# Hybrid Random Forest, XGBoost and LSTM Models with Explainable AI for Respiratory Risk Classification in Poultry Housing and Smart Ventilation Control

**Oklo David Anzaku[1]*, Ridwan Kolapo[1], Ibrahim Anka Salihu[2]**

**[1]Department of Information Technology, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria**

**[2]Department of Software Engineering, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria**

***Corresponding Author**

**DOI: https://doi.org/10.51244/IJRSI.2026.1310000094**

**Received: 15 October 2026; Accepted: 20 October 2026; Published: 02 November 2026**

## ABSTRACT

Respiratory disease is one of the most expensive problems in commercial poultry production, and on most farms it is still handled after the fact. Birds are treated once they are already coughing, gasping or dying, by which point the tracheal lining has usually been damaged. This study asks a narrower question than clinical diagnosis: whether machine-learning classifiers can recover predefined environmental risk states from the housing variables that poultry farms already measure, namely temperature, relative humidity, carbon dioxide, ammonia and airflow, and whether a recovered state can be mapped onto a ventilation instruction. A dataset of 10,000 consecutive hourly records was generated programmatically and labelled into four environmental risk states: healthy, heat stressed, mild respiratory risk and severe respiratory risk. The labels are derived from thresholds on the environmental variables themselves rather than from veterinary assessment, so the task is the classification of environmental risk and not the detection of clinical disease. Random Forest reached 99.75% test accuracy and XGBoost 99.70%, neither missing a severe respiratory case. A two-layer LSTM reached 59.66%, the majority-class rate, and class weighting did not rescue it. Soft voting across the two tree models matched the best single model at 99.75% without improving on it, and adding the LSTM lowered accuracy to 99.70%, so Random Forest is the preferred classifier under these conditions. Two baselines place that in context: a single decision tree reached 99.60% and logistic regression 98.05%. SHAP and LIME show both tree models relying mainly on carbon dioxide and ammonia, which is expected because the generator was parameterised from the literature linking those gases to respiratory damage. Accuracy at this level reflects recovery of a known generating process and is not an estimate of field performance. All results come from simulated data and require validation against instrumented houses and veterinary records.

**Keywords:** Random Forest; XGBoost; Explainable AI; poultry housing environment; environmental risk classification; simulation study

## INTRODUCTION

Poultry is the cheapest reliable source of animal protein in Nigeria and across much of West Africa, and for a large number of rural households it is also the main source of cash income. That makes flock health an economic question as much as a veterinary one. Respiratory infection is the part of that question farmers lose most money on. Poor air in the house is a large part of that loss. Kristensen and Wathes [1] review evidence that atmospheric ammonia irritates the respiratory tract and raises susceptibility to disease at or below the 25 ppm limit commonly applied, a limit set for human safety rather than bird welfare, and Miles et al. [2] found that broilers reared at 50 and 75 ppm ammonia weighed 6% and 9% less at four weeks.

The mechanism is well documented and, importantly for this work, it is environmental. Ammonia released by decomposing litter accumulates when airflow is inadequate and irritates the lining of the respiratory tract [1].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

Carbon dioxide rises in the same conditions and indicates that the air in the house is not being exchanged. High temperature combined with high humidity pushes birds into heat stress, since chickens rely heavily on panting to shed heat [3]. None of this is sudden. Conditions inside a house degrade over hours or days, and by the time a bird shows visible signs the exposure has already been going on for a while.

Disease management on most farms is reactive: the farm manager acts on coughing, nasal or ocular discharge, lethargy or a jump in the mortality count. Sensing itself is no longer the obstacle. Elwakeel [4] built an IoT unit that measures temperature, humidity, ammonia and methane for about 12% of the cost of a commercial reference device, but like most such systems it acts on fixed thresholds and not on any prediction of bird health. Machine learning is an obvious candidate for closing that gap [5]. Random forests and boosted trees are well established for structured agricultural data [6], and recurrent networks have detected livestock disease from sensor time series before clinical signs: Teixeira et al. [7] predicted anaplasmosis in dairy calves up to three days before diagnosis using LSTM models on rumination and activity data.

Two problems remain. The first is trust. A model that returns a risk score without saying why will not be acted on by a farm manager who has no background in machine learning, and Manikandan and Neethirajan [8] list insufficient model interpretability among the main obstacles to deploying AI monitoring on poultry farms. The second is that most published systems stop at prediction. They tell the farmer that risk is elevated and leave the response entirely to human judgement, which defeats the purpose of early warning when the required action, moving more air through the house, is mechanical and can be specified precisely.

This study develops and evaluates a hybrid framework that classifies environmental respiratory risk states from housing variables, explains each classification in terms a farm operator can check, and maps the classification onto a ventilation instruction. Random Forest, XGBoost and an LSTM are combined by soft voting for four-class classification; SHAP and LIME supply both dataset-wide and case-level attribution; and a decision layer maps the predicted class and its confidence onto airflow, temperature and humidity adjustments. The fusion is tested against its own components and against two conventional baselines, a step most published hybrid systems omit. Two boundaries on the work should be stated at the outset. The risk labels are derived from thresholds on the environmental variables rather than from veterinary assessment, so the study cannot establish early detection of clinical disease, only the recovery of predefined environmental states. And all data is simulated, with the ventilation response modelled rather than measured, so no hardware was built and no birds were involved. Section V sets out what those boundaries limit.

## RELATED WORK

### A. Environmental drivers of poultry respiratory disease

The interior of a poultry house directly shapes flock health. Ammonia from decomposing litter accumulates when airflow is inadequate. Kristensen and Wathes [1] review its effects, which include respiratory irritation, greater susceptibility to disease and reduced growth, with harm evident at or below the 25 ppm limit commonly applied. Miles et al. [2] measured the production cost in modern broilers, with body weight 6% and 9% lower at 50 and 75 ppm and mortality rising at the highest concentration, and Al Homidan et al. [9] review the combined effects of ammonia and dust, noting that high levels of inspirable dust cause respiratory problems in broilers. On this evidence, respiratory health in a poultry house is largely a question of air management.

### B. Machine learning for agricultural disease prediction

Liakos et al. [6] review machine learning across crop, livestock, water and soil applications, including ensemble methods such as random forests and boosting. In poultry, image-based diagnosis is the most developed strand: Machuve et al. [10] classified coccidiosis, salmonella and Newcastle disease from faecal images with up to 98.24% accuracy, Nakrosis et al. [11] sorted droppings into six classes with 91.78% accuracy, and Sadeghi et al. [12] detected avian influenza and Newcastle disease from thermal images within 24 hours of infection. Environmental data has mostly been used to predict the air rather than the birds: Rawdhan et al. [13] predicted carbon dioxide, methane and hydrogen sulphide in poultry houses from

temperature and humidity using tree-based and linear models. Random Forest [14] aggregates many decision trees to produce a more stable prediction than any single tree, which matters when training data is sparse or imbalanced. XGBoost [15] builds its trees sequentially instead, each one correcting the residual error of the last, and includes regularisation and class-weighting mechanisms that suit disease datasets where healthy observations vastly outnumber sick ones.

### C. Temporal modelling

Some risk signals in a poultry house arrive gradually. Ammonia accumulation, slow humidity drift and persistent temperature dysregulation unfold over hours, which makes detection partly a temporal problem. LSTM networks [16] are designed for exactly this, retaining information across long sequences without the vanishing gradient behaviour that limits standard recurrent architectures. Teixeira et al. [7] applied LSTM networks to wearable sensor data from dairy calves and predicted anaplasmosis up to three days before clinical diagnosis, with accuracy and sensitivity above 80%. In poultry, Manikandan and Neethirajan [8] review acoustic monitoring and describe a move from hand-crafted audio features to CNN and LSTM models.

### D. Explainable AI

High accuracy is not sufficient for adoption. A model that cannot say why it raised an alert will not change behaviour on a farm where the decision carries real financial consequences. SHAP [17] and LIME [18] address this from opposite ends. SHAP assigns each input feature a contribution to the prediction based on cooperative game theory, giving a dataset-wide picture of what the model is actually using. LIME approximates the decision boundary around a single case with a simpler interpretable model, so a specific alert can be interrogated on its own terms. Ryo [19] shows on agricultural data that these methods expose which variables and interactions drive a prediction, something accuracy figures alone cannot. In a poultry context, the difference is between telling a manager that risk is high and telling them that ammonia in house three has crossed the level that historically precedes respiratory illness.

### E. Intelligent ventilation

Airflow management in poultry houses has traditionally been manual and episodic. Automated systems now keep conditions inside set bands. Elwakeel [4] switches fans and evaporative cooling when readings cross thresholds and sends SMS alerts on equipment failure, Godinho et al. [20] propose a wireless sensor network with a central hub that applies rule-based control, and Lashari et al. [21] monitor temperature, humidity, oxygen, carbon dioxide, carbon monoxide and ammonia in broiler houses. Chen et al. [22] move toward predictive control by forecasting carbon dioxide in a layer house with GRU and LSTM networks, so that ventilation can act before concentrations rise. In each case the controlled quantity is the air itself; none links the control action to a predicted health state of the flock.

### F. Summary and research gap

Table 1 sets the reviewed studies against the four capabilities relevant here. Each study delivers one or two of them and none delivers all four. Predictive accuracy is well covered. Temporal modelling appears in a few studies. Explainability is largely absent from the poultry-specific work. Integration between prediction and ventilation actuation is the rarest of all, and where control exists it is threshold-based and not model-driven. This study addresses that combination.

**Table 1. Summary of related works and identified limitations**

| Author (Year) | Method and contribution | Reported outcome | Limitation |
|---|---|---|---|
| Shwetha et al. [23] (2024) | Review of ML and DL techniques in poultry health and disease management, 2010 to 2023 | Mapped applications in disease detection, health monitoring, behaviour, lameness and stress | Review only; no deployment and no link to a control system |

| | | | |
|---|---|---|---|
| Sadeghi et al. [12] (2023) | Thermal imaging with SVM and Dempster-Shafer fusion for avian influenza and Newcastle disease | 97.2% (avian influenza) and 100% (Newcastle) accuracy within 24 h of infection | Visual modality only; no environmental input, no ventilation response |
| Manikandan and Neethirajan [8] (2025) | Systematic review of AI analysis of poultry vocalisations | Reports disease and thermal-stress detection from sound alone | Inconsistent evaluation and few annotated datasets; no environmental variables or actuation |
| Nakrosis et al. [11] (2023) | Computer vision classification of dropping images into six classes | YOLOv5 reached 91.78% accuracy | Droppings only; no environmental input or actuation |
| Elwakeel [4] (2025) | Low-cost IoT monitoring and control unit with SMS early warning | Sensors agreed with reference devices ($r > 0.96$) at about 12% of commercial cost | Threshold control only; no predictive model; laboratory tested |
| Chen et al. [22] (2024) | GRU and LSTM prediction of CO2 for ventilation control | Accurate CO2 forecasting enabling better air quality control | Predicts a gas concentration, not bird health status |
| Godinho et al. [20] (2025) | Conceptual wireless sensor network with central control hub | Rule-based control of temperature, humidity, CO2 and ammonia | Conceptual; no ML and no link to disease state |
| Lashari et al. [21] (2023) | IoT monitoring of temperature, humidity, O2, CO2, CO and ammonia in broiler houses | Real-time air quality monitoring | Monitoring only; no predictive model |
| Machuve et al. [10] (2022) | Deep learning on faecal images for coccidiosis, salmonella and Newcastle disease | Up to 98.24% accuracy (Xception) | Image modality only; no environmental input or actuation |
| Rawdhan et al. [13] (2026) | Tree-based and linear models predicting poultry-house gases from temperature and humidity | Predicted CO2, methane and H2S from climate readings | Predicts air quality, not bird health; no actuation |
| Teixeira et al. [7] (2022) | LSTM on wearable rumination and activity data in dairy calves | Predicted anaplasmosis up to 3 days before diagnosis (>80% accuracy) | Different species and disease; no environmental variables or control output |

# MATERIALS AND METHODS

## A. System architecture

The framework is organised as four layers, shown in Figure 1. Raw environmental readings enter at the bottom, pass through three parallel classifiers, are combined by a fusion mechanism, and leave as a disease risk status paired with a ventilation instruction. Each layer is defined independently, so a component can be replaced without rebuilding the rest of the system.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

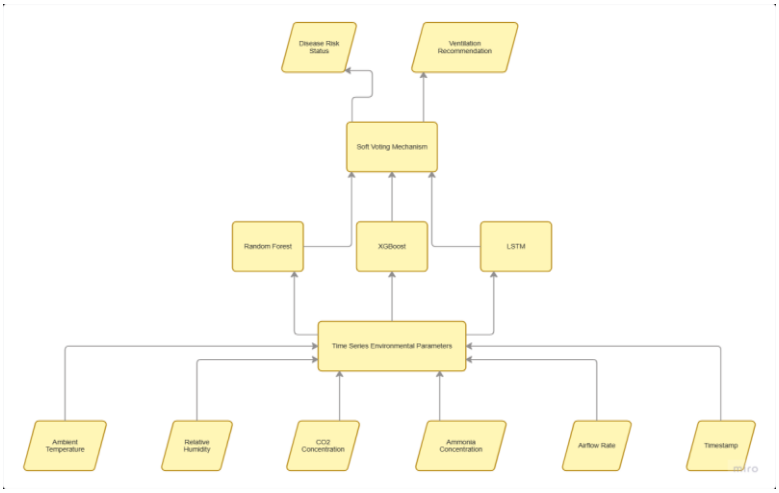



**Figure 1. System architecture of the hybrid prediction and ventilation framework**

The input layer takes hourly time-series records with six fields: ambient temperature in degrees Celsius, relative humidity as a percentage, carbon dioxide concentration in ppm, ammonia concentration in ppm, airflow rate in metres per second, and a timestamp. These five measured variables were chosen because the literature reviewed in Section II links each of them directly to respiratory outcomes.

The processing layer holds the three models. Random Forest and XGBoost operate on the flattened feature matrix and handle the threshold effects and feature interactions in the tabular data. The LSTM receives the same data restructured into overlapping sequence windows, where it can pick up how conditions evolve rather than where they currently sit.

The fusion layer aggregates the three outputs by soft voting [24]. Unlike hard voting, where each model casts one vote and the majority decides, soft voting averages the predicted probability distributions across all four classes and selects the class with the highest mean probability. This preserves the confidence information in each model's output instead of discarding it, which matters when two classes are close.

The output layer converts the winning class into a disease risk status and maps that status onto a ventilation action. Figure 2 shows the same flow from the operator's point of view, from sensor reading through risk category and explanation to a recommended adjustment.

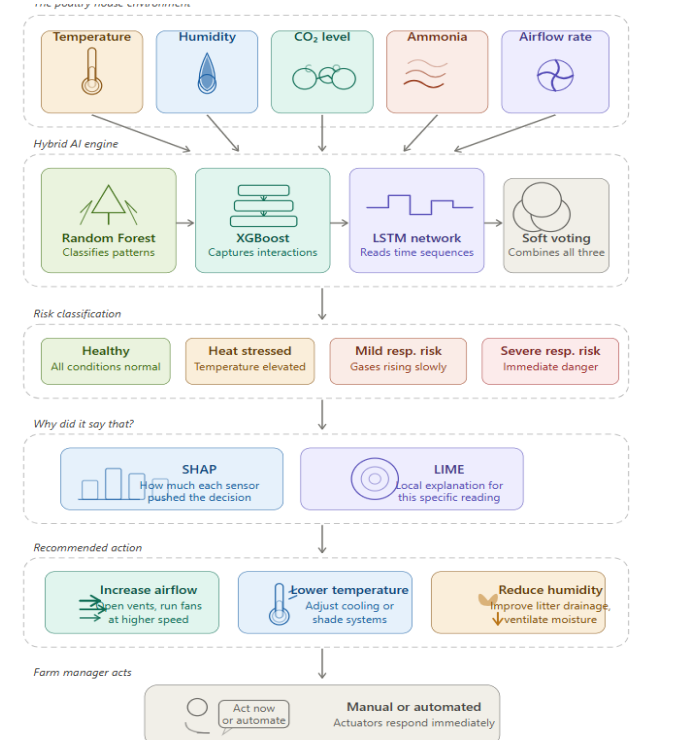



**Figure 2. Conceptual framework showing the path from sensor reading to recommended action**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

### B. Data generation and preprocessing

No public dataset pairs continuous poultry house environmental readings with labelled respiratory disease states at the resolution this study required, so a dataset of 10,000 hourly records was generated programmatically. This is a limitation. It was also a deliberate design choice: generating the data removed hardware variability as a confounding factor and made the experiments fully reproducible.

The dataset is 10,000 consecutive hourly timestamps beginning 1 January 2024, which spans 416.6 days. An earlier version of this work described the period as 30 days; that was an arithmetic error and the correct span is stated here. The records represent a single simulated house. There is no production-cycle structure, no placement or depopulation event and no bird-age variable, which Section V identifies as a substantive restriction.

Generation proceeds in two stages per record. Base conditions are drawn from independent normal distributions: temperature N(25, 3) degrees Celsius, relative humidity N(60, 10) per cent, carbon dioxide N(1000, 300) ppm, ammonia N(15, 5) ppm and airflow N(1.5, 0.5) metres per second. A risk state is then assigned by a uniform draw with probabilities 0.60 healthy, 0.15 mild respiratory, 0.10 severe respiratory and 0.15 heat stress, and each variable receives a class-specific normal offset before being clipped to a class-specific interval. Table 2 gives every offset and interval, so the generator can be reproduced exactly. The clipping intervals are what define the labels, and because they are drawn from the same literature that motivates the study, the class structure is threshold-derived by construction. Table 3 shows the first five generated records and Table 4 the resulting class distribution.

**Table 2. Complete generation parameters. Each cell gives the class-specific offset as mean (standard deviation), followed by the clipping interval applied after the offset.**

| Variable | Healthy | Mild respiratory | Severe respiratory | Heat stress |
|---|---|---|---|---|
| Temperature (°C) | 0 (1), clip 20–28 | 0 (2), clip 22–30 | +2 (2), clip 24–32 | +8 (2), clip 30–35 |
| Humidity (%) | 0 (3), clip 45–70 | +5 (5), clip 60–75 | +10 (5), clip 70–80 | 0 (5), clip 50–70 |
| CO2 (ppm) | 0 (100), clip 400–1200 | +300 (150), clip 1200–2000 | +800 (200), clip 2000–3000 | 0 (150), clip 800–1500 |
| Ammonia (ppm) | -5 (2), clip 5–15 | +5 (3), clip 15–25 | +10 (4), clip 25–35 | 0 (3), clip 10–20 |
| Airflow (m/s) | +0.2 (0.1), clip 1.0–2.5 | -0.3 (0.15), clip 0.5–1.5 | -0.5 (0.2), clip 0.5–1.2 | -0.4 (0.2), clip 0.5–1.3 |
| Class probability | 0.6 | 0.15 | 0.1 | 0.15 |

**Table 3. Sample of the generated dataset (first five records)**

| Temperature (°C) | Humidity (%) | CO2 (ppm) | Ammonia (ppm) | Airflow (m/s) |
|---|---|---|---|---|
| 26.26 | 59.45 | 1200 | 15 | 1.53 |
| 21.76 | 48.04 | 400 | 15 | 2.09 |
| 25.58 | 45 | 931.43 | 11.14 | 1.06 |
| 22.71 | 64.43 | 1200 | 15.99 | 1.11 |
| 27.78 | 70 | 2000 | 25 | 0.5 |

**Table 4. Environmental risk state distribution**

| Disease status | Count | Percentage |
|---|---|---|
| Healthy | 5,963 | 59.63% |
| Mild respiratory | 1,535 | 15.35% |

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

| | | |
|---|---|---|
| Heat stress | 1,480 | 14.80% |
| Severe respiratory | 1,022 | 10.22% |
| Total | 10,000 | 100% |

The distribution was weighted toward the healthy class on purpose, because that is what a real flock looks like. Most hours in a managed house are uneventful. A balanced dataset would have been easier to model but less realistic.

Preprocessing consisted of label encoding of the target, standardisation of the five numeric features with StandardScaler, and a stratified 80/20 train-test split with a fixed random seed of 42. Stratification kept the class proportions identical across both partitions. For the LSTM, the scaled data was reshaped into overlapping 12-hour windows, producing 7,988 training sequences and 1,988 test sequences. No features were removed; all five carried signal.

**C. Model configuration, explainability and evaluation**

Random Forest used 100 estimators and a maximum depth of 20. XGBoost used 100 estimators, a maximum depth of 6, a learning rate of 0.1 and multiclass log loss as its evaluation metric. The LSTM was a two-layer stack, 64 units returning sequences into 32 units, each followed by 0.3 dropout, then a 32-unit dense layer with ReLU activation and a softmax output over the four classes. It held 31,524 trainable parameters and was trained with the Adam optimiser and categorical cross-entropy on a 20% validation split, with early stopping on validation loss at a patience of 10 epochs and best-weight restoration. All three models saw the same split, and random seeds were fixed at 42 for NumPy and TensorFlow. Implementation was in Python on Google Colab, using scikit-learn, XGBoost, TensorFlow and Keras, with SHAP and LIME for explanation and Matplotlib and Seaborn for plotting.

SHAP TreeExplainer was applied to both tree models over a 100-sample slice of the scaled test set to produce global feature attributions. LIME's tabular explainer was fitted on the scaled training set and used to explain individual test instances against the Random Forest's predicted probabilities, with five features reported per explanation. The two methods answer different questions. SHAP says what the model relies on across the dataset; LIME says what drove one particular alert.

The decision layer takes the predicted class and the associated confidence and returns a set of adjustments to airflow, temperature and humidity, with a confidence floor for each class. Floors are lower for the more serious classes, on the reasoning that a false alarm on severe respiratory risk costs a few minutes of extra fan runtime while a missed case costs birds. Table 9 gives the full mapping. The magnitudes in that table were chosen to illustrate the form of a confidence-gated mapping. They were not derived from ventilation engineering and they have not been validated against welfare guidance, and they should be read as placeholders for values that a deployment would have to calibrate.

Three constraints govern any safe application of such a mapping, and the present implementation represents none of them. A minimum ventilation rate per unit of liveweight must be maintained whatever the predicted class, so that the airflow reduction implied by a healthy classification cannot starve the house of fresh air. An upper limit on air speed at bird level applies irrespective of gas concentrations. And during brooding the two objectives conflict directly, because clearing ammonia requires air exchange while young birds require a house temperature some 10 to 15 degrees Celsius above the finishing target; nothing in this framework resolves that conflict, since the dataset carries no bird-age variable.

Each model was assessed on accuracy, per-class precision, recall and F1-score, a confusion matrix, and macro-averaged one-versus-rest ROC-AUC. Macro-averaged and weighted-averaged values are both reported, because with a 59.63% majority class the weighted average alone would flatter a model that had learned nothing but the majority. Two conventional baselines [25] were then fitted on the same split for reference: multinomial logistic regression, and a single decision tree with the same maximum depth of 20 used inside the forest. The soft-voted ensemble was evaluated in two configurations, one averaging the two tree models and one averaging all three, which isolates what each component contributes. Because the LSTM

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

operates on 12-hour windows, its test partition holds 1,988 sequences against the 2,000 records available to the tabular models; the three-model fusion is therefore reported on that aligned subset, alongside the two-model fusion recomputed on the same subset for a fair comparison. The ventilation logic was assessed separately, by checking whether predicted risk levels mapped consistently onto actions across the test set and whether the simulated environment responded as intended over a 12-hour window of rising ammonia.

## RESULTS AND DISCUSSION

### A. Random Forest

Random Forest classified 1,995 of 2,000 test records correctly, an accuracy of 99.75%. Per-class precision, recall and F1 rounded to 1.00 for healthy and severe respiratory, with heat stress at 0.99 precision and mild respiratory at 0.99 F1. The confusion matrix in Figure 3 shows where the five errors sat: one heat stress record classified as mild respiratory, and four mild respiratory records split between healthy and heat stress. Every severe respiratory case was caught. Operationally, this is the preferred error profile, since the class that matters most for bird welfare had no misses.

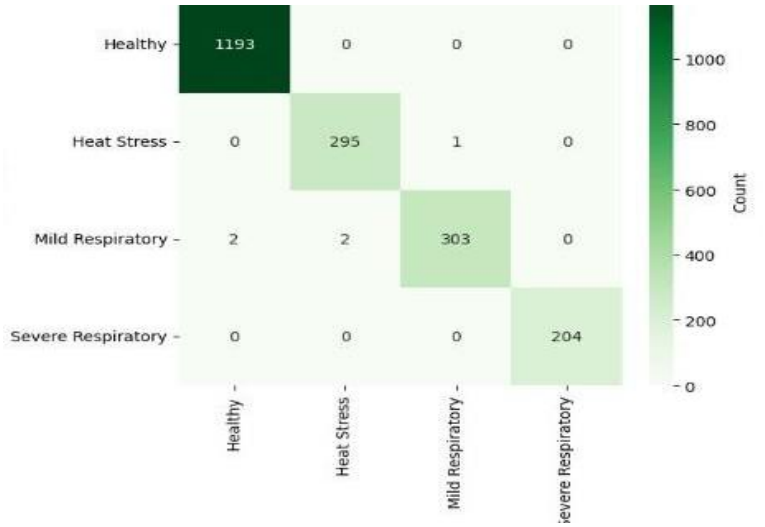



**Figure 3. Random Forest confusion matrix**

### B. XGBoost

XGBoost classified 1,994 of 2,000 correctly, an accuracy of 99.70%, statistically indistinguishable from Random Forest on a test set of this size. The confusion matrix in Figure 4 shows six errors: five misclassified mild respiratory records, and one heat stress record assigned to mild respiratory. Mild respiratory recall was 0.98 against Random Forest's 0.99. Training took roughly three seconds against two for Random Forest.

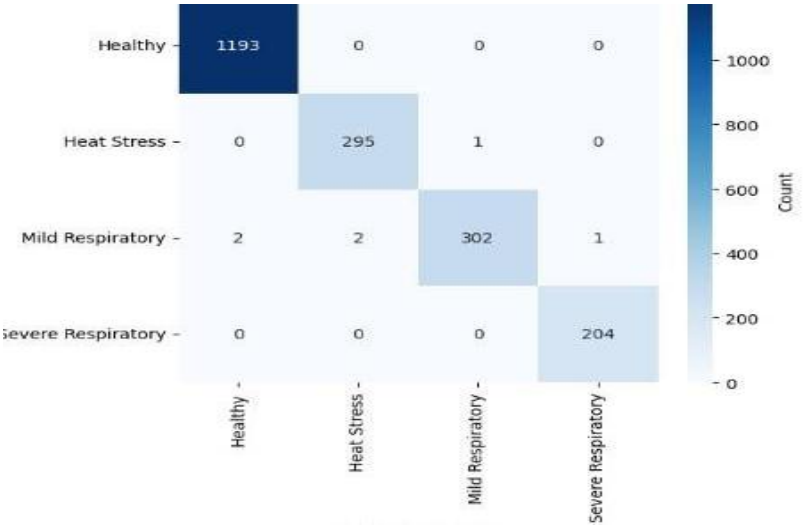



**Figure 4. XGBoost confusion matrix**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

The two tree models also agree on which cases are hard. Both stumble on the mild respiratory class, and both confuse it with adjacent classes instead of with severe respiratory. That is consistent with feature overlap at the boundary between mild elevation and normal variation, not with either model failing.

## C. LSTM

The LSTM reached 59.66% accuracy, which is almost exactly the proportion of healthy records in the test set. Figure 5 explains why: the model assigned every one of the 1,988 test sequences to the healthy class. Precision, recall and F1 for heat stress, mild respiratory and severe respiratory were all zero, macro F1 was 0.19, and macro ROC-AUC was 0.5032, which is chance. That last figure matters, because a model can post a poor accuracy while still ranking cases usefully. This one does not. Figure 6 shows the training history, where validation loss flattens after epoch two and validation accuracy sits pinned at 0.592 from the first epoch.

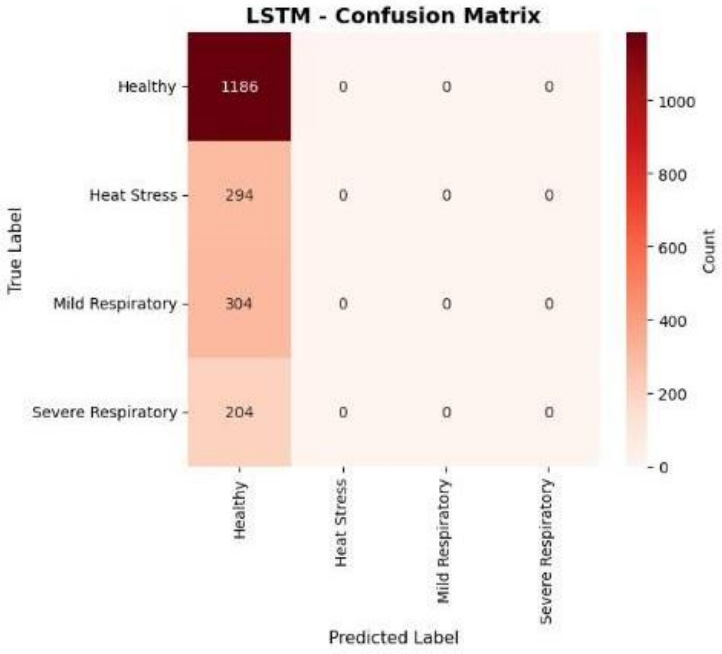



**Figure 5. LSTM confusion matrix showing complete collapse onto the majority class**

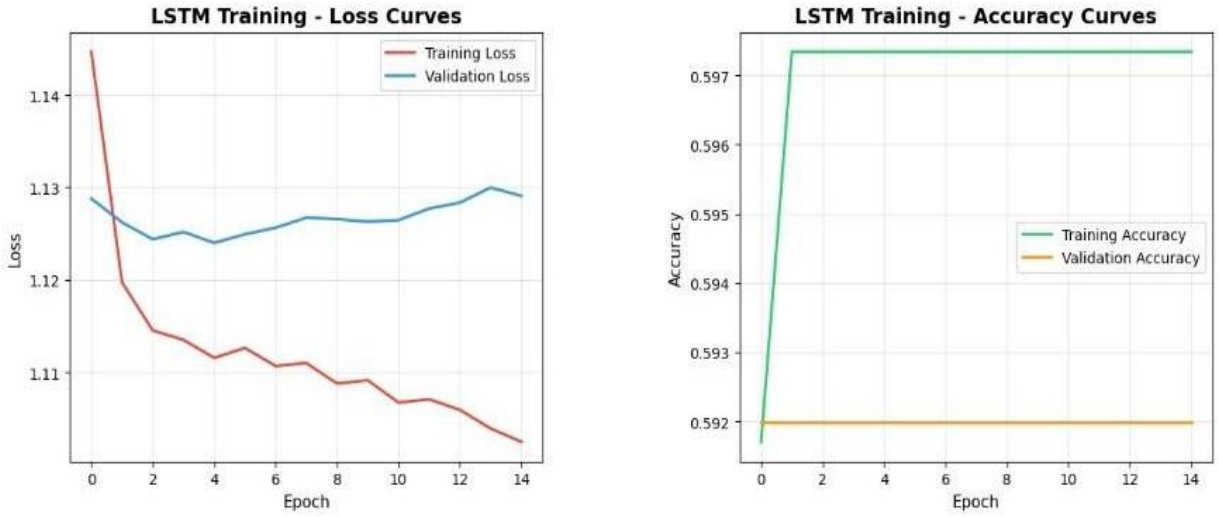



**Figure 6. LSTM training and validation loss and accuracy curves**

Before the causes are discussed, one aspect of the implementation must be disclosed, because it determines how the result should be read. The stratified train-test split was applied to the records first, and the 12-hour windows were constructed from the split partitions afterwards. Because the split shuffles the records, each window therefore contained twelve records drawn from unrelated hours in arbitrary order. The network was

not presented with a time series. Its collapse onto the majority class follows from the data preparation and cannot be attributed to the architecture, the window length or the training schedule. We report the result rather than withdraw it because the underlying dataset is in any case memoryless: records are drawn independently, so no ordering of them would have carried the temporal dependency an LSTM is built to exploit. Both facts point the same way, and neither supports presenting the LSTM as a test of early prediction.

With that stated, four factors bear on the outcome.

- The labels do not depend on sequence history. A record is labelled by the environmental values at that hour, so the temporal context an LSTM is built to exploit carries little additional information. This is a feature-based classification problem wearing the clothes of a sequence problem.
- 7,988 training sequences is thin for a 31,524-parameter recurrent network. Goodfellow et al. [26] give a rough rule of thumb of around 5,000 labelled examples per class for acceptable supervised deep learning performance; the three minority classes here have roughly 800 to 1,250 training sequences each.
- No class weighting or resampling was applied, so predicting healthy for everything was a locally optimal strategy for the loss function. The training curves show the network finding it in the first epoch and staying there.
- The windows were built over shuffled records, as disclosed above, so the sequence axis carried no chronological information. A chronological split with ordered, non-overlapping windows is a prerequisite for any future test of temporal structure.
- A 12-hour window may be the wrong horizon. If the discriminative change happens over two or three hours, a 12-hour window dilutes it.

The third of these was tested directly. The network was refitted with balanced class weights (0.419 for healthy, 1.687 for heat stress, 1.630 for mild respiratory and 2.447 for severe respiratory) and everything else held constant. Weighting broke the collapse, in that the model stopped predicting a single class, but it did not produce a usable classifier: accuracy fell to 25.15% and macro F1 moved only from 0.1868 to 0.1874. Predictions scattered across healthy, heat stress and severe respiratory with mild respiratory almost never chosen. So class imbalance explains the shape of the failure without explaining the failure itself. There is little discriminative information in the sequence windows for the network to find, and removing the majority-class shortcut simply left it guessing. That supports the first explanation above over the third, and it is a more useful finding than the collapse on its own.

### D. Comparative performance

**Table 5. Per-class performance of the three models on the test set**

| Class | Metric | Random Forest | XGBoost | LSTM |
|---|---|---|---|---|
| Healthy | Precision | 1 | 1 | 0.6 |
| Healthy | Recall | 1 | 1 | 1 |
| Healthy | F1-score | 1 | 1 | 0.75 |
| Heat stress | Precision | 0.99 | 0.99 | 0 |
| Heat stress | Recall | 1 | 1 | 0 |
| Heat stress | F1-score | 0.99 | 0.99 | 0 |
| Mild respiratory | Precision | 1 | 1 | 0 |
| Mild respiratory | Recall | 0.99 | 0.98 | 0 |
| Mild respiratory | F1-score | 0.99 | 0.99 | 0 |
| Severe | Precision | 1 | 1 | 0 |

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

| | | | | |
|---|---|---|---|---|
| respiratory | | | | |
| Severe respiratory | Recall | 1 | 1 | 0 |
| Severe respiratory | F1-score | 1 | 1 | 0 |

**Table 6. Overall model comparison**

| Model | Accuracy | Macro precision | Macro recall | Macro F1 | Weighted F1 | Training time |
|---|---|---|---|---|---|---|
| Random Forest | 99.75% | 1 | 1 | 1 | 1 | ~2 s |
| XGBoost | 99.70% | 1 | 1 | 1 | 1 | ~3 s |
| LSTM | 59.66% | 0.15 | 0.25 | 0.19 | 0.45 | ~30 s |

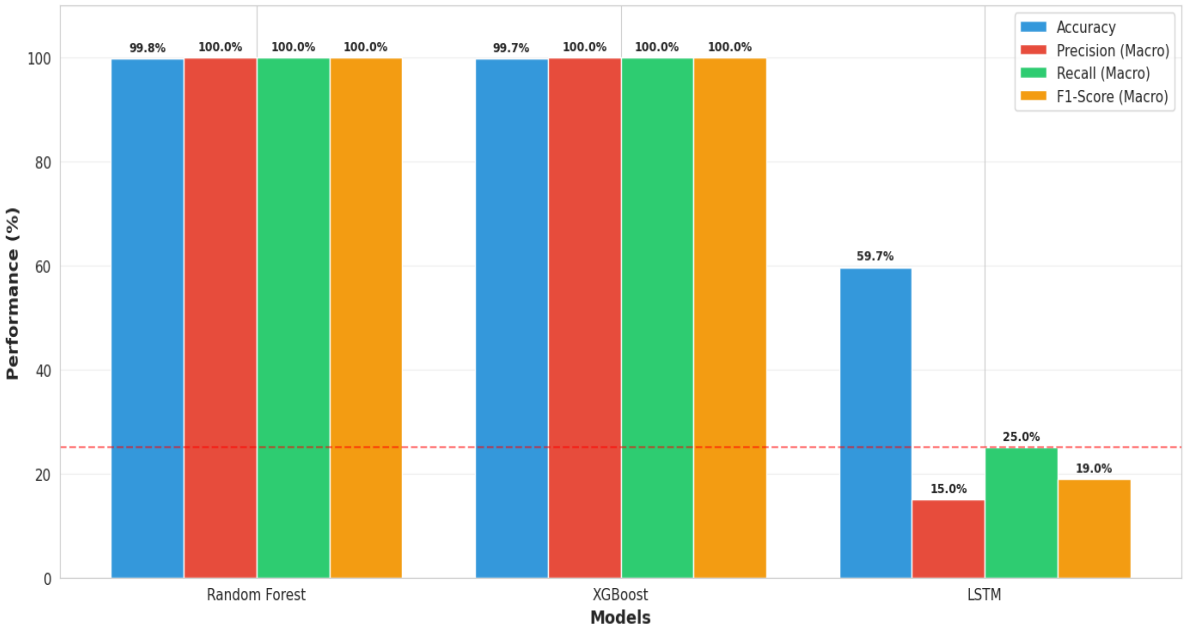



**Figure 7. Accuracy, macro precision, macro recall and macro F1 across the three models**

The gap between the weighted F1 of 0.45 and the macro F1 of 0.19 for the LSTM is the clearest single indicator in Table 6. A model that has learned only the majority class can still post a respectable weighted score. Reporting both averages is what makes the failure visible, and it is the reason multi-class agricultural work should not report accuracy alone.

### E. Fusion and baseline comparison

The point of a hybrid architecture is that the combination should beat its parts. That claim is asserted more often than it is tested, so it was tested here. Table 7 reports the soft-voted ensemble in both configurations alongside the component models and two conventional baselines fitted on the same split.

**Table 7. Fusion and baseline comparison on the same train-test split**

| Model | Accuracy | Macro F1 | ROC-AUC | Test set |
|---|---|---|---|---|
| Logistic regression | 98.05% | 0.9728 | 0.999 | 2,000 |
| Decision tree (depth 20) | 99.60% | 0.9941 | 0.9959 | 2,000 |
| LSTM | 59.66% | 0.1868 | 0.5032 | 1,988 |

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

| | | | | |
|---|---|---|---|---|
| Random Forest | 99.75% | 0.9965 | 1 | 2,000 |
| XGBoost | 99.70% | 0.9963 | 1 | 2,000 |
| Soft voting (RF + XGBoost) | 99.75% | 0.996 | 1 | 2,000 |
| Soft voting (RF + XGBoost), aligned subset | 99.75% | 0.9959 | 1 | 1,988 |
| Soft voting (RF + XGBoost + LSTM) | 99.70% | 0.9954 | 1 | 1,988 |

Three results in Table 7 deserve attention, and none of them flatters the hybrid framing this study started from.

First, fusion bought nothing. Soft voting across the two tree models matched Random Forest on accuracy at 99.75% and came in marginally below it on macro F1, 0.9960 against 0.9965. This is the expected outcome when component models are already close to the ceiling and their errors are correlated, which Figures 3 and 4 showed they are. Both trees fail on the same mild respiratory boundary cases, so averaging their probabilities cannot recover what neither model got right.

Second, adding the LSTM made the ensemble worse. On the 1,988-sequence subset where all three models can be compared directly, two-model fusion scores 99.75% and three-model fusion 99.70%, with macro F1 falling from 0.9959 to 0.9954. The loss is small because the tree models are confident enough to outvote a near-uniform third opinion, but the direction is unambiguous. In a soft vote, a component that carries no signal still pulls the averaged probabilities away from the correct class. Equal weighting is what allows that, so weighted voting or stacking is worth testing. It is also why the ventilation controller in Section IV.I was driven by Random Forest alone.

Third, and most awkward for the premise of the study, a single decision tree reached 99.60% and multinomial logistic regression 98.05%. The ensemble's advantage over a plain decision tree is 0.15 percentage points, three test records out of 2,000. Against logistic regression it is under two points. On data of this separability, the sophisticated model is not earning much, and reporting only the ensemble result would have hidden that. An ensemble accuracy in the high nineties says little about the ensemble unless simple baselines are fitted on the same data.

Random Forest, XGBoost and their fusion all reach a ROC-AUC of 1.0000 while the decision tree returns 0.9959 despite a competitive accuracy. ROC-AUC measures how well a model ranks cases, not whether its predicted probabilities are calibrated, so this difference does not establish that the ensemble produces better confidence estimates and no such claim is made here. Since the ventilation layer in Section III.C gates its actions on confidence thresholds, the calibration of the predicted probabilities is a functional requirement of the design, and establishing it would require reliability diagrams, Brier scores or an equivalent calibration measure. That assessment was not performed and is listed in Section V as work required before any confidence-gated controller is deployed. On the evidence actually available, Random Forest is the preferred classifier and the ensemble has no demonstrated advantage over it.

The pattern across Table 7 has one underlying cause, and it needs to be read with the data generation process in mind. The class boundaries in the generated dataset are driven by parameter ranges that, although overlapping and noise-perturbed, are ultimately separable by threshold rules of the kind a decision tree learns naturally. Accuracy at this level should not be read as an estimate of field performance. It says the models learned the generating process, which is the most a simulation study can establish. Whether real houses are this separable is an empirical question that needs real sensor data to answer.

### F. Behaviour under changed generating assumptions

The results above are obtained by splitting a single generated dataset, so they measure how well a model recovers one generating process rather than how it behaves when the process changes. To probe that distinction a Random Forest trained on the dataset described in Section III.B was applied, without retraining, to further datasets generated under altered assumptions. Table 8 reports the outcome.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

**Table 8. Accuracy of a Random Forest trained on the default generator and applied without retraining to data generated under altered assumptions. The majority-class baseline is 59.6%.**

| Generating condition | Accuracy (%) |
|---|---|
| In-distribution test set (as reported above) | 99.75 |
| Base temperature raised 3 °C | 99.82 |
| Base temperature lowered 3 °C | 99.84 |
| Base ammonia raised 8 ppm | 99.82 |
| Base carbon dioxide raised 400 ppm | 99.78 |
| Base airflow reduced 0.4 m/s | 99.78 |
| Multiplicative sensor error of 15% on all readings | 68.83 |
| Gas concentrations coupled to ventilation rate | 76.01 |

Shifting the centre of any base distribution leaves accuracy essentially unchanged, because the clipping intervals that define the labels are untouched and the classifier has learned those intervals. Two perturbations break it. Adding 15 per cent multiplicative sensor error reduces accuracy to 68.83 per cent, and making the gas concentrations depend on the ventilation rate reduces it to 76.01 per cent; both are within roughly ten points of the 59.6 per cent majority-class baseline. The practical reading is that accuracy near 100 per cent on this dataset is compatible with a loss of 24 to 31 points under modest and physically reasonable changes to the generating assumptions. It is therefore a statement about the recovery of a known process, not an estimate of field performance, and the two should not be conflated when results of this kind are reported.

## G. Explainability results

SHAP analysis of the Random Forest ranked carbon dioxide highest, ammonia second, then temperature, humidity and airflow, as shown in Figure 8. The XGBoost attributions in Figure 9 place temperature marginally ahead of ammonia and carbon dioxide overall, with the three separated by a small margin, and the per-class breakdown shows carbon dioxide and ammonia dominating the classes that matter here while temperature carries most of its weight in the heat stress class. The two models therefore agree on the substance, that gas concentrations drive the respiratory classes, while ordering the top three slightly differently. Running SHAP on both models is what makes this difference visible.

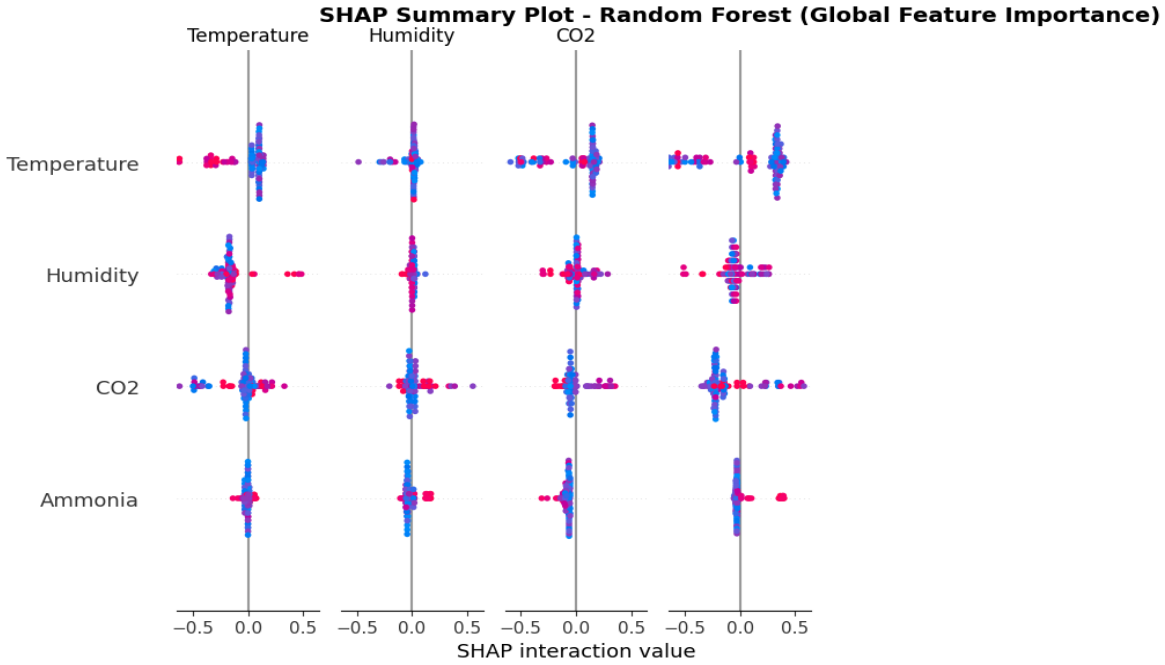



**Figure 8. SHAP summary plot for Random Forest**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

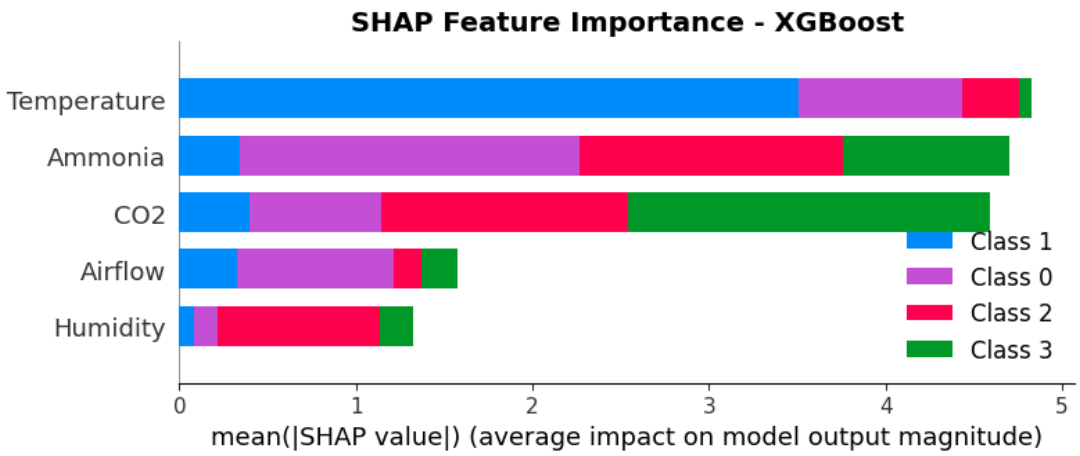



**Figure 9. SHAP feature importance for XGBoost, broken down by class**

LIME was applied to individual predictions. Figure 10 shows the explanation for test instance 42, correctly classified as healthy. Each bar is the signed change in the predicted probability of the predicted class attributable to that feature falling in its observed quartile bin; the contributions are signed and do not sum to one. For this instance, low ammonia and low carbon dioxide are what the model treated as evidence of a healthy state. A companion chart of normalised absolute contributions appeared in an earlier version of this work and has been withdrawn, because its values could not be reconciled with the explainer output they were said to represent.

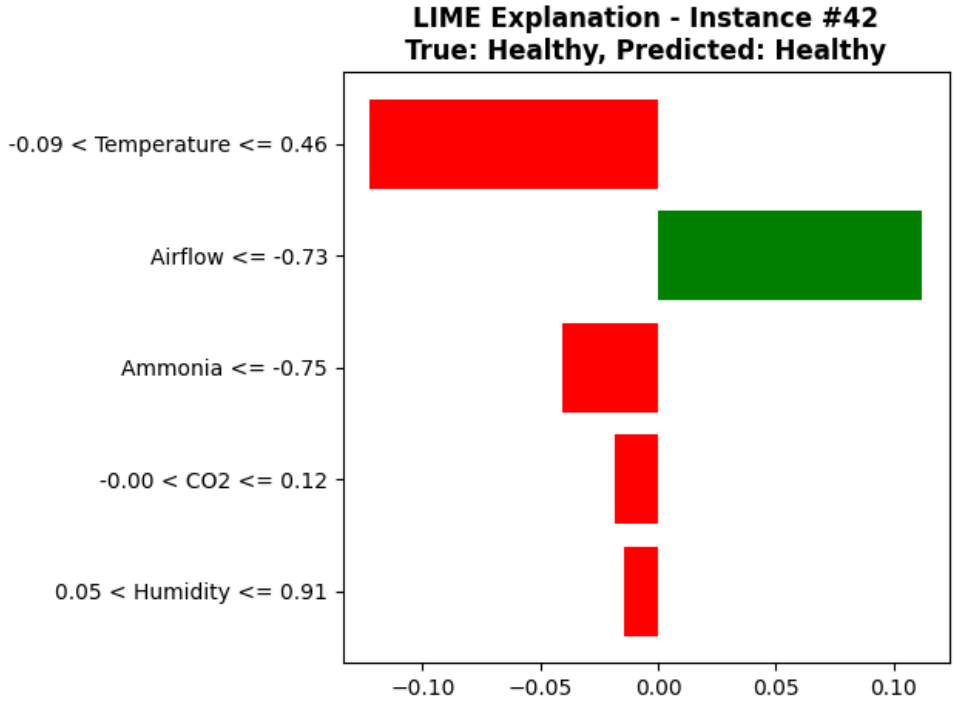



**Figure 10. LIME explanation for test instance 42**

For a farm operator, an alert that says “severe respiratory risk” invites the question of whether to believe it. An alert that says the classification was driven mostly by ammonia at a stated concentration can be checked against the litter condition in the house, and either confirmed or dismissed on the spot.

### H. Correspondence between the attributions and the generator

The attribution rankings correspond to the variables that veterinary evidence identifies. Carbon dioxide above roughly 1,500 ppm indicates that the house is not exchanging air. Ammonia irritates the respiratory tract and raises susceptibility to disease at or below 25 ppm [1], and depresses growth at 50 ppm and above [2].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

Temperature enters mainly through heat stress, since birds depend heavily on panting [3], and elevated temperature combined with high humidity removes that route.

This correspondence should be interpreted carefully, and it is weaker evidence than it may appear. The clipping intervals that define the risk states in Table 2 were themselves parameterised from that same literature. The models were therefore shown data in which carbon dioxide and ammonia determine the label by construction, and recovering that dependence confirms that the pipeline implements the generator faithfully. It does not constitute biological validation and it discovers nothing about poultry physiology. A claim that these attributions reflect clinically meaningful relationships would require the models to be fitted to measured environmental data paired with independent veterinary assessment.

**I. Ventilation control simulation**

The trained Random Forest was embedded in the control function described in Section III.C. Table 9 gives the mapping from predicted class and confidence to control action.

**Table 9. Prediction-driven environmental control strategy**

| Predicted class | Confidence floor | Control action |
|---|---|---|
| Healthy | > 90% | Maintain current conditions |
| Heat stress | > 80% | Airflow +1.0 m/s (maximum), temperature -5 °C, activate cooling |
| Mild respiratory | > 75% | Airflow +0.3 m/s (+20%), temperature -1 °C, humidity -5% |
| Severe respiratory | > 70% | Airflow +0.8 m/s (+50%), temperature -2 °C, humidity -10%, alert operator |

Table 10 shows the controller running on five randomly drawn test samples. Predictions carried full confidence in each case, and the actions taken follow the mapping without exception.

**Table 10. Sample predictions and resulting control actions**

| Temp (°C) | Humidity (%) | CO2 (ppm) | Predicted class | Confidence | Action | Adjustments |
|---|---|---|---|---|---|---|
| 28 | 70 | 1014 | Healthy | 100% | Maintain | None |
| 23.3 | 73.7 | 1686 | Mild respiratory | 100% | Increase ventilation | Airflow +0.3, temp -1 °C, humidity -5% |
| 22 | 65 | 1861 | Mild respiratory | 100% | Increase ventilation | Airflow +0.3, temp -1 °C, humidity -5% |
| 24.7 | 53.1 | 1098 | Healthy | 100% | Maintain | None |
| 21.4 | 47 | 650 | Healthy | 100% | Maintain | None |

Figure 11 traces the simulated environmental response over the hour following an intervention. Carbon dioxide falls from around 2,500 ppm and crosses the 1,500 ppm threshold at roughly the twenty minute mark. Temperature converges on the 27 degree target over about forty minutes, and humidity approaches 70% over a similar period. These trajectories come from an exponential decay model, not from a mass and heat balance of a building, so they describe the shape the decay model was given rather than how a house responds. The model carries no gas production term, no heat balance and no energy cost, and it has not been compared against conventional threshold-based ventilation. Establishing whether the controller offers any advantage over a temperature-led threshold controller requires at minimum a physically grounded house model, and is listed in Section V.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

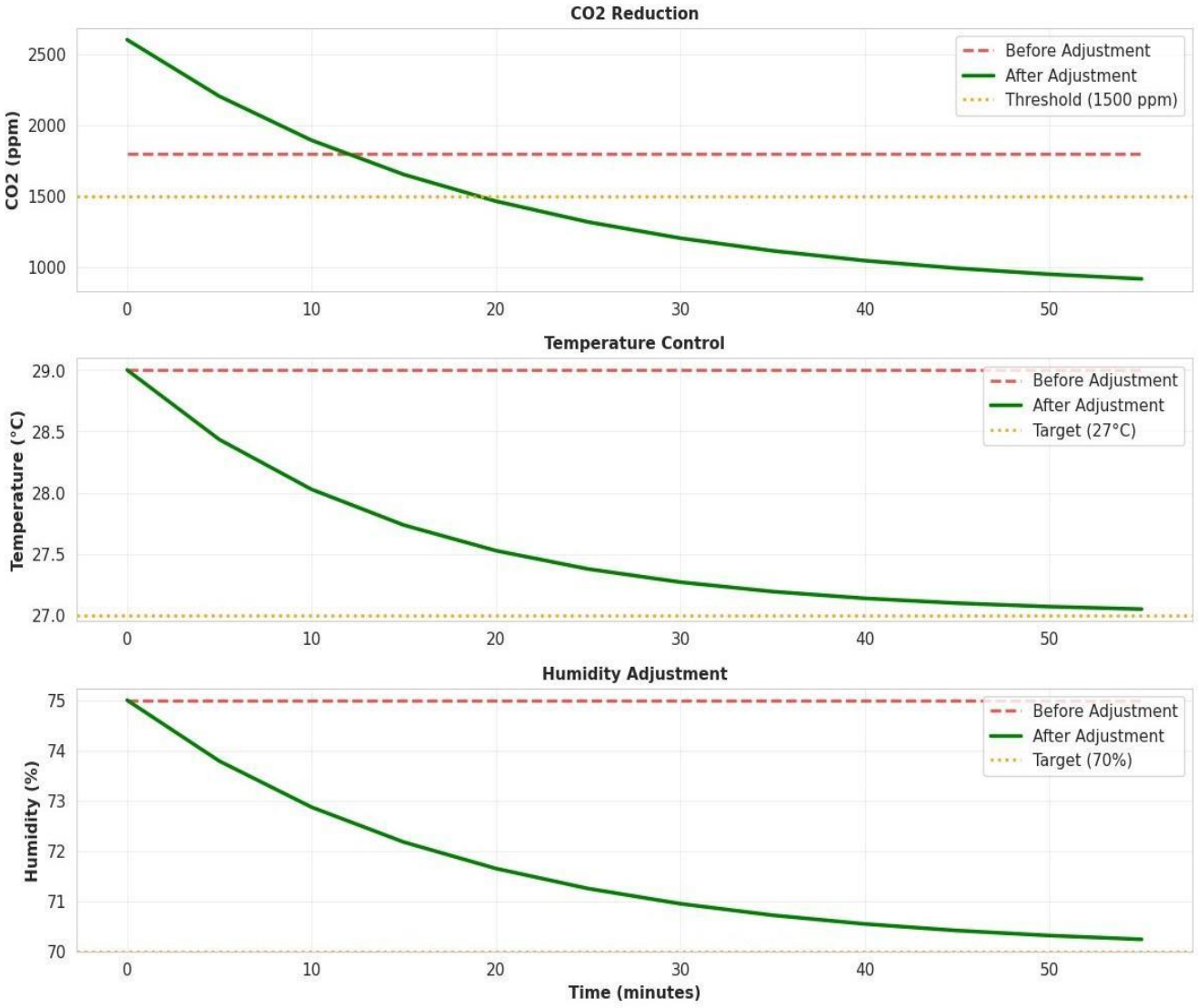



**Figure 11. Simulated environmental response following a ventilation adjustment**

Inference latency was under 50 ms per sample. This bears on computational feasibility only; it says nothing about whether a farm can act on a prediction, which depends on the actuators installed and on the labour available, neither of which this study examined. Whether any economic case follows is likewise untested. Continuous automated monitoring may displace some of the labour currently spent on manual inspection and may reduce losses, but both are hypotheses requiring a costed field trial, and nothing in these results establishes either.

## CONCLUSION AND RECOMMENDATIONS

This study tested whether machine-learning classifiers can recover predefined environmental respiratory risk states from poultry housing variables under simulated conditions, and whether a recovered state can be mapped onto a ventilation instruction. The findings below are separated into what the experiments demonstrate, what the simulation assumes, and what is proposed but not yet tested.

### A. What the results demonstrate

On this dataset the tree-based classifiers recover the generating process almost completely and the recurrent model does not. Random Forest reached 99.75 per cent accuracy and XGBoost 99.70 per cent, neither missing a severe respiratory case, while the LSTM returned the majority-class rate of 59.66 per cent. Soft voting matched its best component without improving on it, adding the LSTM reduced accuracy by 0.05 percentage points, and a single decision tree on the same split reached 99.60 per cent with multinomial logistic regression at 98.05 per cent. Random Forest is therefore the preferred classifier under the reported conditions, and the

hybrid architecture that motivated the work has no demonstrated advantage over it. Stated at the level the evidence supports, the defensible finding is that when risk categories are generated from structured threshold assumptions, simple tree classifiers match substantially more complex ensembles, and a recurrent architecture adds nothing when the generating process carries no temporal dependence.

### B. What the simulation assumes

Four assumptions bound every number above. The risk labels are defined by thresholds on the environmental variables rather than by veterinary assessment, so the classification task is the recovery of those thresholds and not the detection of clinical disease. The records are drawn independently, so the dataset contains no trend, no diurnal cycle and no exposure history, which is why the LSTM result should be read as a property of the data rather than a finding about recurrent models. The dataset represents one house with no production cycle and no bird-age variable, although thermal comfort targets in a real house move by 10 to 15 degrees Celsius across a cycle. And accuracy measured by splitting a single generated dataset is not a measure of field performance: Table 8 shows the same model falling to between 68.8 and 76.0 per cent under modest changes to the generating assumptions, against a majority-class baseline of 59.6 per cent.

### C. What is proposed but not tested

Every operational claim associated with systems of this kind remains a hypothesis here. Improved air quality is expected to reduce environmental stress and the susceptibility of the respiratory epithelium, but this study measured no bird, so no effect on mortality, growth or disease incidence is established. The ventilation mapping has not been compared against conventional threshold-based control, its adjustment magnitudes are illustrative, and the house response is a decay curve rather than a mass and heat balance. Nor does better ventilation substitute for veterinary practice: it does not prevent pathogen introduction, and it stands alongside diagnosis, vaccination, infection control, biosecurity and treatment rather than in place of any of them. An airborne pathogen introduced from a neighbouring premises is the clearest case the framework cannot address, since inter-farm distance and biosecurity practice determine that outcome and no interior reading anticipates the event.

### D. Recommendations

The priority is measured data. Environmental sensor records should be collected from operational houses and paired, where feasible, with veterinary assessment, documented respiratory signs, laboratory confirmation and flock outcomes, since only such data can establish whether these models identify clinically meaningful risk or provide useful warning time. Alongside that, four methodological steps follow directly from the results reported here. Calibration should be assessed explicitly through reliability diagrams or Brier scores before any confidence-gated controller is deployed, since the present evidence does not establish it. Performance should be reported over repeated independently generated datasets with confidence intervals, and under deliberate perturbation of the generating assumptions, rather than from a single split. Temporal modelling should be revisited only once a generating process exists in which exposure history influences subsequent state, with chronological, non-overlapping sequence construction and a comparison against tabular models using lagged features. And the ventilation layer should be evaluated against a physically grounded house model and a conventional threshold controller, on time spent beyond environmental limits, fan runtime, energy demand and temperature stability.

## ACKNOWLEDGEMENT

The authors thank Prof. Prema Kirubakaran, Head of the Department of Information Technology, Nile University of Nigeria, for institutional support.

## REFERENCES

1. Kristensen, H. H., & Wathes, C. M. (2000). Ammonia and poultry welfare: A review. World's Poultry Science Journal, 56(3), 235–245. https://doi.org/10.1079/WPS20000018

2. Miles, D. M., Branton, S. L., & Lott, B. D. (2004). Atmospheric ammonia is detrimental to the performance of modern commercial broilers. Poultry Science, 83(10), 1650–1654. https://doi.org/10.1093/ps/83.10.1650
3. Lara, L. J., & Rostagno, M. H. (2013). Impact of heat stress on poultry production. Animals, 3(2), 356–369. https://doi.org/10.3390/ani3020356
4. Elwakeel, A. E. (2025). A smart automatic control and monitoring system for environmental control in poultry houses integrated with earlier warning system. Scientific Reports, 15, 31630. https://doi.org/10.1038/s41598-025-17074-2
5. Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. Sensing and Bio-Sensing Research, 29, 100367. https://doi.org/10.1016/j.sbsr.2020.100367
6. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674
7. Teixeira, V. A., Lana, A. M. Q., Bresolin, T., Tomich, T. R., Souza, G. M., Furlong, J., Rodrigues, J. P. P., Coelho, S. G., Gonçalves, L. C., Silveira, J. A. G., Ferreira, L. D., Facury Filho, E. J., Campos, M. M., Dorea, J. R. R., & Pereira, L. G. R. (2022). Using rumination and activity data for early detection of anaplasmosis disease in dairy heifer calves. Journal of Dairy Science, 105(5), 4421–4433. https://doi.org/10.3168/jds.2021-20952
8. Manikandan, V., & Neethirajan, S. (2025). AI-powered vocalization analysis in poultry: Systematic review of health, behavior, and welfare monitoring. Sensors, 25(13), 4058. https://doi.org/10.3390/s25134058
9. Al Homidan, A., Robertson, J. F., & Petchey, A. M. (2003). Review of the effect of ammonia and dust concentrations on broiler performance. World's Poultry Science Journal, 59(3), 340–349. https://doi.org/10.1079/WPS20030021
10. 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
11. Nakrosis, A., Paulauskaite-Taraseviciene, A., Raudonis, V., Narusis, I., Gruzauskas, V., Gruzauskas, R., & Lagzdinyte-Budnike, I. (2023). Towards early poultry health prediction through non-invasive and computer vision-based dropping classification. Animals, 13(19), 3041. https://doi.org/10.3390/ani13193041
12. Sadeghi, M., Banakar, A., Minaei, S., Orooji, M., Shoushtari, A., & Li, G. (2023). Early detection of avian diseases based on thermography and artificial intelligence. Animals, 13(14), 2348. https://doi.org/10.3390/ani13142348
13. Rawdhan, S. A., Anwer, A. A., Shaaban, F. M., & Abdelhamid, M. (2026). An IoT and machine learning approach for predicting air quality parameters in poultry houses. Journal of Biosystems Engineering. Advance online publication. https://doi.org/10.1007/s42853-026-00300-8
14. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
15. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
16. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
17. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
18. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
19. Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artificial Intelligence in Agriculture, 6, 257–265. https://doi.org/10.1016/j.aiia.2022.11.003
20. Godinho, A., Vicente, R., Silva, S., & Coelho, P. J. (2025). Wireless environmental monitoring and control in poultry houses: A conceptual study. IoT, 6(2), 32. https://doi.org/10.3390/iot6020032

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

21. Lashari, M. H., Karim, S., Alhussein, M., Hoshu, A. A., Aurangzeb, K., & Anwar, M. S. (2023). Internet of Things-based sustainable environment management for large indoor facilities. PeerJ Computer Science, 9, e1623. https://doi.org/10.7717/peerj-cs.1623
22. Chen, X., Yang, L., Xue, H., Li, L., & Yu, Y. (2024). A machine learning model based on GRU and LSTM to predict the environmental parameters in a layer house, taking CO2 concentration as an example. Sensors, 24(1), 244. https://doi.org/10.3390/s24010244
23. Shwetha, V., Maddodi, B. S., Laxmi, V., Kumar, A., & Shrivastava, S. (2024). Latest trend and challenges in machine learning- and deep learning-based computational techniques in poultry health and disease management: A review. Journal of Computer Networks and Communications, 2024, 8674250. https://doi.org/10.1155/2024/8674250
24. Zhou, Z. H. (2012). Ensemble methods: Foundations and algorithms. CRC Press.
25. Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249–268.
26. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.