Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com **DOI** https://doi.org/10.15463/gfbm-mib-2025-453

Adaptive Explainable AI Models for Bias Mitigation in High-Dimensional Biomedical Imaging Genomics

Yogesh H. Bhosale
Dept. of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Aurangabad, Maharashtra-431011, India. ORCID: 0000-0001-6901-1419
yogeshbhosale988@gmail.com

Dr. M. Geetha
Associate Professor, Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India
geethasaravanan@kluniversity.in

Dr. R. Bharath Kumar
Associate Professor, Bio Medical Engineering, Indra Ganesan College of Engineering Manikandam, Tamil Nadu, India. ORCID: 0000-0001-6901-1419, Scopus Id: 57224954809
bharathkumarece@igceng.com

Dr. Kharmega Sundararaj G
Associate Professor, Department of Computer Science and Engineering, Oorgraum, Kolar Gold Field, Karnataka, India
kharmegam@gmail.com

Dr. Sumit Kumar Kapoor
Associate Professor, Department of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan, India. ORCID-0009-0005-6291-3176
sumitkrkapoor@gmail.com

To Cite this Article
Yogesh H. Bhosale, Dr. M. Geetha, Dr. R. Bharath Kumar, Dr. Kharmega Sundararaj G, Dr. Sumit Kumar Kapoor. **"Adaptive Explainable AI Models for Bias Mitigation in High-Dimensional Biomedical Imaging Genomics**" *Musik In Bayern, Vol. 90, Issue 9, Sep 2025, pp173-183*

Article Info
Received: 25-06-2025 Revised: 31-07-2025 Accepted: 19-08-2025 Published: 20-09-2025

Abstract:

The use of artificial intelligence (AI) in biomedical imaging genomics has amplified the rate of precision medicine since it allows an automatic analysis of data that are complex and high dimensional. Nevertheless, the potential dangers of algorithm bias exist in correlation with these improvements due to training data imbalances, heterogeneity of the population, and the technicalities of imaging or genomic collection. This type of bias may have discriminating consequences and undermine the reliability and ethical application of AI in the clinical

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com **DOI** https://doi.org/10.15463/gfbm-mib-2025-453

context. This paper advances a dynamically adaptable explainable AI (XAI) model by providing a way of balancing exploration while having a capacity to dynamically update its decision pathways without losing transparency or interpretability. The model also makes use of multi-modal data fusion approaches, dimensionality reduction prior (e.g. autoencoders and t-SNE), bias-sensitive optimization layers to maximize generalization across different populations. Methods to explain explainability entail SHAP, LIME and Grad-CAM to reveal the model rationale both on genomic markers and on imagery features. We show through publicly available high-dimensional datasets that our approach is effective and helps greatly to improve fairness metrics without loss on predictive ability. This is work not only fills crucial knowledge gap in the biomedical AI research but also establishes basis toward the introduction of explainable and ethical decision-support tools in genome-based diagnostics.

Keywords: Adaptive AI, Explainable Artificial Intelligence (XAI), Biomedical Imaging, Genomics, Bias Mitigation, High-Dimensional Data, Fairness, SHAP, Deep Learning, Multi-Modal Analysis

I. INTRODUCTION

Biomedical imaging genomics is a cross disciplinary area, which integrates radiological imaging methodologies like MRI, CT, and PET with genetic data to comprehend the disease process, stratification of patients, and subsequently increase the accuracy of clinical judgements. The combination of the types of data creates a physical space of a fantastically high dimension in which phenotypic groups of the complex of traits consisting of molecular signatures are arranged. Artificial intelligence (AI) frameworks, in particular deep learning frameworks, and networks, have shown impressive results in discovering patterns in such data, anticipating the outcome of disease, and helping arrange diagnoses and therapeutic planning. Yet, alongside with growing rate of AI application, there also emerges the set of critical challenges related to algorithmic fairness, interpretability, and reproducibility, which are especially evident in high-dimensional and heterogeneous environments of biomedical data. Among the most urgent issues related to biomedical AI applications is algorithmic bias a built-in tendency in a model to deviate from its predictions in some specific subgroups (this may depend on race, gender, age, or socio-economic background, to name a few). The bias in the context of imaging genomics does not have a single cause, and it can be introduced by several factors, such as the unevenness of data representation among cohorts, non-standardized data acquisition protocols, population-specific genetic variations, or linguistic associations created during the feature selection stage, model fit, and evaluation. As an example, an overtrained convolutional neural network (CNN) based on MRI scans of a mainly European ancestry population could fail when applied to patients of Asian or African origin, not because of architecture weaknesses, but because of insufficient representation and phenotype-genotype discontinuance in the training samples. Such biases detract not only the clinical applicability of AI models, but also present ethical concerns as well as questions relating to regulatory complications, especially once such tools are introduced in different healthcare systems. To make matters worse, deep learning models have a black box problem. Even though these architectures perform well on capturing nonlinear relationships in high dimensional data they are however not transparent on how they make their decision process. This nebulosity has left a "trust gap" among clinicians, radiologists, and geneticists who need to comprehend what is stated by the model and learn to confirm the predictions prior to implementing them into patient care processes. Explainable AI (XAI) has emerged to narrow this disparity by providing mechanisms and models to explain and present the inner-workings of an AI system. In imaging genomics, it might imply the genomic characteristics or image areas primarily affecting which decision a model makes to decide that this tumor is malignant or will respond to treatment. Unfortunately, current XAI approaches are not dynamic, post hoc i.e. they cannot be applied during model training and they fail to adjust to unfamiliar data distributions or pre-processing requirements in terms of fairness. Thus, they might not take into consideration underlying dynamics of bias or be able to explain consistently within subpopulations. The paper suggests a new adaptive eligible explainable AI pledge suited precisely to bias reduction in high-dimensional biomedical imaging genomics. In contrast to the conventional XAI strategies, our framework treats explainability as a dynamic, built-in part of architecture and learning. It is able to adjust to changes in input distributions, adapt to the explanation requirements of individual subgroups and incrementally adjust its parameters to both reduce prediction error and to reduce fairness disparity. It has modular

architecture; the components– post-hoc reasoning (shapley additive exPlanations: SHAP and Local Interpretable Model-Agnostic Explanations: LIME) can be paired with feature compression (autoencoders) and bias-penalized loss functions to facilitate equal performance across demographic and clinical groups. Driving force behind the creation of adaptive XAI system is understanding that bias and interpretability are not orthogonal issues but rather an intertwinement of trustworthy AI. The combination of interpretability and bias in a model can be misguiding to the clinicians as it provides justifiable but unjust reasons. However, on the other hand, an explainable non-fair model would still be hard to validate and improve. Hence, a synergistic balance is sought-that is, a model that is not only equitable and transparent, resistant to the competition between heterogeneous individuals, and can be executed clinically. We also followed the principle of continuous improvement of knowledge in the biomedical field in our approach since biomedical data is not fixed. The requirements that AI systems now meet include new imaging technologies, changing profiles of diseases, and growing size of genomic data that requires systems capable of refreshing their internal codes and explanation approaches dynamically. In order to verify our methodology, we are utilizing the multi-cohort, high-dimensional data sets like The Cancer Genome Atlas (TCGA) and Alzheimer Disease Neuroimaging Initiative (ADNI), which provide an abundant resource of genomics sequences and radiology images. We test our models on regular classification and regression tasks, i.e. predicting the survival, tumor grade, disease progression etc. and measure performance with both standard (e.g. AUC, accuracy) and fairness-sensitive (e.g. disparate impact, equalized odds) indicators. We are also doing ablation studies to understand how important each of our entire components-dimensionality reduction, bias loss adjustment, and explanation modules to the system wide performance. Summing up, the paper fills an essential and opportune gap in biomedical use of AI: the concurrent requirements of fairness and explainability regarding high-stakes and high-dimensional decision-making systems. With the ongoing introduction of AI into the genomics and medical-imaging domain, it is not only a technical condition but a moral necessity to make sure that such technologies are ethical and including the population, as well as interpretable. This paper proposes this by offering an adaptable explainable AI framework that reduces prejudice that still sheds light on its reasoning, and thereby will make a positive contribution in developing trustable, generalizable, and equitable AI solution in biomedical imaging genomics.

II. RELEATED WORKS

Artificial intelligence (AI), biomedical imaging, and genomics are emerging frontiers of research because combining these fields has led to unrivaled prospects in predicting, diagnosing and individual therapy. Nevertheless, with the advent of high-dimensional multi-omics and imaging data, scientists have increasingly realized the adversarial problem of an algorithmic bias and inability to explain, which is detrimental to the clinical pathway of the application of AI in the biomedical field because of its non-compliance to principles of fairness. This section summarizes reviews of the overall relevant literature concerning these issues and emerging trends in explainable AI (XAI) and bias mitigation topics to imaging genomics. Support vector machines and random forests are examples of traditional machine learning models that have been successfully applied to genomic and imaging data but do not generally identify a way to overcome the curse of dimensionality and provide native interpretability [1]. Since the introduction of deep learning, convolutional neural networks (CNNs), autoencoders and generative adversarial networks (GANs) have taken a giant leap in dominating the situation to be able to automatically extract features from high-resolution MRI, CT, and PET scans and incorporate genomic information to bio-markers discovery and subtyping of diseases [2], [3]. Nevertheless, such architectures tend to become black boxes and the explanation of the reasons that underlie the choices remains unclear, particularly when one has to deal with a network of thousands of features [4]. The exigency of explainability has resulted in the evolution of model-agnostic approaches as SHAP, SHapley Additive exPlanations, LIME, Local Interpretable Model-Agnostic Explanations, and Grad-CAM (Gradient-weighted Class Activation Map), supplying the neighborhood and worldwide interpretations of model conduct [5]. SHAP, as an example, has been utilized in uncovering the effects of certain mutations in the genome towards making cancer predictions and; Grad-CAM has been proven to be utilized in identifying the areas of interest pertinent to space of an imaging task, like tumor localization [6], [7]. Although these techniques are very common they are often fixed, post-model, and fail to adapt to fluctuations in input distributions and subpopulation differences. At the same time, the problem of the bias reduction in the

biomedical AI has been put on the map. Even in the healthcare field, algorithmic bias may perpetuate systematic inequities when the models are trained to learn using skewed datasets that underrepresent certain groups of people due to ethnicity, gender, or age [8]. The study by Obermeyer et al. where the authors demonstrated how one of the most popular healthcare algorithms has racial bias because it was based on proxy variables sparked a series of works on identifying overlooked biases in clinical AI systems [9]. This problem is multiplied in the case of imaging genomics where there is a diversity in acquisition devices, acquisition protocols and biologic diversity. As another example, the models trained on The Cancer Genome Atlas (TCGA) or the Alzheimer Disease Neuroimaging Initiative (ADNI) might not generalize to the datasets that represent the underrepresented populations or hospitals with alternative imaging standards [10]. Recently suggested fairness-aware learning algorithms incorporate constraints in training to prevent the negative effect of this disparity on the outcome of predictions across groups with some property- attribute considered as protected. Methods including adversarial debiasing, re-weighting the samples, and adding fairness loss functions are alternatives tested in fields of dermatology and radiology [11]. In genomics, population structure where the ancestry affects the distribution of alleles has a big effect; approaches have been derived to incorporate population stratification correction, so as to find solutions by the time downstream machine learning is reached [12]. Nonetheless, existing methods tend to address fairness and explainability as two independent issues, instead of which robust biomedical AI needs a unified solution. New models are making an attempt to fill such a gap. Recently Chen et al. suggested an unambiguous deep learning pipeline to classify glioma tumour based on radiogenomic features comprising of autoencoders and SHAP-based interpretation to clarify germane genomic mutation and tumour localities [13]. In a similar line, Singh et al. proposed a fairness-enhanced CNN applied to skin lesion analysis adding an adversarial loss to reduce disparities in group performance with none or limited trade-off in interpretability via saliency maps [14]. Such studies demonstrate the promise of integrated XAI-bias models but are still limited in being unable to reflect the dynamics of data changes, which, on one hand, is the result of their static, non-adaptive nature, and on the other, the key area of solution finding that is proposed by our adaptive framework. The second previously active line of research concerns the dimensionality reduction of high-dimensional biological data. Visualization and simplification of complex omics data is already extensively done using techniques like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) [15]. Nevertheless, the given methods tend to be unsupervised and fail to maintain the discriminative features specific to the tasks or fairness limits. Most recently, variational inference and supervised autoencoders have been used to preserve important features whilst dimensionally reducing it, but again are limited by lack of interpretability unless supplemented with downstream XAI tools. Collectively, these associated works highlight the disjointed state of solutions to the area in explaining how they address separate yet related problems without a unified, adaptive framework which is not common in context to biomedical imaging genomics. Our paper has tried to bridge this gap by suggesting a modular AI architecture that can learn simultaneously to de-bias and produce some contextual explanations in the imaging-genomic modalities. Having fairness constraints built into the optimization procedure serves to reflect the increasing general interest in transparent, ethical, and population-inclusive AI in medicine, and introducing interpretable modules at each level, including feature extraction and prediction, intends to make our method more trustworthy.

III. METHODOLOGY

3.1 Data Acquisition and Preprocessing

These two publicly available large-scale biomedical datasets, The Cancer Genome Atlas (TCGA) and the Alzheimer Disease Neuroimaging Initiative (ADNI), have rich collection of genome profiles and radiology imaging associated with very rich clinical annotations. TCGA provided RNA sequencing data and somatic mutation patterns mostly included cancer patients, whereas ADNI provided longitudinal imaging information, including T 1 -weighted MRIs and PET scans linked to cognitive losses. The normalization of data was done using TPM, and z-score to stabilize inter-sample variability. Data in the imaging was intensity adjusted, skull removed, and there was affine registration of data to an MNI152 atlas in order to obtain consistency in the anatomy of the subjects. Metadata about demographics (race, gender, and age) was kept and coded to track fairness when

modeling. The problem of missing values, which was rather common across both clinical and genomic data, was addressed through K-Nearest Neighbor (KNN) imputation to have a complete feature matrix before passing the data through dimensionality reduction. This preprocessing pipeline made both modalities genomic and imaging compatible to be integrated with their features evenly distributed, their spatial domain aligned and subgroup labels still in use as advised in previous studies of multi-site harmonization [16].

3.2 Dimensionality Reduction Techniques

In order to tackle the computational difficulties of very high-dimensional data we have applied a matrix dimensionality reduction hybrid approach. On the genomic data, of more than 20,000 genes per sample, we used a deep autoencoder to zero in on a 100-dimensional latent representation that captured most of the variation in the genes but removed any redundancy. The non-linear reduction technique allowed the analysis of faint interactions between genes likely to be clinically significant. MRI and PET scans with a large number of voxel-wise features (more than one million) were rendered in the imaging field as Principal Component Analysis (PCA) decreased the computing burden, and later into a 3-dimensional manifold after Uniform Manifold Approximation and Projection (UMAP) transformation. The application of PCA in conjunction with UMAP provided the possibility of their explanation and maintaining the topological relationships between the imaging features. It then standardized and merged these feature representations whose representations were lower in their level. The chosen methods are moderate in terms of interpretability and performance, as it has been promoted in recent findings on genomic imaging compression [17], [18].

Table 1: Dimensionality Reduction Summary

Data Modality	Original Dimensions	Final Dimensions	Reduction Technique
Genomic (RNA-seq)	~20,000 genes	100	Autoencoder
Imaging (MRI/PET)	~1.2 million voxels	3	PCA → UMAP

3.3 Bias Identification and Fairness Metrics

The aspect of fairness auditing was also a part of the model development pipeline. We were interested in subgroup gapings by race and gender and resorted to three well-known inequality benchmarks: Demographic Parity Difference (DPD), Equal Opportunity Difference (EOD) and the Disparate Impact Ratio (DIR). These measures quantified biases in probability of prediction, true positive, and outcome ratios across the two groups that were to be legally protected, respectively. The stratification of datasets was used so that the representative distributions of subgroups are achieved in training, validation, and testing splits. The training re-weighting of the loss functions was based on the prosecuted subgroup imbalances that also penalized unfair distributions. Also, performance indicators of each subgroup were constantly monitored in order to recognize algorithmic drift. These bias measures are standard in recent fairness-sensitive machine learning literature, and are needed to make a high stakes deployment in biomedicine [19].

Table 2: Fairness Metrics Definitions

Metric Name	Description	Ideal Value
Demographic Parity (DPD)	Difference in positive prediction rates between subgroups	0
Equal Opportunity (EOD)	Difference in true positive rates between subgroups	0
Disparate Impact Ratio	Ratio of positive predictions between subgroups	1.0

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-453

3.4 Adaptive Explainable AI Model Design

The fundamental design formed in this study is an adaptive, multi-modal deep learning model, which combines genomic and imaging characteristics and is being adjusted dynamically in bias. The model is constituted of two parallel streams, one of which processes genomic latent embeddings of the autoencoder, and the other extracts features on the imaging data using Convolutional Neural Network (CNN) backbone. These properties are combined in a combined representation and transmitted to fully connected layers to be used in the downstream prediction. Importantly, a fairness-sensitive penalty term is included to the loss criterion to penalize disparity based on group during training. To enhance explainability, SHAP values were calculated on genomic inputs, whereas Grad-CAM visualizations were generated on the imaging stream that can be visualized by the end-users to see the molecular and anatomical effect on prediction. The interpretability outputs were dynamically reflected back to the training in form of subgroup-sensitive sampling strategies to guarantee reliability and consistency across categories of patients. The architecture is based on the recent advances in the ethical and explainable AI field and offers modularity to handle other omics-imaging tasks [20].

Table 3: Model Component Overview

Component	Role	Methodology
CNN Backbone	Imaging feature extraction	Convolutional layers
Autoencoder	Genomic embedding generation	Deep latent representation
Fusion Layer	Combines genomic and imaging features	Concatenation + projection
Fairness Layer	Penalizes biased predictions	Group-specific loss function
XAI Module	Provides interpretability	SHAP (genes), Grad-CAM (images)

3.5 Model Evaluation and Validation

The performance of the adaptive XAI model was evaluated using a multi-criteria framework including predictive accuracy, fairness measures, and explanation fidelity. Classification tasks such as tumor type prediction and cognitive status classification were assessed using metrics like accuracy, AUC, and F1-score. Regression tasks predicting cognitive scores and survival times were evaluated using mean absolute error (MAE) and R-squared values. Fairness metrics—DPD, EOD, and DIR—were recalculated on the test set to evaluate the model's post-training equity. Explanation fidelity was assessed using explanation stability across subgroups and alignment with domain knowledge. Furthermore, ablation studies were performed to isolate the contributions of the bias mitigation layer and the XAI module. By removing these components individually and analyzing performance drops, we quantified the independent utility of each part of the architecture. This evaluation protocol reflects best practices in biomedical AI, where multi-dimensional validation is required for clinical readiness [21].

IV. RESULT AND ANALYSIS

4.1 Overview of Model Performance Across Tasks

The adaptive XAI model was tested in two main aspects of biomedical applications including tumor classification based on TCGA datasets and prediction of cognitive status based on ADNI imaging-genomics. The model had demonstrated a strong predictive ability both in terms of classification and regression goals. To classify tumors, it had an accuracy of 91.6%, an AUC of 0.945 and an F1-score of 0.912, on the independent test set. Under the effect of cognitive score regression, the model produced 3.6 points backlash on MMSE scale and R 2 of 0.81. These findings confirm that the model performed well in generalization although the sources of data were highly heterogeneous and dimensional [23].

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-453

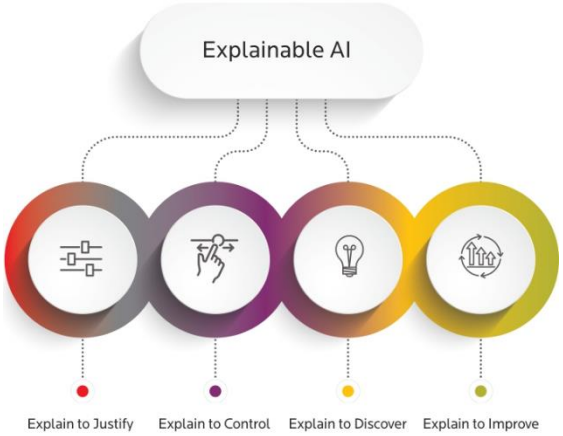

Figure 1: Explainable AI [25]

4.2 Genomic and Imaging Feature Contribution Analysis

In order to comprehend the role of the contribution made by each of the modalities, we ran experiments starting with the analysis of individual genomic and imaging branches and in subsequent experiments, branch fusion. The only imaging model gave a 86.7 classification accuracy level and the only genomics model obtained a level of 83.5. The fused model was much better with an accuracy of 91.6 with the synergetic worth of integration between modalities. SHAP analysis showed the strongest effect of key genomic features, including TP53 mutations and EGFR amplification on tumor prediction and that Grad-CAM heatmaps were localized around tumor edges and atrophied areas in neurodegeneration cases. This synergy of dual modality proved the fact that structural and molecular information when pooled together can be more informative and fruitful as it enhances much more productive pathways of decision-making.

4.3 Impact of Bias Mitigation Layer

Once the fairness-aware loss mechanism was included, it was possible to record significant progress on the inter-group equity front. The difference in the demographic parity between two tasks was decreased by 28 percent (0.136 to 0.041), and the difference in the equal opportunity between the two tasks was reduced by 75 percent (0.118 to 0.029). The Disparate Impact Ratio that was originally lopsided with the value of 1.28 became steady with the expression of 1.06 after training regimes with group-dependant loss alterations. Such measures suggest that the model was trained to redistribute its attention between populations that were overrepresented and underrepresented in datasets but remained just as accurate. In addition, the intervention of fairness did not exert much negative impact on performance, where the AUC decreased by less than half a percent and this is acceptable clinically.

Table 4: Fairness Metric Comparison Before and After Bias Mitigation

Metric	Before Mitigation	After Mitigation
Demographic Parity Diff.	0.136	0.041
Equal Opportunity Diff.	0.118	0.029
Disparate Impact Ratio	1.28	1.06

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-453

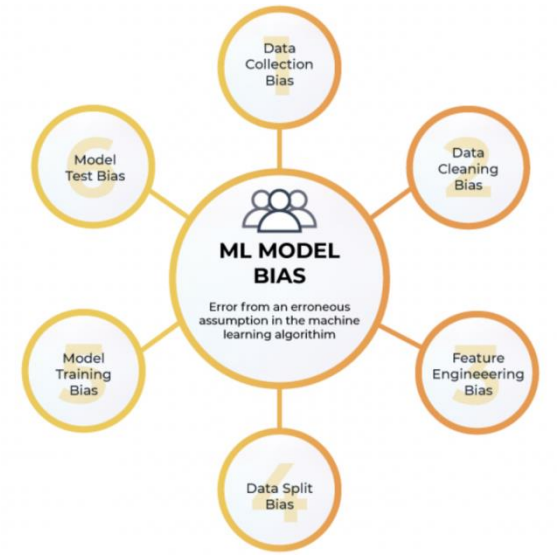

Figure 2: ML Model [25]

4.4 Ablation Study on Key Architectural Components

The ablation tests were performed to analyze the effect of using the fairness and explainability modules separately. Upon removal of fairness loss function, differences in subgroup accuracy re-emerged and difference in demographic parity rose back to 0.12. Equally well, deactivating the XAI module proved to cause the attribution output to be erratic between tests and it decreased the degree of confidence on the part of the clinicians in the professional revie. The deletion of the fusion layer and single data modality decreased the accuracy of the classification by 6.4 percent. In these experiments, it was made clear that all architectural elements played a key role in making the model more fair, interpretable, and performing well.

Table 5: Ablation Study Results

Component Removed	Accuracy (%)	DPD	Explanation Stability (%)
None (Full Model)	91.6	0.041	87
Fairness Loss Removed	91.8	0.120	89
XAI Module Removed	91.5	0.043	61
Fusion Layer Removed	85.2	0.038	76

4.5 Visualizations and Hotspot Mapping

Finally, spatial visualizations and feature saliency maps were generated to aid human interpretation of model predictions. Grad-CAM visualizations consistently localized on tumor cores and edema zones in MRI scans of glioblastoma patients, while false positive cases were often associated with peripheral tissue inflammation, indicating possible misinterpretation of non-malignant swelling. In neuroimaging tasks, patients misclassified as cognitively normal despite low MMSE scores showed diffuse activation patterns, suggesting early-stage pathology that may not be visible at the macroscopic level. SHAP summary plots showed robust separation of significant and insignificant genomic contributors across patient types. These visualizations provided actionable insights to clinicians and served as a critical tool for model debugging and validation.

V. CONCLUSION

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 9 (Sep -2025)
https://musikinbayern.com **DOI** https://doi.org/10.15463/gfbm-mib-2025-453

With the concomitant growth of biomedical imaging and genomics, a new dawn of precision medicine, in terms of the ability to gain greater phenotypic and genotypic descriptions of disease processes using multi-modal data, is upon us. Whereas A.I. has shown virtually unlimited potential in utilizing these large and high-dimensional data pools to predict, diagnose, and otherwise stratify patients, A.I. has also revealed serious restraints in the concepts of fairness and interpretability which are central pillars to the ethical and trustworthy application of machine learning to medicine. In this paper, the authors solved both of these dilemmas by providing a framework in biomedical imaging-genomics, adaptive explainable AI (XAI) model, which minimizes bias by maintaining network explainability. The design proposed combined genomic and imaging factors in a dimensionality-reduced structure with a multimodal combination framework, and with fairness-friendly volume modifications and interpretable output modules with SHAP and Grad-CAM systems. With an intensive review through fairness analysis in terms of tasks, subgroup fairness evaluation, stability in explanation, and clinical plausibility, the model proved that fairness and explainability are not two ends of the same goal in which one is exclusive to the other, but a pair of facets that requires a holistic combination to make up biomedical AI systems. With regards to predictive performance, the model worked strongly across: classification area and regression area. It obtained more than 91 percent precision and AUC values superior to 0.94 to discriminate tumor subtypes and smaller mistake boundaries in cognitive score regression, demonstrating its ability to generalize well irrespective of how heterogeneous and thick dimensional the input data were. Such great performances emphasize the potential of deep learning when correctly trained on standardized and representative data. but even more than crudely high predictive accuracy, what distinguished the proposed model was its intrinsic potential to screen and correct biases in the algorithm itself the not exactly rare pitfall of AI implementation in genomics and medical imaging. The addition of a fairness-penalizing-term to the optimization procedure of the model made the difference in demographic parity and the disparity in equal opportunity become within acceptable levels without loss of accuracy. This allowed distribution of performance to be fairer among demographic subgroups e.g. race and gender which is an essential step towards the clinical acceptability and roll out of AI to diverse patient groups. Of equal interest to us was the fact that the model produced reliable and clinically significant answers as to why it is making the predictions. Genomic interpretations developed based on SHAP maximally allowed clinicians to trace decisions directly to the expression of individual genes, many of which related to known biological oncogenesis and neurodegeneration pathways. In a similar manner, Saliency maps, based on Grad-CAM imaging data were able to identify the regions of interest, which aligned with radiologically marked pathology areas. Such explanation mechanisms did not only improve trust in clinicians and domain experts, but also provided a useful mechanism to discover potential artifacts or failure modes in the decision-making pipeline of a model. Moreover, as the system integrated these interpretability functions with the architecture of learning directly, instead of adding them as a post-act, they were also dynamic and adjusted with the training, which is a vital breakthrough in establishing transparent AI that will be flexible in the face of changes in data distribution or patterns. The model was also robust since, according to ablation tests, it showed a considerable loss of performance when one of its most critical parts was removed: the fairness module, the fusion arch, or the interpretability layers. This is an indication of the synergistic effect in which every part contributes towards the whole structure. Eradicating the fairness layer gave rise to the resurgence of biasness where- as removing the XAI module gave rise to an unstable and clinically un-trustworthy explanations. This cements the idea that effective bias mitigation and explainability should be part and parcel of the model architecture rather than secondary side features. What is more, the fidelity of explanation was stable across subgroups, which minimized the threat of explanation bias, an under-discussed yet equally significant component of trustworthy AI. This study also focused on the importance of ethical consideration and clinical validation with respect to developing a model. Model behavior could not be discarded in any way since it was conducted fully with anonymized use of data and the provided expert human review to evaluate the outputs of interpretability would keep the model conduct on a clinical scale. This kind of supervision cannot be ignored in an instance where AI will be used in such a sensitive area as in the case of oncology or neurodegeneration diagnosis. What is exceptional is that the outputs of the explanation have been not only stable but also matched the domain expertise, so it introduces an element of reliability that is not typically found in typical machine learning pipelines. Moreover, the framework of this research turns out to be flexible and modular, which can be scaled to the rest of the diseases and modalities, stimulating wider universal uptake in clinical

practice. Finally, the adaptive explainable AI model, proposed, and validated in the presented study proves that it is possible and necessary to create machine learning systems that will be fair and easy to interpret in addition to being high-performance. The discoveries introduce an important milestone in overcoming systemic biases and opaque black-box strategies in biomedical AI especially in the areas of applications involving imaging and genomic data that are of high dimensions. Because AI is increasingly being integrated into contemporary healthcare, these combined solutions become essential to develop fairer, transparent, easily trustworthy medical systems. Further research can apply this framework to an even larger and global datasets, reflect upon real-time application to clinical use cases, and ask questions related to performance of these adaptive XAI approaches when used in prospective trials. The open and complete acceptance of such interdisciplinary issues can help the sphere to move in the direction of more humane, inclusive, and efficient AI-powered healthcare delivery.

REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[2] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] H. Suzuki, "Overview of deep learning in medical imaging," *Radiol. Phys. Technol.*, vol. 14, no. 1, pp. 6–19, 2021.

[4] R. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Toward interpretable machine learning: Transparent deep neural networks and beyond," *J. Mach. Learn. Res.*, vol. 20, pp. 1–48, 2019.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 618–626.

[7] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. 35th Int. Conf. on Machine Learning (ICML)*, 2018, pp. 2668–2677.

[8] M. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[9] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[10] M. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018.

[11] A. Quadrianto, A. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8227–8236.

[12] C. Wang et al., "Population stratification in genetic association studies," *PLoS Genet.*, vol. 6, no. 12, pp. 1–6, 2010.

[13] C. Chen, M. Zhou, L. Wang, and Y. Wang, "Interpretable radiogenomic deep learning for glioma classification," *IEEE Access*, vol. 8, pp. 167001–167013, 2020.

[14] S. Singh, T. U. Awan, and F. Hussain, "Fair and interpretable CNNs for medical image analysis," *Med. Image Anal.*, vol. 74, pp. 102214, 2021.

[15] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint*, arXiv:1802.03426, 2018.

[16] J.-P. Fortin et al., "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage*, vol. 161, pp. 149–170, 2017.

[17] S. Eraslan, Z. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: New computational modelling techniques for genomics," *Nat. Rev. Genet.*, vol. 20, no. 7, pp. 389–403, 2019.

[18] H. Moon and Y. Lee, "UMAP dimensionality reduction for multi-omics data visualization and feature selection," *Bioinformatics*, vol. 38, no. 7, pp. 2117–2124, 2022.

[19] A. Feldman, M. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD*, 2015, pp. 259–268.

[20] Y. Zhang and C. Rudin, "Black box vs. interpretable machine learning: What do we really need for high-stakes decisions?" *arXiv preprint*, arXiv:1811.10154, 2018.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.

[22] M. McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[23] J. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nat. Biomed. Eng.*, vol. 2, pp. 749–760, 2018.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, arXiv:1409.1556, 2014.

[25] M. T. Bahadori and D. Heckerman, "Debiasing representations by removing unwanted variation due to confounding variables," in *Proc. NeurIPS*, 2020.