

Detection of Newcastle Disease in Chicken from Faecal Images using a Hybrid Xception-XGBoost Framework

^{*1}Bashar Raheemat Abiola, ¹Gbolagade Morufat Damola,
^{1,2}Adefila Opeyemi Ademola and ¹Bakare Mopelola Mistura

¹Department of Computer Science, Faculty of Computing, Engineering and Technology, Al-Hikmah University, Ilorin, Kwara State, Nigeria

²Agriculture and Rural Management Training Institute (ARMTI), KM 1 Ilorin, Ajase-Ipo road, Kwara State.

*Corresponding authors' email: basharraheemat@gmail.com

ABSTRACT

Newcastle disease (ND) is a serious viral disease of poultry. The high contagious rate and economic cost is mainly alarming due to its high mortality rate and negative economic impact on farmers. This study put forward a hybrid framework based on the poultry (chicken) faecal images to detect Newcastle disease. The framework (Xception+XGBoost) extract features using a pre-trained Xception CNN and classifies the diseased samples using an XGBoost classifier. For tackling the class imbalance issue between ND and Non-ND (Healthy and other diseased samples) multiple image augmentation techniques, class-weighted loss function, and feature-level oversampling of ND samples were used. It was specifically tuned to classify accurate diseases diagnosis with a decision threshold of 0.429. The test set results show an overall accuracy of 98.10% set in a held test set. The model showed a recall of 88.24% (75 true positive cases were detected out of 85) and precision of 85.23% for the minority Newcastle disease class and an F1-score of 86.71%. The Matthews Correlation Coefficient (MCC) of 85.70% confirmed balanced performance across both classes. The further evaluation resulted a promising Precision-Recall Area Under the Curve (PR-AUC) of 95.50%, which indicates potential to tackle class imbalance. The results suggest that the suggested structure offers effective solutions to the poultry disease detection and can be a useful computer-aided screening tool.

Received: 27 Aug. 2026

Accepted: 18 Sept. 2026

Published: 29 Oct. 2026

Keywords:

Newcastle Disease, Poultry Disease Detection, Class Imbalance, Xception, XGBoost, Deep Learning

INTRODUCTION

Poultry farming is one of the lucrative agricultural enterprises in Nigeria, contributing tremendously to the nation's food supply, employment, and overall rural development. According to Poultry Association of Nigeria (PAN, 2025), the Nigerian poultry industry is valued with a net worth of about ₦12 trillion (about \$8 billion United States dollars), over 20 million jobs and contributes to 25% of Nigeria's agricultural GDP. However, losses due to poultry diseases have persisted and Newcastle disease (ND) is one of the most contagious and deadly diseases. An avian paramyxovirus type 1 is responsible for ND, a highly virulent virus in birds, such as chickens, which can lead to respiratory, gastroenteric and neurological issues, often resulting in high mortality rate and substantial economic losses in poultry production (Berihulay et al., 2025).

The current diagnosis of ND is done by laboratory methods like the Polymerase Chain Reaction (PCR) assays with results that can be obtained in 3-7 days after analysis, which causes delays in the treatment that becomes critical for the onset of treatment and can cause mortalities, which ranges from 10-30 percent (Abdisa & Tagesu, 2017; Wong et al., 2017). The use of deep learning based approaches, specifically deep convolutional neural networks (CNNs) architecture, such as Xception (Chollet, 2017), has shown promising results for the automated diagnosis of faecal diseases based on image analysis, but the studies are still affected by the extreme class imbalance; ND images containing a significantly lower number of samples than healthy and other disease images, and hence, suffer bias, resulting in an "accuracy paradox", where

the classifier achieves high overall accuracy by focusing its attention on the majority classes, while failing to accurately identify many true ND images. For instance, the overall accuracy of 93% is reported by Raji-Lawal et al. (2026) with only 46% of ND cases identified from faecal images, missing over 50% of the infected birds. Akomolafe and Medeiros (2021) also had issues with class imbalance, despite their high accuracy (98%). The study of class-imbalanced learning has shown that hybrid level (data level and algorithm level) methods are effective to solve class imbalance issues (Leevy et al., 2018).

In recent studies, hybrid CNN-XGBoost methods have been investigated, integrating CNN image feature extraction with XGBoost classification. For smart agriculture, Kaviarasu et al. (2025) proposed a hybrid CNN-XGBoost framework, which was used to classify the disease using a gradient-boosted tree classification system and deep-learning-based feature extraction. Similarly, Rana et al. (2026) proposed a CNN-XGBoost model for disease detection in leafy plants with overall accuracy of 97.2% combining CNN-based feature extraction and XGBoost classification. The studies show the feasibility of using CNN-based representation learning and gradient-boosted tree classification in agriculture image-based disease detection. The applications of hybrid CNN-XGBoost framework in Newcastle Disease detection from poultry faecal images, however, have been limited, especially when the minority class makes up just 7% of the dataset for severe class imbalance. In another study, Mukhtar et al. (2025) fused the features extracted by ResNet-50 and VGG16 with a Random Forest classifier for age estimation

from images of the sclera. A similar example of combining deep image representation and traditional machine-learning classification system, where the application is different than poultry disease detection.

This study proposes a novel hybrid Xception-XGBoost framework, which utilizes targeted image augmentation, class-weighted cost-sensitive fine-tuning and feature-level oversampling to specifically address the need for an enhanced recall rate and other imbalance-sensitive metrics for the minority ND class without compromising classification performance. The specific objectives were to:

- address class imbalance in Newcastle disease detection using a multi-layered balancing strategy (image augmentation, class-weighted cost-sensitive learning, and feature-level oversampling);
- fine-tuned a pre-trained Xception convolutional neural network (CNN) and extracted feature representations from poultry faecal images;
- classify the extracted features in (ii) using XGBoost classifiers; and
- evaluate the model's performance using Recall, Precision, F1-score, Area Under the Precision-Recall Curve (AUPRC) and Matthews Correlation Coefficient (MCC).

MATERIALS AND METHODS

Research Design

The study used experimental research design involving classification model for Newcastle disease (positive class) and all other classes (negative class) while developing and testing the model on images of poultry faeces with the gold standard label using non-randomized and retrospective study design.

Proposed Framework

The proposed framework for the detection of Newcastle Disease is shown in Figure 1. Poultry diseases images dataset was collected from the Zenodo repository. Following this, the images were preprocessed by resizing and normalizing pixel values. The dataset was then divided into training, validation and test set using stratification to maintain class distribution. The training ND images set underwent image augmentation, which facilitated the selective fine-tuning of Xception backbone, then the feature vectors of 2,048 dimensions are extracted through global average pooling and enhanced via feature-level sampling. An XGBoost classifier is trained on this balanced feature set and lastly, the model's performance is evaluated on the test set using various metrics including Recall, Precision, F1-score, AUPRC and MCC.



Figure 1: Proposed Framework for Detection of Newcastle Disease

Dataset

The study employed the Zenodo poultry faeces image database that contained two main datasets: the Farm-Collected Dataset and the PCR-Validated Dataset (Machuve et al., 2021; Machuve et al., 2022). Farm-Collected contained 6,812 images of poultry faeces collected in Tanzania and PCR-Validated consisted of 1,255 images of poultry faeces with laboratory confirmed diagnosis, resulting in 8,067 images of four classes: Healthy (2,404), Coccidiosis (2,476),

Salmonella (2,625) and Newcastle Disease (562). Thus, the prevalence rate of ND was 7.0% with a ratio of 1:14 of the number of ND cases to all the other conditions (Table 1). The current study combined the disease classes into Healthy (+0), Coccidiosis (+0) and Salmonella (+0) classes ($n = 7,505$) and Newcastle Disease (+1) class ($n = 562$), where +0 represents negative and +1 represents positive. Table 1 below presents a summary of the distributions of the classes.

Table 1: Dataset Schema and Class Distribution

Class	Description	Count	Percentage
Healthy	Normal poultry faecal images	2,404	29.8%
Coccidiosis	Coccidiosis-infected faecal images	2,476	30.7%
Salmonella	Salmonella-infected faecal images	2,625	32.5%
Newcastle Disease	ND-infected faecal images	562	7.0%
Total	Complete dataset	8,067	100.0%

Newcastle Disease samples have a greenish or yellowish-green mucoid discolouration, compared to the brown solid healthy faeces and the bloody/creamy discharge of Coccidiosis and Salmonella, respectively, the sample texture is watery and diarrheic, and the margins are irregular, with

liquid boundaries scattered around. A binary classification was used for classification: ND = 1 (positive class), Salmonella, Coccidiosis, and Healthy = 0 (negative/non-ND class).



Figure 2: Healthy vs. Newcastle Disease (and Other Disease Class) Faecal Images

Data Splitting

To prevent data leakage and maintain the extreme class imbalance in all subsets, we stratified and randomly sampled the dataset to create a training set (70%, $n \approx 5,647$, ND = 394, non-ND = 5,253), a validation set (15%, $n \approx 1,210$, used for hyperparameter and threshold selection, ND = 83, non-ND = 1,127), and a held-out test set (15%, $n \approx 1,210$, ND = 83, non-ND = 1,127) (Kohavi, 1995). Following the split, 393 real ND images were used for training (one image discarded following the pre-processing). All of the images were resized to the input size of Xception.

Importantly, dataset splitting was done prior to data augmentation and feature level oversampling. Here, oversampling was only applied to the training partition and no feature vectors were added to the validation or test partition, and no oversampling operation was performed on the samples. The class distributions and the original images were kept in both partitions. The validation set was used for hyperparameter tuning and optimising the decision threshold, but the held-out test set was not used until the end of the model evaluation to evaluate generalisation to unseen data.

Data Preprocessing

All images were resized to 299×299 pixels to ensure that they could be fed to the input of Xception, with bicubic interpolation to preserve the chromatic gradients relevant to mucoid ND faeces. All the pixels were then scaled to 0 to 1 by dividing by 255 to prevent fluctuations in convergence during fine tuning.

Image Augmentation and Class-Imbalance Handling

A multi-layered approach, implemented at the image, model and feature level, was used to deal with the class imbalance. Targeted image augmentation was performed only on the Newcastle Disease training images, including random rotation ($\pm 15^\circ$), horizontal flipping ($p = 0.5$), random zoom

($0.8\text{--}1.2\times$) and brightness scaling ($0.8\text{--}1.2\times$), but bounded to ensure the greenish-yellow ND chromatic signature was not altered, as is the case in the other classes. Images that were not NDS were not enhanced.

Second, the Xception backbone (pre-trained on ImageNet) was maintained frozen, except for the last two convolutional blocks which were fine-tuned with a class-weighted binary cross-entropy loss, the learning rate of 1×10^{-5} , and early stopping according to the validation loss. This was an attempt to preserve generic representations from ImageNet in the first layers of the network and to fine-tune the final layers using training data from diseases.

Thirdly, after fine-tuning, an augmentation-based feature level oversampling technique was conducted entirely on the train partition. Oversampling in feature spaces has been explored in previous studies to solve the class imbalance problem (Pérez-Ortiz et al., 2016), and Ando and Huang (2017) introduced Deep Over-sampling (DOS), which is based on synthetic embedding targets in a CNN derived feature space. The approach used here differs: the additional feature vectors were created by passing the real Newcastle Disease (ND) training images through a fine-tuned Xception network under random augmentations. Each of the 393 original ND training images were passed through the network 10 times, resulting in 3,930 feature vectors of 2,048 dimensions, while majority-class images were processed once. Unlike SMOTE, no interpolation between samples was performed.

The aim of the feature-level oversampling method was to increase the minority class by increasing the number of feature vectors, but simultaneously also making sure that the extracted feature vectors were always correlated with a real training image, rather than using a new generative model. The study therefore does not claim superiority over SMOTE, GAN-based, or any other oversampling approach, as comparison with these methods were outside the scope of the

reported experiments and are reserved for future work. The 3,930 vectors are not independent observations, but are derived from only 393 source images, which can have an effect on how the XGBoost classifier learns and could be a limitation of the approach.

Furthermore, the validation and test sets were made up of only real images, non were augmented. Therefore, any overfitting due to the training vectors being correlated should be expected to decrease the generalisation to these real held-out images, and not artificially increase the test performance. The validation set was only used for the selection of models and the optimization of the decision threshold; the test set was reserved for the final evaluation of the performance of the models. XGBoost was not weighted by the algorithm level and scale_pos_weight was kept as default value 1.0.

Xception Feature Extraction

Xception (Chollet, 2017) was used as a static feature extractor. It features architecture based on depthwise separable convolutions that capture spatial features per

channel and then combine channel information in a pointwise (1×1) operation to reduce the number of parameters and the risk of overfitting. This network consists of 36 convolutional layers which are structured as Entry flow, Middle flow (repeated 8 time with skip connections), and Exit flow. Each of these $299 \times 299 \times 3$ images were flattened into a 2,048-dimensional feature vector $f = \text{GAP}(M_{\text{Xception}}(I_{\text{norm}}))$, which was used as an input representation for the classification head.

Classification Using XGBoost

An Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) high dimensional data efficient classifier was used to classify the 2048-dimensional feature vectors, as it is user-friendly even for high dimensional inputs, interpretable with respect to end-to-end deep network, and has built-in regularisation. Hyperparameters were hand chosen, according to the common approach and some experimentation (Table 2) with L2 regularisation ($\lambda = 1$, $\gamma = 0$) and early stopping based on the validation AUPRC (rather than accuracy or log loss, due to the severe imbalance).

Table 2: XGBoost Hyperparameter Configuration

Hyperparameter	Search Space	Selected Value	Description
n_estimators	100, 200, 300, 500	300	Number of boosting rounds (trees)
max_depth (d)	3, 4, 5, 6, 8	6	Maximum tree depth
learning_rate (η)	0.01, 0.05, 0.1, 0.2	0.05	Shrinkage per boosting round
subsample (s)	0.6, 0.7, 0.8, 1.0	0.8	Row subsampling ratio per tree
colsample_bytree (c)	0.5, 0.7, 0.8, 1.0	0.8	Column (feature) subsampling ratio
scale_pos_weight	1, 3, 5	1.0 (fixed)	Left unweighted; feature set already rebalanced by oversampling
max_delta_step	1, 2, 3, 4	1	Constrains per-iteration weight updates
objective	binary:logistic, reg:logistic	binary:logistic	Binary classification
eval_metric	logloss, aucpr, error	aucpr	Evaluation metric during training
random_state	0,42	42	Fixed seed for reproducibility
tree_method	hist	hist	Histogram-based tree construction

Max_Depth = 6 was Preferred over Deeper Trees (9, 12) to Mitigate Overfitting Given the Limited Original ND Sample Size ($n = 393$). The Decision Threshold was Optimised on the Validation Set via a Threshold Sweep, Selecting the Maximum Threshold for which ND Recall Remained $\geq 85\%$, which Yielded a Calibrated Threshold of 0.429 (rather than the Default 0.5), thereby Maximising Precision Subject to the Minimum-Recall Constraint Appropriate for Infectious-Disease Screening.

Evaluation Metrics

Standard accuracy is potentially misleading since it is clearly imbalanced (ND = 7% of samples, approximately 1:14 ratio) with potential false negatives jeopardising farm-wide outbreaks, and false positives impacting on correct identification of ND cases and the framework was evaluated using metrics that focus on the performance of the minority class: Recall/Sensitivity ($\text{TP}/(\text{TP}+\text{FN})$); Precision ($\text{TP}/(\text{TP}+\text{FP})$); F1-score (harmonic mean of precision and recall); Area Under the Precision–Recall Curve (AUPRC), preferred over ROC-AUC for imbalanced data because it emphasizes minority-class ranking performance; the Matthews Correlation Coefficient (MCC), a single balanced measure, that includes all four quadrants of the confusion matrix (range -1 to $+1$), and 0 indicates random guessing; and the confusion matrix itself, where True Positive/Negative and

False Positive/Negative are defined by correct/incorrect identification of ND and non-ND cases.

Random seeds were set in all experiments for reproducibility, and all experiments were run with the following versions of Python, TensorFlow, XGBoost, Scikit-Learn, OpenCV and Albumentations: 3.10, 2.20.0, 3.2.0, 1.6.1, 4.13.0 and 2.0.8, respectively.

RESULTS AND DISCUSSION

Image Augmentation and Feature Representation

Augmentation was applied only to minority training image with rotating (by $\pm 15^\circ$) and horizontally flipping minority training images, and scaling brightness (0.8-1.2) and zoom (0.8-1.2) was added, as this would increase the variation of the minority class while preserving the disease specific features. Some of the augmented ND samples are shown in Figure 3.



Figure 3: Representative Image Augmentations (Rotation, Brightness, Flip, Zoom)

The training and validation loss levels dropped from around 0.60 to around 0.02 and training and validation accuracy levels increased from around 70% to 99.3% and 90.0% to 97.4%, respectively (Figure 4) with no signs of overfitting.

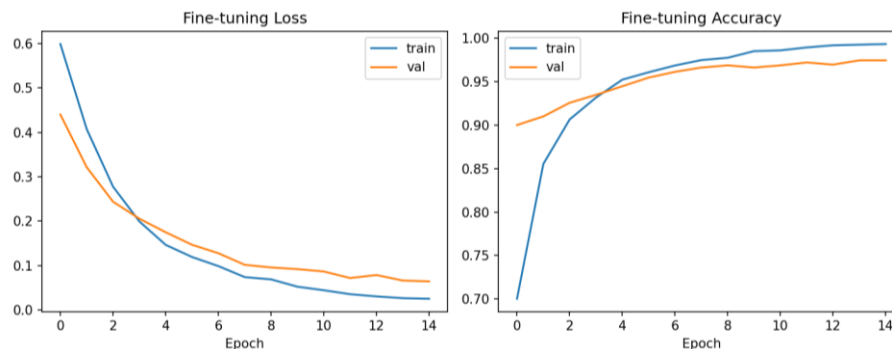


Figure 4: Training and Validation Loss and Accuracy Curves for Selective Xception Fine-Tuning

Since the number of the minority classes was small, the number of feature vectors was increased by applying the augmentation pipeline to the ND training images and fine-tuning the Xception network 10 times to increase the number

of feature vectors for the minority classes from 393 to 3,930 (Figure 5) without adding additional class weighting during the training of the network for the classification.

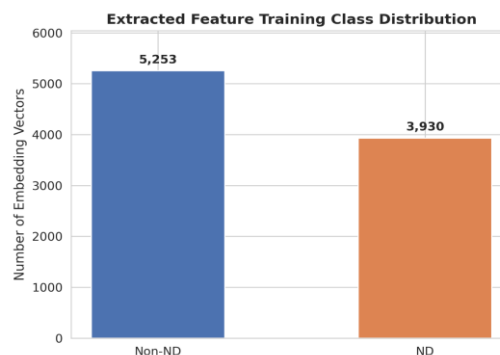


Figure 5: Minority-Class Representation Before and After Feature-Level Oversampling

XGBoost Hyperparameters, Threshold, and Feature Importance

The final setup (Table 2) was max_depth = 6, subsampling on a row/column basis of 0.8, L2 regularisation ($\lambda = 1$), and early stopping that stopped the training after the validation AUPRC failed to increase any further. A threshold sweeps on the validation set found that 0.429 was the optimum threshold which gave a minimal false-positive rate while maintaining an ND recall of 85% or higher.

The feature-importance analysis (Figure 6) showed that the Xception_Dim_981 dimension had a strong contribution followed by Xception_Dim_1979 dimension (0.071) and Xception_Dim_829 dimension (0.024) and a few more contributed slightly. The spread of this information over multiple complementary high-level visual elements (features) and not a single dominant feature that reflects disease-relevance, suggests that the fine-tuned Xception network acquired disease-relevant image representations.

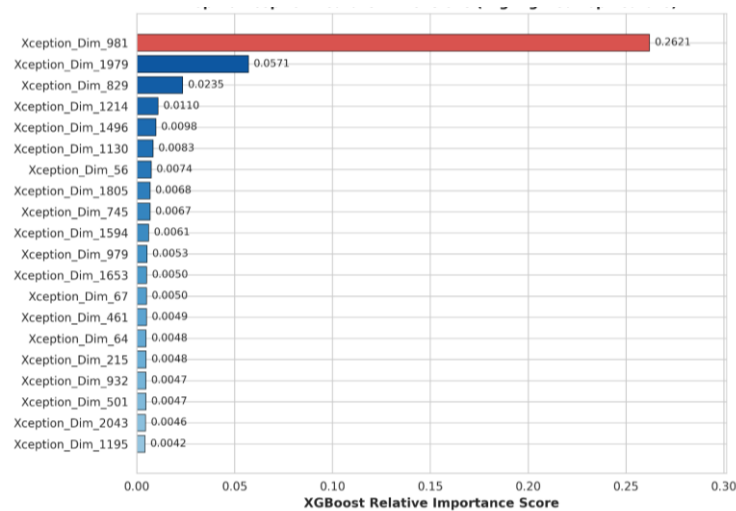


Figure 6: Top 20 XGBoost Feature Importance Rankings

Overall and Per-Class Test Set Performance

On the held-out test set (1,211 images: 1,126 Non-ND, 85 ND), the framework achieved 85.23% precision, 88.24% recall, an F1-score of 86.71%, a PR-AUC (AUPRC) of 95.50%, and an MCC of 85.70% at the calibrated 0.429

threshold (Table 3), correctly identifying 75 of 85 true ND cases (missing 10) while one in six ND alerts would be a false positive pending confirmatory testing, a deliberate trade-off, since in infectious-disease screening the cost of missing an outbreak outweighs the cost of false alarms.

Table 3: Final Test Set Performance Metrics (Decision Threshold = 0.429)

Metric	Value	Interpretation
Precision	85.23%	Of all images flagged as ND, 85.23% were truly ND
Recall (Sensitivity)	88.24%	Of all genuine ND images, the model correctly caught 88.24%
F1-score	86.71%	Balanced average of precision and recall
PR-AUC (AUPRC)	95.50%	Ranking performance focused on minority-class (ND) detection
MCC	85.70%	Single balanced score accounting for all four outcome types

Per-class performance (Table 4) shows non-ND precision/recall/F1 of 0.9911/0.9885/0.9898 (support 1,126) against ND precision/recall/F1 of 0.8523/0.8824/0.8671 (support 85), with overall accuracy of 0.9810, a macro-average F1 of 0.9284, and a weighted-average F1 of 0.9812.

The macro-averaged F1 (0.9284) is not the lowest, but it is very close to the per-class ND F1 (0.8671) indicating that a good performance of the model is not just an artefact of the accuracy paradox.

Table 4: Classification Report on the Test Set, by Class

Class	Precision	Recall	F1-score	Support (test images)
Non-ND (healthy/other disease)	0.9911	0.9885	0.9898	1,126
ND (Newcastle disease)	0.8523	0.8824	0.8671	85
Overall Accuracy	—	—	0.9810	1,211
Macro Average (unweighted)	0.9217	0.9354	0.9284	1,211
Weighted Average	0.9814	0.9810	0.9812	1,211

This confusion matrix (Table 5, Figure 7) verifies 75 true positives, 10 false negatives, 13 false positives and 1,113 true negatives. The results of 1,126 non-ND samples show a

1.24% false alarm rate, with a specificity of 98.5% which indicates a balance between sensitivity and unnecessary notifications.

Table 5: Confusion Matrix of the Proposed Model (Test Set)

Actual \ Predicted	Non-ND	ND	Row Total
Non-ND	1,113 (True Negative)	13 (False Positive)	1,126
ND	10 (False Negative)	75 (True Positive)	85
Column Total	1,126	85	1,211

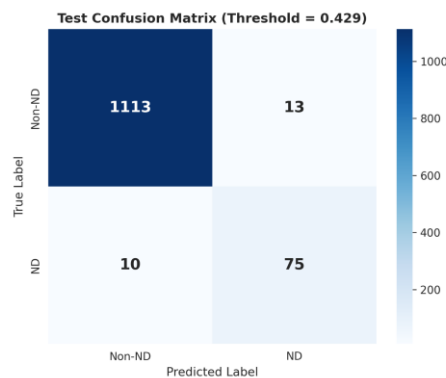


Figure 7: Confusion Matrix

The Precision-Recall curve (PR-AUC) in Figure 8 achieved a good ranking performance with PR-AUC of 95.50%, with higher precision at higher recall, which is beneficial in

situations where the data set is highly imbalanced and targets the minority class.

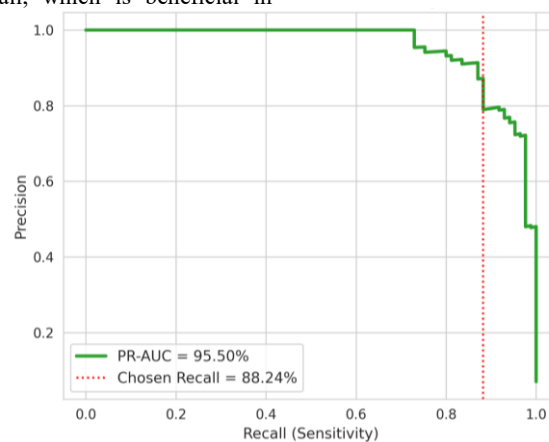


Figure 8: Precision-Recall Curve on the Held-Out Test Set

Inference Efficiency

The deployment constraints were limited by the feature-extraction step (Xception), and not the decision step, with a

saved model size of just 0.61 MB (Table 6), and an inference time of just 0.0184ms per image (Table 6).

Table 6: Model Inference Efficiency

Metric	Value
Inference Speed (XGBoost classification stage)	0.0184 milliseconds per image
Saved XGBoost Model Size	0.61 MB

Comparison with Baseline and Related Studies

Table 7 compares the proposed framework, before and after applying the multi-layered imbalance-handling strategy. In the pre-balance scenario, the classifier obtained a better precision (98.36%) but a significantly low recall (70.59%) this is a typical accuracy paradox. The accuracy of recall

stayed at 88.24% after balance and F1 score and MCC increased, respectively, from 82.19% to 86.71% and 82.35% to 85.70%, while precision experienced a slight decrease (98.36% to 85.23%) but is still a good trade-off, due to the higher clinical and economic burden of missed ND cases.

Table 7: Comparative Performance (Baseline vs. Proposed)

Configuration	Recall	Precision	F1	MCC
Before Balancing (Baseline)	70.59%	98.36%	82.19%	82.35%
After Balancing (Proposed)	88.24%	85.23%	86.71%	85.70%

The results are compared with similar studies in the literature given in Table 8. Machuve et al. (2022) reported an overall accuracy of 98.24%, however they did not present full details of precision/recall on the ND. Raji-Lawal et al. (2026) obtained high precision (96.00%) and low recall (46.00%) thus, there were many cases of ND that were missed. Akomolafe and Medeiros (2021) "Recalling and F1 score levels were higher, under different experimental conditions.

The overall classification performance by the proposed Xception–XGBoost frame work was found with a balanced profile with precision (85.23%), recall (88.24%), F1-score (86.71%) and accuracy (98.10%), which shows that the frame work is good for improving ND detection under severe class imbalance and maintaining high overall classification performance.

Table 8: Comparison with Other Studies

Author (Year)	Model Architecture	ND Precision	ND Recall	ND F1-Score	Overall Accuracy
Akomolafe & Medeiros (2021)	CNN (Model 2)	100.00%	93.20%	96.50%	96.60%
Machuve et al. (2022)	Xception (Fine-tuned)	Not Stated	>85.0%	98.00%	98.24%
Raji-Lawal et al. (2026)	EfficientNetB4	96.00%	46.00%	63.00%	92.69%
Proposed Study	Xception + XGBoost	85.23%	88.24%	86.71%	98.10%

Discussion

From the outcome results, it is observed that the proposed hybrid Xception–XGBoost model can identify the Newcastle Disease in poultry faecal pictures with no issue even if the class class imbalance is 1:14. Targeted image augmentation, selective Xception fine-tuning, feature-level oversampling and XGBoost classification significantly boosted the minority-class detection while not significantly decreasing the overall classification results. The accuracy paradox you observe in the baseline configuration is due to it having a high precision (98.36%) and low recall (70.59%) which means that the model has a tendency to predict the majority class and is susceptible to the omission of active infections. The multi-layered balancing approach improved recall (88.24 %) which is the most clinically relevant indicator for such a highly contagious disease as ND, with a lower false negative rate being important to prevent the disease spreading like a wildfire throughout the flock and almost to the end of the farm. This decrease in precision with calibration was deliberate: the threshold (0.429) was chosen to increase clinical sensitivity, and from a veterinary point of view, the relatively small number of false alarms (1.24 % false alarm rate) is not as important as the economic impact of missing an outbreak.

CONCLUSION

This study revealed that class imbalance is the key issue for accurate Newcastle Disease diagnosis using faecal images from poultry rather than the image-based deep learning approach. The proposed Xception-XGBoost framework resulted in an over 40 percentage point improvement in ND recall when compared with a literature baseline (46% recall) and had an overall reliability of 85.70% on a held-out test set with high precision (85.23%), recall (88.24%), F1-score (86.71%), AUPRC (95.50%) and MCC (85.70%). The results indicated that this framework can be a reliable computer-aided screening tool to identify, intervene and improve the management of the disease in poultry production. The framework is still to be validated with a larger, more geographically diverse collection of data, and will extend to support multi-class classification, testing for generalisability to the real world, deployment to the edge to reduce processor memory and power consumption, and ultimately to integrate into the explainable AI approach and multimodal disease-detection strategies, but is a promising, practical, and memory- and power-efficient way of automated Newcastle Disease surveillance in resource-limited poultry farming settings.

REFERENCES

Abdisa, T., & Tagesu, T. (2017). Review on Newcastle disease of poultry and its public health importance. *Journal of Veterinary Science & Technology*, 8(3), 441. <https://doi.org/10.4172/2157-7579.1000441>

Akomolafe, O. P., & Medeiros, F. B. (2021). Image detection and classification of Newcastle and avian flu diseases infected poultry using machine learning techniques. *University of*

Ibadan Journal of Science and Logics in ICT Research, 6(1–2), 121–131.

Ando, S., & Huang, C.-Y. (2017). Deep over-sampling framework for classifying imbalanced data. In *Advances in Knowledge Discovery and Data Mining* (pp. 770–785). Springer. <https://arxiv.org/abs/1704.07515>

Berihulay, H., Luo, W., Lao, A., Ji, J., Cai, M., Shu, D., & Luo, C. (2025). Exploring the genetic basis of Newcastle disease virus in chickens: A comprehensive review. *Frontiers in Immunology*, 16, 1614794. <https://doi.org/10.3389/fimmu.2025.1614794>

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1800–1807). IEEE. <https://doi.org/10.1109/CVPR.2017.195>

Kaviarasu, K., Immanuel, J. P., Sanjay, R., & Mohan, R. (2025). Revolutionizing crop health: CNN and XGBoost for accurate disease detection in smart agriculture. In *2025 International Conference on Advanced Computing Technologies (ICoACT)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICoACT63339.2025.11005266>

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143). Morgan Kaufmann Publishers.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5, 42. <https://doi.org/10.1186/s40537-018-0151-6>

Machuve, D., Nwankwo, E., Lyimo, E., Maguo, E., & Munisi, C. (2021). *Machine Learning Dataset for Poultry Diseases Diagnostics - PCR annotated [Dataset]*. Zenodo. <https://doi.org/10.5281/zenodo.4495535>

Machuve, D., Nwankwo, E., Mduma, N., & Mbelwa, J. (2022). Poultry diseases diagnostics models using deep learning. *Frontiers in Artificial Intelligence*, 5, 733345. <https://doi.org/10.3389/frai.2022.733345>

- Mukhtar, A. S., Ahmad, M. A., Ibrahim, M., Abdulkadir, S., Mu'azu, A., Dauda, S., & Diso, A. (2025). Enhancing age estimation from sclera images using ResNet-50, VGG16, and Random Forest. *FUDMA Journal of Sciences*. <https://doi.org/10.33003/fjs-2025-0904-3487>
- Pérez-Ortiz, M., Gutiérrez, P. A., Tiño, P., & Hervás-Martínez, C. (2016). Oversampling the minority class in the feature space. *IEEE Transactions on Neural Networks and Learning Systems*, 27(9), 1947–1961. <https://doi.org/10.1109/TNNLS.2015.2461436>
- Poultry Association of Nigeria. (2025, February 6). *The Poultry Association of Nigeria at IPPE 2025: Expanding horizons for the Nigerian poultry industry*.
- Raji-Lawal, H. Y., Aiyeniko, O., Sotonwa, K., Adegbuyi, D. G., & Ajisegiri, R. O. (2026). Poultry disease detection using deep learning for fecal image classification. *Dutse Journal of Pure and Applied Sciences (DUJOPAS)*, 12(1b), 228–238. <https://dx.doi.org/10.4314/dujopas.v12i1b.17>
- Rana, S., Paul, S., Ghosh, S., Kumar, S., Chakraborty, B., & Moitra, S. (2026). Hybrid CNN model for detection of diseases in leafy plants. *ITM Web of Conferences*, 86, 01015. <https://doi.org/10.1051/itmconf/20268601015>
- Wong, J. T., de Bruyn, J., Bagnol, B., Grieve, H., Li, M., Pym, R., & Alders, R. G. (2017). Small-scale poultry and food security in resource-poor settings: A review. *Global Food Security*, 15, 43–52. <https://doi.org/10.1016/j.gfs.2017.04.003>

