BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Joint explainable and fair AI in healthcare

Reza Shahbazian[1], Irina Trubitsyna[1*] and Seyedeh Leili Mirtaheri[1]

Abstract

The nature of decisions in the healthcare domain necessitates accurate, interpretable, and reliable AI solutions. Explanation Guided Learning (EGL) explores the integration of explanation annotations into learning models to align human and model explanations. In this paper, we propose Explanation Constraints Guided Learning (ECGL), a novel approach inspired by the augmented Lagrangian method that integrates domain-specific explanation constraints directly into model training. The goal is to enhance both predictive accuracy and interpretability, making machine learning models more trustworthy. Experimental results on both tabular and image datasets demonstrate that ECGL maintains high accuracy while incorporating fairness and interpretability constraints. Specifically, ECGL improves predictive accuracy on the diabetes dataset compared to the base model and enhances feature alignment, with SHAP analysis. On average, a 36.8% increase in SHAP importance demonstrates that ECGL effectively aligns model explanations with domain knowledge. Furthermore, ECGL improves the identification of clinically significant regions in pneumonia X-ray images, as validated by both improved Equalized Odds Ratio (EOR) and GradCAM visualizations. ECGL achieves a 13% improvement in the EOR fairness metric, indicating better consistency of predictive performance across different groups. These results confirm that ECGL successfully balances performance, fairness, and interpretability, positioning it as a promising approach for trustworthy healthcare AI applications.

Keywords Artificial intelligence, Explainable guided learning (EGL), Fairness, Augmented lagrangian, Healthcare

Introduction

The nature of decisions in the healthcare domain necessitates accurate, interpretable, and reliable AI solutions. Explainable artificial intelligence (XAI) techniques aim to explain the inside of *a black box* in deep neural networks (DNNs) [1]. The majority of the research body in XAI focuses on generating such explanations, while questions like *whether the generated explanations are reasonable and accurate* have received less attention [2]. In response to such questions, a research direction referred to as *Explanation Guided Learning* (EGL) studies the integration of explanation annotations into the learning models in order to align human and model explanations [2–6]. This ultimately enhances machine learning models' interpretability and generalizability. This is due to a better grasp of the correct rationale and becoming more robust to artifacts and noises that interfere with model predictions [2]. Thus, in EGL, explanations are not just outputs used for human interpretation but are protagonists in guiding the model's training process with the aim of building models that are not only accurate but also transparent and trustworthy. This is particularly important in domains where understanding the reasoning behind decisions is critical, such as healthcare and medical image analysis [7, 8].

The success of EGL depends on the quality and relevance of the explanations used. Moreover, integrating explanation feedback into the training process can

*Correspondence:
Irina Trubitsyna
i.trubitsyna@dimes.unical.it
[1]Department of Computer Engineering, Modeling, Electronics and Systems (DIMES), University of Calabria, Rende 87036, Italy

be computationally intensive, especially if explanations need to be generated and evaluated at every iteration [9]. The literature addresses the computational issue by discussing the scalability and introducing the trade-offs between explainability and performance by leveraging the approximation techniques [9], and utilizing adaptive learning algorithms that select only a prominent subset of explanation constraints [3]. Pre-computed explanation modules or partial explanation constraints have also been proposed to reduce the computational complexity [2]. Rieger et al. penalize the explanations that do not align with existing human knowledge to reduce the overhead with generating explanations at every iteration [5]. Although these pre-computed explanations or partial explanation constraints have been effective in reducing the computational complexity, they may fail in adapting to the model's training. This is because in deep learning models, the decision boundaries continuously change as the training progresses, while the pre-computed explanation constraints are static and remain unchanged. This might reduce the effectiveness of such constraints to provide high-quality explanations. It also limits the ability of the model to optimize interpretability along with accuracy. Adaptive learning algorithms address the issue of pre-computed explanations. However, these methods might overlook interactions between multiple constraints by selecting a prominent subset of explanation constraints. In cases where the existing knowledge is incomplete or inconsistent, penalizing the non-aligned explanations might also introduce biases.

The main idea of this paper is to formulate specific constraints on the explanations and to adopt the rewriting used in the Augmented Lagrangian method to integrate these constraints directly into model training. More specifically, unlike traditional and standard EGL where explanation annotations are taken as a fixed supervisory signal, we start by viewing a new formulation for guiding learning through explanations as Explanation Constraints Guided Learning (ECGL). Novelty in ECGL occurs in three areas: 1) supporting a large variety of specific constraints from different domains such as fairness, relevance of features, etc. in a common optimization framework; 2) employing adaptive Lagrange multipliers and penalty parameters which enable the model to balance accuracy, fairness, and interpretability during training and not using fixed hyperparameters; and 3) formulation such that the constraints are iteratively refined by the model's state and thus are capable of filling available gaps instead of being static. These constraints will be used to design specific criteria, such as background knowledge satisfaction and fairness requirements, that the explanations should meet. The importance assigned to each constraint is refined iteratively based on its contribution to optimizing the overall objective function. This ensures a comprehensive inclusion of constraints, adaptability during training, and a more balanced optimization of multiple objectives.

The Lagrangian method takes the objective function and the constraints as input and transforms the original objective function by integrating components that quantify the constraints' satisfaction. The method involves adding extra variables, called Lagrange multipliers, to represent the importance of each constraint. Thus, the violations of the more important constraints are penalized more harshly than violations of the less important constraints. The modified objective function is optimized over the original variable space, making sure that the Lagrange multipliers are set correctly to find the best upper and lower bounds that can be reached for the original problem [10]. Therefore, the importance of the constraints is not defined a priori but is discovered by design.

By integrating constraints on the explanation directly into model training, we aim to perform an optimization on both the prediction accuracy and the quality of explanations simultaneously and we try to ensure that the model does not sacrifice interpretability for accuracy or vice versa. In a technical sense, adaptive Lagrange multipliers and penalty parameters let the model constantly find the best balance between fitting the training data and meeting the explanation constraints. We focus on the healthcare domain to better highlight the performance of the proposed method on both tabular and medical image datasets.

Organization

The rest of this paper is organized as follows: Section 2 reports the related works dealing with explainability and fairness issues in the context of the healthcare domain. ECGL is presented in Sect. 3, where we discuss the problem formulation, the incorporation of explanation constraints, and the construction of the Augmented Lagrangian function. The experimental setup, utilized datasets, and evaluation metrics, followed by a detailed discussion of the results, are provided in Sect. 4. We performed experiments on healthcare datasets for diabetes risk prediction, pneumonia, and multi-label Chest X-ray classification, showing improvements in accuracy, fairness, and interpretability. Finally, Sect. 5 concludes the paper with a summary of findings and potential future research directions.

Related works

It is crucial for humans to understand the decision-making processes of complex learning models, particularly in domains such as healthcare [8]. Several studies have focused on developing trust, transparency, and interpretable techniques. In the following, we provide an overview of the state-of-the-art methods dealing with

explainability and fairness issues in the context of the healthcare domain.

Explainability

XAI methods are commonly categorized into three main groups, providing a clearer understanding of their application and orientation. A widely used taxonomy of XAI methods classifies them into:

1. **Gradient-based methods**, which propagate backward the influence of input features on the output. Representative techniques include Saliency Maps [11], Integrated Gradients [12], and GradCAM [13], along with later variants. These methods are computationally efficient and directly tied to the model parameters, but they are often noisy and sensitive to perturbations [14].
2. **Perturbation-based methods**, which provide explanations by observing the responses to modification or occlusion of input features. LIME [15] and SHAP [16] are prominent examples. They can produce more intuitive explanations but are computationally expensive and may lack stability across runs [17].
3. **Surrogate-model methods**, which approximate the behavior of a complex black-box model with a simpler interpretable model, such as decision trees or linear models [18, 19]. These methods are easy to interpret but risk oversimplifying the underlying model, leading to fidelity issues.

Compared to perturbation- or surrogate-based approaches, gradient-based methods are especially appealing in medical imaging contexts because they scale well to high-dimensional data, provide explanations tightly coupled with the learned representations, and can be directly integrated into training to guide optimization. For these reasons, our work is closest in scope to the first category, as we use gradient-based explanations to influence the training process itself. Building upon this taxonomy, we next review the studies tailored to XAI, particularly focusing on medical data analysis.

Van der Velden et al. [8] provide a comprehensive overview of existing XAI methods within the context of medical image analysis. The review highlights the importance of understanding the underlying rationale behind model predictions. Dong et al. [20] attempt to address the issue of bias and lack of transparency in deep neural networks by proposing a graph-based object-guided consistency interpretation enhancement (OCIE) method. Their approach aims to improve the interpretability and ensure fair feature learning. Their experimental results show the effectiveness of focused explanations on the reliability of the network estimations. Zhao et al. [21] focused on lung cancer diagnosis and developed a robust and explainable detection framework for pulmonary nodule detection, incorporating explanation supervision and imputation methods. Similarly, the authors in [22] and [23] demonstrated the practical applications of XAI in the medical domain, with a system being developed for brain tumor surgery. These studies underscore the critical role of explainability in deep learning models, especially in the medical domain. By providing insights into model decision-making, XAI can foster trust between clinicians and AI systems.

Pukdee et al. [3] proposed a theoretical framework for learning from explanation constraints, shedding light on the potential benefits of integrating explanations into the learning process. The authors in [5, 24] have delved into the interpretability and regularization of deep learning models, further emphasizing the growing significance of explainability in this field. In [24] the authors focus on explaining and regularizing models by examining and selectively penalizing their input gradients, resulting in efficient explanation generation and regularization. Whereas the method proposed in [5] enables practitioners to leverage existing explanation methods to improve predictive accuracy, prioritizing model performance over explanation generation. While both studies demonstrated the effectiveness of their respective approaches in providing explanations and enhancing model performance, their suitability may vary depending on specific application requirements and prioritizing either model explainability or prediction performance.

Several studies have focused on improving the quality and effectiveness of explanations [25–28]. For instance, Zhang et al. [25] investigated the challenges posed by noisy and missing annotations in explanation supervision, while Yan et al. [26] utilized human explanations to identify meaningful categories within social images. In contrast, Gao et al. [29] addressed the limitations of human explanations by handling inaccurate boundaries, incomplete regions, and inconsistent data distributions. Additionally, the authors in [28] improved visual explanations through the incorporation of an attention mechanism. Selvaraju et al. [13] introduced GradCAM, a widely adopted technique for generating visual explanations that emphasize crucial regions within images. The authors in [30] explored the use of explanation-guided training to enhance the focus of deep learning models on relevant image features. Some researchers have also delved into the potential risks and challenges associated with XAI. For instance, the authors in [31, 32] indicate that there is a need for careful consideration of privacy concerns, in addition to considering the trade-off between model transparency and faithfulness, in order to improve explanation plausibility and user trust.

Overall, even though perturbation- and surrogate-based XAI provide interpretation and intuitive explainability, they are rather expensive and tend to suffer from unmatched fidelity when applied to extremely high-dimensional tasks such as those found in medical imaging. In contrast to these two methods, gradient-based methods, despite being sensitive to noise, can provide a fair mix between scalability, faithfulness, and trainability that encourages their application in the present study.

Fairness

Fairness addresses the identification and mitigation of biases in learning systems that may result in discriminatory outcomes affecting specific groups. This is particularly important in domains such as healthcare, where biased predictions can have significant negative consequences [33, 34]. Several studies have highlighted the biases that can arise in deep learning models, leading to discriminatory outcomes against certain groups [33, 34]. Tian et al. [34] provide a comprehensive survey on fairness protection methods, emphasizing the importance of addressing fairness in machine learning models. These methods are generally categorized into three main approaches: data-level, model-level, and post-processing techniques. Data-level techniques aim to address biases in the training data, while model-level approaches focus on modifying the model architecture or training process, and post-processing techniques involve modifying the model's predictions after training [35–38]. The authors in [33] explore various bias-reducing methods for fairness-aware neural networks in the context of vision and language, stressing the importance of addressing algorithmic biases and proposing a novel taxonomy to organize the literature on bias neutralization methods. This survey offers valuable insights into the state-of-the-art de-biasing techniques and challenges in achieving fairness in deep learning.

The authors in [39] evaluated the impact of popular fairness models on overcoming biases across subgroups in medical image analysis. Their findings demonstrate the effectiveness of these models in mitigating fairness issues but also highlight the potential trade-offs with uncertainty estimation. This underscores the importance of carefully considering the implications of fairness models in real-world applications. Moreover, there are several challenges in relation to considering fairness in deep learning models. Namely, the literature lacks a single, universally accepted definition of fairness in machine learning [40]. Common notions of evaluating fairness include demographic parity, equal opportunity, and disparate impact. These definitions can conflict, making it difficult to choose the most appropriate one for a given application. Similarly, many deep learning models exhibit an inherent complexity that renders them difficult to interpret. This complexity poses significant challenges in comprehending how fairness is being attained or undermined, as well as in recognizing and addressing biases [15]. It has also been shown that biased training data (data bias) may potentially lead to discriminatory outcomes against marginalized groups, resulting in algorithmic discrimination [41–43]. Furthermore, ensuring fairness frequently might necessitate making compromises with accuracy, as a model may need to forfeit some predictive capability in order to prevent discrimination against specific groups [44]. Addressing these challenges requires a multi-faceted approach, including meticulous data curation, thoughtful and deliberate algorithm design, and continual monitoring and assessment of model performance.

ECGL in the context of explanation guided learning

ECGL, while inspired by the EGL paradigm, gives a completely different optimization philosophy. In most cases, traditional methods of EGL involve using human explanations as a kind of soft regularization, therefore, adding a specific loss term dedicated to penalizing the variations with respect to a given explanation map (like $\mathcal{L}_{Exp}$ in Eq. 1). Usually, this term comes with a weight-fixed trading-off (α), making it very challenging to tune. In contrast, ECGL puts the explanations in a promise set within the Augmented Lagrangian formalism. Hence, it has more flexibility and is adaptable. Deciding whether an explanation constraint should be satisfied or not becomes dynamic instead of using a fixed hyperparameter, and it is handled by the Lagrange multipliers λ and the penalty term μ. In this manner, ECGL is much more suitable to have many conflicting constraints inserted and to adapt these constraints during the training phase.

Proposed method

We assume that the reader is familiar with the base machine learning and explainability techniques and start defining the problem of Explanation Guided Learning (EGL). Intuitively, EGL integrates human explanation annotations into the learning models in order to align human and model explanations. The problem has been formalized in [2] as follows:

Definition 1 (EGL) Let f be a differentiable model with the parameterized θ that learns to fit inputs $X \in \mathbb{R}^{N \times D}$ and the corresponding one-hot class labels $Y \in \mathbb{R}^{N \times K}$, where N denotes the total number of data samples, D denotes the input dimension, and K denotes the number of classes. An explainer g is chosen to extract the explanation M from the model f given its parameters θ and a set of data points $\langle X, Y \rangle$. The objective function of the EGL problem is:

$$\min \underbrace{\mathcal{L}_{Pred}(f(X),Y)}_{\text{TaskSupervision}} + \underbrace{\alpha\mathcal{L}_{Exp}(g(f,\langle X,Y\rangle),\hat{M})}_{\text{ExplanationSupervision}} + \underbrace{\beta\Omega(g(f,\langle X,Y\rangle))}_{\text{ExplanationRegulization}} \quad (1)$$

where $\hat{M}$ incorporates the right explanation provided by human annotations α and β are regulator parameters.

In practice, the explainer function g is usually realized by the existing differentiable XAI explainers like GradCAM [13]. In task supervision, a loss such as cross-entropy is used, while in explanation supervision and explanation regularization, some general properties, such as maintaining the sparsity of the explanation, are forced.

Let's present Explanation Constraints Guided Learning (ECGL). The idea is to see the same problem of EGL under different perspectives: define it as the combination of the task supervision problem with the so-called *explanation constraints*, able to design specific criteria such as background knowledge satisfaction, fairness requirements, sparsity, or relevance that the explanations should meet.

Definition 2 (Explanation Constraint) An explanation constraint is a specific constraint of the form:

$$g_i(\theta) \leq 0, \quad i \in \{1,\ldots,m\} \quad (2)$$

where m is the number of conditions and g_i represents a specific i-th condition on the generated explanations.

We propose to incorporate the explanation constraints into the learning process. This integration introduces high computational complexity, especially when constraints are incorporated into the model during each iteration. To handle these tasks efficiently, we adopt the rewriting used in the Augmented Lagrangian method. Following it, the constrained optimization problem is relaxed into an unconstrained problem whose objective function (called *the Augmented Lagrangian function*) has two additional components that take into account the satisfaction of constraints, whose importance is determined by the additional variables called Lagrangian multipliers. The formal statement is as follows:

Definition 3 (Augmented Lagrangian [45]) Given a constrained optimization problem:

$$\text{Minimize } L(\theta) \quad \text{subject to } g_i(\theta) \leq 0, \text{for } i \in \{1,\ldots,m\} \quad (3)$$

where $L(\theta)$ is an objective function and $g_i(\theta)$ with $i \in [1..m]$ are constrains, the Augmented Lagrangian function $\mathcal{L}_A$ is formulated as follows:

$$\mathcal{L}_A(\theta,\lambda,\mu) = L(\theta) + \sum_{i=1}^{m} \lambda_i g_i(\theta) + \frac{\mu}{2}\sum_{i=1}^{m} \max(0, g_i(\theta))^2 \quad (4)$$

where θ represents the parameters or decision variables, λ_is are the Lagrange multipliers associated with each constraint $g_i(\theta)$, μ is a positive penalty parameter that intensifies the penalty for violated constraints as it increases, and $\max(0, g_i(\theta))^2$ represents the penalty terms for each constraint, squared and applied only when constraints are not met (i.e., when $g_i(\theta) > 0$).

Such a method provides a robust framework for solving constrained optimization problems. The very premise of these methods would make one combine the Lagrangian with a quadratic penalty on constraint violations. If we think of convex optimization, where objective functions and the constraints g_i are convex, convergence is assured to the global optimum by these methods. The penalty parameter $\mu \to \infty$ will, in fact, cause the solution to the unconstrained problem in Eq. 4 to converge to the solution of the original constrained problem [46]. The Lagrange multipliers λ_i would then correspond to dual variables, estimating the shadow price of each constraint. In the case of non-convex constraints (non-convex functions g_i), the augmented Lagrangian remains highly effective, although some degree of global convergence failure is present. For this reason, the quadratic penalty term $\max(0, g_i(\theta))^2$ smooths the landscape to better accommodate gradient optimization methods. Iterative updates to λ and μ then steer the search towards a local minimum that satisfies the constraints. And the term $\max(0, g_i(\theta))^2$ is a continuously differentiable penalty, which is critical for the stability of adaptive optimizers, including Adam. The rule for updating the Lagrange multipliers, $\lambda_i \leftarrow \lambda_i + \mu.\max(0, g_i(\theta))$, denotes dual ascent in this case. This stage modifies the price of constraint violations according to the current state of violation. The adaptive updating of the penalty parameter μ is important in making sure that constraints become more critical later on in the optimization process to avoid the algorithm being trapped at points optimal for the primary loss but which violate the constraints.

Algorithm 1 ECGL algorithm

1: **Input**: Training data (X, Y), explanation requirements $\mathcal{R}$, initial model parameters θ
2: **Initialize**: Lagrange multipliers $\lambda_i = 0$ for all i, penalty parameter $\mu > 0$, maximum iterations N, learning rate η
3: **Define**: Loss function $L(\theta)$, domain-specific knowledge $\mathcal{K}$
▷ Dynamic Constraint Formulation
4: Define initial explanation constraints $g_i(\theta)$ based on $\mathcal{R}$ and $\mathcal{K}$, $g_i = g_i(\theta)$
▷ Begin Optimization Loop
5: **for** $k = 1$ to N **do**
▷ Augmented Lagrangian Construction
6: $\mathcal{L}_A(\theta, \lambda, \mu) \leftarrow L(\theta) + \delta\|\theta\|_2^2 + \sum_{i=1}^{m} \left(\lambda_i g_i + \frac{\mu}{2} \max(0, g_i)^2\right)$
▷ Optimization Step
7: $\theta \leftarrow \text{Optimizer}(\nabla_\theta \mathcal{L}_A, \theta, \eta)$
▷ Lagrange Multipliers Update
8: **for** $i = 1$ to m **do**
9: $\lambda_i \leftarrow \lambda_i + \mu \max(0, g_i(\theta))$
10: **end for**
▷ Adaptive Penalty Parameter Update
11: $\mu \leftarrow \text{update\_mu}(\mu, \theta)$
▷ Dynamic Adjustment of Constraints
12: $g_i \leftarrow \text{adjust\_constraints}(g_i(\theta), \theta, \lambda_i)$
13: **end for**
14: **Return:** optimized model parameters θ

The proposed ECGL method is presented as Algorithm 1. It takes in input the training data (X, Y), explanation requirements $\mathcal{R}$ (if they exist), and initial model parameters θ, initializes the Lagrange multipliers $\lambda_i = 0$ for all i, the penalty parameter $\mu > 0$, the number of maximum iterations N, and the learning rate η (lines 1–2).

In the system setup, we define the loss (e.g., cross-entropy loss for classification) and the prior knowledge $\mathcal{K}$ that guides the setting of explanation constraints (line 3). The initial explanation constraints are defined on line 4 and are updated during the training process on line 12. Explanation requirements $\mathcal{R}$ could be considered as high-level qualitative goals. For instance, the request is that the *predictions should not be in favor of one gender over another*.

The algorithm iteratively updates the parameters of the model, the Lagrange multipliers, the penalty parameters, and the explanation constraints on lines 5–13. Line 6 defines the Augmented Lagrangian $\mathcal{L}_A$ based on the total loss function $L(\theta)$ and the explanation constraints. The total loss function $L(\theta)$ can be realized as $L(\theta) = \beta_1 L_{\text{accuracy}}(\theta) + \beta_2 L_{\text{fairness}}(\theta)$ $\beta_3 L_{\text{interpretability}}(\theta)$, where $\beta_1, \beta_2, \beta_3$ weights are balancing the importance of accuracy, fairness, and interpretability. Observe that in our formulation, the Augmented Lagrangian $\mathcal{L}_A$ also has an $l2$-regularization term $\delta\|\theta\|_2^2$ to enhance robustness against noisy data.

The optimization step is performed on line 7. We utilized *the Adam optimizer* in our experiments. The Lagrange multipliers are updated on lines 8–10. The adaptive penalty parameter μ is updated in case the constraints are not sufficiently satisfied on line 11. In our experiments, we realize the $update\_mu$ function as follows:

$$\begin{cases} \mu \cdot \zeta & \text{if } \sum_{i=1}^{m} \max(0, g_i(\theta))^2 > \epsilon \\ \mu & \text{otherwise} \end{cases} \tag{5}$$

where there $\zeta > 1$ is an increase factor and ϵ is a threshold for constraint violation tolerance. The $adjust\_constraints$ function on line 12 can be realized as $g_i(\theta) + \gamma \cdot \nabla_\theta \mathcal{L}_A$, where γ is a learning rate specific to the constraint adjustment. Thus, the explanation constraints are updated, taking into account the decision boundaries of the model.

At the end, the optimized model parameters are returned on line 14.

Experiments

To evaluate our proposal, we considered three well-known and publicly available healthcare datasets containing data on diabetes patients, pneumonia, and NIH Chest X-ray images, respectively. We have made the evaluation codes available.[1]

We conducted a comprehensive evaluation using widely adopted fairness metrics and explainability techniques to evaluate the efficacy of ECGL in enhancing the overall fairness and explainability of machine learning models. We evaluated the efficacy of our proposed ECGL in (i) mitigating biases that may lead to discriminatory outcomes and (ii) enhancing transparency into the model's decision-making process.

Evaluation

We utilized two prominent techniques to interpret and explain the predictions of models: SHAP [47] and GradCAM [48]. We employed *SHAP (SHapley Additive exPlanations) values* to interpret the importance of features in the models' predictions (first experiment with the diabetes dataset). SHAP values provide a measure of how much each feature contributes to the difference between a prediction and the average prediction of the model. Therefore, SHAP values can reveal the most influential features in a model's prediction, feature interaction, and global and local explanations.

For the second experiment with the pneumonia X-ray image dataset, we used *GradCAM (Gradient-weighted Class Activation Mapping)* to visualize the regions of the images that the model focused on when making predictions. GradCAM generates heatmaps that highlight the areas of the image that contribute most to the model's classification. These heatmaps offer insights into how the models are making decisions, potentially revealing biases or limitations. By analyzing the regions highlighted by the heatmaps, we can identify which features (e.g., lung nodules, anatomical structures) are most important for the models' predictions.

The utilized metrics on fairness are presented in [33, 34, 49] and reported below. The insights derived from fairness metrics can effectively guide the model development process toward mitigating biases. In our experiments, dealing with medical datasets, we considered the *Equalized Odds Difference (EOD) and Ratio (EOR)* as the fairness metric. This metric evaluates the difference and ratio of the true positive and false positive rates between protected groups, respectively. A difference value of 0 and a ratio closer to 1 suggest that the model is equally likely to classify individuals from different groups correctly or incorrectly. EOD and EOR focus on the fairness of the model's predictions rather than the overall outcomes, ensuring that the conditional probability of a positive outcome given a positive or negative prediction is equal across different demographic groups. Moreover, equalized odds does not enforce the same positive prediction rates across different groups, allowing it to better reflect differences in base rates of outcomes. This is particularly important in domains where certain conditions may be more prevalent in specific demographic groups.

Additionally, we have utilized common metrics, including *Area Under Curve (AUC)* and *Accuracy* to measure the model's performance. Overall, higher AUC and Accuracy values indicate a more accurate model, while lower EOD and higher EOR represent a fairer model.

Experimental settings

To evaluate our proposal, we considered three well-known and publicly available healthcare datasets containing data on diabetes patients and pneumonia and NIH Chest X-ray images, respectively. The experiments on diabetes and pneumonia datasets have been conducted on a system with 2 Intel Xeon virtual CPUs and 12 GB of RAM, using the Python programming language and TensorFlow library. The experiment on the NIH Chest X-ray images was conducted using 2 T4 GPUs, 30 GB of RAM, and 16 GB of GPU RAM. "Fairlearn" and "SHAP" libraries have been used for fairness and explainability evaluation purposes, respectively. The GradCAM visualization has been implemented by the authors using the "Keras" library. We have also used the "Scikit-learn" library for data preprocessing and standardization.

Experiment I: diabetes dataset

The diabetes dataset[2] is a widely used benchmark in the healthcare domain. This dataset comprises 768 instances with eight attributes, including gender, body mass index (BMI), blood pressure, skin thickness, insulin level, glucose level, and diabetes pedigree function. The target variable is a binary class indicating the presence or absence of diabetes. We evaluated the performance of the proposed model across a range of different metrics, including EOD, EOR, Accuracy, and AUC.

Model selection

The model architecture of the examined diabetes deep learning network is illustrated in Fig. 1. We utilized a sequential neural network, consisting of three dense layers with decreasing units ($128 \rightarrow 64 \rightarrow 32$), interspersed with dropout layers to prevent overfitting. The final layer has two output units with *a softmax* activation function, producing probabilities for the two classes, referring to the positive or negative case of diabetes.

[1] https://github.com/ShahbazianR/AL-EGL

[2] available at: https://data.mendeley.com/datasets/wj9rwkp9c2/1

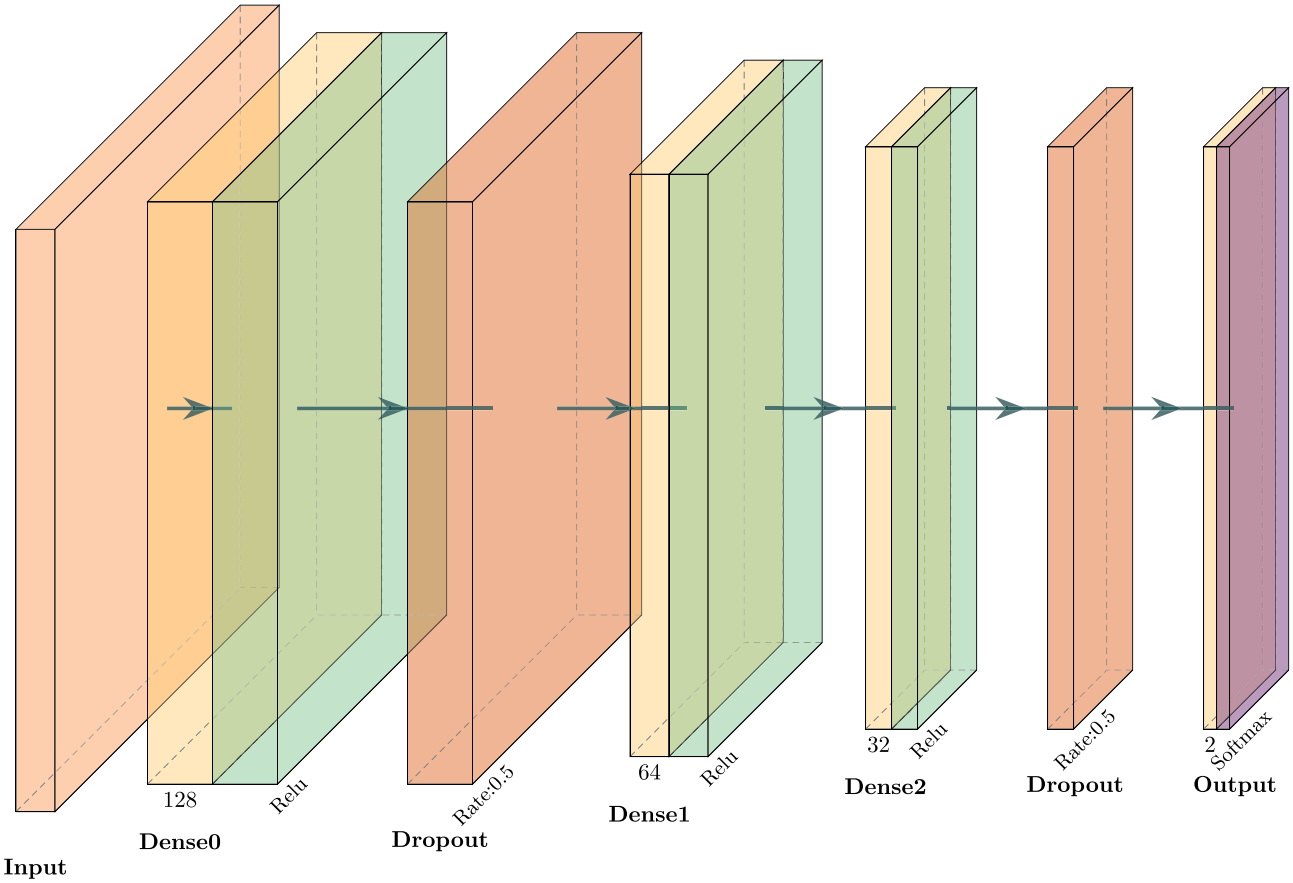


Fig. 1 The architecture of the neural network was utilized for the experiment on the diabetes dataset

Table 1 Specific conditions used in experiment I. LB and HB refer to lower and higher bounds of feature values, respectively. *P* is the empirical probability of diabetes for the specified range. Convexity indicates whether the constraint region defined by LB and HB is convex

	Positive			Negative			Convexity
Feature	**LB**	**HB**	**P**	**LB**	**HB**	**P**	
Pregnancies	0	5	0.2572	0	5	0.7428	Convex
Glucose	170	–	0.8406	50	75	1.0000	Convex
Glucose (cont.)	–	–	–	30	100	0.9271	Convex
Blood Pressure	100	–	0.6154	–	–	–	Convex
Skin Thickness	60	–	1.0000	20	40	0.6311	Convex
Insulin	200	–	0.5412	0	200	0.7279	Convex
BMI	50	60	0.7143	0	20	1.0000	Convex
DPF	–	–	–	0	0.25	0.7463	Convex

As for the training, we used the sparse categorical cross-entropy loss function and the "Adam" optimizer. The model has been trained considering 100 epochs in total, with early stopping applied to avoid the potential of overfitting.

Explanation constraints

To effectively define explanation constraints in our model, we conducted a comprehensive analysis of the dataset and leveraged expert domain knowledge. This analysis allowed us to identify the relevant ranges and the correlations of data features with the likelihood of a positive diabetes diagnosis. We identified the potential probability of positive and negative diabetes cases corresponding to the specified feature ranges reported in Table 1. We defined the corresponding constraints and used them to refine and guide the model's predictions towards classification labels aligning with the determined probabilities. Noteworthy, we consider the normalized values of these conditions to better align with loss values while training the model.

Table 2 Performance of ECGL. *base model* is the model without any explanation constraint. *X model* incorporates explanation constraints for *X*, where $X \in$ {Pregnancies, glucose, blood pressure, skin thickness, Insulin, BMI, DPF, and all Constraints}. Bold represents the best in each column

Model	Accuracy	EOD	EOR	AUC
Proposed ECGL				
Pregnancies	0.7597	**0.1931**	0.3428	**0.7646**
Glucose	0.7532	0.3412	0.3117	0.7232
Blood Pressure	0.7597	0.3888	0.2222	0.7101
Skin Thickness	0.7662	0.3029	**0.3636**	0.7454
Insulin	0.7727	0.3029	0.2380	0.7505
BMI	0.7727	0.3029	0.3117	0.7505
DPF	0.7597	0.3386	0.2449	0.7444
All Constraints	**0.7857**	0.3399	0.3174	0.7566
Base Model	0.7468	0.3399	0.3265	0.7162

In particular, each condition in Table 1 specifies a corresponding constraint violation function $g(\theta)$, which evaluates the inconsistency between the model predictions and expert probabilities under a given range of a feature. For a sample whose features lay in a specified range (say Glucose > 170), this constraint violation is assessed concerning the probabilities output by the model. For instance, should domain knowledge indicate a high degree (e.g., 0.84) of a positive diagnosis for the high glucose class, then this constraint term penalizes the model when it predicts low probabilities for this class. The violation itself is expressed as a smooth and differentiable function of model outputs, which measures the model's failure to conform with the expert knowledge. Such a formulation easily lends itself to gradient-based optimization in the Augmented Lagrangian framework.

The set of constraints in Experiment I is all defined over closed intervals or threshold-based conditions for individual features; hence, they define convex feasible regions in the input space. In Table 1, the classification of each condition with respect to convexity is presented. This makes it compatible with convex constraint optimization techniques; however, our framework remains applicable even in non-convex scenarios because of its iterative penalty-based scheme.

Results and discussion

The *Base model* was trained to optimize the output accuracy without any explanation constraints. Table 2 compares the *base* network with *ECGL* variants that impose explanation constraints on individual features or on all features simultaneously. The All-constraints model achieves the highest accuracy (0.7857, +**3.89**% over *Base Model*) but does not improve fairness (EOD = 0.3399, identical to the *Base model*; EOR = 0.3174 vs. 0.3265). In contrast, several single-feature variants provide meaningful fairness gains: the Pregnancies-only model reduces EOD by 43% (0.1931) and raises AUC by 4.84% (0.7646). Overall, imposing carefully chosen explanation constraints can lower bias while preserving—or slightly improving—predictive performance; combining all constraints maximizes accuracy at the cost of fairness.

Explainability analysis by using the SHAP values is presented in Table 3. As shown in Table 3, the feature importance values generally experience an increase when

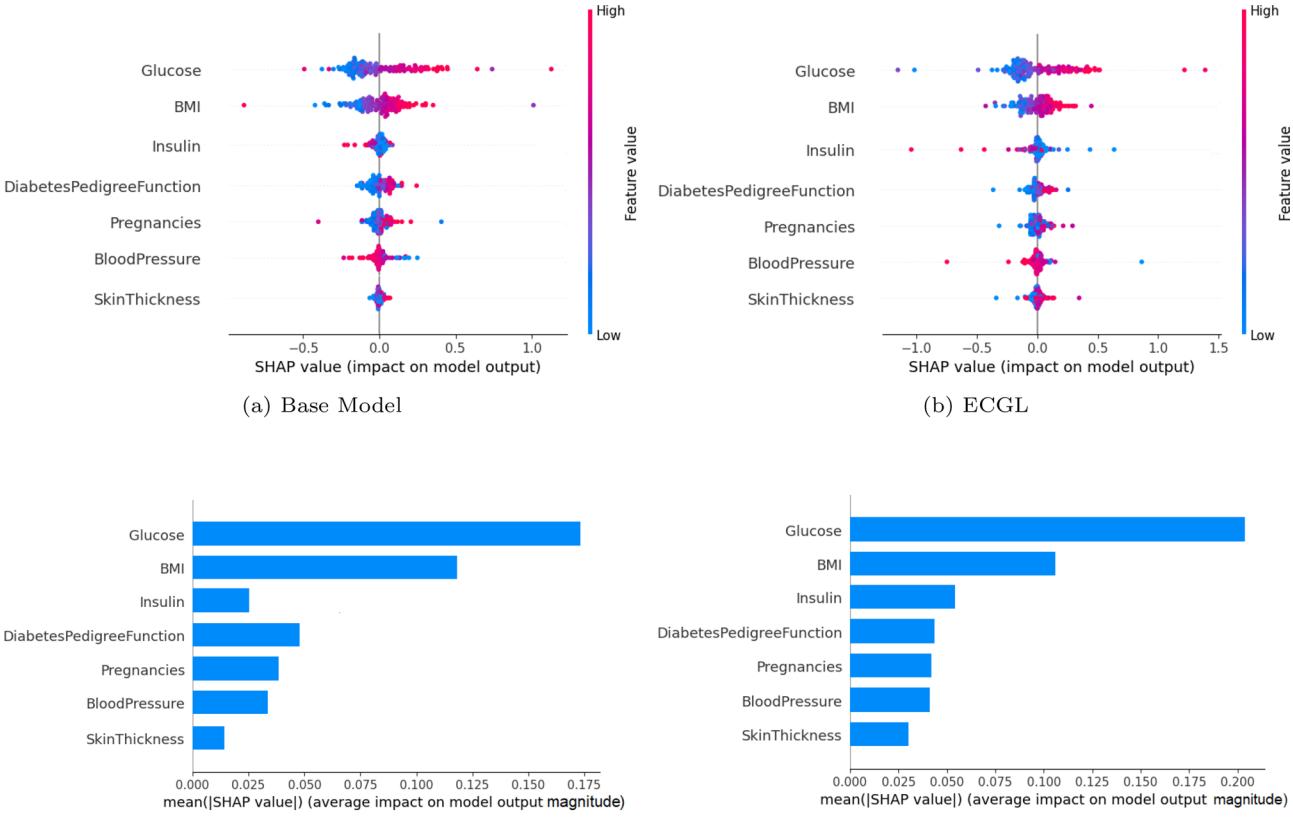
Table 3 Comparing the importance of SHAP values obtained by applying the constraint-based ECGL models with the base Model, and the corresponding improvement percentages

Model	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF
ECGL Models - SHAP Values							
Pregnancies	0.0823	0.2201	0.0507	0.0620	0.0580	0.1217	0.0665
Glucose	0.0324	0.2009	0.0329	0.0335	0.0604	0.0873	0.0496
Blood Pressure	0.0303	0.1725	0.0312	0.0277	0.0377	0.1093	0.0450
Skin Thickness	0.0351	0.1702	0.0249	0.0179	0.0295	0.0952	0.0430
Insulin	0.0269	0.1713	0.0408	0.0122	0.0274	0.1068	0.0400
BMI	0.0837	0.3412	0.0654	0.0297	0.0740	0.2333	0.1605
DPF	0.0377	0.2055	0.0529	0.0199	0.1193	0.1138	0.0691
All Constraints	0.0421	0.2037	0.0410	0.0303	0.0543	0.1060	0.0437
Base Model	0.0384	0.1733	0.0338	0.0142	0.0253	0.1182	0.0479
Percentage Improvements Compared to Base Model (%)							
Pregnancies	+114.3%	+27.1%	+50.0%	+336.6%	+129.2%	+2.9%	+38.9%
Glucose	−15.6%	+15.9%	−2.7%	+135.9%	+138.3%	−26.1%	+3.5%
Blood Pressure	−21.1%	−0.5%	−7.7%	+95.1%	+49.0%	−7.5%	−6.0%
Skin Thickness	−8.6%	−1.8%	−26.3%	+26.1%	+16.6%	−19.4%	−10.2%
Insulin	−29.9%	−1.1%	+20.7%	−14.1%	+8.3%	−9.7%	−16.5%
BMI	+118.0%	+96.8%	+93.5%	+109.2%	+192.5%	+97.4%	+234.9%
DPF	−1.9%	+18.6%	+56.4%	+40.1%	+371.9%	−3.7%	+44.3%
All Constraints	+9.6%	+17.5%	+21.3%	+113.4%	+114.8%	−10.3%	−8.8%

the associated explanation constraints are applied during model training. For instance, integrating the constraint associated with *Pregnancies* led to an approximate 9.6% rise in the importance of this feature in the model's final prediction. Similarly, several features, such as *Skin Thickness* and *Insulin*, exhibited substantial increases in their SHAP importance values. However, a few features, such as *BMI* and *DPF*, showed a slight decrease. Moreover, by combining all constraints, the *All Constraints* model achieves an overall improvement in feature importance alignment compared to the *Base Model*, supporting the effectiveness of ECGL in steering the model towards clinically relevant explanations.

SHAP values can also help identify feature interactions, which may be crucial for understanding the model's behavior. This means that the constraints, including one feature, may also affect how the model adheres to another one. As an example, applying constraints associated with *BMI* indicates massive improvements across almost all constraints (+93.5% to +234.9%). This suggests that *BMI* could be considered a highly sensitive parameter to constraints with a critical role in predictions. The *Insulin* under the *Diabetes Pedigree Function (DPF)* constraint shows the highest value in the table. *BMI* and *Insulin* might be seen as the most "constraint-sensitive" features, since the gains are larger compared to other constraints.

Figure 2 presents SHAP summary plots for the *Base model* and the *proposed ECGL (All Constraint model)*, which provide a visual representation of the feature's contribution to the prediction. The color and size of the dots in SHAP summary plots represent the feature's contribution, while their position (being positive or negative) indicates whether the feature increases or decreases the prediction outcome. In Fig. 2(a) and 2(b), the *Base model* highlights *Glucose* and *BMI* as the most influential features, followed by *DPF*. Features like *Pregnancies, Blood Pressure, Insulin,* and *Skin Thickness* show relatively lower importance. However, the *All Constraint model* demonstrates a shift in feature importance. While *Glucose* and *BMI* remain the most significant features, the importance of *Insulin* has increased significantly due to the integration of explanation constraints. For instance, the *Base model* assigns an importance value close to 0 or negative for *Insulin*, whereas the *All Constraint model* shows a larger range with more positive and extreme values. Similarly, the features *Pregnancies, Blood Pressure,* and *Skin Thickness* exhibit increased importance, indicating that the constraints effectively guide the model to focus on these relevant features.



(a) Base Model (b) ECGL

(c) Base Model (d) ECGL

Fig. 2 SHAP value evaluation of the base model and ECGL (constraint model). (**a**) base Model. (**b**) ECGL. (**c**) base Model. (**d**) ECGL

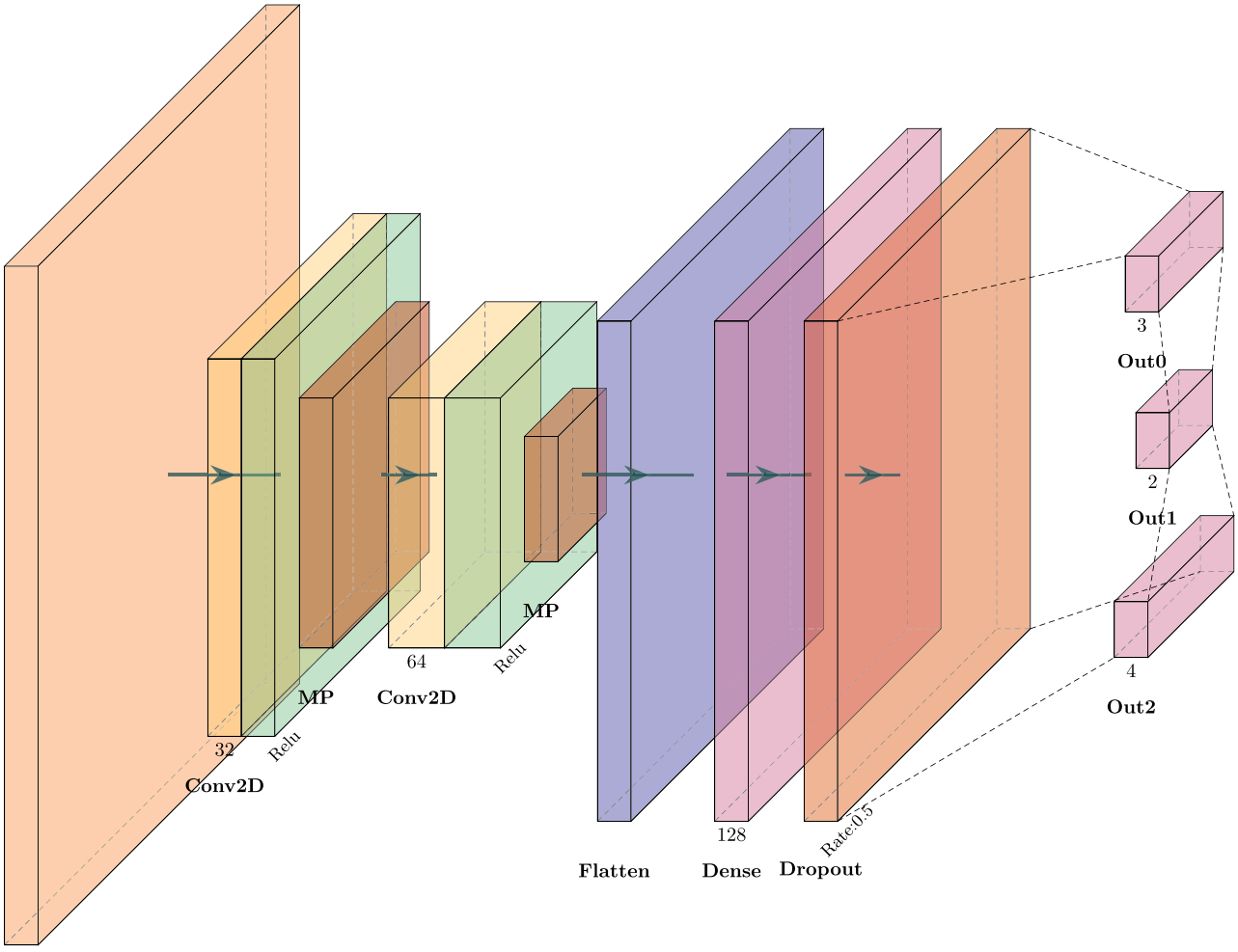


Fig. 3 Architecture of the model to evaluate the performance of the proposed *ECGL* and *base model* on pneumonia X-ray image classification

Figures 2(c) and 2(d) illustrate the mean absolute SHAP values for different features, highlighting their relative importance in predicting the target variable for the *Base model* and the *All Constraint model*. These plots show that applying constraints has slightly changed the overall feature importance. While *Glucose* remains the most influential feature, the importance of other features has increased, indicating that the constraints have guided the model to focus more on these features. For example, the importance value of *Glucose* increased from 0.175 in the *Base model* to 0.20 in the *All Constraint model*. Additionally, the constraints had a significant effect on features such as *Insulin, Pregnancies,* and *Skin Thickness*, which were not utilized properly in the *Base model*. This means that the constraints enhanced the model's learning process, allowing it to better capture the relevance of these features, resulting in improved predictions and interpretations.

Experiment II: pneumonia X-ray images

In this section, we explore the fairness and explainability performance of the proposed ECGL on the pneumonia classification task, using an X-ray image dataset.[3] This dataset contains 30,000 frontal-view chest radiographs from the 112,000-image public National Institutes of Health (NIH). We employed GradCAM to gain insights into the model's decision-making process, enabling us to understand the rationale behind its predictions and identify the improvements gained by applying explanation-guided learning on the model.

Model selection

The overall architecture of the model for *Experiment II* is presented in Fig. 3. Considering the image data, we utilized 2D convolutional layers with *the ReLU* activation

[3] Available at: https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018

function, accompanied by two max pooling layers with a kernel size of 2. The model employs three output layers with a variety of sizes to predict different features. The primary task output is responsible for predicting the main classification labels, meaning pneumonia classification. The model also yields two more outputs, used to integrate the gender-based and bounding box-based explanations into the model and to guide the model towards predictions more aligned with these constraints. Moreover, the bounding box-based explanation layer predicts the information that can be used to explain the model's focus on specific regions of the input X-ray image, which increased the model's transparency regarding the decision-making process and the explainability of the final predictions.

Each sample was converted to a 3-channel RGB image with a size of 128 by stacking the original image three times along the last axis while ensuring the image remains clear with anti_aliasing enabled. Then, the data were split into training and test data, considering an 80%-20% ratio for training and validation, respectively. The *Base model* was trained without any explanation constraints, while for the *ECGL model,* the explanation constraints were defined and taken into account. Both ECGL and base models were trained for 30 epochs, using 2000 X-ray image samples of the pneumonia dataset, using the Adam optimizer, and sparse categorical cross-entropy and mean absolute error as loss functions for label and bounding-box classification and gender prediction, respectively.

Explanation constraints

To integrate fairness and explainability, we incorporated two primary constraints, including gender-based and bounding box-based constraints. By imposinga gender-based explanation, we aimed to mitigate potential gender bias in our model's predictions. These explanation constraints are designed to ensure that the model's outputs are not systematically biased towards or against specific genders. Moreover, the bounding box-based explanation aims to enhance the explainability of our model's predictions. These explanation constraints are intended to guide the model to focus on relevant regions of the input data, providing insights into the factors that influence its decisions.

Table 4 Comparing the accuracy and fairness of the base model and proposed ECGL

Model	Accuracy	EOD	EOR	AUC
Base Model	0.8700	**0.01428**	0.51412	**0.88833**
ECGL Model	0.8700	0.04761	**0.64835**	0.84666

Results and discussion

Table 4 presents a comparative analysis of the *Base* and *ECGL models* in terms of fairness, accuracy, and AUC. The overall performance of the models reveals that both models demonstrate comparable levels of overall accuracy of 87%, suggesting that the constraints imposed on the *ECGL model* did not significantly compromise its predictive power. By contrast, the *Base model* could achieve a slightly higher AUC value, with a relatively small difference of about 4%, while the *ECGL model* exhibits relatively high performance in mitigating biases and disparities.

Evaluating the EOD and EOR values demonstrates slightly different results. While the *Base Model* achieves a 3% lower EOD compared to the *ECGL Model*, the *ECGL Model* exhibits a noticeable 13% improvement in EOR. As these metrics assess whether the model's predictions are equally distributed across different groups, the results suggest that the ECGL-enhanced model achieves greater uniformity and fairness in its predictive performance when guided by explanation constraints.

Figures 4 and 5 depict the GradCAM *Heatmaps* and *Superimposed X-ray images* of correctly and incorrectly predicted pneumonia instances, respectively. The areas in images that are highlighted in color represent the regions that the model deemed most important for its decision. Moreover, the overlay of the heatmaps on the original X-ray images provides valuable information, allowing us to assess whether the models are focusing on relevant features. The superimposed images refer to the original images combined with the heatmaps, providing a more comprehensible visual representation of the focused regions. The illustrated heatmaps depict three classification outputs corresponding to the labels 0, 1, and 2, arranged from top to bottom, respectively. Label 0 indicates "No Lung Opacity," which refers to patients without diagnosed pneumonia. The other labels, 1 and 2, on the other hand, signify *Lung Opacity* and the presence of fuzzy white clouds in the lungs, associated with pneumonia.

Likewise, Figs. 6 and 7 compare the proposed *ECGL model* and the *Base model* in terms of explainability by using GradCAM heatmaps. While both models generally seem to be focusing on similar areas (central region) of the chest X-rays, the color intensity in the heatmaps and superimposed images obtained from the *ECGL model* clearly shows a greater emphasis on certain regions that are related to the respiratory system and pneumonia disease. As it can be seen in Fig. 4(a) and 4(b) (correctly labeled instances), the highlighted regions of the *Base model* are more blurred, hazy, and spread out across the image, suggesting that the model's decisions are less interpretable and apparent to the experts. Whereas the proposed *ECGL model* presents more focused and

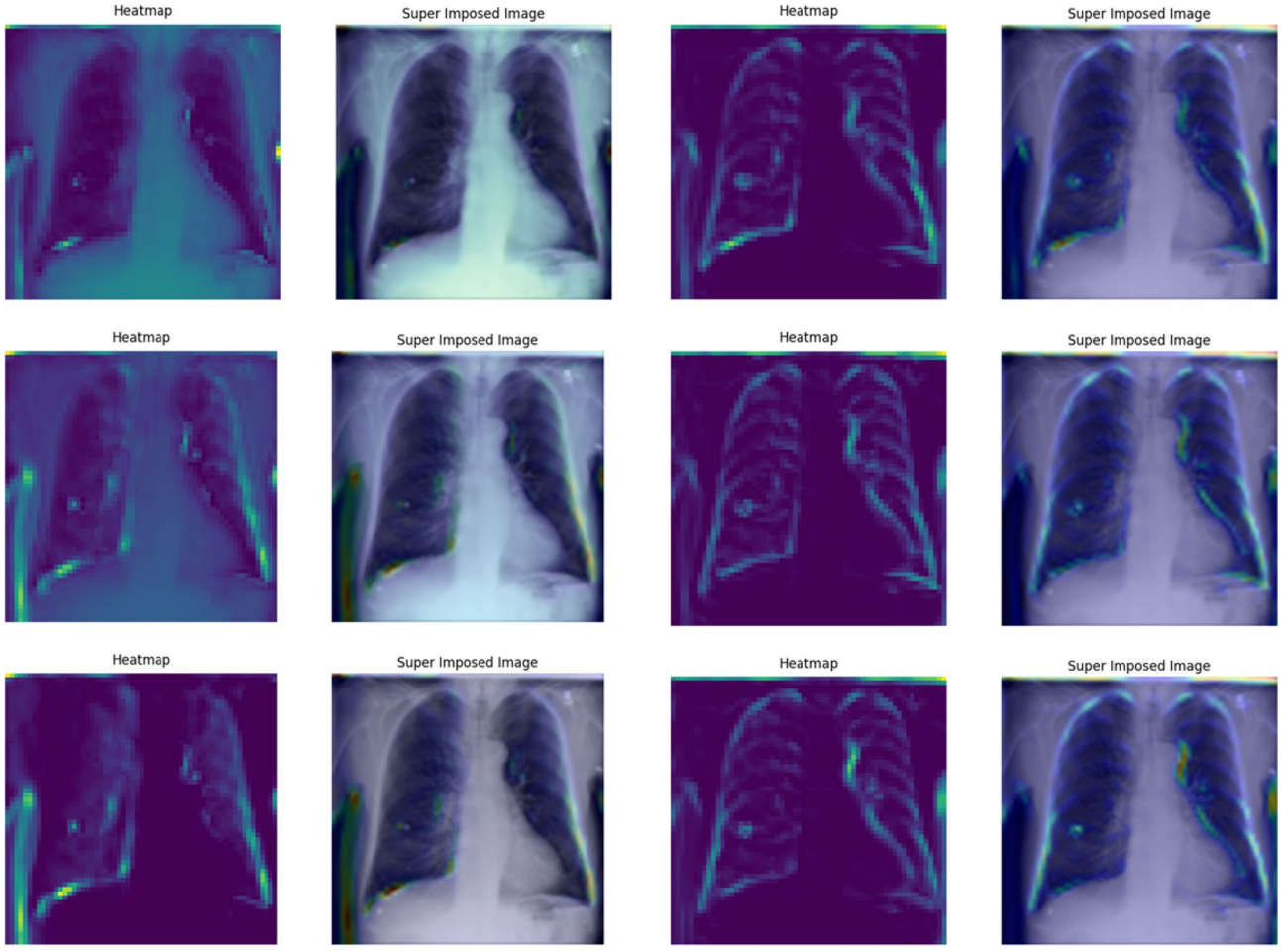


Fig. 4 GradCAM *Heatmaps* and *superimposed images* of the (**a**) *base model* and (**b**) *ECGL model* for correctly labeled instances. The illustrated heatmaps correspond to GradCAM visualizations for labels 0 (normal), 1 (pneumonia present), and 2 (pneumonia absent, abnormalities exist), arranged from top to bottom, respectively. (**a**) base Model. (**b**) ECGL model

localized highlighted regions, showing that the model is primarily relying on specific areas of the image, relating to the particular region attacked by the disease, to make its prediction. The *ECGL model* makes it easier to understand the model's reasoning and therefore presents more explainability. In addition, the *ECGL model* produces a more concentrated heatmap, indicating more confidence in prediction.

Moreover, in spite of both models' similar performance illustrated in Fig. 5(a) and 5(b), the GradCAM visualizations clearly demonstrate noticeable differences. The *Base model* resulted in a more scattered visualization that may not correspond to the actual pathology. This indicates that the *Base model's* prediction can be based on irrelevant features and not focused on the main required regions. The proposed*ECGL model*, on the other hand, highlighted regions that are more relevant to the considered medical condition, suggesting that the constraints have helped the model focus on more meaningful features.

Unlike the *Based model*, the *ECGL model's* heatmap for the incorrect prediction is more localized, highlighting a specific region in the lung. The overlay shows that there is indeed an abnormality in this area, but it could have been insufficient to confidently classify the image, which resulted in an incorrect final prediction. While such false predictions highlight the complexity of the task, the findings suggest that adding the constraints and guiding the model with domain-specific explanations through proposed ECGL can enhance the model's explainability even when overall accuracy remains quite the same. Using such explainability, the experts can better identify the model's deficiencies and limitations, leading to efficient mitigation of them.

A bounding box (BBOX) is a rectangular region that is commonly used to localize (highlight) the interest area. Figure 8 illustrates the BBOX of some pneumonia disease

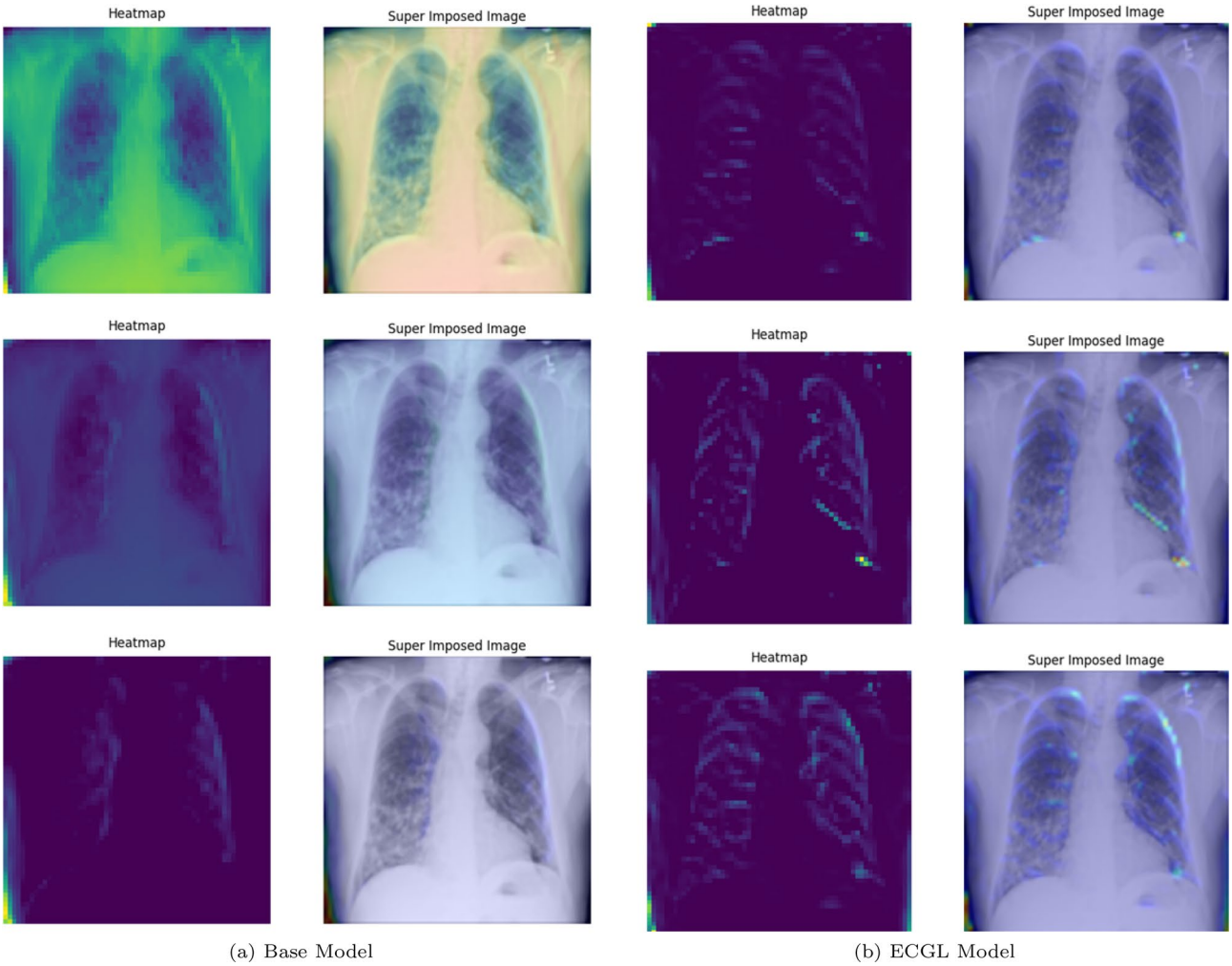


Fig. 5 GradCAM *Heatmaps* and *superimposed images* of the (**a**) *base model* and (**b**) *ECGL model* for incorrectly labeled instances. The illustrated heatmaps correspond to GradCAM visualizations for labels 0 (normal), 1 (pneumonia present), and 2 (pneumonia absent, abnormalities exist), arranged from top to bottom, respectively. (**a**) base Model. (**b**) ECGL model

instances predicted by the proposed ECGL model. The red box refers to the ground truth BBOX stated for the X-ray image, and the blue box demonstrates the predicted box by the *ECGL model*. It is important to note that the *Base mode*l cannot generate BBOX due to the lack of explanation constraints required for BBOX prediction.

Looking at Fig. 8(a) and 8(b), it is evident that the proposed ECGL correctly predicted the BBOX, aligning with the true box, even though the model estimated a larger area. This points out that considering the BBOX constraints in the model as integrated explanation constraints is effective in highlighting the attention to more relevant regions. Figure 8(c) presents a case where the predicted BBOX is partially within the lung area and may highlight the need for more refined BBOX constraints to better capture the spatial extent of pneumonia.

ECGL seems promising; however, it also has many limitations. First, we've only tried it on three datasets-one for diabetes numbers and two for chest x-ray images. As a result, we don't yet know how it will behave with data from other hospitals or patient groups. Second, the "explanation rules" that steer the model aren't produced automatically; they have to be written by clinicians using their know-how plus SHAP plots, and that can be slow and error-prone. Automated methods for discovering constraints could be the theme of future research towards improving scalability. For example, through multi-center data to infer constraints using techniques such as meta-learning, or through unsupervised approaches to detect consistent patterns in model explanations that can, in turn, be formalized as constraints. Third, we looked at fairness for just one binary trait (male vs. female) and one fairness metric (equalized odds). Real-world bias is messier than that. Fourth, the quality of ECGLs explanations

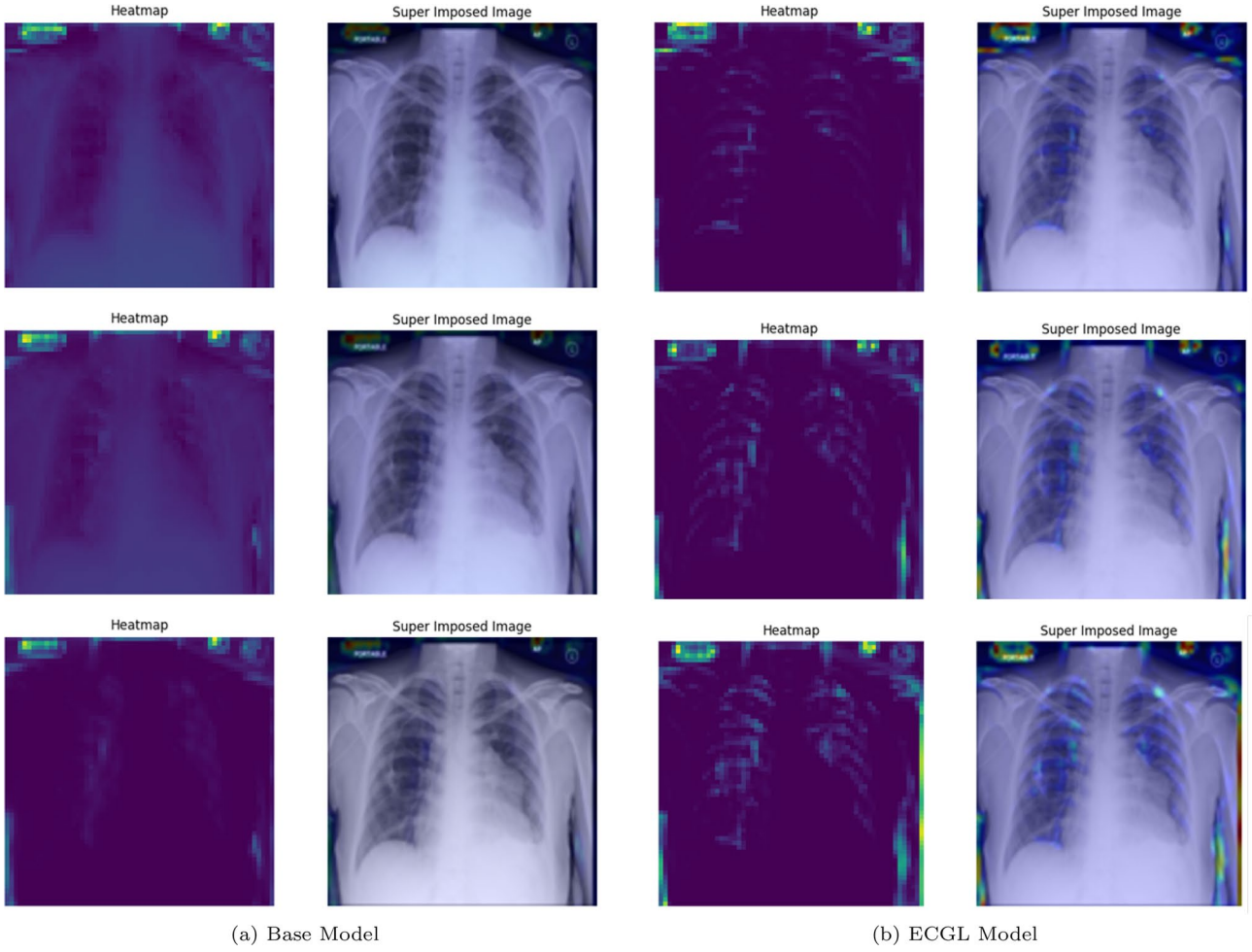


Fig. 6 GradCAM *Heatmaps* (left column) and *Superimposed images* (right column) of the (**a**) *Base model* and (**b**) *ECGL model* for correctly labeled instances. The illustrated heatmaps correspond to GradCAM visualizations for labels 0 (normal), 1 (pneumonia present), and 2 (pneumonia absent, abnormalities exist), arranged from top to bottom, respectively

is tied to Grad-CAM and SHAP; if those tools point in the wrong direction, the constraints can too. Finally, training with extra fairness-and-explanation losses takes more computational resources and a bit of extra hyperparameter tuning. All of these gaps give us a clear to-do list for the next round of work.

Experiment III: NIH chest X-ray images

We assess the performance of our proposed ECG model with an evaluation on the NIH Chest X-ray dataset.[4] This is a large-scale anonymized collection of 112,120 frontal views of chest X-ray images labeled via NLP techniques from radiology reports across 14 thoracic disease categories. From this dataset, we have created a balanced subset for a multi-label classification with at least 700 positive samples per class, where we obtain 8,307 training images, 4,973 validation images, and 7,156 test images. Each image is resized to 128×128 pixels and normalized to [0, 1]. Horizontal flips, brightness, and contrast perturbations are used for image augmentation in training. Gender is included as an auxiliary label to perform fairness analysis. In order to conduct a fairness assessment on demographic groups, especially the case with gender, we will present multiple group fairness metrics: Demographic Parity Difference (DPD), Demographic Parity Ratio (DPR), Equalized Odds Difference (EOD), Equalized Odds Ratio (EOR), and Statistical Parity Difference (SPD). These metrics assess the disparities concerning model decision outputs and are computed across all the disease labels in the multi-label setting.

Model selection

The architecture of the model has been given in Fig. 9. The architecture has a sequential arrangement of four convolutional blocks, each containing one 2D convolutional layer with ReLU activation, batch normalization,

[4] Available at: https://www.kaggle.com/datasets/nih-chest-xrays/data

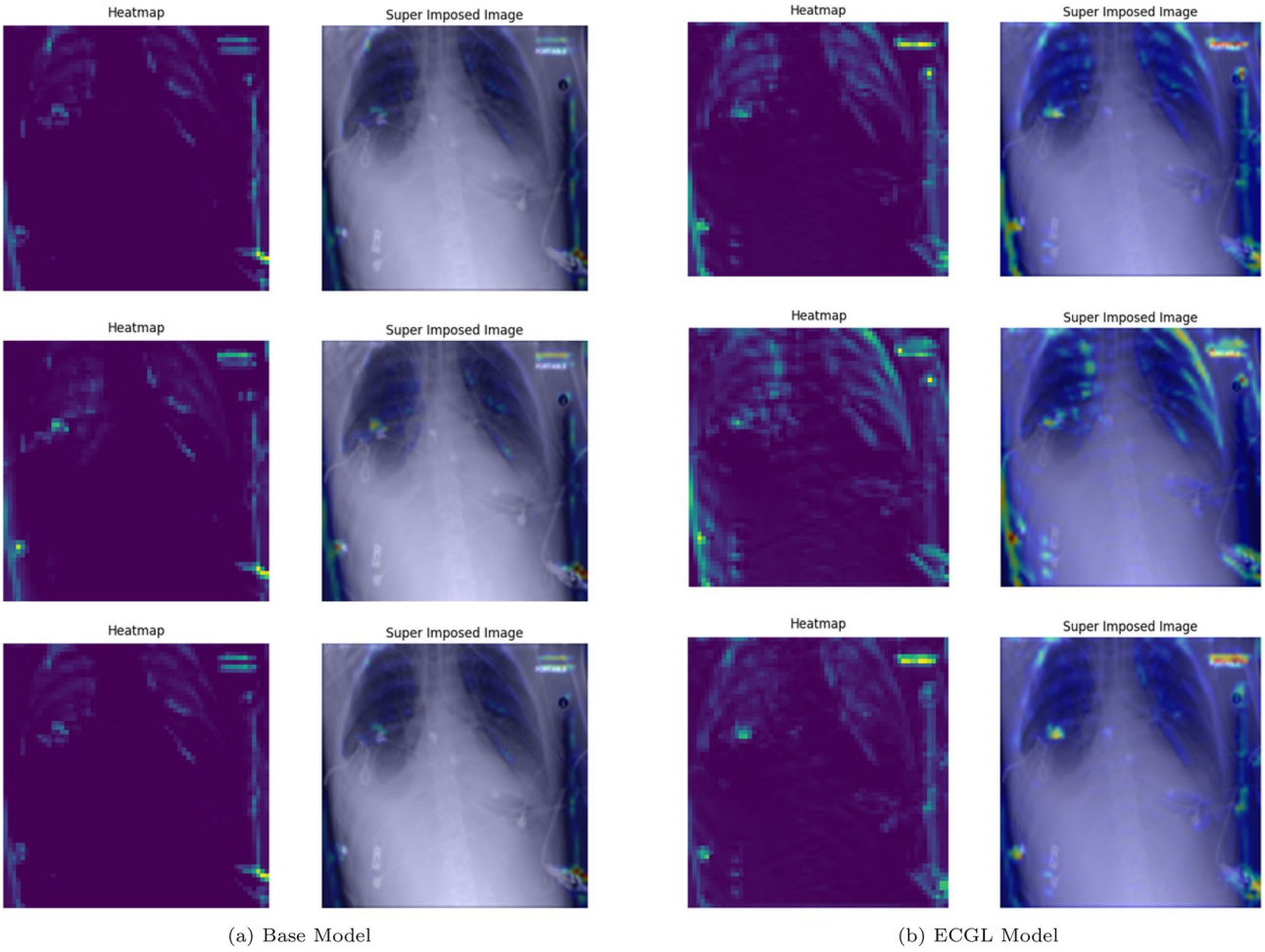


Fig. 7 GradCAM *Heatmaps* (left column) and *Superimposed images* (right column) of the (**a**) *Base model* and (**b**) *ECGL model* for incorrectly labeled instances. The illustrated heatmaps correspond to GradCAM visualizations for labels 0 (normal), 1 (pneumonia present), and 2 (pneumonia absent, abnormalities exist), arranged from top to bottom, respectively

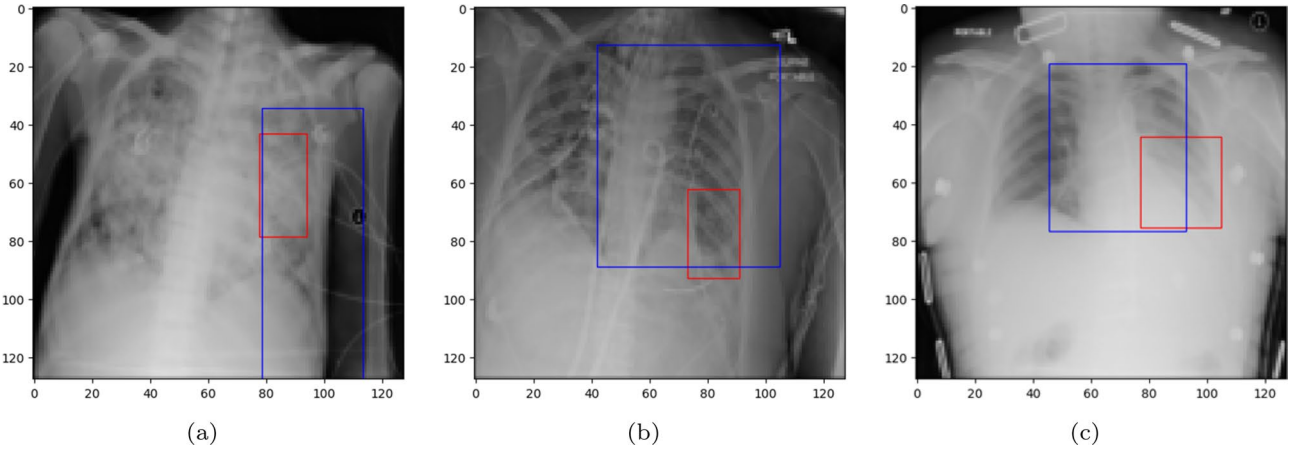


Fig. 8 Ground truth (red) and predicted (blue) BBOX regions in X-ray images: (**a**) and (**b**) represent the samples where the model predicts an incorrect label, while (**c**) corresponds to the case of the correct prediction

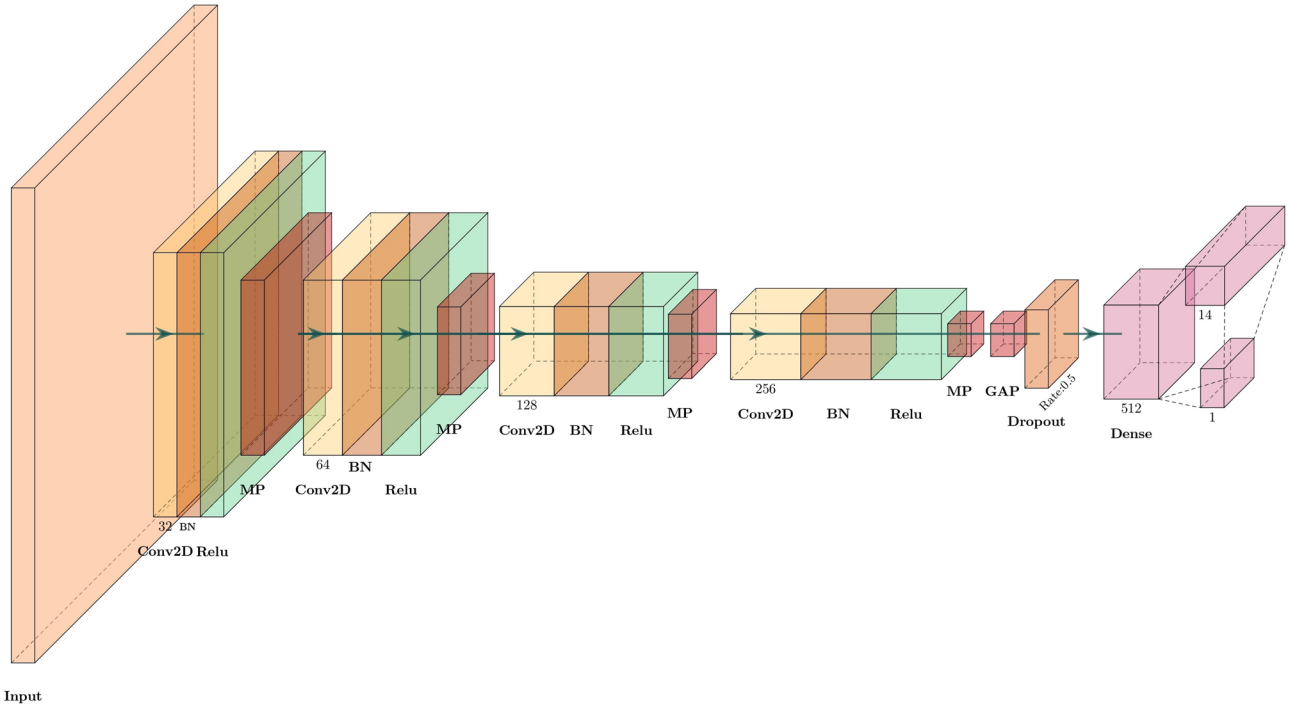


Fig. 9 Architecture of the model to evaluate the performance of the proposed *ECGL* and *base model* on NIH X-ray multi-label image classification

Table 5 Comparing the accuracy of the base model and proposed ECGL on NIH X-ray images

Model	Accuracy		AUC	
	Validation	Test	Micro	Macro
Base	0.8805	0.7130	0.7006	0.6202
ECGL	0.8805	0.7130	0.6737	**0.6208**

and max pooling. The number of filters in the end convolution outputs gradually increases in powers of 2, namely, 32, 64, 128, and 256. This allows the model to extract more and more abstract spatial characteristics of inputs from NIH chest X-ray images. The output generated by the last convolutional layer is passed through a global average pooling layer before going through dropout (at a dropout rate of 0.5) to deal with overfitting. This is then flattened and fully connected to a dense output line with 512 neurons and L2 regularization prior to a last output layer containing sigmoid-activated neurons, indicating the multi-label classification of diseases. This addresses fairness at this level by adding constraints of explanation based on gender, which causes the model to give similar feature attributions for both genders, reducing demographic bias within the learned representations and predictions. Finally, the model is trained with a batch size of 128 and for 20 epochs.

Explanation constraints

Similar to experiment II, we introduced a gender-based explanation bias as one of the constraints to reduce possible gender bias in model predictions. Such a constraint would ensure that the predictions produced by the model are not systematically affected by gender, promoting equal treatment with respect to demographic groups. By enforcing the same gender-based explanation consistency, we guide the model to align certain attribution patterns between genders, thereby lowering the likelihood of any biased feature attributions arising in its decision-making processes.

Results and discussions

Experimental results on the NIH Chest X-ray dataset are shown in Table 5, while Table 6 reports a detailed comparison between our proposed ECGL model and the baseline CNN model in predictive performance and fairness across demographic groups. With respect to classification performances, both models attain identical validation and testing accuracy (0.8805 and 0.7130, respectively). This means that including explanation constraints in ECGL does not harm classification accuracy. Nevertheless, AUC scores reveal that ECGL has a slightly lower micro-AUC (0.6737 vs. 0.7006) but a marginally better macro-AUC (0.6208 vs. 0.6202). This small increase in macro-AUC may indicate more equitable performance amongst disease classes, especially for rarer conditions, which is one of the more desirable qualities for multi-labels in medical diagnosis.

The major strength of the ECGL lies in the fairness metrics. For most disease types, the ECGL model shows

Table 6 Percentage improvement of ECGL model over base model across fairness metrics, on NIH X-ray images. Negative values of improvements indicate degradation

Disease Type	Base Model				ECGL Model				Improvement over Base Model (%)			
	DPD	DPR	EOD	EOR	SPD	DPD	DPR	EOD	EOR	SPD	DPD	DPR	EOD	EOR	SPD
Atelectasis	2.8e-5	0.99	0.07	0.85	−0.63	0.003	0.99	0.05	0.89	0.39	–	0.00	−28.57	4.71	−161.90
Cardiomegaly	0.06	0.76	0.11	0.74	−0.75	0.03	0.89	0.03	0.88	−0.75	−50.00	17.11	−72.73	18.92	0.00
Consolidation	0.02	0.93	0.15	0.73	0.29	0.009	0.96	0.11	0.78	0.28	−55.00	3.23	−26.67	6.85	−3.45
Edema	0.007	0.95	0.008	0.94	0.15	0.003	0.98	0.04	0.90	0.16	−57.14	3.16	400.00	−4.26	6.67
Effusion	0.08	0.86	0.09	0.82	−0.48	0.06	0.90	0.08	0.86	0.56	−25.00	4.65	−11.11	4.88	−216.67
Emphysema	0.02	0.90	0.02	0.91	0.23	0.06	0.86	0.05	0.86	0.42	200.00	−4.44	150.00	−5.49	82.61
Fibrosis	0.004	0.97	0.07	0.77	0.121	0.01	0.88	0.07	0.77	0.118	150.00	−9.28	0.00	0.00	−2.48
Hernia	0.002	0.93	0.06	0.00	0.02	0.007	0.41	0.007	0.39	0.009	250.00	−55.91	−88.33	–	−55.00
Infiltration	0.005	0.99	0.008	0.98	0.62	0.01	0.98	0.01	0.98	0.67	100.00	−1.01	25.00	0.00	8.06
Mass	0.03	0.92	0.02	0.92	0.34	0.01	0.96	0.01	0.97	−0.57	−66.67	4.35	−50.00	5.43	−267.65
Nodule	0.06	0.78	0.07	0.75	0.27	0.05	0.83	0.05	0.81	−0.73	−16.67	6.41	−28.57	8.00	−370.37
Pleural Thickening	0.02	0.94	0.11	0.80	−0.59	0.001	0.99	0.07	0.84	−0.66	−95.00	5.32	−36.36	5.00	−11.86
Pneumonia	0.01	0.86	0.01	0.85	0.06	0.03	0.84	0.03	0.84	0.17	200.00	−2.33	200.00	−1.18	183.33
Pneumothorax	0.03	0.87	0.06	0.88	0.27	0.01	0.96	0.06	0.85	0.23	−66.67	10.34	0.00	−3.41	−14.81

Table 7 Comparing the average computational cost of a single training epoch

	CPU		GPU
Model	**Diabetes**	**Pneumonia X-ray**	**NIH Chest X-ray**
Base	0.84 (s)	78.85 (s)	405.36 (s)
ECGL	3.40 (s)	84.64 (s)	409.82 (s)

a significant improvement in DPD, EOD, and SPD. For example, the ECGL reduced the EOD for *cardiomegaly* from 0.11 to 0.03 (73% improvement) and enhanced the EOR by nearly 19%. The same applies to *Pneumonia* and *Pleural Thickness*, where ECGL performs well in reducing demographic disparities across various fairness metrics. There are cases where classes have improved or degraded to extremes. For instance, Mass and Nodule have very high drops in SPD, meaning that there has been a large change in the predicted positive rates between gender groups. However, the sign flip in SPD from positive to negative in these classes can be interpreted to mean that fairness improved in one view but was perhaps overcompensated. For Hernia, while DPD and SPD improved, the EOR becomes undefined due to an original zero denominator in the base model, indicating instability in fairness metrics for extremely underrepresented classes. Importantly, some diseases (*Consolidation*, *Edema*, *Effusion*, *Infiltration*, *Pneumonia*, *Pneumothorax*) have seen ECGL reducing fairness gaps or making EOR more balanced. This shows that the model is robust at addressing gender inequalities, presumably across medical model decisions. These results demonstrate that explanation-based fairness constraints can help the model easily engage in fair behavior, especially in predicting cases that might be biased toward some demographic attributes such as gender. In all, ECGL model successfully balances fairness and accuracy, most particularly by making the same level of achievement as the base model while ensuring significant improvements in demographic fairness. This gives credence to the view that integrating explanations as a constraint indeed can work as additional regularization for fairer decision-making in medical imaging applications entailing high stakes, such as multi-labels.

Computational overhead analysis

Table 7 compares the average training time per epoch of the proposed ECGL method against the base model for the three different dataset-and-machinery configurations: CPU-based training for the Diabetes and Pneumonia X-ray datasets, and GPU-based training for the NIH Chest X-ray dataset. Thus, since ECGL imposes explanation penalties according to the Augmented Lagrangian method, its complexity mainly impacts the training stage; average epoch time becomes a relevant measure of the overhead.

The results showed that ECGL incurs additional computational costs compared to the base models in all datasets and hardware configurations. On the rather small Diabetes dataset, using CPU, ECGL increased the average epoch duration from 0.84 seconds to 3.40 seconds, thus increasing the training time by approximately fourfold. This larger relative overhead can be laid against the extra efforts required for computing constraint penalties, updating the Lagrange multipliers, and integrating explanation guidance within each training iteration. All of which form a larger fraction of the total computation being performed on smaller models and smaller datasets. In contrast, the overhead appears less for the Pneumonia X-ray dataset (also CPU-based); average epoch time goes up from 78.85 seconds to 84.64 seconds increase of about 7%. This more modest relative increase reflects that the base training cost for the deeper CNN model has a major role in defining the total training time, thus subsidizing the additional cost of ECGL constraints. Finally, on the NIH Chest X-ray dataset being trained on a GPU, which is a large-scale and more complex scenario, the average epoch time changes negligibly from 405.36 seconds (base) to 409.82 seconds (ECGL), representing less than a 2% increase. This observation shows that in very large deep learning settings that already have a substantial baseline training cost, the extra overhead imposed by the explanation constraints and augmented Lagrangian updates is very small and practically negligible.

The findings say that although relatively measurable computational overhead is introduced by ECGL, with respect to the model and dataset complexity, its relative cost decreases. Such behavior argues for the scalability and efficiency of ECGL for large-scale medical imaging tasks in real-world expectations, wherein the marginal cost is more than compensated by improvements in fairness and explainability. An increase in training time by a small percentage is a worthy trade-off when you take into account the advantages offered by combining domain knowledge with explanation constraints that lead to the generation of trustworthy and equitable models.

Limitations

Although the ECGL framework proposal constitutes promising advances in the incorporation of explanation constraints for fairness and interpretability for a spectrum of datasets, some limitations are also present. First, the experimental evaluation is at present limited to only three datasets while looking primarily at gender as the protected attribute. This can restrict the generalizability of the results to other domains, and thereby, protected groups like race or age is to be addressed in future work. Another point is that while ECGL is able to scale to larger models and datasets with reasonable computational cost, in some extreme large-scale scenarios or resource-constrained environments, the complexity resulting from the augmented Lagrangian methods can be a bottleneck. In particular, the iterative updates of multipliers and penalties. An in-depth inquiry into the computational efficiency and potential optimization methods remains an untapped venue for research.

Thirdly, the explanation constraints in ECGL require a degree of expert knowledge in defining pertinent conditions, which may prevent the scalability and applicability of the method in settings where such input from domain experts is not available. Future works may look into automated or data-driven methods to learn constraints to weaken the need for manual annotation. Finally, the current study concentrates on comparisons against baselines without contrasting with any state-of-the-art fairness and interpretability methods-including adversarial debiasing or alternative explanation frameworks. Extensions on comparisons will be critical to position ECGL firmly within a more extensive scope of methods for explainable and fair AI rigorously. By explicitly acknowledging these limitations, we aim to provide future research directions concerning making explanation-constrained learning frameworks more robust, much more applicable, and much more efficient.

Conclusions and future directions

In this paper, we explored how integrating explanation constraints directly into the training process can help optimize both predictive accuracy and explanation quality. The proposed *ECGL model* shows strong potential for improving the interpretability and fairness of deep learning models, particularly in healthcare applications where understanding model reasoning is critical. Through experiments on publicly available real-world medical datasets, we demonstrated that ECGL improves model behavior across multiple dimensions. On the tabular diabetes dataset, ECGL achieved a higher predictive accuracy compared to the *Base model*, along with stronger alignment between model explanations and clinical knowledge, as shown by SHAP analysis. On pneumonia X-ray images, although the predictive accuracies of the *Base model* and *ECGL* were similar, GradCAM visualizations suggest that ECGL provides more clinically meaningful explanations by better highlighting relevant regions of interest. Additionally, fairness evaluations showed a 13% improvement in Equalized Odds Ratio (EOR), reflecting more consistent performance across demographic groups. Similarly, the ECGL maintained comparable performance, while showing up to 100% on Demographic Parity Difference (DPD) and Equally Odds Difference (EOD), and over 17% improvement on Demographic Parity Ratio (DPR) and EOR. While these results are promising, ECGLs effectiveness depends heavily on the availability and quality of domain-specific

constraints. Future work could focus on developing methods for automatically learning such constraints from data and on validating ECGL across broader domains to ensure its scalability and generalization. Overall, the findings suggest that ECGL is a valuable step toward building more trustworthy, interpretable, and fair AI models for healthcare and beyond.

Abbreviations

ECGL	Explanation Constraints Guided Learning
EGL	Explanation Guided Learning
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
GradCAM	Gradient-weighted Class Activation Mapping
AUC	Area Under the Receiver Operating Characteristic Curve
EOD	Equalized Odds Difference
EOR	Equalized Odds Ratio

Acknowledgements
This work was supported by PNRR MUR projects FAIR (PE-0000013) and SERICS (PE-00000014). One of the authors was also funded by the Next Generation EU -Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of 'Innovation Ecosystems,' building 'Territorial R&D Leaders' (Directorial Decree n. 2021/3277) - project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009. This work reflects only the authors' views and opinions; neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Author contributions
R.S. carried out Conceptualization, Methodology, Software development, Formal Analysis, and Writing – Original Draft; S.L.M. led Data Curation, Simulation design, Investigation, Resources, and contributed to Writing – Review & Editing; I.T. oversaw Supervision, Project Administration, and Funding Acquisition, and was responsible for Writing – Review & Editing. All authors reviewed and approved the final manuscript.

Funding
Not applicable.

Data availability
The datasets used and/or analyzed during the current study are publicly available at the following links: Diabetes dataset: https://data.mendeley.com/datasets/wj9rwkp9c2/1. Pneumonia X-ray image dataset: https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018. NIH Chest X-ray dataset: https://www.kaggle.com/datasets/nih-chest-xrays/data. The evaluation code is available at: https://github.com/ShahbazianR/AL-EGL.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 April 2025 / Accepted: 1 October 2025
Published online: 31 October 2025

References

1. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, et al. Explainable ai (xai): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1–33.
2. Gao Y, Gu S, Jiang J, Hong SR, Yu D, Zhao L. Going beyond xai: a systematic survey for explanation-guided learning. ACM Comput Surv. 2024;56(7):1–39.
3. Pukdee R, Sam D, Kolter JZ, Balcan M-F-F, Ravikumar P. Learning with explanation constraints. Adv Neural Inf Process Syst. 2024.
4. Stacey J, Belinkov Y, Rei M. Supervising model attention with human explanations for robust natural language inference. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, pp. 11349–57, vol. 36.
5. Rieger L, Singh C, Murdoch W, Yu B. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. International Conference on Machine Learning. PMLR; 2020, pp. 8116–26.
6. Ismail AA, Corrada Bravo H, Feizi S. Improving deep learning interpretability by saliency guided training. Adv Neural Inf Process Syst. 2021;34:26726–39.
7. Zhang Y, Gu S, Gao Y, Pan B, Yang X, Zhao L. Magi: multi-annotated explanation-guided learning. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023, pp. 1977–87.
8. Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Med Image Anal. 2022;79, 102470.
9. Saeed W, Omlin C. Explainable ai (xai): a systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst. 2023;263, 110273.
10. Lemaréchal C. Lagrangian relaxation. Computational combinatorial optimization: Optimal or provably near-optimal solutions. 2001;112–56.
11. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013).
12. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. International Conference on Machine Learning. PMLR; 2017, pp. 3319–28.
13. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 618–26.
14. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Advances in neural information processing systems 31. 2018.
15. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, pp. 1135–44.
16. Lundberg SM, Lee. S.-I: a unified approach to interpreting model predictions. Advances in neural information processing systems 30. 2017.
17. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling lime and shap: adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020, pp. 180–86.
18. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015, pp. 1721–30.
19. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv (CSUR). 2018;51(5):1–42.
20. Dong L, Chen L, Zheng C, Fu Z, Zukaib U, Cui X, Shen Z. Ocie: augmenting model interpretability via deconfounded explanation-guided learning. Knowl-Based Syst. 2024;302, 112390 (https://doi.org/10.1016/j.knosys.2024.112390.
21. Zhao Q, Chang C-W, Yang X, Zhao L. Robust explanation supervision for false positive reduction in pulmonary nodule detection. Med Phys. 2024;51(3):1687–701.
22. Zeineldin RA, Karar ME, Burgert O, Mathis-Ullrich F. Neuroign: explainable multimodal image-guided system for precise brain tumor surgery. J Educ Chang Med Syst. 2024;48(1):25.
23. Tinauer C, Heber S, Pirpamer L, Damulina A, Schmidt R, Stollberger R, Ropele S, Langkammer C. Interpretable brain disease classification and relevance-guided deep learning. Sci Rep. 2022;12(1), 20254.
24. Ross AS, Hughes MC, Doshi-Velez F. Right for the right reasons: training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017).
25. Zhang Y, Gu S, Gao Y, Pan B, Yang X, Zhao L. Magi: multi-annotated explanation-guided learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023, pp. 1977–87.
26. Yan X, Mao Y, Ye Y, Yu H, Wang F-Y. Explanation guided cross-modal social image clustering. Inf Sci. 2022;593:1–16. https://doi.org/10.1016/j.ins.2022.01.065.
27. Gu S, Zhang Y, Gao Y, Yang X, Zhao L. Essa: explanation iterative supervision via saliency-guided data augmentation. Proceedings of the 29th ACM

SIGKDD Conference on Knowledge Discovery and Data Mining. 2023, pp. 567–76.
28. Fukui H, Hirakawa T, Yamashita T, Fujiyoshi H. Attention branch network: learning of attention mechanism for visual explanation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 10705–14.
29. Gao Y, Sun TS, Bai G, Gu S, Hong SR, Liang Z. Res: a robust framework for guiding visual explanation. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022, pp. 432–42.
30. Šefčík F, Benesova W. Improving a neural network model by explanation-guided training for glioma classification based on mri data. Int J Inf Technol. 2023;15(5):2593–601.
31. Yan A, Huang T, Ke L, Liu X, Chen Q, Dong C. Explanation leaks: Explanation-guided model extraction attacks. Inf Sci. 2023;632:269–84. https://doi.org/10.1016/j.ins.2023.03.020.
32. Liu G, Zhang J, Chan AB, Hsiao JH. Human attention guided explainable artificial intelligence for computer vision models. Neural Netw. 2024;177, 106392 https://doi.org/10.1016/j.neunet.2024.106392.
33. Parraga O, More MD, Oliveira CM, Gavenski NS, Kupssinskü LS, Medronha A, Moura LV, Simões GS, Barros RC. Fairness in deep learning: a survey on vision and language research. ACM Comput Surv. 2023.
34. Tian H, Zhu T, Liu W, Zhou W. Image fairness in deep learning: problems, models, and challenges. Neural Comput Appl. 2022;34(15):12875–93.
35. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst. 2012;33(1):1–33.
36. Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR. Optimized pre-processing for discrimination prevention. Adv Neural Inf Process Syst. 2017.
37. Kilbertus N, Rojas Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B. Avoiding discrimination through causal reasoning. Adv Neural Inf Process Syst. 2017.
38. Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A. The case for process fairness in learning: feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law. Barcelona, Spain: 2016, p. 11, vol. 1.
39. Mehta R, Shui C, Arbel T. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In: Medical imaging with deep learning. PMLR; 2024. p. 1453–92.
40. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR). 2021;54(6):1–35.
41. Buolamwini J, Gebru T. Gender shades. Intersectional accuracy disparities in commercial gender classification. Conference on Fairness, Accountability and Transparency. PMLR; 2018, pp. 77–91.
42. Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Adv Neural Inf Process Syst. 2016.
43. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183–86.
44. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems 29. 2016.
45. Birgin EG, Martínez JM. Practical augmented lagrangian methods for constrained optimization. SIAM, ???; 2014.
46. Nocedal J, Wright SJ. Numerical optimization. Springer, ???; 1999.
47. Lundberg SM, Lee. S.-I. Consistent feature attribution for tree ensembles. arXiv preprint arXiv:1706.06060 (2017).
48. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017, pp. 618–26). doi: https://doi.org/10.1109/ICCV.2017.74.
49. Chen RJ, Wang JJ, Williamson DF, Chen TY, Lipkova J, Lu MY, Sahai S, Mahmood F. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023;7(6):719–42.

Publisher's Note