ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

A Machine Learning-Driven Smart Livestock Monitoring and Risk-Stratified Health Management System: A Prototype Evaluated on Simulated Data

Alirat Ibrahim[1*]; Umar Iliyasu[2]; Rosemary Mrumun Dima[3]

[1]Faculty of Computing, Federal University Dutsin-Ma, Katsina State, Nigeria
[2]Faculty of Computing, Federal University Dutsin-Ma, Katsina State, Nigeria
[3]Department of Cybersecurity, Faculty of Computing, Federal University Dutsinma

Corresponding Author: Alirat Ibrahim[1*]

Publication Date: 2026/11/03

Abstract: Late detection of disease is a persistent problem for cattle keepers in Nigeria, where veterinary visits are few and most health checks still rely on what a farmer notices. This paper describes a prototype health-monitoring system that places each animal in one of four risk tiers (low, medium, high or critical) using 15 indicators that a farm worker can record without laboratory tests and compares six classifiers on the same task. No labelled field data were available, so a simulated dataset of 2,500 records was generated from 48 reference profiles drawn from the veterinary literature. The records cover cattle, pigs, poultry and goats, with cattle making up 43.8%. Each classifier was tested twice, once on a random split and once on a split that kept all records from a profile on the same side. On the random split five of the six models classified every test record correctly, which showed that the split was the weak point and not the models. On the profile-held-out split (520 test records), Gradient Boosting and the Support Vector Machine performed best, each with 90.0% accuracy and a macro F1-score of 0.914, ahead of Random Forest (80.2%), Decision Tree (80.0%), Logistic Regression (74.4%) and a multilayer perceptron (67.7%). Gradient Boosting classified every low, medium and high record correctly but placed 52 of 156 critical records in the high tier. A PHP and MySQL web platform with a rule-based risk engine, alerts and reports were built alongside the models. The work is a prototype. The tier labels were not confirmed by a veterinarian, the trained models are not yet connected to the live platform, and nothing has been tested on real animals or with farmers, so the results do not show how the system would perform on farms.

Keywords: *Livestock Health Monitoring; Machine Learning; Risk Stratification; Data Leakage; Gradient Boosting; Cattle.*

How to Cite: Alirat Ibrahim; Umar Iliyasu; Rosemary Mrumun Dima (2026) A Machine Learning-Driven Smart Livestock Monitoring and Risk-Stratified Health Management System: A Prototype Evaluated on Simulated Data.
International Journal of Innovative Science and Research Technology, 11(10), 2554-2565.
https://doi.org/10.38124/ijisrt/26oct997

I. INTRODUCTION

Livestock farming is a vital component of rural livelihoods, economic growth, and food security in emerging economies such as Nigeria. Nigeria's livestock population reached approximately 273.8 million animals in 2022–2023, including 54.81 million cattle, 64.93 million sheep, 138.95 million goats, and more than 683 million poultry birds (National Bureau of Statistics, 2023). Approximately 40% of the protein consumed in Nigerian households comes from the livestock sector, which sustains the livelihoods of nearly 75 million people and accounts for 5% of the country's gross domestic product (Federal Ministry of Livestock Development, 2024). Despite its scale, the sector continues to face disease outbreaks, delayed diagnosis, poor veterinary infrastructure, and low productivity, with local cattle still yielding less than 10% of the world average in milk output (AU-IBAR, 2024). Across Africa more broadly, the Food and Agriculture Organization projected approximately 299 million cattle by 2022, rising to between 325 and 330 million by 2025, although disease-related mortality continues to constrain sector growth (FAO, 2024).

Manual observation is still the main way cattle health is checked on smallholder farms. It depends on what the farmer notices and when, so by the time an animal looks ill the disease may be well advanced, treatment costs more, and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

other animals may already be exposed. Machine learning applied to sensor and farm-record data is one response. Reddy et al. (2024) reviewed the livestock literature and found that Random Forest, Support Vector Machine and Logistic Regression, as well as CNN and LSTM networks, detect disease earlier than manual inspection, although few of the systems they reviewed go beyond a healthy or diseased output. Surana and Sharma (2024) reported 94% accuracy for cow health status with a CNN-LSTM model on IoT sensor data, and Lasser et al. (2021) predicted several diseases in Austrian dairy herds from routine farm data.

Two problems limit the use of such work on smallholder farms. Many systems assume sensors, connectivity and computing power that are not reliably available, and a prediction that only says an animal is sick does not tell the farmer which animal to attend to first. A tiered output is closer to the decision a farm worker faces each morning. A third problem concerns evidence. Heinen et al. (2025) point to data quality, data access and rare outcomes as recurring difficulties in cattle prediction studies, and Kapoor and Narayanan (2023) show that leakage between training and test data inflates results in many fields. Both concerns apply here, because labelled cattle health data are scarce in Nigeria and small datasets that are expanded by augmentation are easy to leak.

This paper reports a prototype that addresses these points at a small scale. Its objectives were (i) to define a four-tier risk classification built on 15 indicators that can be recorded without laboratory support; (ii) to compare six classifiers on a simulated dataset under two evaluation protocols, one that allows leakage and one that prevents it; (iii) to report accuracy, precision, recall, F1-score, AUC-ROC and confusion matrices; (iv) to build a web platform that combines record keeping, rule-based risk tiers, alerts and reports; and (v) to state clearly what the simulated data can and cannot show. Within the studies we reviewed, we did not find a cattle system that joins a tiered risk output with record keeping and alerts, but our search was not systematic and such work may exist. We therefore present the contribution as a worked prototype and an evaluation caution, and not as a new standard.

II. RELATED WORK

➢ *Machine Learning For Cattle Health*

Studies of cattle health monitoring fall into three broad groups. Sensor-based work uses pedometers, accelerometers, automated feeders and milking-system records to detect a single condition, for example lameness (Taneja et al., 2020), mastitis (Post et al., 2020; Faris et al., 2024; Tian et al., 2024) or respiratory disease in calves (Cantor et al., 2022; Bushby et al., 2024). Record-based work predicts disease occurrence or risk from routine data, such as respiratory disease in feedlot cattle (Rojas et al., 2022) and bovine leukemia virus seropositivity in beef cattle (Megahed et al., 2025). A third group describes system designs, such as the smart cattle farm architecture proposed by El Moutaouakil and Falih (2023). Garcia et al. (2020) reviewed machine learning in precision livestock farming and found the field at an early stage, with most systems stopping at disease detection. Table 1 summarises twelve of the cattle studies. The selection is illustrative and is not the result of a systematic search.

Table 1 Summary of Selected Studies on Machine Learning for Cattle Health

S/N	Study	Method used	Contribution	Limitation
1	Taneja et al. (2020)	IoT pedometers on a 150-cow farm; activity time series analysed for anomalies, with fog-based pre-processing.	Detected lameness about three days before it was visible (87% accuracy) and cut data sent to the cloud by 84%.	One farm; lameness only; requires leg-mounted pedometers.
2	Post et al. (2020)	Logistic Regression, SVM, KNN, Gaussian Naïve Bayes, Extra Trees and Random Forest on 40 months of daily sensor data from 167 cows, with feature selection and resampling.	Classified mastitis and lameness treatment events daily and identified the most useful predictors.	Labels are treatment records; small herd-level dataset; two conditions only.
3	Cantor et al. (2022)	K-nearest neighbour algorithm on automated feeder and pedometer data in a retrospective cohort of preweaned calves.	Showed that feeding and activity behaviour can indicate clinical respiratory disease status up to six days before a bout.	Retrospective data from calves only; one disease; needs automated feeders and activity sensors.
4	Rojas et al. (2022)	Predictive modelling of respiratory disease risk using feedlot records from the first 45 days after arrival.	Framed the task as risk prediction and not only diagnosis.	Feedlot population; one disease complex.
5	El Moutaouakil & Falih (2023)	Design of a smart farm system that collects cattle data.	Gave an architecture for capturing cattle data and presenting it to the farmer.	Emphasis is on data capture and display rather than prediction or risk stratification.
6	Kim et al. (2024)	Machine learning classification of sensor data on the behaviour of beef calves.	Showed that several behaviours can be monitored automatically in beef calves.	Classifies behaviour and does not predict disease or risk; one rearing system.
7	Bushby et al.	Machine learning classifiers on 229	Found that distance walked per	The authors state that

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

	(2024)	feeding, movement and social features from milk feeders and location sensors for 172 group-housed calves.	day was the most important feature for predicting respiratory disease events.	further improvement is needed before reliable prediction; calves only.
8	Faris et al. (2024)	SVM, Logistic Regression, Gaussian Naïve Bayes, KNN, decision tree (CART) and Random Forest on milking system records of 1,493 mastitic and 2,387 healthy cows.	Compared six classifiers for early detection of clinical mastitis.	Modest accuracy (62% to 74%); single herd; one disease.
9	Tian et al. (2024)	Nine machine learning algorithms on activity, rumination, milk yield and milk conductivity data from 447 mastitic and 2,146 healthy cows on five farms.	Built sensor-based models for naturally occurring clinical mastitis across four stages of lactation.	One region and one disease; needs monitoring collars and milking system data.
10	Girdauskaitė et al. (2025)	Decision Tree, Logistic Regression, MLP, Random Forest and Gradient Boosting on milk and behaviour sensor data from 206 Holstein cows.	Random Forest gave the highest accuracy (0.857) for early, non-invasive detection of health disorders.	Two farms; first 100 days of lactation; needs commercial sensor systems.
11	Heinen et al. (2025)	Mixed-methods review of published predictive models for cattle disease, covering inputs, algorithms and outcome definitions.	Identified data quality, data access and rare outcomes as recurring challenges.	A review; does not build or test a deployable system.
12	Megahed et al. (2025)	Supervised algorithms applied to ten years of retrospective records.	Predicted bovine leukemia virus seropositivity in beef cattle from routine records.	One disease; one state; retrospective records.

Three patterns stand out in Table 1. Most studies predict one disease or behaviour and not a general level of risk. Many rely on commercial sensors or automated milking systems and are tested on one or a few farms, which limits their use on herds without such equipment. And, as far as the sources consulted show, none of them reports a tiered risk output together with record keeping, alerts and reporting for farm staff.

➢ *Classification Algorithms*

Six classifiers were used. Random Forest (Breiman, 2001) averages many decision trees, each trained on a random sample of the records. Gradient Boosting (Friedman, 2001) builds trees one after another, with each tree correcting the errors of those before it. The Support Vector Machine (Cortes & Vapnik, 1995) looks for the boundary with the widest margin between classes. Logistic Regression (Hosmer et al., 2013) estimates class probabilities with a linear model. A Decision Tree splits the records by simple rules and is easy to read but unstable. The multilayer perceptron (Goodfellow et al., 2016) is a small neural network. Panigrahy et al. (2026) compared several conventional algorithms on animal disease data and reported Random Forest as the best (95.77% accuracy) and the Support Vector Machine as the weakest (81.69%). Their data and task differ from ours, so the figures are not directly comparable. Grinsztajn et al. (2022) found that tree-based models often outperform neural networks on tabular data of moderate size, which is the situation in this study.

➢ *Evaluation and Data Leakage*

A model is only useful if its measured performance predicts how it will behave on animals it has not seen. Kapoor and Narayanan (2023) reviewed papers from many fields and found leakage, where information from the test data reaches the training process, to be a common cause of results that are too good and cannot be reproduced. Random splitting is especially risky when many records come from one animal, one farm or, as here, one reference profile. Ezanno et al. (2021) also note that standardised, shared animal-health datasets are scarce, which pushes researchers towards small or augmented data. For these reasons we report results under a random split and under a split grouped by profile, and we do not use the first to rank the models.

➢ *Gap Addressed*

In the studies reviewed here the outcome is usually a single disease or behaviour. Few combine a predictive model with the record keeping, alerts and reporting that a farm needs, tiered risk outputs are less common than disease labels, and evaluation is often confined to a few farms. This paper explores these points at prototype level. It does not claim to close them, since field validation is still outstanding.

III. MATERIALS AND METHODS

➢ *Study Design*

The work had two linked parts. The first was an offline experiment in which six classifiers were trained and compared under identical conditions. The second was a web platform that applies risk tiers and alerts to animal records. Both use the same indicators and the same four tiers, but they run separately: the models were developed and tested in Python, whereas the platform assigns live tiers with a rule-based engine (Section 3.7). Fig. 1 shows the overall workflow.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

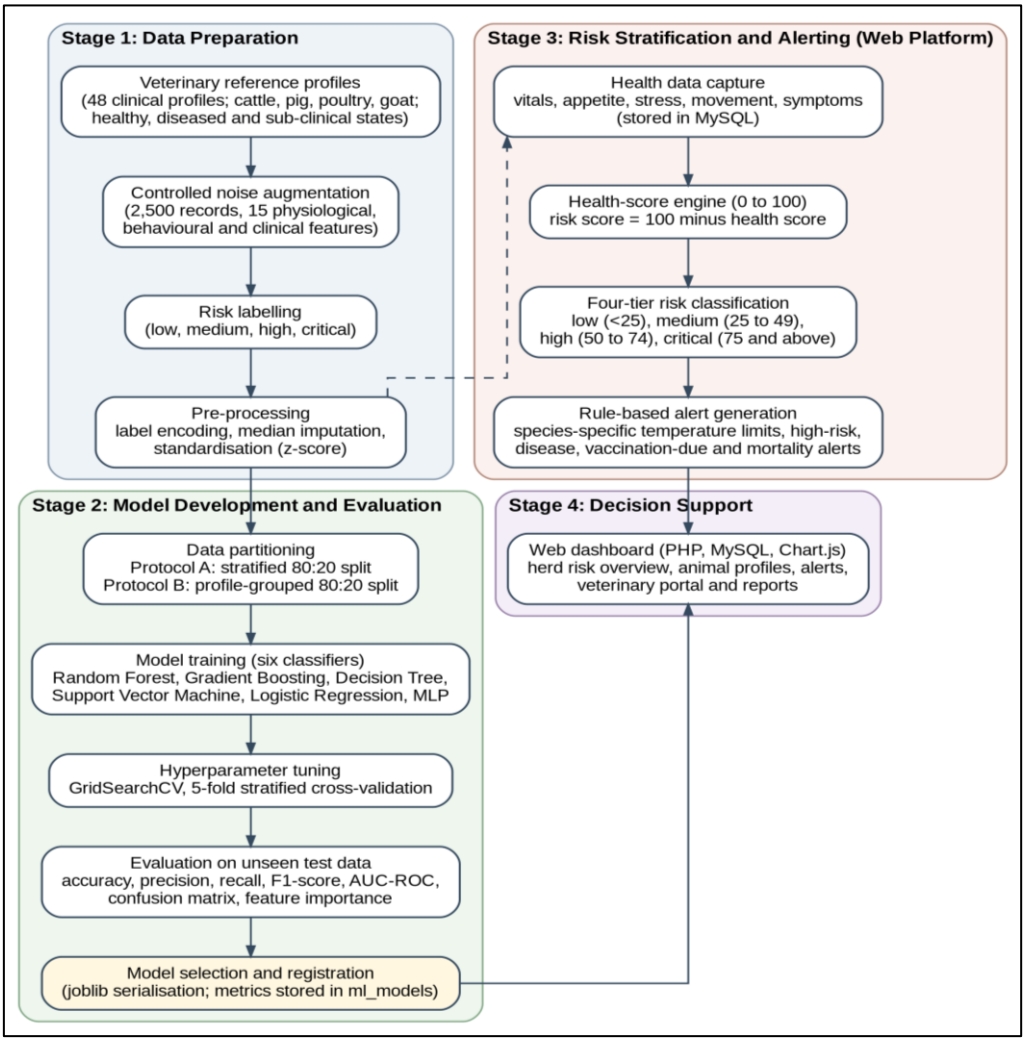


Fig 1 Workflow of the Study, from Data Preparation to Decision Support

➢ *Dataset*

Field records labelled with risk tiers were not available, so the dataset was built from 48 reference profiles compiled by the first author from clinical reference ranges and disease descriptions in the veterinary literature. Each profile represents one animal condition for one species, for example a healthy adult cow or a cow with mastitis. Profiles were grouped as healthy, sub-clinical or diseased and given a risk tier according to clinical severity. They were then expanded to 2,500 records by adding bounded random variation to body temperature, heart rate, respiratory rate, body weight and days since last treatment. The other ten indicators were copied from the parent profile. A fixed random seed was used, so the dataset can be regenerated exactly. Table 2 shows its composition. The profiles cover four species, and cattle account for 1,096 of the 2,500 records. The models were trained on the combined data and not on cattle alone.

Table 2 Composition of the Simulated Dataset

Category	Group	Records	Share (%)
Species	Cattle	1,096	43.8
	Pig	520	20.8
	Poultry	468	18.7
	Goat	416	16.6
Risk tier	Low	1,200	48.0
	Medium	468	18.7
	High	416	16.6
	Critical	416	16.6
Total		2,500	100.0

The dataset has limits that affect how the results should be read. The profiles were compiled by one researcher, and the tier given to each profile is that researcher's judgement of clinical severity, which has not been confirmed by a veterinarian. Only five of the 15 indicators were varied, so records from one profile share the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

other ten values and every record inherits its profile's tier. Boundaries between tiers are therefore likely to be cleaner than they would be in field data. The data are simulated and carry no information about real animals.

➢ *Indicators and Risk Tiers*

Fifteen indicators were used (Table 3). All of them can be recorded by a farm worker or veterinarian without laboratory analysis. Records were assigned to four ordered tiers. Low risk describes animals with normal indicators. Medium risk describes mild deviations or sub-clinical signs that call for closer observation. High risk describes marked abnormalities consistent with active disease. Critical risk describes severe, possibly life-threatening presentations that need immediate veterinary attention.

Table 3 Indicators Used for Risk Classification

Indicator	Type	Description
age_years	Numeric	Age of the animal in years
weight_kg	Numeric	Body weight in kilograms
body_temp_c	Numeric	Rectal temperature in degrees Celsius
heart_rate_bpm	Numeric	Heart rate in beats per minute
respiratory_rate	Numeric	Breaths per minute
appetite_encoded	Ordinal	Appetite level, from poor to good
stress_encoded	Ordinal	Observed stress level, from low to high
body_condition	Numeric	Body condition score on a five-point scale
movement_score	Ordinal	Movement or activity score from 1 to 5
feed_intake_kg	Numeric	Daily feed intake in kilograms
water_intake_l	Numeric	Daily water intake in litres
milk_production_l	Numeric	Daily milk yield in litres (zero where not applicable)
symptoms_present	Binary	Whether clinical symptoms were observed
species_encoded	Categorical	Species, label-encoded
days_since_last_treatment	Numeric	Days since the last treatment

➢ *Pre-Processing*

Pre-processing was placed inside a scikit-learn pipeline (Pedregosa et al., 2011) so that every step was fitted on the training portion only. Appetite, stress level and species were label-encoded, with ordinal indicators kept in their natural order. Missing numeric values were replaced by the training-set median, and all features were then standardised to zero mean and unit variance. Standardisation matters for the Support Vector Machine, Logistic Regression and the neural network, which are sensitive to feature scale, and it does no harm to the tree-based models.

➢ *Models and Tuning*

The six classifiers were implemented with scikit-learn 1.8.0 in Python 3.12. Hyperparameters were chosen by exhaustive grid search (Table 4), scoring each combination with five-fold stratified cross-validation on the training data and using accuracy as the criterion. The folds were stratified by tier but not grouped by profile, so a validation fold could contain records from the same profile as the training folds. Cross-validation scores were therefore used only to choose settings and are not reported as performance. All reported figures come from held-out test sets.

Table 4 Hyperparameter Search Grids

Algorithm	Hyperparameters searched
Random Forest	Trees: 50, 100, 200; maximum depth: 8, 12, 15, none; minimum samples per leaf: 1, 2, 4
Gradient Boosting	Trees: 50, 100, 150; maximum depth: 3, 5, 8; learning rate: 0.05, 0.1, 0.2
Decision Tree	Maximum depth: 5, 10, 15, none; minimum samples to split: 2, 5, 10; criterion: Gini, entropy
Support Vector Machine	C: 0.1, 1, 10; kernel: RBF, linear; gamma: scale, auto
Logistic Regression	C: 0.01, 0.1, 1, 10; solver: lbfgs, liblinear
Multilayer Perceptron	Hidden layers: (64, 32), (128, 64, 32); activation: ReLU, tanh; L2 penalty: 0.0001, 0.001

➢ *Evaluation Protocols and Metrics*

Two protocols were applied to the same data and the same models. Protocol A is the usual approach: a stratified random split with 2,000 records for training and 500 for testing. Because the data come from 48 profiles, records from one profile can fall on both sides of this split. Protocol B removes that overlap. Records were grouped by the profile that generated them, and whole groups were assigned to either training (1,980 records) or testing (520 records) with a grouped shuffle split. The test set then contains profiles the model has never seen in any form, which is closer to meeting a new animal. Results are single estimates from one test set per protocol, and no confidence intervals were computed.

Performance was measured with accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score and macro-averaged one-versus-rest AUC-ROC, together with confusion matrices. Macro averaging gives each tier equal weight, which matters because the medium tier is the smallest. For a given tier, TP, TN, FP and FN are the true positives, true negatives, false positives and false negatives:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

Precision asks how many of the animals placed in a tier belong there, and recall asks how many of the animals in a tier were found. For the critical tier, low precision means false alarms and low recall means seriously ill animals are missed, which is the more dangerous error. AUC-ROC does not depend on one decision threshold. It measures how well predicted probabilities rank animals of one tier above the rest, so a model can have a high AUC-ROC and a modest accuracy when its ranking is sensible but its tier boundaries are in the wrong place.

➢ *Web Platform and Rule-Based Risk Engine*

The platform has three layers. The interface is written in HTML, CSS and JavaScript with Chart.js 4.4. The application layer is written in PHP 8.3 and handles sign-in, validation, risk scoring, alerts and reports. The data layer is a MySQL-compatible database (MariaDB 10.11 was used in testing) that holds farms, animals, health records, vaccinations, treatments, alerts and users. Five user roles (super administrator, farm administrator, veterinarian, farm worker and researcher) see only the modules they need, and passwords are stored as hashes.

Live risk tiers come from a transparent health-score engine. Each assessment starts at 100 points and loses points for departures from normal: abnormal body temperature (up to 25), heart rate (up to 15), respiratory rate (12), body condition (15), movement (15) and activity (10), with further adjustments for poor appetite and high stress. The risk score is 100 minus the health score, and the tier is read from it: low below 25, medium from 25 to 49, high from 50 to 74 and critical from 75 upward. A rule-based engine was chosen because it needs no trained model and lets a veterinarian see exactly why an animal was placed in a tier. The trained classifiers serve as a benchmark for the same task and are not yet connected to the live endpoint. Alerts are triggered by threshold rules on temperature, tier, active disease, overdue vaccination and mortality, and each alert carries a severity that users can filter and resolve.

➢ *Ethical Status*

This study is a software prototype tested on simulated data. No human participants took part, no data were collected from live animals, and no farm or veterinary records were used. The demonstration records in the platform are fictitious, and no ethical clearance was sought for the work reported here. Any follow-up that collects farm records or tests the platform with farmers and veterinarians will need approval from the relevant institutional ethics committee, informed consent from participants, and arrangements to protect farm and personal data.

IV. RESULTS

➢ *Dataset Characteristics*

Low-risk records formed the largest group (48.0%) and the medium tier the smallest (18.7%), as Fig. 2 shows. The imbalance is moderate, and macro-averaged metrics were used so that the smaller tiers were not overshadowed. Although there are 2,500 rows, they come from only 48 profiles, so each profile contributes about 52 records that differ only in the five varied measurements. This structure explains the results in the next section.

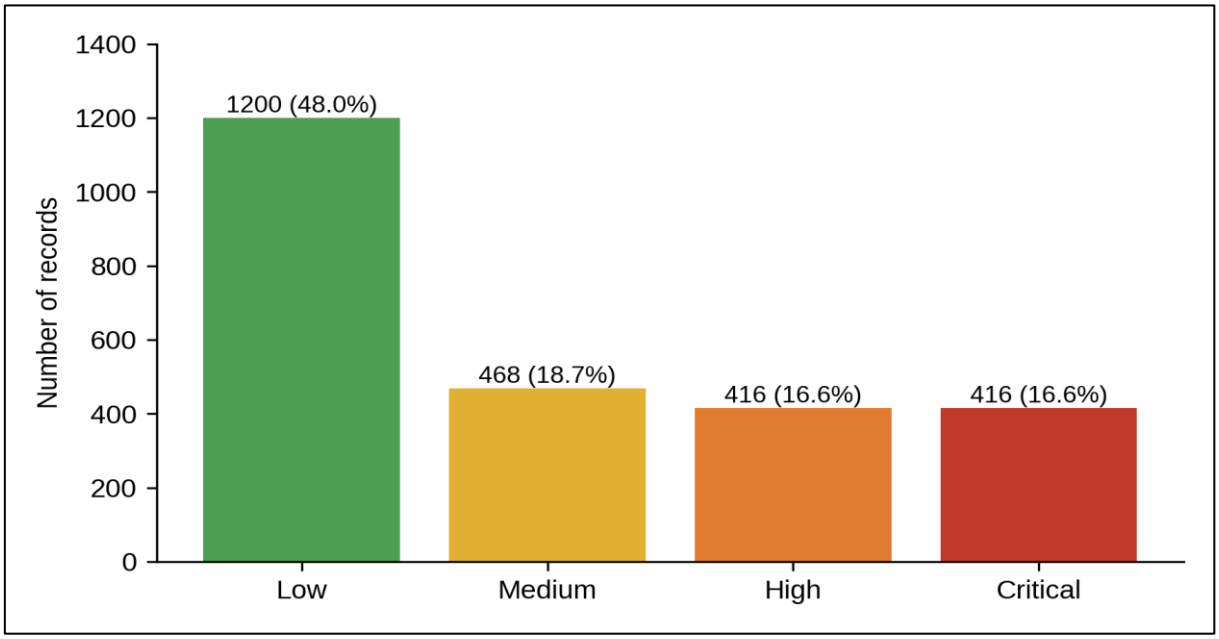


Fig 2 Distribution of Records by Risk Tier

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

➢ *Results Under Protocol A (Random Split)*

Table 5 gives the test results when records were split at random. Five of the six classifiers classified every test record correctly, and the multilayer perceptron was close behind.

Table 5 Test-Set Performance Under Protocol A (Stratified Random Split, 500 Test Records)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
Random Forest	100.0	100.0	100.0	100.0	1.000
Gradient Boosting	100.0	100.0	100.0	100.0	1.000
Decision Tree	100.0	100.0	100.0	100.0	1.000
Support Vector Machine	100.0	100.0	100.0	100.0	1.000
Logistic Regression	100.0	100.0	100.0	100.0	1.000
Multilayer Perceptron	98.6	98.1	97.9	97.9	1.000

A score of 100% from five different algorithms, including a heavily regularised linear model, does not show that the problem has been solved. It shows that the test set contains nothing the models have not already seen. Under this split every test record has near-identical siblings from the same profile in the training set, so a model can succeed by recognising the profile. Cross-validation scores used in tuning were affected in the same way, and these figures were not used to rank the models.

➢ *Results Under Protocol B (Profile-Held-Out Split)*

When whole profiles were withheld, performance separated the models clearly (Table 6 and Fig. 3). Gradient Boosting and the Support Vector Machine tied for the best accuracy at 90.0% and a macro F1-score of 0.914. Random Forest and the Decision Tree followed at about 80%, Logistic Regression reached 74.4%, and the multilayer perceptron was weakest at 67.7%.

Table 6 Test-Set Performance Under Protocol B (Profile-held-Out Split, 520 Test Records)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
Random Forest	80.2	83.5	83.5	83.5	0.976
Gradient Boosting	90.0	93.8	91.7	91.4	0.940
Decision Tree	80.0	79.2	83.3	78.3	0.885
Support Vector Machine	90.0	93.8	91.7	91.4	0.976
Logistic Regression	74.4	80.1	78.7	69.0	0.998
Multilayer Perceptron	67.7	58.6	69.2	61.7	0.993

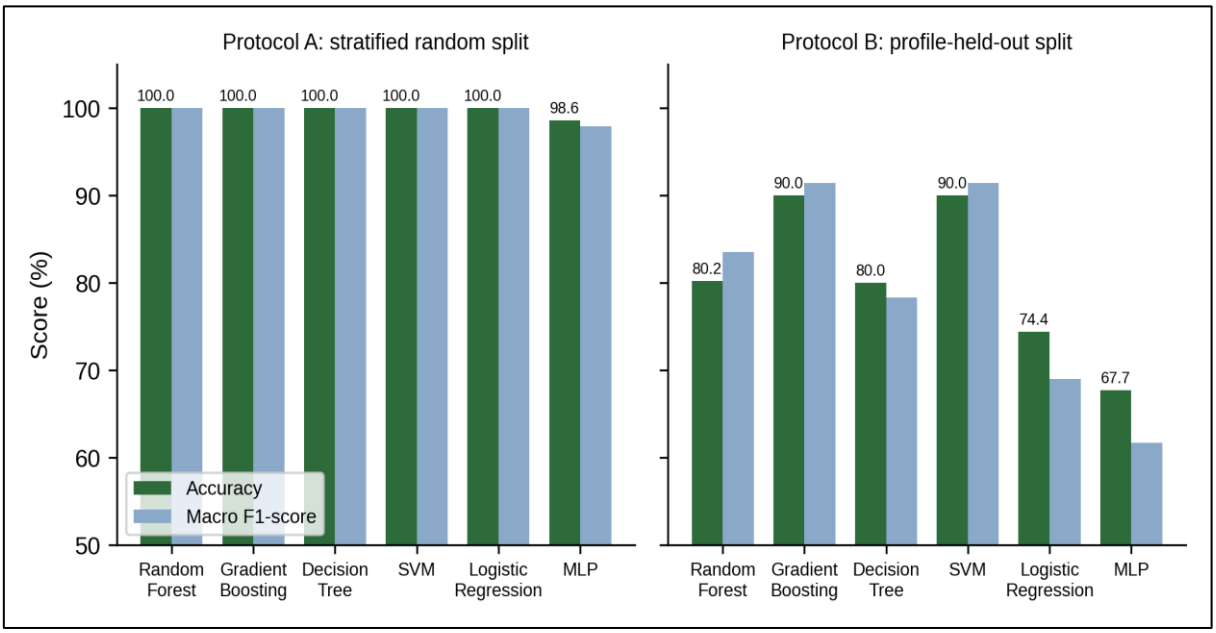


Fig 3 Accuracy and Macro F1-Score of the Six Classifiers Under the Two Protocols

The fall from Protocol A to Protocol B, for example from 100% to 80.2% for Random Forest, shows how much of the first result came from leakage. AUC-ROC stayed high for every model, and Logistic Regression had the highest value (0.998) despite one of the lowest accuracies. That pattern means the model ranked animals sensibly by risk but put the tier boundaries in the wrong places for unseen profiles, which points to threshold calibration and not ranking as the weak spot.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

➢ *Confusion Matrix and Error Analysis*

Fig. 4 and Table 7 show the confusion matrix for Gradient Boosting under Protocol B. Every low, medium and high record was classified correctly. The only errors were 52 critical records assigned to the high tier, which is 52 of 156 critical records (33%). No record moved by more than one tier, and no critical animal was labelled low or medium. Table 8 gives the per-tier measures worked out from the matrix. They show that the cost of the errors falls on the critical tier, whose recall was 66.7%, while the precision of the high tier fell to 75.0% because it absorbed the missed critical animals.

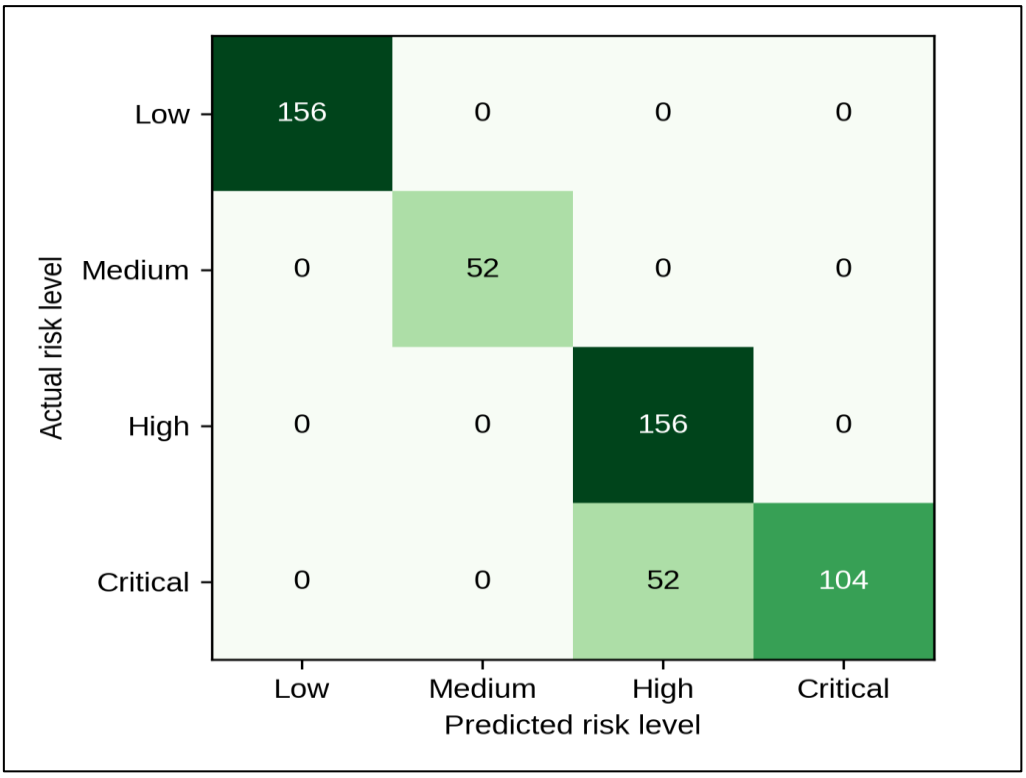


Fig 4 Confusion Matrix for Gradient Boosting Under Protocol B (520 Test Records)

Table 7 Confusion Matrix for Gradient Boosting Under Protocol B

Actual \ Predicted	Low	Medium	High	Critical
Low	156	0	0	0
Medium	0	52	0	0
High	0	0	156	0
Critical	0	0	52	104

Table 8 Per-Tier Measures for Gradient Boosting Under Protocol B (Derived from Table 7)

Tier	Precision (%)	Recall (%)	F1-score (%)	Records
Low	100.0	100.0	100.0	156
Medium	100.0	100.0	100.0	52
High	75.0	100.0	85.7	156
Critical	100.0	66.7	80.0	156

Random Forest made errors in both directions between the high and critical tiers, with 51 critical records predicted as high and 52 high records predicted as critical. Confusion between neighbouring tiers is expected because severity is a continuum, and it is less harmful than confusion between distant tiers. It is still a caution for use. A critical animal predicted as high would be flagged, but possibly with less urgency than it needs, so a high-tier alert should prompt a check for signs of critical deterioration.

➢ *Feature Importance*

Under Protocol A, Random Forest relied most on appetite (24.9%), stress level (18.8%) and movement score (16.1%), as Fig. 5 shows. These three behavioural indicators made up about 60% of total importance, while body temperature and respiratory rate contributed only 3.5% and 3.2%. Under Protocol B, Gradient Boosting concentrated even more on appetite (81%) and stress (15%).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

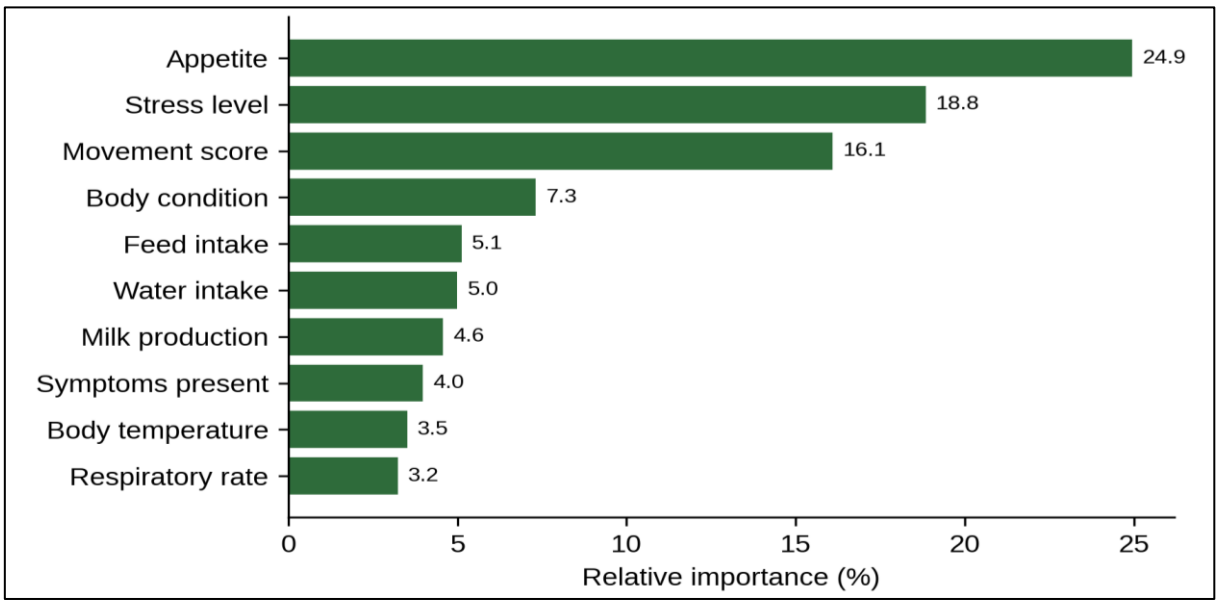


Fig 5 Ten Most Important Features for Random Forest (Protocol A)

This should be read with care. Only five measurements were varied by augmentation, whereas appetite and stress were fixed within each profile and differ cleanly between healthy and diseased profiles, so a model can use them as near-perfect identifiers. In real animals these indicators are noisier and body temperature carries more information than shown here. Appetite and behavioural change are still well-recognised early signs of illness, in line with the interest in behaviour monitoring in the sensor literature (Kim et al., 2024).

➢ *Platform Implementation*

The platform was run against a demonstration database of three farms, 23 animals (8 cattle, 5 pigs, 5 poultry and 5 goats), 198 health records and the alerts generated from them. The dashboard (Fig. 6) summarises farms, animals, healthy and sick animals and vaccinations due, and a herd risk panel showed 14 animals in the low tier (61%), 4 medium (17%), 4 high (17%) and 1 critical (4%). The risk stratification module (Fig. 7) counts animals in each tier and lists those that need follow-up in order of risk score. Eleven modules were opened in turn with the populated database (sign-in, dashboard, livestock registry, health monitoring, risk stratification, disease prediction, alert centre, veterinary portal, vaccination management, reports and the machine learning dashboard), and each loaded and displayed data without errors. This was a functional check by the first author and not a usability study with farmers.



Fig 6 Dashboard with Herd Risk Overview

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

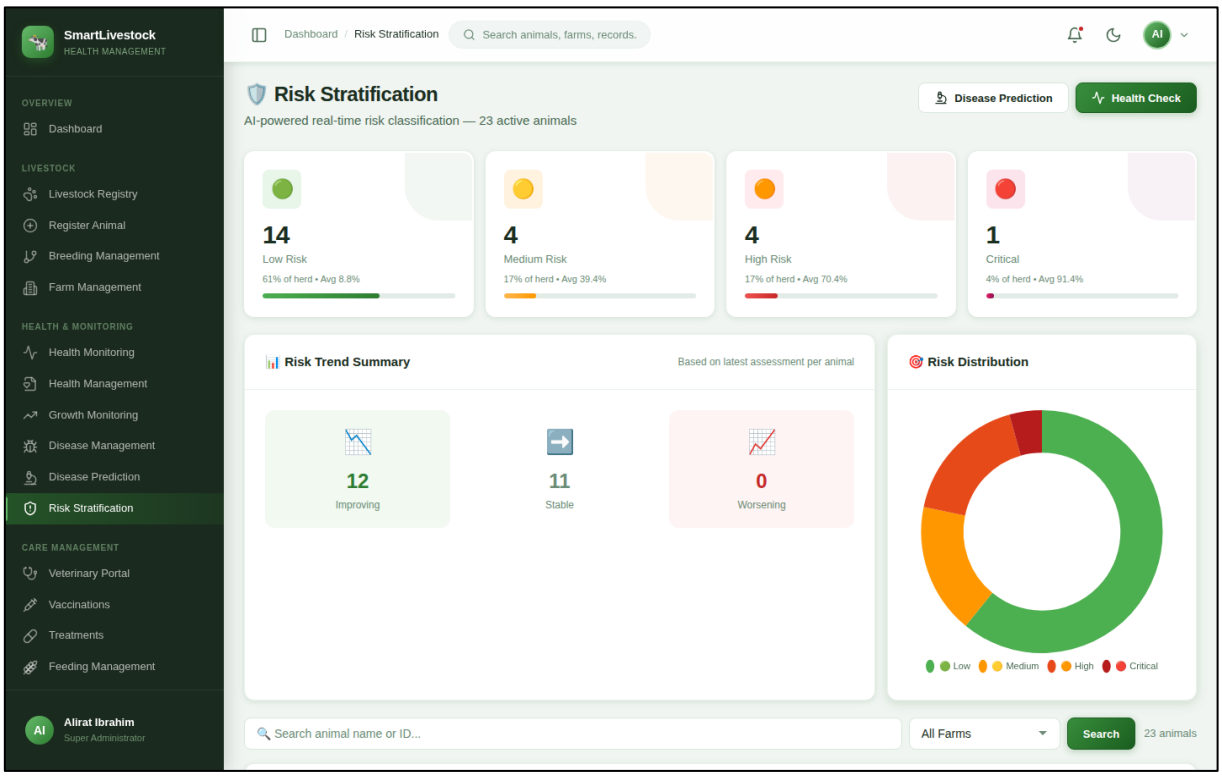


Fig 7 Risk Stratification Module

V. DISCUSSION

The most useful observation from this prototype concerns evaluation. The same six models and the same data gave near-perfect scores under a random split and a spread of 68% to 90% accuracy when whole profiles were held out. This matches the warning of Kapoor and Narayanan (2023) that leakage can make machine learning results look far better than they are, and it supports the concern of Heinen et al. (2025) about data quality in cattle prediction work. Studies that augment small datasets and then split at random should be read with this in mind, and that applies to Protocol A in this paper as much as to anyone else's work.

On model choice, Gradient Boosting and the Support Vector Machine were the most dependable on this dataset, which fits the finding of Grinsztajn et al. (2022) that tree-based models usually do well on tabular data, while the neural network, with the least data per parameter, did worst. The ranking should not be treated as settled. The held-out test set holds only a small number of profiles, and the gap between the two leading models and Random Forest could change with a different grouping. A reliable ranking would need many more independent animals than were available here.

Our accuracy figures cannot be set against those of Panigrahy et al. (2026) or Megahed et al. (2025). Those studies used different data, different tasks and different splits, and ours comes from simulated profiles. The comparison that matters in this paper is between the two protocols on the same data. The result for Gradient Boosting also has a practical meaning. Its errors were confined to one tier boundary, but the missed animals were the most serious ones. If a model like this is ever used, the critical tier needs a threshold that favours recall over precision, and the rule-based engine should stay in place as a transparent baseline. That transparency is valuable where farmers and veterinarians need to trust and act on the output. The cost is that the thresholds are set by judgement, and the classifiers could help to check and refine them once they are trained on field records with veterinarian-confirmed labels.

VI. LIMITATIONS

This is a prototype built on simulated data, and its findings should be read as such. The 2,500 records come from 48 profiles compiled by one researcher, so the effective sample is much smaller than the record count suggests, and the differences between the better models should not be treated as established. The tier labels were not independently confirmed by a veterinarian. The dataset covers four species, and cattle make up 43.8% of it, so the results are not specific to cattle, and retraining and testing on cattle records alone is still to be done. Tuning folds were not grouped by profile, results are single estimates without confidence intervals, and the demonstration database is fictitious. The trained models are not connected to the live platform, which uses a rule-based engine, and the platform has not been evaluated by farmers or veterinarians. The study therefore does not establish field accuracy, clinical validity or usability.

VII. CONCLUSION

This prototype shows that a four-tier health-risk system with record keeping, alerts and reports can be built with ordinary web technology and with indicators that farm

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

staff can record without special equipment. On the simulated dataset, Gradient Boosting and the Support Vector Machine were the most dependable classifiers on unseen profiles and the multilayer perceptron the least. The clearer lesson is that reported accuracy depended heavily on how the data were split, and a perfect score should prompt a check for leakage before it is taken as evidence of a good model. Because the models were trained on augmented profiles that a veterinarian has not confirmed, are not yet connected to the live system and have not been tested on real animals, the work demonstrates a method and a working prototype. It does not establish field accuracy, clinical usefulness or acceptance by farmers and veterinarians.

RECOMMENDATIONS

Farmers and cooperatives should treat tools of this kind as an aid to routine checks and veterinary advice and not as a replacement, and none of the output of this prototype should guide treatment decisions. Veterinary services should review and confirm the tier definitions and thresholds before any pilot and should take part in labelling a field dataset. Researchers working with augmented or repeated-measure data should split by animal, farm or profile, report random and grouped splits side by side, and give AUC-ROC and confusion matrices along with accuracy. Developers should keep the rule-based engine as a transparent baseline, connect a trained model only after field validation, and tune the critical tier for recall. Government and development bodies, such as the Federal Ministry of Livestock Development, could support small pilots with real herds, with ethical approval and informed consent in place from the start.

➢ *Highlights*

- A four-tier risk classification was built on 15 indicators that need no laboratory tests.
- Six classifiers were compared under a random split and a profile-held-out split.
- Accuracy was 100% for five models on the random split but ranged from 67.7% to 90.0% on held-out profiles.
- Gradient Boosting classified every low, medium and high record correctly but placed 52 of 156 critical records in the high tier.
- A working web platform with rule-based tiers, alerts and reports was built. All results come from simulated data.

ABBREVIATIONS

Term	Meaning	Term	Meaning
AUC-ROC	Area under the receiver operating characteristic curve	MLP	Multilayer perceptron
CNN	Convolutional neural network	RF	Random Forest
GB	Gradient Boosting	SVM	Support Vector Machine
IoT	Internet of Things	TP, TN, FP, FN	True and false positives and negatives
LR	Logistic Regression	LSTM	Long short-term memory

REFERENCES

[1]. AU-IBAR. (2024). Livestock productivity report: Africa regional overview. African Union Interafrican Bureau for Animal Resources.

[2]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3]. Bushby, E. V., Thomas, M., Vázquez-Diosdado, J. A., Occhiuto, F., & Kaler, J. (2024). Early detection of bovine respiratory disease in pre-weaned dairy calves using sensor based feeding, movement, and social behavioural data. *Scientific Reports*. https://doi.org/10.1038/s41598-024-58206-4

[4]. Cantor, M. C., Casella, E., Silvestri, S., Renaud, D. L., & Costa, J. H. C. (2022). Using machine learning and behavioral patterns observed by automated feeders and accelerometers for the early indication of clinical bovine respiratory disease status in preweaned dairy calves. *Frontiers in Animal Science, 3*, Article 852359. https://doi.org/10.3389/fanim.2022.852359

[5]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[6]. El Moutaouakil, K., & Falih, N. (2023). A design of a smart farm system for cattle monitoring. *Indonesian Journal of Electrical Engineering and Computer Science, 32*(2), 857–864. https://doi.org/10.11591/IJEECS.V32.I2.PP857-864

[7]. Ezanno, P., Picault, S., Beaunée, G., Bailly, X., Muñoz, F., Duboz, R., Monod, H., & Guégan, J.-F. (2021). Research perspectives on animal health in the era of artificial intelligence. *Veterinary Research, 52*, 40. https://doi.org/10.1186/s13567-021-00902-4

[8]. Faris, D. N., Gad, A. M., El-Tarabany, M. S., Ramadan, S. I., Afifi, G. G., & Manaa, E. A. (2024). Machine learning algorithms for clinical mastitis prediction in a dairy herd use automated milking system data. *Journal of Advanced Veterinary Research, 14*(6), 975–981.

[9]. Federal Ministry of Livestock Development. (2024). Nigeria livestock sector report 2023/2024. Government of Nigeria.

[10]. Food and Agriculture Organization. (2024). Livestock and disease in Africa: Economic and food security dimensions. FAO Publications.

[11]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[12]. Garcia, R., Aguilar, J., Toro, M., Pinto, A., & Rodriguez, P. (2020). A systematic literature review on the use of machine learning in precision livestock

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26oct997

farming. *Computers and Electronics in Agriculture, 179*, 105826.

[13]. Girdauskaitė, A., Grigė, S., Džermeikaitė, K., Krištolaitytė, J., Malašauskienė, D., Televičius, M., Šertvytytė, G., Lembovičiūtė, G., & Antanaitis, R. (2025). Supervised machine learning approaches for early detection of metabolic and udder health disorders in dairy cows using sensor-derived data. *Frontiers in Veterinary Science, 12*, Article 1726719. https://doi.org/10.3389/fvets.2025.1726719

[14]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[15]. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems, 35*. https://arxiv.org/abs/2207.08815

[16]. Heinen, L., Larson, R. L., & White, B. J. (2025). Disease prediction in cattle: A mixed-methods review of predictive modeling studies. *Animals, 15*(17), Article 2481. https://doi.org/10.3390/ani15172481

[17]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

[18]. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

[19]. Kim, S.-J., Jin, X.-C., Bharanidharan, R., & Kim, N.-Y. (2024). Monitoring multiple behaviors in beef calves raised in cow-calf contact systems using a machine learning approach. *Animals, 14*(22), Article 3278. https://doi.org/10.3390/ani14223278

[20]. Lasser, J., Matzhold, C., Egger-Danner, C., Fuerst-Waltl, B., Steininger, F., Wittek, T., & Klimek, P. (2021). Integrating diverse data sources to predict disease risk in dairy cattle: A machine learning approach. *Journal of Animal Science, 99*(12), skab294. https://doi.org/10.1093/jas/skab294

[21]. Megahed, A. A., Bommineni, Y. R., Short, M., Galvão, K. N., & Bittar, J. H. J. (2025). Using supervised machine learning algorithms to predict bovine leukemia virus seropositivity in Florida beef cattle: A 10-year retrospective study. *Journal of Veterinary Internal Medicine, 39*(3), Article e70070. https://doi.org/10.1111/jvim.70070

[22]. National Bureau of Statistics. (2023). Agricultural sector summary report 2022/2023. NBS Nigeria.

[23]. Panigrahy, B., Subudhi, A., Harichandan, T., Padhy, N., & Panigrahi, R. (2026). Comparative analysis of machine-learning and deep-learning approaches for accurate animal disease prediction and health risk assessment. *Engineering Proceedings, 124*(1), Article 52. https://doi.org/10.3390/engproc2026124052

[24]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[25]. Post, C., Rietz, C., Büscher, W., & Müller, U. (2020). Using sensor data to detect lameness and mastitis treatment events in dairy cows: A comparison of classification models. *Sensors, 20*(14), Article 3863. https://doi.org/10.3390/s20143863

[26]. Reddy, S. S., D R, M., S, J., & C, N. (2024). A comprehensive review of machine learning approaches in livestock health monitoring. *Journal of Big Data Technology and Business Analytics, 3*(3), 12–19. https://doi.org/10.46610/JBDTBA.2024.v03i03.002

[27]. Rojas, H. A., White, B. J., Amrine, D. E., & Larson, R. L. (2022). Predicting bovine respiratory disease risk in feedlot cattle in the first 45 days post arrival. *Pathogens, 11*(4), Article 442. https://doi.org/10.3390/pathogens11040442

[28]. Surana, J., & Sharma, S. K. (2024). Predicting cow health with a smart framework: A big data and deep learning-based IoT approach. *International Journal of Intelligent Systems and Applications in Engineering, 12*(4), 3201–3210.

[29]. Taneja, M., Byabazaire, J., Jalodia, N., Davy, A., Olariu, C., & Malone, P. (2020). Machine learning based fog computing assisted data-driven approach for early lameness detection in dairy cattle. *Computers and Electronics in Agriculture, 171*, Article 105286. https://doi.org/10.1016/j.compag.2020.105286

[30]. Tian, H., Zhou, X., Wang, H., Xu, C., Zhao, Z., Xu, W., & Deng, Z. (2024). The prediction of clinical mastitis in dairy cows based on milk yield, rumination time, and milk electrical conductivity using machine learning algorithms. *Animals, 14*(3), Article 427. https://doi.org/10.3390/ani14030427