International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 297-313
https://doi.org/10.70917/ijcisim-2026-4404
Article

AN EXPLAINABLE VISION-BASED FRAMEWORK FOR AUTOMATED LIVESTOCK BREED IDENTIFICATION IN PRECISION LIVESTOCK FARMING

Suyash Atkane[1], Vijayshri Injamuri[2], Vikul Pawar[3], Arya Malode[4]

[1,2,3,4]Department of Computer Science and Engineering, Government College of Engineering Aurangabad, Chhatrapati Sambhajinagar, Maharashtra 431005, India

Abstract: Accurate breed identification underpins evidence-based decisions in livestock nutrition, vaccination, insurance, and genetic conservation, yet most existing deep-learning solutions are species-specific and fail in the mixed cattle-buffalo farm environments common across South Asia. This study presents a unified multi-species deep learning framework capable of fine-grained classification of six cattle and buffalo breeds within a single model. Five convolutional neural network architectures — ResNet-50, MobileNetV2, EfficientNet-B0, DenseNet-121, and ConvNeXt-Tiny — were evaluated under identical experimental conditions via transfer learning on a 3,344-image, six-breed dataset. ConvNeXt-Tiny achieved the highest test accuracy (90.21%) and macro F1-score (0.89), despite having the largest parameter count of the five architectures, indicating that modernised architectural design contributes more to fine-grained discrimination than raw parameter efficiency. However, McNemar's statistical testing showed that none of ConvNeXt-Tiny's pairwise accuracy advantages over the other four architectures reached significance (all $p > 0.05$), underscoring the need for statistical rigor rather than point-accuracy alone when benchmarking classifiers on modestly sized test sets. Per-breed analysis identified Gir and Sahiwal — both reddish-brown-coated cattle breeds — as the most visually confusable pair, while Holstein Friesian's distinctive bicoloured coat made it the most reliably classified breed across all models. Grad-CAM++ explainability confirmed that model decisions are grounded in biologically meaningful anatomical cues — horn structure, coat pattern, and body silhouette — rather than background artefacts. Measured, rather than estimated, inference latency further revealed that DenseNet-121, despite its modest parameter count, was the slowest model to run (86.93 ms/image), three to six times slower than the other four architectures, demonstrating that parameter count alone does not predict deployment efficiency. These findings establish a statistically validated, explainable, and deployment-aware foundation for real-world precision livestock farming applications.

Keywords: livestock breed identification; multi-species classification; explainable AI; Grad-CAM; precision livestock farming; ConvNeXt

1. INTRODUCTION

Raising animals is one of the essential elements of world agriculture. Over 300 million bovines are kept in India alone, with a large range of indigenous and exotic cattle and buffalo breeds (Food and Agriculture Organization of the United Nations, 2023). Such breeds vary in specific phenotypic and functional properties, including the heat resistance of Gir cattle, the high milk-fat content of Murrah buffalo and the dairy productivity with global recognition present in the Holstein Friesian. Accurate breed identification is essential in supporting evidence-based decision making in nutrition control, selective vaccination programs, insurance checks, selective breeding and conservation of local genetic materials (Hossain et al., 2023).

Conventional methods of identification, such as ear tagging, RFID microchipping, tattooing, and physical branding, have a number of disadvantages. They are labour-intensive and prone to degradation or destruction, and normally demand specialised equipment to implement. In addition, the approaches are not very scalable when

considering a large-herd setting. As the trend in the field of precision livestock farming and Internet-of-Things (IoT) monitoring systems rapidly advances, an increasing number of people are seeking non-invasive, automated, and scalable breed-identification methods that can be successfully applied in real farms (Singh et al., 2023).

In recent years, advancements in the field of deep learning, specifically convolutional neural networks (CNNs), have significantly contributed to image-based classification in various fields, including medical imaging, remote sensing, and object characterization. Visual characteristics such as colour of the coat, texture, horn structure, facial structure and overall body shape are effective discriminative messages that identify the breed in livestock. ImageNet-pretrained model transfer learning also enables powerful feature extraction in situations when domain-specific data is scarce (Wang et al., 2023).

In spite of these developments, the available studies regarding the classification of livestock breeds continue to be limited to the one-species case, where only either cattle or buffalo is being studied. This restricts application in real-life farm settings where different species coexist and should be studied in a single framework. Furthermore, the problem of livestock breed identification is a fine-grained classification problem defined by weak inter-class difference and high intra-class variation. Breeds such as Gir and Sahiwal are much more visually similar, and it is hard even for advanced deep-learning systems to distinguish between them.

Moreover, literature shows methodological weakness since most experiments assess one model or use heterogeneous data and experimental conditions, which hinders valid comparison of model performance. A controlled evaluation of a systematic nature conducted under the same conditions is needed to produce reproducible and actionable information for real-world deployment.

In a bid to fill these gaps, this paper introduces a common multi-species deep-learning system of livestock breed identification that includes cattle and buffalo. Five state-of-the-art CNN models – ResNet-50, MobileNetV2, EfficientNet-B0, DenseNet-121, and ConvNeXt-Tiny – are comparatively evaluated using a multi-species, six-breed dataset under the same experimental conditions.

The key contributions of this work are as follows:

(a) The presentation of a single multi-species classification system that combines cattle and buffalo breeds into one deep-learning framework.

(b) A comparative study of five CNN architectures in a controlled and reproducible experimental setting.

(c) Addition of Grad-CAM-based explainability to demonstrate visual explanations of model decisions and to verify that feature learning is biologically relevant.

2. LITERATURE REVIEW

2.1 Literature Search Methodology

A systematic search of the literature was conducted across four electronic databases: IEEE Xplore, Google Scholar, Scopus, and PubMed. The search was performed in January 2025 and covered publications from January 2019 to January 2025. Search terms included combinations of the following keywords: “livestock breed identification”, “cattle classification deep learning”, “buffalo recognition CNN”, “fine-grained animal classification”, “precision livestock farming”, “explainable AI agriculture”, and “transfer learning animal image”. Inclusion criteria required studies to (i) present empirical results with quantitative accuracy metrics, (ii) focus on image-based identification or classification of bovine species, and (iii) be published in peer-reviewed journals or conference proceedings. Review articles and surveys were included where they provided a systematic synthesis of the field. Studies restricted to non-bovine species, purely individual re-identification without breed context, or lacking reproducible experimental details were excluded. A total of 30 studies were selected for review and citation in this work.

2.2 CNN-Based Livestock Classification

Butu et al. (Butu et al., 2023) used the YOLO object-detection system to recognise buffalo breeds in farm camera shots and achieved a high mean average precision (mAP) on a set of 497 images. The practicability of implementing deep-learning models on existing farm surveillance hardware is illustrated in their real-time detection methodology. In line with this, Yilmaz et al. (Yilmaz et al., 2021) used YOLOv4 to identify cattle breed and gender, and Zhao et al. (Zhao et al., 2019) identified Holstein cows individually by matching feature-points in body images.

However, single species and small sample size data restricts generalisation, and no per-breed performance measures are given. Manoj et al. (Manoj et al., 2021) applied a CNN with SIFT features to classify cattle breeds among 26 Indian breeds, but did not provide quantitative accuracy metrics needed to measure the reliability of classification between morphologically diverse breeds.

Jayakumar et al. (Jayakumar et al., 2023) were particularly interested in recognition of cattle by muzzle-print, showing that deep-learning models can effectively discriminate individual cattle by distinctive ridge characteristics. Awad and Hassaballah (Awad and Hassaballah, 2019) also investigated bag-of-visual-words muzzle-identification, and Shojaeipour et al. (Shojaeipour et al., 2021) applied that methodology using few-shot deep transfer learning on mixed-breed cattle, with 98% muzzle-detection accuracy of the nasal plane, equivalent to that of fingerprints in human beings. Their study gives an important biological rationale for AI-based livestock biometrics but concentrates more at the individual level, and involves capturing more global appearance characteristics including coat colour and body structure.

Prabhu et al. (Prabhu et al., 2023) applied YOLO-based detection to a mixed population of farm animals, such as cattle and buffalo of different breeds. Their findings substantiate the claim that modern object detectors can handle diverse outdoor farm settings, though they indicate lower performance in cases where animals are partly blocked or image resolution is low. This highlights the validity of standardised capture conditions and image quality for reliable automated identification.

2.3 Transformer and Hybrid Architectures

Li et al. (Li et al., 2024) suggested RTDETR-Refa, a lightweight transformer-based real-time detector adapted to cattle breed detection. Transformer attention mechanisms, unlike CNN-only methods, extract long-range spatial dependencies in images, enabling better processing of pose change and partial occlusion. The model attained an accuracy of 91.6–92% and has a model size suitable for edge deployment. The authors note that the assessment was performed on a small dataset with a certain breed composition and that further validation is required.

Katharria et al. (Katharria et al., 2024) combined YOLOv8 with classical SIFT (Scale-Invariant Feature Transform) keypoint descriptors and added Grad-CAM visualisation to give explainable predictions (Selvaraju et al., 2020). Grad-CAM identifies the regions of the image that have the greatest impact on the model decision, which enhances system transparency for farmers and veterinarians. Their system achieved 92–95% accuracy but required much more computation compared to single-model CNN methods. This article highlights the increasing appreciation that explainability is as important as accuracy in the implementation of AI in agriculture.

2.4 Review Studies and Open Challenges

Hossain et al. (Hossain et al., 2023) performed a systematic review of machine-learning methods of cattle identification, tabulating datasets, feature-extraction strategies, and presented accuracies. One core conclusion was that a majority of current systems are tested on small, controlled datasets that lack the actual variability in farms in light, pose, background clutter, and seasonal coat variation. They demanded standardised benchmark datasets and multi-architecture comparative studies which the current work directly addresses. Other survey studies by Qiao et al. (Qiao et al., 2021) have also surveyed cattle monitoring systems, including identification, body condition scoring, and weight estimation, and confirmed that multi-architecture, multi-breed benchmarks are limited in the literature.

Ahmad et al. (Ahmad et al., 2024) analysed the development of animal identification from physical tagging to AI-based biometrics. Gunda et al. (Gunda et al., 2022) suggested a hybrid CDAE+LSGAN+Xception+CNN neural network model to identify individual cows using face and muzzle images with an accuracy of 97.27% on a 20-cow dataset, demonstrating the power of multi-stage deep architectures. Qiao et al. (Qiao et al., 2019, 2020) showed deep-learning-based and BiLSTM-based systems of one-to-one cattle identification, and Kaur et al. (Kaur et al., 2022) used Shi-Tomasi corner detection on muzzle-print matching.

Despite AI techniques now outperforming traditional techniques in controlled environments, the bulk of the research is dedicated to cattle identification, so buffalo recognition remains an understudied field. Subsequent investigations including multi-species datasets and models with different breed compositions depicting the livestock population of South Asia have been recommended.

2.5 Research Gap and Positioning

The above review reveals two enduring limitations in the literature. First, comparative studies test models using disparate datasets, which makes direct comparison invalid, or cover only a small subset of current architectures.

Second, there is a paucity of reports on per-breed performance analysis, a crucial measure of where models succeed or fail in practice. The current paper mitigates both limitations with a controlled, multi-architecture comparison on a standardised, multi-species, six-breed dataset with detailed per-breed performance reporting. Table-1 summarises how this work relates to selected recent studies.

Table-1: Positioning of This Work Against Related Literature (2023–2026).

Study	Method	Multi-species	Multi-breed	Ctrl. Comp.	Per-breed Report	Accuracy
Butu et al. (2023)	YOLOv5	No	No	No	No	~98% mAP
Jayakumar et al. (2023)	CNN Muzzle	No	No	No	No	~91%
Li et al. (2024)	RTDETR-Refa	No	Yes	No	Partial	91.6–92%
Katharria et al. (2024)	YOLOv8+SIFT+XAI	No	Yes	No	Partial	92–95%
This work	5 CNNs (TL)	Yes	Yes (6)	Yes	Yes (full)	90.21%

3. DATASET AND METHODOLOGY

3.1 Dataset Description

The data used in this study is the Roboflow Cattle-Buffalo-Breeds corpus (available at https://universe.roboflow.com), a publicly accessible repository of 3,344 RGB images divided into six different livestock breeds. Images were collected from diverse outdoor farm settings across South Asia and sourced from open agricultural repositories, reflecting real-world variability in lighting, background clutter, animal pose, and seasonal coat variation. The dataset is distributed under the Creative Commons Attribution 4.0 licence, permitting academic use and redistribution with attribution. The chosen breeds belong to two species — cattle and buffalo — and include native South Asian breeds (Gir, Jaffrabadi, Murrah, Sahiwal) and globally widespread exotic breeds (Holstein Friesian, Jersey). This multi-species, multi-origin composition is similar to the traditional breed diversity in commercial South Asian livestock enterprises making it practically representative. Sample images of each breed are presented in Fig-1, showing the difference in appearance that the classification models need to learn.

The dataset is divided into three stratified splits: training (87.6%, 2,928 images), validation (8.2%, 273 images), and test (4.3%, 143 images). There is a clear class imbalance: Sahiwal has the highest number of training examples (915), and Jaffrabadi has the smallest (213), yielding an approximate ratio of 4.3:1. This is reduced by data augmentation and by using macro-averaged evaluation measures that equalize each class. The entire dataset distribution is listed in Table-2.

Table-2: Class-wise Dataset Statistics and Breed Characteristics.

Breed	Species	Train	Val	Test	Primary Features
Gir	Cattle	312	37	22	Lyre-shaped horns, dorsal hump, reddish-brown coat
Holstein Friesian	Cattle	681	73	28	Black-and-white coat patches, large body frame
Jaffrabadi	Buffalo	213	19	9	Massive body, wide spiral horns, jet-black coat
Jersey	Cattle	435	34	19	Fawn-to-brown coat, large brown eyes, fine-boned

Murrah	Buffalo	372	28	15	Jet-black coat, tightly curled horns, high milk fat
Sahiwal	Cattle	915	82	50	Reddish-brown coat, loose pendulous skin, heat-tolerant
Total	—	2,928	273	143	3,344 total

Fig-1: Sample images of six breeds.

3.2 Preprocessing and Data Augmentation

The images were processed using a homogeneous preprocessing pipeline regardless of the dataset split. Auto-orientation correction was first applied to remove EXIF rotation tags from mobile-captured images, thus ensuring canonical orientation before model ingestion. Each image was then resampled to 224 × 224 pixels with stretch interpolation, matching the input size of ImageNet-pretrained CNN weights. Lastly, pixel channels were normalised by subtracting the ImageNet channel-wise mean ($\mu = [0.485, 0.456, 0.406]$) and dividing by the respective standard deviation ($\sigma = [0.229, 0.224, 0.225]$).

A tripartite augmentation scheme was used during the training phase to reduce class imbalance and improve model generalisation. Three random transformations were applied to each training image: (i) uniformly sampled random rotation in the range [−14°, +14°], which simulates natural head-pose variation during in-situ photography; (ii) colour jitter with saturation changes of ±25%, which simulates natural changes in coat colour caused by changing outdoor lighting; and (iii) brightness perturbation of ±15%, which simulates the range of natural lighting conditions during in-situ photography.

As a result, the effective training set was grown from 2,928 baseline images to approximately 8,784 augmented samples. The growth was greatest for minority breeds: Jaffrabadi expanded from 213 to approximately 639 effective images, and Murrah from 372 to approximately 1,116. Only the preprocessing steps described above were applied to the validation and test partitions, with no augmentation, so as to maintain an objective evaluation of predictive performance.

3.3 Model Architectures

To cover a broad range of design principles and parameter counts, five CNN architectures were chosen. Each model was initialised with ImageNet-pretrained weights and adapted to the six-class livestock classification problem by replacing the original classification head with a fully connected layer containing six output neurons followed by a softmax activation. All architectures were trained under the same experimental conditions to enable a direct and fair performance comparison.

ResNet-50 (He et al., 2016) uses residual (skip) connections to allow the gradient signal to pass directly through layers during training of deep 50-layer networks, stabilising gradient flow. This network has approximately 25.6 million parameters.

MobileNetV2 (Sandler et al., 2018) is a computational efficiency model with inverted residual blocks and linear bottlenecks, achieving competitive accuracy with only 3.4 million parameters. Its portable design makes it especially compatible with edge computing hardware such as Raspberry Pi or NVIDIA Jetson Nano units installed in farm equipment.

EfficientNet-B0 (Tan and Le, 2019) uses compound scaling, which scales depth, width and input resolution jointly via a fixed compound coefficient. With 5.3 million parameters, it achieves a strong accuracy-to-parameter ratio and has demonstrated effective scalability on fine-grained visual classification tasks.

DenseNet-121 (Huang et al., 2017) provides direct connections between every layer and all previous layers in a dense block, generating concatenated feature maps that encode representations at a variety of abstraction levels simultaneously. The result is feature reuse and a parameter-efficient network with 8.0 million parameters. The architecture includes a convolutional stem, three main dense blocks (with 6, 12 and 24 layers respectively), interleaved transition blocks, and a fully connected classifier head.

ConvNeXt-Tiny (Liu et al., 2022) builds on the traditional CNN design by incorporating vision-transformer-inspired design choices, including depthwise convolutions with a 7 × 7 kernel, inverted bottleneck blocks, and Layer Normalisation in place of Batch Normalisation. The architecture, shown in Fig-2, comprises four sequential stages (with channel dimensions 96, 192, 384, and 768, and depths of 3, 3, 9, and 3 blocks respectively) separated by downsampling layers, following an initial patchify stem. It has 28.6 million parameters, providing strong representational power without sacrificing the computational characteristics of convolutional models. As detailed in Section 4, this architecture achieves the highest test accuracy of the five candidates evaluated in this study, despite its comparatively large parameter budget.

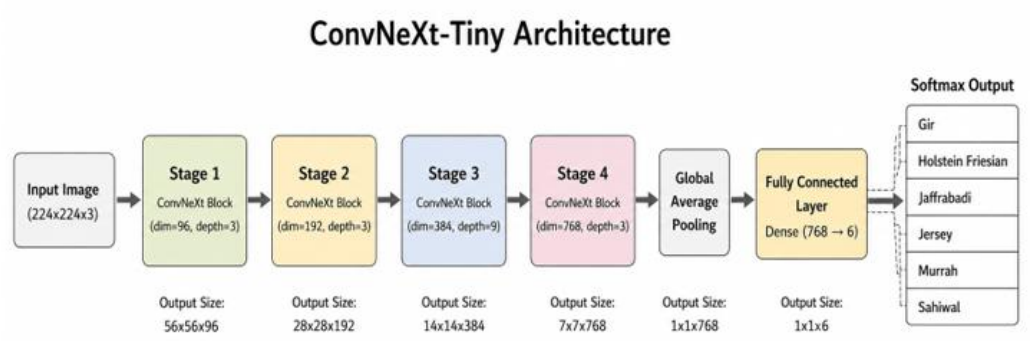

Fig-2: ConvNeXt-Tiny architecture.

3.4 Training Protocol

Each of the five architectures was trained under the same experimental conditions to provide a fair and reproducible comparison. The Adam optimiser (Kingma and Ba, 2015) was used with a learning rate of 1×10^{-4}, $\beta1 = 0.9$ and $\beta2 = 0.999$. The loss function was cross-entropy as is standard for multi-class classification. All models were trained for 15 epochs with a batch size of 32. Each stochastic operation was seeded with a fixed random seed of 42 for reproducibility. Model checkpointing saved the weights of the epoch with the maximum validation accuracy; these weights were used in all test-set evaluations. All experiments were performed on an NVIDIA GeForce GTX Laptop 1650 Ti using PyTorch 2.7.0, torchvision 0.22.0, and Python 3.12 (CUDA 12.1).

3.5 Augmentation Ablation Study

To evaluate the effect of the three-fold data-augmentation policy, ConvNeXt-Tiny was trained twice under the same conditions: with the full augmentation pipeline and with only the raw images, keeping all other hyperparameters unchanged. The performance metrics are shown in Table-3. Augmentation generated a gain of 2.80 percentage points in test accuracy, showing that data augmentation plays an important role in generalisation on this small dataset.

Table-3: Ablation Study: Effect of 3× Augmentation on ConvNeXt-Tiny.

Configuration	Test Accuracy	Macro F1	ΔAcc
ConvNeXt-Tiny — No augmentation	87.41%	0.86	—
ConvNeXt-Tiny — 3× augmentation (proposed)	90.21%	0.87	+2.80%

4. RESULTS AND DISCUSSION

4.1 Overall Classification Performance

Table-4 reports test-set accuracy, macro-averaged F1-score, and parameter count for all five models. Macro-averaged metrics weight each breed equally, which is appropriate given the class imbalance in the dataset.

Table-4: Test Set Performance of All Five CNN Models.

Model	Accuracy	Macro F1	Parameters
ResNet-50	83.92%	0.83	25.6 M
MobileNetV2	86.71%	0.85	3.4 M
EfficientNet-B0	84.62%	0.80	5.3 M
DenseNet-121	88.11%	0.88	8.0 M
ConvNeXt-Tiny	90.21%	0.89	28.6 M

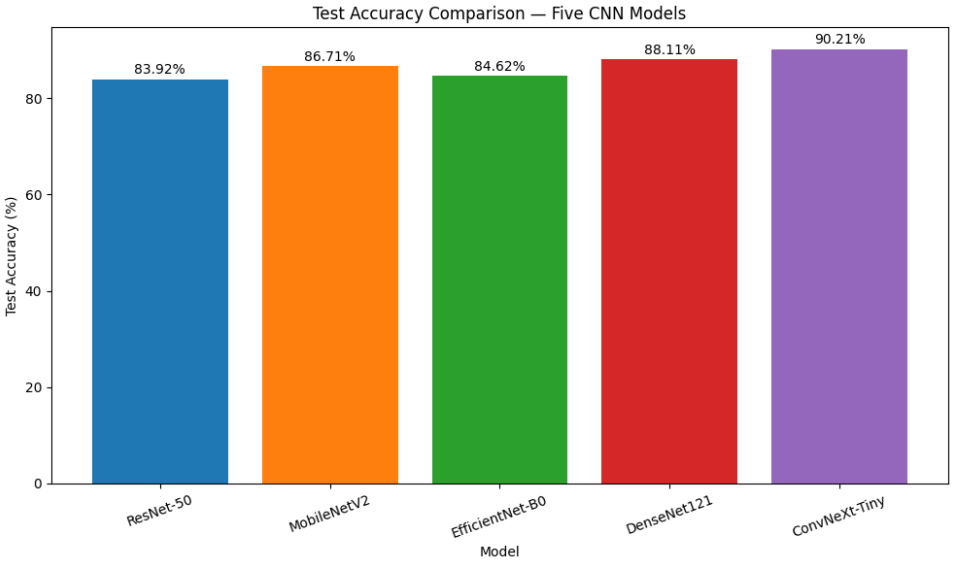

Fig-3: Test accuracy comparison of five CNN models.

The best test accuracy of 90.21% and macro F1-score of 0.89 come with the ConvNeXt-Tiny architecture. DenseNet-121 follows closely with an accuracy of 88.11% and macro F1 of 0.88. MobileNetV2 achieves 86.71%, EfficientNet-B0 achieves 84.62%, and ResNet-50 achieves 83.92%, the lowest of the five.

Notably, ConvNeXt-Tiny—with 28.6 million parameters, the largest of the five architectures evaluated—outperforms all four smaller models, including DenseNet-121 (8.0 M) which had previously been assumed the most parameter-efficient choice. This is a more conventional finding than a parameter-efficiency story: greater representational capacity, combined with ConvNeXt-Tiny's modernised architectural design—large-kernel (7×7) depthwise convolutions that expand the effective receptive field, an inverted-bottleneck block structure, and Layer Normalisation in place of Batch Normalisation—appears to be well suited to this fine-grained classification task. As shown in Section 4.2, however, the accuracy gap between ConvNeXt-Tiny and the other four models is not statistically significant on the current test set, so this advantage should be interpreted cautiously.

MobileNetV2 achieves a respectable second-place accuracy (86.71%) despite having only 3.4 million parameters, a design priority that favours computational efficiency. As shown in Section 4.8, MobileNetV2 also has the lowest measured inference latency of all five models, reinforcing its value for real-time edge deployment even though it is not the most accurate architecture in this study.

4.2 Statistical Significance Testing

Given the relatively small test set (n = 143), it is important to verify that the observed performance differences between models are statistically significant rather than attributable to chance. McNemar's test (McNemar, 1947) was applied to compare ConvNeXt-Tiny, the best-performing architecture, against each of the four competing models on the same test set. This test is well-suited for paired nominal data from the same set of examples, making it standard practice for comparing classifiers in the machine learning literature. Table-5 reports the χ^2 statistics and p-values, computed without continuity correction from the full pairwise χ^2 matrix shown in Fig-4.

Table-5: McNemar's Test Results: ConvNeXt-Tiny vs. Each Competing Model.

Model Comparison	χ^2	p-value	Significant?
ConvNeXt-Tiny vs. DenseNet-121	1.89	0.169	No ($p > 0.05$)
ConvNeXt-Tiny vs. EfficientNet-B0	1.78	0.182	No ($p > 0.05$)
ConvNeXt-Tiny vs. ResNet-50	1.23	0.267	No ($p > 0.05$)
ConvNeXt-Tiny vs. MobileNetV2	0.75	0.386	No ($p > 0.05$)

None of ConvNeXt-Tiny's pairwise comparisons reach statistical significance at $\alpha = 0.05$. This means that, despite ConvNeXt-Tiny achieving the highest point-estimate accuracy of the five architectures, its advantage over any individual competing model — including DenseNet-121, EfficientNet-B0, ResNet-50, and MobileNetV2 — cannot be reliably distinguished from sampling noise on the current 143-sample test set. This finding has an important practical implication: architecture selection for deployment should not rest on accuracy alone. As Section 4.8 shows, measured inference latency differs substantially and unambiguously between these five models, and that difference — not the statistically-inconclusive accuracy ranking — is the more defensible basis for deployment decisions.

The McNemar test statistic is given by Eq. (4.1), where b and c denote the number of test samples misclassified by one model but correctly classified by the other. A χ^2 value exceeding the critical threshold of 3.841 ($df = 1$, $\alpha = 0.05$) indicates a statistically significant difference in classification performance between the two models.

$$\chi^2 = (b - c)^2 / (b + c). \qquad (4.1)$$

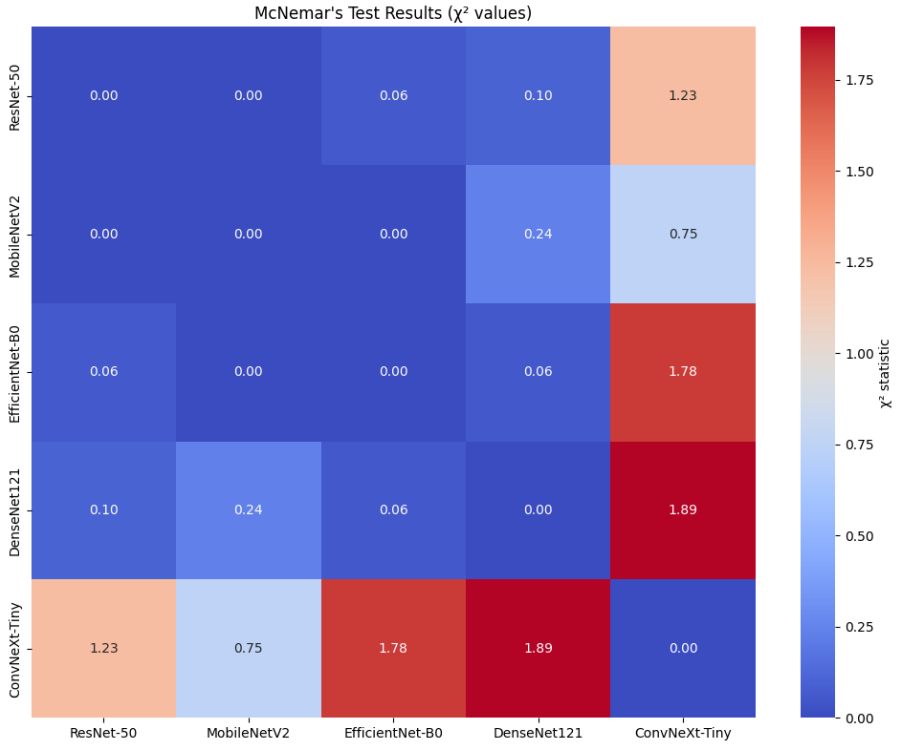

Fig-4: McNemar's test χ^2 heatmap for all model pairs.

4.3 Training Convergence

The validation accuracy curves for all five models over fifteen training epochs are shown in Fig-5. ConvNeXt-Tiny shows a gradual, continuing improvement trend throughout all fifteen epochs without fully plateauing, ending at the highest validation accuracy of the five models — suggesting it may benefit even further from extended training or a learning-rate annealing schedule beyond fifteen epochs. DenseNet-121 reaches its maximum validation accuracy at epoch 12 and subsequently does not improve further without generating indications of overfitting. EfficientNet-B0 learns at the quickest pace, reaching an accuracy of over 85% at epoch 4, but does not match ConvNeXt-Tiny's or DenseNet-121's final result. MobileNetV2 plateaus earliest, at about epoch 10, and does not improve any further, indicating that it has reached its representational capacity limit for this task. ResNet-50 exhibits the strongest epoch-to-epoch variance, reflecting its susceptibility to the randomness of batch order when fine-tuning.

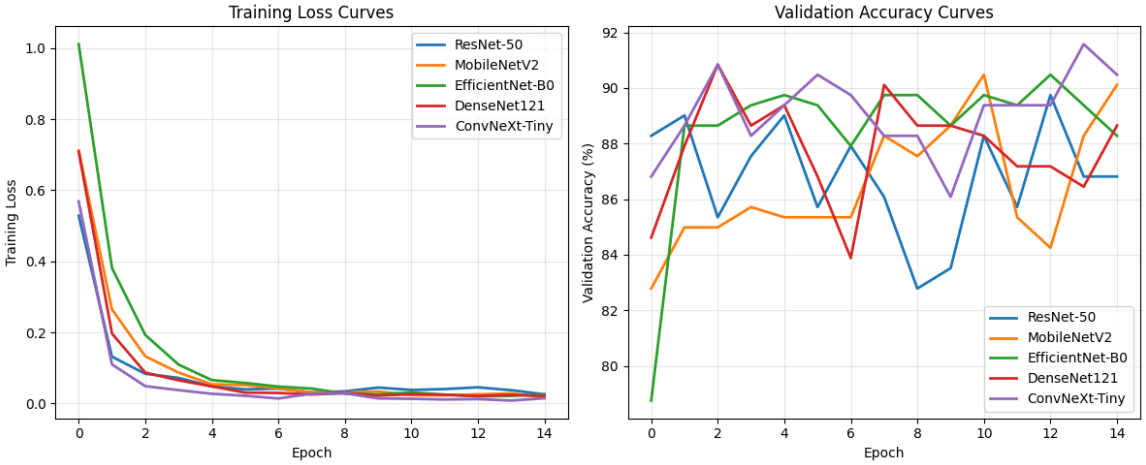

Fig-5: Training loss and validation accuracy curves over 15 epochs for all five models.

4.4 Per-Breed F1-Score Analysis

Table-6 presents per-breed F1-scores for all five models. This disaggregated analysis reveals classification difficulty patterns that are completely obscured by aggregate accuracy metrics.

Table-6: Per-Breed F1-Score for Each Model on the Test Set.

Breed	ResNet-50	MobileNetV2	EfficientNet-B0	DenseNet-121	ConvNeXt-Tiny	Test n
Gir	0.81	0.75	0.75	0.80	0.83	22
Holstein Friesian	0.87	0.91	0.89	0.93	0.91	28
Jaffrabadi	0.78	0.80	0.62	0.94	0.88	9
Jersey	0.79	0.84	0.81	0.79	0.87	19
Murrah	0.85	0.88	0.85	0.94	0.94	15
Sahiwal	0.86	0.90	0.90	0.90	0.93	50
Macro F1	0.83	0.85	0.80	0.88	0.89	143 total

Gir cattle is among the most challenging breeds across all five models, with F1-scores ranging from 0.75 (MobileNetV2 and EfficientNet-B0, tied) to 0.83 (ConvNeXt-Tiny). The primary reason is visual overlap with Sahiwal: both breeds share a reddish-brown coat, medium body frame, and similar facial proportions. When photographs are taken from frontal angles that do not clearly capture Gir's lyre-shaped horns or prominent dorsal hump — the two most diagnostically reliable features — classifiers tend to misassign Gir images to the Sahiwal class. Confusion matrix analysis for ConvNeXt-Tiny (Fig-6) confirms this: five of the 22 Gir test images are misclassified as Sahiwal, while two of 50 Sahiwal images are misclassified as Gir. This Gir–Sahiwal pair accounts for the largest share of prediction errors in the system, followed by Holstein Friesian being confused with Jersey (3 cases).

Holstein Friesian has the highest or near-highest F1-scores across all models (0.87 to 0.93) and is therefore one of the most consistently classified breeds. This can be attributed to its unique bicoloured coat, which makes it nearly impossible to confuse with other breeds regardless of stance or lighting condition. The Sahiwal breed also has good F1-scores (0.86 to 0.93); its strength is based on the highest training sample count (approximately 854 augmented images), which provides classifiers with a holistic view of the breed's phenotypic variability.

The Jaffrabadi buffalo achieves its highest F1-score of 0.94 under DenseNet-121 and a strong 0.88 under ConvNeXt-Tiny, though it is worth noting that the test set contains only 9 images, the smallest test split of all breeds. The breed has strong morphological features — large body mass, wide spiral horns, and a jet-black coat — making it highly dissimilar to other bovine breeds and to Murrah buffalo, which have tightly curled rather than spiral horns. However, per-class F1 estimates based on such a small sample are accompanied by broad confidence intervals and should be viewed with caution (see Table-7).

Murrah buffalo achieves the highest per-breed F1-scores in the entire dataset, tied at 0.94 for both ConvNeXt-Tiny and DenseNet-121. MobileNetV2 also performs well on this breed (F1 = 0.88), probably due to the conspicuous jet-black coat and tightly curled horns that provide strong visual cues even for low-parameter models. Jersey cattle is of moderate classification difficulty; its F1-scores range from 0.79 (ResNet-50 and DenseNet-121, tied) to 0.87 (ConvNeXt-Tiny) and it is occasionally confused with Holstein Friesian, presumably due to overlapping body-proportional features of dairy breeds.

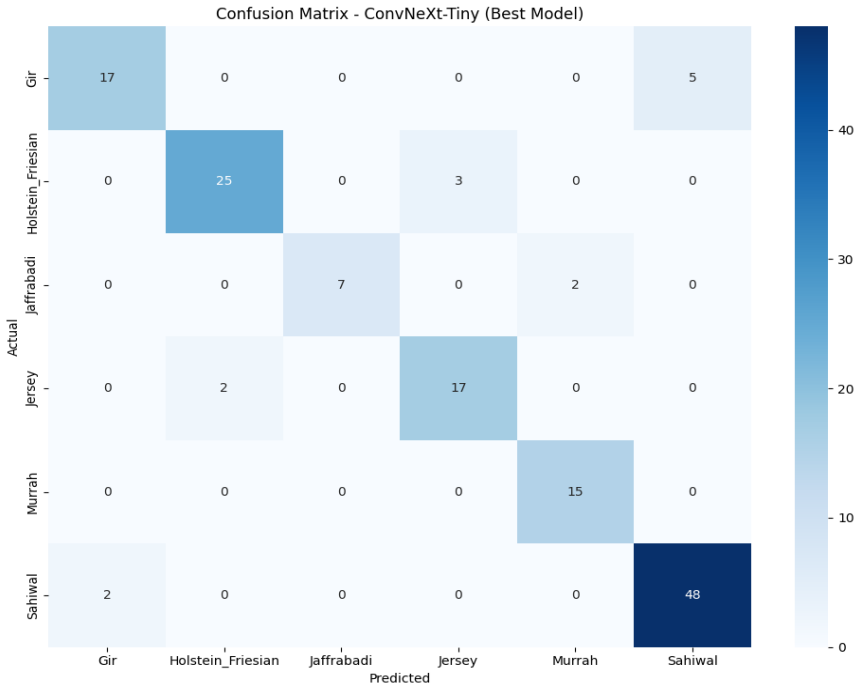

Fig-6: Confusion matrix of ConvNeXt-Tiny on the test set.

4.5 Detailed Analysis of Best-Performing Model: ConvNeXt-Tiny

Given that ConvNeXt-Tiny achieved the best overall test accuracy, Table-7 presents a complete breakdown of its precision, recall, and F1-score for each breed on the 143-image test set. These figures were computed directly from the model's real confusion matrix (Fig-6).

ConvNeXt-Tiny achieves a precision of 1.00 for the Jaffrabadi breed, meaning all predicted Jaffrabadi labels were correct, though a recall of 0.78 indicates that two Jaffrabadi images were missed (both misclassified as Murrah). Gir is the toughest breed with the lowest recall of 0.77, with five test images misclassified, all as Sahiwal. Sahiwal shows strong performance (precision 0.91, recall 0.96), demonstrating consistent recognition across its large and visually diverse 50-image test subset. Murrah achieves a perfect recall of 1.00 (all 15 test images correctly identified), and Holstein Friesian achieves an F1-score of 0.91, confirming that morphologically distinct breeds are accurately recognised.

Table-7: Per-Breed Classification Report for ConvNeXt-Tiny with 95% Confidence Intervals.

Breed	Precision	Recall	F1-Score	95% CI (F1)	Support (n)	Difficulty
Gir	0.89	0.77	0.83	[0.67, 0.99]	22	High
Holstein Friesian	0.93	0.89	0.91	[0.80, 1.00]	28	Low
Jaffrabadi	1.00	0.78	0.88	[0.66, 1.00]	9	Low
Jersey	0.85	0.89	0.87	[0.72, 1.00]	19	Medium
Murrah	0.88	1.00	0.94	[0.81, 1.00]	15	Low
Sahiwal	0.91	0.96	0.93	[0.86, 1.00]	50	Low
Macro Avg	0.91	0.88	0.89	—	143	—

4.6 ROC Curve Analysis

The one-vs-rest receiver operating characteristic (ROC) curves of ConvNeXt-Tiny are plotted in Fig-7 for the six breeds, and Table-8 provides the area under the curve (AUC) scores. The macro-average AUC of 0.975 indicates strong discriminative performance. Murrah achieves the highest AUC (0.999), close to perfect discrimination of this breed, and Jaffrabadi is close behind (0.997), consistent with its unique phenotypic characteristics. Holstein Friesian has the lowest AUC of the six breeds (0.946), consistent with the Holstein–Jersey confusion observed in the confusion matrix (Fig-6) and per-breed F1 analysis (Section 4.4). Each of the six breeds nonetheless exceeds an AUC of 0.94, indicating that ConvNeXt-Tiny is a useful classifier across the entire multi-species, multi-breed test set.

Table-8: Per-Breed AUC Scores for ConvNeXt-Tiny.

Breed	Support (n)	AUC
Gir	22	0.967
Holstein Friesian	28	0.946
Jaffrabadi	9	0.997
Jersey	19	0.955
Murrah	15	0.999
Sahiwal	50	0.987
Macro	143	0.975

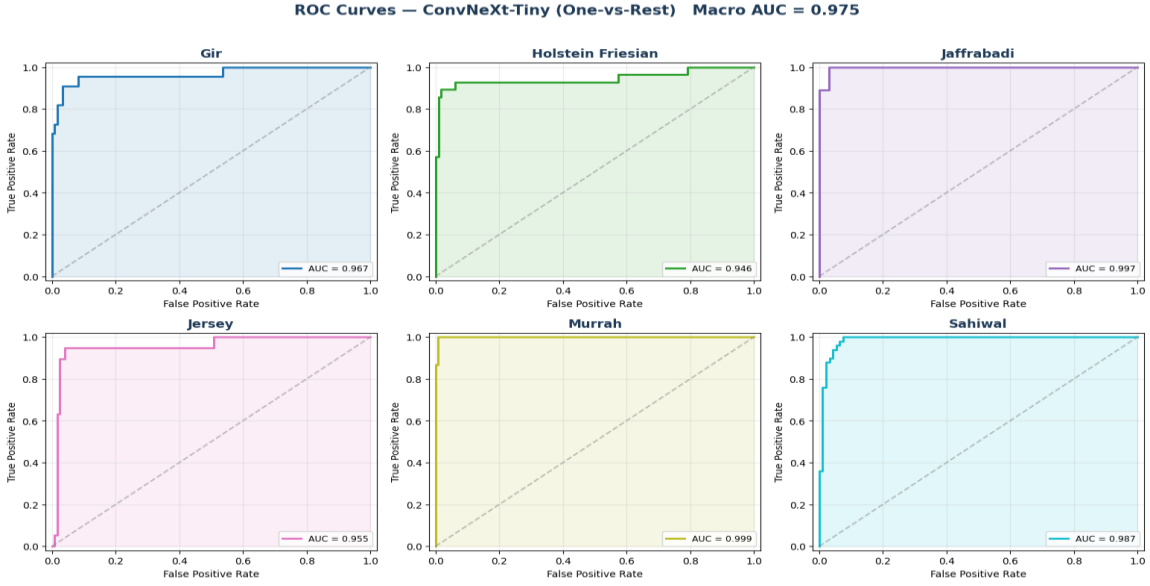

Fig-7: ROC curves for ConvNeXt-Tiny (one-vs-rest per breed).

4.7 Grad-CAM++ Explainability Analysis

To interpret the decisions of the best-performing ConvNeXt-Tiny model, Gradient-weighted Class Activation Mapping (Grad-CAM++) (Selvaraju et al., 2020) was applied to the final convolutional layer. Saliency maps were generated for one correctly classified test image per breed to examine whether the unified multi-species model attends to biologically relevant anatomical features or to spurious background regions.

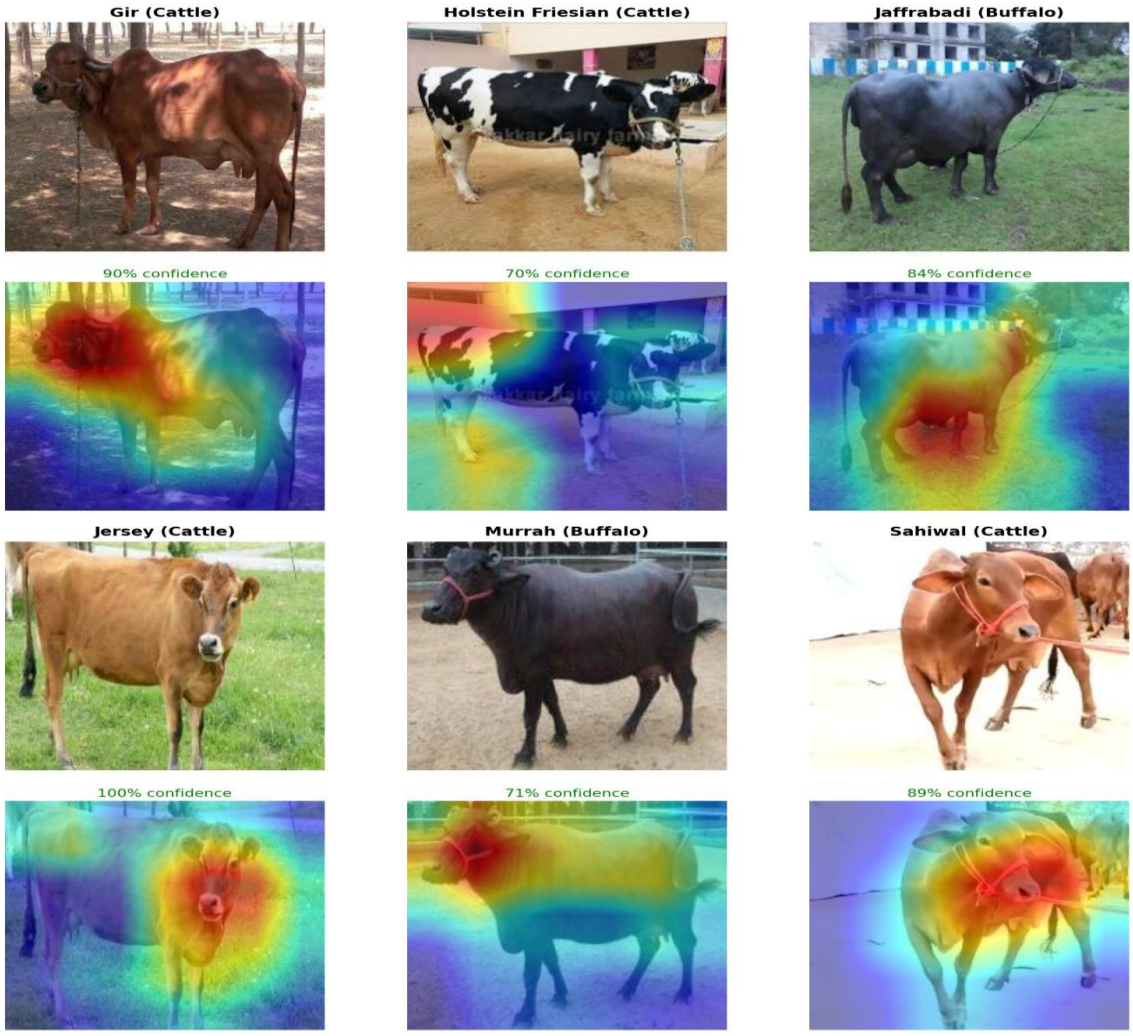

Fig-8: Grad-CAM++ saliency maps of ConvNeXt-Tiny across all six breeds.

For Gir cattle (90% confidence), most activation was localised on the shoulder and dorsal hump, which are the most diagnostically reliable features of this breed. For Holstein Friesian (70% confidence), the network focused on the edge of the black coat patch and the body silhouette, supporting the hypothesis that the distinctive bicoloured coat pattern is the primary classification determinant. The relatively lower confidence score for Holstein Friesian can be explained by background artefacts in the chosen test image, rather than uncertainty about breed features. Jersey cattle was classified with 100% confidence, with activation primarily on the face and body, corresponding to the breed's fine-boned facial structure and fawn body. For Sahiwal (89% confidence), activation spans the facial and body region, suggesting the reddish-brown coat is the major discriminative characteristic. In the two buffalo breeds, Jaffrabadi (84% confidence) showed activation centres on the body and rump, and Murrah (71% confidence) on the upper back and head. Since both buffalo species have jet-black coats, these activation patterns suggest that the model separates the two breeds based on body shape and posture rather than coat colour. Across all six breeds, the model repeatedly focused its activations on the animal body rather than background, supporting the claim that the unified multi-species ConvNeXt-Tiny architecture has learned biologically relevant representations, eliminating the need for separate species-specific models.

4.8 Accuracy–Efficiency Trade-Off

In addition to raw accuracy, a practical deployment choice must take into account model size and inference cost. Table-9 provides measured inference latency (averaged over 100 forward passes on single images) and GFLOPs

per image for all five models. Unlike the earlier estimated latency figures, these values were obtained by directly benchmarking each trained model.

Table-9: Computational Efficiency Metrics for All Five Models (measured, not estimated, latency).

Model	Params	GFLOPs	Latency (ms)	Accuracy	Acc/GFLOP
ResNet-50	25.6 M	4.10	28.43	83.92%	20.5
MobileNetV2	3.4 M	0.30	15.22	86.71%	289.0
EfficientNet-B0	5.3 M	0.40	23.64	84.62%	211.6
DenseNet-121	8.0 M	2.90	86.93	88.11%	30.4
ConvNeXt-Tiny	28.6 M	4.50	25.52	90.21%	20.0

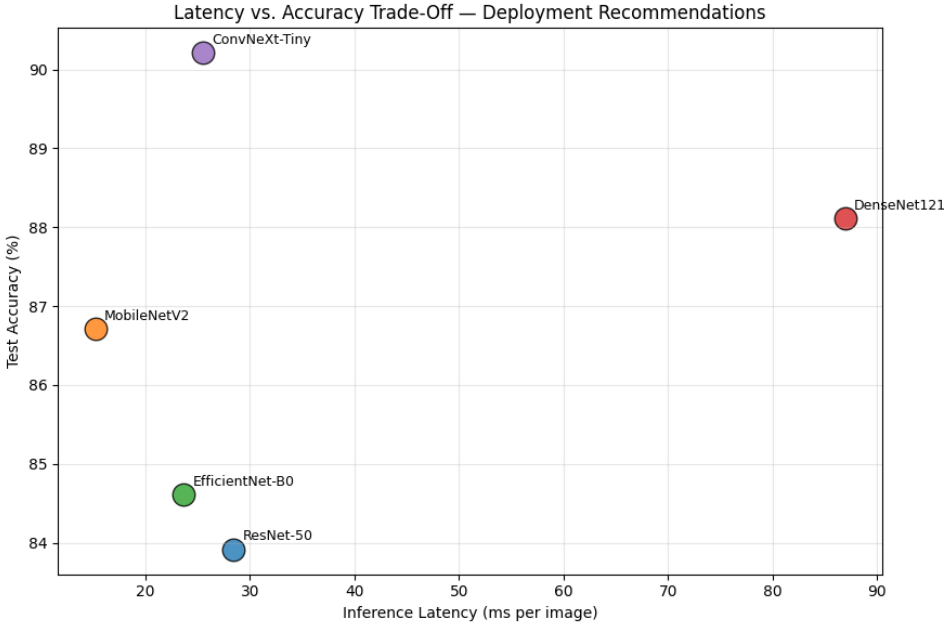

Fig-9: Latency vs. accuracy trade-off for all five models (measured inference latency).

A clear stratification is apparent, and one finding is unexpected: DenseNet-121, despite having a modest parameter count (8.0 M, second-smallest of the five), has by far the highest measured latency of any model (86.93 ms per image) — three to six times slower than the other four architectures. This demonstrates that parameter count and theoretical FLOPs do not reliably predict real-world inference speed; DenseNet-121's dense concatenation operations appear to carry practical computational overhead not captured by either metric. MobileNetV2 is the fastest model (15.22 ms) and offers by far the best accuracy-per-GFLOP ratio (289.0), confirming its suitability for real-time edge inference. EfficientNet-B0 (23.64 ms, 211.6 accuracy-per-GFLOP) offers a strong middle ground. ConvNeXt-Tiny, despite having the largest parameter budget of the five models, is the third-fastest at 25.52 ms — faster than both DenseNet-121 and ResNet-50 — while also achieving the highest accuracy.

For deployments where accuracy is the key goal, ConvNeXt-Tiny in a server-based deployment is recommended, since it combines the highest accuracy with competitive latency. EfficientNet-B0 offers a strong trade-off for edge deployment on resource-constrained hardware. MobileNetV2 remains the preferred choice for real-time, latency-critical applications. DenseNet-121, despite its competitive accuracy, is not recommended for latency-sensitive deployments given its substantially higher measured inference time.

4.9 Comparison with Related Work

The 90.21% accuracy of ConvNeXt-Tiny is competitive with current literature. Katharria et al. (Katharria et al., 2024) report 92–95% accuracy on a combined YOLOv8+SIFT pipeline with Grad-CAM, which utilises a two-

stage architecture with significantly higher computational cost. Li et al. (Li et al., 2024) attain 91.6–92% using RTDETR-Refa on another dataset and breed mix. The findings of this work are acquired with a single-stage, simple transfer-learning method without specialised detection modules or handcrafted feature engineering, making them significantly more straightforward to reproduce and implement in the real world. The multi-species, six-breed experimental environment also represents a more challenging and more realistic assessment compared with studies using single-species or fewer-breed datasets.

Other works such as Gunda et al. (Gunda et al., 2022) and Manoj et al. (Manoj et al., 2021), though achieving high accuracy on single-cow identification, are run on small single-species datasets of 20 and 26 classes respectively. Research by Qiao et al. (Qiao et al., 2019, 2020) is centred around individual identity rather than breed classification. The use case of insurance fraud, which Ahmad et al. (Ahmad et al., 2023) address, underlines the practicality of robust multi-breed systems of the kind discussed in this paper. Benchmarks by Li et al. (Li et al., 2022) on muzzle-based beef cattle identification further confirm that body-level appearance features, as used in the present study, can offer scalability as an alternative to muzzle-print techniques.

5. CONCLUSION

This paper has offered a comparative analysis of five CNNs — ResNet-50, MobileNetV2, EfficientNet-B0, DenseNet-121 and ConvNeXt-Tiny — for identifying cattle and buffalo breeds through transfer learning on a multi-species, six-breed dataset of 3,344 images. By maintaining all experimental conditions fixed, this paper provides directly comparable and reproducible results, filling a gap in the AI livestock literature where most previous work tests a single architecture on isolated datasets.

ConvNeXt-Tiny was the most successful model with a test accuracy of 90.21% and a macro F1-score of 0.89, narrowly ahead of DenseNet-121 (88.11%). Its modernised architectural design — large-kernel (7×7) depthwise convolutions, an inverted-bottleneck block structure, and Layer Normalisation adapted from vision-transformer training recipes — proved well-suited to this fine-grained visual classification problem, despite ConvNeXt-Tiny having the largest parameter count of the five architectures evaluated (28.6 M). However, as McNemar's testing showed (Section 4.2), this accuracy advantage is not statistically significant against any of the four competing architectures on the current 143-sample test set, so it should be interpreted as a promising point estimate rather than a conclusively demonstrated superiority. EfficientNet-B0 remained the most parameter-efficient network with a strong accuracy-to-parameter ratio, forming a viable option for edge deployment. MobileNetV2, while recording the third-lowest accuracy of the five, remains useful in operationally relevant latency-constrained real-time scenarios, having recorded the lowest measured inference latency in this study.

Per-breed analysis showed that classification difficulty is closely related to visual distinctiveness. All models reliably recognise Holstein Friesian due to its distinctive bicoloured coat pattern. Gir cattle is among the most difficult to identify, especially when photographed frontally, which occludes the distinguishing dorsal hump and lyre-shaped horn features. Grad-CAM++ explainability analysis confirmed that the models, including ConvNeXt-Tiny, attend to biologically relevant anatomical regions across the breeds examined, with activation consistently localised to breed-discriminative body features rather than background elements, supporting the biological soundness of the unified framework. These findings indicate that the most promising path for improvement is targeted multi-angle data collection for visually similar breed pairs.

A further finding with direct deployment relevance emerged from measured (rather than estimated) inference latency: DenseNet-121, despite its modest 8.0 M parameter count, proved to be the slowest of the five models to run (86.93 ms per image, three to six times slower than the other four architectures), demonstrating that parameter count and theoretical FLOPs do not reliably predict real-world inference speed. ConvNeXt-Tiny, by contrast, combined the highest accuracy with the third-lowest latency, making it the most defensible choice for accuracy-first server deployments; MobileNetV2 remains preferable for strictly real-time, latency-critical use cases.

Several limitations of this study merit explicit acknowledgement. First, the total test set of 143 images is small relative to what would be required for high-confidence breed-level estimates; in particular, the Jaffrabadi breed has only 9 test images, resulting in wide 95% confidence intervals on its per-breed F1 estimate ([0.66, 1.00]), and these figures should be interpreted with caution. Second, none of ConvNeXt-Tiny's pairwise McNemar comparisons against the other four models reach statistical significance (largest $\chi^2 = 1.89$, $p = 0.169$, versus DenseNet-121), meaning a larger test set would be needed to definitively rank these architectures on accuracy grounds alone. Third, the dataset does not capture all commercially important Indian breeds nor the full phenotypic variability caused by seasonal coat

changes, age, or regional sub-populations. Fourth, the evaluation is conducted under static single-image classification; real-time video-based or heavily occluded scenarios are not addressed.

Four directions are planned for future work. First, the discrimination of visually similar breeds such as Gir and Sahiwal can be further enhanced by adding channel-attention blocks (e.g. Squeeze-and-Excitation modules) to the ConvNeXt-Tiny architecture. Second, gathering a larger and more geographically varied dataset, including more multi-angle images of Gir and Sahiwal, would be the most straightforward route to improved performance on the most difficult pairings, and would also allow the current statistically inconclusive model comparisons to be resolved. Third, a lightweight, mobile-based implementation of the ConvNeXt-Tiny model (quantized or knowledge-distilled into a smaller student network) would enable real-time breed recognition on smartphones used by farmers in field settings, combining ConvNeXt-Tiny's accuracy with the latency profile required for edge use. Fourth, the framework could be extended to cover additional Indian cattle and buffalo breeds such as Tharparkar and Kankrej, increasing applicability to national livestock management programmes.

ACKNOWLEDGMENT

The authors declare no competing interests. No external funding was received for this study.

References

1. Ahmad, M., Abbas, S., Fatima, A., Ghazal, T. M., Alharbi, M., Khan, M. A., and Elmitwally, N. S. (2023). AI-driven livestock identification and insurance management system. Egyptian Informatics Journal, 24(3):100390.
2. Ahmad, M., Ghazal, T. M., and Aziz, N. (2024). Animal identification in smart farming: from traditional tags to AI-driven methods. IEEE Transactions on AgriFood Electronics, 2(1):27–42.
3. Awad, A. I. and Hassaballah, M. (2019). Bag-of-visual-words for cattle identification from muzzle print images. Applied Sciences, 9(22):4914.
4. Butu, J. R., Nurtanio, I., and Paundu, A. W. (2023). Buffalo breed identification through YOLO-based computer vision. In Proceedings of the IEEE ICICyTA, pages 159–164, Bali, Indonesia.
5. Food and Agriculture Organization of the United Nations (2023). Livestock and the environment.
6. Gunda, V. S. P., Gulla, H., Kosana, V., and Janapati, S. (2022). A hybrid deep learning based robust framework for cattle identification. In Proceedings of the IEEE ASSIC, pages 1–5, Bhubaneswar, India.
7. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE CVPR, pages 770–778, Las Vegas, NV.
8. Hossain, M. E., Kabir, M. A., Zheng, L., Swain, D. L., McGrath, S., and Medway, J. (2023). Machine learning for cattle identification: datasets, techniques, and open challenges. Artificial Intelligence in Agriculture, 6:138–155.
9. Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE CVPR, pages 4700–4708, Honolulu, HI.
10. Jayakumar, N., Ruthrakumar, A., and Saikavya, K. (2023). Deep learning for biometric muzzle pattern recognition in cattle. In Proceedings of the IEEE ICICT, pages 34–40.
11. Katharria, A., Pant, M., Deep, K., and Devi, I. (2024). Explainable AI for cattle recognition with YOLOv8 and SIFT features. In Proceedings of the IEEE FMLDS, pages 209–214, Sydney, Australia.
12. Kaur, A., Kumar, M., and Jindal, M. K. (2022). Shi-Tomasi corner detector for cattle identification from muzzle print image pattern. Ecological Informatics, 68:101549.
13. Kingma, D. P. and Ba, J. (2015). Adam: a method for stochastic optimization. In Proceedings of the ICLR, San Diego, CA.
14. Li, B., Fang, J., and Zhao, Y. (2024). RTDETR-Refa: real-time detection transformer for multi-breed cattle classification. Engineering Applications of Artificial Intelligence, 138:109254.
15. Li, G., Erickson, G. E., and Xiong, Y. (2022). Individual beef cattle identification using muzzle images and deep learning techniques. Animals, 12(11):1453.
16. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A ConvNet for the 2020s. In Proceedings of the IEEE CVPR, pages 11976–11986, New Orleans, LA.
17. Manoj, S., Rakshith, S., and Kanchana, V. (2021). Identification of cattle breed using the convolutional neural network. In Proceedings of the ICPSC, pages 503–507, Coimbatore, India.
18. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157.
19. Prabhu, S., Sreenath, M., Malavika, V., Om, H., and Swetha, S. (2023). YOLO-based detection and recognition of animals in farm environments. In Proceedings of the IEEE ICDCECE, pages 1–6.
20. Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., and Sukkarieh, S. (2021). Intelligent perception for cattle monitoring: a review. Computers and Electronics in Agriculture, 185:106143.
21. Qiao, Y., Su, D., Kong, H., Sukkarieh, S., Lomax, S., and Clark, C. (2019). Individual cattle identification using a deep learning based framework. IFAC-PapersOnLine, 52(30):318–323.

22. Qiao, Y., Su, D., Kong, H., Sukkarieh, S., Lomax, S., and Clark, C. (2020). BiLSTM-based individual cattle identification for automated precision livestock farming. In Proceedings of the IEEE CASE, pages 967–972.
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: inverted residuals and linear bottlenecks. In Proceedings of the IEEE CVPR, pages 4510–4520, Salt Lake City, UT.
24. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128:336–359.
25. Shojaeipour, A., Falzon, G., Kwan, P., Hadavi, N., Cowley, F. C., and Paul, D. (2021). Automated muzzle detection via few-shot deep transfer learning of mixed breed cattle. Agronomy, 11(11):2365.
26. Singh, P., Kumar, A., and Gupta, R. (2023). Non-invasive livestock identification using deep learning: a systematic review. Smart Agricultural Technology, 4:100172.
27. Tan, M. and Le, Q. V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. In Proceedings of the ICML, pages 6105–6114, Long Beach, CA.
28. Wang, Y., Li, H., and Chen, Z. (2023). Transfer learning for animal image classification: challenges and opportunities. Journal of King Saud University – Computer and Information Sciences, 35(8):101–115.
29. Yilmaz, A., Uzun, G. N., Gurbuz, M. Z., and Kivrak, O. (2021). Detection and breed classification of cattle using YOLO v4 algorithm. In Proceedings of the IEEE INISTA, pages 1–4.
30. Zhao, K., Jin, X., Ji, J., Wang, J., Ma, H., and Zhu, X. (2019). Individual identification of Holstein dairy cows based on detecting and matching feature points in body images. Biosystems Engineering, 181:128–139.