BioMedInformatics

MDPI

Article

Automatic Interpretation of RPR Tests Using Lightweight Hybrid Architectures for Binary and Ternary Classification: A Preliminary, Single-Device Proof-of-Concept Study

Enmanuel Abilheira [1,*], Bruno Silva [1], Ljiljana Dukanovic [2], Afonso Pinheiro [2] and Vitor Carvalho [1,*]

[1] 2Ai—Applied Artificial Intelligence Laboratory, School of Technology, Polytechnic University of Cávado and Ave (IPCA), Campus of IPCA, Vila Frescaínha S. Martinho, 4750-810 Barcelos, Portugal
[2] CRIAM—Knowledge and Innovation, Industrial Park of Tortosendo, Street G, Lot 60, 6200-823 Tortosendo, Portugal
* Correspondence: eabilheira@ipca.pt (E.A.); vcarvalho@ipca.pt (V.C.)

Abstract

This study evaluates a lightweight, edge-deployable artificial intelligence pipeline to assist, not replace, trained human readers in the classification of RPR test reactions. Two separate and non-directly comparable experimental configurations were investigated: a binary task (Reactive vs. Non-Reactive) using 243 original images and a ternary task (Reactive, Minimally Reactive, Non-Reactive) using a distinct dataset of 293 original images. Because the datasets were acquired using a single device and laboratory protocol, and because deterministic augmentation generates highly correlated transformations rather than independent clinical samples, the reported results should be interpreted as preliminary internal evidence of feasibility rather than proof of clinical generalizability. In the augmented internal test evaluation, the binary model achieved 99.98% accuracy (25,137/25,200), while the ternary model achieved 91.12% accuracy (14,417/15,822). In the original-image deployment evaluation, binary performance remained 100% (58/58) across FP32, FP16, and INT8; ternary performance was preserved under FP32/FP16 at 95.24% (80/84) but decreased to 76.19% (64/84) after INT8 quantization. An additional stochastic augmentation experiment for ternary INT8 deployment restored performance to 95.24% (80/84) and 0.9444 Macro-F1, but external validation remains mandatory before any clinical adoption.

Keywords: RPR test; syphilis screening; deep learning; edge AI; lightweight models; medical image classification; ONNX

Check for updates

Academic Editor: Carson K. Leung

Received: 1 June 2026
Revised: 10 June 2026
Accepted: 18 June 2026
Published: 24 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Syphilis remains a significant global public health challenge despite the availability of effective antibiotic treatment. Caused by *Treponema pallidum*, the disease continues to affect millions of individuals annually, with a growing resurgence reported in both low- and high-income regions [1,2]. Early detection is essential to prevent disease progression, reduce transmission, and avoid severe complications, including neurological and cardiovascular damage, as well as adverse pregnancy outcomes [3]. As a result, large-scale screening programs play a critical role in syphilis control strategies.

The Rapid Plasma Reagin (RPR) test is a widely used non-treponemal serological assay for syphilis screening and treatment monitoring. It is inexpensive, fast, and suitable for diverse environments such as clinical laboratories, antenatal care settings, and public health campaigns [4,5]. The test relies on the visual detection of flocculation patterns formed

by antigen–antibody interactions. Depending on the intensity and distribution of these patterns, results are typically classified as Reactive or Non-Reactive [6], and sometimes Minimally Reactive if it requires further confirmation [7]. However, manual interpretation of RPR tests is inherently subjective and depends on the observer's experience, environmental conditions, and subtle visual cues. This subjectivity can lead to inter- and intra-observer variability, particularly in borderline cases such as Minimally Reactive samples, which are clinically important but visually ambiguous [8]. Additionally, automated RPR methods have been shown to provide more consistent and reproducible results, making them better suited for high-throughput laboratory settings [6,8,9]. Conversely, these are also costly and bulky devices limited to RPR testing, not suited for point-of-care (POC) testing and settings with limited resources.

Recent advances in artificial intelligence, particularly deep learning, have demonstrated strong capabilities in medical image analysis [10–12]. Convolutional neural networks (CNNs) can automatically extract discriminative visual features from raw data, while transformer-based architectures enable modeling of global spatial relationships within images. Hybrid CNN–Transformer models combine these strengths, making them especially suitable for tasks involving both fine-grained texture patterns and global structural information, such as RPR flocculation analysis. In parallel, the emergence of edge artificial intelligence has enabled the deployment of models directly on embedded or portable devices, allowing real-time inference without reliance on cloud infrastructure [13–15]. This paradigm is particularly relevant in medical applications, where data privacy, latency, and connectivity constraints are critical considerations.

Despite these advances, existing studies on automated interpretation of serological tests often rely on limited datasets, lack standardization in preprocessing and evaluation, and rarely address deployment constraints. Moreover, most prior work focuses on binary classification, overlooking the clinically relevant challenge of distinguishing Minimally Reactive cases.

Despite recent advances in artificial intelligence for medical image analysis, several limitations remain in the automated interpretation of serological assays. Existing studies frequently rely on limited datasets, highly controlled acquisition environments, and insufficient external validation. In many cases, large performance gains are obtained through extensive augmentation strategies that increase the number of training samples but do not necessarily introduce independent biological variability. Consequently, reported performance metrics may overestimate real-world generalization, particularly when all images are acquired using the same device, optics, illumination conditions, and laboratory workflow.

Moreover, most previous studies focus primarily on binary classification problems, while comparatively little attention has been given to clinically relevant borderline categories such as Minimally Reactive RPR reactions. These cases are especially challenging because faint or ambiguous flocculation patterns may vary substantially between operators and acquisition conditions, contributing to inter- and intra-observer variability in routine clinical practice.

In this context, the present study investigates a lightweight, edge-deployable artificial intelligence system intended to support, rather than replace, trained laboratory personnel during RPR interpretation. Two complementary but methodologically independent experimental configurations were evaluated: a binary classification task (Reactive vs. Non-Reactive) and a ternary classification task (Reactive, Minimally Reactive, Non-Reactive). Importantly, these experiments were conducted using different original datasets and therefore should not be interpreted as direct comparative evaluations between binary and ternary classification performance.

The present manuscript extends the authors' recent conference publications [16,17] by providing an integrated and more conservative biomedical-engineering analysis of the complete RPR decision-support workflow rather than reporting two isolated feasibility experiments. Specifically, while [16] addressed binary Reactive/Non-Reactive classification using 243 original images and [17] addressed a preliminary three-class formulation using 293 original images, the present work: (i) explicitly frames the binary and ternary configurations as methodologically independent experiments that must not be directly compared; (ii) consolidates the acquisition, preprocessing, augmentation, model-selection, ONNX-export and reduced-precision deployment pipeline in a single reproducible study; (iii) reports deployment-oriented behaviour on original images across FP32, FP16 and INT8 inference, including the clinically relevant degradation observed for ternary INT8 classification; and (iv) provides a critical interpretation of deterministic augmentation, single-device acquisition, absence of external validation, and the decision-support-only nature of the system. Therefore, the main scientific added value of this manuscript is the unified validation-oriented analysis of feasibility, deployment robustness and clinical-readiness limitations for edge AI-assisted RPR interpretation, rather than a claim of definitive clinical generalizability.

The proposed workflow integrates standardized image acquisition, region-of-interest extraction, deterministic rotation-based augmentation, lightweight deep learning architectures, and deployment-oriented optimization through ONNX export and reduced-precision inference. However, because all data were collected under controlled single-device laboratory conditions, and because augmented samples remain highly correlated transformations of a limited number of original acquisitions, the results should be interpreted as preliminary internal evidence of technical feasibility rather than proof of clinical generalizability.

Accordingly, the primary objective of this work is not to establish clinical readiness, but rather to evaluate whether lightweight edge-oriented deep learning models can provide reproducible internal performance under controlled acquisition conditions while identifying the principal limitations that must be addressed in future multi-center validation studies.

The remainder of this article is organized as follows: Section 2 describes the materials and methods; Section 3 presents the experimental results observed; Section 4 discusses the findings, limitations, and implications of this work; and Section 5 concludes the paper and addresses future work.

2. Materials and Methods

This section describes the materials and methods used in this study, including study design, dataset acquisition, preprocessing, augmentation, model architectures, training protocol, evaluation metrics, statistical analysis, and deployment optimization. The methodology was designed for reproducibility under controlled conditions, while recognizing that controlled internal validation is not equivalent to real-world clinical generalization.

2.1. Study Design

This study evaluates an automated decision-support system for RPR test interpretation using lightweight deep learning models optimized for edge deployment. Two separate experimental configurations were investigated:

- Binary classification approach: Reactive and Non-Reactive classes, on a dataset of 243 original images [16].
- Ternary classification approach: Reactive, Minimally Reactive and Non-Reactive classes, on an extended dataset of 293 images [17].

Both configurations followed a comparable methodological pipeline including standardized image acquisition, preprocessing, region-of-interest extraction, deterministic augmentation, model training, evaluation, and deployment-oriented optimization.

However, the binary and ternary experiments were conducted using different original datasets with different class distributions and sample sizes. Consequently, these experiments should be interpreted as independent internal evaluations addressing related but distinct classification problems rather than as direct comparative analyses between binary and ternary performance.

The study was designed as a preliminary proof-of-concept under controlled single-device laboratory conditions. The primary objective was to evaluate the technical feasibility of lightweight edge-oriented deep learning models for automated RPR interpretation support while identifying the principal limitations associated with dataset size, acquisition variability, and reduced-precision deployment.

2.2. Dataset and Image Acquisition

Images were acquired using CRIAM's portable device (based in Lisbon, Portugal) [18], integrating LED ring illumination, fixed focal distance, and standardized exposure settings. Acquisition metadata (exposure, white balance, lens focal length, and LED intensity) were recorded to ensure reproducibility. RPR testing and image acquisition were performed at the University of Minho in a biosafety level 2 (BSL-2) laboratory, as shown in Figure 1.



Figure 1. RPR testing and image acquisition in a BSL-2 lab.

The acquisition system integrates controlled LED illumination, fixed focal distance, and standardized exposure parameters to reduce variability. Standardization of image acquisition is essential to ensure reproducible computer vision performance in medical imaging systems [16,17].

Two datasets were constructed:

- Binary dataset (n = 243) [16]:
  Reactive: 130; Non-Reactive: 113.
  Split: 170 training, 36 validation, 37 test.

- Ternary dataset (n = 293) [17]:
  Reactive: 70; Minimally Reactive: 96; Non-Reactive: 127.
  Split: 203 training, 45 validation, 45 test.

Ground-truth labels were assigned during acquisition and subsequently reviewed by qualified laboratory experts according to established diagnostic criteria. Reactive samples corresponded to clear flocculation patterns, Minimally Reactive samples represented faint or borderline reactions requiring additional attention, and Non-Reactive samples showed no visible flocculation.

No independent external reader panel, prospective clinical validation, or multi-center acquisition protocol was available during this study. Consequently, the reported results should be interpreted as controlled internal evaluations rather than evidence of generalized clinical performance across heterogeneous acquisition environments.

2.3. Preprocessing and Region-of-Interest Extraction

Preprocessing was applied to standardize image inputs and isolate diagnostically relevant features.

The main steps included:

- ROI detection: The circular reaction disk was detected using classical computer vision techniques such as the Hough Circle Transform [19].
- Cropping and resizing: Square bounding box centered on the disk with a peripheral margin; resized to 640 × 640 px.
- Normalization: Per-channel normalization based on training set statistics.
- Background removal: A circular mask isolated the reaction disk for each original image, producing two base variants (original + background-removed) per sample. Specular reflections and small bubbles were retained as real-world acquisition noise.

No artificial artifact removal techniques were applied. Preserving real-world artifacts improves model robustness and generalization [20].

2.4. Data Augmentation

Data augmentation was used to improve robustness and mitigate the limited dataset size. Since RPR assays are circular, rotation is a label-preserving transformation.

Before augmentation, each original image was transformed into two base variants: (i) the original image and (ii) a background-removed version obtained by applying a circular mask to isolate the reaction disk. This process effectively doubled the number of base images (×2 factor).

- Binary configuration [16]:
  Rotation: 1° increments (0–359°).
  Total: 243 × 2 × 360 = 174,960 images.
- Ternary configuration [17].
  Rotation: 2° increments (0–358°).
  Total: 293 × 2 × 180 = 105,480 images.

All augmented samples derived from the same original image were assigned to the same dataset split, preventing direct train-test leakage across transformations of the same source image. Offline deterministic augmentation was selected to ensure reproducibility and orientation invariance. However, the augmented images remain highly correlated transformations of a limited number of original acquisitions and therefore should not be interpreted as independent biological or clinical observations.

Importantly, augmentation increases geometric variability but does not reproduce the diversity expected across different acquisition devices, laboratories, operators, reagent

lots, illumination conditions, or patient populations. Consequently, the reported internal performance metrics may overestimate real-world generalization capability under heterogeneous clinical conditions and may partly reflect adaptation to the specific acquisition device, optics, illumination setup, and laboratory protocol.

This limitation is central to the interpretation of the reported results. Although deterministic rotation generated large, augmented datasets, the effective biological and acquisition diversity remained limited to 243 original binary images and 293 original ternary images, all obtained under controlled single-device conditions [20,21].

2.5. Model Architectures

The primary architecture was a hybrid CNN–Transformer model adapted from the **YOLOv8** classification variant.

The model combines:

- Convolutional layers for local feature extraction.
- Transformer encoder blocks for global spatial modeling.

CNNs are effective at capturing local texture features [11], while transformers enable modeling of long-range dependencies [12]. Hybrid approaches combine both advantages [22].

The output layer was adapted to:

- 2 classes (binary) [16].
- 3 classes (ternary) [17].

Baseline models included:

- MobileNet [23].
- EfficientNet [24].
- DenseNet [25].
- ResNet [26].

All models were trained under identical conditions for fair comparison.

Additionally, feature extraction using embeddings from the YOLOv8n-cls model was explored as an alternative approach. In this configuration, deep features obtained from the penultimate layer were used as input to classical machine learning and shallow neural network classifiers, including Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and ensemble-based methods.

Furthermore, fine-tuning strategies were evaluated by adapting the pre-trained YOLOv8n-cls model to the specific dataset, leveraging transfer learning to refine feature representations for the RPR classification task. Fine-tuning enables the adjustment of model parameters to better capture domain-specific patterns while preserving general visual features learned during pre-training.

However, experimental results demonstrated that neither the embedding-based approaches nor additional fine-tuning strategies outperformed the original end-to-end YOLOv8n-cls model under the evaluated internal conditions. These findings suggest that the integrated architecture was well-suited to capturing both local texture patterns and global spatial relationships relevant to the RPR classification task.

Therefore, the original end-to-end YOLOv8n-cls model was selected as the final approach because it demonstrated the strongest internal performance while maintaining a comparatively simple deployment pipeline for edge-oriented inference. The embedding-based experiments are reported as secondary analyses to document model-selection decisions and exploratory evaluations; however, they are not central to the primary clinical decision-support claim of the present study.

2.6. Training Protocol

All models were trained under standardized experimental conditions to ensure fair comparison across architectures.

- Learning rate: 0.01.
- Weight decay: 0.0005.
- Batch size: 256.
- Image size: 640 × 640 px.
- Epochs: 100.
- Automatic Mixed Precision (AMP) training was enabled to improve computational efficiency and reduce GPU memory consumption during training [27].

To efficiently manage the large augmented datasets, a checkpoint-based segmented training strategy was employed during both model training and storage phases.

- Categorical cross-entropy was employed as the loss function for both binary and ternary classification tasks.

The final model was selected based on validation performance. Early stopping was not used; instead, the best checkpoint across the fixed 100-epoch schedule was retained according to validation-set performance. No learning-rate scheduler was applied unless otherwise specified in the implementation configuration.

To improve reproducibility, fixed random seed initialization was applied whenever supported by the corresponding framework components.

2.7. Evaluation Metrics

Performance was evaluated using standard classification metrics [28]:

Binary classification:

- Accuracy.
- Precision.
- Recall (Sensitivity).
- Specificity.
- F1-score.
- AUC-ROC.

Ternary classification:

- Per-class precision, recall, F1-score.
- Macro-F1 and weighted-F1.
- One-vs-rest ROC curves.

Statistical comparisons were conducted using bootstrap confidence intervals and paired tests such as McNemar's test [29].

Bootstrap confidence intervals were estimated using 1000 resampling iterations with replacement. Statistical significance was assessed at $p < 0.05$.

Because the augmented samples are highly correlated transformations of a limited number of original acquisitions, and because no independent external validation cohort was available, the resulting confidence intervals and paired statistical analyses should be interpreted as exploratory internal stability assessments rather than confirmatory evidence of clinical generalizability.

Where applicable, exact prediction counts are reported alongside percentages to reduce potential overinterpretation of rounded metrics.

For transparency, the principal accuracy estimates are accompanied by exact counts:

- Binary augmented internal test: 25,137/25,200
- Ternary augmented internal test: 14,417/15,822

- Binary original-image deployment test: 58/58
- Ternary original-image deployment test:
  - FP32/FP16: 80/84
  - INT8 before stochastic augmentation: 64/84

Formal inferential statistics derived from deterministic augmentation should be interpreted cautiously because the independence assumption underlying paired statistical procedures may not be fully satisfied under highly correlated augmentation conditions.

2.8. Model Optimization and Deployment

Models were optimized for edge deployment using a multi-stage pipeline:

- ONNX export [30].
- FP16 conversion to reduce memory footprint.
- INT8 quantization for efficient integer inference.

Quantization techniques can reduce model size and potentially improve inference efficiency in resource-constrained deployment environments [14,31]. In the present study, however, INT8 quantization was treated primarily as an empirical deployment stress test rather than being assumed to preserve classification accuracy. This consideration is particularly important in borderline multi-class medical-image interpretation tasks, where reduced numerical precision may affect subtle decision boundaries associated with Minimally Reactive cases.

2.9. Inference Verification

To ensure consistency after optimization:

- Outputs of optimized models were compared to the FP32 baseline.
- Numerical differences were evaluated using absolute and relative error thresholds.
- Full test set evaluation was repeated.

Quantization-aware validation is important for assessing model reliability after numerical optimization and deployment [14]. However, Quantization-Aware Training (QAT) was not implemented in the present study. Given the performance degradation observed in the ternary INT8 configuration, QAT is identified as a relevant direction for future work to improve robustness under low-precision inference and to better preserve classification performance in borderline multi-class scenarios.

2.10. Ethical and Data Considerations

All data were acquired under controlled laboratory conditions with appropriate handling of biological samples. Only de-identified images were used. The proposed workflow is intended for edge-based inference to reduce data transmission requirements and support privacy preservation [15].

The system is not intended to provide a standalone clinical diagnosis. Instead, it is designed as a decision-support tool to assist laboratory personnel by suggesting likely reaction categories and highlighting borderline cases that may require confirmatory review according to local clinical protocols.

2.11. Implementation Details and Reproducibility

All experiments were conducted using Python 3.11, PyTorch, CUDA-enabled GPU acceleration, and the Ultralytics YOLOv8 framework. Training and inference were performed on a high-performance computing system equipped with two AMD EPYC ROME 7452 processors (32 cores/64 threads each, 2.35 GHz), 512 GB DDR4 ECC RAM, and four NVIDIA Tesla A100 GPUs with 40 GB VRAM each. The exact software environment,

including Python 3.11, PyTorch 2.7.0+cu128, CUDA 12.8, ONNX/ONNX Runtime 1.23.2, and Ultralytics 8.4.55 versions, was recorded to support reproducibility.

The proposed architecture was based on the lightweight YOLOv8n-cls model. Mixed precision training (AMP) was enabled to improve computational efficiency. NVMe SSD storage supported high-throughput data loading and checkpointing. Random seeds were fixed for dataset splitting and model initialization whenever supported by the corresponding software components in order to improve experimental reproducibility. Models were trained for a fixed 100-epoch schedule with checkpoint selection based on validation performance; training time was recorded for reproducibility and should be reported together with the final computational environment in the production version of the manuscript.

3. Results

3.1. Binary Classification Performance

The performance of all evaluated models on the binary classification task (Reactive vs. Non-Reactive) is summarized in Table 1. In the augmented internal test set, the proposed YOLOv8n-cls model achieved the highest performance, with 99.98% accuracy (25,137 correct predictions out of 25,200 augmented test images). This very high internal performance must be interpreted cautiously because the augmented test images are deterministic transformations of a small number of original test acquisitions obtained using the same device and laboratory protocol.

Table 1. Binary classification models' metrics on the test set, ordered by accuracy [16].

Models/Metrics	Accuracy	Precision	Recall	Loss
YOLOv8n-CLS	0.9998	0.9999	0.9967	0.0003
MobileNEtV2	0.9831	0.9779	0.9791	0.0097
Xception	0.9809	0.9745	0.9799	0.0307
NASNetLarge	0.9799	0.9735	0.9732	0.0287
InceptionV3	0.9783	0.9636	0.9719	0.0703
NASNetMobile	0.9721	0.9536	0.9689	0.1142
DenseNet121	0.9703	0.9798	0.9673	0.0493
InceptionResNetV2	0.9678	0.9388	0.9598	0.1144
DenseNet201	0.9659	0.9145	0.9559	0.1261
EfficientNetV2M	0.6981	0.7881	0.6218	0.6462
VGG16	0.6763	0.6977	0.5999	0.6986

The proposed YOLOv8n-cls model achieved the highest internal performance, with 99.98% accuracy (25,137 correct predictions out of 25,200 augmented test images). Because the augmented test set consists of deterministic transformations derived from a limited number of original acquisitions, this result should be interpreted as evidence of internal consistency under controlled conditions rather than proof of clinical generalization.

The confusion matrix (Figure 2) shows near-perfect internal classification performance, with only negligible misclassifications. Nevertheless, the result should not be interpreted as proof of real-world clinical generalization, because independent images from other devices, sites, or operators were not available.

The three best models in Table 1 achieved strong internal performance (accuracy > 98%); however, YOLOv8n-cls consistently outperformed them. This finding supports the selection of YOLOv8n-cls for further deployment experiments, while the unusually high binary performance remains subject to the dataset and acquisition constraints described above.

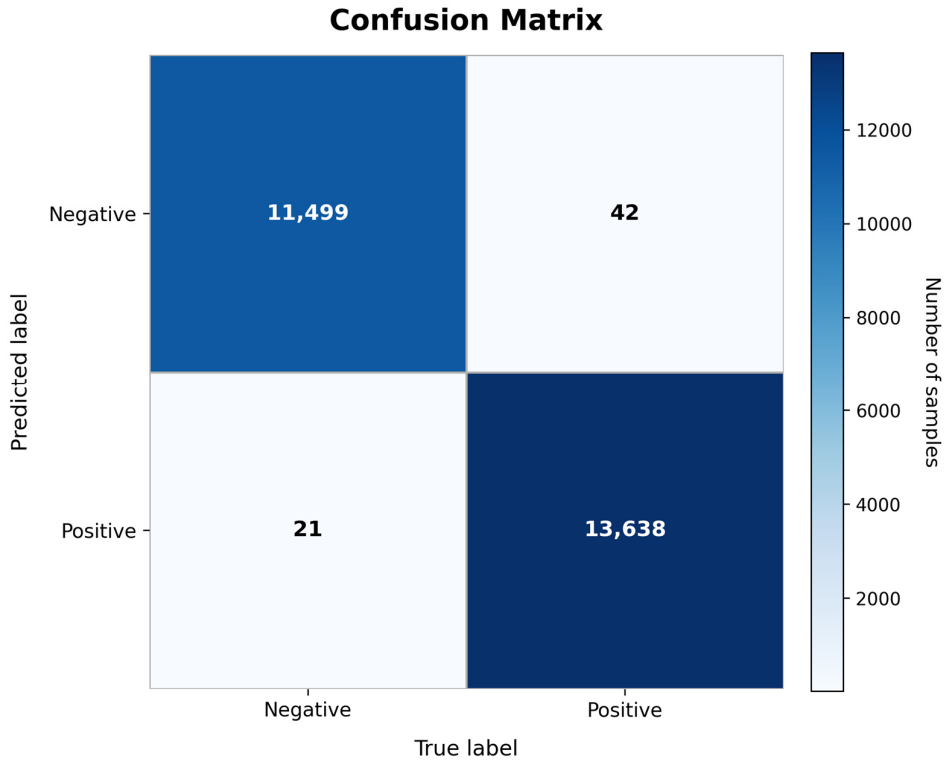


Figure 2. Confusion matrix for the best performing model YOLOv8n-cls for binary classification on the test set [16].

3.2. Ternary Classification Performance

The results for the ternary classification task (Reactive, Minimally Reactive, Non-Reactive) are presented in Tables 2 and 3. In the augmented internal test set, YOLOv8n-cls achieved 91.12% accuracy (14,417 correct predictions out of 15,822 augmented test images). Because the ternary dataset differs from the binary dataset in both size and class composition, these results should be interpreted independently and not as a direct comparison with the binary classification task. Given the structural differences between the binary and ternary datasets, their respective evaluation metrics are not directly comparable.

Table 2. Ternary classification models' metrics on the test set, ordered by accuracy [17].

Models/Metrics	Accuracy	Loss
YOLOv8n-CLS	0.9112	0.0002
MobileNEtV2	0.8831	0.0097
Xception	0.8809	0.0307
NASNetLarge	0.8799	0.0287
InceptionV3	0.8783	0.0703
NASNetMobile	0.8721	0.1142
DenseNet121	0.7703	0.0493
InceptionResNetV2	0.7678	0.1144
DenseNet201	0.7659	0.1261
EfficientNetV2M	0.7981	0.6462
VGG16	0.7763	0.6986

As shown in Table 2, YOLOv8n-cls consistently surpassed all evaluated deep learning architectures, including MobileNetV2, Xception, and NASNet variants. While most baseline models achieved competitive performance (accuracy above 0.87), the proposed model demonstrated a clear advantage in terms of both accuracy and loss, indicating better generalization and more stable predictions. Notably, EfficientNetV2M and VGG16 exhibited

significantly lower performance, suggesting that heavier architectures do not necessarily translate into improved results for this task.

Table 3. ML classifiers trained on YOLOv8n-cls embeddings [17].

Models/Metrics	Accuracy	Macro Precision	Macro Recall	Macro F1
YOLOv8n-cls + LinearSVM	0.8889	0.8718	0.8816	0.87417
YOLOv8n-cls + Ensemble	0.8889	0.8718	0.8816	0.87417
YOLOv8n-cls + MLP_2_layers	0.8778	0.8602	0.8697	0.86105
YOLOv8n-cls + ExtraTrees	0.8778	0.8612	0.8697	0.86260
YOLOv8n-cls + Stacking	0.8778	0.8612	0.8697	0.86260
YOLOv8n-cls + RidgeClassifier	0.8667	0.8509	0.8578	0.85085
YOLOv8n-cls + LogisticRegression	0.8667	0.8492	0.8578	0.84786
YOLOv8n-cls + SVM_RBF	0.8667	0.8492	0.8578	0.84786
YOLOv8n-cls + RandomForest	0.8667	0.8509	0.8578	0.85086
YOLOv8n-cls + MLP_3_layers	0.8667	0.8415	0.8545	0.84793
YOLOv8n-cls + CalibratedClassifierCV	0.8667	0.8509	0.8578	0.85086
YOLOv8n-cls + GradientBoosting	0.8556	0.8389	0.8394	0.83778
YOLOv8n-cls + AdaBoost	0.8556	0.8326	0.8394	0.83465
YOLOv8n-cls + GaussianNB	0.8556	0.8407	0.8459	0.83890
YOLOv8n-cls + KNN	0.8444	0.8232	0.8307	0.82277
YOLOv8n-cls + LightGBM	0.8444	0.8205	0.8275	0.82169
YOLOv8n-cls + NearestCentroid	0.8444	0.8275	0.8339	0.82333
YOLOv8n-cls + LDA	0.8444	0.8262	0.8307	0.82585
YOLOv8n-cls + RadiusNeighbors	0.8333	0.8169	0.8221	0.81072
YOLOv8n-cls + HistGradient	0.8000	0.7748	0.7799	0.77228
YOLOv8n-cls + xgboost	0.4444	0.1481	0.3333	0.20513

Per-class analysis showed that Reactive and Non-Reactive classes achieved higher precision and recall than the Minimally Reactive class, which remained the primary source of classification uncertainty. This pattern is clinically plausible because Minimally Reactive samples correspond to faint or borderline flocculation patterns.

The confusion matrix presented in Figure 3 corresponds to the best-performing model, YOLOv8n-cls, for the ternary classification task. This model achieved the highest overall accuracy of 91.12%, outperforming all baseline architectures reported in Table 2, including MobileNetV2 (0.8831), Xception (0.8809), NASNetLarge (0.8799), and other evaluated models.

Table 3 presents the performance of classical machine learning and shallow deep learning classifiers trained on feature representations extracted from the YOLOv8n-cls model. In this approach, the YOLOv8n-cls network was used as a feature extractor, and the resulting embeddings were used as input to different classifiers, including Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and ensemble-based models such as Random Forest, Gradient Boosting, and Stacking.

Among the evaluated methods, LinearSVM and ensemble classifiers achieved the highest performance, reaching an accuracy of 0.8889 and a macro F1-score of approximately 0.87. However, none of these approaches outperformed the original end-to-end YOLOv8n-cls model.

These results indicate that the integrated feature learning and classification mechanism of YOLOv8n-cls is more effective than decoupled approaches based on fixed feature extraction followed by external classifiers. Therefore, the end-to-end model was retained as the final approach due to its superior performance and simplicity.

The confusion matrix for the best-performing ternary model (YOLOv8n-cls) is presented in Figure 3. Most misclassifications occurred between the Minimally Reactive and Reactive classes. Non-Reactive samples were clearly separated in this internal test setting. The lower performance for the Reactive and Minimally Reactive classes reflects the inherent ambiguity of borderline reactions, where subtle flocculation patterns may overlap visually.

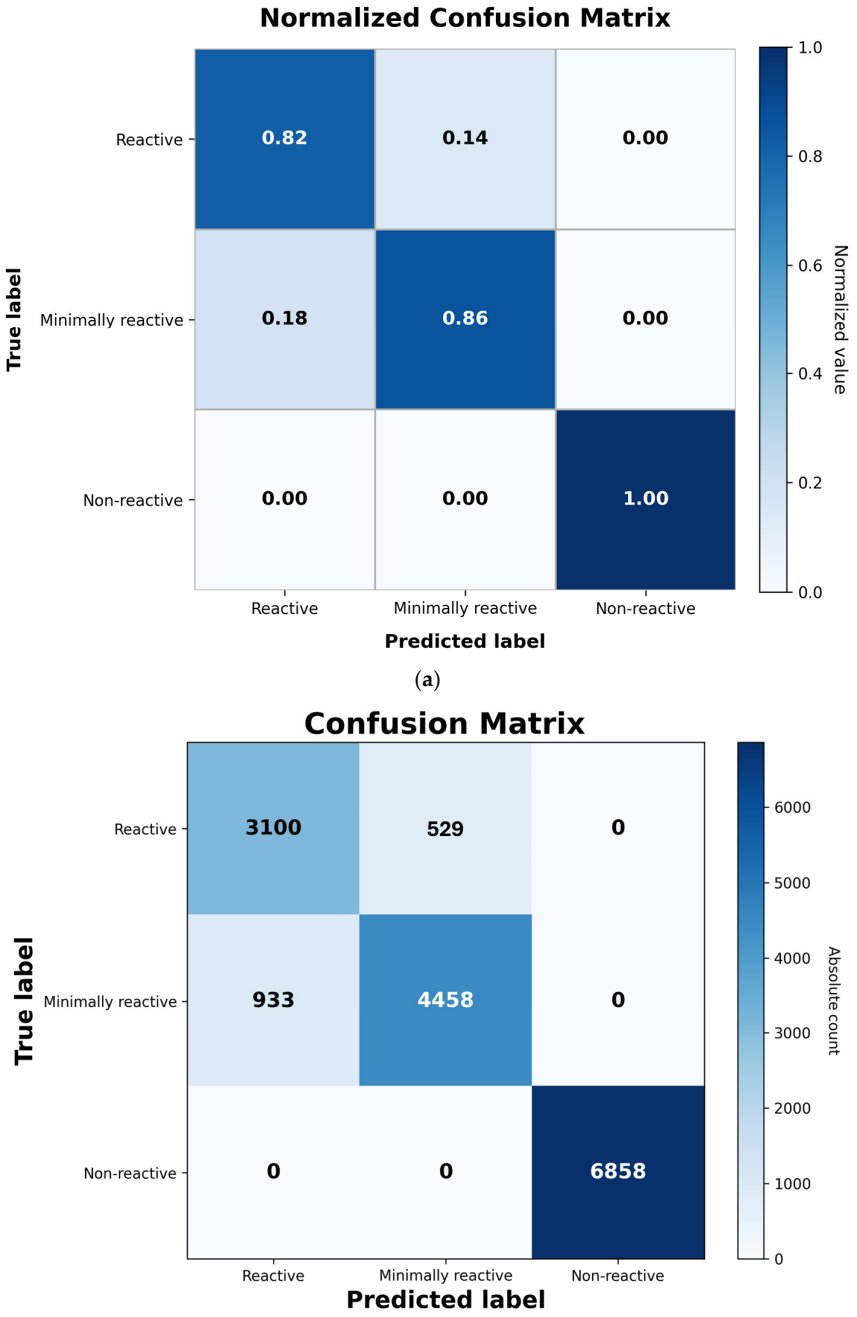


Figure 3. Confusion matrix normalized (**a**) and absolute sample counts (**b**) for ternary classification approach [17].

Overall, ternary classification introduced additional complexity compared with the binary configuration. The model showed promising internal ability to separate Non-Reactive samples from Reactive/Minimally Reactive samples, but most errors were concentrated in the clinically challenging borderline region. This reinforces the need for human review and confirmatory testing, particularly for Minimally Reactive outputs.

3.3. Impact of Model Optimization

The effect of deployment-oriented optimizations (FP32, FP16, INT8) on YOLOv8n-cls performance is summarized in Tables 4 and 5. This deployment evaluation used the original-image test sets, consisting of 58 images for the binary task and 84 images for the

ternary task, and is therefore distinct from the augmented internal test-set evaluations reported in Sections 3.1 and 3.2.

Table 4. Performance comparison for Binary Classification Approach YOLOv8n model under different precision formats (58 test images) [16].

Models	Accuracy	Precision	Recall	F1-Score	Latency (ms/img)	Image per Second	Total Time (s)
PyTorch (FP32) (YOLOv8n-cls)	1.0000	1.0000	1.0000	1.0000	865.77	1.16	50.21
ONNX FP32	1.0000	1.0000	1.0000	1.0000	1017.13	0.98	58.99
ONNX FP16	1.0000	1.0000	1.0000	1.0000	907.37	1.10	52.63
ONNX INT8	1.0000	1.0000	1.0000	1.0000	783.43	1.28	45.44

Table 5. Performance comparison for Ternary Classification Approach YOLOv8n model under different precision formats (84 test images) [17].

Models	Accuracy	Macro Precision	Macro Recall	Macro-F1	Latency (ms/img)	Image per Second	Total Time (s)
PyTorch (FP32) (YOLOv8n-cls)	0.9524	0.9487	0.9487	0.9444	932.03	1.07	78.29
ONNX FP32	0.9524	0.9487	0.9487	0.9444	905.12	1.10	76.03
ONNX FP16	0.9524	0.9487	0.9487	0.9444	1089.40	0.92	91.51
ONNX INT8	0.7619	0.7890	0.7203	0.7323	864.93	1.16	72.65

These results correspond directly to the original end-to-end YOLOv8n-cls model, without the use of external classifiers based on feature embeddings. As shown in Table 3, embedding-based approaches using classical machine learning and shallow deep learning classifiers did not outperform the original model. Therefore, the end-to-end YOLOv8n-cls architecture was selected for deployment and optimization experiments.

For binary classification, the model maintained 100% accuracy (58/58 original test images) across FP32, FP16, and INT8 formats, with no observable degradation in F1-score. This suggests that the binary decision boundary was stable under reduced precision in this controlled internal evaluation.

For ternary classification, FP32 and FP16 configurations achieved comparable performance, preserving 95.24% accuracy (80/84 original test images) and a Macro-F1 of 0.9444. In contrast, INT8 quantization decreased accuracy to 76.19% (64/84 original test images) and Macro-F1 to 0.7323, indicating that aggressive quantization substantially affected the borderline multi-class decision problem. The observed degradation suggests that low-precision inference may disproportionately affect decision boundaries associated with visually subtle Minimally Reactive cases.

This degradation was primarily associated with increased errors involving the Minimally Reactive class, which is the most visually subtle and clinically ambiguous category. These results indicate that INT8 quantization may be acceptable for binary screening support under controlled conditions but is not yet sufficiently robust for ternary deployment without additional mitigation, such as quantization-aware training, calibration-set expansion, or mixed-precision inference.

3.4. Inference Performance and Deployment

To explore a mitigation strategy for the ternary performance degradation observed after INT8 quantization (Section 3.3), an additional stochastic augmentation experiment was conducted.

For each original image in the ternary dataset (n = 293), three stochastic variants were generated using random combinations of transformations. Each augmented sample included a sequence of four randomly selected transformations from a predefined set, including rotation, background masking, intensity variation, and geometric perturbations. All variants derived from the same original image remained within the same dataset split.

This experiment aimed to increase acquisition-like variability and improve model robustness under low-precision inference conditions. Examples of the applied transformations are illustrated in Figure 4.

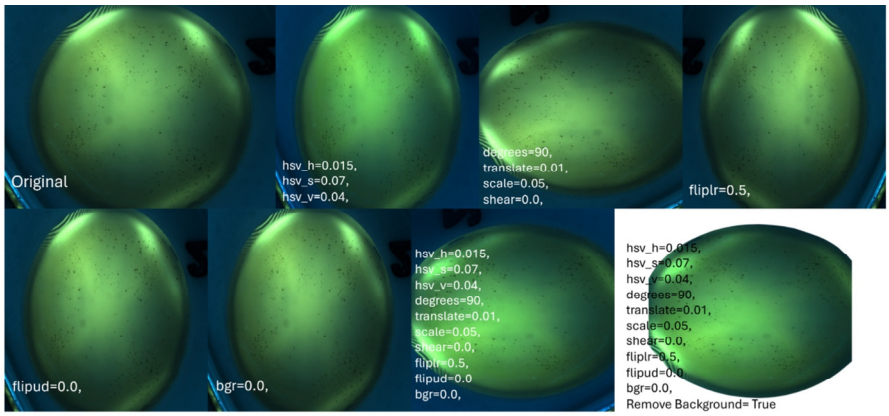


Figure 4. Examples of stochastic data augmentation transformations applied after quantization.

After stochastic augmentation, the ternary INT8 model achieved 95.24% accuracy (80/84 original test images), Macro-Precision of 0.9487, Macro-Recall of 0.9487, and Macro-F1 of 0.9444. This represents an improvement over the non-augmented INT8 deployment result of 76.19% accuracy (64/84 images) and 0.7323 Macro-F1. However, because this was still an internal single-device evaluation, the result should be interpreted as a promising mitigation signal rather than a definitive solution for clinical deployment.

Although stochastic augmentation restored the original-image test performance of the INT8 model, this result was obtained using data derived from the same acquisition environment and therefore cannot be interpreted as evidence that the mitigation strategy will generalize to external datasets.

A dedicated summary row/table for this stochastic-augmentation INT8 experiment should be retained in the final layout to make the mitigation result auditable by readers and reviewers.

All models were successfully exported and deployed using the ONNX format and evaluated under edge inference conditions.

The proposed model achieved:

- Latency of approximately 0.92–1.28 ms/image, corresponding to approximately 783–1089 images/s under the tested computing conditions.
- Reduced memory footprint under FP16 and INT8.
- Consistent predictions across FP32 and FP16 deployment configurations, with ternary INT8 requiring additional mitigation.

No major discrepancies were observed between FP32 and FP16 inference validation; INT8 discrepancies were clinically relevant for ternary classification and are therefore explicitly reported as a deployment limitation.

3.5. Comparative Summary

A comparative analysis across all experiments highlights three key findings:

1. The proposed hybrid model consistently outperformed all baseline architectures in both classification tasks.
2. Binary classification achieved near-perfect internal performance, demonstrating clear separability between Reactive and Non-Reactive samples under controlled single-device acquisition conditions.

3. Ternary classification introduced additional complexity, with the Minimally Reactive class representing the main source of classification uncertainty.

4. Discussion

The results obtained in this study indicate that lightweight deep learning models can support automated interpretation of RPR tests under controlled acquisition conditions. The near-perfect binary classification outcomes indicate that the visual patterns distinguishing Reactive from Non-Reactive samples were highly linearly separable within the internal dataset. However, the unusually high performance should be interpreted cautiously because the dataset was small, acquired with a single standardized device, and expanded mainly through highly correlated deterministic rotations. The unusually high binary performance may partly reflect adaptation to acquisition-specific characteristics of the imaging device, illumination system, optics, and laboratory workflow, rather than exclusively disease-related visual biomarkers.

A key internal finding was the consistent superiority of YOLOv8n-cls compared with other lightweight baseline architectures. This performance advantage may be related to the combination of convolutional layers for fine-grained local texture patterns and transformer components for global spatial relationships. In RPR images, both microtexture (flocculation granularity) and global distribution patterns are diagnostically relevant. Nevertheless, this conclusion remains limited to the evaluated internal datasets and requires external confirmation.

In contrast to the binary task, the ternary classification problem introduces a significantly higher level of complexity. The observed reduction in performance is primarily associated with the Minimally Reactive class, which consistently exhibited lower recall and F1-score. This behavior reflects the intrinsic ambiguity of borderline cases and is consistent with known inter-observer variability in manual RPR interpretation. The overlap in visual characteristics between Minimally Reactive and adjacent classes reduces the separability of the feature space, leading to increased misclassification rates.

Importantly, in the internal ternary confusion matrix, no Minimally Reactive samples were directly classified as Non-Reactive, which is encouraging, as such misclassifications could otherwise lead to undetected early or borderline infection signals. However, misclassification between Minimally Reactive and Reactive samples may still lead to unnecessary confirmatory testing, overtreatment risk, or patient anxiety. Both error directions are therefore clinically relevant. This balance further supports positioning the proposed system as a decision-support tool that flags results for expert review rather than as a fully autonomous diagnostic system. Nevertheless, because only 84 original test images were available for deployment evaluation, larger independent validation cohorts are required before any conclusions regarding clinical safety can be drawn.

Another important aspect of this study is the evaluation of deployment-oriented optimizations. FP16 precision preserved performance comparable to FP32 across the evaluated internal tests. By contrast, INT8 quantization introduced a clinically relevant degradation in the ternary task, particularly around the Minimally Reactive class. Although stochastic augmentation restored ternary INT8 performance in the internal original-image test set, this does not eliminate the need for quantization-aware training, larger calibration sets, and external validation before low-precision ternary deployment can be considered reliable [14,31].

Quantization-aware training was not implemented in the current study. This is a central limitation for the edge-deployment claim, particularly for ternary classification, and should be addressed in future work before deploying an INT8 ternary model in clinical workflows. Therefore, the reported INT8 mitigation results should be considered

preliminary, and future studies should evaluate whether quantization-aware training can maintain performance under independent external validation conditions.

Deterministic rotation-based augmentation is technically appropriate for circular RPR assays because rotation is label-preserving. However, it primarily improves orientation invariance and reproducibility; it does not create independent biological samples or reproduce the variability expected across devices, operators, laboratories, lighting conditions, or reagent batches.

Although deterministic augmentation substantially increased the number of training samples, the augmented images are not independent observations. The model may have learned acquisition-specific characteristics of the CRIAM device, controlled illumination, optics, and laboratory setup. Consequently, the reported internal metrics may overestimate performance under heterogeneous real-world clinical conditions.

The proposed workflow could be integrated into a point-of-care or laboratory workflow as a screening-support module. After image acquisition by a standardized portable reader, the model would provide a suggested category and confidence score to the operator. Reactive and Minimally Reactive outputs would still require laboratory interpretation and confirmatory procedures according to local clinical protocols. In this intended-use scenario, the system supports triage, documentation consistency, and reader standardization; it does not replace professional judgment.

Several limitations must be acknowledged. First, the original dataset size is small, especially for ternary classification, and augmentation cannot substitute for independent real-world samples. Second, all images were acquired using a single device in a single laboratory under controlled conditions, and no external validation dataset was available; consequently, the study cannot determine whether the model learned disease-specific visual biomarkers or acquisition-specific imaging characteristics. Third, the study did not include repeated cross-validation, prospective testing, or a direct comparison against independent expert human interpretation. Fourth, the statistical analyses are exploratory because of the dependence structure introduced by deterministic augmentation. Finally, robust ternary INT8 deployment remains unresolved because quantization-aware training and external calibration were not implemented.

Future work should prioritize multi-center external validation with independent devices, operators, reagent lots, and acquisition environments; prospective comparison with expert readers; uncertainty estimation for borderline cases; explainability methods such as Grad-CAM; and quantization-aware or mixed-precision training strategies for robust low-precision deployment.

Accordingly, the proposed system should be interpreted as an early proof-of-concept decision-support tool for assisting laboratory personnel, not as evidence of clinical readiness or as a standalone diagnostic system.

5. Conclusions and Future Work

This paper presented a lightweight, edge-deployable AI pipeline for decision support in RPR syphilis test interpretation, addressing separate binary (Reactive vs. Non-Reactive) and ternary (Reactive, Minimally Reactive, Non-Reactive) configurations. The main findings are:

- Binary classification achieved 99.98% accuracy in the augmented internal test evaluation (25,137/25,200 augmented test images) and 100% accuracy in the original-image deployment evaluation (58/58 images). However, these results were obtained under controlled single-device conditions and should not be interpreted as evidence of clinical generalization.

- Ternary classification achieved 91.12% accuracy in the augmented internal test evaluation (14,417/15,822 augmented test images). In the original-image deployment evaluation, FP32/FP16 reached 95.24% accuracy (80/84 images), whereas INT8 initially decreased to 76.19% (64/84 images), highlighting the vulnerability of borderline multi-class classification to aggressive quantization. The ternary task remains considerably more challenging because of the ambiguity associated with Minimally Reactive reactions.
- Deterministic rotation-based augmentation improved orientation robustness for circular assays but generated highly correlated samples and therefore cannot replace independent external data.
- All configurations achieved low-latency inference (approximately 0.92–1.28 ms/image; approximately 783–1089 images/s under the tested computing conditions), supporting technical feasibility for on-device decision support under controlled acquisition conditions.

Future work should incorporate larger and more diverse datasets, including multi-center data acquisition, independent device validation, and prospective workflow testing. Additional work should integrate uncertainty estimation or confidence scoring for borderline Minimally Reactive cases, compare model outputs with expert human interpretation, and implement quantization-aware training or mixed-precision strategies for robust ternary edge deployment.

Therefore, the proposed system should be regarded as a preliminary decision-support tool intended to assist trained laboratory personnel. External multi-center validation, comparison with expert readers, uncertainty estimation, and robust low-precision deployment strategies remain necessary before clinical adoption can be considered.

The limited dataset size, high correlation among augmented samples, absence of external validation, lack of independent human-reader comparison, and incomplete robustness of ternary INT8 deployment remain important limitations of the current study.

Author Contributions: Conceptualization, E.A., B.S., A.P., L.D. and V.C.; methodology, E.A., B.S., A.P., L.D. and V.C.; software, E.A., B.S., A.P., L.D. and V.C.; validation E.A., B.S., A.P., L.D. and V.C.; formal analysis, E.A., B.S., A.P., L.D. and V.C.; investigation, E.A., B.S., A.P., L.D. and V.C.; resources, A.P., L.D. and V.C.; data curation, E.A., A.P. and L.D.; writing—original draft preparation, E.A., B.S., A.P., L.D. and V.C.; writing—review and editing, A.P., L.D. and V.C.; visualization, E.A., B.S., A.P., L.D. and V.C.; supervision, A.P., L.D. and V.C.; project administration, A.P., L.D. and V.C.; funding acquisition, A.P., L.D. and V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by national funds and FCT/MCTES (PIDDAC), through the Foundation for Science and Technology, I. P. (FCT), under the scope of the projects UID/05549/2025 (https://doi.org/10.54499/UID/05549/2025) and LASI-LA/P/0104/2020 (https://doi.org/10.54499/LA/P/0050/2020).

Data Availability Statement: The datasets presented in this article are not readily available because they are intellectual property of CRIAM, provided under a legal agreement. Requests to access the datasets should be directed to Criamknowledge, S.A.

Conflicts of Interest: Ljiljana Dukanovic and Afonso Pinheiro are affiliated with CRIAM (A.P.; L.D.), the company whose prototype device was used in the present study. This affiliation is disclosed for transparency purposes. The authors confirm that this relationship did not influence the study design, data collection, data analysis, interpretation of results, manuscript preparation, or final conclusions. The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AMP	Automatic Mixed Precision

AUC	Area Under the Curve
CLS	Classification
CNN	Convolution Neural Network
FP32	32-bit Floating Point
FP16	16-bit Floating Point
INT8	8-bit Integer Quantization
KNN	K-Nearest Neighbours
ML	Machine Learning
ONNX	Open Neural Network Exchange
POC	Point-of-Care
ResNet	Residual Network
ROI	Region of Interest
ROC	Receiver Operating Characteristic
RPR	Rapid Plasma Reagin
SVM	Support Vector Machine
YOLO	You Only Look Once

References

1. World Health Organization. *Syphilis*; World Health Organization: Geneva, Switzerland, 2025. Available online: https://www.who.int/news-room/fact-sheets/detail/syphilis (accessed on 2 November 2025).
2. Rowley, J.; Vander Hoorn, S.; Korenromp, E.; Low, N.; Unemo, M.; Abu-Raddad, L.J.; Chico, R.M.; Smolak, A.; Newman, L.; Gottlieb, S.; et al. Chlamydia, gonorrhoea, trichomoniasis and syphilis: Global prevalence and incidence estimates, 2016. *Bull. World Health Organ.* **2019**, *97*, 548–562P. [CrossRef] [PubMed]
3. Kojima, N.; Klausner, J.D. An update on the global epidemiology of syphilis. *Curr. Epidemiol. Rep.* **2018**, *5*, 24–38. [CrossRef] [PubMed]
4. Peeling, R.W.; Mabey, D.; Kamb, M.L.; Chen, X.S.; Radolf, J.D.; Benzaken, A.S. Syphilis. *Nat. Rev. Dis. Primers* **2017**, *3*, 17073. [CrossRef] [PubMed]
5. Mudanyali, O.; Dimitrov, S.; Sikora, U.; Padmanabhan, S.; Navruz, I.; Ozcan, A. Integrated rapid-diagnostic-test reader platform on a cellphone. *Lab Chip* **2012**, *12*, 2678–2686. [CrossRef] [PubMed]
6. Sanfilippo, A.M.; Freeman, K.; Schmitz, J.L. Comparison of manual and fully automated AIX1000 rapid plasma reagin assays for laboratory diagnosis of syphilis. *J. Clin. Microbiol.* **2018**, *56*, e00214–e00218. [CrossRef] [PubMed]
7. Becton, Dickinson and Company. *BD Macro-Vue™ RPR Card Tests, 18 mm Circle, Qualitative and Quantitative: Instructions for Use*; BD: Franklin Lakes, NJ, USA, 2021.
8. Herring, A.J.; Ballard, R.C.; Pope, V.; Adegbola, R.A.; Changalucha, J.; Fitzgerald, D.W.; Hook, E.W.; Kubanova, A.; Mananwatte, S.; Pape, J.W.; et al. A multi-centre evaluation of nine rapid, point-of-care syphilis tests using archived sera. *Sex. Transm. Infect.* **2006**, *82*, v7–v12. [CrossRef] [PubMed]
9. Lee, J.H.; Lim, C.S.; Lee, M.G.; Kim, H.S. Comparison of an automated rapid plasma reagin test with the conventional RPR card test in syphilis testing. *BMJ Open* **2014**, *4*, e005664. [CrossRef] [PubMed]
10. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
13. Chen, T.; Ran, X. Deep learning with edge computing: A review. In Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; pp. 610–614. [CrossRef]
14. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2704–2713.
15. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
16. Abilheira, E.; Silva, B.; Dukanovic, L.; Pinheiro, A.; Carvalho, V. AI-based classification of rapid plasma reagin test reactions for syphilis detection: A preliminary approach. In Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2026), Marbella, Spain, 2–4 March 2026.

17. Abilheira, E.; Silva, B.; Dukanovic, L.; Pinheiro, A.; Carvalho, V. A preliminary study on AI-based three-class classification of syphilis RPR test reactions. In Proceedings of the 10th International Young Engineers Forum on Electrical and Computer Engineering (YEF-ECE 2026), Lisbon, Portugal, 19 June 2026.
18. CRIAM. CRIAM Knowledge. Available online: https://www.criamknowledge.pt (accessed on 3 November 2025).
19. Duda, R.O.; Hart, P.E. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM* **1972**, *15*, 11–15. [CrossRef]
20. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
21. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
24. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
25. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; García, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed precision training. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
28. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
29. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]
30. ONNX Community. Open Neural Network Exchange—ONNX. Available online: https://onnx.ai (accessed on 2 November 2025).
31. Krishnamoorthi, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv* **2018**, arXiv:1806.08342.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.