CentaurMD: Confidence-Aware Human–AI Decision Fusion for Multi-Label Disease Diagnosis via Label-Specific MoE

Youcheng Zhang
Northwestern Polytechnical University
Xi'an, China
zhangyoucheng@mail.nwpu.edu.cn

Hui Wang
Harbin Engineering University
Harbin, China
hui.wang@hrbeu.edu.cn

Jiaqi Liu*
Northwestern Polytechnical University
Xi'an, China
jqliu@nwpu.edu.cn

Yao Zhang
Northwestern Polytechnical University
Xi'an, China
yaozh.g@nwpu.edu.cn

Zhiwen Yu*
Harbin Engineering University &
Northwestern Polytechnical University
Harbin, China
zhiwenyu@nwpu.edu.cn

Bin Guo
Northwestern Polytechnical University
Xi'an, China
guob@nwpu.edu.cn

ABSTRACT

Multi-label disease diagnosis is prevalent in clinical applications, such as chest X-rays that may indicate multiple coexisting diseases. Despite advances in AI, current models remain insufficient for reliably addressing such complexity. Human–AI synergy thus emerges as both a necessary and promising approach, motivating our focus on effective decision fusion for multi-label disease diagnosis. There are two challenges. Confidence, a key factor in decision fusion, is often unrecorded in human annotations, making its estimation nontrivial. Moreover, label-specific variations in human and model expertise must be considered to achieve effective fusion. To address these challenges, we propose **CentaurMD**, a confidence-aware human–AI decision fusion framework based on label-specific Mixture-of-Experts (MoE). We first present a novel multi-label confusion matrix construction method that employs maximum entropy modeling to capture label correlations, enabling more accurate confidence estimation and weight allocation. Then, we develop a label-specific MoE module with dedicated gating networks and thresholds, which dynamically adjust expert weights using information extracted from the confusion matrix via a Transformer encoder. Extensive experiments on three real-world clinical datasets demonstrate that our method reduces Hamming loss by 39.14% and improves MMR (missed-misdiagnosis reduction) by 17.38%, achieving substantial diagnostic improvements.

KEYWORDS

Human-AI Decision Fusion; Mixture of Experts; Medical Diagnosis; Multi-Label Classification; Confusion Matrix

ACM Reference Format:
Youcheng Zhang, Hui Wang, Jiaqi Liu*, Yao Zhang, Zhiwen Yu*, and Bin Guo. 2026. CentaurMD: Confidence-Aware Human–AI Decision Fusion for Multi-Label Disease Diagnosis via Label-Specific MoE. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems*

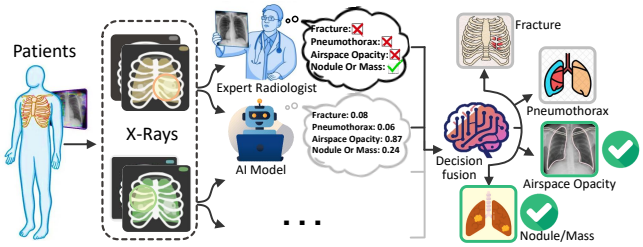


Figure 1: An example of collaborative human-AI disease diagnosis on the multi-label dataset ChestX-ray. Each medical image in this dataset may have multiple pulmonary diseases.

(AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 9 pages. https://doi.org/10.65109/DWYB7830

1 INTRODUCTION

Multi-label classification (MLC) [26] is the task of predicting all relevant labels for a given instance, where multiple labels often co-occur. This is particularly common in medical diagnosis [34, 40], as a report may indicate several diseases simultaneously. For example, a Chest X-ray (CXR) image may reveal fractures, pneumothorax, airspace opacity, and nodules or masses [25]. To improve MLC performance, recent AI models have been explored from different perspectives, including leveraging latent contextual information [3, 4, 15, 17], and modeling label or image correlations [5, 37, 38, 43]. These methods work effectively for typical MLC scenarios. However, in clinical disease diagnosis, where disease types are complex and diagnostic accuracy demands are stringent, relying solely on AI models yields suboptimal results and may lead to misdiagnosis or missed diagnosis. Therefore, incorporating human experts proves more effective for complex disease diagnosis [31]. Human-AI collaboration has been widely applied across medical domains, including online disease diagnosis [30, 42], skin cancer recognition [27], and decision-making in sepsis diagnosis [39].

Human-AI collaborative decision fusion refers to the weighted combination of human and AI decision outcomes. The Bayesian approach [12] is widely adopted and effective for label classification. However, it relies on fixed statistical assumptions to support rigorous mathematical reasoning. As a result, it cannot dynamically adapt to varying label difficulty or expert proficiency across

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DWYB7830

cases, potentially leading to suboptimal weighting. Data-driven Mixture-of-Experts (MoE) frameworks weight human and model predictions based on confidence. AI models typically produce both binary outputs and probabilistic confidence scores, whereas human experts often provide only binary labels without explicit confidence. Accurate estimation of human confidence is therefore critical for effective fusion. In MLC, individual labels differ in difficulty and class balance, which yields label-wise variability in prediction reliability. For a given case, a human may be more confident and accurate on label A while the model outperforms on label B, so it is essential to identify and exploit these variations in decision fusion.

Motivated by these considerations, achieving effective human–AI collaboration in multi-label disease diagnosis entails two major challenges. First, existing confidence estimation methods commonly rely on confusion matrices but overlook label correlations, making it difficult to accurately assess human expert confidence. Second, variations across labels introduce heterogeneity in decision fusion. To tackle these challenges, we propose two novel modules: (i) A human expert confidence assessment module that leverages a maximum entropy model to capture label correlations and constructs a multi-label confusion matrix (MLCM) using a weighted probabilistic rule for confidence estimation. (ii) A label-specific MoE module with dedicated gating networks and thresholds, augmented by a Transformer encoder with multi-head attention to model correlations in the MLCM and refine gating weights for accurate fusion.

In this paper, we propose **CentaurMD**, a confidence-aware human–AI decision fusion framework for multi-label disease diagnosis via label-specific MoE. CentaurMD first estimates human expert confidence via the MLCM. Next, the label-specific MoE module produces initial weights for human and AI models via dedicated gating networks, which are then refined by the estimated confidence to generate the final probabilistic outputs. Finally, label-specific thresholds convert these probabilistic outputs into binary decisions.

Our main contributions are summarized as follows:

- **Assessing Human Expert Confidence:** We propose an MLCM construction method that captures label correlations without assuming independence, enabling more accurate and unbiased estimation of human expert confidence.
- **Label-Specific MoE:** We introduce the application of a Mixture-of-Experts to human–AI collaborative decision fusion for the first time. By employing label-specific gating networks and thresholds and calibrating initial weights based on confidence, our method improves diagnostic accuracy while reducing misdiagnosis and missed diagnosis rates.
- **Experimental Study:** Experiments on three real-world clinical datasets show that our approach consistently outperforms human-only, AI-only, and existing human–AI collaborative methods. Compared with fusion baselines, it reduces Hamming loss by 39.14% and improves MMR by 17.38% on average, yielding markedly better diagnostic performance.

2 RELATED WORK

2.1 Multi-Label Disease Classification

In recent years, MLC has received increasing attention for its ability to model label correlations, address class imbalance, and ensure scalability [11]. Zhu et al. [43] extracted label correlations via scene detection and co-occurrence matrices, yet depends on contextual information with high computation cost. AdaBoost [16] optimizes label correlations to minimize Hamming loss, though implementation remains challenging. Siahroudi et al. [24] introduced a partial multi-label learning method based on constrained clustering, transforming MLC into a clustering problem. In medical diagnosis, MLC commonly arises in applications such as Chest X-ray and ECG analysis. Chen et al. [3] developed a semantic graph embedding framework to enhance visual embeddings for classification. HydraViT [21] employs self-attention on key regions while maintaining lesion co-occurrence awareness, enhancing CNN-based classification. In medical MLC, disease relationships are often more complex, making label correlation modeling especially critical [**?**]. Despite these advances, existing methods ignore the complementary role of human expertise, motivating research on human–AI decision fusion.

2.2 Human–AI Fusion for Diagnostic Tasks

Existing works [20, 22] show that human–AI teams can outperform either agent alone. Most existing human–AI fusion methods, however, are designed for single-label scenarios [18, 41, 42]. Kerrigan et al. [12] proposed a Bayesian approach that models annotator reliability via confusion matrices and calibrates probabilities under conditional independence. Other research [2, 8, 13] introduced allocation systems that assign each case to either a classifier or a human expert to mitigate human limitations. Current deferral algorithms [1, 33] allow AI to decide whether to predict or defer to humans, but they often require extensive expert annotations. Hemmer et al. [9] addressed this issue with a three-stage approach while maintaining performance. Despite these advances, existing methods largely ignore multi-label correlations in diagnostic tasks.

2.3 MLCM for Confidence Assessment

The construction of MLCM generally follows two approaches: transformation from single-label confusion matrices and direct MLCM construction. Single-label confusion matrices are well-established, treating each label as an independent class and aggregating the results to form a multi-label matrix. Existing work [28] has reformulated MLC as a multi-class problem using the label power-set transformation, but this approach incurs exponential computational complexity for large label sets and suffers from severe data sparsity. In MLC tasks, certain label pairs often exhibit strong co-occurrence patterns and misclassification correlations [10], making methods that ignore these correlations theoretically inadequate. For direct MLCM construction, several strategies have been explored. An ontology-driven method [29] measures feature-level semantic similarity between labels but performs poorly when explicit semantic relationships are absent. Krstinić et al. [14] proposed a generic MLCM construction framework, yet it cannot capture varying dependency strengths across label pairs, limiting its practical effectiveness.

3 METHODOLOGY

3.1 Problem Statement

In multi-label diagnostic tasks, each instance may contain multiple diseases (up to n labels) with strong label correlations. For a given instance, a human expert (e.g., a doctor) provides a binary prediction $Y^h = (y_1^h, ..., y_n^h) \in \{0, 1\}^n$, while the AI model produces

probabilistic predictions $Y^m = (y_1^m, ..., y_n^m) \in [0,1]^n$ based on the instance features $V \in \mathbb{R}^d$. To assess human expert confidence, we construct an MLCM $C \in \mathbb{R}^{(n+1)\times(n+1)}$ that accounts for label correlations. Each element C_{ij} represents the confidence of a human expert assigning label j given true label i, with diagonal entries C_{ii} indicating confidence in correct classification. An additional row and column are incorporated to handle cases where No True Label (NTL) or No Predicted Label (NPL) exists. Using an MoE approach, we fuse human-AI predictions to produce a refined probabilistic output $\hat{Y} = (\hat{y}_1, ..., \hat{y}_n) \in [0,1]^n$. In this context, the MoE framework dynamically assigns weights to experts based on input-specific features, allowing the system to leverage their complementary strengths for improved decision making. This refined output is then converted to a binary prediction $Y = (y_1, ..., y_n) \in \{0,1\}^n$ via label-specific thresholds $\theta = (\theta_1, ..., \theta_n) \in [0,1]^n$. The objective of our method is to jointly minimize the Hamming loss and the residual risk of missed diagnosis and misdiagnosis:

$$\min_{\tilde{G}_h, \tilde{G}_m, \theta} \text{HammingLoss} + (1 - \text{MMR}), \tag{1}$$

where $\tilde{G}_h$ and $\tilde{G}_m$ denote the weights of the human expert and the AI models, respectively. MMR (Missed-Misdiagnosis Reduction) represents the ability to reduce missed diagnosis and misdiagnosis, defined as the harmonic mean of anti-missed diagnosis capability $(1 - FNR)$ and anti-misdiagnosis capability $(1 - FPR)$:

$$\text{MMR} = \frac{2 \times (1 - FNR) \times (1 - FPR)}{(1 - FNR) + (1 - FPR)}, \tag{2}$$

where FNR and FPR denote the missed diagnosis rate and the misdiagnosis rate, respectively.

3.2 Framework Overview

Motivated by these challenges, we propose a human–AI decision fusion framework using MLCM and label-specific MoE. For each instance, both human experts and AI models provide their predictions. To assess expert reliability, we construct an MLCM to serve as a proxy for confidence estimation. Decision fusion is performed via label-specific gating networks, which adaptively combine human and AI predictions. This approach assigns distinct weights to the predictions of human experts and AI and computes a weighted combination to yield the final decision. Moreover, label-specific thresholds are applied to convert the probabilistic outputs into binary (0/1) decisions. As illustrated in Figure 2, CentaurMD consists of two core components: human expert confidence assessment and a label-specific MoE, described in detail below.

a) Assessing Human Expert Confidence. To assess human expert confidence, we propose a method for constructing an MLCM that captures label correlations. Specifically, we define label correlation as the probability of label i being misclassified as label j, and model this relationship using a maximum entropy model without prior assumptions to derive a transition probability matrix. Subsequently, we introduce a probability-weighted MLCM construction method that integrates this transition probability matrix to quantify expert confidence. This will be detailed in §3.3.

b) Label-Specific MoE. To fuse predictions from human experts and AI models, we design dedicated gating networks for each label, which generate initial weights based on predictions and image

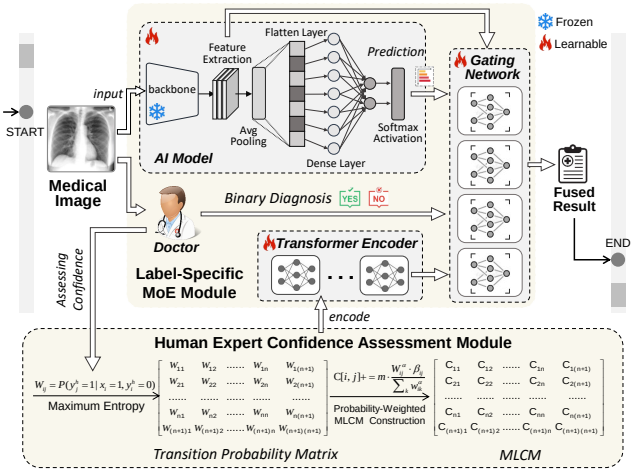


Figure 2: Framework Overview of CentaurMD.

features. To further refine these weights, a Transformer encoder equipped with multi-head self-attention is employed to extract information from MLCM. The extracted representation is then linearly transformed and fed into gating networks to recalibrate the human expert's weights. A weighted combination is subsequently computed based on the recalibrated weights to produce the probabilistic output. This will be detailed in §3.4.

3.3 Assessing Human Expert Confidence

Our human confidence assessment approach consists of two key modules: 1) modeling label correlations via a maximum entropy model, and 2) constructing a probability-weighted MLCM. We first estimate label prediction error probabilities using a maximum entropy model without distributional assumptions, encoding them in a transition probability matrix that captures all possible errors. These probabilities are then grouped into four categories to construct the MLCM. By explicitly modeling label correlations, our method yields a more faithful confusion matrix, facilitating principled and accurate confidence estimation and providing a robust foundation for subsequent human-AI decision fusion.

3.3.1 Modeling Label Correlations via Maximum Entropy. We define label correlations as the probability of misclassifying class-i as class-j, which is quantified using a transition probability matrix,

$$W_{ij} = P(y_j^h = 1 | x_i = 1, y_i^h = 0). \tag{3}$$

When $i = j$, the diagonal element $W_{ii} = P(y_i^h = 1 | x_i = 1)$ denotes the probability of a correct prediction for class i. Traditional approach [6] computes conditional probabilities using the Bayes method, which relies on the strong assumption of label independence (i.e.,$P(y_i|x_i) = P(y_i), \forall i \neq j$), expressed as,

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{k=1}^{n} P(y_k|x_k). \tag{4}$$

Diagnostic tasks often involve complex label correlations, rendering the independence assumption biased in evaluation. This issue is alleviated through a maximum entropy model. The model selects the distribution with maximal entropy subject to the known constraints, thereby reducing unwarranted assumptions. For each

Algorithm 1 Construction of MLCM

Require: x: Ground-truth labels, y^h: Human expert diagnosis
Ensure: MLCM
1: Extract label co-occurrence features from (x, y^h)
2: Train a Maximum Entropy Model to estimate transition probabilities $P(y_j^h \mid x)$
3: Construct transition probability matrix W
4: **for** each label i **do**
5: Update MLCM$[i, :]$ based on $W[i, :]$
6: **end for**
7: **return** MLCM

label pair (i, j), an independent binary logistic regression model is trained to estimate y_j^h, as detailed below:

Constructing Feature Representations. To avoid redundant computation, we focus on regions of interest by extracting all samples where the ground truth is $x_i = 1$ and human prediction is $y_i^h = 0$, which correspond to misclassified regions. For each selected sample, an input feature vector $\mathbf{x} = [x_i, y_i^h]$ is constructed to capture contextual correlations and potential confusion patterns across classes. The resulting feature-label pairs are subsequently used to train the maximum entropy model, which estimates inter-class confusion and dependency relationships.

Defining Prediction Targets. For each label pair (i, j), we define the prediction target as y_j^h, which represents the human prediction for class j. This serves as the supervision signal for training the maximum entropy model, enabling it to capture how likely class-j is predicted when class-i is misclassified.

Estimating Conditional Probabilities. We use a logistic regression function to model the conditional probability $P(y_j^h = 1|\mathbf{x})$. The estimation formula of the maximum entropy model is as follows,

$$P(y_j^h = 1|\mathbf{x}) = \sigma(w_{j0} + w_{j1} \cdot x_j + w_{j2} \cdot y^h), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, and $w_{j\cdot}$ denotes the parameters trained for class j.

Training the Maximum Entropy Model. We optimize the model using L-BFGS, learning parameters $\mathbf{w}$ and bias b by minimizing an L2-regularized logistic loss,

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \log\left(1 + e^{-y_i^h \mathbf{w}^T \mathbf{x}^i}\right), \tag{6}$$

where n is the number of instances, and C is a regularization hyperparameter that controls the strength of the L2 penalty.

Inference with the Maximum Entropy Model. After training, inference is performed with the fixed input $\mathbf{x} = [1, 0]$, representing the condition $x_i = 1, y_i^h = 0$. The resulting probability $P(y_j^h = 1 \mid \mathbf{x})$ is recorded in the conditional probability matrix at position (i, j), indicating the likelihood of predicting class j when class i is missed.

3.3.2 Constructing a Probability-Weighted MLCM.. To address the limitation of assuming uniform label correlations [14], the transition probability matrix (§3.3.1) is used to weight and construct a more realistic MLCM. Following prior work, an additional row and column are retained for NTL and NPL states to accommodate boundary conditions. Empirical analysis of human expert diagnosis relative to ground-truth labels reveals four characteristic scenarios: (1) perfect match, (2) missed detection, (3) over-prediction, and (4) mixed errors. To avoid redundancy and enable more precise

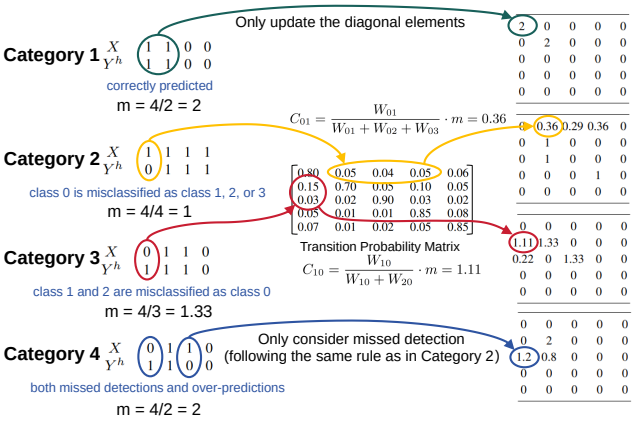


Figure 3: Category Contributions to MLCM (see §3.3).

assessment of human expert confidence, these four patterns are categorized during human expert confusion matrix construction.

Category 1: Perfect Match (true labels match predicted labels). When the true labels and predicted labels are the same (i.e., $X = Y^h$), the confusion matrix is updated as follows:

$$C[i, i] \leftarrow C[i, i] + m, \tag{7}$$

where m is the weight for the perfect match and is computed as,

$$m = \frac{n}{max(NT, NP)}, \tag{8}$$

where NT and NP denote the numbers of true and predicted labels for a given instance. When both are zero ($X = 0, Y^h = 0$), the confusion matrix is updated as follows,

$$C[NTL, NPL] \leftarrow C[NTL, NPL] + n, \tag{9}$$

which handle cases with no true or predicted labels.

Category 2: Missed Detection (unpredicted true labels exist). When the human expert misses a true label (i.e., $\exists i(x_i = 1 \wedge y_i^h = 0) \wedge \forall j(y_j^h = 1)$), the confusion matrix is updated as follows,

$$C[i, j] \leftarrow C[i, j] + \frac{W_{ij} \cdot m}{\sum_{k=0}^{n} W_{ik}}, \tag{10}$$

where W_{ij} denotes the probability of misclassifying category i as j in the transition probability matrix. When predicted labels are zero ($Y^h = 0$), the confusion matrix is updated as follows,

$$C[i, \text{NPL}] \leftarrow C[1, \text{NPL}] + m. \tag{11}$$

Category 3: Over-prediction (extra false positives exist). When the expert predicts a non-existent condition (i.e., $\exists i(x_i = 0 \wedge y_i^h = 1) \wedge \forall k(y_k^h = 1)$), the confusion matrix is updated as follows,

$$C[k, i] \leftarrow C[k, i] + \frac{W_{ki}}{\sum W_{ki} \cdot m}. \tag{12}$$

Category 4: Mixed Errors (both missed and over-predicted labels). When both missed detections and over-predictions occur (i.e., $\exists i\ (x_i = 1 \wedge y_i^h = 0) \wedge \exists j\ (y_j^h = 1)$ and $\exists p\ (x_p = 0 \wedge y_p^h = 0) \wedge \exists q\ (y_q^h = 1)$), only the missed detection case is considered, following Category 2's rule, to avoid redundant calculations. Following these steps, we constructed an MLCM for assessing the confidence of human experts, as detailed in Algorithm 1.

To provide an intuitive illustration of our method, the detailed calculation process is presented in Figure 3. We show one representative example for each of the four categories, including the ground

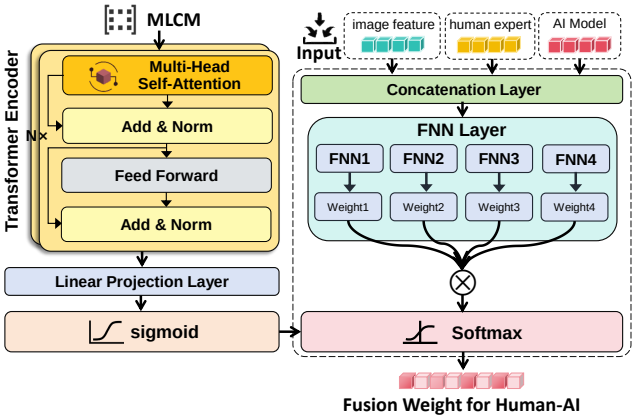


Figure 4: The implementation details of the MoE module.

truth labels (X), human expert diagnosis (Y^h), and the corresponding transition probability matrix (W):

① Category 1: Both class 0 and class 1 are correctly predicted. To minimize the overall error, we normalize the MLCM such that the sum of all entries equals the number of labels n. Accordingly, $m = \frac{4}{2} = 2$, yielding $C_{00} = 2$ and $C_{11} = 2$.

② Category 2: Class 0 is misclassified as class 1, 2, or 3, while the other classes are correctly predicted. First, we compute $m = \frac{4}{4} = 1$. The correctly predicted labels have the following updates: $C_{11} = 1$, $C_{22} = 1$, $C_{33} = 1$. For the mispredicted labels, the transition probability matrix elements are $W_{01} = 0.05$, $W_{02} = 0.04$, $W_{03} = 0.05$. The value for C_{01} is computed as:

$$C_{01} = \frac{W_{01}}{W_{01} + W_{02} + W_{03}} \cdot m = 0.36. \tag{13}$$

Similarly, we obtain $C_{02} = 0.29$ and $C_{03} = 0.36$.

③ Category 3: Class 1 and class 2 are mostly correctly predicted, with some misclassified as class 0. First, we compute $m = \frac{4}{3} = 1.33$. The correctly predicted classes are updated as: $C_{11} = 1.33$, $C_{22} = 1.33$. For the mispredicted labels, the transition probabilities are $W_{10} = 0.15$, $W_{20} = 0.03$. The value for C_{10} is computed as:

$$C_{10} = \frac{W_{10}}{W_{10} + W_{20}} \cdot m = 1.11. \tag{14}$$

In a similar manner, we obtain $C_{20} = 0.22$.

④ Category 4: Both missed detections and over-predictions occur. To avoid redundant computation, only the missed detection is considered, following the same rule as in category 3.

3.4 Label-Specific MoE

We propose a label-specific MoE decision fusion module that dynamically selects and aggregates multiple expert models according to input relevance. It is particularly suitable for multi-label classification (MLC), where different labels benefit from distinct expert combinations. By integrating human expertise with AI models, the proposed module enhances diagnostic performance. Our methodology comprises two steps: 1) recalibrating human-AI weights via MLCM and 2) label-specific fusing human-AI decisions. We first extract image features using a backbone network (ResNet-18). Features are processed by a Transformer encoder that leverages MLCM information to model human-specific decision tendencies and uncertainties. This learned information is then used to adaptively weight the human decisions via gating networks. Modeling label correlations allows effective human–AI integration, reducing individual weaknesses and improving diagnostic accuracy.

Figure 4 illustrates the detailed architecture of our proposed MoE module. At the concatenation layer, aggregated input features are fed into independent Feedforward Neural Networks (FNNs) per label to produce initial weights. The MLCM undergoes iterative information extraction through an N-layer Transformer block, after which the refined representation is projected via a linear transformation layer to recalibrate the initial weights, specifically adjusting the human expert contributions. The final fusion weights for both human experts and AI models are then generated. The implementation details will be elaborated in §3.4.1 and §3.4.2.

3.4.1 Recalibrating Human Weights via MLCM.. To capture and enrich relational information in the confusion matrix C (obtained in §3.3), we compute its self-correlation C^TC, producing a comprehensive representation of confusion characteristics. The enhanced matrix is processed by a Transformer encoder with two multi-head self-attention layers, capturing label correlations, dependencies, and expert-specific biases to facilitate the subsequent decision fusion stage. It is formally expressed as

$$M = TransformerEncoder(C^TC) \in \mathbb{R}^{n \times d}, \tag{15}$$

where M is a matrix, with each row as a d-dimensional embedding of a label's enriched features for human-AI fusion.

Within the Transformer encoder, an attention mechanism quantifies misclassification relationships. For each label i, the model computes an attention score for every label j based on learned query and key vectors, capturing how strongly label i attends to label j and revealing confusion patterns.

$$\alpha_{ij} = \frac{\exp(q_i^T k_j / \sqrt{d})}{\sum_{k=1}^{n} \exp(q_i^T k_k / \sqrt{d})}, \tag{16}$$

where $q_i, k_j \in \mathbb{R}^d$ are the query and key vectors for labels i and j. The attention weights α_{ij} measure the likelihood that label i is confused with label j, capturing misclassification patterns and correlations. These distributions are used to compute contextualized label embeddings, forming the final output M in Eq. (15).

During human-AI fusion, to determine the allocation of weights for human experts, we recalibrate expert confidence based on the information extracted by the Transformer encoder. Specifically, the output features $M \in \mathbb{R}^{n \times d}$ are projected into an expert-specific decision space via a linear transformation $W_p \in \mathbb{R}^{d \times E}$. This produces confusion influence factors $\lambda \in \mathbb{R}^{n \times E}$, which quantify the trust assigned to human expert decisions for each label:

$$\lambda = W_p M + b_p, \tag{17}$$

where $\lambda \in \mathbb{R}^{n \times E}$, E denotes the number of experts, and b_p is the corresponding bias term.

The confusion influence factors λ are input into the gating networks to recalibrate expert weights. This amplifies expert influence on confident labels and reduces it on error-prone labels, improving the reliability and performance of human-AI fusion.

3.4.2 Label-Specific Fusing Human-AI Decisions. In MLC, human experts and AI models exhibit inherent heterogeneity, as their differing feature focus leads to varying prediction accuracy. We adopt

Algorithm 2 CentaurMD: Human-AI Decision Fusion Process

Require: Ground truth labels $\mathbf{X}$, human predictions $\mathbf{Y}^h$, AI predictions $\mathbf{Y}^m$, medical image;
Ensure: Final predictions $\mathbf{Y}$;
Phase 1: Assessing Human Expert Confidence
Compute transition probability matrix W
Construct MLCM $C[i, j] += m\frac{W_{ij}\cdot\beta_{ij}}{\sum W_{ik}}$
Phase 2: MoE-Based Fusion Training
while early stopping criterion not met **do**
  Extract image features $\mathbf{x}$ using CNN
  Encode C using Transformer to obtain latent representation
  Generate initial weights $G_0(\mathbf{x})$ via gating networks
  Recalibrate weights $\tilde{G} = Softmax(G_0(\mathbf{x}) \odot \sigma(\lambda))$
  Fuse expert predictions: $\hat{\mathbf{Y}} = \tilde{G}_h \odot \mathbf{Y}^h + \tilde{G}_m \odot \mathbf{Y}^m$
end while
Phase 3: Inference
Given test image, $\mathbf{x}$ and MLCM representation
Output final predictions: $\mathbf{Y} = \mathbb{I}[\hat{\mathbf{Y}} > \theta]$, $\theta = \{\theta_1, \ldots, \theta_n\}$

an MoE architecture for human-AI fusion, inspired by the HQS approach [36]. The HQS approach combines task-specific and shared experts to improve MLC performance. Central to MoE are gating networks that adaptively select and weight experts. Our gating network dynamically assigns weights to human and AI decisions using input features and decision outputs, enabling personalized, context-aware fusion. Specifically, we implement disease-specific four-layer neural gating networks with feature compression and confusion matrix recalibration to generate adaptive weights.

Initial Weight Generation. Due to divergent prediction capabilities of human experts and AI models across labels, each label category employs a dedicated gating network. A four-layer fully connected network maps the d-dimensional image features to expert weight space, using LeakyReLU ($\alpha = 0.01$) activation to enhance gradient flow. The initial weights are computed as

$$G_0(\mathbf{x}) = Softmax(W_{g3} \cdot LeakyReLU(W_{g2} \cdot LeakyReLU(W_{g1} \cdot \mathbf{x} + b_1) + b_2)), \tag{18}$$

where W_{gi} is the i^{th} fully connected layer weights.

We apply Layer Normalization (LN) after the second linear transformation and Dropout after the first nonlinearity during training to improve generalization and training stability.

Confusion Matrix Recalibration. The confusion influence factor λ (obtained in §3.4.1) is fused with dynamically generated weights G_0 via Hadamard product to adjust the recalibrated expert weights:

$$\tilde{G} = Softmax(G_0(\mathbf{x}) \odot \sigma(\lambda)), \tag{19}$$

where $\odot$ denotes element-wise multiplication with the broadcasting mechanism, and $\sigma(\cdot)$ represents the Sigmoid function, which constrains the confusion factor to the interval $(0, 1)$.

Human-AI Fusion Mechanism. The fused prediction combines the expert outputs y^h and y^m with learned weights to obtain the final probability fusion result $\hat{y} \in [0, 1]$:

$$\hat{y} = \tilde{G}_h \odot y^h + \tilde{G}_m \odot y^m, \tag{20}$$

where $\tilde{G}_h$ is the first column of $\tilde{G}$, $\tilde{G}_m$ is the second column of $\tilde{G}$. Specifically, $\tilde{G}_h$ represents the human expert weights for each label, and $\tilde{G}_m$ represents AI model weights for each label.

Label-Specific Thresholds. To convert the probabilistic outputs $\hat{y}_i$ into binary predictions, we adopt a label-specific thresholding strategy. Instead of applying a uniform threshold (e.g., 0.5) across all labels, which may lead to suboptimal performance due to label imbalance and varying difficulty, we optimize a separate threshold θ_i for each label i. This approach enables more fine-grained decision boundaries tailored per class. Formally, the set of thresholds are denoted as $\theta = \{\theta_1, ..., \theta_n\} \in [0, 1]^n$. These thresholds are optimized using a grid search strategy on a held-out validation set to minimize the Hamming loss, a common metric for multi-label tasks:

$$\theta_i = \arg\min_{\theta\in[0,1]} \text{HammingLoss}\left(y_i^{\text{val}}, \mathbb{I}(\hat{y}_i^{\text{val}} \geq \theta)\right). \tag{21}$$

The final binary prediction for label i is then made as follows,

$$y_i = \begin{cases} 1, & \hat{y}_i \geq \theta_i \\ 0, & otherwise \end{cases}. \tag{22}$$

This method ensures binary decision boundaries are empirically calibrated for each label, improving overall multi-label performance.

During human-AI decision fusion, weights are assigned by jointly considering image features and decisions. Moreover, due to the heterogeneity of multi-label classification, it is necessary to design label-specific gating networks and thresholds. In summary, the training process of our framework is presented in Algorithm 2.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Real Datasets. To validate the effectiveness of our method, we conduct experiments on three public multi-label clinical datasets.

The ChestX-ray[19, 32] dataset, released by the National Institutes of Health (NIH), is a widely used for multi-label disease classification. We focus on four clinically significant conditions: Fracture, Pneumothorax, Airspace Disease, and Nodule or Mass. Each radiograph may exhibit multiple abnormalities. The dataset contains 4,375 chest X-rays with 13,080 annotations from 22 radiologists, with ground truth set by consensus of three radiologists.

The S12L-ECG[23] dataset from the Telehealth Network of Minas Gerais (TNMG) contains 827 12-lead ECG recordings with six cardiac conditions: 1dAVb, RBBB, LBBB, SB, AF, and ST. Annotations were independently done by three experienced cardiologists, with final labels set by consensus following the protocol in [23].

The LID-FFA [35] dataset includes 5,435 fundus fluorescein angiography (FFA) images from 500 patients using the Spectralis HRA+OCT system. It covers six categories: Leakage (L), Transmission and Pooling (TP), Staining (ST), Shadowing (SH), Non-Perfusion (NP), and Vessel Abnormality (VA). Original annotations from ophthalmologists are not available, and only the ground truth labels have been retained. For experiments, simulated expert annotations were generated with an accuracy of 90%.

4.1.2 Baselines. We compare our method with several baselines: (1) human experts only, (2) AI models only (using different neural networks like ResNet18, VGG19 and AlexNet), and (3) existing human-AI joint decision methods CHM [12], HAIT [8], JSF [13] and L2D-CL [35]. The CHM [12] develops a probabilistic combination strategy that fuses the model's confidence scores with human predictions. The HAIT method [8] introduces a joint training

Table 1: Performance comparison on ChestX-ray, S12L-ECG, and LID-FFA datasets.

Methods	ChestX-ray Dataset				S12L-ECG Dataset				LID-FFA Dataset			
	Hamming Loss	AUC	MAP	MMR	Hamming Loss	AUC	MAP	MMR	Hamming Loss	AUC	MAP	MMR
Human Only	0.0708	0.7946	0.5119	0.6639	0.0067	0.8731	0.9492	0.7626	0.1083	0.8955	0.8907	0.9052
AI model Only [7]	0.0830	0.7976	0.5189	0.4411	0.0093	0.9753	0.9990	0.6889	0.1711	0.8218	0.8521	0.8332
CHM [12]	0.0647	0.8803	0.6994	0.6648	0.0027	-	-	0.7848	0.0873	0.9463	0.9576	0.9123
HAIT [8]	0.1062	0.6913	0.4115	0.4666	0.0067	-	0.7276	0.7626	0.1083	0.8689	0.9009	0.9021
JSF [13]	0.1133	0.7374	0.3967	0.3118	0.0067	-	0.7276	0.7626	0.1558	0.8378	0.8887	0.8587
L2D-CL [35]	0.0684	0.7097	0.4264	0.8864	0.0064	0.9242	0.8509	0.7523	0.1040	0.8257	0.8602	0.8827
CentaurMD (Ours)	**0.0611**↓	**0.8886** ↑	**0.7024** ↑	**0.7172** ↑	**0.0015** ↓	**0.9753** ↑	**0.9990** ↑	**0.8070** ↑	**0.0848** ↓	**0.9627** ↑	**0.9638** ↑	**0.9202** ↑
w/o MLCM	0.0707	0.7986	0.5186	0.6641	0.0097	0.9885	0.7910	0.6884	0.1012	0.9402	0.9522	0.9108
w/o Gating Network	0.0731	0.8669	0.6063	0.5460	0.0293	0.3798	0.2630	0.0000	0.1104	0.9484	0.9479	0.8968
w/o Label-Spec. Thres.	0.8231	0.8883	0.6986	0.2667	0.9707	0.9990	0.8128	0.0560	0.3645	0.9548	0.9572	0.7615

Table 2: Performance of CNN-based AI Models

Methods	Hamming Loss	AUC	MAP	MMR
AI Model (ResNet18)	0.0830	0.7976	0.5189	0.4411
AI Model (VGG19)	0.0942	0.7988	0.5042	0.4555
AI Model (AlexNet)	0.1018	0.7603	0.4223	0.2053

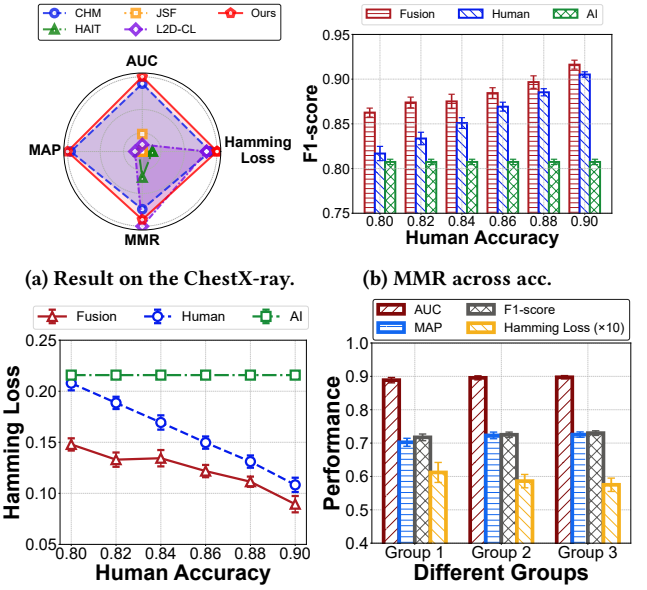
framework that simultaneously learns a classifier and an instance allocation strategy, assigning each sample to the most suitable decision maker, whether a human expert or an AI model. The JSF introduces a joint sparse framework for multi-label classification, where a set of $m + 1$ binary sigmoid classifiers is trained. Each instance is assigned to the human or model with the highest activation score. The L2D-CL [35] proposes a new loss function based on the learn-to-defer framework to mitigate underfitting.

4.1.3 Metrics. We evaluate our method using four widely adopted multi-label classification metrics: Hamming Loss, AUC, MAP, and MMR. Hamming Loss computes the proportion of misclassified labels across samples and disease categories, while AUC measures the probability that a randomly chosen positive instance ranks above a negative one. MAP evaluates ranking quality by computing the average precision (AP) for each label and averaging over labels. MMR quantifies the ability to jointly reduce the FNR and FPR, as defined in Equation 2. These metrics enable a comprehensive evaluation of our method's performance in multi-label disease diagnosis.

4.2 Overall Performance

Table 1 shows consistent improvements of our framework across all datasets. Compared to human experts alone, it reduces Hamming loss by 13.70%, 77.61%, and 21.70%, while improving MMR by 8.03%, 5.82%, and 1.66%, respectively. Relative to AI-only baselines (ResNet18), Hamming loss decreases by 26.39%, 83.87%, and 50.44%, with MMR gains of 62.59%, 17.14%, and 10.44%. Furthermore, our framework consistently outperforms existing fusion methods (CHM, HEIT, JSF, L2D-CL), achieving an average Hamming loss reduction of 39.14% and an average MMR improvement of 17.38%. These results highlight its effectiveness in reducing diagnostic errors via human-AI collaboration. In addition, variance analysis further shows significantly improved robustness over existing approaches across datasets of varying scales and clinical settings.

To intuitively compare methods, we present a multi-metric radar chart for the ChestX-ray dataset in Figure 5(a). All metrics are normalized, with Hamming loss inverted so larger polygon areas indicate better performance. Our method yields the largest, most rounded area, demonstrating balanced and superior performance.



(a) Result on the ChestX-ray. (b) MMR across acc.

(c) Hamming loss across acc. (d) Performance on diff. groups

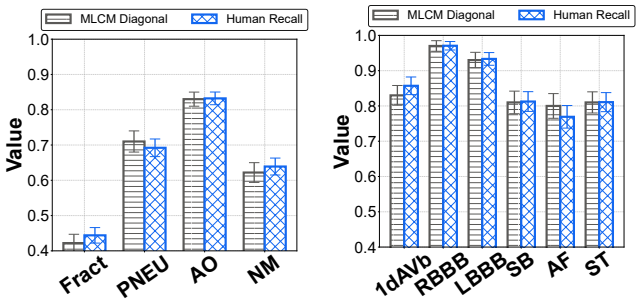
Figure 5: Performance on Multiple Perspectives. (d) Group 1 (Human + ResNet18), Group 2 (Human + ResNet18 + VGG19), Group 3 (Human + ResNet18 + VGG19 + AlexNet).

4.3 Performance Across Human Accuracy

To evaluate the effectiveness of our framework across varying levels of human expertise, we simulated human experts with accuracies ranging from 0.80 to 0.90 on the LID-FFA dataset. These decisions fused with a ResNet-18 model to evaluate performance in terms of the MMR and Hamming loss (see Fig. 5 (b) and (c)). The results show that the fusion approach outperforms both human-only and AI-only methods. As human expert accuracy increases, MMR steadily rises while Hamming loss correspondingly decreases, demonstrating the framework's adaptability to varying expertise levels.

4.4 Ablation Study

To comprehensively assess each module's contribution to overall performance, we conduct ablation experiments across all datasets, analyzing the effects of the Multi-Label Confusion Matrix (MLCM), gating network, and label-specific thresholds. Removing MLCM causes substantial performance degradation, with Hamming loss increasing by 13.58%, 84.54%, and 89.88%, and MMR decreasing by 8.00%, 17.23%, and 1.02%, confirming its key role in modeling expert



(a) MLCM recall On ChestX-ray. (b) MLCM recall On S12L-ECG.

Figure 6: Comparison of MLCM Diagonal and Human Recall.

bias and label correlations. Replacing the proposed gating network with conventional stacking methods further raises Hamming loss to 0.0731, 0.0293, and 0.1104, while reducing MMR to 0.5460, 0.0000, and 0.8968, underscoring the necessity of this design. The sharpest performance drop occurs when label-specific thresholds are replaced by a fixed 0.5, leading to drastic increases in Hamming loss and decreases in MMR across all datasets, highlighting label heterogeneity and the critical importance of threshold customization. Collectively, these results demonstrate that MLCM corrects expert bias, the gating network adaptively adjusts human–AI weights, and label-specific thresholds ensure reliable clinical predictions, jointly enabling the robust performance of the proposed framework.

4.5 Scalability Analysis

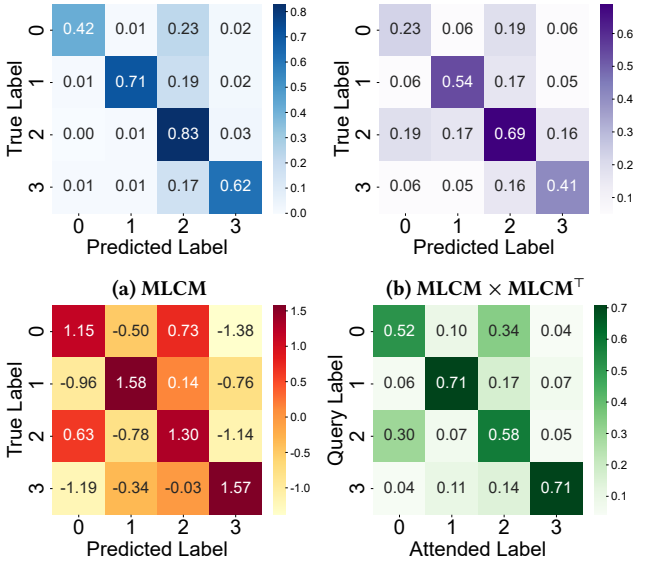
To validate the scalability, we configured various human-AI groups by sequentially incorporating ResNet18, VGG19, and AlexNet into the baseline human expert setting on the ChestX-ray dataset. As shown in Table 2, the performance of individual AI models progressively decreased, with Hamming Loss values of 0.0830, 0.0942, and 0.1018, and corresponding MMR of 0.4411, 0.4555, and 0.2053, respectively, indicating inferior performance compared to the human expert. Nevertheless, as illustrated in Figure 5(d), model performance improves with the addition of AI models. Specifically, compared to Group 1, the inclusion of VGG19 yields an approximate 2% improvement in MMR and a 4% reduction in Hamming Loss. Further addition of AlexNet achieves an additional 2.8% increase in MMR and a 6% overall decrease in Hamming Loss relative to the baseline. These results demonstrate the method's scalability.

4.6 Validating the Effectiveness of MLCM

To assess whether the MLCM reliably reflects human expert confidence, we compare its diagonal elements with label recall. As shown in Equation (23), recall is equivalent to the normalized MLCM diagonal, making this comparison a principled evaluation. Figures 6(a) and (b) show that their difference is below 0.025 for all labels and that they exhibit a strong linear correlation (Pearson r = 0.984), confirming that the MLCM accurately reflects expert confidence. The horizontal axis in Figure 7(a) corresponds to the labels, which are Fracture, Pneumothorax, Airspace Opacity, and Nodule Or Mass.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} = Normalized(C_{ii}), \tag{23}$$

where C_{ii} denotes the diagonal element of the row-normalized MLCM, corresponding to the recall for class i.



(c) Transformer Encoder (d) Attention Weights

Figure 7: Visualizations of the evolution of MLCM across different modules on the ChestX-ray dataset.

4.7 Visualizing the Evolution of MLCM

To investigate the evolution of the MLCM within our framework, we visualize its key transformation steps, as illustrated in Figures 7. Figure 7(a) depicts the MLCM derived from the ChestX-ray dataset, with diagonal elements MLCM$[i, i]$ represent correctly predicted instances for label i. Figure 7(b) shows the autocorrelation matrix C^TC, where the diagonal elements measure the strength of each label's relationship with itself. Figure 7(c) illustrates the semantic correlation matrix generated by the Transformer encoder, with diagonal elements indicating the semantic correlation strength of each label. Figure 7(d) shows SoftMax-normalized attention weights refining expert contributions, with diagonals indicating each label's final confidence. In summary, the transition from Figures 7(a) to (d) demonstrates the model's structural interpretability.

5 CONCLUSION

In this paper, we propose a novel Human-AI Decision Fusion framework that integrates experienced human experts with the global modeling capabilities of AI models to improve multi-label medical diagnosis. Specifically, we introduce a probability-weighted MLCM construction approach to accurately assess human expert confidence, providing a foundation for informed weight allocation in decision fusion. Furthermore, we develop an MoE-based fusion strategy that calibrates the initial gating weights via the MLCM and refines the final predictions using label-specific thresholds. A Transformer encoder is employed to effectively extract information from the MLCM for more precise weight assignment. Extensive experiments on three publicly available real-world datasets demonstrate the effectiveness and efficiency of our method. Future work will address label imbalance to further improve fusion performance.

ACKNOWLEDGMENTS

This work was supported in part by the National Key RD Program of China (No. 2024YFB4505502) and the National Natural Science Foundation of China (No. 62372381).

REFERENCES

[1] Jean V Alves, Diogo Leitão, Sérgio Jesus, Marco OP Sampaio, Javier Liébana, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2025. A benchmarking framework and dataset for learning to defer in human-AI decision-making. *Scientific data* 12, 1 (2025), 506.

[2] Susan C Athey, Kevin A Bryan, and Joshua S Gans. 2020. The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, Vol. 110. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 80–84.

[3] Bingzhi Chen, Zheng Zhang, Yingjian Li, Guangming Lu, and David Zhang. 2021. Multi-label chest X-ray image classification via semantic similarity graph embedding. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 4 (2021), 2455–2468.

[4] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. 2022. Structured semantic transfer for multi-label recognition with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 339–346.

[5] Zhaomin Chen, Quan Cui, Ruoxi Deng, Jie Hu, and Guodao Zhang. 2024. Modeling Label Correlations with Latent Context for Multi-label Recognition. In *European Conference on Computer Vision*. Springer, 218–234.

[6] Walter Gerych, Tom Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke A Rundensteiner. 2021. Recurrent bayesian classifier chains for exact multi-label classification. *Advances in Neural Information Processing Systems* 34 (2021), 15981–15992.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming effective human-AI teams: building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948* (2022).

[9] Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. 2023. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6002–6011.

[10] Gregory Holste, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, Dongkyun Kim, Trong-Hieu Nguyen-Mau, Minh-Triet Tran, et al. 2024. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Medical Image Analysis* 97 (2024), 103224.

[11] Sajjad Kamali Siahroudi, Zahra Ahmadi, and Daniel Kudenko. 2024. Effectively Capturing Label Correlation for Tabular Multi-Label Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1060–1069.

[12] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems* 34 (2021), 4421–4434.

[13] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 154–165.

[14] Damir Krstinić, Maja Braović, Ljiljana Šerić, and Dunja Božić-Štulić. 2020. Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology* 1 (2020), 1–14.

[15] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16478–16488.

[16] Jiaxuan Li, Xiaoyan Zhu, and Jiayin Wang. 2023. AdaBoost. C2: boosting classifiers chains for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8580–8587.

[17] Lanting Li, Peng Cao, Jinzhu Yang, and Osmar R Zaiane. 2022. Modeling global and local label correlation with graph convolutional networks for multi-label chest X-ray image classification. *Medical & Biological Engineering & Computing* 60, 9 (2022), 2567–2588.

[18] Hefei Liang, Jiaqi Liu, Zhiwen Yu, and Bin Guo. 2024. Utilizing Machine Experience: Reinforcement Learning in Automated Diagnosis. In *International Conference on Neural Information Processing*. Springer, 327–340.

[19] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294, 2 (2020), 421–431.

[20] Nastaran Okati, Abir De, and Manuel Rodriguez. 2021. Differentiable learning under triage. *NeurIPS* 34 (2021), 9140–9151.

[21] Şaban Öztürk, M Yiğit Turalı, and Tolga Çukur. 2025. Hydravit: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. *Biomedical Signal Processing and Control* 100 (2025), 106959.

[22] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[23] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. 2020. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications* 11, 1 (2020), 1760.

[24] Sajjad Kamali Siahroudi and Daniel Kudenko. 2023. Partial Multi-label Learning via Constraint Clustering. In *International Conference on Neural Information Processing*. Springer, 453–469.

[25] Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, et al. 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine* 3, 1 (2020), 70.

[26] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition* 118 (2021), 107965.

[27] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature medicine* 26, 8 (2020), 1229–1234.

[28] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3, 3 (2007), 1–13.

[29] Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. 2020. Knowledge-based construction of confusion matrices for multi-label classification algorithms using semantic similarity measures. *arXiv preprint arXiv:2011.00109* (2020).

[30] Hu Wang, David Butler, Yuan Zhang, Jodie Avery, Steven Knox, Congbo Ma, Louise Hull, and Gustavo Carneiro. 2024. Human–AI collaborative multi-modal multi-rater learning for endometriosis diagnosis. *Physics in Medicine & Biology* 70, 1 (2024), 015008.

[31] Hui Wang, Zhiwen Yu, Yao Zhang, Yanfei Wang, Fan Yang, Liang Wang, Jiaqi Liu, and Bin Guo. 2024. hmos: An extensible platform for task-oriented human–machine computing. *IEEE Transactions on Human-Machine Systems* 54, 5 (2024), 536–545.

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2097–2106.

[33] Zixi Wei, Yuzhou Cao, and Lei Feng. 2024. Exploiting human-ai dependence for learning to defer. In *Forty-first International Conference on Machine Learning*.

[34] Jiayin Xiao, Si Li, Tongxu Lin, Jian Zhu, Xiaochen Yuan, David Dagan Feng, and Bin Sheng. 2024. Multi-label chest x-ray image classification with single positive labels. *IEEE transactions on medical imaging* 43, 12 (2024), 4404–4418.

[35] Jianyang Xie, Xiuju Chen, Yitian Zhao, Yanda Meng, He Zhao, Anh Nguyen, Xiaoxin Li, and Yalin Zheng. 2024. Multi-disease Detection in Retinal Images Guided by Disease Causal Estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 743–753.

[36] Zihao Yin, Chen Gan, Kelei He, Yang Gao, and Junfeng Zhang. 2024. Hybrid sharing for multi-label image classification. In *The Twelfth International Conference on Learning Representations*.

[37] Jialu Zhang, Jianfeng Ren, Qian Zhang, Jiang Liu, and Xudong Jiang. 2023. Spatial context-aware object-attentional network for multi-label image classification. *IEEE Transactions on Image Processing* 32 (2023), 3000–3012.

[38] Jialu Zhang, Qian Zhang, Jianfeng Ren, Yitian Zhao, and Jiang Liu. 2022. Spatial-context-aware deep neural network for multi-class image classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1960–1964.

[39] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2024. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[40] Yuan Zhang, Yutong Xie, Hu Wang, Jodie C Avery, M Louise Hull, and Gustavo Carneiro. 2025. A Novel Perspective for Multi-modal Multi-label Skin Lesion Classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 3549–3558.

[41] Xuehan Zhao, Jiaqi Liu, Zhiwen Yu, and Bin Guo. 2024. HADT: Human-AI diagnostic team via hierarchical reinforcement learning. In *SIAM International Conference on Data Mining*. SIAM, 860–868.

[42] Xuehan Zhao, Jiaqi Liu, Yao Zhang, Zhiwen Yu, and Bin Guo. 2024. Haiformer: Human-ai collaboration framework for disease diagnosis via doctor-enhanced transformer. In *European Conference on Artificial Intelligence*. IOS Press, 1495–1502.

[43] Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, and Jiuxin Cao. 2023. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1473–1482.