IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 3, Ser. II (May. – June. 2026), PP 01-09
www.iosrjournals.org

Development of a Smart Disease Diagnostic Model for Farm Animals Using Machine Learning and Deep Learning

Md. Meem Mursalin Chowdhury[1], Md. Morshedul Arefin[1], Md. Rokanujjaman[1]
[1](Department of Computer Science and Engineering, University of Rajshahi, Bangladesh)

Abstract:
Background : *Livestock plays a critical role in the agro-based economy of Bangladesh, yet production is frequently hampered by diseases and inadequate veterinary healthcare services. To overcome these constraints, this study proposes a smart disease diagnostic model that utilizes machine learning (ML) and deep learning (DL) algorithms to process clinical symptoms and field images for early disease detection.*
Materials and Methods : *Using a synthetic clinical dataset of 120 instances and a large-scale image dataset of 8,014 annotated images, we evaluated multiple classical classifiers and object detection models.*
Results : *The results indicate that the CatBoost algorithm achieved a 100% testing accuracy for symptom-based classification, while the YOLOv12 architecture demonstrated robust field-image classification with an overall precision of 0.779.*
Conclusion : *This system promises to aid in better risk assessment, precise decision-making, and enhanced livestock productivity.*
Key Word : *Machine Learning, Deep Learning, Disease Diagnosis, Farm Animals, YOLOv12, CatBoost, Livestock Management*

Date of Submission: 25-05-2026 Date of Acceptance: 03-06-2026

I. INTRODUCTION

Bangladesh is an agro-based country, and her economy is largely dependent on agriculture. Livestock is an important subsector contributing 13.46% to the agricultural GDP and 1.47% to the national GDP with an increment of 3.47% annually. It plays pivotal roles in supplying essential nutrients in terms of meat, milk, and eggs for healthy national development, income generation, poverty alleviation, and women's empowerment. In 2019-2020, there were about 243.91 million cattle, 14.93 million buffaloes, 36.07 million sheep, and 264.35 million goats in Bangladesh[1]. Farm animals such as cattle, buffalo, goats, and sheep frequently suffer from various diseases resulting in increased morbidity and mortality, which ultimately reduces overall productivity. This is largely because of inadequate veterinary health care services due mainly to shortage of veterinarians, unavailability of necessary modern equipment, lack of trained technicians and large number of animal population under each veterinary livestock offices and veterinary hospitals of a large administrative area.

The development and introduction of smart veterinary diagnostic and healthcare services present a promising approach to overcoming these constraints. A smart disease diagnostic model is a software based diagnostic system which uses algorithmic and machine learning models for data processing, training and predicting diseases. In such a system clinical history, clinical signs, epidemiological data and images of disease and deformity are uploaded in the system which is processed by an algorithmic method to give a precise diagnostic clue and suggestions. The model will work better in diseases where the disease lesions and deformity are visible on the body surface such as foot and mouth disease (FMD), lumpy skin disease (LSD), peste des petits ruminants (PPR), rabies, bottle jaw, mastitis, papillomatosis, etc.

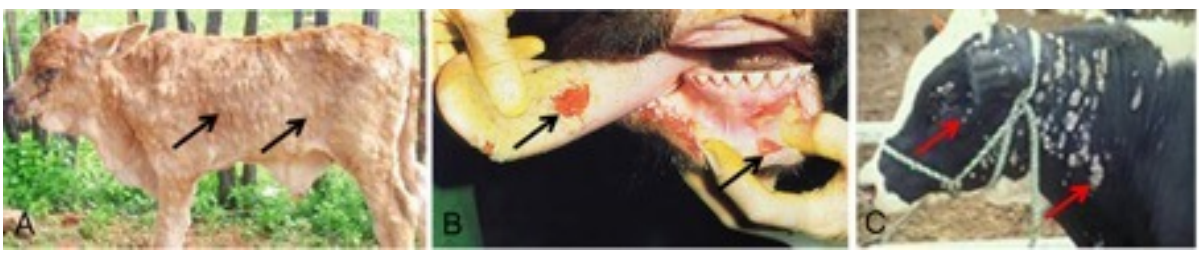


Figure 1: Cattle with nodules of LSD (A), wound on the tongue of FMD (B) and cauliflower-like growth of papillomatosis (C) on the external surface.

Building on this foundation, this study aims to develop a comprehensive algorithmic model combining both classical machine learning and deep learning for the early detection of common diseases with external manifestations. The system is structured into two distinct prediction pipelines:

- **Machine Learning (Symptom-Based) Prediction:** The classical supervised ML model is trained on clinical data to predict six specific diseases: Papillomatosis, Lumpy Skin Disease, Humpsore, Infectious Bovine Rhinotracheitis, Ringworm, and Lumpy Jaw Disease.
- **Deep Learning (YOLO-Based) Detection:** For the visual detection of diseases from field images, Computer Vision techniques like YOLO architectures are utilized to detect and classify distinct visual categories, including Infected Foot Image, Mouth Disease Infected, Normal Healthy Cow, Normal Mouth Image, and Lumpy Skin.

II. RELATED WORK

Machine learning has been widely applied in medical diagnosis, including both human dermatology and veterinary sciences. Foundational concepts of supervised learning algorithms and pattern recognition provide the essential probabilistic approaches and classification techniques necessary for modern diagnostic tools. Recent reports in journals such as *Nature Medicine* highlight numerous AI-based diagnostic studies that demonstrate significantly improved clinical accuracy using ensemble models. Ensemble models, such as Random Forest[2] and gradient boosting techniques like XGBoost[3], have consistently demonstrated superior performance in structured medical datasets due to their robust ability to capture nonlinear relationships among clinical symptoms. Additionally, Support Vector Machines (SVM)[4] have proven highly effective in high-dimensional classification problems, though they can be sensitive to feature scaling and dataset size.

In the specific domain of veterinary medicine, modern technology has enabled new applications to assist veterinarians in making better informed decisions about the health and disease of animals. Researchers are working on many algorithms that may be accustomed to sight different types of diseases. Varun Garg, et. al.[5] has provided a methodology based on machine learning that can detect cattle diseases which can provide a vital economical and medical solution for regions with scarce medical facilities for farm animals. This system facilitates the early detection of diseases to prevent costly delays in identification.

Furthermore, Wan and Bao[6] proposed a method where Support Vector Machine (SVM) is used in the animal disease diagnostics expert system. They have designed the model of animal disease diagnoses expert system which was used to diagnose the cow diseases. Expanding on text-based symptom classification, Lijing Niu, Chenhao Yang, et. al.[7], proposed a method which analyses the data of many electronic medical records, and used the SVM algorithm in machine learning to classify texts. They also used the data mining association algorithm to correlate the disease of the cattle according to the symptoms of the cattle and give corresponding diagnosis and treatment suggestions in time.

Developing novel and intelligent approaches to detect and predict animal diseases is now central to well-informed risk management plans and with respect to the use of medicines. However, there is still an essential need to develop smart models that can help technical experts in predicting health and disease related events prior and during manifestation of the disease signs and symptoms. The dual-modality approach presented in this study addresses this need by processing both structured symptoms and digital images simultaneously.

III. MATERIAL AND METHODS

This section outlines the developmental framework of the smart disease diagnostic model. The core of this research focuses on developing an algorithmic pipeline to process clinical data for accurate disease diagnosis. It details the system architecture, data acquisition, preprocessing steps, and the dual-pipeline approach utilizing both classical machine learning and deep learning algorithms.

Proposed System Architecture & Workflow

The proposed model consists of a data pre-processing module, a training module, and a predictor module for disease classification. The architecture is designed to handle multimodal inputs: users can input pictures of diseased animals or manually enter structured clinical symptoms.

After data ingestion, the pre-processing module standardizes the inputs (e.g., image enhancement, region of interest extraction, and feature scaling). The trained machine learning and deep convolutional neural network (DCNN) models then classify the disease and categorize its severity. If the result of the diagnosis is severe, the farmer is advised to contact veterinarians; otherwise, the system advises precautionary measures. Figure 2 shows the block diagram of overall workflow.

Dataset Preparation and Feature Engineering

To support the dual-input architecture, two distinct datasets were prepared.

Structured Clinical Dataset (Symptoms): Clinical signs and epidemiological data characteristics were mapped to six specific diseases: Papillomatosis, Lumpy Skin Disease (LSD), Humpsore, Infectious Bovine Rhinotracheitis (IBR), Ringworm, and Lumpy Jaw Disease.

- A synthetic dataset of 120 records was engineered based on 10 binary clinical features, including "Cauliflower Growth," "Smooth Lesion," "Nodular Lesion," "Lesion Rupture," "Location Head," "Location Neck," "Age < 2.5," "Cross-breed," "Muzzle Red," and "Maggot Formation". Table 1 represents a sample of symptom dataset.
- This dataset was split into training (70%), validation (15%), and testing (15%) subsets to prevent data leakage and ensure robust model evaluation. Figure 3 shows the data partitioning for the structural clinical dataset.

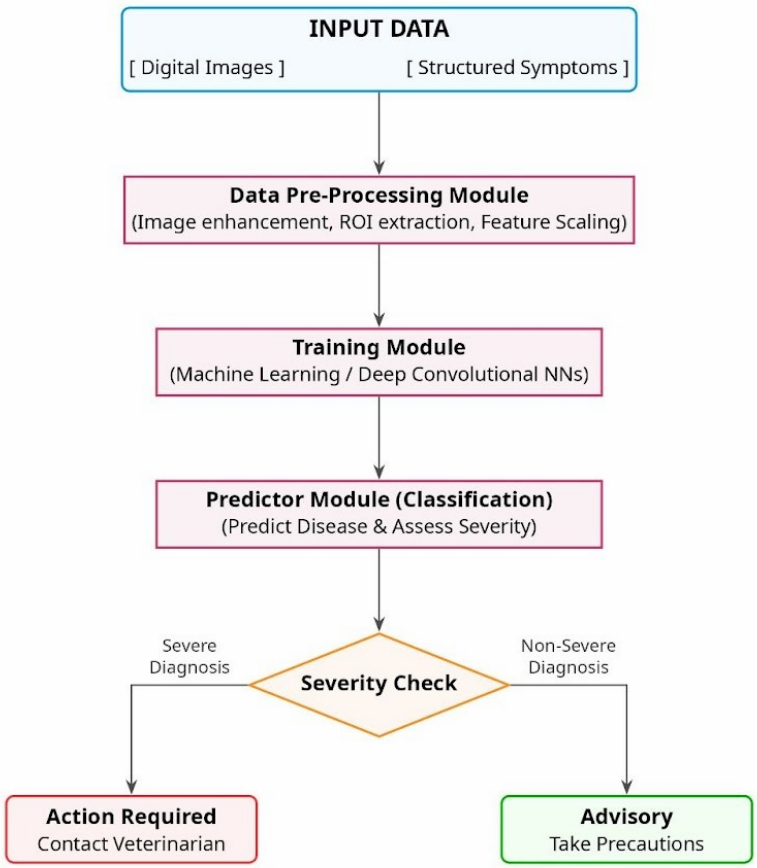


Figure 2: Block Diagram of the System Workflow.

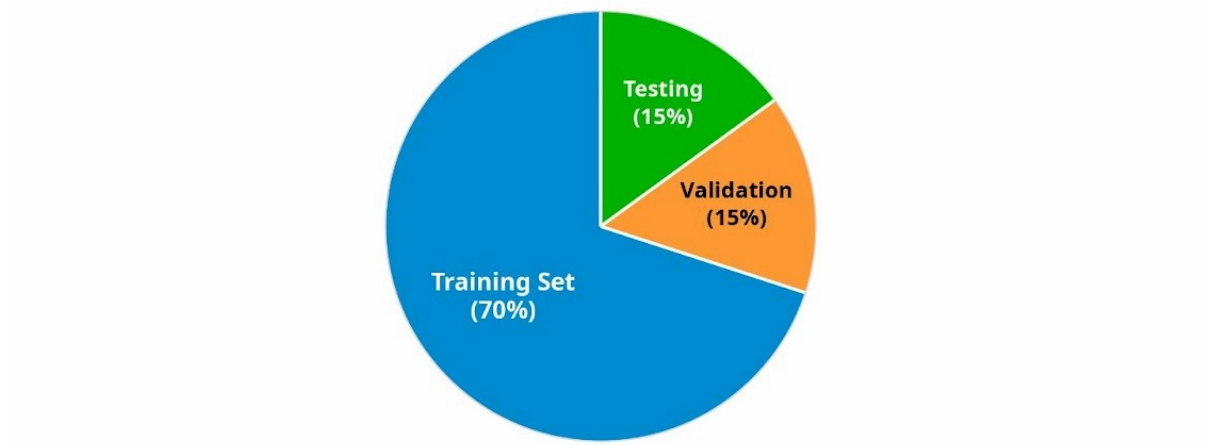


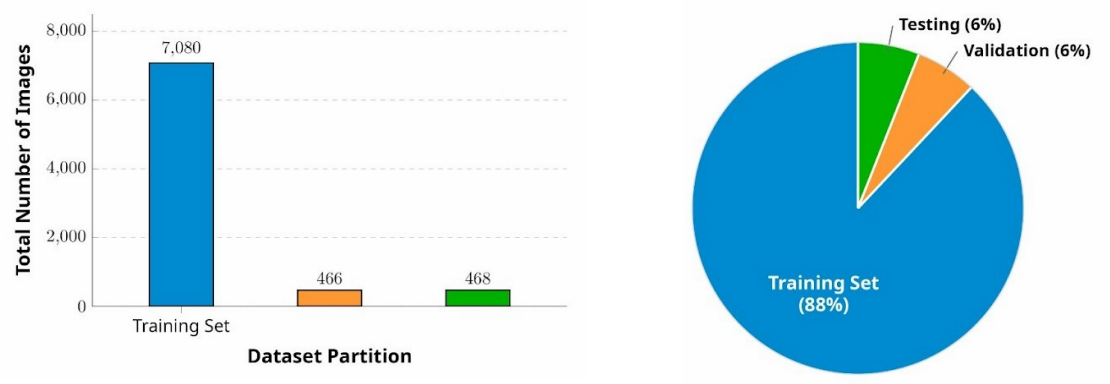
Figure 3: Data Partitioning for the Structured Clinical Dataset (120 Total Instances).

Table 1: Sample Encoded Symptom Dataset.

Label	Disease	Cauliflower Growth	Smooth Lesion	Nodular Lesion	Lesion Rupture	Location Head	Location Neck	Age < 2.5	Cross-breed	Muzzle Red	Maggot Formation
1	**(Papillomatosis)**	1	1	0	0	1	0	1	1	0	0
2	**(LSD)**	0	0	1	1	0	0	1	0	0	0
3	**(Humpsore)**	0	0	0	0	0	1	0	1	0	1
4	**(IBR)**	0	0	0	0	0	0	1	1	1	0

Unstructured Image Dataset (Visuals): For the visual detection of diseases, an open-source dataset titled "Cattle_disease Detection Dataset" was sourced from Roboflow Universe[8]. Figure 4 shows the image dataset distribution.

- The dataset comprises a total of 8,014 annotated field images.
- The image data was partitioned into a Train Set of 88% (7,080 Images), a Validation Set of 6% (466 Images), and a Test Set of 6% (468 Images).
- The data was labeled into five detection classes: Infected Foot Image, Mouth Disease Infected, Normal Healthy Cow, Normal Mouth Image, and Lumpy Skin.



(a) (b)

Figure 4: Image Data Distribution. (a) Total Number of Annotated Images per Partition (Total: 8014 Images), (b) Percentage Distribution of the Unstructured Image Dataset.

To optimize the dataset for the YOLO object detection architecture, a rigorous preprocessing and augmentation pipeline was applied by the dataset creator prior to model training:

- **Image Preprocessing:** All images were auto-oriented to correct original camera angles and resized (stretched) to a standard resolution of 640x640 pixels, which is the optimal input dimension for YOLO models to balance inference speed and spatial feature extraction. Furthermore, Auto-Adjust Contrast using Histogram Equalization was applied to standardize lighting variations across different field environments.
- **Data Augmentation:** To improve the model's robustness and prevent overfitting, the training data was artificially expanded by the creator generating 3 outputs per original training example (resulting in the final training set size of 7,080 images). Spatial and pixel-level transformations included horizontal and vertical flipping, random rotations between -26° and +26°, and the injection of up to 5% pixel noise. Additionally, Mosaic augmentation was applied—a critical technique for YOLO models that combines multiple images into a single view to improve the detection of smaller disease lesions in complex contexts.

Machine Learning Pipeline (Symptom-Based)

For the text-based predictor module, six classical supervised machine learning algorithms were selected to evaluate the structured clinical data. The chosen models were Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and Support Vector Machines (SVM). The workflow of machine learning models is presented in figure 5.

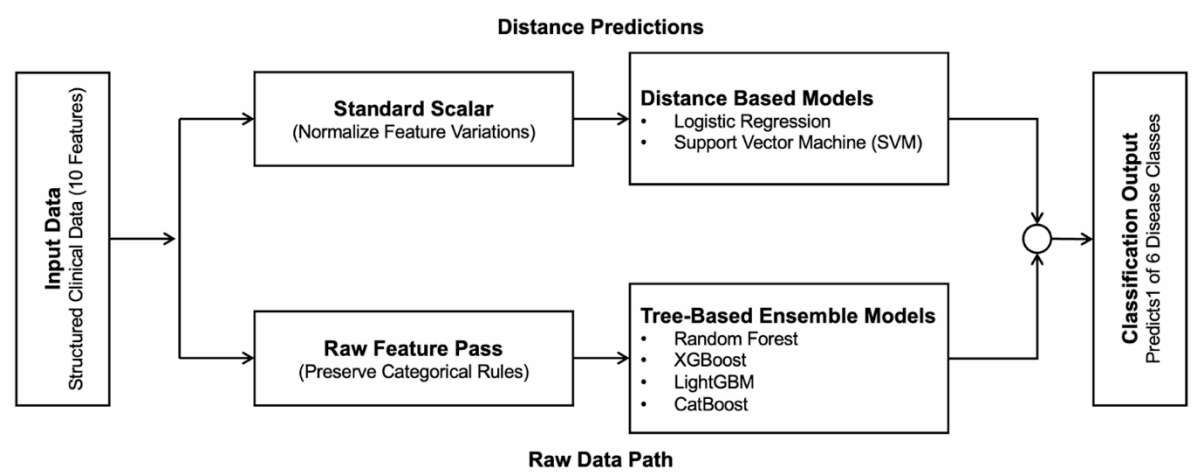


Figure 5: Workflow of Machine Learning Models.

To accommodate the mathematical requirements of distance-based models (such as SVM and Logistic Regression), a standard scaler was applied to normalize the feature variations. Tree-based ensemble models (like Random Forest, XGBoost, and CatBoost) utilized the unscaled data to preserve categorical rule interpretation. The models were iteratively trained and validated to optimize classification accuracy.

Deep Learning Pipeline (Image-Based)

To localize and classify surface lesions from digital images, Deep Convolutional Neural Networks (DCNN) were employed. Specifically, the state-of-the-art YOLO (You Only Look Once) object detection architectures—YOLOv11 and YOLOv12—were utilized for object detection and classification. The workflow of YOLO models is presented in figure 6.

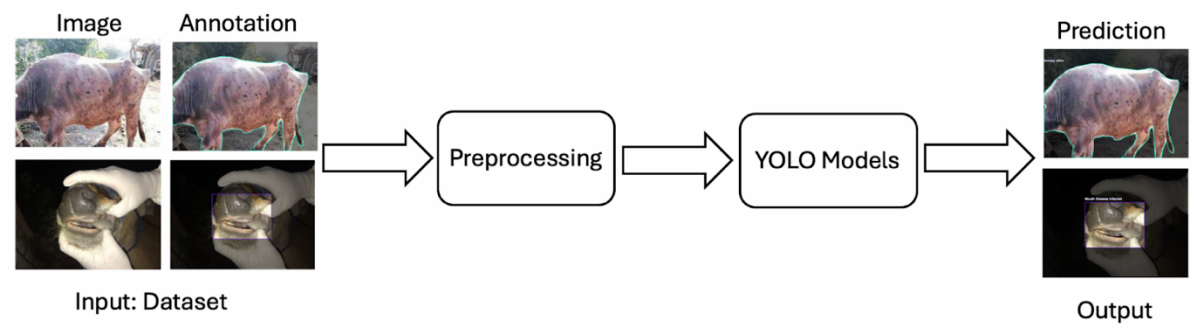


Figure 6: Workflow of YOLO Models.

The models were tasked with learning bounding box coordinates (Region of Interest) and class probabilities for the disease indicators. The training process involved multiple iterations to evaluate precision (P), recall (R), and Mean Average Precision (mAP50). YOLO models were selected due to their real-time processing capabilities, which align with the end goal of deploying the model as a mobile application for farm-level diagnostics.

IV. EXPERIMENTAL RESULTS

Machine Learning Model Performance on Clinical Data

The predictive performance of the machine learning pipeline was evaluated exclusively on the unseen testing dataset to ensure a robust assessment of each model's real-world generalizability. Table 1 presents the testing accuracy achieved by the six different algorithms evaluated in this study.

Table 2: Testing Accuracy of Machine Learning Models on Clinical Data

Model	Testing Accuracy
Logistic Regression	0.8333
Random Forest	0.8889
XGBoost	0.8333
LightGBM	0.7222
CatBoost	**1.0000**
SVM	0.7778

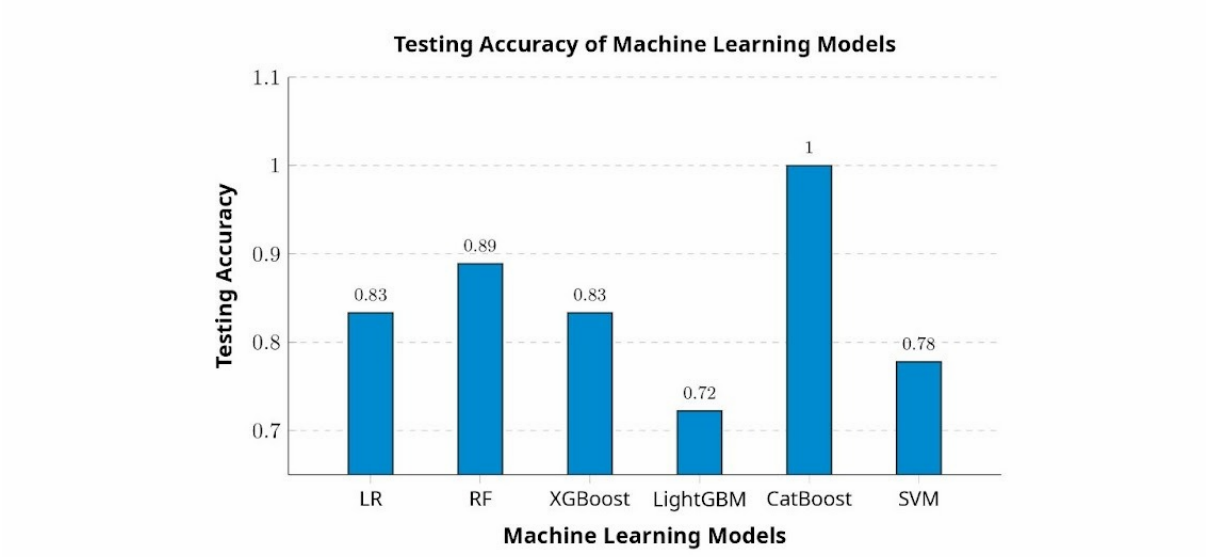


Figure 7: Comparison of the testing accuracies achieved by the six machine learning algorithms. (LR = Logistic Regression, RF = Random Forest).

The experimental results highlight significant variances in how well different algorithms generalized to the test data. **CatBoost** emerged as the superior model, achieving a flawless testing accuracy of 1.0000 (100%). This indicates that the CatBoost algorithm perfectly captured the underlying patterns in the clinical features and successfully applied those rules to completely unseen instances without any misclassifications.

Other tree-based ensemble models also demonstrated strong predictive capabilities. **Random Forest** achieved the second-highest testing accuracy at 0.8889, proving its robustness in handling the dataset's variance. **XGBoost** followed closely with a testing accuracy of 0.8333. Conversely, **LightGBM** proved to be the weakest

performer in this specific experiment, yielding the lowest testing accuracy at 0.7222. Because LightGBM is highly optimized for very large-scale datasets using leaf-wise tree growth, its lower performance here suggests its default splitting strategies may not be optimally suited for the specific size and distribution of this dataset.

Finally, the traditional classifiers, **Logistic Regression** and **Support Vector Machine (SVM)**, demonstrated moderate and stable performance. Interestingly, Logistic Regression achieved a testing accuracy of 0.8333, matching the performance of the more complex XGBoost model. This suggests that a significant portion of the dataset's underlying variance can be effectively separated using linear decision boundaries. SVM yielded a testing accuracy of 0.7778, establishing a solid baseline but falling short of the top ensemble methods.

Overall, CatBoost is the highly recommended algorithm for the symptom-based classification task, offering unmatched accuracy on unseen data.

YOLO Deep Learning Model Performance on Image Data

To evaluate the visual diagnostic pipeline, two iterations of the YOLO architecture (YOLOv11 and YOLOv12) were tested. The performance was measured using Precision (Box P), Recall (R), Mean Average Precision at 50% IoU (mAP50), and mAP at 50-95% IoU. Table 2 presents the aggregated testing results formatted as Mean ± Standard Deviation, detailing both the overall performance and the breakdown across all individual disease classes.

Table 3: Testing Performance of YOLO Models on Image Data (Mean ± SD)

Model	Precision	Recall	mAP50	mAP50-95
YOLOv11	0.581 ± 0.032	0.176 ± 0.013	0.378 ± 0.024	0.250 ± 0.020
YOLOv12	**0.779**	**0.160**	**0.471**	**0.270**

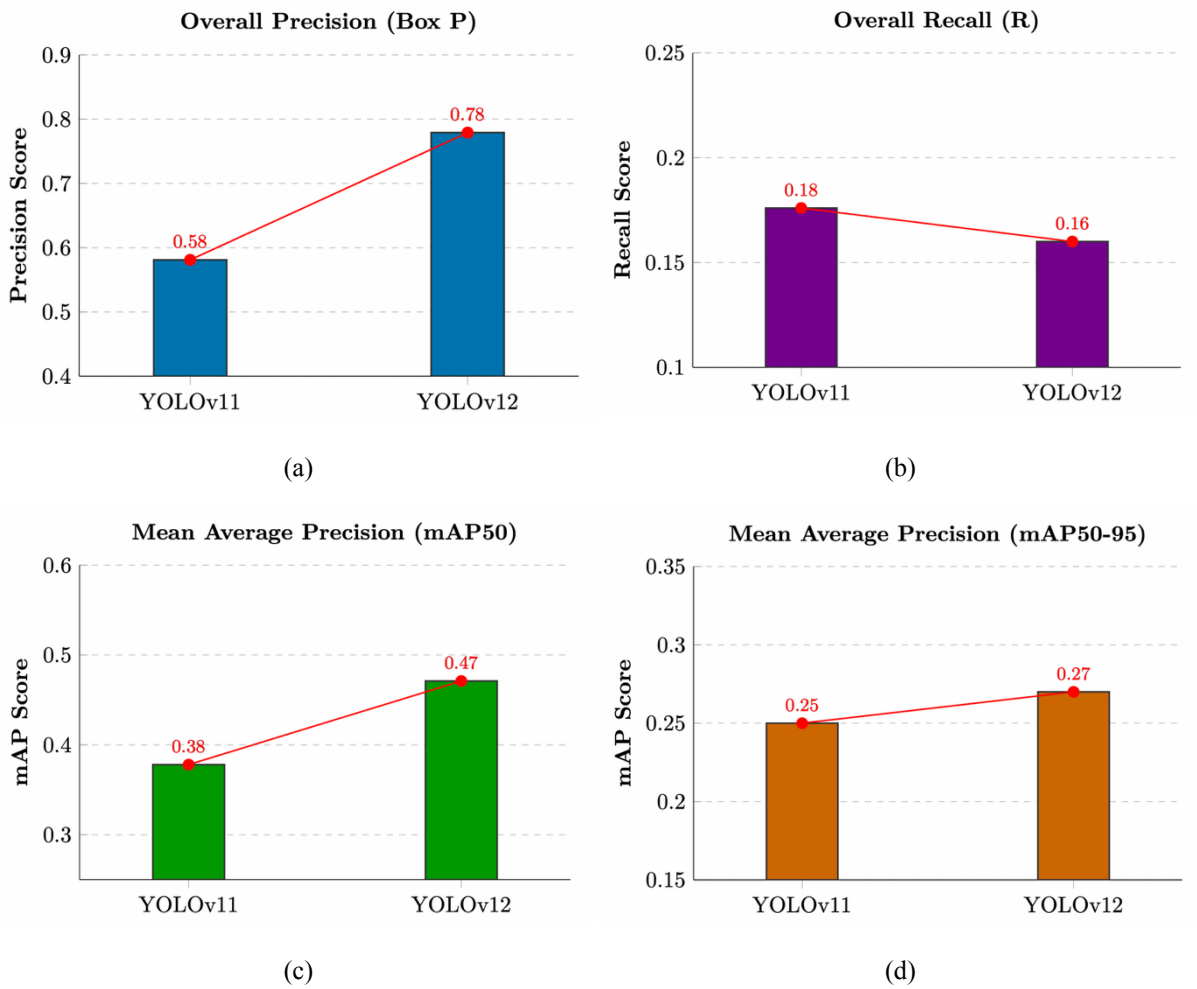


Figure 8: Performance metric comparisons between YOLOv11 and YOLOv12 on the testing dataset. (a)

Precision, (b) Recall, (c) mAP50, (d) mAP50-95. The red trend lines highlight the trajectory of model improvement.

The evaluation clearly establishes **YOLOv12** as the superior architecture for visual disease detection. Upgrading to YOLOv12 resulted in a substantial leap in overall **Precision**, rising from an average of 0.581 to 0.779. This indicates that YOLOv12 is significantly more reliable at ensuring that the bounding boxes it predicts actually contain a true disease instance, vastly reducing false positives.

Furthermore, the overall detection accuracy improved, with **mAP50** increasing from 0.378 to 0.471, and **mAP50-95** improving from 0.250 to 0.270. Interestingly, there was a minor trade-off observed in **Recall**, which dropped slightly from 0.176 to 0.160. This suggests that while YOLOv12 misses slightly more ground-truth instances than YOLOv11, the predictions it does make are vastly more accurate and confident.

Analyzing the class-wise performance reveals the distinct visual challenges of the dataset. The improvements in YOLOv12 were particularly beneficial for distinguishing subtle visual features between similar classes, proving its enhanced spatial feature extraction capabilities.

V. CONCLUSION

The results of this study successfully validate the implementation of a multi-modal diagnostic pipeline. For structured clinical data, **CatBoost** provided an optimal solution, yielding a perfect 1.0000 testing accuracy. It proved highly capable of discerning complex rules from tabular symptoms without succumbing to overfitting.

Concurrently, the image-based diagnostic pipeline demonstrated that **YOLOv12** effectively extracts complex visual features from field images, outperforming its predecessor by a wide margin in precision and mean average precision across the defined classes. By combining the perfect structured-data classification of CatBoost with the rigorous visual localization capabilities of YOLOv12, this dual-pipeline approach provides a highly robust, complementary, and reliable framework for comprehensive agricultural disease diagnostics.

REFERENCES

[1] Department of Livestock Services (DLS), "*Livestock Economy at a Glance (2019-2020)*," Ministry of Fisheries and Livestock, Govt. of the People's Republic of Bangladesh, 2020. [Online]. Available: dls.gov.bd.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324.

[3] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: https://doi.org/10.1007/BF00994018.

[5] V. Garg and K. Garg, "*Early Stage Disease Detection Platform in*," vol. 3, no. 11, pp. 13–15, 2016.

[6] L. Wan and W. Bao, "Animal Disease Diagnoses Expert System Based on SVM," *IFIP Advances in Information and Communication Technology*, pp. 539–545, 2010, doi: https://doi.org/10.1007/978-3-642-12220-0_78.

[7] L. Niu, C. Yang, Y. Du, L. Qin, and B. Li, "Cattle Disease Auxiliary Diagnosis and Treatment System Based on Data Analysis and Mining," *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, May 2020, doi: https://doi.org/10.1109/icccs49078.2020.9118590.

[8] M. rahaman Rohan, "Cattle_disease Detection Dataset," *Roboflow Universe*, 2023. [Online]. Available: https://universe.roboflow.com/mdzillur-rahaman-rohan/cattle_disease-detection.

Funding: This research is funded by "Faculty of Engineering, University of Rajshahi, Bangladesh" grant number A/5/52/R.U./Research/Engg-19/2023-2024