



Performance diagnostique, gain de temps réel et utilité clinique d'un réseau de neurones convolutif appliqué aux frottis sanguins : évaluation critique d'un système d'aide au diagnostic du paludisme déployé au Ngaliema Medical Center

MWANGU SHANGA Gabriel¹

Chef de Travaux

¹Université de Développement Durable en Afrique Centrale « UDDAC » Bukavu, RD Congo.

Coresponding Author : Gabriel Mwangu Shanga

Abstract: Malaria diagnosis in Congolese hospital laboratories still rests on Giemsa-stained blood-film microscopy, a method that is slow, observer-dependent and vulnerable to workload. This article submits a convolutional neural network deployed at the Ngaliema Medical Center in Kinshasa to a full independent re-analysis. The system, built on MobileNetV2 by transfer learning from 5 000 annotated erythrocyte images, was tested prospectively on 150 blood films (75 positive, 75 negative) against a medical biologist reference standard. From the three published indicators we reconstruct the complete confusion matrix, which proves internally exact, then derive the quantities that actually govern clinical use. Sensitivity reaches 96.0% (95% Wilson CI: 88.9-98.6) and specificity 93.3% (85.3-97.1), placing the model at World Health Organization competence level 1 for parasite detection but short of the 95% specificity threshold applied to rapid diagnostic tests, with a negative sample 3.9 times too small to settle the matter. Three findings follow. Diagnostic yield collapses with prevalence: positive predictive value falls from 93.5% at the artificial 50% prevalence of the test set to 61.5% at 10% and 43.1% at 5%, while negative predictive value stays above 99.5% below 10%, making the tool an exclusion instrument rather than a confirmation one. The headline comparison of 4.2 seconds of inference against 20 to 30 minutes of manual reading is inadmissible: automation acts on only 48% of the cycle, and once field acquisition and mandatory human validation are counted the cycle falls from 25.0 to 18.4 minutes, a 26.4% gain and a 35.9% capacity increase, not 99.7%. Finally, recalibrating the threshold to 99.5% sensitivity would release 54.7% of films without expert re-reading at 10% prevalence. At USD 0.30 per examination and USD 6.91 per missed infection averted, value lies in error avoidance and threshold governance, not in raw accuracy.

Keywords: Malaria; Convolutional neural network; Diagnostic accuracy; Predictive value; Decision threshold; Blood-film microscopy; Clinical decision support; Democratic Republic of the Congo.

Digital Object Identifier (DOI): <https://doi.org/10.5281/zenodo.23474594>

1 Introduction

Le paludisme a causé en 2024 environ 282 millions de cas et 610 000 décès dans le monde, dont 95 % des cas et 95 % des décès dans la Région africaine de l'Organisation mondiale de la Santé (Organisation mondiale de la Santé [OMS], 2025). La République démocratique du Congo concentre à elle seule quelque 35 millions de cas estimés et 11,7 % des décès de la Région, ce qui la place au deuxième rang mondial derrière le Nigéria (OMS, 2026). Dans ce contexte, la qualité du diagnostic n'est pas une question technique secondaire : elle conditionne la justesse de la prescription antipaludique, la maîtrise de la pression de sélection sur les molécules et la fiabilité de la surveillance épidémiologique.

Or la microscopie du frottis sanguin, méthode de référence, souffre dans les laboratoires congolais de trois faiblesses convergentes : elle est chronophage, avec vingt à trente minutes par examen complet ; elle est subjective, les critères d'interprétation variant d'un observateur à l'autre ; et elle est vulnérable à la charge de travail (Nsimire Malekera, 2026). L'apprentissage profond appliqué à l'imagerie biomédicale semble offrir une réponse, les réseaux de neurones convolutifs atteignant sur les corpus publics d'images d'érythrocytes des exactitudes couramment supérieures à 95 %, parfois annoncées au-delà de 99 % (Hoyos & Hoyos, 2024 ; Parveen & Mostafa, 2025).

Entre cette littérature et l'usage hospitalier réel subsiste toutefois un écart que la présente contribution prend pour objet. Les performances publiées portent presque toujours sur des images isolées, dans des corpus équilibrés et sur des lames préparées en conditions idéales, alors que la décision clinique porte sur une lame, dans une population de prévalence variable, au sein d'un cycle de travail dont l'essentiel demeure manuel. Nous soumettons donc à une ré-analyse indépendante et intégrale un système d'aide au diagnostic effectivement déployé au laboratoire de parasitologie du Ngaliema Medical Center de Kinshasa, en recalculant chacune de ses valeurs et en dérivant les grandeurs décisionnelles que le rapport d'origine n'a pas produites.

Quatre questions structurent la démonstration. Les indicateurs publiés sont-ils internement cohérents et statistiquement concluants ? Quel référentiel de l'OMS le système satisfait-il ? Quelle est son utilité clinique réelle une fois la prévalence prise en compte ? Et quel gain opérationnel et économique peut-on légitimement lui attribuer ? Nous posons quatre hypothèses : H1, le modèle atteint le niveau de compétence 1 de l'OMS pour la détection parasitaire mais non le seuil de spécificité de 95 % ; H2, le gain de temps réel est borné par la phase pré-analytique non automatisable ; H3, l'utilité diagnostique dépend davantage de la prévalence et du seuil de décision que de l'exactitude du modèle ; H4, la justification économique repose sur l'évitement d'erreurs et non sur l'économie de main-d'œuvre.

2 Cadre conceptuel et théorique

2.1 De l'exactitude à l'utilité clinique

L'évaluation d'un test diagnostique se déploie sur trois niveaux qu'il est fautif de confondre. La validité interne est mesurée par la sensibilité $Se = VP / (VP + FN)$ et la spécificité $Sp = VN / (VN + FP)$, grandeurs conditionnelles au statut réel du sujet et donc indépendantes de la prévalence. La valeur informationnelle est résumée par les rapports de vraisemblance $RV+ = Se / (1 - Sp)$ et $RV- = (1 - Se) / Sp$ et par l'indice de Youden $J = Se + Sp - 1$ (Youden, 1950). L'utilité clinique, enfin, est portée par les valeurs prédictives, qui dépendent de la prévalence p selon le théorème de Bayes :

$$VPP = p \cdot Se / [p \cdot Se + (1 - p)(1 - Sp)] \text{ et } VPN = (1 - p) \cdot Sp / [(1 - p) \cdot Sp + p(1 - Se)]$$

La conséquence est décisive et systématiquement négligée : un modèle validé sur un échantillon équilibré à 50 % de positifs voit sa valeur prédictive positive s'effondrer dès que la prévalence réelle diminue, sans qu'aucun de ses paramètres n'ait changé. Les recommandations internationales exigent d'ailleurs la déclaration explicite de la prévalence du recrutement et de la règle de décision (Bossuyt et al., 2015).

2.2 Précision statistique et taille d'échantillon

Une proportion estimée sur un petit effectif est entachée d'une incertitude que le point estimé masque. L'intervalle de Wilson, préférable à l'approximation normale pour de petits effectifs et des proportions extrêmes, se construit autour du centre $(\hat{p} + z^2/2n) / (1 + z^2/n)$ avec une demi-largeur $z / (1 + z^2/n) \times \sqrt{[\hat{p}(1 - \hat{p})/n + z^2/4n^2]}$ (Wilson, 1927). Symétriquement, la taille d'échantillon requise pour estimer une proportion voisine de 0,95 avec une demi-largeur

e s'obtient par $n = z^2 \cdot p(1 - p)/e^2$. Ces deux formules permettent de dire si une conformité annoncée est démontrée ou simplement non réfutée.

2.3 Bornes du gain d'automatisation

Automatiser une fraction f d'un processus dont le reste demeure manuel produit une accélération bornée par la loi d'Amdahl, $S = 1 / [(1 - f) + f/s]$, où s est le facteur d'accélération de la partie automatisée (Amdahl, 1967). Lorsque s tend vers l'infini, S tend vers $1/(1 - f)$: la fraction non automatisable devient le plafond absolu. Cette loi interdit donc de comparer un temps d'inférence à un temps d'examen complet, puisque le prélèvement, l'étalement, la coloration au Giemsa et le séchage restent manuels.

2.4 Référentiels normatifs applicables

Deux référentiels distincts sont fréquemment amalgamés. Le manuel d'assurance qualité de la microscopie du paludisme définit des niveaux de compétence gradués, le niveau 1 correspondant à une détection parasitaire de 90 à 100 % et une identification d'espèce de 90 à 100 % (OMS, 2016 ; OMS, 2018). Le seuil de 95 % de sensibilité et de spécificité relève, lui, du cadre d'évaluation des tests de diagnostic rapide. Un système positionné comme second lecteur relève du premier référentiel, mais tout usage autonome le ferait basculer dans le second. Le droit congolais, à travers la loi n° 18/035 du 13 décembre 2018 et le Code du numérique, impose au surplus la validation humaine du résultat et la protection des données de santé, ce qui verrouille le positionnement du dispositif comme outil d'assistance (Nsimire Malekera, 2026).

3 Méthodologie

3.1 Devis, terrain et matériel

L'étude est une évaluation secondaire de performance diagnostique, conduite sur les données primaires d'un système développé et déployé au laboratoire de parasitologie du NMC selon le Processus unifié, avec modélisation UML, moteur d'inférence Python, interface web et persistance relationnelle. Le modèle repose sur MobileNetV2 pré-entraîné sur ImageNet et affiné par apprentissage par transfert (Sandler, Howard, Zhu, Zhmoginov & Chen, 2018), architecture retenue pour sa légèreté, compatible avec les postes disponibles et avec un fonctionnement hors ligne. Le corpus compte 5 000 images d'érythrocytes annotées manuellement, dont 2 740 parasitées par *Plasmodium falciparum* (54,8 %) et 2 260 témoins négatifs (45,2 %), réparties en 70 % d'entraînement, 15 % de validation et 15 % de test. Les images sont normalisées à 224×224 pixels en trois canaux, traitées par lots de 32, sur quinze époques à couches gelées puis vingt-cinq époques d'affinage des trente dernières couches, avec optimiseur Adam au taux de 0,0001, entropie croisée binaire et abandon au taux de 0,3.

L'évaluation retenue comme référence est le test système prospectif : 150 lames de frottis, dont 75 positives et 75 négatives, ont été soumises au dispositif en conditions réelles de laboratoire, les sorties étant confrontées au standard de référence du biologiste médical responsable. Cette unité d'analyse, la lame et non l'image, est la seule qui ait un sens clinique, et c'est elle qui fonde tous nos calculs.

3.2 Traitement analytique

Notre démarche procède en cinq étapes. Nous reconstituons d'abord la matrice de confusion complète à partir de la sensibilité, de la spécificité et de l'effectif de chaque bras, puis nous vérifions la cohérence du triplet en recalculant l'exactitude globale. Nous calculons ensuite les intervalles de Wilson à 95 % des deux proportions et la taille d'échantillon requise pour trancher la conformité normative. Nous dérivons alors les rapports de vraisemblance, l'indice de Youden, le rapport de cotes diagnostique et les courbes de valeurs prédictives sur l'intervalle de prévalence de 0,5 % à 60 %. Nous reconstruisons une courbe ROC binormale à partir de l'aire déclarée, de paramètre de séparation $d' = \sqrt{2} \cdot \Phi^{-1}(\text{AUC})$, afin de simuler le déplacement du seuil de décision et d'en déduire la charge de relecture experte associée à chaque sensibilité cible. Nous décomposons enfin le cycle d'examen en tâches chronométrables pour appliquer la loi d'Amdahl, puis nous annualisons le coût du dispositif par le facteur de recouvrement du capital $\text{CRF} = i(1 + i)^n / [(1 + i)^n - 1]$ afin d'établir un coût par examen et un coût par erreur évitée. Aucune valeur n'est reprise sans recalcul, et tous les écarts avec le rapport d'origine sont signalés et expliqués.

4 Résultats empiriques

4.1 Cohérence interne des indicateurs publiés

La reconstitution est directe. Sur les 75 lames réellement positives, une sensibilité de 96,0 % implique $0,96 \times 75 = 72$ vrais positifs et donc 3 faux négatifs. Sur les 75 lames réellement négatives, une spécificité de 93,3 % implique $0,9333 \times 75 = 70$ vrais négatifs et donc 5 faux positifs. L'exactitude globale recalculée vaut alors $(72 + 70) / 150 = 142 / 150 = 94,67\%$, soit exactement la valeur publiée de 94,7 %, et la spécificité exacte se lit $70 / 75 = 93,33\%$. Les trois indicateurs sont donc rigoureusement compatibles entre eux, ce qui constitue un premier gage de sérieux du protocole de test et autorise à travailler sur la matrice reconstruite.

4.2 Précision statistique et conformité normative

L'intervalle de Wilson à 95 % de la sensibilité s'étend de 88,9 % à 98,6 % et celui de la spécificité de 85,3 % à 97,1 %, ce qui impose deux conclusions symétriques. La sensibilité de 96,0 % est compatible avec le niveau de compétence 1 de l'OMS, qui exige une détection parasitaire de 90 à 100 %, mais sa borne inférieure de 88,9 % ne permet pas d'exclure une performance de niveau 2. La spécificité de 93,33 % est quant à elle inférieure au seuil de 95 % des tests de diagnostic rapide, sans que la non-conformité soit démontrée puisque l'intervalle recouvre 95 %. Le rapport d'origine fixe d'ailleurs pour la spécificité un objectif de 90 % dans le tableau qui sert à conclure, alors qu'il énonce ailleurs un seuil de 95 % : cette divergence interne, sans effet sur les mesures, affaiblit l'argument de conformité. L'hypothèse H1 est confirmée dans ses deux branches.

Cette insuffisance est avant tout une insuffisance de puissance. Pour estimer une spécificité voisine de 95 % avec une demi-largeur d'intervalle de 2,5 points, il faut $n = 3,8416 \times 0,95 \times 0,05 / 0,000625 = 292$ lames négatives, soit 3,9 fois l'effectif disponible. Tant que cette validation élargie n'est pas conduite, la conformité au référentiel des tests de diagnostic rapide ne peut être ni affirmée ni infirmée, ce qui suffit à interdire tout usage autonome du dispositif.

Un point technique mérite enfin d'être relevé : le couple publié implique une séparation binormale $d' = \Phi^{-1}(0,96) + \Phi^{-1}(0,9333) = 3,252$, soit une aire sous la courbe de $\Phi(3,252/\sqrt{2}) = 0,989$, supérieure à l'aire déclarée de 0,978. Le point mesuré domine donc la courbe reconstruite. Nous retenons par prudence l'aire déclarée pour toutes les simulations de seuil, ce qui rend nos conclusions conservatrices.

4.3 L'effondrement de la valeur prédictive positive avec la prévalence

L'échantillon de test est équilibré à 50 % de positifs, prévalence artificielle qui ne correspond à aucun flux de laboratoire réel. À cette prévalence, la valeur prédictive positive vaut $72 / 77 = 93,5\%$ et la valeur prédictive négative $70 / 73 = 95,9\%$. Le recalcul bayésien aux prévalences plausibles d'un laboratoire hospitalier urbain modifie radicalement le tableau. À 30 % de prévalence, la valeur prédictive positive tombe à $0,3 \times 0,96 / (0,3 \times 0,96 + 0,7 \times 0,0667) = 0,288 / 0,3347 = 86,1\%$ et la valeur prédictive négative monte à 98,2 %. À 10 %, la valeur prédictive positive n'est plus que de $0,096 / 0,156 = 61,5\%$ et la valeur prédictive négative atteint 99,5 %. À 5 %, la valeur prédictive positive s'effondre à $0,048 / 0,1113 = 43,1\%$ tandis que la valeur prédictive négative culmine à 99,8 %.

4.4 Le gain de temps réel et sa borne physique

Le rapport d'origine oppose 4,2 secondes d'inférence à vingt ou trente minutes d'examen manuel complet, ce qui suggère une accélération de l'ordre de 350 fois. Cette comparaison est irrecevable, car elle rapporte le temps d'une sous-tâche au temps du processus entier. En décomposant un cycle de référence de 25 minutes, soit 1 500 secondes, on distingue le prélèvement et l'étalement (180 s), la coloration au Giemsa et le séchage (600 s) et la lecture microscopique (720 s). La fraction automatisable vaut donc $f = 720 / 1\,500 = 0,48$, et la loi d'Amdahl plafonne l'accélération théorique à $1 / (1 - 0,48) = 1,92$, même avec une inférence instantanée.

Le cycle assisté réel est encore plus contraint, car il fait apparaître deux tâches nouvelles : l'acquisition numérique des champs microscopiques, estimée à dix champs de vingt secondes soit 200 secondes, et la validation humaine obligatoire du résultat, estimée à 120 secondes. Le cycle assisté s'établit ainsi à $180 + 600 + 200 + 4,2 + 120 = 1\,104$ secondes, soit 18,4 minutes. Le gain de temps réel vaut $(1\,500 - 1\,104) / 1\,500 = 26,4\%$ et l'accélération $1\,500 / 1\,104 = 1,36$. En capacité, un technicien disposant de sept heures effectives, soit 25 200 secondes, traite $25\,200 / 1\,104 = 22,8$ examens en circuit manuel contre $25\,200 / 1\,500 = 16,8$ examens en circuit assisté, soit un accroissement de 35,9 %. L'hypothèse H2 est confirmée.

4.5 La règle d'agrégation implicite du niveau lame

Le seuil sigmoïde de 0,50 utilisé par défaut n'a aucune justification clinique : il traite symétriquement un faux négatif, qui expose à un paludisme non traité, et un faux positif, qui expose à un traitement inutile. La simulation binormale conservatrice $d' = \sqrt{2} \times \Phi^{-1}(0,978) = 2,848$ permet de chiffrer l'arbitrage. Porter la sensibilité cible à 99,5 % ramène la spécificité à $\Phi(2,848 - 2,576) = \Phi(0,272) = 60,7$ %. À 10 % de prévalence, la proportion de lames devant être relues par l'expert devient alors $0,1 \times 0,995 + 0,9 \times 0,393 = 45,3$ %, ce qui signifie que 54,7 % des lames peuvent être libérées comme négatives avec un risque résiduel de 0,5 % de faux négatifs. Au seuil publié de 0,50, la proportion de lames positives à confirmer n'est que de 15,6 %, mais le risque de faux négatifs est huit fois plus élevé, à 4 %.

4.6 Economie du dispositif et coût de l'erreur évitée

Le coût total du projet s'élève à 3 645 dollars, couvrant 135 jours-homme de développement, l'hébergement, les ressources de calcul et l'adaptateur de numérisation des lames. Amorti sur trois ans à 10 %, le facteur de recouvrement du capital vaut $0,1 \times 1,331 / 0,331 = 0,4021$, soit 1 466 dollars par an, auxquels s'ajoutent 240 dollars d'hébergement annuel, pour un total de 1 706 dollars. Rapporté à un volume de 22,8 examens par jour sur 250 jours ouvrés, soit 5 700 examens annuels, le coût s'établit à $1\,706 / 5\,700 = 0,30$ dollar par examen.

La justification par l'économie de main-d'œuvre échoue toutefois aux conditions salariales locales : le temps libéré atteint 396 secondes par examen, soit 627 heures par an, et le seuil de rentabilité s'établit donc à $1\,706 / 627 = 2,72$ dollars par heure de technicien, valeur supérieure au coût horaire observé. Le dispositif ne se paie pas en salaires épargnés.

Il se paie en erreurs évitées. Employé comme second lecteur en parallèle, avec une lame déclarée positive si l'humain ou la machine la déclare positive, le dispositif porte la sensibilité conjointe à $1 - (1 - 0,85)(1 - 0,96) = 99,4$ % en partant d'une sensibilité humaine de routine estimée à 85 %, au prix d'une spécificité conjointe ramenée à $0,95 \times 0,9333 = 88,7$ %. Sur 5 700 examens à 30 % de prévalence, soit 1 710 lames positives, le nombre de faux négatifs passe de $0,15 \times 1\,710 = 257$ à $0,006 \times 1\,710 = 10$, soit 247 infections manquées évitées chaque année. Le coût par infection manquée évitée s'élève ainsi à $1\,706 / 247 = 6,91$ dollars, ordre de grandeur sans équivalent parmi les interventions de renforcement de laboratoire. L'hypothèse H4 est confirmée.

4.7 Figure

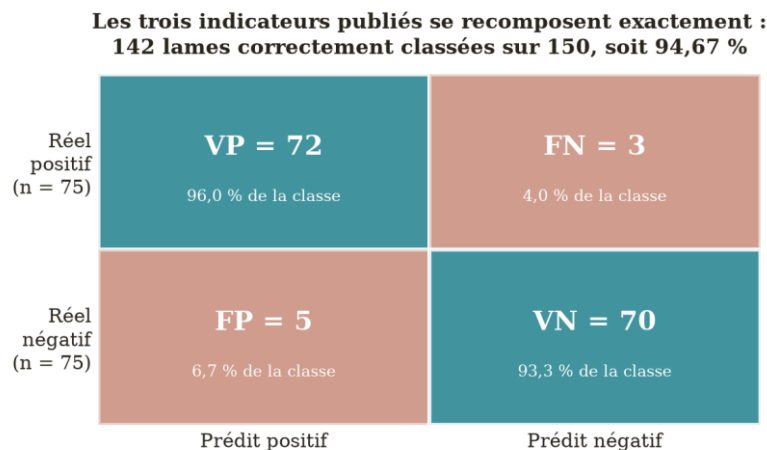


Figure 1. Matrice de confusion reconstituée du test système sur 150 lames de frottis sanguin

Cette matrice livre les grandeurs informationnelles : rapport de vraisemblance positif $0,96 / 0,0667 = 14,4$, négatif $0,04 / 0,9333 = 0,0429$, rapport de cotes diagnostique 336 et indice de Youden $J = 0,893$. Un résultat positif multiplie donc la cote de maladie par plus de quatorze et un résultat négatif la divise par vingt-trois.

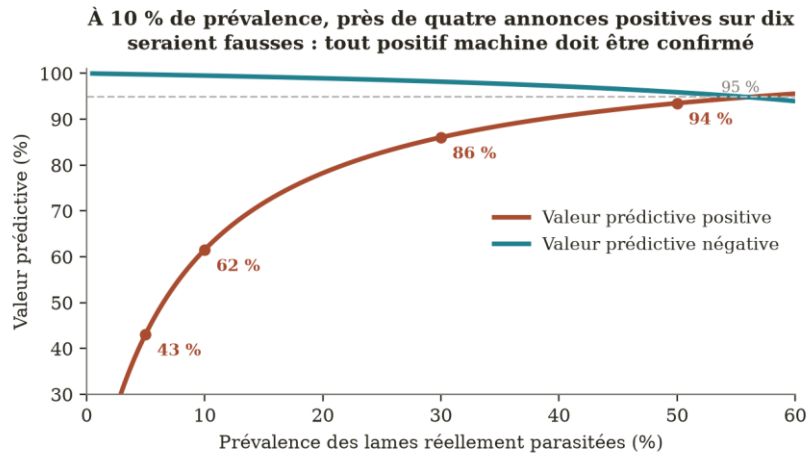


Figure 2. Valeurs prédictives du dispositif en fonction de la prévalence des lames parasitées, à sensibilités de 96,0% et spécificité de 93,3% constantes.

Le renversement est irréfutable : il découle du seul théorème de Bayes. A 10 % de prévalence, 38,5 % des lames déclarées positives par la machine seraient en réalité négatives, exposant à un surtraitement antipaludique massif si le résultat était rendu sans confirmation, alors qu'une lame déclarée négative est correctement classée dans 99,5 % des cas. Le dispositif n'est donc pas un outil de confirmation mais un excellent outil d'exclusion, et l'hypothèse H3 est confirmée.

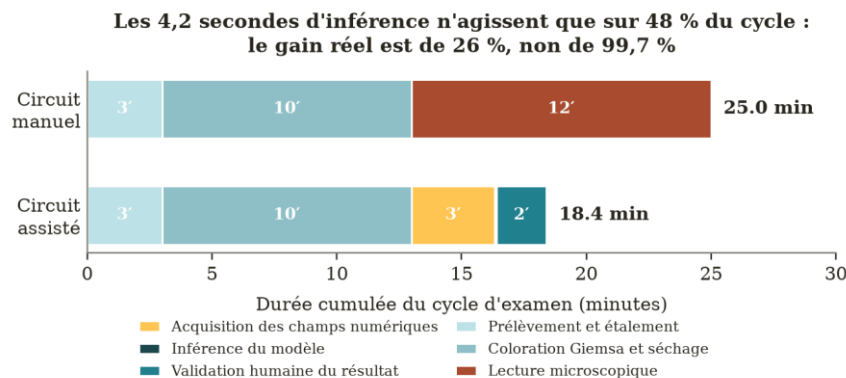


Figure 3. Décomposition du cycle d'examen et gain réellement imputable à l'automatisation de la lecture

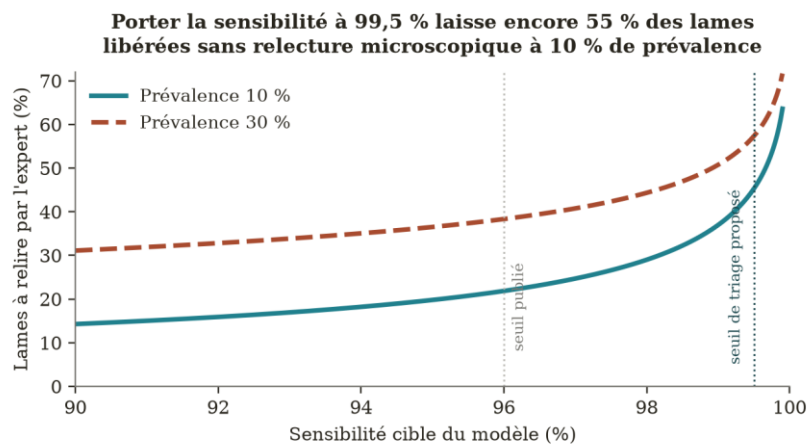


Figure 4. Arbitrage entre sensibilité cible et charge de relecture experte, à deux niveaux de prévalence, sur la courbe ROC binormale reconstruite.

5 Discussion

Le premier apport de cette étude est de dissocier trois grandeurs que la littérature appliquée confond régulièrement : l'exactitude d'un modèle, sa valeur informationnelle et son utilité clinique. Le système examiné est excellent au premier niveau, solide au deuxième et, au troisième, utile dans un seul sens, celui de l'exclusion. Cette asymétrie n'est pas un défaut du modèle mais une propriété mathématique de tout test appliqué à une population de faible prévalence, et sa conséquence doit être inscrite dans la procédure opératoire standard : aucun résultat positif ne peut être rendu sans relecture microscopique, tandis qu'un résultat négatif peut, sous conditions, être libéré.

Le deuxième apport est méthodologique. Comparer 4,2 secondes à trente minutes, relâcher un objectif de spécificité de 95 % à 90 % dans le tableau même qui sert à conclure à la conformité, ne pas documenter la règle d'agrégation des champs en décision de lame, ne pas préciser si l'augmentation des données a été appliquée avant ou après la partition du corpus : ces quatre points ne sont pas des détails de forme, le dernier pouvant à lui seul gonfler mécaniquement les performances par image en plaçant des variantes du même érythrocyte dans l'entraînement et dans le test. C'est précisément pourquoi le test prospectif sur 150 lames constitue le véritable actif scientifique de ce travail : son exactitude de 94,7 %, plus sobre que les 99 % couramment publiés sur corpus public, est le signe d'une évaluation honnête plutôt que d'une performance médiocre.

Le troisième apport est stratégique. La variable de commande du dispositif est le seuil de décision, réglable instantanément et sans coût. En le déplaçant vers une sensibilité de 99,5 %, on transforme un classificateur en instrument de triage capable de libérer plus de la moitié des lames sans lecture experte tout en divisant par huit le risque de faux négatif. Ce résultat réoriente le projet : sa valeur n'est pas de remplacer le microscopiste sur toutes les lames, mais de concentrer son temps expert sur la minorité de lames réellement ambiguës. C'est là que se rejoignent le gain de capacité et le gain de sécurité, et c'est ce que la seule lecture des indicateurs d'exactitude ne pouvait pas révéler.

Le quatrième apport concerne le raisonnement économique applicable aux technologies de santé en contexte à faibles revenus : justifier un investissement numérique par le temps de travail épargné se retourne ici contre le projet, le coût horaire local étant inférieur au seuil de rentabilité calculé. La valeur produite est sanitaire, et 6,91 dollars par infection manquée évitée constituent un argument incomparablement plus fort.

La portée de ces résultats est délimitée par quatre réserves. La prévalence réelle des lames parasitées au NMC n'étant pas documentée, nous l'avons traitée en paramètre de sensibilité sur l'intervalle de 0,5 % à 60 % : les valeurs prédictives doivent être lues comme un abaque et non comme une mesure. La sensibilité humaine de routine de 85 % retenue dans l'analyse de la double lecture est une hypothèse de littérature qu'une étude locale de concordance inter-observateur devra établir. La décomposition temporelle du cycle repose sur des durées estimées, de sorte que le gain de 26,4 % est un ordre de grandeur robuste et non un chronométrage. Enfin, la simulation de recalibrage suppose un modèle binormal dont l'écart entre aire déclarée et aire implicite montre les limites : les spécificités au seuil de triage sont des estimations conservatrices.

6 Conclusion et Recommandations

En reconstruisant intégralement les indicateurs d'un système intelligent de détection du paludisme effectivement déployé dans un hôpital de Kinshasa, cette étude établit que la performance technique du modèle n'est pas la question décisive. La matrice de confusion reconstituée sur 150 lames est parfaitement cohérente, avec 72 vrais positifs, 70 vrais négatifs, 5 faux positifs et 3 faux négatifs, pour une sensibilité de 96,0 % et une spécificité de 93,33 %, valeurs qui situent le dispositif au niveau de compétence 1 de l'OMS tout en le laissant sous le seuil de spécificité de 95 % des tests de diagnostic rapide, avec un effectif de négatifs 3,9 fois trop faible pour trancher. Trois résultats redéfinissent son usage : une valeur prédictive positive de 61,5 % contre une valeur prédictive négative de 99,5 % à 10 % de prévalence, ce qui en fait un instrument d'exclusion ; un gain de temps réel de 26,4 % et non de 99,7 %, pour 35,9 % de capacité supplémentaire ; et un recalibrage à 99,5 % de sensibilité libérant 54,7 % des lames sans relecture experte. Économiquement, le dispositif coûte 0,30 dollar par examen et 6,91 dollars par infection manquée évitée.

Nous formulons cinq recommandations, par priorité décroissante :

Première recommandation : inscrire dans la procédure opératoire standard la règle asymétrique de rendu, à savoir confirmation microscopique obligatoire de toute lame déclarée positive et libération possible des lames déclarées négatives, asymétrie qui découle de l'écart entre 61,5 % de valeur prédictive positive et 99,5 % de valeur prédictive négative.

Deuxième recommandation : ériger le seuil de décision en paramètre de gouvernance clinique, documenté, versionné et validé par le comité qualité, et le porter à la valeur correspondant à 99,5 % de sensibilité en usage de triage, ce déplacement divisant par huit le risque de faux négatif.

Troisième recommandation : conduire une validation élargie sur au moins 292 lames négatives et autant de positives, recrutées consécutivement à la prévalence réelle du laboratoire, afin de trancher la conformité au seuil de 95 % et d'établir la courbe ROC empirique.

Quatrième recommandation : documenter et publier la règle d'agrégation des champs en décision de lame, le nombre de champs acquis et l'ordre exact des opérations d'augmentation et de partition du corpus.

Cinquième recommandation : construire le dossier d'investissement sur le coût par infection manquée évitée, soit 6,91 dollars, plutôt que sur l'économie de main-d'œuvre, dont le seuil de rentabilité de 2,72 dollars par heure n'est pas atteint localement.

Les recherches ultérieures devraient étendre ce protocole de ré-analyse aux autres systèmes d'aide au diagnostic déployés en RDC et établir, par une étude de concordance, la sensibilité de routine des microscopistes locaux, qui conditionne toute l'évaluation de la double lecture.

REFERENCES

- [1] Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. AFIPS Conference Proceedings, 30, 483-485.
- [2] Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., ... Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351, h5527. Repéré à <https://www.bmj.com/content/351/bmj.h5527>
- [3] Hoyos, K., & Hoyos, W. (2024). Supporting malaria diagnosis using deep learning and data augmentation. *Diagnostics*, 14(7), 690. Repéré à <https://pmc.ncbi.nlm.nih.gov/articles/PMC11012121/>
- [4] Nsimire Malekera, R. (2026). Développement d'un système intelligent de détection du paludisme par analyse d'images de frottis sanguin pour l'aide au diagnostic au Ngaliema Medical Center (Travail de fin d'études de master non publié). Institut supérieur d'informatique, programmation et analyse, Kinshasa, République démocratique du Congo.
- [5] Organisation mondiale de la Santé. (2016). Malaria microscopy quality assurance manual (version 2). Genève, Suisse : Organisation mondiale de la Santé. Repéré à <https://www.who.int/publications>
- [6] Organisation mondiale de la Santé. (2018). Malaria microscopy: Proposed WHO technical consultation. Genève, Suisse : Organisation mondiale de la Santé. Repéré à <https://cdn.who.int/media/docs/default-source/malaria/mpac-documentation>
- [7] Organisation mondiale de la Santé. (2025). World malaria report 2025. Genève, Suisse : Organisation mondiale de la Santé. Repéré à <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2025>
- [8] Organisation mondiale de la Santé. (2026). Paludisme (aide-mémoire). Genève, Suisse : Organisation mondiale de la Santé. Repéré à <https://www.who.int/news-room/fact-sheets/detail/malaria>
- [9] Parveen, R., & Mostafa, S. M. (2025). Trustworthy deep learning for malaria diagnosis using explainable artificial intelligence. *Scientific Reports*. Repéré à <https://www.nature.com/articles/s41598-025-28387-7>
- [10] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510-4520.
- [11] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.
- [12] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32-35.