A Step towards Interpretable Multimodal AI Models with MultiFIX

Mafalda Malafaia
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Mafalda.Malafaia@cwi.nl

Thalea Schlender
Leiden University Medical Center
Leiden, The Netherlands
T.Schlender@lumc.nl

Tanja Alderliesten
Leiden University Medical Center
Leiden, The Netherlands
T.Alderliesten@lumc.nl

Peter A. N. Bosman
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands
Peter.Bosman@cwi.nl

ABSTRACT

Real-world problems are often dependent on multiple data modalities, making multimodal fusion essential for leveraging diverse information sources. In high-stakes domains, such as in healthcare, understanding how each modality contributes to the prediction is critical to ensure trustworthy and interpretable AI models. We present MultiFIX, an interpretability-driven multimodal data fusion pipeline that explicitly engineers distinct features from different modalities and combines them to make the final prediction. Initially, only deep learning components are used to train a model from data. The black-box (deep learning) components are subsequently either explained using post-hoc methods such as Grad-CAM for images or fully replaced by interpretable blocks, namely symbolic expressions for tabular data, resulting in an explainable model. We study the use of MultiFIX using several training strategies for feature extraction and predictive modeling. Besides highlighting strengths and weaknesses of MultiFIX, experiments on a variety of synthetic datasets with varying degrees of interaction between modalities demonstrate that MultiFIX can generate multimodal models that can be used to accurately explain both the extracted features and their integration without compromising predictive performance.

CCS CONCEPTS

• **Computing methodologies → Genetic programming**; **Learning latent representations**; **Unsupervised learning**.

KEYWORDS

Genetic Programming, Interpretability, Multimodality

ACM Reference Format:
Mafalda Malafaia, Thalea Schlender, Tanja Alderliesten, and Peter A. N. Bosman. 2025. A Step towards Interpretable Multimodal AI Models with MultiFIX. In *Genetic and Evolutionary Computation Conference (GECCO '25 Companion), July 14–18, 2025, Malaga, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3712255.3734292

GECCO '25 Companion, July 14–18, 2025, Malaga, Spain

ACM ISBN 979-8-4007-1464-1/2025/07
https://doi.org/10.1145/3712255.3734292

1 INTRODUCTION

Data availability is significantly expanding across numerous domains, not only in volume but also in diversity, translating to a heterogeneous data landscape [2]. Contrary to the oftentimes unimodal nature of Artificial Intelligence (AI) approaches, domain experts rely on multiple data modalities in their decision-making processes. For instance, in the healthcare domain, medical experts typically consider medical imaging exams, demographics, blood analysis, and further clinical information to make an informed decision. It is known that Multimodal Machine Learning (ML) can outperform single-modality approaches [15], offering increased robustness and the ability to leverage complementary information [12].

Despite the state-of-the-art performance of Deep Neural Networks (DNNs) in various unimodal and multimodal tasks [26], their opaque nature can present challenges in high-stakes domains, where interpretability and trust are paramount. Thus, to be employed in real-world situations, AI frameworks must be human-verifiable, and in some cases interpretable [17], considering not only ethical but also legal and privacy aspects.

Interpretable multimodal approaches increase transparency, can lead to knowledge discovery, and enable verifiability. Moreover, these promote vital and constant interaction between AI and domain experts, interpreting how models work, and providing input on if and how models should be adjusted [10]. Especially for predictive models in several high-stakes fields, this is of key importance.

In this work, we demonstrate MultiFIX: a Multimodal Feature engIneering approach to eXplainable AI. MultiFIX is a framework that is aimed at interpretability through the discovery of key features for different modalities, the combination of which is used to make predictions. By providing explanations of the features and the final predictions, MultiFIX provides a unique, novel approach to explainable multimodal AI. To develop MultiFIX models, the powerful learning potential of Deep Learning (DL) and the interpretability of symbolic expressions generated with Genetic Programming (GP) are leveraged. The latter is used to create models that can readily be analyzed and interpreted as a surrogate for the deep learning model. Additionally, other modality-specific post-hoc explanation techniques can be used to study model components in MultiFIX - for instance, the use of Grad-CAM for image processing neural networks.

Additionally to the MultiFIX pipeline, a key contribution of this paper is the examination of different training strategies. Preliminary work on MultiFIX [14] illustrates initial favorable outcomes in multimodal integration, primarily using end-to-end training. While end-to-end training has the potential to effectively leverage joint optimization of all components in MultiFIX, the complexity of the joint learning task may present certain limitations. We therefore also consider sequential and hybrid training, in combination with different pre-training procedures, providing flexibility in the most suitable training strategy according to the nature of the problem and the preferred architectural blocks in practice. This paper provides the first comprehensive analysis of the MultiFIX pipeline and various training methods, making it the most complete and detailed study on the subject. To demonstrate the versatility and potential of MultiFIX, in this paper we perform an exploratory study on synthetic problems with various degrees of dependence between modalities that are representative of real-world scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the MultiFIX methodology and experimental design; Section 4 presents the staged problems and outcomes, namely the resulting interpretable models; Section 5 reflects a discussion on the exploratory study; and Section 6 concludes the present work with a discussion of future directions.

2 RELATED WORK

We specifically focus on techniques that combine image and tabular data - two modalities that are commonly incorporated in multimodal AI pipelines [21].

Multimodal learning methods are commonly categorized by fusion strategies: early fusion, which concatenates input data before feature extraction; intermediate fusion, which combines modality-specific features after extraction; and late fusion, which aggregates predictions from unimodal models [9, 22]. Late fusion is the dominant strategy due to its simplicity and effectiveness in handling heterogeneous data [21]. However, recent advancements in intermediate fusion methods provide enhanced ways of capturing interactions between modalities, especially with the development of DL techniques for feature extraction [7].

Despite these advancements, most multimodal systems still lack interpretability, a crucial aspect for building trust in AI. Recent literature highlights the importance of interpretability in multimodal models, particularly in high-stakes sectors where transparency and clarity on how the models work are critical to using them [19]. Current state-of-the-art literature on explainable multimodal approaches relies substantially on modality-specific post-hoc methods.

Post-hoc explainability methods for image data have been primarily used in unimodal approaches, where the exclusive contribution of the image is mapped to the prediction. Grad-CAM is arguably the most-used method among gradient-based methods that rely on the backpropagation process of Convolutional Neural Networks (CNNs) to generate attention maps in the images [1].

The most common explainability method for tabular data analysis is the post-hoc use of SHAP values to analyze feature importance [1]. Another approach is GP, which can be used to evolve higher-level features in the form of symbolic expressions [24, 27], or to directly evolve prediction models that are fully white-box [3]. In comparison to post-hoc methods like SHAP, GP directly generates interpretable models without additional approximations, reducing the risk of misleading explanations. In Evans et al. [5], GP is used to generate compact symbolic expressions that are used as a post-hoc method to approximate ML estimators without compromising the predictive performance significantly.

Utilizing single modality, post-hoc, explainability methods, Chen et al. [4] propose an explainable multimodal pipeline using an intermediate fusion approach to provide prognostic predictions across 14 types of cancer. Specifically, they use attention mechanisms for histology images and SHAP values for genomic data to find correlations between input features and the target prediction.

Post hoc explanations are frequently chosen to explain black-box models, but may provide misleading and unreliable explanations [16]. Hence, Swamy et al. [23] describe an inherently interpretable multimodal approach that indicates the cumulative contributions of each modality to the prediction. Additionally, the modularity of the pipeline allows multi-task predictions while handling potentially missing modalities, due to its sequential training strategy. However, the proposed pipeline does not include interpretability on a feature extraction level, i.e., the feature contributions within each modality remain a black box, which arguably limits explainability, including knowledge discovery.

In contrast to existing literature, MultiFIX introduces innovative feature engineering interpretability with explicit contributions of each modality to the final prediction. Specifically, we address two major challenges: to incorporate *inherently* interpretable fusion techniques within an intermediate fusion pipeline by using GP to generate symbolic expressions; and to use sparse embedded feature engineering to extract a narrow bottleneck of modality-specific features that capture patterns that are potentially relevant to the prediction.

3 METHODOLOGY

We describe the MultiFIX pipeline, including methods used and the experimental setup. Additionally, we provide a general description of the input data used.

3.1 MultiFIX: Multimodal Feature engIneering for eXplainable AI

In MultiFIX, DL architectures are trained to generate black-box models that are designed to learn representative, potentially complex features from each modality. Whenever possible, these models are replaced by GP-generated symbolic expressions that are interpretable by design, and otherwise are explained using Grad-CAM to generate visual explanations. While we use Grad-CAM in this work, other image explainability methods could be used.

We demonstrate our pipeline using two data modalities: image and tabular data. However, the modularity of the pipeline supports straightforward integration of other data modalities by adding modality-specific feature engineering blocks. An overview of MultiFIX is provided in Figure 1.

3.2 Feature Engineering Blocks

The structure of each feature engineering block is very similar, regardless of the modality to be analyzed, comprising a DL architecture to process the input data and extract meaningful features.

We introduce the concept of sparse feature engineering, where the feature engineering bottleneck is deliberately restricted to a small number of features (we use at most three in this paper), in order to increase the chance of obtaining (easily) interpretable models.

For image data, we use a Convolutional Neural Network (CNN). In this paper, we use a pre-trained Resnet [8], but the architecture may be adjusted to suit specific needs or be optimized using neural architecture search. Besides this, an autoencoder (AE) can be trained (with the same architecture) and used as a pre-trained image processing block in some training strategies.

For tabular data, we use a Multi-Layer Perceptron (MLP) with three hidden layers, all of which are 128 nodes wide. Dropout and batch normalization are used to prevent overfitting and promote fast convergence.

3.3 Fusion Strategy

MultiFIX employs an intermediate fusion strategy to combine the engineered features from each modality to make the final prediction. However, the unique enforcement of a bottleneck with up to three features per modality makes the fusion analysis simpler and, thus, interpretability-focused.

The architecture used for the fusion block in this paper consists of the same MLP architecture used for tabular processing. Here also, alternative architectures could be used, or architecture search could be applied.

3.4 Training Strategies

We study different training strategies with MultiFIX, to explore the potential benefits of different strategies specific to different types of problems and synergies between modalities. Additionally, single-modality approaches were used as a baseline comparison, as well as pre-trained blocks for the respective modalities.

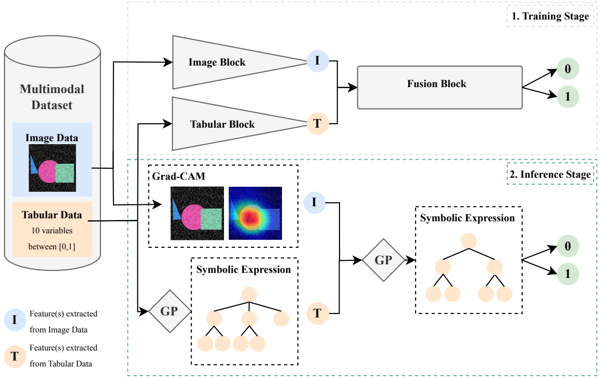


Figure 1: Overview of MultiFIX. Data passes into the feature engineering blocks. Feature vectors I and T are concatenated and passed to the fusion block to make the final prediction in the Training Stage (top). In the Inference Stage, image features are explained through Grad-CAM, and symbolic expressions are obtained for both the tabular features and the target prediction with GP-GOMEA, replacing their NN counterparts.

Population size	initially 64 (using IMS [25])
Number of generations	512
Operators	numeric $[+, -, *, /, .^2, .^3,]$ Boolean $[==, \neq, >, <, AND, OR]$ if-then-else
Maximum tree depth	[2, 3]

Table 1: GP-GOMEA settings.

Six different training strategies were considered, each varying either in pre-training weights for feature engineering blocks, sequential or parallel training, or partial temporary freezing of the architectural blocks:

1. *End-to-end training (End):* train the entire architecture simultaneously.
2. *Sequential training with AE weights (Seq AE):* use encoder weights from the trained AE in the image feature engineering block; freeze the latter while training the tabular feature engineering and the fusion blocks.
3. *Sequential training with AE weights and De-freezing (Seq AE Temp Freeze):* use encoder weights from the trained AE in the image feature engineering block; freeze image block for 15 epochs while training remaining blocks and then train the whole architecture simultaneously.
4. *Sequential training with single modality weights (Seq Single):* use weights from each single modality model for respective feature engineering blocks; train fusion block sequentially, while freezing the remaining blocks.
5. *Hybrid training with AE weights (Hyb AE):* use encoder weights from the trained AE in the image feature engineering block; train the whole architecture simultaneously using the pre-trained block.
6. *Hybrid training with single modality weights (Hyb Single):* use weights for each single modality model for respective feature engineering blocks; train whole architecture simultaneously using the pre-trained blocks.

3.5 Interpretability Techniques

In the inference stage, the DL model blocks are either explained using post-hoc explainable methods or replaced by a symbolic expression obtained using inherently interpretable methods. For images, Grad-CAM [20] leverages the gradient information from convolutional layers to generate visualizations that portray which parts of an image contribute (more) to a specific prediction. GP-GOMEA [25] is a model-based evolutionary algorithm for GP known for its effectiveness in evolving small and potentially interpretable symbolic expressions [13]. We use a recent adaptation of GP-GOMEA that allows the use of both numeric and Boolean operators, as well as if-then-else statements [18].

The explainability method used for feature engineering blocks depends on the nature of the data: Grad-CAM is used for images; GP-GOMEA is used for tabular data. GP-GOMEA is also used to replace the fusion block. Grad-CAM is applied to the activations from the last residual convolutional block of the ResNet, as suggested in guidelines [6]. For GP-GOMEA, we followed the settings described in Table 1 for all experiments.

3.6 Experimental Setup

The experiments were designed to evaluate the performance of the MultiFIX pipeline and the interpretability of the resulting models for different multimodal datasets and training strategies. Single-modality performance is used as a baseline to study whether improvements are obtained when data modalities are combined.

A standard experimental setup is used for all experiments, following the configurations described in Table 2. We use 5-fold cross-validation to assess the generalization capability of the models and a stratified 80/20 data split between train and validation sets. The Adam optimizer [11] is used with the Cross-Entropy (CE) or Binary Cross-Entropy (BCE) loss, depending on the nature of the target label. Early stopping is used with a patience of 5 epochs, with a maximum training period of 75 epochs. The batch size is 32. Hyper-Parameter Optimization (HPO) is performed using a grid-search strategy to choose the optimal Learning Rate, Weight Decay, and number of extracted features (up to three) for the image and tabular inputs (Image Bottleneck and Tabular Bottleneck, respectively). The optimal configuration can be unimodal if a bottleneck of zero features for either one of the modalities is chosen. This configuration will, however, differ from the unimodal baseline, since it also includes the fusion block, in this case combining features from a single modality.

For each of the studied problems, all training strategies mentioned in Subsection 3.4 are used, in addition to single modality approaches, resulting in eight trained models, all following the same protocol. The autoencoder, needed for three out of the six training strategies, is trained following the configuration of Table 2. The trained models are evaluated using Balanced Accuracy (BAcc) for performance purposes, with loss and AUC-ROC metrics also being evaluated and available in the Supplementary Material. The final interpretable models are evaluated considering two aspects: predictive performance and interpretability. For the former, we compare BAcc values of the DL model with the interpretable model. Additionally, we compare the performance of the studied training strategies using statistical tests on cross-validation results. For each pair of strategies, we performed a paired t-test to determine if the performance differences were statistically significant. Following common statistical practice, we set the significance level at $\alpha = 0.05$ and applied Bonferroni correction to control for inflated Type I error due to multiple pairwise comparisons. Interpretability is evaluated through a manual analysis.

AE settings		**Optimiser**	Adam
	AE Settings	**Loss Function**	MSE
		Learning Rate	0.0001
		No. of Epochs	100
MultiFIX Settings		**Optimiser**	Adam
	MultiFIX Settings	**Loss Function**	CE or BCE
		No. of Epochs	75
		Early Stopping Patience	5 epochs
Grid-Search		**Learning Rate**	[1e-3, 1e-4, 1e-5]
	HPO Grid	**Weight Decay**	[1e-3, 1e-4, 0]
		Image Bottleneck	[0, 1, 2, 3]
		Tabular Bottleneck	[0, 1, 2, 3]

Table 2: Settings used in the MultiFIX pipeline. The best parameters are chosen according to the loss (lowest average ± standard deviation over the 5 folds).

3.7 Dataset

We created a synthetic dataset with images and tabular data. For the imaging modality, 1,000 samples were automatically generated, each with a size of 200 × 200 pixels. Each image can contain the following shapes: a circle, a rectangle and/or a triangle. Each shape can either be present or absent in the image, with the possibility of having none or up to three shapes. None of the shapes can appear twice in the same sample, and all are generated randomly with a 50% chance, in different sizes and colors. Random noise was introduced to all images by mutating the color of 10,000 random pixels. An illustration of possible samples is presented in Figure 2. Tabular data consists of 1,000 samples with ten numerical features, uniformly sampled between 0 and 1, that are then used in each problem to synthetically engineer tabular features that combine two or more of the numerical input features.

4 EXPERIMENTS AND RESULTS

In this section, we present the results of training the proposed MultiFIX pipeline using various training strategies for five synthetic problems that feature different dependencies between modalities.

Each subsection pertains to one of the proposed problems, including: the problem description; DL performance results for single modality approaches and six multimodal approaches with different training strategies; and the explainable model for the best performing strategy.

4.1 AND Problem

4.1.1 Problem Description.

4.1.1 Problem Description. This problem comprises the binary operation AND between the presence of a circle ($circle$) in the image data, and the tabular feature $x_1 > x_2$. The target label is thus the output of $AND(circle, x_1 > x_2)$. This problem shows a moderate level of dependence between the two modalities to predict the target, since the target label 1 is only possible if both engineered features are present, but the target label 0 correlates with any feature being 0.

4.1.2 DL Performance Results. Figure 3 presents the performance results for the AND Problem using a single modality and MultiFIX with different training strategies. Statistical testing highlights that all multimodal approaches significantly outperform single modality approaches. The multimodal approaches perform generally similarly from a statistical significance perspective, excluding the comparison between end-to-end and sequential AE approaches, in which the former is significantly better than the latter.

4.1.3 Interpretable Model. Figure 3(a) illustrates the resulting interpretable model for the AND problem using the end-to-end training strategy, which obtained the highest average BAcc results. With the interpretable model, one can analyze the prediction in a block-by-block fashion: the image visual explanations along with the feature values indicate a high correlation between I_2 and the presence of a circle, denoted by I_{GT}; the tabular feature T_1 is obtained with a piecewise symbolic expression that inversely correlates with the required feature $x_1 > x_2$; the prediction Y_{pred} is obtained with a symbolic expression with a binary output that is 1 generally for very high values of I_2 and low values of T_1, excluding feature I_1,

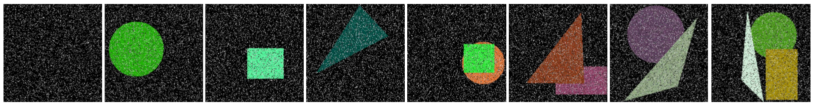


Figure 2: Representative samples for the image modality.

	Image Only	Tabular Only	End-to-End	Hybrid AE	Hybrid Single	Sequential AE	Sequential AE Defreeze	Sequential Single
Image Only	0.607 ± 0.041	not significant	worse	worse	worse	worse	worse	worse
Tabular Only	not significant	0.693 ± 0.034	worse	worse	worse	worse	worse	worse
End-to-End	better	better	0.939 ± 0.029	not significant	not significant	better	not significant	not significant
Hybrid AE	better	better	not significant	0.925 ± 0.022	not significant	not significant	not significant	not significant
Hybrid Single	better	better	not significant	not significant	0.923 ± 0.014	not significant	not significant	not significant
Sequential AE	better	better	worse	not significant	not significant	0.881 ± 0.032	not significant	not significant
Sequential AE Defreeze	better	better	not significant	not significant	not significant	not significant	0.914 ± 0.038	not significant
Sequential Single	better	better	not significant	not significant	not significant	not significant	not significant	0.901 ± 0.029

Table 3: AND Problem - BAcc Results with Statistical Testing: average BAcc and standard deviation over 5 folds are highlighted in blue; each row indicates statistical significance comparison with the remaining approaches: better, worse or not significant.

which is not relevant for the model. The truth table presented binarizes the feature values according to each explained extracted feature and the corresponding output. Although the model does not include the intended modality feature values, the final prediction is correct. The model found, while not appearing as anticipated, is an *equivalent* model, since the induced tabular feature is just the complement of the true hidden tabular feature. Only by explaining all components in MultiFIX can we actually see this (and realize that such equivalent reasonings exist). The predictive power of the interpretable model is very similar to the black-box (DL) model, with a difference of 0.006 in BAcc.

4.2 XOR Problem

4.2.1 Problem Description. This problem comprises the binary operation XOR between the presence of a circle ($circle$) in the image data, and the tabular feature $x_1 > x_2$. The target label is thus the output of $XOR(circle, x_1 > x_2)$. The XOR problem reflects an extreme dependence between the two modalities since neither the image feature nor the tabular feature alone gives any information about the target label.

4.2.2 DL Performance Results. Figure 4 presents the performance results for the XOR Problem using single modality learning and MultiFIX with different training strategies. Statistical testing highlights that all multimodal approaches significantly outperform single modality approaches. Although, similarly to the AND problem, the performance values between multimodal approaches are mostly similar from a statistical perspective, their spread is larger, with the hybrid single training strategy leading to the highest average BAcc values.

4.2.3 Interpretable Model. Figure 3(b) illustrates the interpretable model for the XOR problem using the hybrid single training strategy. The image extracted features and respective explanations indicate a high correlation between I_2 and the presence of a circle (I_{GT}); the tabular feature T_1 is obtained with a piecewise symbolic expression inversely correlated with $x_1 > x_2$ (T_{GT}); Y_{pred} is obtained with a symbolic expression that outputs 1 if the binarization of each useful feature using different thresholds is equal, and 0 otherwise. The truth table presented binarizes the feature values according to each evolved threshold (0.3 for the tabular feature and 0.7 for the image feature) and the corresponding prediction. Again, the complete model is correct, but the intermediate features are inverted, which leads to an equivalent model, as we can now see with MultiFIX. The predictive power of the interpretable model is higher than the black-box (DL) model, with an increase of 0.035 in BAcc.

4.3 Multifeature Problem

4.3.1 Problem Description. In this problem, complexity is increased by scaling the number of required features per modality. It entails the following relationship:

$$OR(AND(circle, x_1 > x_2), AND(!triangle, x_3 > x_4)) \qquad (1)$$

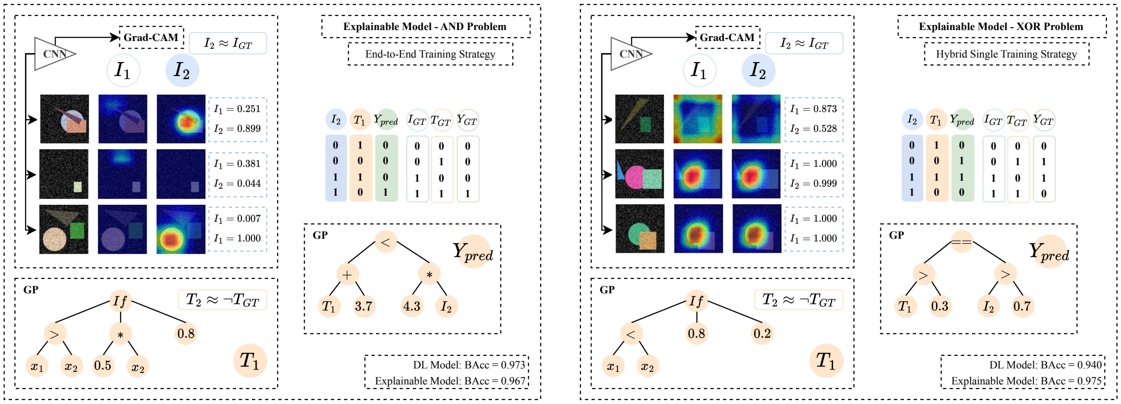
Predictions of whether or not a circle and whether or not a triangle are in the image are now needed as separate features. For the tabular data, two features are also required: $x_1 > x_2$ and $x_3 > x_4$. The Multifeature Problem reflects a pairwise dependence between features from each modality. While each AND operation is partially dependent on both modalities, the real added difficulty of this problem comes from the need to learn each intermediate feature individually.

4.3.2 DL Performance Results. Figure 5 presents the performance results for the Multifeature Problem using single modality learning and MultiFIX with different training strategies. Average BAcc values and respective standard deviations show worse performance with single modality approaches and better performance for end-to-end and hybrid training approaches. Statistical testing, however, does not show significant differences between most of the approaches, excluding the comparison between the image approach and the sequential training approach using single modality weights. This can be justified by the extremely high variance of all multimodal approaches across different folds, and the low sample size that can reduce the sensitivity of the statistical test, despite the clear separation in BAcc ranges for different approaches.

4.3.3 Interpretable Model. Figure 4(a) illustrates the resulting interpretable model for the Multifeature problem using the hybrid training strategy with encoder weights for the pre-trained image feature engineering block, which obtained the highest average performance over the remaining approaches. The inherent added

	Image Only	Tabular Only	End-to-End	Hybrid AE	Hybrid Single	Sequential AE	Sequential AE Defreeze	Sequential Single
Image Only	0.502 ± 0.027	not significant	worse	worse	worse	worse	worse	worse
Tabular Only	not significant	0.552 ± 0.020	worse	worse	worse	worse	worse	worse
End-to-End	better	better	0.899 ± 0.023	not significant	not significant	not significant	not significant	not significant
Hybrid AE	better	better	not significant	0.902 ± 0.020	not significant	not significant	not significant	not significant
Hybrid Single	better	better	not significant	not significant	0.918 ± 0.017	not significant	not significant	not significant
Sequential AE	better	better	not significant	not significant	not significant	0.812 ± 0.031	not significant	not significant
Sequential AE Defreeze	better	better	not significant	not significant	not significant	not significant	0.894 ± 0.046	not significant
Sequential Single	better	better	not significant	not significant	not significant	not significant	not significant	0.766 ± 0.041

Table 4: XOR Problem - BAcc Results with Statistical Testing: average BAcc and standard deviation over 5 folds are highlighted in blue; each row indicates statistical significance comparison with the remaining approaches: better, worse or not significant.



I_2	T_1	Y_{pred}	I_{GT}	T_{GT}	Y_{GT}
0	1	0	0	0	0
0	0	0	0	1	0
1	1	0	1	0	0
1	0	1	1	1	1

(a) AND Problem End-to-End Training

I_2	T_1	Y_{pred}	I_{GT}	T_{GT}	Y_{GT}
0	1	0	0	0	0
0	0	1	0	1	1
1	1	1	1	0	1
1	0	0	1	1	0

(b) XOR Problem Hybrid Single Training

Figure 3: Interpretable Models: Grad-CAM heatmaps explain the image input contributions for each extracted feature. GP-GOMEA symbolic expressions explain the tabular features and the fusion of both modalities to make the prediction. Learned features and predictions are compared to their Ground Truth (GT) counterparts.

	Image Only	Tabular Only	End-to-End	Hybrid AE	Hybrid Single	Sequential AE	Sequential AE Defreeze	Sequential Single
Image Only	0.655 ± 0.006	not significant	not significant	not significant	not significant	not significant	not significant	worse
Tabular Only	not significant	0.674 ± 0.028	not significant	not significant	not significant	not significant	not significant	not significant
End-to-End	not significant	not significant	0.798 ± 0.044	not significant	not significant	not significant	not significant	not significant
Hybrid AE	not significant	not significant	not significant	0.799 ± 0.054	not significant	not significant	not significant	not significant
Hybrid Single	not significant	not significant	not significant	not significant	0.798 ± 0.058	not significant	not significant	not significant
Sequential AE	not significant	not significant	not significant	not significant	not significant	0.703 ± 0.046	not significant	not significant
Sequential AE Defreeze	not significant	not significant	not significant	not significant	not significant	not significant	0.766 ± 0.077	not significant
Sequential Single	better	not significant	not significant	not significant	not significant	not significant	not significant	0.743 ± 0.018

Table 5: Multifeature Problem - BAcc Results with Statistical Testing: average BAcc and standard deviation over 5 folds are highlighted in blue; each row indicates statistical significance comparison with the remaining approaches: better, worse or not significant.

complexity of this problem is reflected in the resulting interpretable model. In the image feature engineering block, complex features were learned. Naively, one would hope that each extracted feature would exclusively relate to one of the shapes of interest (circle and triangle). However, the resulting intermediate features, although comprising relevant information, are not as simple to analyze. Plotting samples with different image characteristics in a 2D space with I_1 and I_2 on the axes revealed that lower values of I_1 are correlated with the presence of a triangle *and* the absence of a circle in the image; very high values of I_2 (approximately higher than 0.8) are correlated with the absence of *both* shapes in the image; samples in which $I_1 > I_2$ are correlated with the presence of a circle in the image; when a circle is present, the presence or absence of a triangle is not well distinguished (which indicates samples for which the model fails). The reason for this happening is that the intermediate image features are set to be real-valued, making it possible to map multiple features that are essentially binary to subranges of one real-valued feature. This analysis is corroborated with the analyzed plot in the Supplementary Material. The fusion symbolic expression with the highest performance is a tree with depth three, while for the problems so far, a tree with depth two sufficed. This increases complexity and arguably decreases interpretability. Despite being still transparent and readable, the obtained expression is not easy to interpret. The condition is related to whether there is a circle in the image and the then-branch is related to the relation between x_1 and x_2. The else-branch is related to the relation between x_3 and x_4, as well as whether there is a triangle in the image, which signals correct logic in terms of features involved. The tabular feature engineering block reveals symbolic expressions that accurately correlate to the two required features: T_1 is inversely correlated with $x_1 > x_2$, presenting values larger than 0.4 when $x_1 < x_2$, and values smaller or equal to 0.4 otherwise; T_2 and T_3 are correlated and inversely correlated (respectively) with $x_3 > x_4$, using different evolved binarization thresholds. Lastly, the predictive power of the interpretable model is higher than the black-box (DL) model, with an increase of 0.020 in BAcc.

4.4 Multiclass Problem

4.4.1 Problem Description. The last problem has a multiclass target with four possible classes rather than a binary target. The classes relate to combinations of the intermediate features of each modality (presence of circle in an image, and $x_1 > x_2$), as demonstrated in Table 6. Despite each modality being correlated with the endpoint, optimal predictions need joint information from both inputs.

ft_{circle}	$ft_{x_1>x_2}$	$Y_{multiclass}$
0	0	0
0	1	1
1	0	2
1	1	3

Table 6: Multiclass Problem using ft_{circle} and $ft_{x_1>x_2}$.

4.4.2 DL Performance Results. In figure 7 the performance results for the Multiclass Problem using single modality learning and MultiFIX with different training strategies are shown. Statistical testing highlights that all multimodal approaches significantly outperform single modality approaches, excluding the sequential training approach using AE weights in the image block. Similarly to the AND and XOR problems, the different multimodal approaches are close in performance from a statistical significance perspective, although the hybrid single training strategy achieves higher average BAcc values.

4.4.3 Interpretable Model. Figure 4(b) illustrates the resulting interpretable model for the Multiclass problem using the hybrid single training strategy. The image visual explanations along with the feature values indicate a high correlation between I_1 and the absence of a circle, which can be inferred from a very low feature value and a low-contribution heatmap when a circle is present; the tabular feature T_1 is obtained with a piecewise symbolic expression that directly correlates with the engineered feature $x_1 > x_2$; the prediction Y_{pred} is calculated with a piecewise symbolic expression with the condition $I_1 < 0.5$ (is a circle present in the image), and with two possible subtrees that use T_1 to assign the values 2 or 3, if a circle is in the image, and 0 or 1, if a circle is absent in the image. The predictive power of the interpretable model is substantially higher than the black-box (DL) model, with an increase of 0.075 in BAcc.

5 DISCUSSION

MultiFIX is a unique approach to explainable multimodal learning that enforces a small number of features to be learned from different modalities. The strengths of DL and GP are leveraged to learn interpretable models that integrate symbolic expressions for both interpretable tabular feature extraction and interpretable fusion of different modalities, and image explanation techniques such as Grad-CAM. Our results showcase the potential for developing models that are interpretable by design without sacrificing performance, specifically in problems that include tabular data or combine heterogeneous data modalities. This work contributes to the growing need for explainable multimodal methods in high-stakes domains, such as healthcare.

In comparison to post-hoc methods like SHAP, GP-GOMEA directly generates interpretable models without additional approximations, reducing the risk of misleading explanations. Furthermore, the incorporation of interpretability at the feature level enables directly understanding which variables are relevant for the task at hand. Lastly, our experiments indicate that end-to-end and hybrid training achieve better performance than sequential training approaches. However, statistical testing indicates that most of the studied training approaches obtain similar performance. Thus, the choice of training strategy should rely on the specifications of the problem. For instance, one may have a powerful feature engineering block for one of the modalities that can be used either as pre-training weights or as a sequential block. Additionally, in many multimodal problems, there are different amounts of samples per modality, which usually translates in removing samples from the oversampled modality. Using hybrid and sequential training, the total amount of samples can be leveraged to learn more accurate feature engineering blocks.

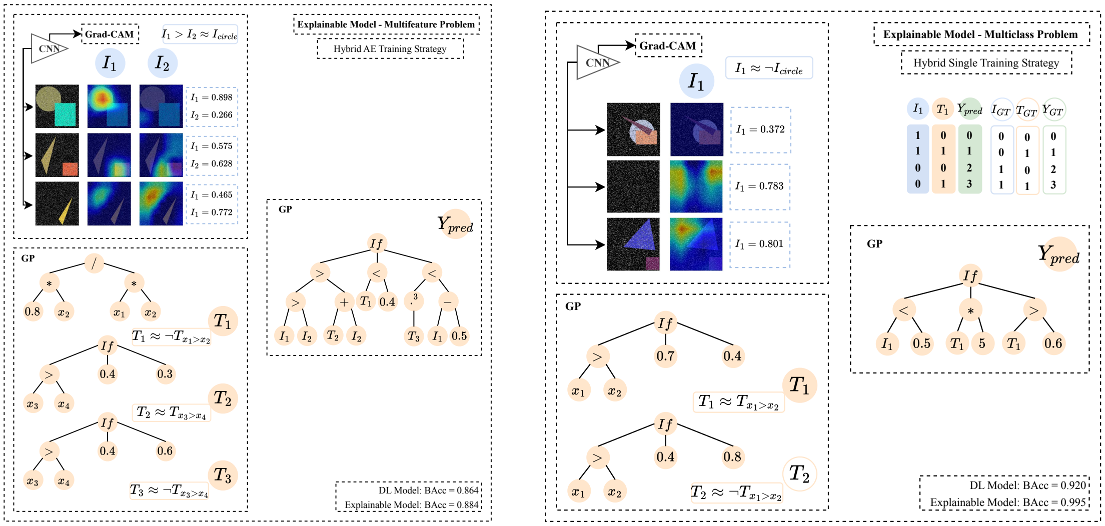
While providing a unique step toward interpretable multimodal learning, the way we have trained and used MutiFIX so far has clear limitations as well. Firstly, we have here considered a limited set of problems with a limited number of modalities. While this is important to be able to fundamentally study the possibilities and limitations of MultiFIX when the ground truth is known, the added value offered by MultiFIX must be corroborated on real-world datasets, which we intend to do in the near future.

When more than one complex feature is needed, as demonstrated in the Multifeature Problem, the complexity of the intermediate image features can hinder interpretability. Moreover, requiring more features in the bottleneck creates possibilities to obtain different, but equivalent models, making it not always easy to interpret what the model is doing, as at first it may be counterintuitive. Forcing the image block to engineer simpler features can reduce the overall complexity and increase interpretability. More generally, the penalization of complexity as an additional objective is likely advantageous. For the discovery of symbolic expressions, using a multi-objective variant of GP-GOMEA could help to find models of different sizes, some of which are more easily interpretable in their own right, or may provide hints as to what slightly larger (and potentially more accurate) expressions are capable of modeling.

The use of Grad-CAM to explain image feature extraction highlights regions used by the DL model that contribute to the engineered feature. However, the post-hoc nature of Grad-CAM limits the symbolic meaning that can be associated with each feature. Having inherently interpretable image blocks would be highly beneficial from an interpretability perspective. Lastly, the interpretable models generated by MultiFIX, although readable and transparent, can further benefit from user-friendly presentations that can

	Image Only	Tabular Only	End-to-End	Hybrid AE	Hybrid Single	Sequential AE	Sequential AE Defreeze	Sequential Single
Image Only	0.485 ± 0.004	not significant	worse	worse	worse	not significant	worse	worse
Tabular Only	not significant	0.453 ± 0.014	worse	worse	worse	worse	worse	worse
End-to-End	better	better	0.823 ± 0.038	not significant	not significant	not significant	not significant	not significant
Hybrid AE	better	better	not significant	0.858 ± 0.036	not significant	not significant	not significant	not significant
Hybrid Single	better	better	not significant	not significant	0.919 ± 0.007	not significant	not significant	not significant
Sequential AE	not significant	better	not significant	not significant	not significant	0.691 ± 0.061	not significant	not significant
Sequential AE Defreeze	better	better	not significant	not significant	not significant	not significant	0.849 ± 0.053	not significant
Sequential Single	better	better	not significant	not significant	not significant	not significant	not significant	0.742 ± 0.055

Table 7: Multiclass Problem - BAcc Results with Statistical Testing: average BAcc and standard deviation over 5 folds are highlighted in blue; each row indicates statistical significance comparison with the remaining approaches: better, worse or not significant.



I_1	T_1	Y_{pred}	I_{GT}	T_{GT}	Y_{GT}
1	0	0	0	0	0
1	1	1	0	1	1
0	0	2	1	0	2
0	1	3	1	1	3

(a) Multifeature Problem Hybrid AE Training

(b) Multiclass Problem Hybrid Single Training

Figure 4: Interpretable Models: Grad-CAM heatmaps explain the image input contributions for each extracted feature. GP-GOMEA symbolic expressions explain the tabular features and the fusion of both modalities to make the prediction. Learned features and predictions are compared to their GT counterparts.

enhance overall interpretability. This includes further visual explanations, approximations, and simplifications for the user, as well as ways to perform interactive and sample-based analyses.

6 CONCLUSION

In this paper, we present the first comprehensive experimental work on MultiFIX, a novel multimodal pipeline to obtain multimodal, interpretable models. The uniqueness of MultiFIX is reflected in its interpretability-focused design by forcing a limited number of features per modality to be automatically engineered and subsequently used to make predictions. DL can be used to perform feature engineering, whereas GP-GOMEA can be used to evolve interpretable symbolic expressions for tabular engineered features and for the final fusion. Modality-specific post-hoc explainability techniques can be used, such as Grad-CAM for images to explain the overall model in a component-wise fashion.

Considering the different multimodal training strategies that we have studied, there seems to be no statistically significant difference in performance for the created benchmark problems. The choice of the most suitable training strategy is up to the specifications of the task at hand. In general, our results demonstrated that MultiFIX can accurately capture multimodal relationships and that learned models have high potential for interpretability.

Despite these advancements, MultiFIX needs further improvements, predominantly interpretability enhancements that minimize complexity and promote inherently interpretable methods for image features, in combination with additional interactive visualization tools to intuitively clarify the mechanics of the learned models. Furthermore, we aim to evaluate our pipeline on real-world datasets that involve more modalities as well as more complex intermodal relationships.

In conclusion, we believe that we have demonstrated the feasibility and potential to perform interpretable multimodal learning by leveraging a unique feature-inducing architecture combined with inherently interpretable methods across heterogeneous data types with MultiFIX.

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.

Code availability.

https://github.com/mafaldamalafaia/MultiFIX_GECCO25_code.git

REFERENCES

[1] Johannes Allgaier, Lena Mulansky, Rachel Draelos, and Rüdiger Pryss. 2023. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine* 143 (09 2023), 102616. https://doi.org/10.1016/j.artmed.2023.102616

[2] Nitin Arora, Anupam Singh, Vivek Shahare, and Goutam Datta. 2023. Introduction to Big Data Analytics. In *Towards the Integration of IoT, Cloud and Big Data: Services, Applications and Standards*. Springer, 1–18.

[3] Jaume Bacardit, Alexander EI Brownlee, Stefano Cagnoni, Giovanni Iacca, John McCall, and David Walker. 2022. The intersection of evolutionary computation and explainable AI. In *Proceedings of the Genetic and Evolutionary Computation conference companion*. 1757–1762.

[4] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 40, 8 (2022), 865–878.

[5] Benjamin P Evans, Bing Xue, and Mengjie Zhang. 2019. What's inside the black-box? a genetic programming method for interpreting complex machine learning models. In *Proceedings of the genetic and evolutionary computation conference*. 1012–1020.

[6] Jacob Gildenblat and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

[7] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. 2024. A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications. *arXiv preprint arXiv:2408.02686* (2024).

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[9] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* 3, 1 (2020), 136.

[10] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. 2021. A review on explainability in multimodal deep neural nets. *IEEE Access* 9 (2021), 59800–59821.

[11] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)* (2015).

[12] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine* 5, 1 (2022), 171.

[13] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabricio de Franca, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason Moore. 2021. Contemporary Symbolic Regression Methods and their Relative Performance. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. Curran.

[14] Mafalda Malafaia, Thalea Schlender, Peter AN Bosman, and Tanja Alderliesten. 2024. MultiFIX: An XAI-friendly feature inducing approach to building models from multimodal data. *arXiv preprint arXiv:2402.12183* (2024).

[15] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion* 81 (2022), 203–239.

[16] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[17] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 16 (2022), 1–85.

[18] Thalea Schlender, Mafalda Malafaia, Tanja Alderliesten, and Peter Bosman. 2024. Improving the efficiency of GP-GOMEA for higher-arity operators. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 971–979.

[19] Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. 2024. Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. *arXiv preprint arXiv:2411.03782* (2024).

[20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[21] William C Sleeman, Rishabh Kapoor, and Preetam Ghosh. 2022. Multimodal classification: Current landscape, taxonomy and future directions. *Comput. Surveys* 55, 7 (2022), 1–31.

[22] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. 2022. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics* 23, 2 (2022), bbab569.

[23] Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. 2024. Multimodn—multimodal, multi-task, interpretable modular networks. *Advances in Neural Information Processing Systems* 36 (2024).

[24] Marco Virgolin, Tanja Alderliesten, and Peter AN Bosman. 2020. On explaining machine learning models by evolving crucial and compact features. *Swarm and Evolutionary Computation* 53 (2020), 100640.

[25] Marco Virgolin, Tanja Alderliesten, Cees Witteveen, and Peter A. N. Bosman. 2021. Improving model-based genetic programming for symbolic regression of small expressions. *Evolutionary Computation* 29, 2 (2021), 211–237.

[26] Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multimodal Data Fusion. *Comput. Surveys* 56, 9 (2024), 1–36.

[27] Ryan Zhou and Ting Hu. 2023. Evolutionary approaches to explainable machine learning. In *Handbook of Evolutionary Machine Learning*. Springer, 487–506.