

Quadratic Loss-Based Optimization for Multiclass Image Classification in Neo Uro Light Vision

Nicholas Renaldo ^{a*}, Joko Pitoyo ^b, Jahrizal Jahrizal ^c, Fajri Marindra Siregar ^b, Nyoto Nyoto ^a, Rizaldi Putra ^a

^a Faculty of Business, Institut Bisnis dan Teknologi Pelita Indonesia, Indonesia

^b Faculty of Medicine, Universitas Riau, Indonesia

^c Faculty of Economics and Business, Universitas Riau, Indonesia

Article History

Received

9 June 2026

Received in revised form

19 July 2026

Accepted

26 August 2026

Published Online

30 September 2026

*Corresponding author

nicholasrenaldo@lecturer.pelitaindonesia.ac.id

Abstract

Artificial intelligence (AI)-based image classification offers potential support for preliminary medical screening. However, multiclass classification systems may exhibit uneven performance across categories, particularly when the dataset contains an imbalanced class distribution. This study proposes a quadratic loss-based optimization framework for multiclass image classification in Neo Uro Light Vision, an AI-based system designed to support preliminary screening of external genital health conditions in children. The study employs a quantitative experimental approach, using precision, recall, and F1-score to evaluate classification performance across ten classes. The initial evaluation involved 179 images and achieved a weighted precision of 89.77%, weighted recall of 88.27%, and weighted F1-score of 88.41%. Class-wise results revealed variations in performance, with the F1-score ranging from 72.73% for Penile_strangulation to 93.33% for Fistula. The dataset also exhibited substantial class imbalance, with 84 images in the Hipospadia class compared with only four in the Penile_strangulation class. Based on these findings, a quadratic objective is formulated to quantify deviations between class-specific F1-scores and predefined performance targets. The proposed framework is intended to support the systematic comparison of model configurations and the evaluation of class-wise performance. However, the effectiveness of the optimization approach has not yet been empirically established and requires comparative experiments and independent validation. This study provides a mathematical foundation for further investigating class-sensitive optimization in AI-based preliminary screening while emphasizing the distinction between computational classification performance and clinical diagnostic validity.

Keywords: Quadratic Loss; Multiclass Image Classification; Medical Image Analysis; Class Imbalance; Artificial Intelligence; Neo Uro Light Vision; F1-Score Optimization

DOI: <https://doi.org/10.35145/z5q4f997>

SDGs: Good Health and Well-Being (3); Industry, Innovation and Infrastructure (9); Reduced Inequalities (10); Quality Education (4)

1.0 INTRODUCTION

Artificial intelligence (AI) and computer vision have increasingly become important technologies for image-based classification and decision-support applications (Jena et al., 2021). In healthcare, image classification systems can assist in identifying visual patterns associated with different conditions and support preliminary screening. However, the performance of these systems cannot be adequately assessed using overall accuracy alone, particularly when a classification task involves multiple classes with different sample sizes and different error patterns. Therefore, the development of appropriate evaluation and optimization methods is important for improving classification performance, identifying weaknesses across individual classes, and supporting more reliable system evaluation.

Multiclass image classification presents several challenges, including variations in image characteristics, imbalanced class distributions, and differences in the ability of a model to distinguish between classes (Tarekegn et al., 2021). A model may achieve satisfactory aggregate performance while still exhibiting relatively weak precision or recall for particular classes. In screening applications, these differences deserve particular attention because false-positive and false-negative predictions have different implications. Consequently, model evaluation should consider multiple performance indicators, including precision, recall, and the F1-score, together with class-

level support and an analysis of classification errors. Such an approach provides a more comprehensive understanding of model performance than a single aggregate metric.

Neo Uro Light Vision is an AI-based image analysis system developed for preliminary screening of external genital health conditions in children. The system integrates digital image processing, computer vision (Renaldo et al., 2025), and AI-based classification to produce class-specific detection results and recommendations for further examination. According to the reported development results, the prototype can process uploaded images and provide detection outputs for multiple classes, including buried penis, chordee, epispadias, phimosis, fistula, hypospadias, micropenis, normal, paraphimosis, and penile strangulation. The system is intended to support preliminary screening rather than replace professional clinical assessment. Its continued development therefore requires systematic evaluation of classification performance, particularly for classes with limited evaluation samples.

Initial evaluation of Neo Uro Light Vision involved 179 images distributed across ten classes. The reported weighted precision was 0.897694, weighted recall was 0.882682, and weighted F1-score was 0.884068. Although these results provide an initial indication of classification performance, the class-level results reveal differences in predictive performance and sample support. For example, the Penile_strangulation class achieved a precision of 0.571429 and a recall of 1.000000 based on four samples, whereas the Hypospadias class achieved a recall of 0.916667 based on 84 samples. These differences demonstrate why aggregate metrics should be complemented by class-specific analysis. However, because the evaluation set is relatively limited and the results alone do not establish independent clinical validity, further testing is needed before drawing conclusions about generalization or clinical usefulness.

A further methodological consideration is the use of loss functions that reflect differences in classification performance across classes (Liu et al., 2026). Conventional classification objectives may not adequately address class imbalance, particularly when some classes contain substantially fewer samples than others. Recent work on medical image classification has investigated loss functions derived from metrics such as the F1-score and Matthew's correlation coefficient, reporting improvements in class-wise performance under imbalanced data conditions. These findings support the importance of evaluating optimization strategies beyond conventional aggregate accuracy.

In addition, medical AI systems require careful validation because performance measured on an internal dataset may not generalize to different data sources or acquisition conditions (Pitoyo et al., 2026). (Abbott et al., 2026) demonstrated that hidden data acquisition biases can substantially affect estimates of generalization across medical AI datasets. This issue is particularly relevant to Neo Uro Light Vision, where the current evaluation includes 179 images distributed unevenly across ten classes. Therefore, an optimization approach should not only seek to reduce classification error but also be evaluated using class-specific metrics and an independent test dataset.

Despite progress in medical image classification, there remains a practical need to investigate simple, interpretable optimization procedures that can be applied to existing AI screening prototypes without requiring a complete redesign of their underlying architectures (Kanyana et al., 2026). In this study, quadratic loss-based optimization is proposed as an approach for quantifying classification errors and investigating potential improvements in multiclass classification performance in Neo Uro Light Vision. The quadratic formulation penalizes larger deviations more strongly than smaller ones, providing a mathematical framework for comparing model configurations. Nevertheless, the effectiveness of the proposed approach must be established empirically, and a quadratic approximation should only be retained if it adequately represents the observed relationship between model configurations and classification performance.

Accordingly, this study aims to formulate and evaluate a quadratic loss-based optimization framework for multiclass image classification in Neo Uro Light Vision. The specific objectives are to: (1) formulate a quadratic loss function for evaluating class-wise classification errors; (2) examine differences in precision, recall, and F1-score across the ten classification classes; (3) investigate whether the proposed optimization approach improves classification performance compared with the baseline model; and (4) assess the consistency of the observed improvements using independent evaluation data. The expected contribution is a simple and interpretable mathematical framework for evaluating classification performance and guiding optimization in an AI-based screening prototype, while maintaining the distinction between computational classification results and clinical diagnosis.

2.0 LITERATURE REVIEW

2.1. Artificial Intelligence for Medical Image Classification

Artificial intelligence (AI), particularly deep learning, has become an important approach to medical image classification because it can learn visual patterns from image data and use those patterns to distinguish between multiple categories (Rayed et al., 2024). In medical screening applications, image classification systems may assist in identifying visual characteristics associated with different conditions and separating them from normal appearances. However, their usefulness depends not only on overall classification performance but also on their ability to recognize individual classes reliably.

In a multiclass screening system, each image is assigned to one of several predefined categories. The model must therefore learn features that distinguish visually similar classes while avoiding confusion between clinically different conditions. This challenge becomes more significant when some categories contain substantially fewer training examples than others. Moreover, high performance on a particular dataset does not necessarily guarantee that a model will perform equally well on new images collected under different conditions. Research on medical AI has highlighted the importance of evaluating generalization and identifying potential sources of performance degradation beyond the development dataset (Ross & McIntosh, 2024).

For Neo Uro Light Vision, the classification task involves ten categories: Buried_penis, Chordee, Epispadia, Fimosis, Fistula, Hipospadia, Micropenis, Normal, Parafimosis, and Penile_strangulation. The system is intended to support preliminary screening rather than replace professional medical examination. Accordingly, the classification model should be evaluated in terms of both overall performance and its ability to identify each category, especially those represented by relatively few images.

2.2. Multiclass Classification and Evaluation Metrics

The performance of a multiclass image classifier can be evaluated using precision, recall, and the F1-score (Kautz et al., 2017). These metrics provide complementary information about classification quality.

Precision measures the proportion of images predicted to belong to a particular class that actually belong to that class. Recall measures the proportion of actual images in a class that the model correctly identifies. The F1-score is the harmonic mean of precision and recall, balancing the two measures.

For class k , precision, recall, and F1-score can be expressed as:

$$P_k = \frac{TP_k}{TP_k + FP_k}$$

$$R_k = \frac{TP_k}{TP_k + FN_k}$$

$$F1_k = 2 \frac{P_k R_k}{P_k + R_k}$$

where TP_k denotes true positives, FP_k denotes false positives, and FN_k denotes false negatives for class k .

For multiclass problems, weighted-average metrics account for differences in the number of samples per class. Although useful for summarizing overall performance, weighted averages can be dominated by classes with larger support. Therefore, they should be interpreted together with class-level results and sample counts.

The initial Neo Uro Light Vision evaluation included 179 images across ten classes. Its reported weighted precision was 0.8977, weighted recall was 0.8827, and weighted F1-score was 0.8841. However, class-level results varied. For example, Hipospadia had 84 samples, whereas Penile_strangulation had only four. The reported precision for Penile_strangulation was 0.5714, despite a recall of 1.0000. This pattern illustrates why an overall score alone is insufficient to characterize the model's behavior across all categories.

These results are treated as baseline evaluation evidence. They do not, by themselves, demonstrate that an optimization method improves performance or establish clinical validity.

2.3. Class Imbalance in Medical Image Classification

Class imbalance occurs when the number of samples differs substantially across categories (Kaur & Singh, 2022). In multiclass medical image classification, this condition can cause a model to perform well on frequent classes while producing less reliable predictions for classes with limited training examples.

(Kocak et al., 2025) investigated imbalance-aware loss functions for medical image classification using loss functions derived from the Matthews correlation coefficient and F1-score. Their experiments showed that combining conventional cross-entropy loss with imbalance-aware objectives could improve balanced accuracy and class-wise F1-scores for minority classes. Their findings suggest that the choice of optimization objective can influence how effectively a model learns from imbalanced medical datasets.

This issue is relevant to Neo Uro Light Vision because its baseline dataset is unevenly distributed. Hipospadia accounts for 84 of the 179 images, while Penile_strangulation accounts for only four. Such differences may affect the stability of class-specific evaluation results. In particular, a metric calculated from only a few examples can change substantially when one prediction changes.

An imbalance-aware evaluation strategy should therefore consider class-specific precision, recall, and F1-score rather than relying exclusively on weighted-average metrics. Nevertheless, the existence of class imbalance does not automatically establish that it caused a particular classification error. Confirming that relationship requires further experiments, such as comparing models trained with different sampling strategies or loss functions.

2.4. Quadratic Loss and Optimization

Quadratic loss is a mathematical approach that penalizes the squared difference between an observed value and a target value (Mukherjee, 2012). For a target y_i and a model output $\hat{y}_i$, a basic quadratic loss can be written as:

$$L_{\text{quad}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

where N represents the number of observations. Squaring the error makes larger deviations contribute disproportionately to the total loss. This property makes quadratic objectives useful when the objective is to minimize deviations from specified targets.

For multiclass classification, however, quadratic loss can refer to different formulations. For example, mean squared error may be calculated between one-hot encoded class labels and predicted class probabilities. Alternatively, a quadratic objective can be defined over aggregate evaluation metrics, such as the deviation of each class's F1-score from a desired performance target. These formulations are not equivalent and should not be treated as interchangeable.

For the present study, a practical formulation is to use quadratic loss as a class-wise performance optimization objective. Let F_k represent the F1-score of class k , and let T_k denote the target F1-score for that class. The proposed objective is:

$$J(\theta) = \sum_{k=1}^K w_k [F_k(\theta) - T_k]^2$$

subject to:

$$\theta \in \Omega$$

where:

- K is the number of classes;
- $F_k(\theta)$ is the measured F1-score for class k under model configuration θ ;
- T_k is the target F1-score for that class;
- w_k is the importance weight assigned to class k ;
- Ω is the set of feasible model configurations.

The model configuration θ may represent parameters that can actually be varied in experiments, such as the class weights used during training, the sampling strategy, or selected training hyperparameters. Minimizing J seeks to reduce deviations from the specified class-level targets.

This formulation is a proposed optimization framework for Neo Uro Light Vision, not a claim that quadratic loss has already been implemented in the model's training process. It also requires care when selecting the target values and class weights. If the targets are arbitrary or the weights are poorly chosen, the optimization may produce a misleading objective. A fair comparison should therefore report the individual class metrics alongside the aggregate objective.

2.5. Generalization and Reliability of Medical AI

Model evaluation on the available dataset is not sufficient to establish that a medical AI system will generalize to unseen data (Mollaie et al., 2026). A model may learn features associated with the image acquisition process, image background, or dataset composition rather than features directly related to the target category. Such behavior is commonly described as shortcut learning.

Ong Ly et al. (2024) examined this problem across 13 healthcare datasets and reported that conventional evaluation could overestimate model performance because of hidden data acquisition biases. Their findings emphasize the importance of evaluating whether models rely on meaningful features and whether their performance remains stable across different data sources.

For Neo Uro Light Vision, this concern supports the need for a disciplined evaluation procedure. Images used to tune the optimization objective should be separated from images used for final performance assessment. Where feasible, an independent test dataset should be collected from a different source or acquisition setting. The evaluation should also report per-class results and examine the confusion matrix to identify which categories are frequently confused.

Because the application concerns sensitive medical images involving children, technical evaluation should be accompanied by appropriate data protection, ethical oversight, and professional review. The system should be presented as a preliminary screening aid rather than an autonomous diagnostic tool.

2.6. Research Gap and Proposed Contribution

The reviewed literature identifies three issues relevant to this study. First, multiclass medical image classification requires more than a single aggregate performance score because class-level errors may differ substantially (Kocak et al., 2025). Second, class imbalance can disadvantage minority categories, making class-sensitive evaluation and optimization important (Ahsan et al., 2024). Third, high performance on a development dataset does not necessarily imply reliable generalization to new data (Mohammed et al., 2025).

Neo Uro Light Vision provides a practical setting in which to investigate these issues. Its initial evaluation includes ten classes and reports a weighted F1-score of 0.8841 across 179 images. However, the uneven class distribution and variation in class-specific metrics indicate that further evaluation is warranted.

Accordingly, this study proposes a quadratic loss-based optimization framework for assessing and improving class-wise classification performance. The central contribution is the formulation of an objective that penalizes deviations between measured class-level performance and predefined targets. Its effectiveness must be established experimentally by comparing the baseline model with optimized configurations under consistent evaluation conditions.

The study should answer the following questions:

1. Can a quadratic objective be used to quantify deviations from class-specific performance targets?
2. Which model configurations minimize the proposed objective while maintaining satisfactory performance across all ten classes?
3. Does the optimized configuration improve class-wise F1-scores and overall classification performance compared with the baseline?
4. Do the observed improvements persist on data that were not used to select the optimized configuration?

These questions distinguish the mathematical formulation from its empirical validation. A successful optimization result should be supported by comparative experiments and independent evaluation, rather than inferred from the baseline classification report alone.

3.0 METHODOLOGY

3.1. Research Design

This study employs a quantitative experimental approach to formulate and evaluate a quadratic loss-based optimization framework for multiclass image classification in Neo Uro Light Vision (Zhang et al., 2026). The research focuses on evaluating class-wise classification performance and determining whether an optimization objective based on quadratic deviations can improve the balance of performance across multiple classes.

The system classifies images into ten categories: Buried_penis, Chordee, Epispadia, Fimosis, Fistula, Hipospadia, Micropenis, Normal, Parafimosis, and Penile_strangulation. The initial classification report contains 179 evaluated images, with a weighted precision of 0.8977, weighted recall of 0.8827, and weighted F1-score of 0.8841.

These initial results serve as the baseline for the proposed study. The methodology distinguishes between the existing baseline evaluation and the subsequent optimization experiments. Because comparative experimental results are not yet available, the effectiveness of the proposed optimization is treated as a research question rather than an established finding.

3.2. Research Workflow

The study is organized into six stages.

Stage 1 — Dataset Preparation. Organize images and verify the ten class labels.

Stage 2 — Baseline Evaluation. Calculate precision, recall, F1-score, and support for each class.

Stage 3 — Quadratic Objective Formulation. Define target performance values and class weights.

Stage 4 — Model Optimization. Evaluate feasible model configurations and minimize the quadratic objective.

Stage 5 — Comparative Evaluation. Compare baseline and optimized performance using the same evaluation protocol.

Stage 6 — Independent Validation. Assess generalization using data not used for optimization.

3.3. Dataset Preparation

The dataset consists of labeled images assigned to ten predefined classification categories. Each image is associated with a ground-truth class that serves as the reference for evaluating model predictions.

The dataset preparation procedure consists of the following steps:

1. Label verification: Ensure that each image is assigned to the correct category and that class labels are consistent.
2. Image preprocessing: Apply the image resizing, normalization, and other preprocessing operations required by the existing model.
3. Dataset partitioning: Separate the images into training, validation, and test sets, where feasible. The test set must remain separate from the optimization process.
4. Class distribution analysis: Record the number of images in each category to identify potential class imbalance.
5. Reproducibility: Record the dataset partition, model configuration, preprocessing settings, and random seed where applicable.

The available classification report indicates that 179 images were evaluated. However, the information currently available does not establish whether these images constitute an independent test set. The study must document the actual data partition before making claims about generalization.

3.4. Baseline Model Evaluation

The existing Neo Uro Light Vision classifier is used as the baseline model. Its predictions are evaluated against the ground-truth labels using precision, recall, and F1-score for each class.

For class k , the evaluation metrics are defined as follows:

$$P_k = \frac{TP_k}{TP_k + FP_k}$$

$$R_k = \frac{TP_k}{TP_k + FN_k}$$

$$F1_k = \frac{2P_k R_k}{P_k + R_k}$$

where TP_k , FP_k , and FN_k represent true positives, false positives, and false negatives, respectively.

The weighted F1-score is calculated as:

$$F1_{\text{weighted}} = \sum_{k=1}^K \frac{n_k}{N} F1_k$$

where n_k is the number of actual samples in class k , N is the total number of evaluated samples, and $K=10$.

The baseline classification report provides the following overall results:

Table 1. Baseline Classification Report

Evaluation metric	Baseline result
Weighted precision	0.8977
Weighted recall	0.8827
Weighted F1-score	0.8841
Number of evaluated images	179
Number of classes	10

These values provide the reference point for subsequent experiments. The class-level results are also retained because weighted metrics alone may conceal weak performance in minority classes.

3.5. Quadratic Loss-Based Optimization Model

3.5.1. Mathematical formulation

The proposed optimization framework minimizes the squared deviation between the measured F1-score of each class and its predefined target value.

The objective function is expressed as:

$$J(\theta) = \sum_{k=1}^K w_k [F1_k(\theta) - T_k]^2$$

subject to:

$$\boldsymbol{\theta} \in \Omega, \quad w_k \geq 0, \quad \sum_{k=1}^K w_k = 1$$

where:

- $J(\boldsymbol{\theta})$ is the quadratic objective value;
- $\boldsymbol{\theta}$ represents the selected model configuration;
- $F1_k(\boldsymbol{\theta})$ is the measured F1-score for class k ;
- T_k is the target F1-score for class k ;
- w_k is the importance weight of class k ;
- Ω represents the feasible set of model configurations.

A smaller value of J indicates that the measured class-level F1-scores are closer to their corresponding targets. The objective does not automatically guarantee a higher overall F1-score; its behavior depends on the targets, weights, and available model configurations.

3.5.2. Selection of target values and class weights

The target values T_k must be specified before evaluating the final test set. They may be based on predefined performance requirements or a documented development-stage protocol. They should not be selected after observing the final test results.

For a simple initial experiment, equal class weights can be used:

$$w_k = \frac{1}{K} = \frac{1}{10} = 0.1$$

This formulation gives every class equal importance in the optimization objective, regardless of its sample count.

Equal weights are particularly relevant when the objective is to avoid allowing a large class to dominate the optimization. Nevertheless, equal weights do not eliminate the uncertainty associated with classes represented by very few images. Confidence intervals or repeated evaluation should be considered where the dataset size permits.

3.5.3. Optimization procedure

Because class-level F1-scores are calculated from discrete predictions, the proposed objective is not directly differentiable with respect to the model configuration in its ordinary form. Therefore, the initial implementation can use a controlled search over feasible configurations.

The procedure is as follows:

1. Define a set of feasible configurations, such as selected class weights, sampling strategies, or training hyperparameters.
2. Train or evaluate the model under each configuration using the same training and validation protocol.
3. Calculate the class-level F1-scores on the validation data.
4. Compute the quadratic objective $J(\boldsymbol{\theta})$ for each configuration.
5. Select the configuration with the smallest validation objective.
6. Evaluate the selected configuration on the held-out test set.

If the model is not retrained and only its output probabilities are available, alternative post-processing strategies may be investigated. However, any such strategy must be explicitly described and compared with the original classifier.

The key distinction is that this method optimizes a quadratic objective over model configurations. It should not be described as replacing the neural network's training loss unless the quadratic objective is actually incorporated into the training procedure.

3.6. Experimental Comparison

The baseline model and optimized model are compared using identical evaluation data and consistent metric definitions.

The primary comparison metrics are:

- Weighted precision.
- Weighted recall.
- Weighted F1-score.
- Macro-average F1-score.
- Class-specific precision, recall, and F1-score.
- Quadratic objective value.
- Confusion matrix.

Macro-average F1-score is calculated as:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K F1_k$$

Unlike weighted F1-score, macro F1 assigns equal weight to each class. Reporting both measures allows the study to examine whether improvements in overall performance are accompanied by improvements in less represented categories.

The comparison table should be completed after the optimization experiments.

Table 2. Comparison Table

Metric	Baseline	Optimized model
Weighted precision	0.8977	To be measured
Weighted recall	0.8827	To be measured
Weighted F1-score	0.8841	To be measured
Macro F1-score	To be calculated	To be measured
Quadratic objective	To be calculated	To be measured

No improvement should be claimed until the optimized model has been evaluated under the same conditions as the baseline.

3.7. Validation and Statistical Reliability

The final evaluation must use images that were not used to select model configurations, tune class weights, or define the optimization parameters. This separation reduces the risk of overestimating performance.

Where feasible, the study should additionally:

1. Repeat training with different random seeds to assess variability.
2. Report per-class support alongside classification metrics.
3. Examine the confusion matrix to identify systematic misclassification.
4. Estimate confidence intervals for relevant metrics, using an appropriate resampling method where feasible.
5. Evaluate the model on an independent dataset when one is available.

Because Neo Uro Light Vision is intended for preliminary screening involving sensitive medical images, classification performance should not be interpreted as proof of diagnostic accuracy. Any future clinical application would require appropriate ethical review, data governance, and validation by qualified medical professionals.

4.0 RESULTS AND DISCUSSION

4.1. Baseline Classification Performance

The initial evaluation of Neo Uro Light Vision was conducted on 179 images belonging to ten classification categories. The classification report provides precision, recall, and F1-score for each category, as well as weighted-average performance metrics.

The baseline model achieved a weighted precision of 0.8977, a weighted recall of 0.8827, and a weighted F1-score of 0.8841. These results indicate that the model demonstrated promising classification performance on the evaluated dataset. However, the aggregate metrics should be interpreted alongside the distribution of samples and the performance of individual classes.

Baseline performance

Precision	89.77%
Recall	88.27%
F1-score	88.41%

Weighted-average metrics from the existing classification report; 179 evaluated images across 10 classes. The weighted F1-score of 0.8841 provides a useful reference for evaluating subsequent optimization experiments. Nevertheless, this value alone does not establish the model's clinical reliability or its ability to generalize to images from different sources.

4.2. Class-Wise Classification Performance

The class-wise results reveal differences in precision, recall, and F1-score across the ten categories. Table 1 summarizes the reported classification metrics.

Table 3. Baseline class-wise classification performance

Class	Precision	Recall	F1-score	Support
Buried_penis	1.0000	0.8000	0.8889	10
Chordee	1.0000	0.7143	0.8333	7
Epispadia	1.0000	0.8571	0.9231	7
Fimosis	1.0000	0.7778	0.8750	18
Fistula	1.0000	0.8750	0.9333	8
Hipospadia	0.8652	0.9167	0.8902	84
Micropenis	1.0000	0.8571	0.9231	7
Normal	0.8182	0.9310	0.8710	29
Parafimosis	1.0000	0.8000	0.8889	5
Penile_strangulation	0.5714	1.0000	0.7273	4

Source: Existing Neo Uro Light Vision classification report.

The highest reported F1-score was obtained by Fistula, at 0.9333, followed by Epispadia and Micropenis, both at 0.9231. In contrast, Penile_strangulation recorded the lowest F1-score, at 0.7273. These differences suggest that overall performance does not represent the classification behavior of every category equally.

The results also illustrate the importance of interpreting precision and recall together. For example, Penile_strangulation achieved a recall of 1.0000 but a precision of 0.5714. This means that all actual instances of this class in the evaluated sample were identified, but some images predicted as Penile_strangulation belonged to other classes. Given the support of only four images, however, this estimate is based on limited evidence and requires further validation.

Similarly, Normal achieved a recall of 0.9310 and a precision of 0.8182. The difference suggests that some images predicted as normal were assigned incorrectly. In a preliminary screening context, such errors deserve careful investigation because a model's output may influence whether a user seeks further examination.

4.3. Effect of Class Imbalance

The dataset exhibits a substantial difference in class support. Hipospadia contains 84 images, representing approximately 46.9% of the evaluated dataset, while Penile_strangulation contains only four images, representing approximately 2.2%.

This distribution is relevant to the interpretation of weighted-average metrics. Because weighted metrics give greater influence to classes with more samples, the results of the largest category contribute substantially to the overall score. Consequently, a satisfactory weighted F1-score can coexist with weaker performance in a minority category.

The contrast between Hipospadia and Penile_strangulation illustrates this issue. Hipospadia achieved an F1-score of 0.8902 across 84 images, whereas Penile_strangulation achieved 0.7273 across only four images. The latter result indicates a potential area for improvement, but the small sample size prevents a reliable conclusion about the model's performance on this category in general.

These observations motivate the proposed quadratic loss-based optimization framework. By assigning explicit weights to class-specific deviations, the objective can represent performance requirements that might otherwise be obscured by aggregate metrics. However, whether this formulation actually improves minority-class performance must be demonstrated through comparative experiments.

4.4. Quadratic Loss-Based Optimization Analysis

The proposed quadratic objective is defined as:

$$J(\theta) = \sum_{k=1}^{10} w_k [F1_k(\theta) - T_k]^2$$

where $F1_k(\theta)$ represents the F1-score of class k , T_k is its predefined target, and w_k represents its importance weight.

The squared deviation penalizes larger differences from target performance more strongly than smaller differences. This characteristic makes the formulation suitable for evaluating whether a model configuration meets class-specific performance objectives.

For Neo Uro Light Vision, the quadratic objective can be used to compare feasible model configurations evaluated on validation data. Configurations with lower objective values are closer to the specified class-level targets according to the chosen weights.

Nevertheless, the current classification report is insufficient to calculate a definitive quadratic optimization result because the target values, class weights, and experimental results from alternative model configurations have not yet been provided. Therefore, no reduction in quadratic loss or improvement after optimization can currently be reported.

The distinction is important: the baseline results justify investigating class-wise optimization, but they do not demonstrate that the proposed method has already improved classification performance.

4.5. Implications for Neo Uro Light Vision

The baseline results provide three practical implications for the development of Neo Uro Light Vision. First, the overall weighted F1-score of 0.8841 provides a useful benchmark for subsequent experiments. Any proposed optimization should be compared against this baseline using a consistent evaluation procedure.

Second, class-wise evaluation should remain a central component of model assessment. Improvements in weighted F1-score alone may not be sufficient if performance deteriorates in categories with fewer samples. Reporting macro F1-score, class-specific recall, and the confusion matrix would provide a more complete assessment.

Third, the results highlight the need for additional data and independent validation. Categories represented by only a few images produce less stable performance estimates. Collecting more appropriately labeled images, where ethically and practically feasible, would improve the evaluation of underrepresented categories.

Because Neo Uro Light Vision is intended for preliminary screening involving sensitive pediatric medical images, the model should not be presented as a substitute for clinical examination. Its outputs require appropriate privacy safeguards, ethical oversight, and professional interpretation.

4.6. Limitations

Several limitations should be considered when interpreting the current findings.

1. Limited evaluation data. The available report contains 179 evaluated images, and several classes have fewer than ten samples.
2. Uneven class distribution. The largest category contains 84 images, whereas the smallest contains four, potentially limiting the reliability of comparisons between classes.
3. Absence of comparative optimization results. No results from an optimized model have been provided, so the effectiveness of the proposed quadratic objective remains unverified.
4. Unconfirmed independence of the test set. The available information does not establish whether the 179 images were completely separated from the data used for training and model selection.
5. No documented external validation. The available metrics do not establish performance across different data sources or acquisition conditions.

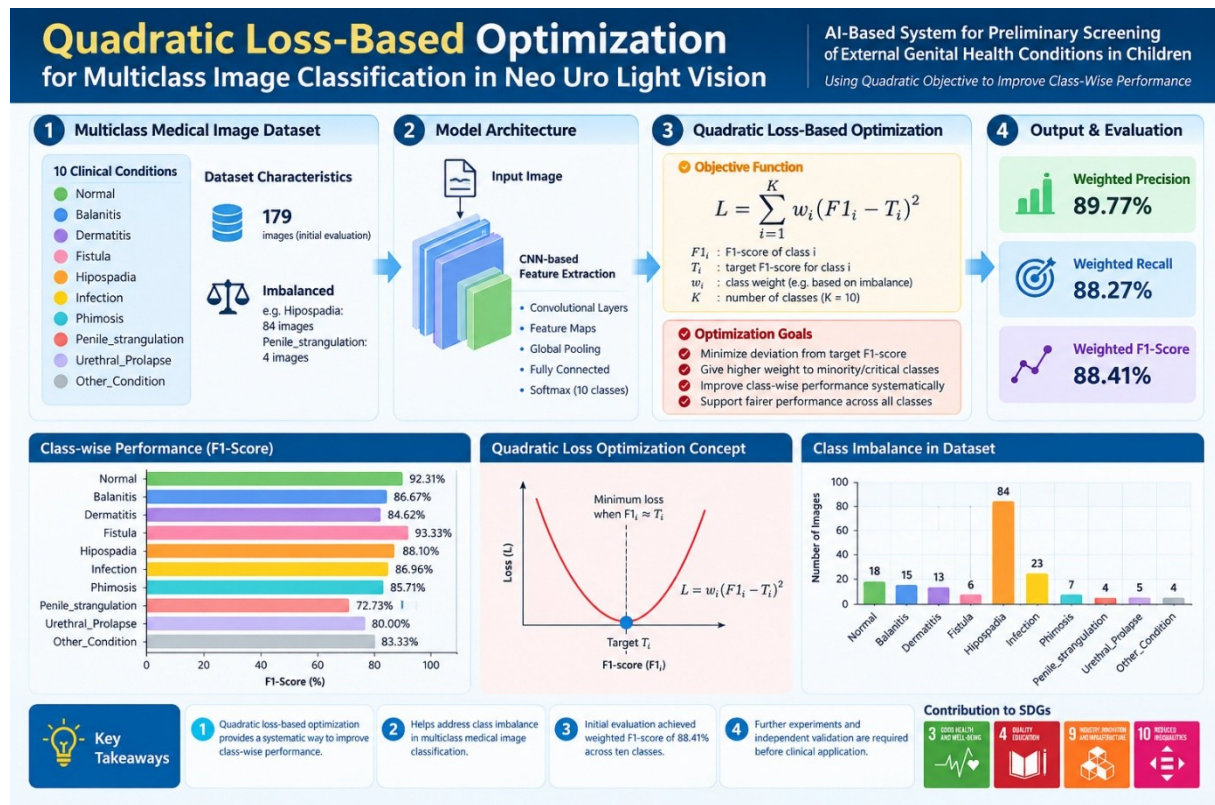
These limitations do not invalidate the baseline results. Rather, they define the additional experiments needed to establish the contribution of the proposed optimization method.

4.7. Discussion Summary

The baseline evaluation indicates that Neo Uro Light Vision achieved a weighted F1-score of 0.8841 across ten image classes. However, differences in class-specific performance and the uneven distribution of samples demonstrate why aggregate metrics alone are insufficient.

The proposed quadratic loss-based framework provides a mathematical approach for measuring deviations from predefined class-level F1-score targets (Yu & Chen, 2026). Its potential value lies in making class-specific performance requirements explicit during model configuration and evaluation.

The principal research question remains whether minimizing this objective produces measurable improvements over the baseline without compromising performance in other categories. Answering that question requires optimization experiments, a clearly defined comparison protocol, and evaluation on data not used for model selection.



5.0 CONCLUSION

5.1. Conclusion

This study presents a quadratic loss-based optimization framework for multiclass image classification in Neo Uro Light Vision. The proposed framework is designed to evaluate class-specific classification performance by measuring the squared deviation between the observed F1-score and a predefined target for each class.

The initial classification report evaluated 179 images across ten categories and achieved a weighted precision of 0.8977, weighted recall of 0.8827, and weighted F1-score of 0.8841. Although these results provide a useful baseline, the class-wise metrics reveal variations in classification performance. In particular, Penile_strangulation achieved an F1-score of 0.7273, while Fistula achieved 0.9333. The number of evaluated samples also varied considerably across classes.

These findings support the need for class-sensitive evaluation rather than relying exclusively on aggregate metrics. The proposed quadratic objective provides a mathematical basis for comparing model configurations according to their deviations from class-level performance targets. However, its effectiveness in improving classification performance has not yet been established experimentally. Further comparative experiments and independent validation are required before the proposed framework can be considered demonstrably effective.

5.2. Implication

The study has methodological and practical implications for the development of AI-based medical image classification systems. From a methodological perspective, the quadratic objective provides an explicit way to incorporate class-specific performance requirements into model evaluation and configuration selection. This approach may help researchers identify configurations that better align with predefined classification targets, particularly when performance differs across classes.

From a practical perspective, the baseline results demonstrate the importance of examining individual classification categories in addition to overall performance. For Neo Uro Light Vision, this means that model assessment should consider precision, recall, F1-score, and sample support for each category.

The framework may also provide a foundation for evaluating other multiclass image classification applications with uneven class distributions. Nevertheless, its applicability to other datasets and tasks must be established through further experimentation.

Because Neo Uro Light Vision is intended for preliminary screening involving sensitive pediatric medical images, the results should be interpreted as technical classification performance rather than evidence of clinical diagnostic accuracy. The system should support, not replace, professional medical assessment.

5.3. Limitation

This study has several limitations. First, the baseline evaluation consists of 179 images distributed unevenly across ten classes. Some categories contain very few samples, which limits the reliability and generalizability of their performance estimates.

Second, the available results provide baseline classification metrics but do not include comparative results from optimized model configurations. Consequently, it is not yet possible to quantify improvements attributable to the proposed quadratic loss-based framework.

Third, the independence of the evaluated images from the model-training and optimization datasets has not been established in the available information. Without a clearly separated test set, the generalization performance of the model cannot be determined confidently.

Fourth, the quadratic objective depends on the selection of class-specific targets, importance weights, and feasible model configurations. Different choices may produce different optimization outcomes. The sensitivity of the results to these choices has not yet been evaluated.

Finally, the current findings do not establish clinical validity, performance across different image acquisition settings, or suitability for autonomous medical decision-making.

5.4. Recommendation

Based on the findings and limitations, several recommendations are proposed. First, the dataset should be expanded where ethically and practically feasible, particularly for categories with limited sample counts. All images should have appropriate labels and comply with relevant privacy and research ethics requirements.

Second, the quadratic optimization framework should be implemented and evaluated using multiple feasible model configurations. The baseline model and optimized model should be compared using the same data partitions and evaluation metrics.

Third, future evaluations should report macro F1-score, weighted F1-score, class-specific precision and recall, and the confusion matrix. These measures would provide a more comprehensive understanding of classification performance across all ten categories.

Fourth, the selection of target F1-scores and class weights should be documented and justified before final testing. The optimization process should use validation data, while the final test set should remain separate from model selection.

Finally, any future application of Neo Uro Light Vision should incorporate appropriate data protection, expert review, and independent evaluation before being considered for use in a clinical setting.

5.5. Future Research

Future research should focus on experimentally validating the proposed quadratic loss-based optimization framework and investigating its effectiveness under different dataset conditions. The first direction is to compare the quadratic objective with alternative optimization approaches, including conventional cross-entropy training, class-weighted training, and other imbalance-aware loss functions. Such comparisons could determine whether the proposed formulation offers measurable benefits over established methods.

The second direction is to investigate different target values and class-weighting strategies. Experiments could examine whether equal class weights or weights adjusted for class importance produce more balanced classification performance.

The third direction is to evaluate the framework using larger datasets and independent test samples. This would help establish whether improvements observed during development persist across unseen images and different acquisition conditions.

The fourth direction is to investigate uncertainty estimation and confidence calibration. These techniques may help identify predictions for which the model's output is insufficiently reliable and should therefore be reviewed by a qualified professional.

Finally, future studies should examine the feasibility of integrating Neo Uro Light Vision into a professionally supervised preliminary screening workflow. Such work would require appropriate ethical approval, privacy protection, and clinical evaluation before any claims about diagnostic usefulness could be made.

Overall, future research should move beyond the mathematical formulation of quadratic optimization toward rigorous comparative testing, independent validation, and responsible evaluation of its potential role in medical image screening.

Acknowledgement

The authors would like to express their sincere gratitude to the Directorate General of Research and Development (Direktorat Jenderal Riset dan Pengembangan), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia) for providing financial support for this research under Research Contract No. 102/SPK/C.C4/PPK.DHK.IV/2026, dated 20 April 2026. The authors also gratefully acknowledge the support provided by Universitas Riau, Institut Bisnis dan Teknologi Pelita Indonesia, and PT Ciera Cahaya Harmoni for facilitating the implementation of this research through academic collaboration, research facilities, and institutional support. Their contributions have been instrumental in the development of the proposed Quantum Vision AI platform for the early detection of pediatric urological disorders. Finally, the authors would like to thank all collaborators, healthcare professionals, and technical staff who contributed valuable expertise, constructive feedback, and technical assistance throughout the research and prototype development process.

References

- Abbott, E. E., Rehman, T., Rosania, A., Lum, D. L., Taylor, T. B., Kirk, A. J., Taylor, R. A., Baker, E. F., Rabin, E., Padela, A., Genes, N., Srivastava, A., Sangal, R. B., & Apakama, D. (2026). Understanding and Addressing Bias in Artificial Intelligence Systems: A Primer for the Emergency Medicine Physician. *JACEP Open*, 7(1), 100311. <https://doi.org/10.1016/J.ACEPJO.2025.100311>
- Ahsan, M. M., Ali, M. S., & Siddique, Z. (2024). Enhancing and improving the performance of imbalanced class data using novel GBO and SSG: A comparative analysis. *Neural Networks*, 173, 106157. <https://doi.org/10.1016/J.NEUNET.2024.106157>
- Jena, B., Saxena, S., Nayak, G. K., Saba, L., Sharma, N., & Suri, J. S. (2021). Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review. *Computers in Biology and Medicine*, 137, 104803. <https://doi.org/10.1016/J.COMPBIOMED.2021.104803>
- Kanyana, R., Ounyesiga, L., & Uzorka, A. (2026). Artificial intelligence-assisted design and optimization of radiation shielding materials for medical applications. *Journal of Radiation Research and Applied Sciences*, 19(3), 102603. <https://doi.org/10.1016/J.JRRAS.2026.102603>
- Kaur, R., & Singh, S. (2022). A comprehensive review of object detection with deep learning. *Digital Signal Processing: A Review Journal*, 132. <https://doi.org/10.1016/j.dsp.2022.103812>
- Kautz, T., Eskofier, B. M., & Pasluosta, C. F. (2017). Generic performance measure for multiclass-classifiers. *Pattern Recognition*, 68, 111–125. <https://doi.org/10.1016/J.PATCOG.2017.03.008>
- Kocak, B., Klontzas, M. E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., Bressem, K. K., Ugga, L., Mercaldo, N., Díaz, O., & Cuocolo, R. (2025). Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, 3, 100030. <https://doi.org/10.1016/J.EJRAI.2025.100030>
- Liu, B., Pedersen, M., & Raja, K. (2026). Loss function assessment: Towards learning hard-to-learn classes. *Neurocomputing*, 686, 133638. <https://doi.org/10.1016/J.NEUCOM.2026.133638>
- Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132, 102549. <https://doi.org/10.1016/J.IS.2025.102549>
- Mollaie, F., Shakoor, M. H., Sepahvand, M., & Boostani, R. (2026). Medical data types in machine learning and their challenges: A comprehensive review. *Intelligence-Based Medicine*, 15, 100427. <https://doi.org/10.1016/J.IBMED.2026.100427>
- Mukherjee, K. (2012). A Review of Robust Estimation under Conditional Heteroscedasticity. *Handbook of Statistics*, 30, 123–154. <https://doi.org/10.1016/B978-0-444-53858-1.00006-5>
- Pitoyo, J., Jahrizal, Siregar, F. M., Nyoto, N., Putra, R., & Renaldo, N. (2026). Quantum Vision Ai For Pediatric Urological Screening Using Hybrid Quantum Machine Learning. *Journal of Information Systems, Digitization and Business*, 5(3), 01–07. <https://doi.org/10.38142/jisdb.v5i3.2116>
- Rayed, M. E., Islam, S. M. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, 47, 101504. <https://doi.org/10.1016/J.IMU.2024.101504>
- Renaldo, N., Susanti, W., Yuliendi, R. R., Marlim, Y. N., Kurniawan, W. J., & Tendra, G. (2025). Innovations in Computer Vision for External Disease Identification. *Journal of Information Systems, Digitization and Business*, 4(1), 1–7.

- Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118, 107965. <https://doi.org/10.1016/J.PATCOG.2021.107965>
- Yu, H. C., & Chen, Y. K. (2026). Improved FTIR-based classification for food authentication using a topological ensemble framework. *Current Research in Food Science*, 12, 101441. <https://doi.org/10.1016/J.CRFS.2026.101441>
- Zhang, X., Wang, H., Zhou, H., Xiao, D., Xiong, H., Cheng, J., He, Y., Feng, Q., & Yang, J. (2026). An AI-powered diagnostic system for grading and invasion of non-muscle-invasive bladder cancer via TURBT specimens: A multicenter study. *IScience*, 29(7), 116651. <https://doi.org/10.1016/J.ISCI.2026.116651>