained models are especially useful when the

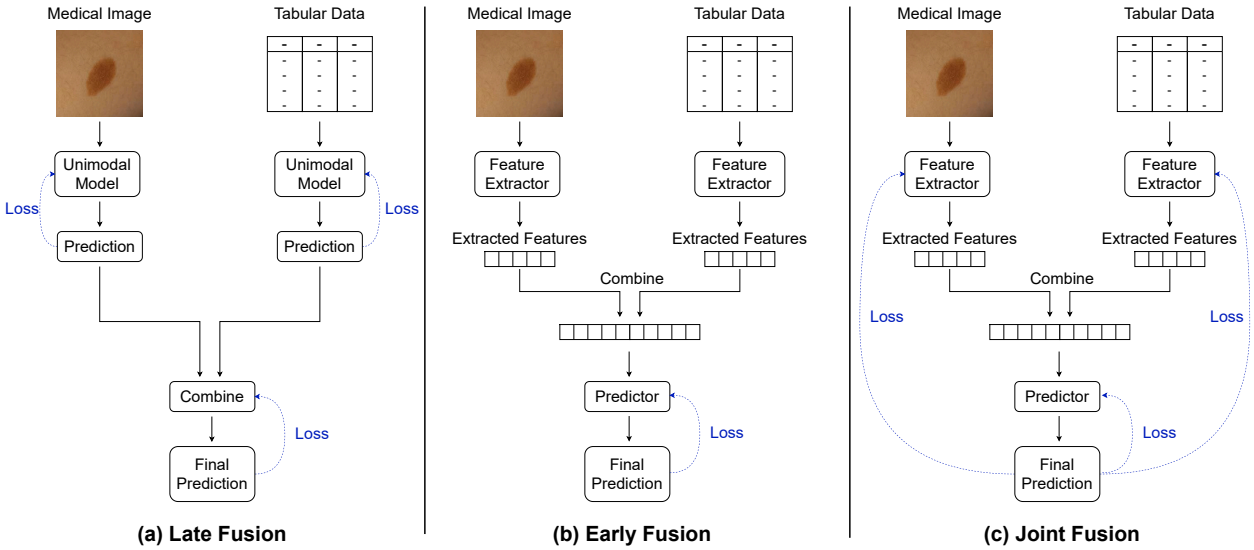


Fig. 3: **Illustration of the three types of multimodal fusion.** (a) Late fusion combines the predictions of separate unimodal models. (b) Early fusion trains a predictive model on the combination of fixed extracted features. (c) Joint fusion flexibly integrates multiple modalities, learning to extract representations and make predictions simultaneously in an end-to-end manner.

available data for a specific task is scarce or when computational resources are limited. They allow one to leverage learned features and patterns from vast datasets, potentially improving performance on related tasks.

While most models are pretrained in a supervised manner, self-supervised pretraining has been shown to improve performance on many classification tasks. Thus, in addition to supervised ViT [36], we also consider DINOv2 [42]. DINOv2 was developed by first training a 1B parameter ViT model on the LVD-142M dataset using self-supervised learning and then distilling this model into a series of smaller models.

One approach to using pretrained models is to extract a fixed representation and train a new classification head on the available task-specific data. However, often these representations are not well adapted for the specific data under consideration, especially when transferring from the natural image to the medical image domain. In this case, we can train these models further by fine-tuning the entire model on the available data. Fine-tuning is most important when the target task or domain is related but not identical to the one on which the model was originally trained by adapting the generalized capabilities of the pretrained model to specific needs without the necessity of training a model from scratch. The optimal training strategy depends on the specific task, availability of data, and computational resources. AutoPrognosis-M can be used to train vision models from scratch, use their existing representations, or fine-tune on the available data.

C. Multimodal data integration

Multimodal ML seeks to effectively integrate multiple types of data to enable more accurate predictions than can be obtained using any single modality. Modalities can exhibit different relationships, including redundancy, depending on the interactions between the information contained in each source [43], which can additionally vary in complexity. Thus a key challenge is discerning the relation between modalities and learning how to integrate modalities most effectively.

Three main multimodal strategies exist [4], [44], [45]: Late Fusion, Early Fusion, and Joint Fusion (Fig. 3). Multimodal architectures can typically be decomposed into two components: modality-specific representations and a joint prediction [43]. Multimodal learning strategies differ primarily in the nature of these components and whether they are jointly or separately learned. We incorporated approaches from each fusion strategy in AutoPrognosis-M.

1) Late fusion: Late fusion is an ensemble-based approach that combines predictions from multiple unimodal models, and thus is sometimes referred to as decision-level fusion [46] (Fig. 3a). Each modality is processed independently using a modality-specific model before the predictions are combined. This allows the user to find the best classifier for each modality independently and evaluate whether each modality has predictive power for the original task. However, this has the drawback of not permitting interactions between modalities before the final output, which could result in suboptimal performance when the relationship between modalities is crucial for making accurate predictions or decisions. One benefit of late fusion is the ability to incorporate new modalities by adding an additional unimodal model and retraining only the ensembling step. We implemented late fusion using a weighted combination of unimodal tabular and imaging models.

2) Early fusion: Early fusion necessitates the translation of each modality into a representation that can be combined using a fusion module, such as concatenation, into a single, unified representation [44] (Fig. 3b). The combined representation is then used as input to train a separate predictive model.

Compared to late fusion, early fusion allows interactions between different modalities to be captured by the predictive model. However, the representations are fixed and translating different data types into effective (latent) representations to

form a unified representation can be challenging, especially when dealing with heterogeneous data sources with differing scales, dimensions, or types of information.

For image data, a common strategy is to use an intermediate (or the last) layer of a vision model that was either trained to solve the relevant prediction task or a general-purpose imaging model. For tabular data, especially if the number of features is relatively modest, the representation step can be skipped and the original features are directly combined with the latent representations extracted from the other modalities. We used concatenation to combine the imaging and tabular features and trained a fully connected neural network on this representation.

3) Joint fusion: The fixed, independent representations used in early fusion may not capture relevant factors for the joint prediction task. Joint fusion [4] (or intermediate fusion [45]) aims to improve these representations using joint optimization to enable both cross-modal relationships and modality-specific features to be learned using end-to-end training (Fig. 3c). The added flexibility of joint fusion can come at the cost of potentially overfitting, especially in the limited data setting.

The most popular approaches for joint fusion use differentiable unimodal models to produce latent representations that are combined via concatenation and passed to a prediction head. The system is then trained in an end-to-end manner, either from scratch or using pretrained unimodal models. We implemented joint fusion similarly to early fusion, except we trained end-to-end.

D. Fusion ensembles

One major challenge is determining which fusion approach is best. Further, no individual fusion approach may be universally best for all patients. Ensembling has repeatedly been shown to improve performance, even when multiple copies of the same model are trained, but can be particularly beneficial when different approaches are combined due to the increased diversity of predictions [47]. The three fusion approaches learn to combine the information from multiple modalities in distinct ways. Thus, combining these different strategies via an ensembling approach could improve both the absolute performance and the robustness of the final model.

We combined the best-performing unimodal and multimodal fusion approaches in a weighted ensemble as follows. We seek to optimize the following objective:

$$\begin{aligned} \arg\max_{w} \quad & \mathbb{E}_{x,y\sim p_{XY}}\left[\rho\left(\sum_i w_i f_i(x), y\right)\right] \\ \text{subject to} \quad & \sum_i w_i = 1,\ w_i \in [0,1], \end{aligned}$$

where ρ is a quantitative measure of performance, f_i are the prediction pipelines that will form the ensemble, and w_i are the ensemble weights for each pipeline. In practice, it is not possible to optimize this expectation directly since we do not have access to the true joint distribution p_{XY}. Instead, we empirically optimized the objective with respect to a validation dataset $D_{\text{val}} = \{(x_i, y_i)\}_{i=1}^{n}$, determining the ensemble weights w_i using Bayesian optimization, specifically a tree-structured Parzen estimator with 50 trials implemented using Optuna [48]. We constructed an ensemble for each modality and multimodal fusion approach using the above procedure. We refer to the multi-ensemble of unimodal models and multimodal fusion models as AutoPrognosis-M in our experiments.

E. Explainability

Models must be thoroughly understood and debugged to validate the underlying logic of the model, engender model trust from both clinical users and patients [26], and satisfy regulatory requirements prior to clinical use [49]. This is particularly true in the multimodal setting, where we wish to understand what information is being used from each modality and for which patients each modality is contributing to the model output. Consequently, AutoPrognosis-M contains multiple classes of explainability techniques to enable ML models to be better understood. We have included feature-based interpretability methods, such as SHAP [50] and Integrated Gradients [51], that allow us to understand the importance of individual features, as well as an example-based interpretability method, SimplEx [52], that explains the model output for a particular sample with examples of similar instances, similar to case-based reasoning.

F. Uncertainty estimation

Quantifying the uncertainty of predictions is another critical component in engendering model trust with clinicians and patients, and can be used both to protect against likely inaccurate predictions and inform clinical decision-making [53], [54]. We adopted the conformal prediction framework, which produces statistically valid prediction intervals or sets for any underlying predictor while making minimal assumptions and capturing the total uncertainty of predictions [55]. We used inductive conformal prediction [56], which uses a calibration set to determine the width of prediction intervals or the size of the prediction sets, with local adaptivity to adjust the interval or set to the specific example [57]–[59]. For our experiments, we used regularized adaptive prediction sets [60]. Further details can be found in Appendix I-F.

III. Experiments

We demonstrate the application of AutoPrognosis-M to multimodal healthcare data with the example of skin lesion diagnosis. This process is inherently multimodal, with primary care physicians, dermatologists, or other clinicians using multiple factors to determine a diagnosis. Visual inspection has formed a crucial element, for example the "ABCD" rule or the ELM 7-point checklist [63]. These approaches have been refined to include other characteristics beyond the appearance of the lesion at a single point in time, such as evolution [64]. Beyond visual examination, clinicians also consider medical history and other factors, such as itching and bleeding [10].

A. Data and experimental setup

Experiments were conducted using PAD-UFES-20 [65]. The dataset contains 2,298 skin lesion images from 1,373

TABLE I: **Unimodal skin legion classification performance.** The best result for each modality is in bold. The best non-ensemble approach for each modality is underlined.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
Tabular								
Log. Reg. [28]	63.6%	63.3%	0.890	0.559	83.0%	0.904	0.814	0.657
Random Forest [29]	65.2%	54.0%	0.865	0.535	83.0%	0.903	0.810	0.662
XGBoost [30]	66.5%	54.4%	0.875	0.545	81.3%	0.885	0.797	0.623
CatBoost [61]	64.3%	57.2%	0.877	0.545	83.4%	0.902	0.822	0.667
MLP [31]	**69.7%**	52.8%	0.878	0.526	83.1%	0.902	0.819	0.663
TANGOS [32]	64.7%	64.0%	0.885	0.569	83.3%	0.905	0.819	0.667
TabTransformer [62]	60.6%	56.4%	0.868	0.507	82.2%	0.887	0.808	0.641
FTTransformer [33]	66.3%	63.7%	0.875	0.585	84.1%	0.903	0.828	0.681
AutoPrognosis [17]	68.7%	**64.3%**	**0.895**	**0.592**	**84.8%**	**0.912**	**0.833**	**0.692**
Imaging								
ResNet18 [37]	59.8%	57.8%	0.885	0.547	81.9%	0.897	0.808	0.637
ResNet34 [37]	57.4%	54.5%	0.873	0.517	80.3%	0.881	0.790	0.603
ResNet50 [37]	60.7%	60.0%	0.888	0.562	82.0%	0.888	0.811	0.637
ResNet101 [37]	60.3%	56.6%	0.886	0.543	81.6%	0.883	0.802	0.628
ResNet152 [37]	63.6%	59.6%	0.895	0.578	82.1%	0.892	0.810	0.640
EfficientNetB0 [38]	64.3%	60.8%	0.899	0.577	82.4%	0.900	0.807	0.645
EfficientNetB1 [38]	65.5%	63.7%	0.901	0.602	82.6%	0.899	0.811	0.648
EfficientNetB2 [38]	65.1%	59.7%	0.899	0.578	81.5%	0.888	0.801	0.628
EfficientNetB3 [38]	64.2%	62.6%	0.902	0.598	81.9%	0.898	0.805	0.635
EfficientNetB4 [38]	66.7%	62.1%	0.899	0.602	81.9%	0.897	0.807	0.635
EfficientNetB5 [38]	66.7%	62.8%	0.904	0.609	82.6%	0.903	0.810	0.649
MobileNetV2 [39]	58.4%	54.5%	0.868	0.512	79.6%	0.877	0.779	0.590
ViTBase [36]	67.1%	65.0%	0.917	0.618	82.9%	0.913	0.816	0.657
ViTLarge [36]	68.1%	65.2%	0.916	0.631	84.1%	0.916	0.831	0.682
DinoV2Small [42]	68.1%	65.0%	0.913	0.630	84.2%	0.912	0.834	0.683
DinoV2Base [42]	68.5%	65.9%	0.914	0.639	84.0%	0.913	0.833	0.679
DinoV2Large [42]	69.0%	65.8%	0.916	0.640	84.4%	0.913	0.834	0.686
Imaging ensemble	**71.6%**	**69.4%**	**0.927**	**0.672**	**85.3%**	**0.927**	**0.845**	**0.706**

patients in Brazil. Images of lesions were captured from smartphones and each image is associated with 21 tabular features, including the patient's age, the anatomical region where the lesion is located, demographic information, and other characteristics of the legion, such as whether it itched, bled, or had grown. An overview of the clinical features can be found in Table IV. Further details can be found in [65].

Skin lesions are classified as one of six different diagnoses, three of which are cancerous (Basal Cell Carcinoma, Squamous Cell Carcinoma, and Melanoma) and three are non-cancerous (Actinic Keratosis, Melanocytic Nevus, and Seborrheic Keratosis). As is common in studies of skin lesions, there are substantial differences in the number of lesions with each diagnosis (Table IV). Aggregating the diagnoses into cancerous and non-cancerous almost eliminates this class imbalance (47% cancerous, 53% non-cancerous). To demonstrate the suitability of our framework for balanced and imbalanced classification scenarios, we explored predicting the specific diagnoses and the binary determination of whether a given lesion is cancerous. We assessed lesion categorization using accuracy (Acc.), balanced accuracy (Bal. Acc.), area under the receiver operating curve (AUROC), and macro F1 score, and assessed cancer diagnosis using accuracy, AUROC, F1 score and Matthew's correlation coefficient (MCC). For each metric, a larger value corresponds to improved performance. Accuracy and balanced accuracy range from 0% to 100%, AUROC, F1 score and macro F1 score range from 0 to 1, and MCC ranges from -1 to 1. Formulae are provided in Appendix I-E.

To account for the presence of multiple images for some patients and the imbalance in incidence of different diagnoses, we conducted 5-fold cross-validation with stratified sampling, ensuring all images from the same patient were contained in the same fold. Additionally, we used 20% of the training set of each fold to optimize hyperparameters and ensemble weights. 13 clinical variables had substantial levels of missingness (c. 35%) and this missingness was often strongly associated with diagnosis. Consequently, we retained only features without this missingness. We retained the six features with entries recorded as "unknown" (occurrence: 0.1% - 17%), corresponding to the patient being asked the question but not knowing the answer. This resulted in eight tabular variables. Categorical variables were one-hot encoded, yielding 27 clinical features and one image for each sample.

B. How predictive is each modality in isolation?

Collecting additional information is never without expense, be it financial, time, or even adverse effects of collecting the information. Thus it is critical to understand whether a modality is necessary and brings additional predictive power. Therefore, we first used AutoPrognosis-M to optimize ML pipelines for each modality separately. Both the clinical variables and images exhibited some predictive power for lesion

TABLE II: **Skin legion classification performance.** All multimodal approaches outperform the unimodal baselines. AutoPrognosis-M achieves the best results.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
Tabular ensemble	68.7%	64.3%	0.895	0.592	84.8%	0.912	0.833	0.692
Best image model	68.5%	65.9%	0.914	0.639	84.4%	0.913	0.834	0.686
Imaging ensemble	71.6%	69.4%	0.927	0.672	85.3%	0.927	0.845	0.706
Best late fusion	76.8%	72.5%	0.938	0.699	89.2%	0.951	0.884	0.784
Late fusion ensemble	78.8%	74.0%	0.940	0.729	89.0%	0.956	0.881	0.779
Best early fusion	70.9%	68.1%	0.918	0.657	85.7%	0.922	0.847	0.713
Early fusion ensemble	74.7%	70.2%	0.930	0.701	86.7%	0.936	0.856	0.731
Best joint fusion	73.8%	71.4%	0.930	0.698	87.5%	0.940	0.866	0.748
Joint fusion ensemble	75.6%	71.9%	0.937	0.716	88.8%	0.951	0.880	0.775
AutoPrognosis-M	80.0%	75.0%	0.947	0.754	89.8%	0.958	0.889	0.794

categorization and cancer diagnosis (Table I), with the best individual imaging models outperforming the tabular models, particularly for lesion categorization.

Tabular. We tested several baseline classification models covering a range of model classes, specifically linear models (Log. Reg. [28]), tree-based approaches (Random Forest [29], XGBoost [30], CatBoost [61]), and deep learning (MLP [31], TANGOS [32], TabTransformer [62], FT-Transformer [33]). We additionally assessed the performance of AutoPrognosis-M on just the tabular modality, which is equivalent to AutoPrognosis [17]. Further details about the baseline models can be found in Section II-B.1, Appendix I-A, and the respective original publications. Overall, AutoPrognosis outperformed any individual tabular model demonstrating the importance of AutoML and ensembling (Table I). However, the relative outperformance over the best models, such as FT-Transformer for cancer diagnosis, was relatively minor, perhaps reflecting the nature of the structured information available.

Imaging. The different imaging architectures displayed significant variability in performance across the lesion categorization and cancer diagnosis prediction tasks; however, several trends could be observed. The vision transformer architectures (ViT and DINOv2) outperformed the CNN-based models (ResNet, EfficientNet, MobileNet) almost universally across both tasks. One explanation beyond the model architecture could be the pretraining set, which differed between the transformer and CNN models (see Table III). Increasing the size of models typically led to improvements in performance for the transformer models, although the largest models did not necessarily improve performance (e.g. DINOv2Large vs. DINOv2Base), while the trend was less clear for the CNN-based models. All model architectures consistently underperformed when trained from initialization, thus all results shown are from fine-tuning pretrained models.

Ensembling the best-performing image models resulted in a substantial increase in performance across all metrics for both prediction tasks. While the transformers outperformed the CNNs individually, many of the ensembles contained CNN-based approaches, demonstrating the importance of diversity in ensemble learning.

C. What benefit does multimodal ML provide?

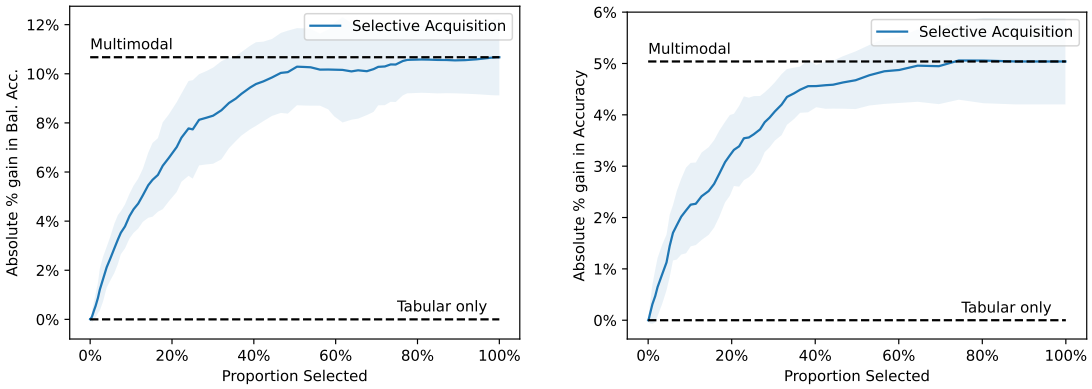
We next sought to quantify what, if any, benefit incorporating both modalities when issuing predictions provides. All three multimodal fusion strategies included in AutoPrognosis-M exhibited significant improvements over the unimodal classifiers, demonstrating the importance of integrating data from multiple sources. In Table II, we report the best single model for each fusion strategy, together with the impact of ensembling the best-performing models (determined using the held-out portion of the training set). The impact of combining the modalities varied across the various model architectures, with the results also differing for each of the fusion strategies, late (Table V), early (Table VI), and joint (Table VII).

Perhaps surprisingly, late fusion outperformed both early and joint fusion, with early fusion the worst-performing fusion approach. This is likely a consequence of the relatively strong predictive power of the tabular features and the number of samples available, but could also reveal the nature of the relationship between the two modalities. Again, ensembling the best-performing models for each fusion strategy provided a relatively small but consistent improvement, except for the cancer diagnosis task for late fusion, where the best individual model performed similarly to the ensemble.

AutoPrognosis-M leverages the power of each fusion strategy and the unimodal models by combining them in an ensemble. This approach performed best across both tasks as measured by any metric, improving the performance over any one fusion approach alone. Despite late fusion outperforming the other multimodal and unimodal approaches, it was not always selected as the most important component in the ensemble. Each fusion approach was given the largest weight in at least one ensemble, and the largest weight assigned to any of the five strategies (two unimodal, three fusion) in the multi-ensemble was 47%, further reinforcing the importance of diversity.

D. When are additional modalities most helpful?

While we have shown multimodal systems significantly outperform unimodal approaches for lesion categorization and cancer diagnosis, we might not require all modalities for all patients. As mentioned previously, there might be downsides



(a) Lesion categorization (b) Cancer diagnosis

Fig. 4: Selective acquisition of images based on conformal prediction. By acquiring images for around 20% of samples with the highest predicted uncertainty based on the tabular features, we capture over 60% of the improvement of the multimodal classifier for (a) lesion categorization and (b) cancer diagnosis, respectively. We approach the performance of the multimodal classifier by acquiring images for around half of all patients.

to collecting additional information, thus identifying when we would benefit is both important and often a key clinical decision, since modalities are typically acquired sequentially.

We demonstrate how AutoPrognosis-M can be used to answer this question. We assumed access initially to the clinical features and wanted to identify for which patients to acquire an image of the lesion. We used conformal prediction to estimate the uncertainty of each prediction and chose to acquire images for the patients with the highest uncertainty.

Acquiring images for around 20% of samples with the highest predicted uncertainty based on the tabular features captured almost two-thirds of the total improvement of the multimodal ensemble classifier for the cancer diagnosis task (Fig. 4b) and over half for the lesion categorization task (Fig. 4a). Acquiring images for around 50% of patients approached the performance of the multimodal classifier, thereby halving the number of images needed to be collected.

E. Understanding the information provided by each modality

Understanding why predictions were issued is incredibly important across medical contexts. We demonstrate how the interpretability techniques included in AutoPrognosis-M can be used to analyze the rationale for predictions across multiple modalities. We used integrated gradients [51] to analyze the predictions from the image-only and joint fusion variants of EfficientNetB4. An example is shown in Fig. 5.

The image-only model incorrectly classified the lesion as Melanocytic Nevus (NEV, non-cancerous), while the joint fusion model correctly identified the lesion as Basal Cell Carcinoma (BCC, cancerous). The image attributions (Fig. 5 center) both placed the most importance on the lesion, although there are minor differences in several areas. Importantly, the clinical variables allowed the multimodal approach to correct the image-only prediction. NEV is typically asymptomatic [66] and more common in younger individuals [67]. The patient reported the lesion had bled, hurt, and itched, which the multimodal model correctly identified made NEV less likely and increased the chance of BCC, offset by the relatively young age of the patient (32). This example clearly demonstrates the importance of incorporating both modalities and the understanding that XAI can provide.

IV. Discussion

Predictive modeling has the potential to support clinical decision-making and improve outcomes. However, incorporating multiple types of data into computational approaches is not yet widespread in medicine and healthcare.

In this paper, we demonstrated the utility of AutoPrognosis-M for developing clinical models from multimodal data using AutoML. Our framework simplifies the application of multimodal fusion strategies, automatically determining the best strategy for the available data and clinical application. We showed that AutoPrognosis-M can also be used to perform unimodal analysis for tabular and imaging data, enabling clinicians to understand when multimodal approaches will provide benefit. Beyond prediction, we used uncertainty estimation to determine for which patients additional information is necessary and XAI to improve model understanding.

While bespoke multimodal ML systems have been developed, few general-purpose frameworks exist. HAIM [68] is an early fusion approach using user-defined pretrained feature-extraction models to extract representations that are concatenated and passed to an XGBoost model. Wang et al. proposed a multimodal approach for esophageal variceal bleeding prediction [69]. They first trained an imaging model and then used automated machine learning to develop a classifier based on structured clinical data and the output of the imaging model. Finally, AutoGluon-Multimodal [70] enables fine-tuning of pretrained models across multiple modalities, combining their outputs via late fusion. In contrast, AutoPrognosis-M incorporates multiple fusion approaches, including both early and late fusion as possible strategies, while our experiments highlight the limitations of only considering a single fusion strategy.

A significant challenge with AutoML frameworks is balancing the size of the search space, which can suffer from

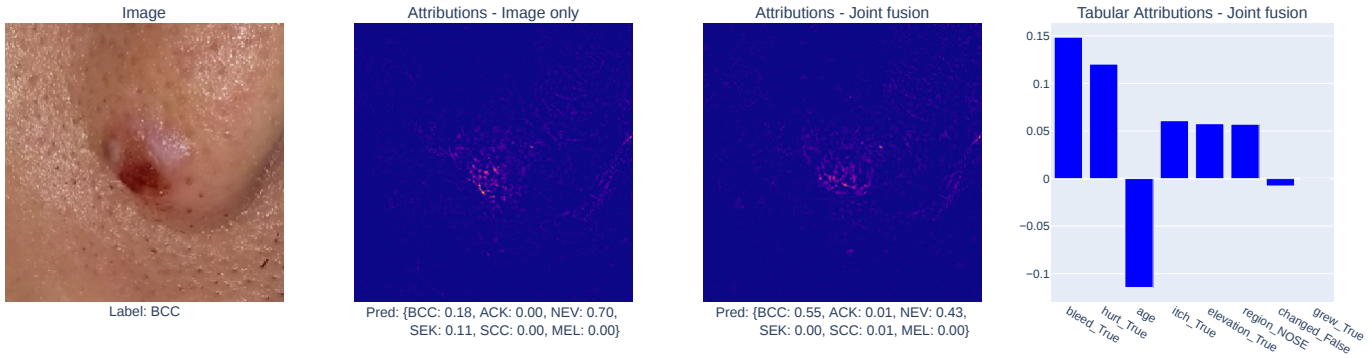


Fig. 5: Comparison of explanations for unimodal and multimodal models using integrated gradients. The original image (left, img_id: PAT_521_984_412) together with attributions for the unimodal (center left) and joint fusion EfficientNetB4 models (center right and right).

combinatorial explosion, with the computational cost required to search it. This problem is particularly apparent in the multimodal setting, where different fusion approaches represent an additional dimension over which to optimize. In our experiments, we managed this by restricting our search space to one approach for each of the three fusion strategies. However, many fusion approaches exist, in particular for joint fusion, and there is no guarantee that the approaches in our framework will always be optimal. To test this, we implemented an additional joint fusion approach, MetaBlock [71]. In general, the results were slightly weaker than our joint fusion approach, and constructing ensembles including the best-performing MetaBlock models did not improve the overall performance (Table VIII). Hence, while we believe our framework offers a balance between capturing a broad range of fusion strategies and the size of the search space, future work could investigate the impact of incorporating more fusion strategies.

Our framework exploits pretrained vision models by fine-tuning them on the available data. Recently, largely due to the size of large language models, parameter-efficient fine-tuning (PEFT) strategies, such as LoRA [72], have been proposed. While offering substantially faster training, full fine-tuning has typically still been found to outperform PEFT [73]. Future work could explore the use of PEFT for multimodal approaches, in particular in the low sample regime.

Finally, AutoML frameworks such as AutoPrognosis-M can aid in model development, but models must still be suitably validated to ensure they exhibit the desired characteristics, such as being accurate, reliable, and fair. As with any learning algorithm, significant care must be taken by the user to ensure appropriate study design and data curation, without which an inaccurate or biased model could be developed which could have adverse effects on patient health.

V. Conclusion

In this study, we introduced an AutoML framework for multimodal learning, AutoPrognosis-M. To the best of our knowledge, ours is the only framework that considers multiple classes of multimodal fusion, incorporating early, joint, and late fusion approaches. Our experiments demonstrated the importance of multimodal approaches in general, with the final AutoPrognosis-M ensemble outperforming all unimodal and multimodal approaches. Notably, combining different multimodal fusion approaches outperformed ensembles of different models employing the same fusion approach.

While we demonstrated the application of AutoPrognosis-M to skin lesion diagnosis using smartphone images and clinical variables, our framework is generally applicable and can naturally be extended to additional modalities, models, and fusion strategies. We believe AutoPrognosis-M represents a powerful tool for clinicians and ML experts when working with data from multiple modalities and we hope our framework aids the adoption of multimodal ML methods in healthcare and medicine.

References

[1] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digit. Med.*, vol. 1, no. 1, p. 39, 2018.

[2] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[3] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digit. Med.*, vol. 5, no. 1, p. 171, Nov 2022.

[4] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digit. Med.*, vol. 3, no. 1, p. 136, 2020.

[5] A. Leslie and P. R. Jones, A. J.and Goddard, "The influence of clinical information on the reporting of CT by radiologists." *Br. J. Radiol.*, vol. 73, no. 874, pp. 1052–1055, 2000.

[6] C. Castillo, T. Steffens, L. Sim, and L. Caffery, "The effect of clinical information on radiology reporting: A systematic review," *J. Med. Radiat. Sci.*, vol. 68, no. 1, pp. 60–74, 2021.

[7] W. W. Boonn and C. P. Langlotz, "Radiologist use of and perceived need for patient data access," *J. Digit. Imaging*, vol. 22, no. 4, pp. 357–362, 2009.

[8] M. Y. Wang, S. Asanad, K. Asanad, R. Karanjia, and A. A. Sadun, "Value of medical history in ophthalmology: A study of diagnostic accuracy," *J. Curr. Ophthalmol.*, vol. 30, no. 4, pp. 359–364, 2018.

[9] M. J. Ombrello, K. A. Sikora, and D. L. Kastner, "Genetics, genomics, and their relevance to pathology and therapy," *Best Pract. Res.: Clin. Rheumatol.*, vol. 28, no. 2, pp. 175–189, 2014.

[10] M. Bergenmar, J. Hansson, and Y. Brandberg, "Detection of nodular and superficial spreading melanoma with tumour thickness $\leq$ 2.0 mm - an interview study," *Eur. J. Cancer Prev.*, vol. 11, no. 1, pp. 49–55, 2002.

[11] P. Li, Y. Hu, and Z.-P. Liu, "Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods," *Biomed. Signal Process. Control.*, vol. 66, p. 102474, 2021.

[12] Y. Liu *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nat. Med.*, vol. 26, no. 6, pp. 900–908, 2020.

[13] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[14] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Improving workflow efficiency for mammography using machine learning," *J. Am. Coll. Radiol.*, vol. 17, no. 1, pp. 56–63, 2020.

[15] J. Wu *et al.*, "Radiological tumour classification across imaging modality and histology," *Nat. Mach. Intell.*, vol. 3, no. 9, pp. 787–798, 2021.

[16] T. Callender and M. van der Schaar, "Automated machine learning as a partner in predictive modelling," *Lancet Digit. Health*, vol. 5, no. 5, pp. e254–e256, 2023.

[17] F. Imrie, B. Cebere, E. F. McKinney, and M. van der Schaar, "AutoPrognosis 2.0: Democratizing diagnostic and prognostic modeling in healthcare with automated machine learning," *PLOS Digit. Health*, vol. 2, no. 6, p. e0000276, 06 2023.

[18] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PLoS One*, vol. 14, no. 5, pp. 1–17, 05 2019.

[19] A. M. Alaa and M. van der Schaar, "Prognostication and risk factors for cystic fibrosis via automated machine learning," *Sci. Rep.*, vol. 8, no. 1, p. 11242, Jul 2018.

[20] A. M. Alaa, D. Gurdasani, A. L. Harris, J. Rashbass, and M. van der Schaar, "Machine learning to guide the use of adjuvant therapies for breast cancer," *Nat. Mach. Intell.*, vol. 3, no. 8, pp. 716–726, Aug 2021.

[21] T. Callender *et al.*, "Assessing eligibility for lung cancer screening using parsimonious ensemble machine learning models: A development and validation study," *PLOS Med.*, vol. 20, no. 10, p. e1004287, 2023.

[22] F. Imrie, P. Rauba, and M. van der Schaar, "Redefining digital health interfaces with large language models," *arXiv preprint arXiv:2310.03560*, 2023.

[23] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," *Adv. Neural Inf. Process. Syst.*, vol. 28, p. 2755–2763, 2015.

[24] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," *Proc. Int. Conf. Knowl. Discovery Data Mining*, p. 847–855, 2013.

[25] R. S. Olson and J. H. Moore, "TPOT: A tree-based pipeline optimization tool for automating machine learning," *In: International Conference on Machine Learning - AutoML Workshop*, pp. 66–74, 2016.

[26] F. Imrie, R. Davis, and M. van der Schaar, "Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare," *Nat. Mach. Intell.*, vol. 5, no. 8, pp. 824–829, Aug 2023.

[27] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artif. Intell. Med.*, vol. 104, p. 101822, 2020.

[28] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1-2, pp. 167–179, 1967.

[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[30] T. Chen, "XGBoost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, 2015.

[31] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.

[32] A. Jeffares, T. Liu, J. Crabbé, F. Imrie, and M. van der Schaar, "TANGOS: Regularizing tabular neural networks through gradient orthogonalization and specialization," *In: International Conference on Learning Representations*, 2023.

[33] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Adv. Neural Inf. Proc. Sys.*, vol. 34, pp. 18 932–18 943, 2021.

[34] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An introduction to statistical learning*. Springer Nature, 2023.

[35] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," *In: Proc. 2010 IEEE Int. Symp. Circuits Syst.*, pp. 253–256, 2010.

[36] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *In: International Conference on Learning Representations*, 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[38] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *In: International Conference on Machine Learning*, pp. 6105–6114, 2019.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, 2018.

[40] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, pp. 1–40, 2016.

[41] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[42] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024.

[43] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[44] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, p. 121, 2021.

[45] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: A review," *Brief. Bioinform.*, vol. 23, no. 2, p. bbab569, 2022.

[46] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, 1998.

[47] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Adv. Neural Inf. Process. Syst.*, vol. 7, 1994.

[48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *In: Proc. Int. Conf. Knowl. Discovery Data Mining*, pp. 2623–2631, 2019.

[49] Medicines & Healthcare products Regulatory Agency, "Software and ai as a medical device change programme - roadmap," 2023, updated 14 June 2023. [Online]. Available: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap

[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[51] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *In: International Conference on Machine Learning*, pp. 3319–3328, 2017.

[52] J. Crabbe, Z. Qian, F. Imrie, and M. van der Schaar, "Explaining latent representations with a corpus of examples," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 154–12 166, 2021.

[53] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *npj Digit. Med.*, vol. 4, no. 1, p. 4, 2021.

[54] M. A. Helou, D. DiazGranados, M. S. Ryan, and J. W. Cyrus, "Uncertainty in decision making in medicine: A scoping review and thematic analysis of conceptual models," *Acad. Med.*, vol. 95, no. 1, 2020.

[55] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[56] V. Vovk, "Conditional validity of inductive conformal predictors," *In: Asian Conference on Machine Learning*, pp. 475–490, 2012.

[57] H. Papadopoulos, V. Vovk, and A. Gammerman, "Regression conformal prediction with nearest neighbours," *J. Artif. Intell. Res.*, vol. 40, pp. 815–840, 2011.

[58] U. Johansson, C. Sönströd, and H. Linusson, "Efficient conformal regressors using bagged neural nets," *In: International Joint Conference on Neural Networks*, pp. 1–8, 2015.

[59] N. Seedat, A. Jeffares, F. Imrie, and M. van der Schaar, "Improving adaptive conformal prediction using self-supervised learning," *In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[60] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," *In: International Conference on Learning Representations*, 2021.

[61] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," *Adv. Neural Inf. Proc. Sys.*, vol. 31, 2018.

[62] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.

[63] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Arch. Dermatol.*, vol. 134, no. 12, pp. 1563–1570, 12 1998.

[64] N. R. Abbasi *et al.*, "Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria," *JAMA*, vol. 292, no. 22, pp. 2771–2776, 2004.
[65] A. G. Pacheco *et al.*, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, p. 106221, 2020.
[66] P. Macneal and B. C. Patel, "Congenital melanocytic nevi," *StatPearls Publishing, Treasure Island (FL)*, 2020.
[67] I. Zalaudek *et al.*, "Frequency of dermoscopic nevus subtypes by age and body site: A cross-sectional study," *Arch. Dermatol.*, vol. 147, no. 6, pp. 663–670, 06 2011.
[68] L. R. Soenksen *et al.*, "Integrated multimodal artificial intelligence framework for healthcare applications," *npj Digit. Med.*, vol. 5, no. 1, p. 149, Sep 2022.
[69] Y. Wang *et al.*, "Automated multimodal machine learning for esophageal variceal bleeding prediction based on endoscopy and structured data," *J. Digit. Imaging*, vol. 36, no. 1, pp. 326–338, Feb 2023.
[70] Z. Tang *et al.*, "AutoGluon-Multimodal (AutoMM): Supercharging multimodal AutoML with foundation models," *In: International Conference on Automated Machine Learning*, 2024.
[71] A. G. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3554–3563, 2021.
[72] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of Large Language Models," *In: International Conference on Learning Representations*, 2022.
[73] D. Biderman *et al.*, "LoRA learns less and forgets less," *Trans. Mach. Learn. Res.*, 2024.
[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009.

APPENDIX I
TRAINING AND EXPERIMENTAL DETAILS

A. Tabular

Tabular models were trained from scratch on the training dataset. We trained logistic regression (Log. Reg.) [28], random forest [29], XGBoost [30], CatBoost [61], and MLP [31] models, using the implementations and hyperparameters from AutoPrognosis [17]. We additionally conducted baselines of three recent deep learning models: TANGOS [32], TabTransformer [62], and FT-Transfomer [33]. TANGOS and FT-Transformer are outlined in Section II-B.1. TabTransformer [62] uses a self-attention-based transformer to process the embeddings of categorical features before concatenating these embeddings with continuous features, which are then fed through an MLP. For TANGOS, we conducted a hyperparameter search using a grid search of the following values $\lambda_1 \in \{0.1, 1, 10\}$, $\lambda_2 \in \{0.01, 0.1\}$, and $p_{\text{dropout}} \in \{0.0, 0.1\}$ and used the validation set for early stopping. All other hyperparameters are as described in the original publication [32]. We trained for a maximum of 200 epochs with patience of 20 and a batch size of 50. For both TabTransformer and FT-Transformer, we again used a grid search to determine hyperparameters from the following values transformer layers $\in \{2, 3\}$, $n_{\text{heads}} \in \{4, 6, 8\}$, and $p_{\text{feed forward dropout}} \in \{0.0, 0.1\}$. Other hyperparameters matched those suggested in the original publication, which in the case of TabTransformer was determined via extensive search over five datasets. We used the validation set for early stopping, with a maximum of 200 epochs, patience of 25, and a batch size of 50.

For AutoPrognosis, we conducted pipeline selection using an internal 4-fold cross-validation on each fold of the training dataset. Since age was the only non-binary clinical variable following one-hot encoding, we considered no feature scaling or min-max scaling and searched over five classifiers, namely logistic regression [28], random forests [29], CatBoost [61], TANGOS [32] and FT-Transformer [33]. We set search parameters $num\_iter = 10$, $num\_study\_iter = 3$. For the 6-way lesion classification task, we optimized both pipelines and the tabular ensemble for macro F1 score, while for the binary cancer diagnosis task we used AUROC. The final tabular ensembles were constructed as weighted combinations of up to five pipelines.

B. Imaging

All imaging models were utilized in a pretrained manner, while DINOv2 was pretrained on the LVD-142M dataset [42] and all other models on ImageNet [74]. We fine-tuned the models in two stages. First, we froze the backbone and only trained the last fully connected layer with a learning rate of 1e-4 for 50 epochs. In the second phase, we unfroze the backbone and continued training with a learning rate of 1e-6. We trained image models using weighted cross-entropy in the 6-class setup, and binary cross-entropy in the binary setup, and used early stopping with a patience of 20 on the validation loss.

C. Fusion strategies

For both the early and late fusion experiments, we first trained the vision models as described in Appendix I-B. In late fusion, the predictions of the vision model and the tabular ensemble were averaged to obtain the final prediction. For early and joint fusion strategies, we employed a two-layer fully connected MLP with 32 units in the hidden layer as the predictor, which takes as input both the extracted vision model features and the tabular features. For early fusion, we trained the MLP with a learning rate of 1e-4, applying early stopping with a patience of 20 based on validation loss. For the joint fusion strategy, we trained the entire pipeline end-to-end, initializing the imaging model as in Appendix I-B. In this case, we initially trained the MLP classifier for 50 epochs at a learning rate of 1e-4, and then reduced the learning rate to 1e-6 when unfreezing the backbone.

D. Multimodal ensembles

We constructed the multimodal ensembles and the image-only ensemble using the following manner. For each fold, we constructed a weighted ensemble of the best five models, based on their performance of the validation set. We used Bayesian optimization for 50 trials to find the weights that maximized the performance on the validation set. We maximized balanced accuracy for the 6-way lesion classification task and AUROC for the binary cancer diagnosis task.

E. Metrics

The metrics used in our experiments are defined as follows:

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

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}$$

$$\text{AUROC} = \int_0^1 TPR(FPR^{-1}(x))\,dx$$

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN}$$

$$\text{Macro F1 score} = \frac{1}{C}\sum_{i=1}^{C}\text{F1 score}_i$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where C is the number of classes, TP and FP are the number of true and false positives, respectively, TN and FN are the number of true and false negatives, respectively, TP_i, FN_i are the number of true positives and false negatives for class i, respectively, TPR is the true positive rate $\frac{TP}{TP+FN}$, FPR is the false positive rate $\frac{FP}{FP+TN}$, and F1 score$_i$ is the per-class F1 score for class i.

F. Uncertainty quantification

To determine for which patients to acquire additional information (see Fig. 4), we used conformal prediction to estimate uncertainty. Specifically, we applied regularized adaptive prediction sets (RAPS) [60] to the predictions from the tabular ensemble model. For each fold, we used the validation set to calibrate predictions. Due to the low number of classes (six and two for the lesion classification and cancer prediction tasks, respectively), we set $k_{reg} = 0$ and used a default value for $\lambda = 0.01$. We varied the desired coverage $1 - \alpha$ and acquired more information for any individual with a prediction set including more than one label. For individuals with only one label in the prediction set, we use the original prediction of the tabular model. For those with more than one label, we use the prediction from the final multimodal ensemble constructed by AutoPrognosis-M.

TABLE III: **Imaging models included in AutoPrognosis-M.** CNN - Convolutional Neural Network. ViT - Vision Transformer.

Model	Type	# Param.	Pretraining Data	Embedding size	Ref.
ResNet18	CNN	11.7 M	ImageNet-1k	512	[37]
ResNet34	CNN	21.8 M	ImageNet-1k	512	[37]
ResNet50	CNN	25 M	ImageNet-1k	2048	[37]
ResNet101	CNN	44.5 M	ImageNet-1k	2048	[37]
ResNet152	CNN	60.2 M	ImageNet-1k	2048	[37]
EfficientNetB0	CNN	5.3 M	ImageNet-1k	320	[38]
EfficientNetB1	CNN	7.8 M	ImageNet-1k	320	[38]
EfficientNetB2	CNN	9.2 M	ImageNet-1k	352	[38]
EfficientNetB3	CNN	12 M	ImageNet-1k	384	[38]
EfficientNetB4	CNN	19 M	ImageNet-1k	448	[38]
EfficientNetB5	CNN	30 M	ImageNet-1k	512	[38]
MobileNetV2	CNN	3.4 M	ImageNet-1k	320	[39]
ViTBase	ViT-B/16	86 M	ImageNet-1k	768	[36]
ViTLarge	ViT-L/16	307 M	ImageNet-21k	1024	[36]
DinoV2Small	ViT-S/14	22 M	LVD-142M	384	[42]
DinoV2Base	ViT-B/14	86 M	LVD-142M	768	[42]
DinoV2Large	ViT-L/14	307 M	LVD-142M	1024	[42]

G. Example-based model explanation

We also used SimplEx [52] to analyse the same example as in Section III-E, using the 50 most similar latent representations from the training set as the corpus. For the image-only model, while eight of the ten most similar images were also from the nose or face, the most similar age was over 20 years older than the individual in question, and none matched their clinical variables, ultimately resulting in an incorrect prediction. In contrast, the diagnoses of the five examples selected for the joint fusion model matched the prediction, with four being very close in age and the clinical variables much more closely matching the test patient, including an exact match (Figure 6), demonstrating the importance of incorporating data from both modalities.

TABLE IV: Clinical variables in the PAD-UFES-20 dataset (n=2,298).

Diagnosis	
Basal Cell Carcinoma (BCC)	845 (36.8%)
Squamous Cell Carcinoma (SCC)	192 (8.4%)
Melanoma (MEL)	52 (2.3%)
Actinic Keratosis (ACK)	730 (31.8%)
Melanocytic Nevus (NEV)	244 (10.6%)
Seborrheic Keratosis (SEK)	235 (10.2%)
Age	
6-29	92 (4.0%)
30-49	386 (16.8%)
50-69	1,098 (47.8%)
70-94	722 (31.4%)
Region	
Face	570 (24.5%)
Forearm	392 (17.1%)
Chest	280 (12.2%)
Back	248 (10.8%)
Arm	192 (8.4%)
Nose	158 (6.9%)
Hand	126 (5.5%)
Neck	93 (4.0%)
Thigh	73 (3.2%)
Ear	73 (3.2%)
Abdomen	36 (1.6%)
Lip	23 (1.0%)
Scalp	18 (0.8%)
Foot	16 (0.7%)
Itch	
Yes	1,455 (63.3%)
No	837 (36.4%)
Unknown	6 (0.3%)
Grew	
Yes	925 (40.2%)
No	971 (42.3%)
Unknown	402 (17.5%)
Hurt	
Yes	397 (17.3%)
No	1,891 (82.3%)
Unknown	10 (0.4%)
Changed	
Yes	202 (0.9%)
No	1,700 (74.0%)
Unknown	396 (17.2%)
Bleed	
Yes	614 (26.7%)
No	1,678 (73.0%)
Unknown	6 (0.3%)
Elevation	
Yes	1,433 (62.4%)
No	863 (37.6%)
Unknown	2 (0.1%)

TABLE V: **Late fusion skin lesion classification performance.** The best result is in bold. The best non-ensemble approach is underlined.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
ResNet18 [37]	74.5%	68.7%	0.924	0.662	88.2%	0.948	0.872	0.763
ResNet34 [37]	73.0%	67.6%	0.923	0.650	87.9%	0.948	0.869	0.757
ResNet50 [37]	73.2%	69.4%	0.927	0.656	87.4%	0.944	0.862	0.746
ResNet101 [37]	73.7%	69.2%	0.927	0.666	87.4%	0.944	0.862	0.745
ResNet152 [37]	73.8%	69.4%	0.929	0.664	87.7%	0.946	0.865	0.751
EfficientNetB0 [38]	74.4%	69.9%	0.925	0.660	88.3%	0.949	0.873	0.765
EfficientNetB1 [38]	75.2%	71.3%	0.930	0.686	88.6%	0.950	0.875	0.770
EfficientNetB2 [38]	75.2%	69.6%	0.929	0.669	88.2%	0.946	0.871	0.763
EfficientNetB3 [38]	74.7%	71.6%	0.931	0.695	88.1%	0.947	0.870	0.760
EfficientNetB4 [38]	76.1%	70.7%	0.930	0.684	88.4%	0.948	0.872	0.765
EfficientNetB5 [38]	77.1%	71.5%	0.931	0.693	88.4%	0.950	0.874	0.767
MobileNetV2 [39]	71.8%	65.7%	0.915	0.626	87.0%	0.940	0.857	0.738
ViTBase [36]	76.8%	72.5%	0.938	0.699	87.5%	0.950	0.863	0.748
ViTLarge [36]	76.6%	71.7%	0.937	0.699	88.4%	0.952	0.874	0.768
DinoV2Small [42]	76.6%	71.3%	0.938	0.696	**89.2%**	0.951	**0.884**	**0.784**
DinoV2Base [42]	75.6%	71.8%	0.935	0.694	88.3%	0.952	0.874	0.765
DinoV2Large [42]	77.3%	72.4%	0.937	0.703	87.8%	0.950	0.866	0.754
Ensemble	**78.8%**	**74.0%**	**0.940**	**0.729**	89.0%	**0.956**	0.881	0.779

TABLE VI: **Early fusion skin lesion classification performance.** The best result is in bold. The best non-ensemble approach is underlined.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
ResNet18 [37]	59.9%	59.2%	0.879	0.558	81.8%	0.899	0.803	0.635
ResNet34 [37]	58.1%	56.3%	0.865	0.541	80.9%	0.884	0.795	0.614
ResNet50 [37]	62.8%	61.7%	0.891	0.590	84.1%	0.908	0.832	0.679
ResNet101 [37]	66.8%	63.4%	0.904	0.617	84.4%	0.921	0.833	0.685
ResNet152 [37]	69.1%	64.6%	0.910	0.636	83.3%	0.904	0.821	0.663
EfficientNetB0 [38]	66.0%	61.2%	0.884	0.601	83.4%	0.903	0.822	0.665
EfficientNetB1 [38]	64.3%	60.8%	0.884	0.586	82.8%	0.903	0.816	0.655
EfficientNetB2 [38]	63.8%	58.7%	0.884	0.577	80.3%	0.882	0.786	0.602
EfficientNetB3 [38]	64.1%	58.8%	0.882	0.569	81.6%	0.896	0.801	0.629
EfficientNetB4 [38]	66.4%	62.2%	0.894	0.607	82.9%	0.902	0.818	0.656
EfficientNetB5 [38]	66.3%	61.2%	0.893	0.603	82.0%	0.898	0.807	0.638
MobileNetV2 [39]	64.4%	59.4%	0.880	0.576	82.1%	0.898	0.808	0.641
ViTBase [36]	58.1%	55.2%	0.820	0.517	80.7%	0.881	0.792	0.612
ViTLarge [36]	70.1%	66.6%	0.915	0.653	84.1%	0.914	0.829	0.680
DinoV2Small [42]	70.9%	68.1%	0.918	0.657	85.7%	0.922	0.847	0.713
DinoV2Base [42]	69.1%	64.8%	0.904	0.635	83.5%	0.912	0.821	0.669
DinoV2Large [42]	70.4%	62.8%	0.906	0.630	83.8%	0.915	0.829	0.674
Ensemble	**74.7%**	**70.2%**	**0.930**	**0.701**	**86.7%**	**0.936**	**0.856**	**0.731**

TABLE VII: **Joint fusion skin lesion classification performance.** The best result is in bold. The best non-ensemble approach is underlined.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
ResNet18 [37]	65.8%	61.6%	0.899	0.588	83.4%	0.911	0.823	0.667
ResNet34 [37]	62.8%	61.6%	0.889	0.575	82.1%	0.891	0.809	0.639
ResNet50 [37]	66.1%	60.1%	0.879	0.568	84.3%	0.909	0.833	0.684
ResNet101 [37]	67.5%	61.5%	0.892	0.595	83.9%	0.911	0.827	0.675
ResNet152 [37]	68.8%	66.3%	0.913	0.642	85.3%	0.921	0.842	0.704
EfficientNetB0 [38]	62.9%	64.0%	0.899	0.605	84.7%	0.921	0.833	0.691
EfficientNetB1 [38]	66.1%	65.1%	0.905	0.624	85.4%	0.926	0.843	0.705
EfficientNetB2 [38]	68.3%	63.9%	0.912	0.620	84.0%	0.914	0.828	0.677
EfficientNetB3 [38]	68.1%	65.3%	0.913	0.634	84.8%	0.920	0.837	0.693
EfficientNetB4 [38]	69.0%	66.2%	0.915	0.644	86.3%	0.933	0.853	0.724
EfficientNetB5 [38]	71.0%	66.2%	0.922	0.648	87.3%	0.936	0.863	0.744
MobileNetV2 [39]	61.9%	60.5%	0.894	0.572	83.3%	0.908	0.821	0.665
ViTBase [36]	64.9%	63.6%	0.898	0.600	83.3%	0.909	0.822	0.665
ViTLarge [36]	72.6%	68.4%	0.930	0.676	87.5%	0.940	0.866	0.748
DinoV2Small [42]	69.6%	66.4%	0.915	0.651	86.2%	0.929	0.856	0.725
DinoV2Base [42]	73.8%	71.4%	0.930	0.698	87.0%	0.935	0.861	0.740
DinoV2Large [42]	73.8%	68.2%	0.933	0.678	86.6%	0.941	0.857	0.733
Ensemble	**75.6%**	**71.9%**	**0.937**	**0.716**	**88.0%**	**0.951**	**0.880**	**0.775**

TABLE VIII: **MetaBlock [71] joint fusion lesion classification performance.** The best result is in bold. The best non-ensemble approach is underlined.

	Lesion Categorization (6-way)				Cancer Diagnosis (Binary)			
Method	**Acc.**	**Bal. Acc.**	**AUROC**	**F1**	**Acc.**	**AUROC**	**F1**	**MCC**
ResNet18 [37]	59.0%	58.4%	0.861	0.516	81.6%	0.897	0.801	0.630
ResNet34 [37]	62.6%	61.8%	0.872	0.547	83.0%	0.903	0.813	0.657
ResNet50 [37]	61.3%	62.4%	0.887	0.539	83.6%	0.906	0.821	0.669
ResNet101 [37]	62.2%	63.2%	0.889	0.548	83.5%	0.906	0.821	0.669
ResNet152 [37]	63.3%	64.2%	0.890	0.559	83.4%	0.906	0.817	0.665
EfficientNetB0 [38]	60.4%	59.8%	0.866	0.528	82.8%	0.904	0.813	0.655
EfficientNetB1 [38]	62.4%	58.4%	0.875	0.532	83.4%	0.906	0.819	0.667
EfficientNetB2 [38]	64.7%	63.4%	0.890	0.569	84.3%	0.910	0.830	0.685
EfficientNetB3 [38]	65.7%	65.7%	0.895	0.583	84.3%	0.913	0.830	0.685
EfficientNetB4 [38]	66.2%	65.4%	0.896	0.583	84.9%	0.915	0.835	0.696
EfficientNetB5 [38]	66.7%	65.7%	0.895	0.590	84.5%	0.920	0.829	0.688
MobileNetV2 [39]	50.2%	51.6%	0.828	0.450	81.7%	0.896	0.799	0.633
ViTBase [36]	64.8%	63.7%	0.885	0.573	84.0%	0.908	0.826	0.678
ViTLarge [36]	68.7%	**68.5%**	**0.907**	0.618	85.8%	0.927	0.846	0.717
DinoV2Small [42]	55.3%	58.2%	0.849	0.495	81.7%	0.892	0.796	0.632
DinoV2Base [42]	65.2%	66.3%	0.896	0.580	83.9%	0.910	0.825	0.676
DinoV2Large [42]	**69.2%**	68.4%	0.906	**0.620**	84.9%	0.920	0.835	0.696
MetaBlock ensemble	67.5%	67.0%	0.902	0.602	**86.2%**	**0.928**	**0.851**	**0.724**

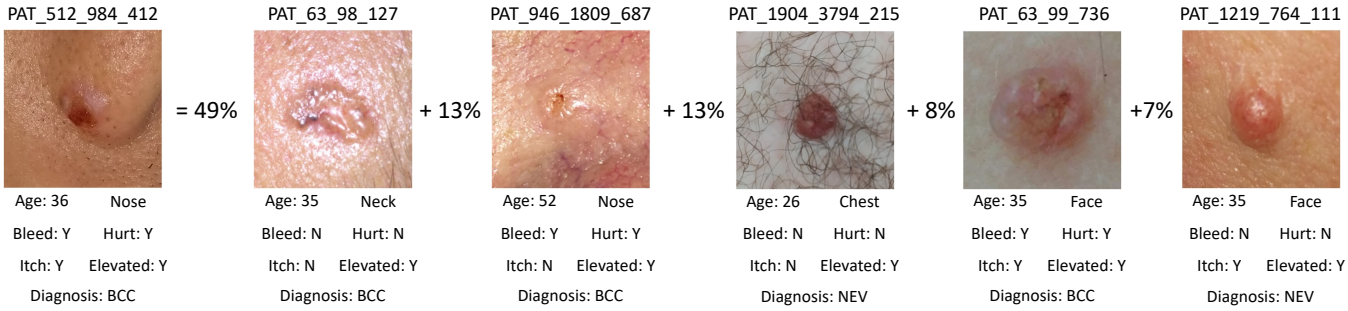


Fig. 6: **Example-based explanation with SimplEx [52].** The five examples with the highest weight are shown, together with several key variables and the diagnosis for each lesion. The diagnoses of the selected examples matches the model prediction (see Figure 5).