Learning Multimodal Explainable AI Models from Medical Images and Tabular Data Proof of Concept

Malafaia, Mafalda; Schlender, Thalea; Bosman, Peter A.N.; Alderliesten, Tanja

DOI
10.1117/12.3040402

Publication date
2025

Document Version
Final published version

Published in
Medical Imaging 2025

Learning Multimodal Explainable AI Models from Medical Images and Tabular Data: Proof of Concept

Mafalda Malafaia [a], Thalea Schlender [b], Peter A. N. Bosman [a,c], and Tanja Alderliesten [b]

[a]Evolutionary Intelligence Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[b]Dept. of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands
[c]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Medical applications often involve several data modalities, particularly medical images and clinical information, which can be combined to enhance the decision-making process by improving accuracy. Multimodal learning approaches can leverage all available data for increased robustness in the resulting models, consequently outperforming unimodal approaches. Furthermore, AI frameworks must be human-verifiable and interpretable to be deployed in real-world situations, considering legal and privacy aspects. Due to the opaque nature of Deep Learning (DL) methods, interpretability is often limited despite their state-of-the-art performance in many tasks. Genetic Programming (GP) can provide compact and interpretable symbolic expressions for tabular data but is less effective for image analysis. We introduce MultiFIX: a new interpretability-focused pipeline for multimodal learning that leverages the strengths of DL and GP to explicitly engineer features from different data types and combine them to make the final prediction. The MultiFIX pipeline comprises two stages: the training stage, where a DL (black-box) model is trained using different training procedures to extract relevant features from each modality; and the inference stage, where the resulting model is transformed to be interpretable. Image features are explained with attention maps by Grad-CAM, and inherently interpretable symbolic expressions evolved with GP fully replace the tabular feature engineering block, and the fusion of the extracted features to predict the target label. To show the application potential of the presented pipeline, we demonstrate MultiFIX with a Melanoma Risk Assessment dataset. Results show that MultiFIX outperforms unimodal models while offering explanations that can be straightforwardly analysed and are consistent with the expectations.

Keywords: explainability, genetic programming, medical image analysis, multimodal learning

1. INTRODUCTION

Within the field of Artificial Intelligence (AI), multimodal fusion has emerged as an important topic of research. Multimodal Machine Learning (ML) approaches often outperform unimodal approaches for problems with a multimodal nature,[1] offering increased robustness and the ability to leverage diverse data sources.[2] This is particularly relevant in domains such as healthcare, where combining medical images and other clinical data can improve predictions and decision-making accuracy.

While Deep Neural Networks (DNNs) achieve state-of-the-art performance across various tasks,[3] their opaque nature poses challenges in high-stakes domains such as healthcare, where interpretability and trust are essential.[4] Genetic Programming (GP) offers promise in addressing this challenge by evolving symbolic expressions, particularly for tabular data.[5,6] Particularly, GP-GOMEA[7] is a model-based evolutionary algorithm for GP known for its efficiency in evolving small and, thus, potentially interpretable symbolic expressions.[8] However, GP is less suited for image analysis, where Deep Learning (DL) models excel. Methods like Gradient-weighted Class Activation Mapping (Grad-CAM)[9] provide post-hoc explainability for DL-based image analysis, but current multimodal approaches remain limited in offering both high performance and interpretable feature analysis

Send correspondence to Mafalda Malafaia (Mafalda.Malafaia@cwi.nl) and/or Tanja Alderliesten (T.Alderliesten@lumc.nl)

Medical Imaging 2025: Image Processing, edited by Olivier Colliot, Jhimli Mitra, Proc. of SPIE Vol. 13406, 1340612 · © 2025 SPIE
1605-7422 · doi: 10.1117/12.3040402

Table 1. Input-output Table for Melanoma Risk Assessment. $Y_{original}$ concerns the original label, $age > 60$ is the constructed tabular feature, and Y_{GT} is the new ground truth label and respective imbalance ratio.

$Y_{original}$	$age > 60$	Y_{GT}
0	0	0 (56%)
0	1	1 (13%)
1	0	2 (16%)
1	1	3 (13%)

across modalities. Notable mentions demonstrate strong performance and interpretability in medical applications.[10,11] However, these works rely on post-hoc explainability restricted to final predictions, without addressing the feature extraction process for each modality or enabling integrated interpretability across modalities.

To address these gaps, we introduce MultiFIX: a Multimodal Feature engineering approach to eXplainable AI. MultiFIX is a new interpretability-focused pipeline for multimodal data that leverages the integration of DL for medical image analysis and feature extraction, and GP-GOMEA to generate interpretable symbolic expressions that replace both the tabular feature engineering block and the fusion block. Thus, the design of the pipeline provides not only the explanation for the prediction but also the individual contribution of each extracted feature. MultiFIX has unique specifications pertaining to the integration of multiple modalities and interpretability:

- **Embedded feature engineering** to extract and optimise the number of representative features for each data modality with modular feature engineering blocks, before fusion.
- **Explainability of engineered features**, including contribution heatmaps for image-engineered features and symbolic expressions for tabular-engineered features.
- **Explainability of multimodal fusion** with symbolic expressions that combine the extracted features to make the final prediction.
- **Flexibility in training procedures** to assess end-to-end, sequential, and hybrid learning to train the DL models.

We demonstrate the novelty and potential of MultiFIX with proof-of-concept experiments for medical applications, through a dataset that involves both medical images and tabular clinical information.

2. DATASET

For the present work, data was sampled from the publicly available ISIC 2020 Challenge Dataset.[12] A stratified subset of 1,000 samples was used to detect melanoma with a class imbalance ratio of 30/70%, containing dermoscopic images with skin lesions and clinical tabular data with the following features: age, anatomical general site of the lesion (torso, lower extremity, upper extremity, head/neck, or palms/soles), and sex. Cases with a malignant diagnosis label were histopathologically verified. Benign-labeled cases went through an expert agreement, longitudinal follow-up, or histopathology. Image data was cropped to be squared and resampled to a standard resolution of 200×200 pixels. Regarding the tabular features, the anatomical location of the lesion was one-hot encoded, and age and sex remained in the original format.

Since multimodal learning on the original dataset has shown a marginal improvement compared to learning using only the image model,[13] we modified the dataset by increasing the strength of the dependence of the tabular part on the final label, so that the workings of our proof-of-principle implementation can be clearly observed. A new class label is constructed so that both image and tabular information are jointly correlated to a new multiclass ground truth label between 0 and 3, Y_{GT}, which can be seen as a melanoma risk factor for the patient. We assume the ground truth image feature, I_{GT}, as the original dataset label, $Y_{original}$, and consider the following constructed tabular feature: $age > 60$, as the ground truth tabular feature, T_{GT}. We build the new melanoma risk problem as presented in Table 1. With this new target label, both the image and the variable age of the clinical data are needed to make the best prediction.

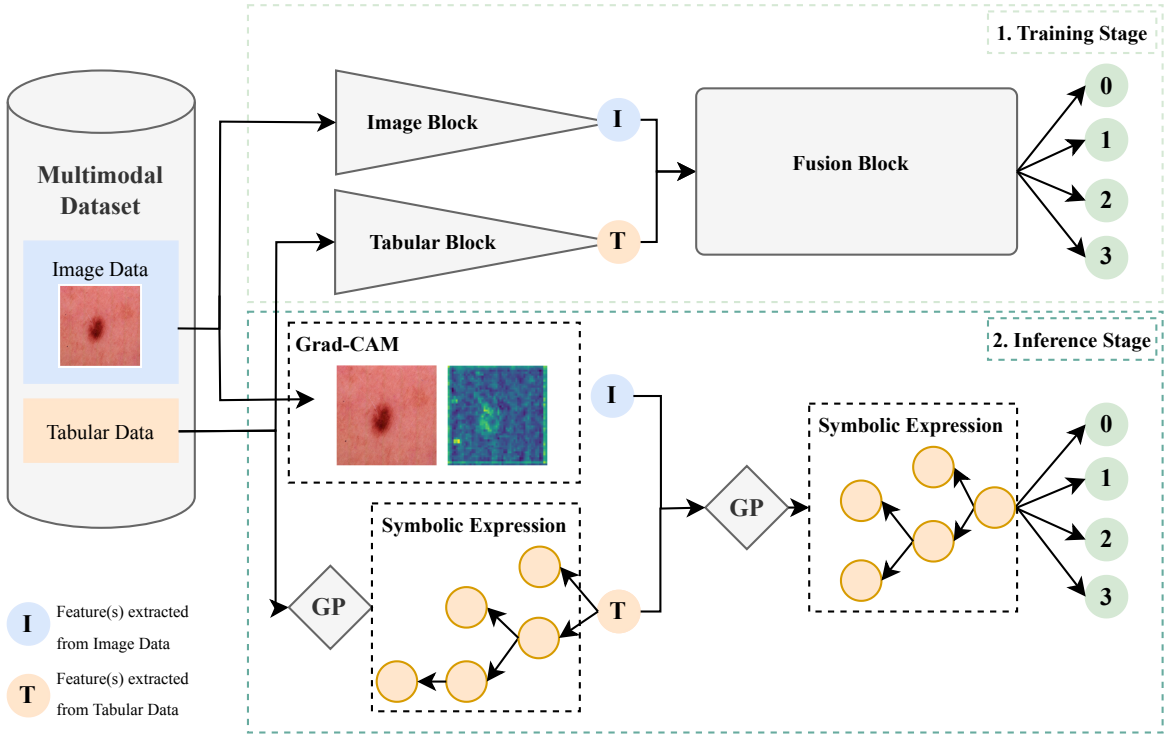

Figure 1. Overview of MultiFIX. I and T are extracted from the input data in the feature-engineering blocks and fed to the fusion block to make the final prediction in the Training Stage (top). In the Inference Stage, image features are explained through Grad-CAM, and symbolic expressions are obtained for both the tabular features and the target prediction with GP-GOMEA, replacing their DL counterparts.

3. METHODS

3.1 Pipeline and Experimental Setup

The MultiFIX pipeline consists of two stages, training and inference, resulting in an explainable model to make the final prediction. An overview of the experimental setup is illustrated in Figure 1.

3.1.1 Training Stage

The training stage comprises the training process of a DL architecture with 3 blocks: image feature engineering, tabular feature engineering, and fusion. Three training processes are studied: end-to-end learning, sequential learning, and hybrid learning. End-to-end learning enables the simultaneous learning of both the feature extraction from each modality and their influence on the prediction in the fusion block. Sequential learning uses feature engineering methods for the feature engineering blocks separately and then uses the engineered features to train the fusion block. Hybrid learning uses the pre-trained feature engineering blocks to train the pipeline end-to-end for integrated optimisation.

For the present work, the following architectures are used for each block: a pre-trained ResNet18[14] for image feature engineering; a two hidden layer MultiLayer Perceptron (MLP) for tabular feature engineering; a two hidden layer MLP for fusion. For the end-to-end learning setup, a bottleneck of representative features is imposed for each feature engineering block. Sequential learning is trained by using the single modality models to extract a latent feature representation of 64 nodes from each modality, from which a bottleneck for each branch is trained in the fusion network. Hybrid learning uses the architecture of the single modality models as encoders with the same latent representation as the sequential setup. These encoders are trained again in combination with the fusion block. Due to the modularity of the pipeline, these architectures can be changed according to the task

Table 2. Optimisation grid for optimisation of representative features and HPO. The best parameters are chosen according to the loss (lowest average ± standard deviation over the 5 folds).

No. Features	**Image features**	[1,2,3]
	Tabular features	[1,2,3]
HPO	**Learning rate**	[1e-2, 1e-3, 1e-4, 1e-5]
	Weight decay	[1e-2, 1e-3, 1e-4, 0]

at hand and the preference of the user. We use grid-search to perform Hyper-Parameter Optimisation (HPO) for the optimal learning rate and weight decay. To provide the most accurate interpretability and representation of the data, we also employ grid search to systematically explore a predefined set of values for the number of extracted features per modality by changing the number of output nodes of each feature engineering block and thereby determine the best setting for the task at hand. For details, see Table 2.

The experimental setup uses stratified 5-fold cross-validation and an 80/20 train-validation split. We use the Adam optimiser, Cross-Entropy loss, and a batch size of 32. Early stopping is used with a patience of 10 epochs, with a maximum training period of 100 epochs. The model with the best hyperparameters is used for the Inference Stage.

3.1.2 Inference Stage

The inference stage comprises the interpretability of the trained DL models. The image feature engineering block is explained in a *post-hoc* fashion with Grad-CAM,[9] resulting in visual explanations that are correlated with image contributions through the gradient information from the convolutional layers. GP-GOMEA,[7] a modern, state-of-the-art, model-based evolutionary algorithm for GP with proven efficiency in evolving small and potentially interpretable symbolic expressions, is recently adapted to include a mixture of numeric and Boolean operators, as well as the inclusion of if-then-else statements to model discontinuities.[15] This new version of GP-GOMEA is used to replace both the tabular feature engineering and fusion DL blocks with representative symbolic expressions.

Grad-CAM is applied on the activations from the first residual convolutional block of the ResNet. GP-GOMEA runs with an initial population size of 64 using an Interleaved Multistart Scheme (IMS),[7] for 256 generations. We provide numeric, $[+, -, *, /, .^2, .^3,]$, and Boolean, $[==, \neq, >, <, AND, OR]$ operators, along with the if-then-else operator. For interpretability purposes, we ensure small final expressions by limiting the maximum tree depth to 2 or 3. We run GP-GOMEA for 5 different seeds to evaluate the robustness of the obtained expressions and select the best performing one for the explainable model.

The models are compared performance-wise using Balanced Accuracy (BAcc), to account for data imbalance. To assess the benefits of multimodal learning, we use the architecture of each feature engineering block to assess the performance of each modality to predict both the ground truth label, Y_{GT}, and the respective ground truth engineered feature, I_{GT} or T_{GT}. The performance of the fusion block when given the I_{GT} and T_{GT} directly as input is also assessed.

4. EXPERIMENTS AND RESULTS

To show the potential of MultiFIX in capturing multimodal relationships, we used the Melanoma Risk Assessment problem, described in Section 2. It is important to highlight that the feature labels, I_{GT} and T_{GT}, are not given, but rather learned by the model in an unsupervised fashion. This can lead to I and T features that, although highly correlated with I_{GT} and T_{GT}, are not exactly the same.

4.1 Baseline Performance

As mentioned in Section 3, we use as baseline single modality results, i.e., models trained exclusively with the image and exclusively with the tabular data, using the same relevant building blocks as used in MultiFIX. Furthermore, to understand how well MultiFIX can capture the engineered features and the combination of the latter, we also train the supervised version of each building block: image input to predict I_{GT} using the image feature-engineering block; tabular input to predict T_{GT} using the tabular feature-engineering block; I_{GT} and T_{GT} as input to predict Y_{GT} using the fusion block. Table 3 describes the presented results, according to the

input given and target label to predict. The results show a BAcc of 0.794 ± 0.0480 when the image is given to detect melanoma (I_{GT}), and a BAcc of 1.000 ± 0.000 when the tabular data is given to predict the feature $age > 60$ (T_{GT}). Regarding unimodal performance results, both image and tabular inputs show low BAcc results (0.541 ± 0.0239 and 0.500 ± 0.000, respectively). These results indicate that unimodal approaches show low predictive potential towards the target label Y_{GT}, and that the image model is not able to fully extract the original dataset label, denoted as I_{GT}, which explains the limitation in performance for the multimodal models.

Table 3. Baseline Performance Results. I_{GT} and T_{GT} represent the intended engineered features from image and tabular inputs. Y_{GT} denotes the ground truth label, described in Table 1.

Input	*Image*		*Tabular*		$I_{GT} + T_{GT}$
Target	I_{GT}	Y_{GT}	T_{GT}	Y_{GT}	Y_{GT}
BAcc	0.794 ± 0.0480	0.541 ± 0.0239	1.000 ± 0.000	0.500 ± 0.000	1.000 ± 0.000

4.2 DL Performance

The performance results presented in Table 4 show a clear benefit from using MultiFIX in comparison to single modality approaches. The performance of MultiFIX with the studied training processes shows that, although a sequential approach cannot capture the representative features as successfully as the other approaches, the end-to-end and hybrid approaches achieve similar average performance values. The standard deviation in BAcc of the hybrid model in combination with the performance of the best model indicates that although less robust than the end-to-end approach, it can achieve higher performance values in some folds. Lastly, both in the end-to-end and the hybrid setups selected, the optimal number of representative features for the problem is found. Contrarily, the sequential approach, which achieved the lowest performance, achieved the best results using the maximum number of features for both modalities. This can be seen as an indicator that sequential models were not able to capture the problem as well as the remaining approaches.

4.2.1 Interpretability

The inference stage of the pipeline generated explainable models for the three setups. End-to-end and hybrid learning setups generated models in which the symbolic expressions for both T and Y are similar to what was to be expected. The explainable sequential model, although achieving similar performances to the DL model, exhibits complex symbolic expressions that are not easily understandable.

Figure 2 demonstrates one of the obtained explainable models trained with the Hybrid learning approach. In this explainable model, we can conclude that the learned image feature, I, correlates with the ground truth feature, I_{GT}, meaning that high values of I correspond to samples where I_{GT} is 1. The corresponding heatmaps display the most important regions in the image samples, which would be analysed by the expert in a real-world setting. Additionally, the feature T can be inherently explained and calculated using the evolved expression $age < 63$, which is highly similar and inversely correlated with the ground truth feature, T_{GT}. As mentioned previously, the model learned the opposite feature of the engineered one, since it was learned in an unsupervised fashion. However, T provides the same information to the model as T_{GT}, with a slightly different constant - 63 instead of 60. This is likely related to the patient samples provided to the model: the threshold evolved by GP is optimised according to the given training samples, which may not have included patients with ages between 60 and 63. Finally, the final prediction Y can be easily calculated using the extracted features and the following symbolic expression: $2 \times I + T - 1$. Given the ground truth expression to predict the label, $2 \times I_{GT} + T_{GT}$,

Table 4. MultiFIX Performance Results in the training stage. The second and third rows show the number of representative features chosen for each modality. The last row refers to the best model, chosen for the Inference Stage.

Training Procedure	**End-to-End**	**Sequential**	**Hybrid**
Image Features	1	3	1
Tabular Features	1	3	1
BAcc (avg ± sd)	0.763 ± 0.019	0.656 ± 0.060	0.746 ± 0.093
BAcc of best model	0.765	0.736	0.838

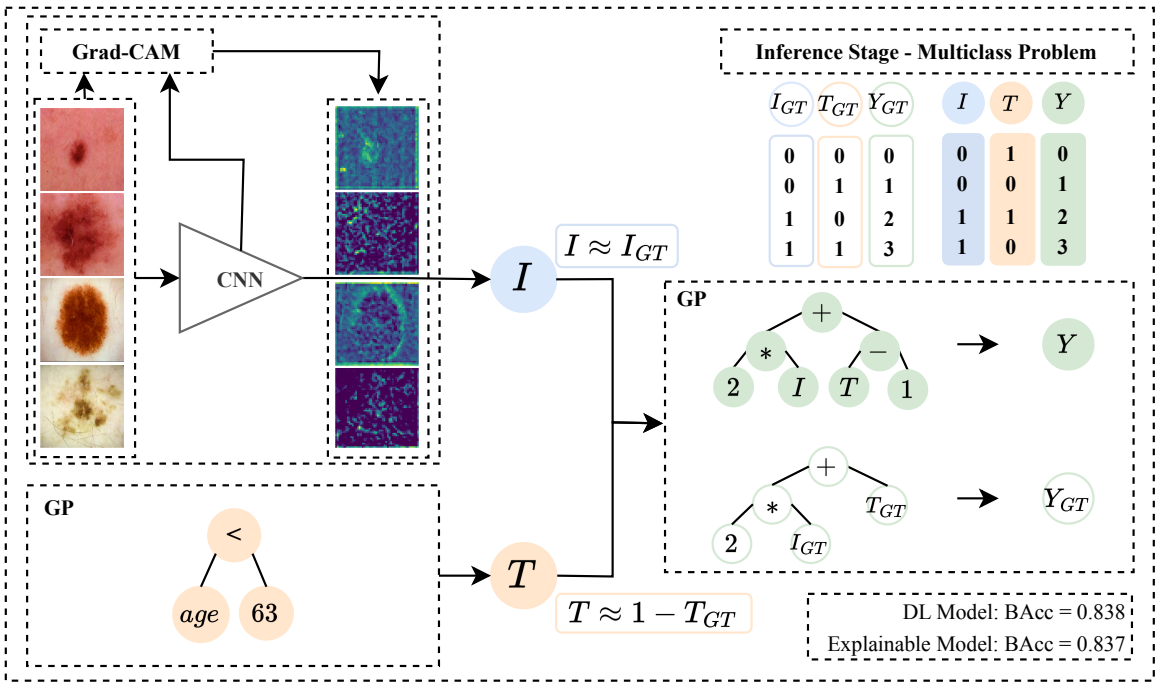

Figure 2. explainable model generated with MultiFIX for the Hybrid learning approach. Each image and respective explanation refer to a different class label, from 0 to 3. Hollow shapes refer to the Ground Truth (GT), whilst filled shapes refer to the outcome of the model. Although an expression is obtained that differs from the ground truth (it is inverted and the target age is slightly off), it can be equivalently used to predict the correct outcome.

and knowing that $T \approx 1 - T_{GT}$, we can see that the obtained symbolic expression matches the ground truth expression.

The performance of the explainable model is very similar to the performance of the surrogate DL model, with a BAcc of 0.837 for the explainable model, and a BAcc of 0.838 for the DL model. This performance confirms the expectations, knowing that the image feature engineering predictive ability was limited, thus limiting the overall multimodal performance. The explainable model shows an approximation of the values, and thus also infers that for image feature values, the binarisation may lead to faulty predictions. Likewise, the optimised constant in the feature T can also lead to faulty predictions for patients aged between 60 and 63.

5. CONCLUSION

The present work showcases the use of MultiFIX using different training procedures for a Melanoma Risk Assessment dataset. Results show a clear performance improvement compared to single-modality models, providing inherently interpretable symbolic expressions, especially for tabular feature engineering. Both end-to-end and hybrid learning approaches were successful. While end-to-end learning may be beneficial to train the full pipeline from scratch and learn features from each modality simultaneously, hybrid learning can be the most successful when application-specific feature extraction methods are available to use as pre-trained feature engineering blocks to re-train simultaneously with the fusion block. GP is integral to the explainability of MultiFIX.

With this proof-of-concept of MultiFIX, we introduce for the first time, a highly flexible interpretability-focused pipeline for integral learning from multimodal data that can be easily tailored to suit the preferences of both the user and the specific application at hand: all the architecture blocks can be replaced, and the training procedure can be adjusted accordingly to the specifications of the architecture.

ACKNOWLEDGMENTS

This research is part of the "Uitlegbare Kunstmatige Intelligentie" project funded by the Stichting Gieskes-Strijbis Fonds. We also thank NWO for the Small Compute grant on the Dutch National Supercomputer Snellius.

REFERENCES

[1] Huang, W., Tan, K., Hu, J., Zhang, Z., and Dong, S., "A review of fusion methods for omics and imaging data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(1), 74–93 (2022).

[2] Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., and Luo, Y., "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine* **5**(1), 171 (2022).

[3] Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., and Gandomi, A. H., "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in Biology and Medicine* **144**, 105253 (2022).

[4] Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B., "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923* (2017).

[5] Tran, B., Xue, B., and Zhang, M., "Using feature clustering for GP-based feature construction on high-dimensional data," in [*Genetic Programming: 20th European Conference, EuroGP 2017, Amsterdam, The Netherlands, April 19-21, 2017, Proceedings 20*], 210–226, Springer (2017).

[6] Virgolin, M., Alderliesten, T., and Bosman, P. A. N., "On explaining machine learning models by evolving crucial and compact features," *Swarm and Evolutionary Computation* **53**, 100640 (2020).

[7] Virgolin, M., Alderliesten, T., Witteveen, C., and Bosman, P. A. N., "Improving model-based genetic programming for symbolic regression of small expressions," *Evolutionary Computation* **29**(2), 211–237 (2021).

[8] La Cava, W., Orzechowski, P., Burlacu, B., de Franca, F., Virgolin, M., Jin, Y., Kommenda, M., and Moore, J., "Contemporary symbolic regression methods and their relative performance," in [*Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*], Vanschoren, J. and Yeung, S., eds., **1**, Curran (2021).

[9] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision* **128**(2), 336–359 (2019).

[10] Wang, S., Yin, Y., Wang, D., Wang, Y., and Jin, Y., "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Transactions on Cybernetics* **52**(12), 12623–12637 (2021).

[11] Azher, Z. L., Suvarna, A., Chen, J.-Q., Zhang, Z., Christensen, B. C., Salas, L. A., Vaickus, L. J., and Levy, J. J., "Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication," *BioData Mining* **16**(1), 23 (2023).

[12] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific Data* **8**(1), 34 (2021).

[13] Ha, Q., Liu, B., and Liu, F., "Identifying melanoma images using EfficientNet ensemble: Winning solution to the SIIM-ISIC melanoma classification challenge," *arXiv preprint arXiv:2010.05351* (2020).

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

[15] Schlender, T., Malafaia, M., Alderliesten, T., and Bosman, P., "Improving the efficiency of gp-gomea for higher-arity operators," in [*Proceedings of the Genetic and Evolutionary Computation Conference*], 971–979 (2024).