Enhancing the performance of multimodal chronic respiratory disease detection via transfer learning with data augmentation techniques

Thanh-Trong Tran[1,2], Dang-Khoa Bui[1,2], Hoang-Anh Ngo[3,4*], and Tuan-Khoi Nguyen[1,2*†]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Linh Xuan Ward, Ho Chi Minh City, Vietnam
{trong.tran588413799, khoa.buidkhoa, tuankhoin}@hcmut.edu.vn
[3] Woolcock Institute of Medical Research, Hanoi, Vietnam
[4] Faculty of Medicine and Health, The University of Sydney, NSW, Australia
hoang-anh.ngo@alumni.polytechnique.org

Abstract. Chronic respiratory diseases (CRDs) impose a substantial global health burden, and their early detection remains challenging due to heterogeneous clinical presentations and limited positive samples for many rare conditions. As such this study develops a CRD diagnosis system by integrating chest X-ray (CXR) image features extracted from a DenseNet model using a transfer learning approach with clinical features derived from electronic health records (EHR). To address class imbalance at the feature level, two data augmentation strategies—Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling (ROS)—are investigated in conjunction with XGBoost-based models. Specifically, SMOTE is applied in combination with both XGBClassifier and XGBRegressor during the modeling of image-derived features and clinical data, while ROS is evaluated with XGBClassifier as a complementary oversampling baseline, particularly in multimodal settings that incorporate EHR features. Experimental results demonstrate that SMOTE significantly improves recall when using image features alone, whereas ROS combined with XGBClassifier yields superior performance when EHR data are incorporated, leading to enhanced precision and recall across multiple rare disease categories.

Keywords: Chronic Respiratory Diseases · Chest X-ray · DenseNet · Electronic Health Record · SMOTE · ROS · XGBoost · Medical Imaging

[1] These authors contributed equally to this work as first authors.
[2] ∗ Co-senior authors.
[3] † Corresponding author.

1 Introduction

Chronic respiratory diseases (CRDs) are a major cause of global morbidity and mortality, especially in regions affected by air pollution and long-term tobacco exposure [1,2]. Early and accurate detection is essential to prevent disease progression, improve quality of life, and reduce healthcare burden [3].
Chest X-ray (CXR) imaging is a widely used, non-invasive diagnostic tool for assessing lung abnormalities [4], but its interpretation requires substantial clinical expertise and is prone to diagnostic errors due to subtle and overlapping pathological patterns [5]. Recent advances in machine learning (ML), particularly deep learning (DL), have shown promising results in automated CXR-based diagnosis. Convolutional neural networks (CNNs) can effectively extract complex visual features and achieve high accuracy in medical imaging tasks [6]. However, DL models often lack interpretability and are sensitive to severe class imbalance, a common issue in CRD datasets where positive samples for rare diseases are scarce [7,8].
Electronic health records (EHR) provide complementary clinical information and offer better interpretability through traditional ML models [9]. Nevertheless, EHR data are frequently sparse and incomplete, and class imbalance remains a challenge when EHR is used alone. Previous work has explored combining EHR variables with CXR features to improve prediction performance [10]; however, class imbalance and augmentation were not explicitly addressed, motivating our use of Synthetic Minority Oversampling Technique (SMOTE) and other Random Over-sampling (ROS) methods in this study. To address these limitations, this work proposes a multimodal CRD prediction framework that integrates CXR image features and structured EHR data. Image features are extracted using transfer learning from pre-trained CNNs, while conventional ML models are applied to clinical variables. To mitigate class imbalance, feature-level data augmentation techniques, including SMOTE and Random Over-Sampling (ROS), are employed.
The main contributions of this study are as follows: (1) we propose a multimodal CRD diagnosis framework that combines CXR and EHR data; (2) we investigate feature-level data augmentation strategies to mitigate class imbalance; and (3) we provide a comprehensive evaluation of both predictive performance and model decision behavior.

2 Method

2.1 Design Choices

This project utilizes two primary data sources: the Medical Information Mart for Intensive Care (MIMIC) and the CheXpert dataset.
MIMIC-IV [11,12] contains 299,712 detailed records of patient hospital admissions at Beth Israel Deaconess Medical Center (BIDMC). This database is used to extract patients' electronic health records (EHRs) and to provide ground truth labels for the task, specifically the International Classification of Diseases (ICD) codes that document diagnoses assigned at discharge.
MIMIC-CXR [13] comprises 227,835 de-identified imaging studies from 64,588

patients at BIDMC and is linked to MIMIC-IV. This dataset is employed to retrieve corresponding chest X-ray (CXR) images for the patients included in the study.

CheXpert [14] is an additional large-scale CXR dataset used in this project. It consists of 224,316 chest X-ray images from 65,240 patients at Stanford Hospital and is entirely independent of BIDMC. In this study, CheXpert is utilized to pre-train the selected deep learning model, enabling transfer learning on CXR data from MIMIC. This strategy enhances the generalization capability of the model to real-world clinical data, including previously unseen cases.

Data Access Control: To comply with ethical requirements related to medical data usage, all authors have completed the necessary training programs to obtain authorized access to the databases. For security and privacy reasons, the datasets used in this project are not publicly released. Instead, the authors may provide data processing scripts that require user authentication to verify training status before accessing the databases and constructing the datasets upon request.

Target CRD Categories: Given the large number of chronic respiratory diseases (CRDs), it is impractical to cover all possible conditions. Therefore, this study focuses on eight representative CRDs that are expected to exhibit diverse clinical characteristics and data distributions [15,16,17,18,19]. A summary of the selected diseases is presented in Table 1. Patients are identified as having CRDs based on the descriptions of International Classification of Diseases (ICD) codes assigned at hospital discharge, which are directly obtained from MIMIC-IV.

EHR Features: Based on the eight selected CRDs, twelve EHR features are extracted from MIMIC-IV. For these features, all temporal information is removed to minimize the risk of patient information leakage and label leakage [15,16,17,18,19].

Deep Learning Image Features: DenseNet is a deep neural network architecture designed to preserve feature representations at multiple levels of abstraction [20]. In this study, DenseNet is selected to extract the most informative patterns from chest X-ray (CXR) images. A CheXpert-pretrained model, available in the `torchxrayvision` package [21], is employed to enable transfer learning. Features are extracted from the second-to-last dense layer, resulting in 18 latent features for each CXR image.

Disease	ICD keywords	Number of cases	% Of cases in cohort
arthritis	arthritis	1101	4.1%
bronchitis	bronchitis	1485	5.1%
fracture	fracture, broken	2078	7.1%
lung_cancer	tumor, cancer	7787	26.8%
lung_infection	infection, infectious	952	3.2%
pneumonia	pneumonia	7980	27.4%
scoliosis	scoliosis, curvature	150	0.5%
tuberculosis	tuberculosis	771	2.7%

Table 1: Statistics of case counts and prevalence rates for CRDs in the dataset

2.2 Data Collection and Organization

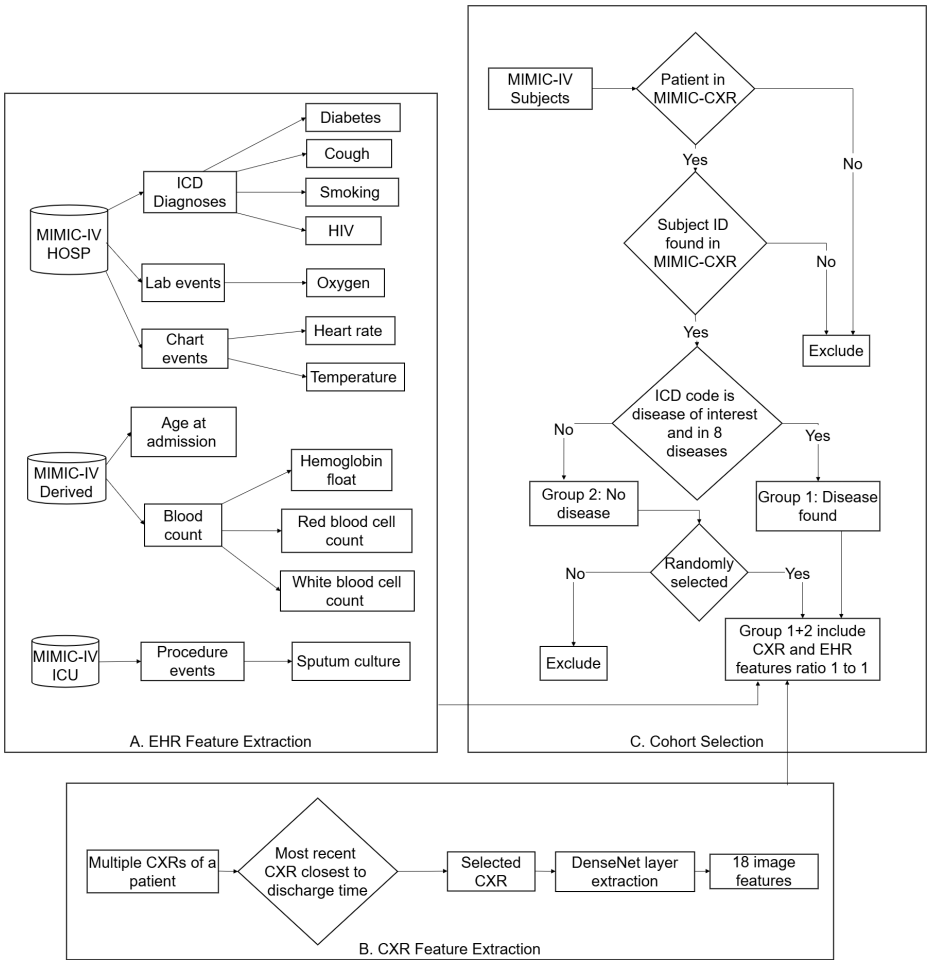

Figure 1: Overview of the cohort extraction and data construction pipeline

Patients in MIMIC-IV are excluded if no chest X-ray (CXR) records are available. The remaining patients are then screened using ICD codes to determine whether they match any of the eight selected CRDs, thereby dividing the population into two groups. All patients with matching ICD codes are included in the study cohort, while the remaining patients—who are substantially more numerous—are randomly sampled to match the size of the CRD-positive group. This cohort construction process is summarized in Figure 1C.

Subsequently, eleven EHR features are extracted from the MIMIC-IV database, as illustrated in Figure 1A. For each selected patient in the study cohort, the most recent CXR prior to hospital discharge is chosen. This strategy simplifies the modeling process and avoids data duplication, as a single patient may have multiple CXR examinations recorded in MIMIC [13]. In the final study dataset,

each patient is expected to have 18 image-derived features and 12 EHR features, resulting in a total of 30 features per patient.
While the image features require no additional preprocessing due to the standardized data acquisition protocol and the comprehensive cleaning procedures already applied in the MIMIC-CXR dataset [13], the EHR data contain several abnormal measurements, such as body temperatures recorded as high as 150°C. Therefore, to ensure data integrity for continuous features, we removed any measurements considered outliers in the feature distribution, defined as values that deviate from the mean by more than three standard deviations.

2.3 Processing Pipeline

The processing pipeline of the proposed approach consists of two main stages: *data augmentation* and *evaluation*. Each model is trained and evaluated on the augmented dataset in order to analyze the impact of data balancing on predictive performance.

2.3.1 Stage 1: Data Augmentation

Method 1: SMOTE with XGBClassifier/XGBRegressor.
In the dataset, there exists a substantial imbalance between negative and positive labels, which adversely affects predictive performance. To address this issue, SMOTE [8] is applied to generate synthetic samples for the minority class, thereby balancing the data distribution and improving the model's learning capability. Specifically, SMOTE synthesizes new minority-class samples by interpolating between existing minority instances and their nearest neighbors in the feature space, rather than simply duplicating samples. This approach expands the decision region of the minority class and reduces the risk of overfitting commonly associated with naive oversampling methods. Subsequently, XGBoost [22] is trained on the augmented dataset, as it effectively handles both class imbalance and missing values.
SMOTE for image features: SMOTE is applied to the 18 image features extracted from chest X-ray images. The objective is to ensure that the numbers of positive and negative samples in the training set are approximately balanced, allowing the model to learn disease-related features more effectively. After augmentation, the newly generated image samples are combined with incomplete EHR data to construct a complete dataset, which is then used for model training.
XGBClassifier and XGBRegressor for EHR features: The EHR data consist of both categorical features (`diabetes`, `HIV`, `smoke`, and `cough`) and numerical features `age`, `oxygen`, `heart_rate`, `temperature`, `hemoglobin`, `rbc`, and `wbc`). To predict these features:

- `XGBClassifier` is used for categorical features to determine binary label values (0 or 1) based on patient characteristics (Fig. 2B).
- `XGBRegressor` is employed for numerical features (Fig. 2D). Prior to regression, a missing-value mask is predicted using `XGBClassifier` (Fig. 2C) to indicate whether a feature value exists. If the predicted mask equals 1, `XGBRegressor` is subsequently applied to estimate the actual value of the corresponding numerical feature (Fig. 2E).

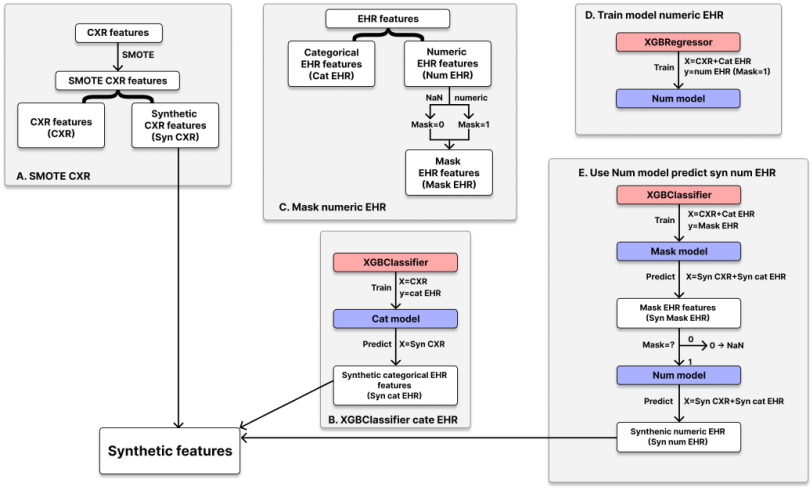

Figure 2: Data augmentation pipeline for integrating CXR and EHR features.

The integration of SMOTE, `XGBClassifier`, and `XGBRegressor` ensures that the input dataset for the final XGBoost model is balanced, complete, and representative of both imaging and EHR information, thereby improving predictive performance for chronic respiratory disease (CRD).

Method 2: ROS with XGBClassifier.

In addition to SMOTE, Random Over-Sampling (ROS) [23] is employed as a simpler class-balancing strategy. Unlike SMOTE, ROS balances the dataset by randomly duplicating minority-class samples in the training set until a more balanced class distribution is achieved. This approach reduces model bias toward the majority class and improves the learning of discriminative patterns for rare disease cases. To avoid data leakage, the dataset is first split into training and testing sets prior to any augmentation. ROS is applied exclusively to the training set, while the test set remains unchanged in order to preserve the real-world data distribution. An `XGBClassifier` [22] is then trained on the ROS-balanced training data under two feature configurations: (i) image features only, and (ii) combined image and EHR features. ROS serves as a strong baseline because it does not introduce synthetic samples and preserves the original missing-data patterns commonly observed in clinical datasets [24].

2.3.2 Stage 2: Evaluation

Two models are evaluated in this stage. First, an XGBoost model is trained and evaluated on datasets augmented using SMOTE, using image features alone and in combination with EHR features, in order to assess the contribution of each data modality. Second, an `XGBClassifier` trained with Random Over-Sampling (ROS) is used as a comparative model to specifically analyze the effect of class balancing. For each of the eight CRDs, models are evaluated under both feature configurations and data settings (baseline and ROS). Performance is reported

using Accuracy, Precision, Recall, and F1 score, which jointly reflect overall classification quality and the ability to detect minority-class cases.

3 Results

3.1 Stage 1: Data Augmentation

Method 1: SMOTE and XGBClassifier/XGBRegressor.
Among the eight diseases investigated, each disease uses an independent train–test split; however, the data augmentation procedure using SMOTE and the EHR feature prediction pipeline remain identical across all diseases. Therefore, in the Results section, we present a representative example to illustrate the effectiveness of the data augmentation stage, while also providing overall statistics to ensure methodological consistency across all diseases. After applying SMOTE to the CXR image features, the number of positive samples increased substantially. Due to varying train–test splits across multiple runs, SMOTE generated different synthetic positive samples in each training set. On average, 8,533 additional positive samples were created, resulting in a more balanced class distribution. This improved balance plays a critical role in reducing model bias toward the negative class and facilitates more effective learning of disease-related features.

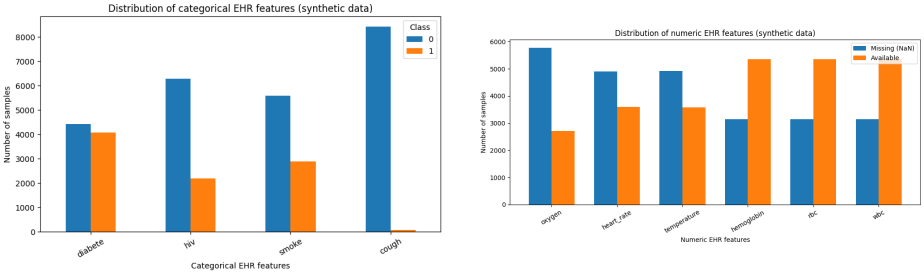

Figure 3: Distribution of categorial (left) and numerical (right) EHR features in SMOTE-generated CXR samples

From Figure 3, it can be observed that the binary values (0/1) of features such as diabetes, HIV, smoke, and cough are no longer missing and are consistently distributed across the entire synthetic dataset. This indicates that the XGBClassifier successfully learned the relationships between image features and categorical clinical variables, enabling it to assign plausible labels to synthetic samples without distorting the original data distribution. The results also demonstrate that the mask prediction performed by `XGBClassifier` effectively identifies the presence of numerical features, while `XGBRegressor` estimates actual values only at positions permitted by the predicted mask. Overall, by preserving the inherent missing-data structure of EHR records and avoiding generating clinically implausible values, this two-step combined pipeline produces an augmented dataset that is both label-balanced and information-complete across multiple data modalities. This dataset serves as a consistent and reliable input for the subsequent model performance evaluation stage.

Method 2: Random Over-Sampling (ROS).

Similar to SMOTE, ROS is employed to address the class imbalance problem in CRD-related tasks. However, instead of generating synthetic samples, ROS balances the dataset by randomly duplicating positive-class samples in the training set. Therefore, to evaluate the effectiveness of the data augmentation stage, we focus on comparing the label distributions before and after applying ROS on the training set. The training set exhibits severe class imbalance across the eight diseases. Negative samples (Class 0) typically exceed 1,000 per disease, whereas positive samples (Class 1) are much fewer, ranging from 150 cases for scoliosis (0.5%) to approximately 8,000 cases for pneumonia and lung cancer (∼27%; Table 1). The imbalance is particularly pronounced for tuberculosis (771 cases, 2.7%), arthritis (1,101 cases, 4.1%), and scoliosis (150 cases, 0.5%), which can bias the model toward negative predictions and lead to low positive-class Recall. Applying ROS achieves its intended objective of reducing class imbalance and enabling the model to observe more positive cases during training, which is expected to improve disease detection performance, particularly Recall, during evaluation.

3.2 Stage 2: Evaluation

Model 1: XGBoost.

Disease	Without Data Augmentation				With Data Augmentation			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
pneumonia	70.3 ± 0.3	34.0 ± 0.9	9.0 ± 0.6	14.2 ± 0.8	58.1 ± 0.4	31.0 ± 0.6	43.0 ± 0.6	36.0 ± 0.4
tuberculosis	97.3 ± 0.1	6.7 ± 14.9	0.1 ± 0.2	0.2 ± 0.3	88.6 ± 0.5	3.7 ± 0.2	13.2 ± 1.2	5.8 ± 0.3
bronchitis	94.8 ± 0.1	10.4 ± 3.4	0.2 ± 0.1	0.4 ± 0.2	81.0 ± 0.6	7.0 ± 0.6	21.8 ± 2.3	10.5 ± 0.9
arthritis	95.9 ± 0.02	10.0 ± 22.4	0.0 ± 0.1	0.1 ± 0.2	83.3 ± 0.5	5.2 ± 0.6	18.0 ± 2.1	8.0 ± 0.8
fracture	92.7 ± 0.1	14.6 ± 4.5	0.6 ± 0.2	1.1 ± 0.4	75.7 ± 0.4	8.5 ± 0.4	25.0 ± 1.6	12.7 ± 0.7
lung cancer	71.5 ± 0.3	35.7 ± 0.9	8.2 ± 0.3	13.3 ± 0.4	58.5 ± 0.5	30.4 ± 0.5	42.4 ± 1.3	35.4 ± 0.8
scoliosis	99.4 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	97.8 ± 0.2	0.7 ± 0.5	2.0 ± 1.3	1.0 ± 0.7
lung infection	96.7 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	86.6 ± 0.2	4.0 ± 0.5	13.4 ± 1.9	6.1 ± 0.8

Table 2: Evaluation results using image features only (in percentages, mean ± SD reported)

Although the model achieves high overall Accuracy (ranging from approximately 70.3% to nearly 100%), Precision and Recall remain very low in the absence of SMOTE, particularly for severely imbalanced diseases such as tuberculosis, arthritis, and scoliosis. After applying SMOTE, Recall improves substantially—for example, pneumonia increases from 9.0% to 43.0%, and lung cancer from 8.2% to 42.4%. This improvement also leads to a marked increase in F1-score, with

pneumonia rising from 14.2% to 36.0% and lung cancer from 13.3% to 35.4%. These results (provided in detail in Table 2) indicate that the model learns minority-class characteristics more effectively, although Precision shows a slight decrease in some cases. Furthermore, the low standard deviation values suggest that the model exhibits stable performance across repeated evaluations.

Disease	Without Data Augmentation				With Data Augmentation			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
pneumonia	78.4 ± 0.3	70.6 ± 0.7	36.4 ± 0.9	48.0 ± 0.6	78.1 ± 0.4	69.5 ± 1.1	35.9 ± 1.1	47.3 ± 1.0
tuberculosis	97.9 ± 0.1	96.2 ± 3.2	20.4 ± 1.4	33.7 ± 2.0	97.9 ± 0.1	98.2 ± 1.3	20.0 ± 2.2	33.2 ± 3.0
bronchitis	95.2 ± 0.2	78.2 ± 3.8	10.6 ± 1.3	18.7 ± 2.0	95.2 ± 0.2	76.6 ± 5.7	9.9 ± 2.5	17.5 ± 4.1
arthritis	96.6 ± 0.2	93.4 ± 2.6	16.3 ± 2.3	27.7 ± 3.4	96.6 ± 0.2	94.2 ± 2.7	16.3 ± 0.7	27.8 ± 1.1
fracture	93.2 ± 0.1	69.4 ± 4.5	8.5 ± 0.7	15.1 ± 1.1	93.2 ± 0.1	68.8 ± 6.2	8.6 ± 0.9	15.3 ± 1.5
lung cancer	77.3 ± 0.2	66.0 ± 0.9	31.8 ± 1.2	42.9 ± 1.1	77.2 ± 0.4	65.8 ± 1.5	31.3 ± 1.0	42.4 ± 1.2
scoliosis	99.6 ± 0.01	100 ± 0.0	20.4 ± 2.0	33.9 ± 2.8	99.6 ± 0.1	100.0 ± 0.0	24.2 ± 4.0	38.9 ± 5.0
lung infection	97.0 ± 0.1	90.6 ± 5.1	10.1 ± 1.7	18.2 ± 2.8	97.0 ± 0.1	94.4 ± 4.2	10.0 ± 0.7	18.1 ± 1.2

Table 3: Evaluation results using both image and EHR features (in percentages, mean ± SD reported)

Table 3 demonstrates that integrating EHR features with CXR image features leads to a significant breakthrough in model performance. Compared to using image features alone (Table 2), the inclusion of EHR data results in a dramatic surge in Precision and Recall for nearly all diseases, particularly for rare and imbalanced conditions. For instance, tuberculosis sees its Precision jump from a mere 6.7% to 96.2%, while scoliosis achieves a perfect Precision of 100%. Consequently, the F1-score increases substantially across the board, reflecting a much more robust balance between Precision and Recall. These results confirm that EHR data provide critical complementary information, effectively addressing the limitations of image-only models and significantly enhancing disease detection performance while maintaining high and stable Accuracy levels (mostly above 90%).

Regarding the performance changes between settings with and without data augmentation, with image features only, the application of SMOTE leads to a substantial increase in Recall and F1-score for most diseases, indicating a more balanced trade-off between Precision and Recall and a markedly improved ability to detect minority-class cases. In contrast, the effects on Accuracy and Precision remain relatively limited. When EHR features are incorporated, changes in Recall and F1-score are generally small or only slightly positive, suggesting that EHR data mainly contribute to stabilizing model performance rather than significantly

enhancing minority-class detection. Overall, these results (Table 4) highlight the critical role of data balancing in the image feature space for improving Recall and F1-score, while demonstrating that EHR features provide complementary information that helps maintain stable and consistent performance of the XGBoost model.

Disease	Image Features Only				Image with EHR Features			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
pneumonia	−12.2 ± 0.5	−3.0 ± 1.1	34.0 ± 0.8	21.8 ± 0.9	−0.3 ± 0.5	−1.1 ± 1.3	−0.6 ± 1.4	−0.7 ± 1.2
tuberculosis	−8.7 ± 0.6	−2.9 ± 14.9	13.1 ± 1.2	5.7 ± 0.4	0.0 ± 0.2	2.0 ± 3.5	−0.4 ± 2.6	−0.5 ± 3.6
bronchitis	−13.8 ± 0.7	−3.4 ± 3.4	21.6 ± 2.3	10.1 ± 0.9	0.0 ± 0.3	−1.6 ± 6.9	−0.7 ± 2.8	−1.2 ± 4.6
arthritis	−12.6 ± 0.5	−4.8 ± 22.4	18.0 ± 2.1	7.9 ± 0.9	0.0 ± 0.3	0.8 ± 3.7	0.0 ± 2.4	0.0 ± 3.5
fracture	−17.0 ± 0.4	−6.0 ± 4.6	24.4 ± 1.6	11.6 ± 0.8	0.0 ± 0.2	−0.6 ± 7.7	0.1 ± 1.1	0.2 ± 1.9
lung cancer	−12.9 ± 0.6	−5.3 ± 1.0	34.2 ± 1.3	22.1 ± 0.9	−0.1 ± 0.4	−0.2 ± 1.7	−0.4 ± 1.5	−0.4 ± 1.6
scoliosis	−1.7 ± 0.2	0.7 ± 0.5	2.0 ± 1.3	1.0 ± 0.7	0.0 ± 0.1	0.0 ± 0.0	3.8 ± 4.4	5.0 ± 5.7
lung infection	−10.1 ± 0.3	4.0 ± 0.5	13.4 ± 1.9	6.1 ± 0.8	0.0 ± 0.1	3.8 ± 6.7	−0.1 ± 1.8	−0.1 ± 3.0

Table 4: Performance differences between models with and without data augmentation (in percentages, mean ± SD reported)

Model 2: XGBClassifier + ROS.

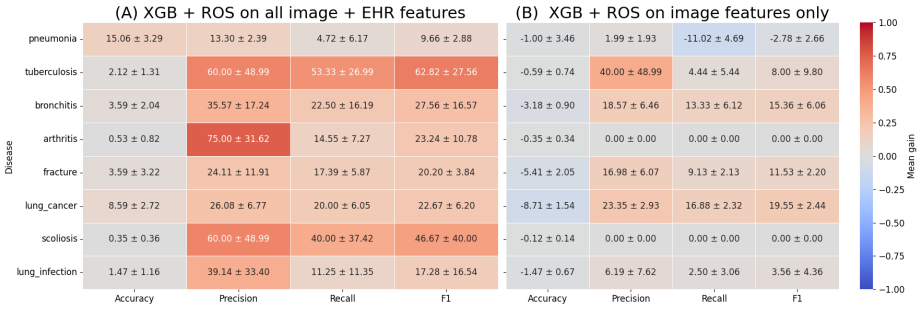

Figure 4: Mean performance gain against Logistic Regression on image features (in percentages, mean ± SD reported)

Incorporating all EHR features into XGB + ROS improves disease-detection performance for most of the eight conditions, with clear gains in Precision and Recall compared with the image-only setting. Moreover, overall Accuracy changes remain marginal, suggesting that the improvements mainly arise from better class-level discrimination rather than global prediction accuracy. Compared with using image-only features with logistic XGB + ROS leads to improvements in

Precision and Recall for several diseases; however, the gains are less consistent than those observed when EHR features are included. Performance remains limited for arthritis and scoliosis, and pneumonia exhibits a decrease in Recall.

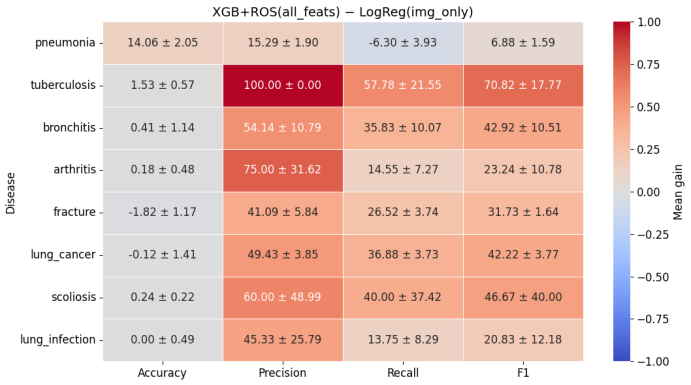

Figure 5: Performance gain of XGB+ROS with image and EHR features over Logistic Regression with image-only features (in percentages, mean ± SD reported)

Last but not least, the full pipeline (XGB + ROS with image+EHR) provides the most consistent Precision/Recall improvements over the baseline, with Accuracy largely unchanged; for pneumonia, the pattern suggests more conservative predictions that may benefit from threshold tuning when Recall is prioritized. Considering the reported standard deviations, tuberculosis, arthritis, and scoliosis show higher variability across folds, likely due to strong class imbalance and the small number of positives per split, which makes Precision/Recall more sensitive to sampling and oversampling randomness. Therefore, their mean gains should be interpreted together with the SD as an indicator of stability.

4 Discussion

4.1 Result Analysis

4.1.1 SMOTE with XGBClassifier/XGBRegressor

Experimental results indicate that when XGBoost relies solely on image features extracted from CXR, overall Accuracy remains relatively high, while Precision and Recall are low for severely imbalanced diseases. This suggests a strong bias toward negative predictions, a common limitation in medical machine learning tasks with rare positive cases, leading to missed diagnoses.

Applying SMOTE to image features substantially improves Recall across most diseases, particularly those with low positive prevalence. By generating synthetic minority samples, the model better captures representative patterns of rare conditions, reducing negative-class dominance. Although Precision slightly decreases in some cases, the Recall improvement highlights the benefit of SMOTE in enhancing disease detection, which is critical in clinical diagnosis.

4.1.2 ROS + XGBClassifier

Heatmap results show that combining ROS with XGBClassifier significantly

enhances positive-class detection when EHR features are incorporated. Compared to image-only settings, adding EHR leads to notable gains in Precision and Recall for several rare diseases such as tuberculosis, scoliosis, and arthritis, indicating the importance of structured clinical variables in reducing noise and improving discriminative power.

When using image features alone, XGB + ROS does not consistently outperform Logistic Regression in terms of Accuracy, but it often achieves higher Precision and Recall for specific diseases. This reflects a typical trade-off: stronger nonlinear models with oversampling become more sensitive to positive cases, improving disease detection at the expense of overall Accuracy.

4.2 Limitations and Recommendations

4.2.1 SMOTE with XGBClassifier/XGBRegressor

Although SMOTE was applied to augmented CXR image features and combined with EHR for training XGBoost models, the overall performance gain was marginal. Precision, Recall, and Accuracy with EHR remained largely unchanged compared to the non-SMOTE setting, indicating limited benefit from this augmentation strategy.

This suggests that image-based oversampling alone cannot compensate for sparsity and limited information in EHR data. While SMOTE increases the number of training samples, the synthetic image features do not introduce new clinical information to EHR, preventing the model from learning more meaningful multimodal relationships. Moreover, the Recall improvement observed in image-only settings often came at the cost of reduced Precision, indicating over-positive predictions, which was not effectively mitigated when EHR was included.

4.2.2 ROS with XGBClassifier

ROS improves class balance by duplicating existing positive samples rather than generating new information, which may lead to overfitting, especially for extremely rare diseases such as *tuberculosis* and *scoliosis*. As a result, performance gains are not uniform across diseases, and Accuracy does not always improve alongside Precision and Recall.

When using image features only, XGB + ROS does not consistently outperform Logistic Regression in terms of Accuracy and may even degrade performance in some cases. This indicates that oversampling in the image feature space can increase sensitivity to positive cases while reducing prediction stability on the negative class. For certain diseases (e.g., *pneumonia*), Recall may even decrease, suggesting that ROS is not universally beneficial.

With the inclusion of EHR features, substantial gains in Precision and Recall are observed for several rare diseases; however, these improvements strongly depend on EHR quality and completeness. To address these issues, our future work plan to (i) evaluate model stability using multiple random seeds or cross-validation, (ii) apply stronger regularization and hyperparameter tuning to mitigate overfitting induced by ROS, and (iii) optimize decision thresholds based on clinical objectives rather than relying on default settings.

5 Conclusion

This study presents a multimodal framework for CRD diagnosis that combines CXR image features and EHR data with class imbalance handling using SMOTE and ROS. Experimental results show that SMOTE improves Recall for rare diseases when applied to image features, while ROS with XGBClassifier provides more consistent gains in Precision and Recall, especially in the multimodal setting. Overall, the XGB + ROS + EHR pipeline achieves more balanced performance and better detection of positive cases, indicating strong potential despite data and validation limitations.

6 Acknowledgement

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, for supporting this study.

Reference

1. Anthony L. Byrne, Ben J. Marais, Carole D. Mitnick, Leonid Lecca, and Guy B. Marks. Tuberculosis and chronic respiratory disease: a systematic review. *International Journal of Infectious Diseases*, 32:138–146, March 2015.
2. Joan B. Soriano, Parkes J. Kendrick, Katherine R. Paulson, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 2020.
3. Anthony Chapron, Emilie Andres, Laure Fiquet, Fabienne Pelé, Emmanuel Allory, Estelle Le Pabic, Aurélie Veislinger, Lisa Le Guillou, Stéphanie Guillot, Bruno Laviolle, and Stéphane Jouneau. Early detection of chronic obstructive pulmonary disease in primary care: a randomised controlled trial. *British Journal of General Practice*, 73(737):e876–e884, July 2023.
4. Liqa A. Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with covid-19 pneumonia. *BMC Pulmonary Medicine*, 20(1), September 2020.
5. R.M Hopstaken, T Witbraad, J.M.A van Engelshoven, and G.J Dinant. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. *Clinical Radiology*, 59(8):743–752, August 2004.
6. Adnane Ait Nasser and Moulay A. Akhloufi. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics*, 13(1):159, January 2023.
7. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 618–626. IEEE, October 2017.
8. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.
9. Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019.
10. Tuan-Khoi Nguyen, Hoang-Anh Ngo, Clemence Mottez, Xiaochen Hou, Ivy Liang, Qichi Liang, and Thu-Anh Nguyen. Enhancing accuracy of chest x-ray diagnosis for

respiratory diseases with transfer learning. *Respirology*, 29(S3):212–213, November 2024.
11. Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2023.
12. Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), January 2023.
13. Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), December 2019.
14. Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019.
15. Daniel Aletaha and Josef S. Smolen. Diagnosis and management of rheumatoid arthritis. *JAMA*, 320(13):1360, October 2018.
16. Guillermo Stegen, Kenneth Jones, and Patricio Kaplan. CRITERIA FOR GUIDANCE IN THE DIAGNOSIS OF TUBERCULOSIS. *Pediatrics*, 43(2):260–263, February 1969.
17. Joseph A Janicki and Benjamin Alman. Scoliosis: Review of diagnosis and treatment. *Paediatrics & Child Health*, 12(9):771–776, November 2007.
18. Dawn E. Jaroszewski, Brandon J. Webb, and Kevin O. Leslie. Diagnosis and management of lung infections. *Thoracic Surgery Clinics*, 22(3):301–324, August 2012.
19. Samuel N. Grief and Julie K. Loza. Guidelines for the evaluation and treatment of pneumonia. *Primary Care: Clinics in Office Practice*, 45(3):485–503, September 2018.
20. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.
21. Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.
22. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.
23. 2. over-sampling — version 0.14.1. https://imbalanced-learn.org/stable/over_sampling.html. Accessed: 2026-1-5.
24. Cynthia Yang, Egill A Fridgeirsson, Jan A Kors, Jenna M Reps, and Peter R Rijnbeek. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J. Big Data*, 11(1), January 2024.