A Comprehensive Analysis of Advanced Deep Learning Techniques for Mitigating Class Imbalance in Medical Image Classification

Section 1: Introduction: The Critical Challenge of Class Imbalance in Medical AI

1.1 Defining Class Imbalance in the Medical Context

In the domain of artificial intelligence (AI) for medical diagnostics, the development of robust and reliable classification models is paramount. A significant and pervasive obstacle to achieving this goal is the phenomenon of class imbalance, a condition where the distribution of classes within a training dataset is highly skewed.[1] This issue is particularly pronounced in medical imaging datasets, where images corresponding to "normal" or healthy cases vastly outnumber those depicting "abnormal" or pathological conditions.[1] This disparity is not a statistical curiosity but a direct reflection of clinical reality; rare diseases, by definition, have low prevalence in the general population, and their representation in medical archives is correspondingly sparse.[5]

The core technical problem arises from the optimization behavior of standard machine learning algorithms. Most deep learning models are trained to minimize a loss function that is aggregated over all training samples, with the implicit assumption that each sample contributes equally to the learning process. When one class dominates the dataset, the model can achieve a low overall loss—and thus high accuracy—simply by developing a strong bias towards predicting the majority class.[3] This optimization strategy, while mathematically sound for balanced problems, is clinically disastrous in the medical context. The model effectively learns to ignore the minority class, which often represents the very conditions of greatest clinical interest.[5]

The severity of this challenge is not uniform; it exists on a spectrum. A moderately imbalanced scenario might involve a condition present in 5% of cases, while an extremely imbalanced one could involve a rare pathology with a prevalence of less than 0.5%. As the abstract for this research outlines, traditional deep learning approaches often struggle to address this spectrum of diagnostic challenges effectively, necessitating the development of advanced, specialized techniques.[8] The ethical and clinical stakes are exceptionally high, as the cost of misclassifying a critical but rare condition (a false negative) is profoundly greater than the cost of incorrectly flagging a healthy case for further review (a false positive).[1]

1.2 Clinical Significance and Diagnostic Implications

The consequences of class imbalance extend far beyond poor statistical performance; they have direct and severe implications for patient care and clinical outcomes. An AI diagnostic system trained on an imbalanced dataset may appear highly accurate on paper but fail catastrophically in a clinical setting by consistently missing the detection of rare but critical diseases.[8] This failure mode undermines the fundamental purpose of medical AI: to augment human expertise, improve diagnostic accuracy, and ultimately enhance patient safety.

Consider a deep learning model trained to detect various thoracic diseases from chest X-rays. If a rare but life-threatening condition like a hernia is present in only 0.2% of the training images, a standard classifier will have very few examples from which to learn its distinguishing features. The overwhelming gradient signals from the 99.8% of non-hernia cases will dominate the training process, leading to a model that is an excellent detector of "not a hernia" but is functionally blind to the actual presence of one. A patient with this critical condition could receive a false "all-clear" from the AI system, leading to delayed diagnosis and potentially fatal consequences.

This elevates the problem of class imbalance from a technical hurdle to a critical issue of **clinical risk management**. The objective is not merely to "balance" a dataset for statistical neatness but to engineer a model whose predictive behavior aligns with clinical priorities and mitigates the risk of the most harmful diagnostic errors. The evaluation of any proposed solution must therefore be viewed through this lens: how effectively does it reduce the probability of missing a critical, low-prevalence disease? This perspective reframes the goal, shifting the focus from maximizing overall accuracy to selectively maximizing sensitivity for the conditions where it matters most.

1.3 A Taxonomy of Mitigation Strategies

To address the multifaceted challenge of class imbalance, researchers have developed a range of mitigation strategies that can be broadly categorized into three families. This taxonomy provides a structured framework for understanding and applying the advanced techniques that will be detailed in this report.[12]

1. **Data-Level Solutions:** These methods focus on modifying the training dataset itself to create a more balanced class distribution before it is presented to the model. The goal is to rebalance the data to reduce the inherent bias of the learning algorithm. This category includes:
   - **Oversampling:** Increasing the number of instances in the minority class. This can be done through simple duplication (random oversampling) or, more effectively, through the generation of new, synthetic data points, as exemplified by the Synthetic Minority Over-sampling Technique (SMOTE).[2]
   - **Undersampling:** Decreasing the number of instances in the majority class by randomly removing samples. While simple, this approach risks discarding potentially useful information.[2]
   - **Data Augmentation:** Artificially expanding the dataset by creating modified versions of existing images (e.g., through rotation, scaling, or brightness adjustments). When applied preferentially to the minority class, this serves as a powerful and effective oversampling technique.[14]
2. **Algorithm-Level Solutions:** These methods modify the learning algorithm's objective function or training process to give more importance to the minority class, without altering the underlying data distribution. This category includes:
   - **Cost-Sensitive Learning:** Assigning a higher misclassification cost (penalty) to errors made on the minority class. This forces the model to pay more attention to correctly classifying these rare instances.[12]
   - **Specialized Loss Functions:** Designing novel loss functions that inherently counteract the effects of imbalance. The most prominent example is Focal Loss, which dynamically adjusts the loss contribution of each sample based on how well it is classified, thereby focusing the training on harder-to-learn examples.[6]
3. **Hybrid Solutions:** These approaches combine data-level and algorithm-level strategies to leverage the strengths of both. For instance, one might first use data augmentation and SMOTE to create a more balanced and diverse dataset, and then train a model on this new dataset using a cost-sensitive or focal loss function. Such hybrid methods often yield the most robust and superior performance.[12]

This report will systematically explore these strategies, evaluating their theoretical underpinnings and practical efficacy in the context of varying degrees of class imbalance in medical image classification.

Section 2: Foundational Components: Datasets and Architectures

Before delving into the specific techniques for mitigating class imbalance, it is essential to establish the foundational components of the analysis: the dataset that serves as the testbed for these methods and the deep learning architecture that provides the model framework. The choice of dataset is critical, as it must realistically represent the challenges encountered in clinical practice. Similarly, the choice of architecture is foundational, as it must possess the capacity to learn the complex visual features characteristic of medical images.

2.1 The NIH ChestX-ray14 Dataset: A Case Study in Real-World Imbalance

The NIH ChestX-ray14 dataset, initially released as ChestX-ray8, is a cornerstone resource for research in medical image analysis and serves as the primary dataset for this study.[8] It is a large-scale, publicly available collection of 112,120 frontal-view chest X-ray images from 30,805 unique patients, making it a powerful benchmark for developing and testing diagnostic AI models.[20]

2.1.1 Labeling Methodology and Its Implications

A defining characteristic of the ChestX-ray14 dataset is its labeling methodology. The 14 disease labels associated with the images were not generated through meticulous manual annotation by radiologists, which is a time-consuming and expensive process. Instead, they were extracted from the associated free-text radiological reports using Natural Language Processing (NLP) techniques.[17] This approach is a form of "weak supervision." While the creators estimate the label accuracy to be greater than 90%, this method inherently introduces a degree of label noise.[21]

This choice of dataset is not merely for its scale but for its **realism in imperfection**. The weakly supervised, NLP-derived labels create a dual challenge that is highly representative of real-world clinical data environments: the model must contend with **class imbalance *and* label noise** simultaneously. A successful framework must therefore be robust not only to infrequent data but also to potentially *incorrect* data. For instance, a model training to detect the rare "Hernia" class must learn its features from a small pool of examples, some of which may be mislabeled, while also learning to distinguish it from "No Finding" cases, some of which might contain an un-labeled hernia. A system that can succeed under these messy, practical conditions is inherently more likely to generalize to the imperfect data found in actual clinical workflows, making this dataset an excellent and challenging testbed.

2.1.2 Statistical Analysis of Pathologies

The dataset is labeled with 14 common thoracic pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, and Pneumothorax.[21] A crucial feature of the dataset is that images can have multiple labels, reflecting the clinical reality of co-existing diseases.[22]

The distribution of these pathologies exhibits a classic long-tail pattern, which is central to this investigation. As detailed in Table 1, the prevalence of these conditions varies dramatically, from common findings like Infiltration (17.7%) to extremely rare ones like Hernia (0.2%). A substantial portion of the dataset, approximately 53.8%, is labeled as "No Finding," representing the healthy majority class.[22] This quantitative evidence makes the abstract problem of "varying degrees of imbalance" tangible and provides the basis for the comparative study.

Table 1: Prevalence of All 14 Pathologies in the NIH ChestX-ray14 Dataset
Data compiled from a study by Li et al. which analyzed the full dataset of 112,120 images.22

Pathology	Absolute Count	Prevalence (%)
Infiltration	19,894	17.7%
Effusion	13,317	11.9%
Atelectasis	11,559	10.3%
Nodule	6,331	5.6%

Mass	5,782	5.2%
Pneumothorax	**5,302**	**4.7%**
Consolidation	4,667	4.2%
Pleural_Thickening	3,385	3.0%
Cardiomegaly	2,776	2.5%
Emphysema	2,516	2.2%
Edema	2,303	2.1%
Fibrosis	1,686	1.5%
Pneumonia	1,431	1.3%
Hernia	**227**	**0.2%**

2.1.3 Case Study Pathologies

As specified in the research abstract, this report will focus on two pathologies from this dataset to represent the spectrum of imbalance.[8] Table 1 clearly situates these choices within the broader context:

- **Pneumothorax:** With 5,302 cases and a prevalence of approximately 4.7%, this condition serves as an exemplar of **moderate imbalance**. While underrepresented, there is a substantial number of examples from which a model can learn.
- **Hernia:** With only 227 cases and a prevalence of just 0.2%, this condition represents a case of **extreme imbalance**. This poses a significant challenge, as the model has a very small and potentially non-representative sample to learn from, making it highly susceptible to overfitting and memorization.

By analyzing techniques against both of these cases, it becomes possible to draw nuanced conclusions about which strategies are most effective at different points along the imbalance spectrum.

2.2 Convolutional Neural Networks (CNNs) in Medical Imaging

Convolutional Neural Networks (CNNs) are the de facto standard for image analysis tasks, having demonstrated state-of-the-art performance in medical image classification, segmentation, and detection.[28] Their architecture is specifically designed to process grid-like data such as images, leveraging properties like spatial hierarchies of features.

2.2.1 Architectural Principles

A CNN is composed of a sequence of specialized layers that progressively extract more complex features from an input image [28]:

- **Convolutional Layers:** These are the core building blocks of a CNN. They apply a set of learnable filters (or kernels) to the input image. Each filter is a small matrix of weights that slides across the image, computing dot products to create a feature map. These filters learn to detect specific low-level features, such as edges, corners, and textures.[28]
- **Activation Functions (ReLU):** After each convolution, an activation function is applied to introduce non-linearity into the model. The Rectified Linear Unit (ReLU), which outputs the input directly if it is positive and zero otherwise, is the most common choice. This non-linearity is crucial for allowing the network to learn complex, non-linear relationships between pixels.[28]
- **Pooling Layers:** These layers are used to reduce the spatial dimensions (width and height) of the feature maps. Max Pooling is a common technique where a filter slides over the feature map and outputs only the maximum value within its receptive field. Pooling makes the feature representations more compact, reduces computational load, and provides a degree of translation invariance, meaning the network becomes more robust to the exact position of a feature in the image.[28]
- **Fully Connected Layers:** After several stages of convolution and pooling, the high-level feature maps are flattened into a one-dimensional vector. This vector is then fed into one or more fully connected layers, which are standard neural network layers where each neuron is connected to all neurons in the previous layer. These layers perform the final classification task by learning to map the extracted features to the output classes.[28]

2.2.2 The Challenge of Depth and Deep Residual Learning (ResNet)

A common intuition in deep learning is that deeper networks (with more layers) have a greater capacity to learn and should therefore perform better. However, early attempts to simply stack more layers onto CNNs encountered a "degradation problem": as the network depth increased, accuracy would first saturate and then rapidly degrade. Counter-intuitively, a deeper "plain" network could have a higher training error than its shallower counterpart, indicating that the deeper model was harder to optimize.[33]

The Deep Residual Learning framework, or ResNet, introduced by He et al. (2016), provided an elegant solution to this problem and is a key reference in the user's abstract.[8] The core innovation of ResNet is the "residual block," which features a "shortcut connection" or "skip connection."

Instead of forcing a set of layers to learn a desired underlying mapping $H(x)$, ResNet reformulates the problem. It lets the layers learn a *residual mapping* $F(x)=H(x)-x$. The original mapping is then recast as $F(x)+x$. This is implemented via a shortcut connection that bypasses the layers and adds the original input x to the output of the layers $F(x)$. The hypothesis, which was proven empirically, is that it is easier for the network to learn to push the residual $F(x)$ towards zero (if an identity mapping is optimal) than it is to learn an identity mapping from scratch with a stack of non-linear layers.[33]

This framework allows for the training of networks that are substantially deeper—up to 152 layers in the original paper—than was previously feasible. This increased depth is not merely an architectural feat; it is synergistic with the challenge of class imbalance. Deeper networks have a higher representational capacity, enabling them to learn a more complex hierarchy of features. This is essential for distinguishing rare pathologies that may be defined by extremely subtle and nuanced visual cues, such as faint texture changes or small structural anomalies that a shallower network might overlook. The ability of ResNet to effectively leverage depth provides the necessary model capacity to capture the faint signals of minority classes, making it a foundational enabler for the imbalance-specific techniques that follow.

Section 3: Algorithmic Interventions: Specialized Loss Functions and Cost-Sensitive Learning

While data-level strategies modify the input to the model, algorithm-level interventions alter the learning process itself. These techniques adjust the model's objective function to explicitly counteract the biasing effects of an imbalanced class distribution. By changing what the model is optimized to do, these methods can compel it to pay greater attention to the underrepresented minority class. This section details two primary algorithmic approaches: the use of a specialized loss function, Focal Loss, and the broader framework of cost-sensitive learning.

3.1 Focal Loss: An In-Depth Analysis

The standard loss function for multi-label classification tasks is the Binary Cross-Entropy (BCE) loss. While effective for balanced datasets, BCE has a significant drawback in imbalanced scenarios. The total loss is a sum of the loss for each sample. In a dataset with a 100:1 imbalance ratio, the loss contribution from the numerous "easy" negative examples (the well-classified majority class) can overwhelm the loss from the few "hard" positive examples (the minority class). Even if the loss for each majority class sample is small, their sheer volume dominates the gradient updates, preventing the model from learning the features of the minority class effectively.[6]

3.1.1 Introduction and Mathematical Formulation

To address this, Lin et al. (2017) introduced the Focal Loss, a dynamically scaled cross-entropy loss designed to down-weight the contribution of easy, well-classified examples.[8] This allows the training process to focus on a sparse set of hard examples, which are more informative for learning. The Focal Loss is defined as:

$FL(pt)=-\alpha t(1-pt)\gamma \log(pt)$
where:

- pt is the model's estimated probability for the ground-truth class. For a positive sample, $pt=p$; for a negative sample, $pt=1-p$.
- $(1-pt)\gamma$ is the **modulating factor**, the core innovation of Focal Loss.[6]
- $\gamma \geq 0$ is a tunable **focusing parameter**.
- αt is a **balancing parameter**, similar to that used in cost-sensitive learning.

The modulating factor is what gives Focal Loss its power. When an example is easily classified (i.e., pt is close to 1), the modulating factor $(1-pt)\gamma$ approaches 0, drastically reducing that sample's contribution to the total loss. Conversely, when an example is misclassified or classified with low confidence (pt is close to 0), the modulating factor is near 1, and the loss is unaffected. This mechanism effectively filters out the noise from easy examples and focuses the model's attention on the hard ones.[6]

The focusing parameter, γ, controls the rate of this down-weighting. As γ increases, the effect of the modulating factor becomes more pronounced, increasing the focus on hard examples. The original paper found that $\gamma=2$ worked well in practice.[6] The balancing parameter,

αt, is a static weight (often set as the inverse class frequency) that addresses the raw numerical imbalance between classes, similar to standard cost-sensitive approaches.[6]

3.1.2 Focal Loss as an Adaptive Learning Regularizer

It is crucial to understand that Focal Loss is more than a simple re-weighting scheme; it functions as an **adaptive learning regularizer**. Standard cost-sensitive learning applies a *static* penalty to all samples of a minority class, regardless of whether an individual sample is easy or hard to classify. Focal Loss, by contrast, applies a *dynamic* penalty based on the model's own confidence (pt). A minority class sample that is very obvious and easily classified will have its loss down-weighted, just like an easy majority class sample.

This adaptive nature is particularly valuable when dealing with datasets that have label noise, such as the NIH ChestX-ray14 dataset.[41] Consider two scenarios:

1. An incorrectly labeled majority class sample (e.g., a "No Finding" image that actually contains a pathology). A standard model would treat this as an "easy negative," but because it is mislabeled, the model will struggle, resulting in a low pt. Focal Loss will identify this as a "hard negative" and maintain its high loss contribution, encouraging the model to learn from it.
2. A correctly labeled minority class sample that has very strong, unambiguous visual features. This would be an "easy positive." A simple cost-sensitive approach would still apply a large weight to its loss, potentially leading to overfitting. Focal Loss, however, would see the high pt and down-weight its loss, allowing the model to focus on more ambiguous minority cases.

By keying its behavior to model confidence rather than just class membership, Focal Loss is better equipped to navigate the complexities of imperfect, real-world data. It provides an efficient, built-in mechanism for "hard example mining," focusing training on the truly informative samples—be they rare positives or mislabeled negatives—making it a more nuanced and robust tool.[16] Recent research has even explored adaptive versions of Focal Loss where the

γ parameter is dynamically adjusted during training to further optimize this process.[42]

3.2 Cost-Sensitive Learning Frameworks

Cost-sensitive learning is a broader and more established framework that directly addresses the problem of unequal misclassification costs.[15] The fundamental principle is to modify the learning algorithm such that it minimizes a cost-based metric rather than a simple error count. In the context of medical diagnosis, this means assigning a higher penalty for a false negative (missing a disease) than for a false positive (a false alarm).

The most common method for implementing cost-sensitive learning in deep neural networks is through the use of **class weights** in the loss function.[15] For a binary classification problem, the weighted cross-entropy loss can be expressed as:

$L=-[wp\cdot y\log(p)+wn\cdot(1-y)\log(1-p)]$
where y is the true label, p is the predicted probability, and wp and wn are the weights assigned to the positive and negative classes, respectively. To counteract imbalance, the weight for the minority class (wp) is set to be higher than the weight for the majority class (wn). A common heuristic is to set the weights to be inversely proportional to the class frequencies in the training data.[15] For example, in a dataset with a 99:1 ratio of negative to positive samples, one might set

$wn=1$ and $wp=99$.

This approach directly increases the magnitude of the loss for any error made on a minority class sample, forcing the model's optimization process to prioritize learning features that can correctly identify this class. While conceptually similar to the α parameter in Focal Loss, cost-sensitive learning is a more direct and less complex mechanism.

The choice between simple cost-sensitive learning and the more complex Focal Loss represents a trade-off between **simplicity and adaptability**. Cost-sensitive learning is straightforward to implement and has only one primary hyperparameter to tune per class (the weight). For problems with moderate imbalance and high-quality, clean labels, this simpler approach may be sufficient and easier to optimize. However, for problems characterized by extreme imbalance and/or significant label noise, as is the case with the Hernia pathology in the NIH dataset, the adaptive nature of Focal Loss is likely superior. Its ability to dynamically modulate the loss based on sample difficulty provides a more granular control over the learning process, which is necessary to prevent the model from overfitting to the few, and potentially noisy, minority samples. The selection between these methods should be considered a deliberate design choice based on the specific characteristics of the dataset and the diagnostic task.

Section 4: Data-Level Strategies: Augmentation and Synthetic Data Generation

In contrast to algorithmic interventions that modify the model's learning process, data-level strategies focus on transforming the training data itself. The goal is to create a more balanced and diverse dataset that mitigates the inherent biases of standard learning algorithms. This section explores two principal data-level techniques: the generation of synthetic data points using the SMOTE algorithm and the creation of plausible variations of existing images through data augmentation.

4.1 The SMOTE Algorithm and its Variants

Simple oversampling of the minority class by duplicating existing samples can lead to overfitting, as the model may simply memorize the repeated examples without learning to generalize.[5] The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. (2002), offers a more sophisticated solution by creating new,

synthetic samples that populate the feature space of the minority class.[8]

4.1.1 Algorithmic Steps

The SMOTE algorithm operates not in the raw data space (e.g., pixel space) but in the *feature space*. It generates new samples by interpolating between existing minority class instances.[5] The process is as follows [47]:

1. **Select a Minority Instance:** Randomly choose a sample from the minority class.
2. **Find Nearest Neighbors:** Identify its *k* nearest neighbors (typically *k*=5) that also belong to the minority class. The distance is usually calculated in the feature space using Euclidean distance.
3. **Choose a Neighbor:** Randomly select one of these *k* neighbors.
4. **Generate Synthetic Sample:** Create a new synthetic data point at a random location along the line segment connecting the original sample and its chosen neighbor in the feature space. This is done by taking the vector difference between the two samples,

multiplying it by a random number between 0 and 1, and adding the result to the original sample's feature vector.

This procedure is repeated until the desired number of synthetic minority samples has been generated, creating a more balanced dataset. By creating synthetic examples along the lines connecting existing ones, SMOTE effectively expands and densifies the decision region for the minority class, encouraging the classifier to learn broader, more generalizable boundaries.[47]

4.1.2 Application to Image Data and Limitations

A critical consideration is that SMOTE is designed to operate on feature vectors, not directly on raw images.[49] Applying interpolation directly to the pixel values of two different medical images would likely result in a nonsensical, blurry artifact that does not represent a plausible anatomical structure. Therefore, for image classification tasks, a multi-step approach is required [50]:

1. **Feature Extraction:** First, a powerful feature extractor, such as a pre-trained CNN (e.g., the convolutional base of a ResNet), is used to convert each image into a high-level, lower-dimensional feature vector.
2. **SMOTE in Feature Space:** The SMOTE algorithm is then applied to these feature vectors. This creates new, synthetic *feature vectors* for the minority class.
3. **Classifier Training:** The final classifier (e.g., the fully connected layers of the network) is then trained on the balanced set of real and synthetic feature vectors.

While powerful, standard SMOTE has limitations. It generates samples without regard to the location of majority class instances. If the minority class region is noisy or heavily overlaps with the majority class, SMOTE can generate synthetic samples that fall within the majority class region, effectively creating label noise and potentially harming classifier performance.[7] This has led to the development of several variants designed to be more discerning:

- **Borderline-SMOTE:** This variant focuses on oversampling only the minority instances that are on the "borderline" (i.e., have both majority and minority class neighbors), as these are considered the most critical for defining the class boundary.[7]
- **Adaptive Synthetic Sampling (ADASYN):** This method generates more synthetic data for minority examples that are harder to learn (i.e., have more majority class neighbors), adaptively shifting the decision boundary to focus on difficult regions.[5]

4.2 Geometric and Photometric Data Augmentation

Data augmentation is a widely used and highly effective technique for artificially increasing the size and diversity of a training dataset.[14] In the context of class imbalance, it is used as a form of intelligent oversampling for the minority class.[55] By applying a series of realistic transformations to the existing minority class images, a multitude of new, unique training examples can be generated, helping the model to learn features that are invariant to these transformations and improving its ability to generalize to unseen data.

4.2.1 Catalogue of Transformations

Augmentation techniques are typically divided into two categories [54]:

- **Geometric Transformations:** These modify the spatial properties of an image, teaching the model invariance to changes in position, orientation, and scale. Common geometric augmentations include [53]:
  - **Rotation:** Rotating the image by a random angle.
  - **Flipping:** Mirroring the image horizontally or vertically.
  - **Scaling:** Zooming in or out on the image.
  - **Translation:** Shifting the image horizontally or vertically.
  - **Shearing:** Skewing the image along an axis.
- **Photometric (Color) Transformations:** These alter the pixel values of an image, teaching the model robustness to variations in imaging equipment and conditions. Common photometric augmentations include [54]:
  - **Brightness Adjustment:** Randomly increasing or decreasing the overall brightness of the image.
  - **Contrast Adjustment:** Randomly changing the difference between light and dark areas.
  - **Hue/Saturation Shifts:** Modifying the color properties of the image (less common for grayscale medical images).
  - **Noise Injection:** Adding random Gaussian noise to simulate sensor noise or other imaging artifacts.

4.2.2 Best Practices in Medical Imaging

When applying data augmentation to medical images, it is imperative that the transformations are **clinically plausible**. An unrealistic augmentation can introduce harmful artifacts that teach the model incorrect features. For example, rotating a chest X-ray by 180 degrees creates an anatomically impossible orientation. Similarly, horizontal flipping may be inappropriate for pathologies where laterality (left vs. right) is a key diagnostic feature. The parameters of each transformation (e.g., the maximum rotation angle) must be carefully chosen to reflect realistic variations. It is also a fundamental principle that augmentation should *only* be applied to the training set. The validation and test sets must remain untouched to ensure an unbiased and accurate evaluation of the model's true performance on unseen data.[52]

There is a fundamental distinction in the type of information generated by these two data-level strategies. Data augmentation creates **plausible variations** of existing data points. A slightly rotated or brightened chest X-ray is still a valid and realistic chest X-ray. In contrast, SMOTE creates **interpolated averages** in a high-dimensional feature space. The resulting synthetic feature vector is a mathematical construct designed to densify the decision boundary; it may not correspond to any plausible real-world medical image.

This distinction has important implications. Augmentation is better suited for teaching a model robustness to perceptual variations like viewpoint, scale, and imaging conditions. SMOTE is better suited for filling gaps in a sparsely populated feature space, making the minority class decision boundary more convex and easier for a classifier to learn. For a case of extreme imbalance like Hernia, where the few samples may be very similar to one another, augmentation alone might only create slightly different versions of the same few examples, risking overfitting. SMOTE could be more powerful by creating novel feature combinations, but it is also riskier if the feature space is not well-behaved. This suggests that the most effective strategies may involve a hybrid approach, such as using an augmented set of minority samples as the input for the SMOTE algorithm.

Section 5: Framework for Empirical Validation: Experimental Design and Evaluation Metrics

To rigorously assess the efficacy of the various techniques for mitigating class imbalance, a well-designed experimental framework is essential. This framework must include a clear baseline for comparison, a methodology for isolating the impact of each technique, and, most importantly, a set of evaluation metrics that accurately reflect model performance on an imbalanced problem. Relying on inappropriate metrics can lead to misleading conclusions and the deployment of clinically ineffective models.

5.1 Designing a Comparative Study

A robust empirical validation process should be structured to allow for clear, unambiguous comparisons between different strategies.

5.1.1 Establishing a Baseline

The first and most critical step is to establish a baseline model. This involves training the chosen architecture (e.g., ResNet) on the original, imbalanced dataset using a standard loss function like Binary Cross-Entropy, with no special handling for class imbalance.[59] The performance of this baseline model serves as the reference point against which all other methods are measured. It quantifies the severity of the problem and provides a benchmark for improvement.

5.1.2 Isolating Variables and Data Splits

To understand the contribution of each technique, experiments should be designed to test them in isolation before exploring combinations. This means creating separate experimental arms for:

- Baseline + Cost-Sensitive Learning
- Baseline + Focal Loss
- Baseline + Data Augmentation (on minority class)
- Baseline + SMOTE

After evaluating each component individually, hybrid approaches (e.g., Augmentation + Focal Loss) can be tested to assess synergistic effects.

Throughout this process, the dataset must be partitioned correctly to avoid data leakage and ensure an unbiased evaluation. The split should be performed at the **patient level**, meaning all images from a single patient must belong to only one set (training, validation, or test). This prevents the model from being tested on images from a patient it has already seen during training. A standard split ratio for the NIH ChestX-ray14 dataset is 70% for training, 10% for validation, and 20% for testing.[22] The training set is used to fit the model parameters, the validation set is used for hyperparameter tuning (e.g., selecting the optimal γ for Focal Loss or the best class weights), and the test set is reserved for a final, one-time evaluation of the best-performing model's generalization ability.

5.2 Beyond Accuracy: Robust Evaluation for Imbalanced Classification

The single most common pitfall in evaluating models on imbalanced datasets is the misuse of the accuracy metric.

5.2.1 The Inadequacy of Accuracy

Accuracy, defined as the proportion of all correct predictions, is a deeply misleading metric for imbalanced problems. This is often referred to as the "accuracy paradox".[5] In the case of the Hernia pathology (0.2% prevalence), a naive model that simply predicts "No Hernia" for every single image would achieve an accuracy of 99.8%. While statistically impressive, this model is clinically useless as it has zero diagnostic power for the condition of interest. Therefore, it is imperative to use evaluation metrics that provide a more nuanced and clinically relevant picture of performance.

5.2.2 The Confusion Matrix and Key Metrics

All meaningful classification metrics are derived from the 2x2 **confusion matrix**, which tabulates the model's predictions against the true labels [30]:

- **True Positives (TP):** The model correctly predicts the positive class.
- **True Negatives (TN):** The model correctly predicts the negative class.
- **False Positives (FP):** The model incorrectly predicts the positive class (a "false alarm").
- **False Negatives (FN):** The model incorrectly predicts the negative class (a "miss").

From these four values, we can derive metrics that are robust to class imbalance. Table 2 provides a guide to the most important metrics for this context.

Table 2: Guide to Evaluation Metrics for Imbalanced Medical Classification
Metrics and interpretations compiled from multiple sources on model evaluation.13

Metric	Formula	Question Answered	Clinical Use Case / When to Prioritize
Accuracy	TP+TN+FP+FNTP+TN	What fraction of all predictions were correct?	Avoid as a primary metric for imbalanced datasets. Can be misleadingly high.
Precision	TP+FPTP	Of all the cases the model flagged as positive, what fraction were actually positive?	Prioritize when the cost of a false positive is high (e.g., to avoid unnecessary, invasive follow-up tests).
Recall (Sensitivity)	TP+FNTP	Of all the patients who actually have the disease, what fraction did we correctly identify?	Prioritize when the cost of a false negative is high (e.g., screening for life-threatening diseases).
Specificity	TN+FPTN	Of all the patients who are healthy, what fraction did we correctly identify?	Prioritize when confirming the absence of a disease is critical.
F1-Score	2·Precision+RecallPrecision·Recall	What is the harmonic mean of Precision and Recall?	A good general-purpose metric for balancing the trade-off between false positives and false negatives.

The choice of the primary evaluation metric should not be a purely technical decision but a **clinically motivated one**. For the case studies in this report, both Pneumothorax and Hernia are critical conditions where a missed diagnosis (a false negative) can have severe health consequences. Therefore, the most important metric for evaluating model performance on these pathologies is **Recall (Sensitivity)**. While a high F1-score is desirable as a balanced measure, a model with slightly lower precision but significantly higher recall would be clinically preferable. The analysis of results must be framed around this clinical priority.

5.2.3 AUC-ROC and Precision-Recall Curves

In addition to threshold-based metrics like Precision and Recall, it is valuable to assess a model's performance across all possible operating thresholds.

- **AUC-ROC Curve:** The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR = FP/(TN+FP)) at various thresholds. The Area Under the Curve (AUC) represents the probability that the model will assign a higher score to a randomly chosen positive instance than to a randomly chosen negative one. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents random guessing.[13]
- **Precision-Recall (PR) Curve:** For severely imbalanced datasets, the ROC curve can be overly optimistic. Because the number of True Negatives (TN) is enormous, even a large increase in False Positives (FP) may result in only a tiny increase in the FPR, making the curve appear close to ideal. The Precision-Recall (PR) curve, which plots Precision versus Recall, is often more informative in these scenarios.[13] Since both metrics in the PR curve focus on the positive class and are not influenced by the vast number of TNs, a drop in performance is much more visually apparent. For the extreme imbalance case of Hernia, the Area Under the Precision-Recall Curve (AU-PRC) should be considered a primary evaluation metric alongside AUC-ROC.

Section 6: Comparative Analysis: Tailoring Solutions to Imbalance Severity

The central hypothesis of this research is that the optimal strategy for addressing class imbalance is not universal but is contingent upon the severity of the imbalance. This section provides a comparative analysis of the techniques discussed, applying them to the two distinct case studies: the moderate imbalance of Pneumothorax and the extreme imbalance of Hernia. By synthesizing the properties of each technique with the challenges posed by each scenario, we can derive recommendations for tailoring solutions to specific diagnostic problems.

6.1 Analysis for Moderate Imbalance (Pneumothorax, ~4.7% prevalence)

With approximately 5,302 positive examples in the full dataset, the Pneumothorax class provides a substantial, albeit minority, set of samples for a model to learn from.

6.1.1 Expected Performance and Effective Techniques

- **Baseline Performance:** A standard ResNet model is expected to perform poorly on this class, likely achieving a low recall. The model's predictions will be biased towards the "No Finding" majority class, but it may not fail completely due to the presence of several thousand positive examples.
- **Algorithm-Level Solutions:** Both **Cost-Sensitive Learning** and **Focal Loss** are expected to be highly effective. By increasing the penalty for misclassifying Pneumothorax cases, these methods will force the model to shift its decision boundary, leading to a significant increase in recall, likely accompanied by a modest decrease in precision. The adaptive nature of Focal Loss may provide a slight advantage by preventing the model from focusing too much on "easy" Pneumothorax examples, leading to a better overall balance.
- **Data-Level Solutions:**
  - **Data Augmentation:** This is likely to be one of the most effective and lowest-risk strategies. Applying a range of clinically plausible geometric and photometric transformations to the ~5,000 Pneumothorax images can create a much larger and more diverse training set, substantially improving the model's robustness and generalization without introducing synthetic artifacts.
  - **SMOTE:** While potentially useful, SMOTE may be less critical in this scenario. The existing data provides a reasonably dense manifold in the feature space, so the risk of generating noisy samples is lower, but the benefit over simple augmentation may be marginal. A direct comparison between a model trained with augmentation alone versus one trained with SMOTE would be highly informative.

6.1.2 Hypothesized Best Approach

For a moderately imbalanced class like Pneumothorax, a hybrid approach combining a powerful data-level technique with a robust algorithm-level technique is hypothesized to be optimal. The most promising combination is the use of extensive **Data Augmentation** to enrich the minority class representation, coupled with **Focal Loss** to guide the training process adaptively. This approach leverages the benefits of increased data diversity while ensuring the model focuses its learning on the most challenging examples, leading to a model with high recall and a well-balanced F1-score.

6.2 Analysis for Extreme Imbalance (Hernia, 0.2% prevalence)

The Hernia class, with only 227 examples in the entire dataset, presents a far more formidable challenge. The primary risks are twofold: (1) the model may fail to learn any meaningful features at all, and (2) any learning that does occur may be highly susceptible to overfitting the tiny set of training examples.

6.2.1 Expected Performance and Necessary Techniques

- **Baseline Performance:** The baseline model is expected to fail completely, achieving a recall of zero for the Hernia class. It will learn to always predict "No Hernia," as this is the optimal strategy for minimizing the overall loss.
- **Data-Level Solutions are Critical:**
  - **Data Augmentation:** This is a necessary first step but is likely insufficient on its own. Augmenting only 227 images, while helpful, may still result in a set of examples that are too similar, leading the model to simply memorize them.[3]
  - **SMOTE:** This technique becomes much more crucial in the extreme imbalance scenario. The primary goal is to move beyond mere variations of existing data and to generate novel synthetic samples that can help the model form a more generalized decision boundary. SMOTE's ability to create new feature combinations is essential for expanding the representation of the Hernia class beyond the sparse cluster of original samples.
- **Algorithm-Level Solutions are Essential:**
  - **Focal Loss:** In this high-risk scenario, Focal Loss is strongly preferred over simple cost-sensitive learning. With so few positive examples, blindly up-weighting all of them with a static cost factor could cause the model to overfit aggressively to any noise or peculiarities within that small set. The adaptive nature of Focal Loss, which

modulates the loss based on model confidence, provides a more stable and nuanced learning signal, which is critical when training on a sparse and potentially fragile data distribution.

6.2.2 Hypothesized Best Approach

Addressing extreme imbalance requires an aggressive, multi-stage hybrid strategy. No single technique will suffice. A robust pipeline would likely involve:

1. **Heavy Data Augmentation:** First, apply a wide range of strong but plausible augmentations to the original 227 Hernia images to create a larger initial pool of minority examples.
2. **Feature-Space SMOTE:** Use a pre-trained feature extractor to convert the augmented Hernia images and a subset of negative images into feature vectors. Then, apply **SMOTE** (or a more robust variant like Borderline-SMOTE) in this feature space to generate a large and diverse set of synthetic minority feature vectors.
3. **Focal Loss Training:** Train the final classifier on the new, balanced dataset of real and synthetic feature vectors using **Focal Loss**. This ensures that the model learns effectively from the newly generated data while maintaining a focus on the most informative examples.

6.3 Discussion of Hybrid Approaches

The analysis of both scenarios underscores a central theme: the most powerful solutions for class imbalance are typically hybrid, combining complementary strategies. Data-level methods address the problem of data representation, while algorithm-level methods address the problem of model optimization. The former provides the model with a better dataset to learn from; the latter ensures the model learns from that dataset in an intelligent way.

The use of SMOTE, in particular, illustrates a critical **risk-reward trade-off** that varies with imbalance severity.

- **For moderate imbalance (Pneumothorax):** The risk of using SMOTE is relatively low, as the ~5,000 real samples provide a stable and representative manifold in the feature space for interpolation. However, the potential reward is also moderate, as extensive data augmentation might be sufficient on its own.
- **For extreme imbalance (Hernia):** The risk of using SMOTE is very high. The ~200 samples may not form a coherent manifold, and interpolation between them could

generate meaningless or noisy feature vectors that pollute the training process.[1] However, the potential reward is also extremely high, as it may be the *only* way to provide the model with enough diversity to learn a generalizable decision boundary.

This trade-off suggests that for extreme cases, the application of SMOTE should be done with care, potentially favoring more conservative variants like Borderline-SMOTE, and always validated rigorously to ensure that the synthetic data is beneficial rather than detrimental. Table 3 provides a comparative summary of the primary techniques discussed.

Table 3: Comparative Framework of Class Imbalance Techniques

Technique	Primary Mechanism	Pros	Cons	Best Suited For
Cost-Sensitive Learning	Statically increases the loss penalty for misclassifying minority class samples.	Simple to implement and interpret; directly addresses unequal misclassification costs.	Can lead to overfitting on noisy minority samples; requires careful tuning of class weights.[3]	Moderate imbalance with clean labels where a simple re-weighting is sufficient.
Focal Loss	Dynamically re-weights the loss based on model confidence to focus training on hard examples.	Adaptive and robust to label noise; provides automatic hard example mining; highly effective.[6]	More complex with two interacting hyperparameters (α, γ); may require more tuning.[3]	All levels of imbalance, especially with noisy labels or when a nuanced learning signal is required.
Data Augmentation	Creates plausible variations of existing images (e.g., rotation, scaling,	Low risk of adding harmful noise; improves model robustness and generalization.	Does not create truly novel data, only variations; may be insufficient for extreme	All levels of imbalance; a fundamental and highly recommended baseline strategy.

	brightness).	[54]	imbalance.[3]	
SMOTE	Generates new, synthetic minority samples by interpolating between existing ones in feature space.	Creates truly novel data points to expand the decision boundary; can be very effective.[5]	Can generate unrealistic or noisy samples, especially if classes overlap; computationally intensive.[3]	Extreme imbalance where new feature combinations are needed to learn a generalizable model.

Section 7: Conclusion: Synthesis, Current Challenges, and Future Trajectories

This report has systematically analyzed the critical challenge of class imbalance in medical image classification, a problem that poses a significant barrier to the clinical deployment of diagnostic AI systems. By examining the issue through the lens of the NIH ChestX-ray14 dataset and its varying degrees of pathological prevalence, a clear set of principles and strategies has emerged.

7.1 Summary of Key Findings and Recommendations

The central conclusion of this analysis is that there is no one-size-fits-all solution to class imbalance. The optimal strategy is context-dependent, requiring a careful consideration of the severity of the imbalance, the quality of the data labels, and the clinical implications of the diagnostic task.

- **For Moderate Imbalance (e.g., Pneumothorax):** A robust and effective approach is a hybrid strategy combining extensive **data augmentation** of the minority class with an advanced, algorithm-level technique like **Focal Loss**. This combination enhances data diversity while ensuring the model's learning process is adaptively focused on the most informative examples.
- **For Extreme Imbalance (e.g., Hernia):** A more aggressive, multi-stage hybrid approach is necessary. This should involve **heavy data augmentation** as a first step, followed by

feature-space SMOTE (or a conservative variant like Borderline-SMOTE) to generate novel synthetic samples, and finally, training with **Focal Loss** to provide a stable learning signal in a sparse data environment.
- **Evaluation is Paramount:** The use of clinically motivated evaluation metrics is non-negotiable. Accuracy is a dangerously misleading metric. Performance should be primarily assessed using **Recall (Sensitivity)** for critical conditions, balanced by the **F1-score** and visualized with both **AUC-ROC** and **Precision-Recall curves**, with the latter being particularly crucial for cases of extreme imbalance.

7.2 Persistent Challenges in Medical Image Classification

While the techniques discussed offer powerful solutions, they operate within a landscape of broader, persistent challenges that continue to shape the field of medical AI.

- **Data Scarcity and Privacy:** The fundamental bottleneck remains the difficulty in acquiring large, high-quality, and well-annotated medical datasets. Patient privacy regulations (such as HIPAA) and the high cost of expert annotation severely limit the data available for training, which is the root cause of many imbalance problems.[52]
- **Model Generalization and Domain Shift:** A model trained on data from one hospital's imaging equipment and patient population may fail to generalize to data from another institution. Differences in scanners, imaging protocols, and demographics create domain shifts that can significantly degrade performance.
- **Annotation Quality and Cost:** The field is heavily reliant on the manual annotation of data by clinical experts, a process that is both expensive and time-consuming. This has driven the use of weakly supervised methods like NLP-based labeling, but these introduce their own challenges of label noise and ambiguity, which can be exacerbated by imbalance-mitigation techniques.[41]
- **Interpretability and Trust:** A major barrier to the clinical adoption of deep learning models is their "black box" nature. Understanding *why* a model made a particular prediction is crucial for building trust with clinicians and for debugging model failures, yet remains a significant research challenge.

7.3 Future Research Directions and Recent Advancements (Post-2017)

The future of addressing class imbalance is moving beyond the corrective techniques detailed in this report and towards new learning paradigms that are inherently more robust to data limitations. The most promising future trajectories involve a shift from **post-hoc correction** to

data-efficient representation learning. Instead of asking, "How do we fix our imbalanced dataset?", emerging research asks, "How can we learn powerful, generalizable features from all available data in a way that makes the final classification of a rare class more tractable?"

Several key areas of research are driving this paradigm shift:

- **Advanced Generative Models:** While SMOTE generates synthetic data in a feature space, more powerful deep generative models like **Generative Adversarial Networks (GANs)** and **Diffusion Models** are now being used to create highly realistic, synthetic medical images.[65] These models can learn the underlying distribution of the minority class data and generate entirely new, high-fidelity images to augment training sets, offering a more powerful alternative to traditional augmentation and interpolation.
- **Self-Supervised and Semi-Supervised Learning:** These approaches aim to leverage the vast quantities of unlabeled medical data that are available. By designing pretext tasks (e.g., predicting a rotated version of an image or using contrastive learning to pull similar images together in a feature space), a model can learn rich and robust feature representations from unlabeled data. This pre-trained model can then be fine-tuned on a small, imbalanced labeled set, where the powerful learned features make the final classification task much easier.[67]
- **Federated Learning (FL):** To overcome data access and privacy barriers, federated learning offers a privacy-preserving framework to train a single, robust model on decentralized data from multiple institutions without ever centralizing the data itself. Recent advancements in FL specifically focus on developing algorithms that can handle class imbalance and non-IID (non-identically and independently distributed) data across different clinical sites, which is a more realistic representation of real-world medical data.[69]
- **Novel Architectures and Loss Functions:** Research continues to evolve in designing network architectures with built-in mechanisms, such as specialized attention modules or dual decoders, that are inherently better at focusing on minority class features.[67] Similarly, the development of new loss functions that generalize or improve upon Focal Loss continues to be an active area of investigation.[71]

In conclusion, while the current state-of-the-art provides a strong toolkit for managing class imbalance, the ultimate solution lies in developing learning paradigms that are fundamentally less dependent on the need for massive, perfectly balanced, and manually annotated datasets. The trajectory of the field points towards a future where AI can learn effectively from the messy, heterogeneous, and privacy-sensitive data that characterizes the real world of medicine, thereby unlocking its full potential to improve patient diagnosis and care.

Works cited

1. Addressing the Class Imbalance Problem in Medical Datasets - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/239608168_Addressing_the_Class_Imb

alance_Problem_in_Medical_Datasets

2. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging | Request PDF - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/344370433_Assessing_and_mitigating_the_effects_of_class_imbalance_in_machine_learning_with_application_to_X-ray_imaging
3. Handling Class Imbalance in Image Classification: Techniques and Best Practices - Medium, accessed September 1, 2025, https://medium.com/@okeshakarunarathne/handling-class-imbalance-in-image-classification-techniques-and-best-practices-c539214440b0
4. Unraveling the Impact of Class Imbalance on Deep-Learning ... - MDPI, accessed September 1, 2025, https://www.mdpi.com/2076-3417/14/8/3419
5. Smote for Imbalanced Classification with Python, Technique - Analytics Vidhya, accessed September 1, 2025, https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/
6. Focal loss | CloudFactory Computer Vision Wiki, accessed September 1, 2025, https://wiki.cloudfactory.com/docs/mp-wiki/loss/focal-loss
7. Anomaly Detection in Medical Images Using SMOTE Algorithm: A Comprehensive Approach - AI Publications, accessed September 1, 2025, https://aipublications.com/uploads/issue_files/2IJEEC-SEP20245-Anomaly.pdf
8. STC Abstract.pdf
9. SMOTE: Synthetic Minority Over-sampling Technique | Journal of ..., accessed September 1, 2025, https://www.jair.org/index.php/jair/article/view/10302
10. (PDF) SMOTE: Synthetic Minority Over-sampling Technique - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/220543125_SMOTE_Synthetic_Minority_Over-sampling_Technique
11. CLASS-WEIGHTED EVALUATION METRICS FOR IMBALANCED DATA CLASSIFICATION - OpenReview, accessed September 1, 2025, https://openreview.net/pdf?id=PBfaUXYZzU
12. A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11300732/
13. Best techniques and metrics for Imbalanced Dataset - Kaggle, accessed September 1, 2025, https://www.kaggle.com/code/marcinrutecki/best-techniques-and-metrics-for-imbalanced-dataset
14. How does data augmentation help with class imbalance? - Milvus, accessed September 1, 2025, https://milvus.io/ai-quick-reference/how-does-data-augmentation-help-with-class-imbalance
15. How to implement cost-sensitive learning in decision trees ..., accessed September 1, 2025,

https://www.geeksforgeeks.org/python/how-to-implement-cost-sensitive-learning-in-decision-trees/

16. Batch-balanced focal loss: a hybrid solution to class imbalance in deep learning - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10289178/
17. (PDF) ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/316736470_ChestX-ray8_Hospital-scale_Chest_X-ray_Database_and_Benchmarks_on_Weakly-Supervised_Classification_and_Localization_of_Common_Thorax_Diseases
18. 2.4. Benchmarking Resources - ML.recipes, accessed September 1, 2025, https://ml.recipes/resources/benchmarking.html
19. [PDF] ChestX-Ray8: Hospital-Scale Chest X-Ray Database and ..., accessed September 1, 2025, https://www.semanticscholar.org/paper/ChestX-Ray8%3A-Hospital-Scale-Chest-X-Ray-Database-on-Wang-Peng/58b6bd06ea58c367c64286126ba14128b45041b8
20. NIH Chest X-ray Dataset, accessed September 1, 2025, https://datasets.activeloop.ai/docs/ml/datasets/nih-chest-x-ray-dataset/
21. NIH Chest X-rays - Kaggle, accessed September 1, 2025, https://www.kaggle.com/datasets/nih-chest-xrays/data
22. Enhancing Multi-disease Diagnosis of Chest X-rays with Advanced ..., accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10054207/
23. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists - PubMed Central, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6245676/
24. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6476887/
25. anshuak100/NIH-Chest-X-ray-Dataset - GitHub, accessed September 1, 2025, https://github.com/anshuak100/NIH-Chest-X-ray-Dataset
26. Efficient Thorax Disease Classification and Localization Using DCNN and Chest X-ray Images - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10669971/
27. Manas2703/chest-xray-14 · Datasets at Hugging Face, accessed September 1, 2025, https://huggingface.co/datasets/Manas2703/chest-xray-14
28. Image Classification using CNN - GeeksforGeeks, accessed September 1, 2025, https://www.geeksforgeeks.org/machine-learning/image-classifier-using-cnn/
29. Image Classification Using CNN with Keras and CIFAR-10 - Analytics Vidhya, accessed September 1, 2025, https://www.analyticsvidhya.com/blog/2021/01/image-classification-using-convolutional-neural-networks-a-step-by-step-guide/
30. Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature - MDPI, accessed

September 1, 2025, https://www.mdpi.com/2504-4990/6/1/33

31. Convolutional neural networks in medical image understanding: a survey - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7778711/
32. Convolutional Neural Networks — Image Classification w. Keras - LearnDataSci, accessed September 1, 2025, https://www.learndatasci.com/tutorials/convolutional-neural-networks-image-classification/
33. Deep Residual Learning for Image Recognition - The Computer Vision Foundation, accessed September 1, 2025, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf
34. [1512.03385] Deep Residual Learning for Image Recognition - arXiv, accessed September 1, 2025, https://arxiv.org/abs/1512.03385
35. [PDF] Deep Residual Learning for Image Recognition | Semantic ..., accessed September 1, 2025, https://www.semanticscholar.org/paper/Deep-Residual-Learning-for-Image-Recognition-He-Zhang/2c03df8b48bf3fa39054345bafabfeff15bfd11d
36. (PDF) Focal Loss for Dense Object Detection (2017) | Tsung-Yi Lin | 20552 Citations, accessed September 1, 2025, https://scispace.com/papers/focal-loss-for-dense-object-detection-j0su4gk9as
37. Focal Loss for Dense Object Detection - CVF Open Access, accessed September 1, 2025, https://openaccess.thecvf.com/content_ICCV_2017/papers/Lin_Focal_Loss_for_ICCV_2017_paper.pdf
38. [PDF] Focal Loss for Dense Object Detection - Semantic Scholar, accessed September 1, 2025, https://www.semanticscholar.org/paper/Focal-Loss-for-Dense-Object-Detection-Lin-Goyal/1a857da1a8ce47b2aa185b91b5cb215ddef24de7
39. [PDF] Focal Loss for Dense Object Detection | Semantic Scholar, accessed September 1, 2025, https://www.semanticscholar.org/paper/Focal-Loss-for-Dense-Object-Detection-Lin-Goyal/79cfb51a51fc093f66aac8e858afe2e14d4a1f20
40. Focal Loss For Dense Object Detection | PDF | Robust Statistics - Scribd, accessed September 1, 2025, https://www.scribd.com/document/409373181/loss
41. Exploring the ChestXray14 dataset: problems - Lauren Oakden-Rayner, accessed September 1, 2025, https://laurenoakdenrayner.com/2017/12/18/the-chestxray14-dataset-problems/
42. An Enhanced Focal Loss Function to Mitigate Class Imbalance in Auto Insurance Fraud Detection with Explainable AI - arXiv, accessed September 1, 2025, https://arxiv.org/html/2508.02283v1
43. Enhancing Semantic Segmentation with Adaptive Focal Loss: A Novel Approach - arXiv, accessed September 1, 2025, https://arxiv.org/html/2407.09828v1
44. Cost-Sensitive Learning (CSL) - Machine Learning with Imbalanced Data - YouTube, accessed September 1, 2025, https://www.youtube.com/watch?v=LbhJ4gYoKxA

45. Understanding Cost Sensitivity in Imbalanced Classification | by Divyesh Bhatt - Medium, accessed September 1, 2025, https://medium.com/@dbhatt245/understanding-cost-sensitivity-in-imbalanced-classification-b51110e873d2
46. Research on expansion and classification of imbalanced data based on SMOTE algorithm, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8674253/
47. View of SMOTE: Synthetic Minority Over-sampling Technique, accessed September 1, 2025, https://www.jair.org/index.php/jair/article/view/10302/24590
48. A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning | Request PDF - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/221139351_A_Novel_Synthetic_Minority_Oversampling_Technique_for_Imbalanced_Data_Set_Learning
49. SMOTE for Imbalanced Classification with Python - MachineLearningMastery.com, accessed September 1, 2025, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
50. Enhancing Classification Accuracy for Imbalanced Image Data Using SMOTE - Medium, accessed September 1, 2025, https://medium.com/@fatimazahra.belharar/enhancing-classification-accuracy-for-imbalanced-image-data-using-smote-41737783a720
51. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare - PMC - PubMed Central, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10131309/
52. Analyzing Data Augmentation for Medical Images: A Case Study in Ultrasound Images, accessed September 1, 2025, https://arxiv.org/html/2403.09828v1
53. Data Augmentation Techniques Applied to Medical Images - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/382869099_Data_Augmentation_Techniques_Applied_to_Medical_Images
54. Data Augmentation in Computer Vision: Techniques & Examples - Lightly AI, accessed September 1, 2025, https://www.lightly.ai/blog/data-augmentation
55. Data Augmentation: A Class Imbalance Mitigative Measure, accessed September 1, 2025, https://blog.paperspace.com/data-augmentation-a-class-imbalance-mitigative-measure/
56. approaches and performance comparison with classical data augmentation methods - arXiv, accessed September 1, 2025, https://arxiv.org/html/2403.08352v3
57. [Literature Review] Enhancing Medical Image Analysis through ..., accessed September 1, 2025, https://www.themoonlight.io/en/review/enhancing-medical-image-analysis-through-geometric-and-photometric-transformations
58. Enhancing Medical Image Analysis through Geometric and Photometric transformations - arXiv, accessed September 1, 2025,

https://www.arxiv.org/pdf/2501.13643
59. Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification - arXiv, accessed September 1, 2025, https://arxiv.org/html/2505.06646v1
60. Class Imbalance and Evaluation Metrics for Medical Image Segmentation with Machine Learning Models - ResearchGate, accessed September 1, 2025, https://www.researchgate.net/publication/377782096_Class_Imbalance_and_Evaluation_Metrics_for_Medical_Image_Segmentation_with_Machine_Learning_Models
61. Classification: Accuracy, recall, precision, and related metrics ..., accessed September 1, 2025, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall
62. Understanding Model Evaluation Metrics for Image Classification - Akridata, accessed September 1, 2025, https://akridata.ai/blog/understanding-model-evaluation-metrics-for-image-classification/
63. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems - PubMed Central, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10688675/
64. Medical image data augmentation: techniques, comparisons and interpretations - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10027281/
65. Recent Advances in Medical Imaging Segmentation: A Survey - arXiv, accessed September 1, 2025, https://arxiv.org/html/2505.09274v1
66. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review - MDPI, accessed September 1, 2025, https://www.mdpi.com/2313-433X/9/4/81
67. CVPR Poster A Semantic Knowledge Complementarity based Decoupling Framework for Semi-supervised Class-imbalanced Medical Image Segmentation, accessed September 1, 2025, https://cvpr.thecvf.com/virtual/2025/poster/33219
68. CVPR Poster DyCON: Dynamic Uncertainty-aware Consistency and Contrastive Learning for Semi-supervised Medical Image Segmentation, accessed September 1, 2025, https://cvpr.thecvf.com/virtual/2025/poster/34987
69. FedIIC: Towards Robust Federated Learning for Class-Imbalanced Medical Image Classification | MICCAI 2023 - Accepted Papers, Reviews, Author Feedback, accessed September 1, 2025, https://conferences.miccai.org/2023/papers/267-Paper2902.html
70. Advances in Deep Learning-Based Medical Image Analysis - PMC, accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10880179/
71. Towards Reliable Healthcare Imaging: A Multifaceted Approach in ..., accessed September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12289783/