Explainable AI for Accurate Multi-Class Leather Defect Recognition Across Diverse Animal Sources

Submitted by
Shafiur Rahman
ID:182-35-2510
Department of Software Engineering
Daffodil International University

Supervised by
Mr. Khalid Been Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

This thesis paper has been submitted in fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering

Summer 2025

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Explainable AI for Accurate Multi-Class Leather Defect Recognition Across Diverse Animal Sources**", submitted by **Shafiur Rahman (ID:182-35-2510)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

This statement confirms that Shafiur Rahman completed this research while working under the direction of Khalid Been Badruzzaman Biplob, Lecturer (Senior Scale) in the department of software engineering at Daffodil International University. Additionally, it states that neither this project nor any component of it has been submitted to another institution for the award of a degree

Submitted By

Student Name : Shafiur Rahman

Student ID: 182-35-2510

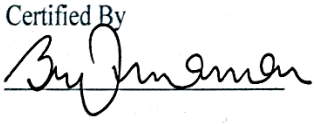
Batch: 26th

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Certified By



Khalid Been Badruzzaman Biplob

Lecturer (Senior Scale),

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Shafiur Rahman
Date of Birth	:	26 December 1998
Title	:	Explainable AI for Accurate Multi-Class Leather Defect Recognition Across Diverse Animal Sources
Academic Session	:	Summer 2025

I declare that this thesis is classified as:

☒	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☒	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

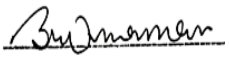
I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

182-35-2510
Student ID
Date: 12/08/2025



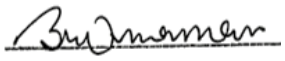
(Supervisor's Signature)

Mr. Khalid Been Badruzzaman Biplob
Name of Supervisor
Date: 12/08/2025

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

Full Name : Mr. Khalid Been Badruzzaman Biplob

Position : Senior Lecturer

Date : 12/08/2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Shafiur Rahman

ID Number : 182-35-2510

Date : 12 August 2025

Explainable AI for Accurate Multi-Class Leather Defect Recognition Across Diverse Animal Sources

SHAFIUR RAHMAN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Software Engineering)

DAFFODIL INTERNATIONAL UNIVERSITY

AUGUST 2025

ACKNOWLEDGEMENTS

I would like to thank my supervisor Mr. Khalid Been Badruzzaman Biplob for his valuable guidance, feedback, and support, which shaped the direction and quality of this research. Their mentorship improved my technical skills and problem-solving approach.

I appreciate the faculty and staff of the Department of Software Engineering at Daffodil International University for providing a supportive academic environment and necessary resources. I am also thankful to the tannery professionals in Dhaka for granting access to their facilities and offering practical insights into leather quality control, enhancing the study's relevance.

I acknowledge my fellow researchers and friends for their constructive discussions and collaboration, which made the research process rewarding. Special thanks to my family for their unwavering love and encouragement throughout this journey.

Finally, I recognize the open-source software communities for developing the tools and frameworks that supported the experimentation and deployment phases of this work. Their commitment to sharing knowledge has inspired m

DEDICATION

This work is dedicated to my parents, whose love and sacrifices made my education possible. I also dedicate it to my family for their unwavering support. To my teachers, mentors, and friends who inspired me, this achievement is as much yours as it is mine. Lastly, to all learners and researchers striving for meaningful change, may this work contribute to that vision.

ABSTRACT

Maintaining consistent leather quality is essential for enhancing product value, minimizing waste, and complying with international standards. Traditional inspection methods often depend on subjective expert judgment, resulting in inconsistencies and slow processing times. This study presents a deep learning approach for the automated detection and classification of leather defects in animals, focusing on four categories: cuts, folds, scratches, and normal leather. We created a balanced dataset from tanneries in Dhaka, which included 1,600 original images alongside an augmented set of 6,400 images. These images were captured using an iPhone 16 under various lighting conditions to accurately reflect real-world inspection scenarios. We fine-tuned five state-of-the-art architectures: Xception, InceptionResNetV2, LeViT, MaxViT, and MobileViT, evaluating their performance using metrics such as accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC). Data augmentation techniques, including rotation, flipping, and color jitter, significantly improved accuracy: an increase of 1.42% for cow leather, 2.15% for goat leather, 1.68% for sheep leather, and 2.21% for buffalo leather. Among the models, MaxViT showed the best performance after augmentation, while MobileViT achieved competitive accuracy with greater computational efficiency, making it ideal for resource-limited environments. To enhance model transparency, we incorporated explainability through GradCAM heatmaps, which allowed for defect localization. Finally, we developed a Flask-based web application for real-time defect classification, complete with visual support. The findings underscore that targeted data augmentation improves classification robustness, presenting an effective solution for leather quality control in the industry.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SBPWM	Simple Boost Pulse Width Modulation
ZSI	Z source inverter

CHAPTER 1

INTRODUCTION

1.1 Background

The leather industry plays a significant role in the global economy, providing raw materials for various sectors including fashion, automotive, upholstery, and industrial applications. Recent market reports indicate that the global leather goods market surpassed USD 400 billion in 2023 and is expected to continue growing, driven by rising consumer demand and expanding export markets [1]. Key players in this industry include countries such as Bangladesh, India, Italy, and China, where leather manufacturing and export serve as vital sources of revenue and employment [2], [3], [4]. Notably, Bangladesh's leather sector ranks among its top export earners, highlighting the importance of maintaining high-quality production standards to remain competitive in global trade.

Quality control in leather manufacturing is critical not only for meeting customer expectations but also for minimizing economic losses caused by defective products. Defects in leather, including cuts, folds, scratches, wrinkles, stains, insect bites, and grain damage, can significantly reduce material usability and lower the market value of products [5], [6]. These defects may arise from the quality of animal hides, improper processing, or mishandling during storage and transportation. Even minor defects, if not detected, can lead to significant waste during manufacturing, damage to brand reputation, and customer dissatisfaction [7].

Currently, defect detection in most tanneries and leather manufacturing facilities relies heavily on manual inspection, with human inspectors visually examining each hide. While human expertise has its value, this process is inherently subjective, labor-intensive, and susceptible to errors due to fatigue [8], [9]. Inconsistent lighting, variations in inspectors' skills, and the subtle nature of certain defects can cause variability in detection accuracy. Additionally, as production scales increase, the limitations of manual

inspection become more pronounced [10], often resulting in missed defects or misclassification.

Recent studies in computer vision and deep learning present an opportunity to transform the leather inspection process by providing consistent, rapid, and highly accurate defect detection. Automated systems powered by convolutional neural networks (CNNs) and vision transformers can analyze high-resolution images, identify defects across multiple categories, and reduce reliance on manual labor [11]. Moreover, the integration of explainable AI (XAI) allows manufacturers to understand the predictions made by these models, fostering trust in AI-assisted quality control systems and encouraging their adoption in industrial settings.

1.2 Problem Statement

The leather industry, despite its economic importance, faces ongoing challenges in defect detection. In many tanneries, human inspectors are still primarily responsible for evaluating each hide for subtle irregularities in texture, grain, and color [12]. While skilled professionals can often identify clear flaws, their assessments can be affected by factors such as fatigue, environmental lighting, and personal bias [13]. As production volumes increase, these limitations become more pronounced, leading to bottlenecks in quality control and, in some cases [14], allowing defective materials to enter the manufacturing process.

The inherent complexity of leather adds to the problem. Variations in hide origin, tanning methods, and surface treatments mean that a defect may look very different depending on the type of leather [15], [16]. For instance, a faint scratch on cowhide may be clearly visible under certain lighting conditions, while it could be nearly undetectable on softer goat leather. Furthermore, folds, cuts, and micro-abrasions can be disguised by surface patterns, dyes, and finishing techniques [17]. This diversity necessitates a detection method that is not only precise but also adaptable across multiple leather types and defect categories.

Existing automated inspection systems [18], [19], [20], though promising, often do not perform well under real-world tannery conditions. Many rely on handcrafted features or limited datasets that do not encompass the full range of defect variability [21]. The lack of large, diverse, and well-labeled datasets has hindered the development of robust and

generalizable models [22]. Additionally, the opaque nature of many AI-based solutions can make them difficult to interpret, leading to a lack of trust among industry stakeholders who need to justify quality decisions to clients and regulators.

1.3 Research Motivation

Leather is one of the most enduring and versatile materials in human history, yet its journey from raw hide to finished product is fraught with quality checkpoints. In today's competitive global market, consumer expectations are higher than ever, making the ability to guarantee flawless materials essential. A single defect overlooked during the early stages of production can have a cascading effect-wasting labor, increasing costs, and damaging the brand's reputation.

In recent years, industries worldwide have adopted Industry 4.0 principles, integrating automation, artificial intelligence, and data-driven decision-making into their operations [23]. Sectors like automotive manufacturing and semiconductor fabrication have already benefited from AI-powered quality inspection, achieving levels of speed and precision that are beyond human capability. However, the leather sector has been slower to embrace these technologies, partly due to the complex visual characteristics of leather surfaces and the absence of comprehensive datasets that represent real-world conditions [24], [25].

Meanwhile, advances in convolutional neural networks and vision transformers, have enabled remarkable capabilities in image classification and object detection. These models not only outperform traditional machine vision in accuracy but also adapt more readily to the inherent variability of natural materials [26]. When paired with explainable AI tools like GradCAM, these systems can offer visual explanations for their predictions, bridging the gap between algorithmic outputs and human understanding.

For Bangladesh, where leather and leather goods rank among the nation’s top exports, the stakes are particularly high. By implementing a robust, AI-driven defect detection framework, local tanneries could enhance consistency, reduce waste, and strengthen their competitive position in the global market. This study is motivated by the vision of integrating cutting-edge data-driven solutions into a traditional industry, ensuring that technological innovation contributes to both academic interests and tangible economic and industrial advancements.

1.4 Research Objectives

Building on the challenges and opportunities outlined above, this research aims to design, implement, and evaluate an AI-driven framework for multi-class leather defect detection that is both accurate and interpretable. The primary goal is to bridge the gap between advanced computer vision research and the practical realities of leather quality inspection in industrial settings. To achieve this goal, the study is guided by the following specific objectives:

- Develop a real-world dataset of leather images from multiple animal sources (cow, goat, sheep, buffalo), ensuring balanced representation across four categories: cut, fold, scratch, and normal.
- Apply targeted augmentation techniques-such as rotation, flipping, scaling, color jitter, and Gaussian noise-to enhance data diversity and improve model robustness under varying inspection conditions.
- Train and fine-tune a combination of convolutional neural networks (Xception, InceptionResNetv2) and vision transformer architectures (LeViT, MaxViT, MobileViT) on both augmented and non-augmented datasets, assessing performance using standard metrics including accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (MCC).
- Incorporate GradCAM-based visual explanation methods to highlight defect regions in model predictions, enabling transparency and interpretability for industrial adoption.
- Integrate the highest-performing model into a Flask-based web application capable of processing user-uploaded leather images, classifying defects, and generating corresponding GradCAM heatmaps in real time.
- Conduct a systematic performance comparison between CNN and transformer models, analyze the impact of data augmentation, and identify strengths, limitations, and potential avenues for further enhancement.

The research aims not only to deliver a high-performing classification system but also to establish a scalable, explainable, and industry-ready solution that aligns with modern manufacturing standards and Industry 4.0 principles.

1.5 Research Questions

This study is guided by the following research questions:

- RQ1: How do convolutional neural networks compare to vision transformer architectures in classifying multi-class leather defects across different animal leathers?
- RQ2: To what extent does data augmentation enhance model accuracy, robustness, and generalization in leather defect detection?
- RQ3: Can GradCAM effectively localize and visualize leather defect regions to improve interpretability for industrial adoption?
- RQ4: Is it feasible to integrate the highest-performing deep learning model into a lightweight, real-time Flask-based web application for tannery environments without compromising detection accuracy?
- RQ5: What performance trends, limitations, and trade-offs emerge from a comparative analysis of CNN and Transformer models in leather defect detection, and how can these inform future research?

1.6 Contributions

Key contributions of the study are as follows:

- Acquired 1,600 original leather images (400 per animal type: cow, goat, sheep, buffalo) across four defect categories (*cut*, *fold*, *scratch*, *normal*), captured under variable lighting using an iPhone 16. Applied targeted augmentation to expand the dataset to 6,400 images per animal, yielding a total of 25,600 images.
- Implemented and fine-tuned five state-of-the-art models (two convolutional neural networks and three vision transformers) from scratch with transfer learning initializations.

- Evaluated models using a multi-metric framework. Employed confusion matrices for class-specific error analysis and comparative performance assessment between augmented and non-augmented scenarios.

- Incorporated XAI-based visualizations were quantitatively and qualitatively assessed to validate model interpretability and support industrial trust in automated decisions.

- Developed a Flask-based inference pipeline integrating the highest-performing model with GradCAM processing. The system processes uploaded images in real time, outputs predicted defect class probabilities, and overlays defect localization heatmaps on original inputs. Designed for local server deployment with minimal latency.

- Performed comparative analyses revealing architecture-specific advantages. Highlighted trade-offs in computational cost, parameter efficiency, and accuracy, providing guidelines for selecting architectures in resource-constrained industrial settings.

1.7 Organization

The remainder of this thesis is organized into six chapters, each building upon the previous one to provide a cohesive overview of the research process, findings, and implications. Chapter 2 offers a comprehensive review of the literature related to automated leather defect detection. It starts with conventional inspection methods and moves forward to recent advancements in computer vision, deep learning, and explainable AI. This chapter also explores the capabilities of CNNs and vision transformers for handling texture-rich datasets, while identifying existing gaps that this study aims to address.

Chapter 3 details the process of dataset development, including the collection of leather images from various animal sources, the distribution of classes, and the environmental factors affecting image capture. It further explains the applied augmentation techniques, the preprocessing pipeline, and the rationale behind these choices. Chapter 4 outlines the experimental setup, including hardware and software configurations, model architectures, training protocols, hyperparameter ranges, and evaluation metrics. The

methodology for integrating Grad-CAM-based explainability into both CNN and Transformer models is also discussed in this chapter.

Chapter 5 presents the experimental results, organized into quantitative and qualitative analyses. It compares the performance of the selected architectures on both augmented and non-augmented datasets using various evaluation metrics, supported by confusion matrices, ROC-AUC curves, PR curves, and Grad-CAM visualizations. This chapter also offers a comparative discussion of the performance between CNNs and Transformers, as well as the impact of data augmentation on defect detection accuracy and generalization. Chapter 6 discusses the broader implications of the results, analyzing model trade-offs, industrial applicability, and the benefits of interpretability. Limitations of the current study are addressed, along with potential directions for future research. Chapter 7 concludes the thesis by summarizing the key findings, reiterating the main contributions, and highlighting the role of AI-driven defect detection in transforming leather quality control practices for Industry 4.0 manufacturing environments.

CHAPTER 2 Literature Review

2.1 Traditional Machine Learning and Feature Engineering Approaches

Before the widespread adoption of deep learning, early attempts at automating leather quality control relied heavily on handcrafted features and traditional machine learning classifiers. These approaches extracted texture, shape, or spectral descriptors from leather images and fed them into algorithms such as Support Vector Machines (SVMs), Random Forests, and k-Nearest Neighbors (KNN). While these methods offered notable improvements over manual inspection in terms of speed and repeatability, they often lacked the adaptability required to handle the wide variability of leather textures, defect types, and imaging conditions.

One such approach was introduced by Jawahar et al. [27], who developed the Fourier Angular Radial Partitioning (FARP) algorithm to capture both spatial and rotational invariance in defect features. By applying the magnitude of the Fourier Transform to sub-regions of leather images, the method could effectively represent defects such as cuts, scars, and pinholes. These descriptors, combined with Gray Level Co-occurrence Matrix (GLCM) features, were classified using a Random Forest ensemble, achieving an accuracy of 88.67% and a ROC AUC of 0.875. While the approach outperformed manual inspection, it depended heavily on high-quality imaging and was computationally complex.

Texture-based statistical analysis was also central to the work of Viana et al. [28], who classified bovine leather defects into four categories-tick marks, brand marks, cuts, and scabies-using features extracted from Interaction Maps and GLCMs. These were fed into an SVM optimized via Simulated Annealing, implemented through the LIBSVM framework. The optimized model achieved an impressive accuracy of 99.59%, outperforming SMO, AdaBoost, and Multi-Layer Perceptron (MLP) baselines. However, the method's reliance on handcrafted features and the relatively small dataset size limited its generalization capability.

In the domain of species classification, Varghese et al. [29] proposed a cost-effective solution combining portable digital microscopy with morphological feature extraction. Hair-pore segmentation was performed using the Circular Hough Transform, and species classification was carried out using a KNN classifier. The system achieved a 92.5% accuracy rate, surpassing MLP-based approaches. While innovative for its portability and low cost, the approach depended on the distinctiveness of pore patterns, which may vary under different processing or imaging conditions.

Machine learning has also been explored in production process optimization, as demonstrated by Baierle et al. [30], who applied multiple regression and ensemble models to predict raw material yield in the tanning industry. Using historical production data from a Brazilian tannery, the study evaluated eight machine learning models, with AdaBoost achieving the best performance (MAE: 0.042, RMSE: 0.057, R^2: 0.331). This work demonstrated the potential of machine learning beyond defect detection, but its success depended heavily on data quality and accurate batch-level tracking.

While these traditional approaches achieved strong results in controlled settings, they share several limitations. Their performance is tightly coupled to the design of handcrafted features, making them sensitive to variations in lighting, texture, and defect presentation. Additionally, they lack the ability to automatically learn hierarchical features from raw data-a strength that deep learning methods have since brought to the field. These limitations created the need for more adaptable, data-driven approaches capable of handling the complex visual variability inherent to leather inspection.

2.2 Deep Learning for Leather Defect Detection

The emergence of deep learning has transformed the field of leather defect detection by enabling models to learn complex, multi-scale representations directly from raw image data. Unlike traditional approaches that rely on handcrafted descriptors, deep neural networks automatically capture hierarchical features, making them more robust to variations in texture, lighting, and defect morphology.

One notable contribution is Chakrabarti et al. [31], who developed *LeatherNet*, a lightweight CNN tailored for defect detection and classification on leather surfaces. The model was trained on a custom dataset expanded from 384 to 12,000 images through data augmentation, achieving a training accuracy of 99.78% and testing accuracy of 97.42%.

Its shallow architecture minimized computational costs while maintaining high precision and recall, outperforming deeper models such as VGG16 and ResNet50. However, the model's reliance on a single, domain-specific dataset limited its cross-domain applicability.

Earlier, Liong et al. [32] investigated both Artificial Neural Networks (ANNs) and CNNs for defect classification using 1,897 leather images. Preprocessing steps included Canny edge detection and block partitioning for feature extraction. The ANN achieved 80.3% accuracy, while a modified AlexNet CNN attained 76% accuracy using a 1:3 defective-to-non-defective ratio. Although innovative in combining classical preprocessing with deep learning, the approach suffered from dataset imbalance and modest classification performance.

A more recent example is Deng et al. [33], who introduced ResNet50-2, an optimized residual network designed for ultra-high-definition (UHD) leather defect images. Using 3,000 images and a sliding patch window of size 536×536 pixels, the model classified five defect types-scratches, rotten surfaces, holes, needle eyes, and defect-free samples-achieving 94.6% accuracy. This performance surpassed Lenet5 and Caffenet, with improvements attributed to patch optimization and data augmentation. Nevertheless, the approach required substantial computational resources and relied solely on supervised learning.

Moving beyond classification, deep learning has also been applied to localize and segment defects. Banduka et al. [34] employed YOLOv11 for dual-side defect detection in leather, targeting grain-side and flesh-side defects such as grubs and suckouts. Using 1,200 images, the model achieved higher accuracy on the flesh side, with detection rates of 93.5% for grubs and 91.8% for suckouts. Innovations included a controlled digitization chamber to ensure consistent imaging conditions. However, the approach struggled with grain-side defects and required specialized imaging setups.

Liong et al. [35] explored automated tick bite defect segmentation using Mask R-CNN integrated with a robotic arm for image acquisition. On a dataset of 584 images, the model achieved 91.5% training accuracy but dropped to 70.35% testing accuracy, indicating potential overfitting. Key innovations included boundary optimization via modified Local Binary Patterns and automated defect marking with precise coordinates.

Limitations included its narrow focus on a single defect type and reduced performance in real-world conditions.

Across these CNN-based classification and object detection studies, several trends emerge. First, augmentation techniques consistently improve performance, as seen in [31] and [33]. Second, lightweight models such as LeatherNet [31] can rival or outperform deeper architectures when designed for domain-specific textures. Third, specialized acquisition environments ([34]) can boost accuracy but limit scalability. Finally, the relative scarcity of large, diverse, multi-animal defect datasets restricts the generalizability of these systems-a gap this thesis aims to address by combining CNN and Transformer architectures with a balanced multi-animal dataset and explainable AI integration.

2.3 Leather Species Identification Studies

While defect detection has been the primary focus of many leather inspection studies, species identification plays a critical role in ensuring product authenticity, regulatory compliance, and accurate grading. Species classification is inherently challenging due to inter-species similarities and intra-species variability in pore patterns, surface textures, and grain structures. Several studies have leveraged convolutional neural networks (CNNs), transfer learning, generative adversarial networks (GANs), and segmentation architectures to address these challenges.

Varghese et al. [36] presented a CNN-based system for large-scale automated leather species identification, using a dataset of 7,600 images from buffalo, cow, goat, and sheep, containing both ideal and non-ideal cases. The authors compared five CNN architectures-ResNet50, MobileNet, DenseNet201, InceptionNetV3, and InceptionResNetV2-reporting that MobileNet achieved the highest accuracy (98.29%) on the full dataset. InceptionNetV3, however, demonstrated better generalization on a smaller subset of 1,200 images (94.07%). When the datasets were combined, the overall accuracy improved to 98.5%. While this work highlighted the potential of CNNs for species identification, it was constrained by dependence on specific imaging conditions.

In another study, authors [37] introduced a novel preprocessing pipeline using Generative Adversarial Networks (GANs) to enhance microscopic leather images before classification. The dataset consisted of 1,200 images (300 per species), and features were

extracted using AlexNet, VGG16, GoogLeNet, and ResNet18, followed by Support Vector Machine (SVM) classification. The GAN-based preprocessing significantly improved pore pixel quality, and ResNet18 achieved the highest accuracy of 99.58%. Despite its strong results, the approach relied on small datasets and a specialized preprocessing stage, which may limit scalability.

Further exploring transfer learning, authors [38] fine-tuned four CNNs-AlexNet, VGG16, GoogLeNet, and ResNet18-for microscopic species identification. Using the same 1,200-image dataset, ResNet18 achieved the best performance with 99.69% accuracy, outperforming traditional machine learning methods by 7%. The study's novelty lay in its handling of texture similarity challenges, but again, dataset size was a limiting factor.

Beyond classification, Varghese et al. [39] proposed ApSnet, a modified Unet architecture for pore segmentation, designed to improve species prediction. By simplifying the Unet structure and introducing a weighted loss function, ApSnet achieved superior segmentation performance compared to Unet, Unet++, and I-Unet. When combined with KNN classification, it reached 97.4% accuracy on species identification tasks. While computationally efficient, the model depended heavily on the availability of pixel-level ground truth data, which can be labor-intensive to obtain.

Finally, authors [29] explored a low-cost, portable imaging approach using a digital microscope for hair-pore segmentation and morphological feature extraction. The features were classified using a KNN model, achieving 92.5% accuracy. This work demonstrated the feasibility of species identification in resource-limited settings but was sensitive to variations in pore visibility and processing conditions.

Collectively, these studies underscore the importance of species classification in leather quality control, particularly for multi-animal datasets. They also reveal persistent limitations-small datasets, specialized imaging requirements, and domain-specific preprocessing-that hinder generalization. These challenges directly motivate the approach in this thesis, which integrates species diversity into defect detection, enabling a unified, multi-class, multi-animal inspection system.

2.4 Specialized Imaging and Sensing Techniques

Advances in imaging and sensing technologies have opened new possibilities for leather inspection, enabling the capture of finer surface details, richer spectral information, and more consistent image quality. These methods aim to address limitations of standard RGB imaging, particularly in detecting subtle or hidden defects, differentiating species, and ensuring consistent environmental conditions during data acquisition.

Chen et al. [40] introduced the Hyperspectral Leather Defect Detection Algorithm (HLDDA) to identify defects in wet-blue leather using hyperspectral imaging. The dataset comprised 20 hyperspectral samples containing five types of defects, analyzed with Hyperspectral Target Detection (HTD) integrated into deep learning models-1D-CNN, 2D-Unet, and 3D-Unet. The 3D-Unet achieved the best overall performance, with all defect types exceeding 96% accuracy. The pixel-level spectral–spatial analysis provided exceptional defect localization capabilities. However, the system's high computational demands and difficulty in detecting smaller defects limited its scalability for real-time industrial use.

Portable microscopy was leveraged in [29] for species classification, where a digital microscope captured detailed pore structures for morphological feature extraction. While cost-effective and adaptable to on-site inspections, the approach was sensitive to variations in pore visibility and image capture conditions, affecting consistency across different environments.

Imaging environment control was a focus in Banduka et al. [34], where a controlled digitization chamber was developed to ensure consistent lighting and positioning for YOLOv11-based dual-side defect detection. This setup reduced environmental variability and improved detection rates for flesh-side defects (93.5% for grubs, 91.8% for suckouts), though performance on the grain side remained lower. The cost and complexity of such controlled setups may limit their adoption in smaller tanneries.

Outside the traditional optical spectrum, Sánchez et al. [41] demonstrated the potential of non-visual sensing for defect detection by applying an Electronic Nose (E-nose) system to detect and classify quality defects in Spanish-style green table olives. While not directly targeting leather, the E-nose-combined with chemometric analysis-achieved R^2 values between 0.74 and 0.86 for predicting sensory ratings. This work illustrates how

non-destructive, low-cost sensing methods can complement visual inspection in quality control contexts.

At the molecular level, Maidment et al. [42] investigated the biochemical origins of leather looseness, a defect affecting durability and texture. Using proteomic analysis of cattle hides, the study identified reduced concentrations of collagen Type I, decorin, and proteoglycan in defect-prone regions, correlating these with disorganized collagen fiber structures observed under confocal microscopy. While not a direct inspection method, such biomarker identification offers potential for early defect detection during pre-processing stages.

Collectively, these studies highlight the potential of specialized imaging and sensing techniques to enrich leather inspection workflows. Hyperspectral imaging excels in spectral–spatial analysis, microscopy provides fine-grained morphological detail, controlled chambers ensure acquisition consistency, non-visual sensing enables chemical-level defect detection, and proteomic analysis offers a molecular perspective. However, limitations-including high equipment costs, complex data processing, and narrow application scopes-underscore the need for solutions that balance accuracy, adaptability, and scalability in real-world tannery environments.

2.5 Related Industrial Applications in Quality Control

Although the majority of research in leather inspection focuses on defect detection or species classification, several related industrial applications provide transferable insights for developing robust, automated quality control systems. These studies span niche inspection tasks, aesthetic grading, and portable solutions, offering methodological innovations that can be adapted to leather manufacturing.

Pazzaglia et al. [43] addressed a specialized inspection challenge-classifying stitching colors on leather products-using the LASCC dataset, which contained 67 images representing two leather colors and seven stitching colors. Three deep CNNs-VGG16, ResNet50, and InceptionV3-were evaluated, achieving accuracies up to 99.9%. While the task differs from defect detection, the study demonstrated how deep learning can handle fine-grained color classification in textured materials. The key limitation was the extremely small dataset, which restricted generalizability and called for dataset expansion.

Moving beyond binary classification, Rosati et al. [44] developed a deep ordinal classification approach for aesthetic quality control (AQC) in wooden stock grading, applicable to Industry 4.0 settings. Using 2,120 images graded into 10 quality levels, the method integrated a cumulative link model (CLM) with VGG16 to penalize distant misclassifications and account for geometric biases. The model achieved a 93.7% quadratic weighted kappa score, outperforming existing state-of-the-art methods. The ordinal approach is particularly relevant to leather grading scenarios, where defect severity or visual appeal is evaluated along a scale rather than as discrete classes.

Automation in defect localization was explored by Liong et al. [35], who integrated Mask R-CNN segmentation with a robotic arm for automated tick bite defect marking on leather. The robotic system not only detected but also provided precise defect coordinates for marking, with training accuracy of 91.5% and testing accuracy of 70.35%. While limited to a single defect type and affected by performance drops in real-world tests, the integration of detection, segmentation, and robotic actuation offers a model for end-to-end automated inspection systems.

Cost-effective and portable solutions have also emerged in species identification contexts. Authors [29] employed a digital microscope with hair-pore segmentation and morphological feature extraction, coupled with a KNN classifier, to achieve 92.5% accuracy in species classification. Although primarily targeting authentication, the combination of portable imaging hardware and lightweight machine learning models suggests potential pathways for low-cost, on-site leather defect detection in smaller-scale operations.

Together, these studies illustrate how innovations from related quality control tasks-fine-grained color classification, ordinal quality grading, integrated detection–marking systems, and portable inspection setups-can inform and enhance the design of automated leather inspection systems. They also emphasize that scalability, environmental robustness, and data availability remain cross-cutting challenges across industrial inspection domains.

2.6 Explainable AI and Model Interpretability

As deep learning models become increasingly prevalent in industrial quality control, their black-box nature poses a critical barrier to adoption. While convolutional neural networks

(CNNs) and vision transformers can achieve exceptional accuracy, their decision-making processes are often opaque to end-users-particularly quality inspectors and production managers in manufacturing environments. Without clear, human-understandable reasoning, stakeholders may be reluctant to fully trust automated decisions, especially in high-value domains such as leather inspection, where misclassifications can lead to significant financial losses.

Baraldi et al. [45] addressed this challenge in the context of Entity Matching (EM) by proposing *Landmark Explanation*, a framework designed to improve model interpretability. Although not directly applied to image classification, the methodology is relevant in that it adapts local post-hoc explanation techniques, such as LIME, to provide token-level justifications for predictions. The framework generates dual explanations through token perturbation and injection, thereby clarifying why a model considers two entities as matching or non-matching. Its novelty lies in enhancing transparency in situations where the decision process is inherently complex. However, the approach remains dependent on the quality of the underlying perturbation methods and the characteristics of the dataset.

In visual inspection contexts, explainable AI (XAI) has increasingly been paired with deep learning to make predictions more transparent. Methods like Gradient-weighted Class Activation Mapping (GradCAM) generate heatmaps that highlight the regions of an image most influential in a model's decision. This is particularly relevant for leather defect detection, where the location, size, and intensity of a defect influence its classification. GradCAM not only enables inspectors to verify whether the model's focus aligns with actual defect areas but also assists in debugging models by revealing biases, such as over-reliance on irrelevant background features.

The integration of interpretability methods into inspection systems serves multiple purposes:

- Providing visual evidence for AI predictions increases confidence among non-technical stakeholders.
- Identifying where the model's attention deviates from expected defect regions helps refine datasets and training strategies.

- In industries where automated quality control may be subject to audits, interpretability supports transparency requirements.

For leather quality control, combining high-performance defect detection models with explainability frameworks like GradCAM bridges the gap between predictive accuracy and human interpretability. This ensures that automated systems not only match but also justify the decisions made by experienced inspectors, fostering wider adoption in Industry 4.0 manufacturing environments.

2.7 Research Gap Analysis

The literature reviewed in this chapter illustrates the significant progress made in automating leather quality control, moving from traditional feature engineering methods to modern deep learning-based inspection systems. Early classical approaches-such as the Fourier Angular Radial Partitioning (FARP) algorithm, handcrafted texture descriptors with SVM optimization, and morphological feature extraction with portable microscopes-demonstrated that machine learning could outperform manual inspection in speed and consistency. However, these systems were constrained by their reliance on handcrafted features, sensitivity to environmental variations, and limited adaptability to diverse leather types and defect patterns.

The advent of deep learning shifted the paradigm by enabling automated feature extraction and multi-scale texture analysis. CNN-based systems such as LeatherNet, modified AlexNet architectures, and optimized residual networks achieved high classification accuracy for specific defect types, while object detection and segmentation models like YOLOv11 and Mask R-CNN introduced localization capabilities. Despite these advancements, most studies were trained on small, domain-specific datasets, often limited to a single species or a narrow set of defect classes, restricting their generalizability in real-world tannery environments.

Leather species identification studies have explored both CNN and hybrid preprocessing techniques to handle inter-species similarity and intra-species variability. While these approaches excelled in species classification, they rarely integrated defect detection into the same framework, missing the opportunity to build unified inspection systems capable of handling both tasks concurrently.

Specialized imaging techniques-including hyperspectral imaging, controlled imaging chambers, electronic sensing, and proteomic analysis-offered fine-grained or molecular-level insights into leather quality. Yet, their high cost, complex setup, and limited scalability hinder adoption in typical production lines. Similarly, cross-domain quality control studies provided transferable innovations such as ordinal grading, fine-grained classification, and integrated detection–marking pipelines, but these remain largely unexplored in leather defect contexts.

A critical observation across all reviewed works is the scarcity of XAI integration. While interpretability frameworks like Landmark Explanation and visual attribution methods such as GradCAM are well-suited for industrial trust-building, few studies have adopted them systematically. This omission creates a gap between high-performance predictions and stakeholder confidence, particularly in quality-sensitive industries.

Based on the literature, the key research gaps are as follows:

- A lack of large-scale, balanced, multi-animal leather defect datasets covering multiple defect classes under real-world variability.
- Limited comparative analysis between CNN and Transformer-based models for leather defect detection, despite the proven advantages of Transformers in handling texture-rich, high-resolution imagery.
- Most models are optimized for a single domain, defect type, or imaging setup, reducing cross-environment adaptability.
- Minimal use of visual explanation tools like GradCAM to enhance interpretability and industry trust.
- Few studies have developed end-to-end, real-time inspection solutions that combine accurate classification, defect localization, and interpretability in a deployable application.

This thesis addresses these gaps by:

- Constructing a balanced, multi-animal leather defect dataset across four species and four defect categories, with both original and augmented images.

- Benchmarking CNN and Transformer architectures to identify performance trade-offs in accuracy, efficiency, and generalization.
- Integrating GradCAM-based XAI to provide visual justifications for defect predictions.
- Deploying the best-performing model in a Flask-based real-time web application, enabling practical use in tannery environments.

CHAPTER 3 Dataset and Preprocessing

3.1 Dataset Collection

Figure 1 depicts the overall proposed methodology.

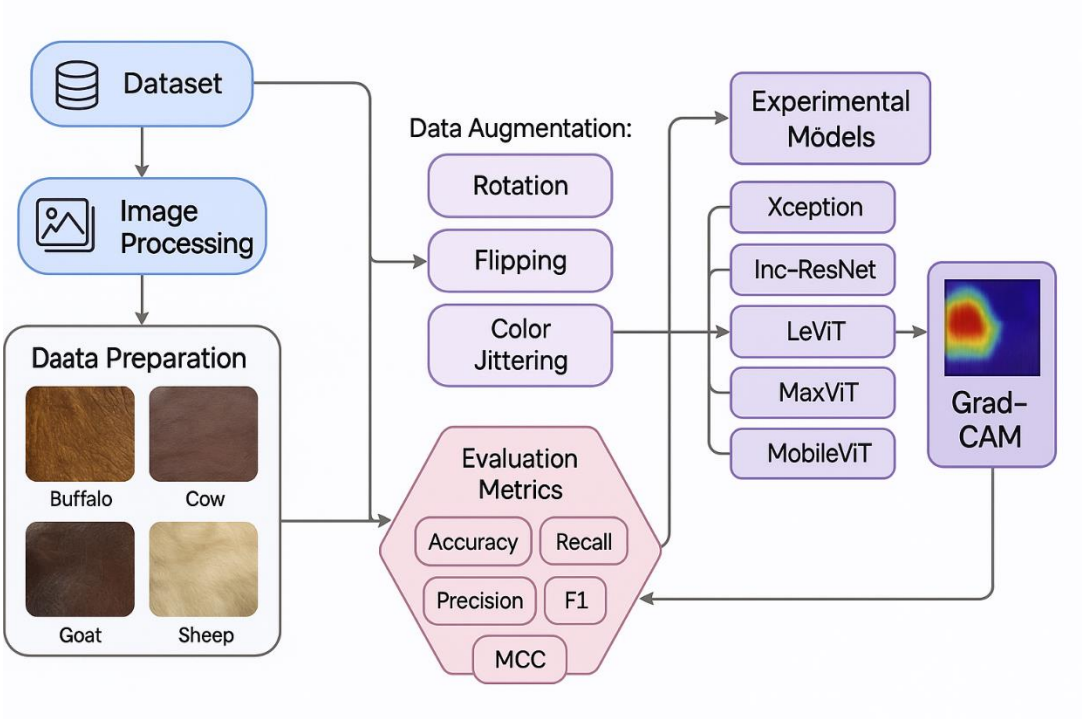


Figure 1. Overview of the proposed methodology.

3.1.1 Data Acquisition Sources

The dataset for this study was collected from multiple tanneries located in Dhaka, Bangladesh, a region recognized as a key hub in the national leather industry. These

facilities handle large-scale processing of hides from various animal sources, making them ideal for capturing a diverse range of leather textures and defect types under realistic industrial conditions.

Image acquisition was performed during different production workflow stages, including post-tanning inspection, drying, and grading phases. This ensured that the dataset encompassed both freshly processed and finished leather surfaces, thereby capturing variations in defect visibility across the production cycle. The multi-stage collection approach was essential for producing a dataset representative of real-world inspection scenarios in tanneries.

3.1.2 Species and Class Distribution

The dataset encompasses four distinct animal species: cow, goat, sheep, and buffalo, each selected for their prevalence in the leather industry and their unique surface grain characteristics. These natural variations in pore structure, fiber density, and texture play a significant role in how defects manifest, making species diversity essential for building a model capable of generalizing across different leather types. Within each species, four defect categories were considered: cut, fold, scratch, and normal. Cuts are linear or jagged surface damages typically caused by sharp tools or machinery during handling and processing. Folds represent permanent creases or wrinkles formed through improper storage or bending during the tanning process. Scratches appear as superficial abrasions resulting from friction with rough surfaces or improper transportation. Normal samples, free from visible defects, serve as a baseline for comparison and help the model distinguish between defective and non-defective leather.

A balanced dataset structure was maintained to avoid classification bias and to ensure equal representation of all species and defect classes. As presented in Table 1, for the original dataset, 400 images were collected for each animal species, with exactly 100 images per class, resulting in 1,600 original images across all species. Through targeted augmentation techniques (described in Section 3.2), the dataset was expanded to create an augmented version containing 1,600 images per species, with 400 images for each class, yielding a total of 6,400 augmented images. This balanced distribution ensures that the learning process is not skewed toward any single class or species, thereby improving the generalization capabilities of the trained models.

Table 1. Distribution of original and augmented images per species and defect class.

Species	Cut (Original/Aug)	Fold (Original/Aug)	Scratch (Original/Aug)	Normal (Original/Aug)	Total (Original/Aug)
Cow	100 / 400	100 / 400	100 / 400	100 / 400	400 / 1600
Goat	100 / 400	100 / 400	100 / 400	100 / 400	400 / 1600
Sheep	100 / 400	100 / 400	100 / 400	100 / 400	400 / 1600
Buffalo	100 / 400	100 / 400	100 / 400	100 / 400	400 / 1600
Total	400 / 1600	400 / 1600	400 / 1600	400 / 1600	1600 / 6400

3.1.3 Imaging Conditions

All images were captured using an iPhone 16 camera, which offers advanced optical and computational imaging capabilities suitable for high-resolution defect analysis. Key specifications include:

- Sensor Size: 1/1.3″ CMOS sensor.
- Resolution: 48 MP primary wide camera.
- Aperture: f/1.78 for optimal light capture.
- Focal Length: 26 mm equivalent.

To ensure dataset diversity, images were taken under both natural lighting (daylight from open workspaces) and artificial lighting (industrial LED and fluorescent sources). This variety in illumination conditions improves the model's robustness to lighting variations encountered in production environments. Positioning was standardized to minimize variability: leather samples were photographed at consistent distances, with minimal tilting, and against non-reflective backgrounds. Angles were adjusted slightly to replicate real-world inspection scenarios while avoiding excessive distortion. Figure 2 depicts sample image from each defect category of animals.

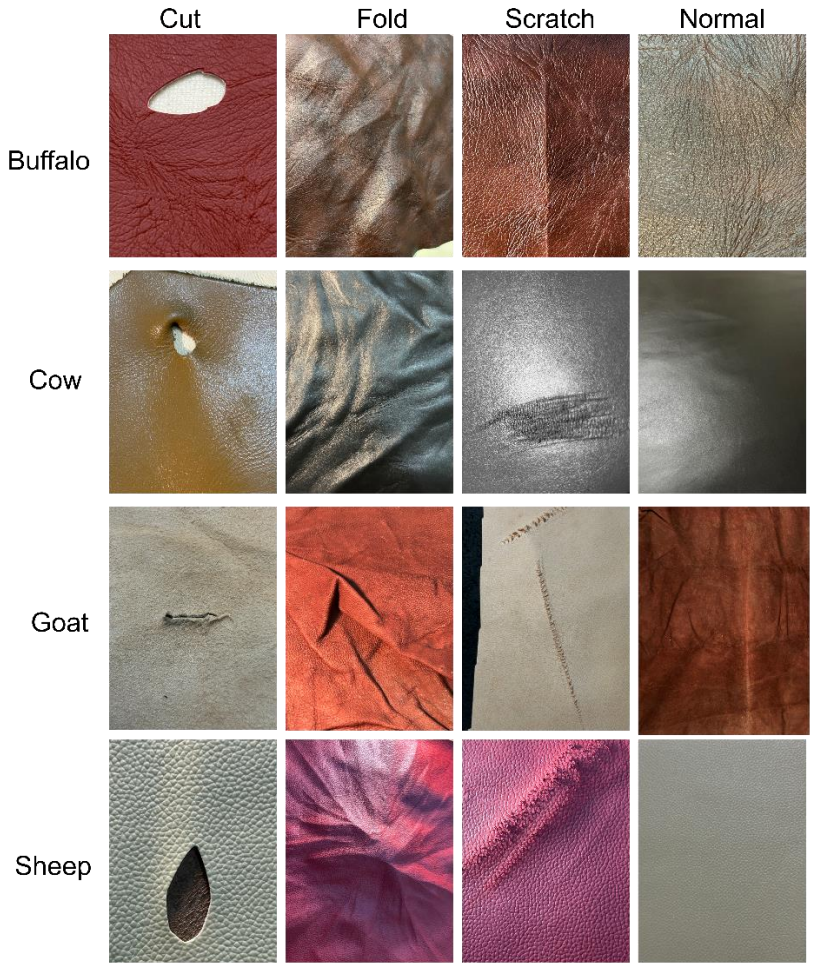


Figure 2. Sample images from each defect category across the four species.

3.2 Data Augmentation

3.2.1 Purpose of Augmentation

Data augmentation was implemented to tackle two significant challenges in leather defect classification: the limited size of the dataset and the variability encountered in real-world inspection conditions. By synthetically generating new image samples from the existing data, augmentation enhances the diversity of the training set without requiring additional manual image collection [46]. This increased diversity helps the model learn features that remain consistent regardless of changes in orientation, lighting, and texture presentation [47]. Since leather defects can manifest in various positions, scales, and lighting environments, augmentation is crucial for preventing overfitting and improving the model's generalization capabilities. Furthermore, incorporating controlled variations into the training data ensures that the model can effectively identify defects under conditions

that differ from those it encountered during training, making it better suited for deployment across diverse tannery environments. Figure 3 illustrates the output of each augmentation technique on a sample leather defect image.

3.2.2 Applied Augmentation Techniques

To achieve the goals outlined above, a series of carefully selected augmentation techniques were applied to all classes and species within the dataset, ensuring that the dataset remained balanced across categories:

- Random rotations of up to 15 degrees were applied to simulate inspection scenarios in which leather samples might be oriented differently. This variation helps the model learn rotational invariance, which is particularly important for linear defects like cuts and scratches.

- Both horizontal and vertical flips were applied to generate mirrored perspectives of each sample. This transformation is especially beneficial for defect detection, as defect patterns may present themselves in reversed orientations depending on the position of the hide during inspection.

- Images were resized within a scale range of 80% to 120% to simulate changes in the distance between the camera and the leather surface. This ensures that the model remains effective, even if defects appear larger or smaller due to variations in acquisition distance.

- Random adjustments were made to brightness, contrast, and saturation to reflect the varying lighting conditions found across different tanneries. These changes train the model to focus on texture and shape cues, rather than being overly sensitive to color or lighting variations.

- A small amount of Gaussian noise was introduced to simulate sensor imperfections and natural surface irregularities in leather. This helps improve the model's resilience against low-quality image capture conditions.

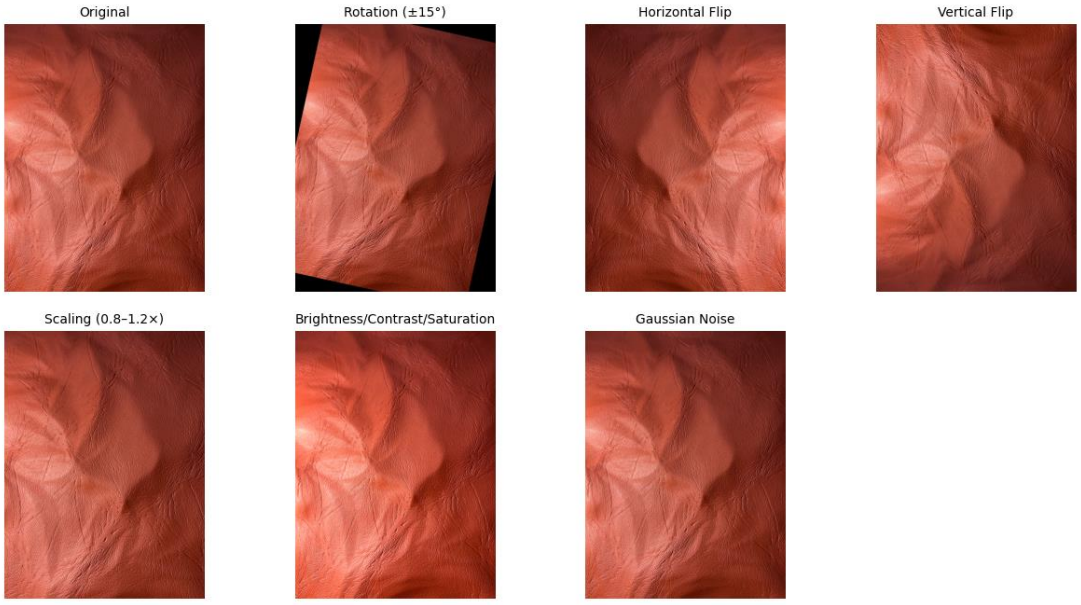


Figure 3. Visual examples showing an original image alongside its augmented variants for each technique.

3.2.3 Augmentation Implementation

The augmentation pipeline was implemented using PyTorch's `torchvision.transforms` module. Transformations were applied stochastically to ensure that each training epoch presented the model with slightly different variations of the same original image, further enhancing its generalization capabilities. To maintain dataset integrity, it was applied uniformly across all classes and species, preventing any single category from having an inflated sample size. The final augmented dataset retained the original balanced distribution, with 400 images per class per species, resulting in 6,400 augmented images in total. By combining targeted augmentation techniques with balanced application, the dataset was enriched with realistic variability while avoiding class bias, thus creating a robust foundation for model training.

3.3 Preprocessing

It is a crucial stage in preparing image datasets for deep learning, as it ensures consistency in input dimensions, pixel value distributions, and class labeling, while removing anomalies that could negatively impact training. For this study, the preprocessing pipeline consisted of image resizing**,** pixel value normalization**,** data integrity checks, and a

deliberate decision to omit certain steps in order to preserve the natural context of leather samples.

3.3.1 Image Resizing

All images were resized to 224×224 pixels to match the input layer requirements of the CNN and Transformer models used in this study. Resizing also ensures a uniform input size across the dataset, reducing computational complexity during training [48].

The resizing transformation is defined as:

$$I_{\text{resized}} = \mathcal{R}\left(I_{\text{original}}, W_t, H_t\right)$$

Where, I_{original} is the original input image, $W_t = 224$ and $H_t = 224$ are the target width and height in pixels, and $\mathcal{R}(\cdot)$ represents the resizing operation, implemented using bilinear interpolation to preserve visual details while minimizing aliasing artifacts.

3.3.2 Normalization

Normalization was applied to ensure that pixel values were on a consistent scale, facilitating faster convergence during training and improving numerical stability in gradient updates [49]. Two normalization strategies were considered:

Min–Max Scaling to [0, 1]**:**

$$I_{\text{norm}} = \frac{I_{\text{resized}}}{255}$$

Where I_{resized} contains pixel values in the range [0, 255].

Mean–Standard Deviation Normalization**:**

$$I_{\text{norm}} = \frac{I_{\text{resized}} - \mu}{\sigma}$$

Where, μ = mean pixel value computed over the training set and σ = standard deviation of pixel values over the training set. For this study, mean–std normalization was chosen to center the data distribution, which benefits both CNNs and Transformer-based architectures.

3.3.3 Data Integrity Checks

To maintain the quality of the dataset after augmentation, two primary integrity checks were performed:

(a) Ensuring that augmented images retained their correct labels was critical to prevent model mislearning. The verification process can be expressed as:

$$\forall I_i \in D_{\text{aug}}, \quad L(I_i) = L(I_{\text{source}})$$

Where, I_i is the augmented image, I_{source} is its original counterpart, and $L(\cdot)$ denotes the class label function.

(b) Duplicate detection was performed by computing image hashes and comparing them:

$$h(I_a) = h(I_b) \quad \Rightarrow \quad I_a \text{ and } I_b \text{ are duplicates}$$

Where $h(\cdot)$ is a perceptual hash function that maps images to fixed-length hash codes. Corrupted images were detected by attempting to load each file and checking for exceptions or invalid pixel dimensions.

3.3.4 Omitted Preprocessing Steps

No cropping or background removal was applied to the images. This decision was made to maintain the natural context in which defects appear in industrial inspection scenarios. Removing backgrounds could potentially strip away contextual cues (such as shadows, surface curvature, or ambient reflections) that may aid the model in distinguishing between defect and non-defect regions. Furthermore, the variation in background textures across species and defect types may help improve generalization by exposing the model to more diverse environmental conditions during training.

CHAPTER 4 Experimental Setup

4.1 Hardware and Software

All experiments were performed on a high-performance workstation configured to handle computationally intensive deep learning workflows. As presented in Table 3, the system was equipped with an NVIDIA GeForce RTX 3060 GPU featuring 12 GB of GDDR6 VRAM, enabling high-throughput parallel computations for both training and inference. An Intel® Core™ i7-12700K processor, operating at 3.60 GHz, was utilized for data preprocessing, augmentation, and efficiently serving inference requests in the deployed web application. The workstation also featured 32 GB of DDR4 memory, ensuring seamless data loading and parallel augmentation without memory bottlenecks. Storage requirements were met by a 1 TB NVMe SSD, providing rapid read/write access to large datasets and model checkpoints. The experiments were conducted in a Windows 11 Pro environment, chosen for its compatibility with development tools and GPU drivers.

Table 2. Hardware and software specifications.

Component	Specification / Version
GPU	NVIDIA GeForce RTX 3060, 12 GB GDDR6
CPU	Intel® Core™ i7-12700K @ 3.60 GHz
RAM	32 GB DDR4
Storage	1 TB NVMe SSD
Operating System	Windows 11 Pro (64-bit)
Python Version	3.10.12
Deep Learning Frameworks	PyTorch 2.0.1, TensorFlow 2.13.0
Libraries	torchvision 0.15.2, OpenCV 4.7.0, scikit-learn 1.3.0

Component	Specification / Version
Web Framework	Flask 2.3.2

The experimental framework was developed using Python 3.10 as the primary programming language. PyTorch served as the main deep learning framework, with TensorFlow used for alternative benchmarking and compatibility verification. The torchvision library facilitated access to model repositories, pretrained weights, and augmentation utilities, while OpenCV was employed for image reading, preprocessing, and visualization tasks. Additionally, scikit-learn was integrated for computing evaluation metrics, generating confusion matrices, and plotting performance curves. To enable real-time accessibility, Flask was used to build a web-based inference system that seamlessly combined model predictions with GradCAM visualizations, allowing intuitive interpretation of the classification outcomes.

4.2 Data Splitting

To ensure robust performance evaluation, the dataset was partitioned into three mutually exclusive subsets: training (80%), testing (10%), validation (10%). The split was performed stratified by class and species (Table 3 and Table 4), ensuring that each subset preserved the same proportional distribution of defect categories and animal types as the full dataset. This approach prevents class imbalance in any single subset and maintains fair evaluation conditions. Cross-validation techniques such as k-fold were not employed in this study. Given the balanced dataset and clearly defined training/validation/test splits, the primary focus was on maintaining fixed partitions for reproducibility and direct comparability across CNN and Transformer architectures.

Table 3. Dataset splitting detail.

Split	Percentage	No. of Images (Original)	No. of Images (Augmented)
Training	80%	1,280	5,120

Split	Percentage	No. of Images (Original)	No. of Images (Augmented)
Validation	10%	160	640
Testing	10%	160	640
Total	100%	1600	6400

Table 4. Model training parameters

Model	Learning Rate	Batch Size	Optimizer	Epochs	Scheduler
Xception	0.0003	32	Adam	50	CosineAnneal
InceptionResNetv2	0.0002	32	AdamW	60	StepLR
LeViT	0.0005	64	SGD (mom=0.9)	50	ReduceLROnPlateau
MaxViT	0.0001	16	AdamW	70	CosineAnneal
MobileViT	0.0004	64	Adam	40	StepLR

4.3 Performance Metrics

The performance of each model was assessed using a comprehensive set of classification metrics to capture various aspects of predictive ability:

- Accuracy: the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: the proportion of positive identifications that were actually correct:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: the proportion of actual positives correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1 Score: the harmonic mean of precision and recall, balancing both measures:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- MCC: a balanced measure accounting for all classes, suitable for multiclass evaluation:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

CHAPTER 5 Methodology

5.1 Model Architectures

5.1.1 Xception

It was proposed by Chollet in 2017, which is a convolutional neural network architecture that extends the Inception paradigm by replacing its conventional modules with depthwise separable convolutions [50]. The network is organized into three principal stages-Entry Flow, Middle Flow, and Exit Flow-that progressively transform input images into discriminative high-level features [51]. In this design, each standard convolution operation is decomposed into two separate steps: a depthwise convolution and a pointwise convolution. Given an input tensor $X \in R^{H\times W\times C_{in}}$ and a kernel $K \in R^{k\times k\times C_{in}\times C_{out}}$, the standard convolution can be expressed as

$$Y_{std}(i,j,o) = \sum_{m=1}^{k}\sum_{n=1}^{k}\sum_{c=1}^{C_{in}} K_{m,n,c,o} \cdot X_{i+m,j+n,c}.$$

Depthwise separable convolution factorizes this into a depthwise convolution,

$$Y_{depth}(i,j,c) = \sum_{m=1}^{k}\sum_{n=1}^{k} K_{m,n}^{(c)} \cdot X_{i+m,j+n,c},$$

followed by a pointwise convolution,

$$Y_{point}(i,j,o) = \sum_{c=1}^{C_{in}} P_{c,o} \cdot Y_{depth}(i,j,c\}.$$

This approach reduces the computational complexity from $k^2 \cdot C_{in} \cdot C_{out}$ to $k^2 \cdot C_{in} + C_{in} \cdot C_{out}$, enabling faster training and inference while maintaining expressive power.

In this study, the Xception model was adapted to address the problem of multi-species leather defect detection. As shown in Figure 4, the input layer was configured to accept preprocessed images of size $224 \times 224 \times 3$. The Entry Flow stage employed three depthwise separable convolutional blocks followed by max pooling, with filter sizes of 32, 64, and 128, designed to extract low-level defect patterns and textural cues from the leather surfaces. The Middle Flow comprised eight identical residual blocks using depthwise separable convolutions to learn complex and fine-grained surface characteristics across multiple species. The Exit Flow expanded the feature representation through a sequence of convolutions with 728, 1024, 1536, and 2048 filters, followed by global average pooling. Finally, a fully connected layer with Softmax activation was employed to classify the extracted features into sixteen classes, corresponding to four defect categories across four species.

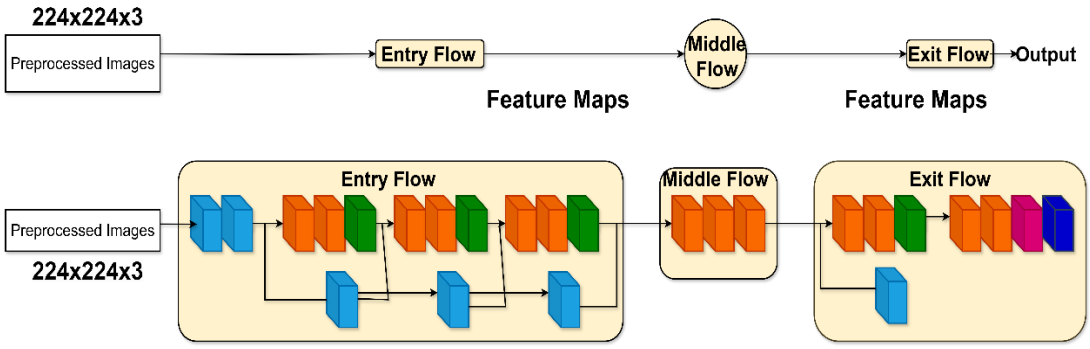


Figure 4. Xception architecture.

The adoption of depthwise separable convolutions significantly reduced the number of parameters and computations compared to conventional CNN architectures, enabling efficient training without sacrificing performance [52]. This efficiency, combined with the model's hierarchical feature extraction process, enhanced its ability to detect subtle texture variations and structural defects that are critical in leather quality assessment. However, its reliance on reduced parameterization also introduced certain limitations. On smaller datasets with insufficient augmentation, the model exhibited a tendency to underfit due to the limited representational capacity [53]. Furthermore, its performance was sensitive to hyperparameter choices [54], such as the learning rate and regularization parameters, necessitating careful tuning to achieve optimal results.

5.1.2 Inception-ResNetv2

It a sophisticated fusion of the Inception modules and residual connections, combining the representational efficiency of Inception with the training stability and gradient flow advantages of residual learning [55]. As depicted in Figure 5, the network begins with a stem block that processes the input image of size $224 \times 224 \times 3$ using a series of convolutions and pooling operations to extract low-level features while reducing spatial dimensions. This is followed by multiple Inception-ResNet-A modules, which are specifically designed for capturing multi-scale spatial correlations through parallel convolutional branches [56]. Each branch processes the feature map with different kernel sizes-such as 1×1, 3×3-to learn both fine-grained and context-aware features. The outputs from these branches are concatenated and projected back to the original dimensionality via a 1×1 convolution, and a residual shortcut connection adds the original input to the processed output, formulated as:

$$y = F(x, W) + x$$

where $F(x, W)$ denotes the transformation applied by the Inception sub-blocks and x is the identity mapping. A non-linear activation function, typically ReLU, follows the summation to enhance the network's non-linearity [57]. After a series of Inception-ResNet-A modules, a Reduction-A block reduces the spatial resolution while expanding the depth of feature maps, enabling the network to capture higher-level abstractions. This is succeeded by multiple Inception-ResNet-B modules, which incorporate asymmetric convolutions [58], such as 1×7 followed by 7×1, to efficiently model long-range spatial dependencies without excessive computational cost. The structure of these modules is optimized to minimize parameter count while maintaining high representational power.

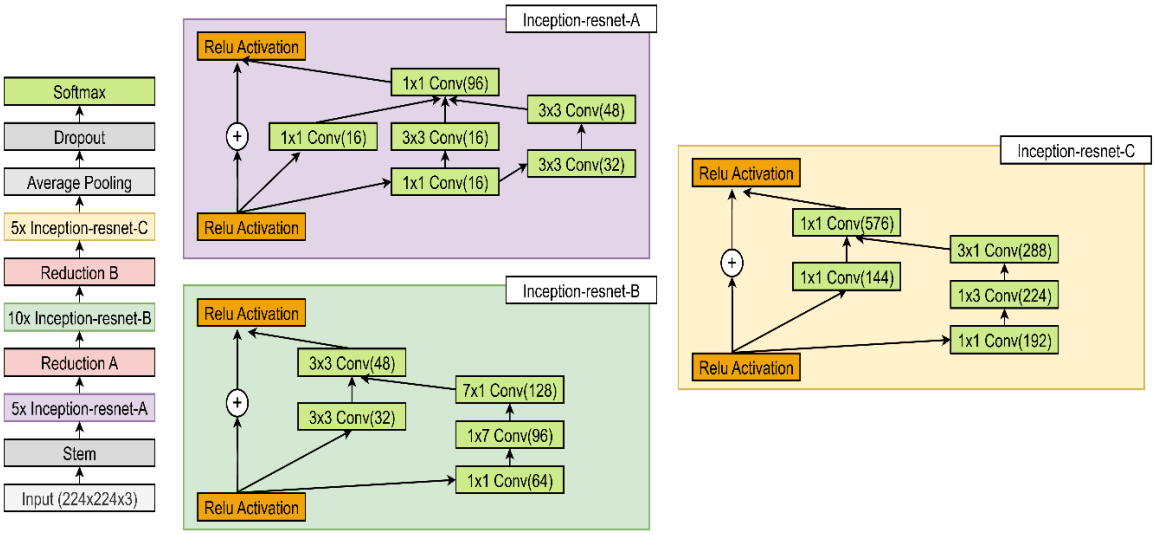


Figure 5. Inception-ResNetv2 architecture

Following the Inception-ResNet-B stage, the network applies a Reduction-B block to further downsample and increase feature dimensionality. This leads into several Inception-ResNet-C modules, which leverage 1×3 and 3×1 convolutional filters to capture anisotropic spatial features and refine complex high-level representations. Throughout these stages, residual connections ensure stable training, prevent vanishing gradients, and facilitate convergence in deeper architectures [59]. The architecture concludes with a global average pooling layer that condenses the spatial dimensions to a single vector per feature map, followed by a dropout layer to reduce overfitting. Finally, a fully connected layer outputs the class probabilities through a Softmax activation function. The model provides exceptional feature extraction capabilities by integrating multi-scale convolutions with residual learning, making it effective for large-scale image recognition tasks. However, its deep and complex design demands substantial computational resources and memory [60], which can limit deployment in real-time or resource-constrained environments.

5.1.3 LeViT

Its architecture represents a hybrid design that integrates CNN layers with the efficiency of ViTs, specifically engineered for high-speed image classification without compromising accuracy. The architecture (Figure 6) begins with a series of 3×3 convolutional layers that extract low-level spatial features from the input image, while progressively reducing its spatial resolution. These initial convolutional stages serve as

an effective inductive bias, capturing local patterns such as edges and textures, and producing a compact feature map that feeds into the transformer stages [61].

Following the convolutional stem, the model transitions into transformer blocks that are structured in a multi-stage hierarchy, each consisting of alternating Multi-Layer Perceptron (MLP) blocks and Multi-Head Self-Attention (MHSA) mechanisms [62]. The self-attention mechanism computes relationships between all token pairs using the equation:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices derived from the input tokens, and d_k is the dimensionality of the keys. This allows the network to capture global dependencies and contextual relationships, which CNNs traditionally struggle to represent. To enhance computational efficiency, LeViT employs a "Shrink Attention" mechanism at the boundaries between stages, where spatial resolution is reduced and embedding dimensions are increased. This downsampling process decreases the number of tokens while enriching their feature representation, striking a balance between accuracy and speed [63]. In early transformer stages, 4-head attention mechanisms focus on fine local patterns, while deeper stages utilize higher head counts (6, 8, and 12 heads) to capture broader contextual relationships. Each attention output is combined with residual connections and passed through an MLP block, expressed as:

$$y = x + \sigma\big(W_2\delta(W_1 x)\big)$$

where W_1 and W_2 are learnable weights, δ represents the GELU activation function, and σ is the dropout regularization. The hierarchical design concludes with an average pooling layer that aggregates spatial information across all tokens, followed by a fully connected classifier head for final category prediction [64].

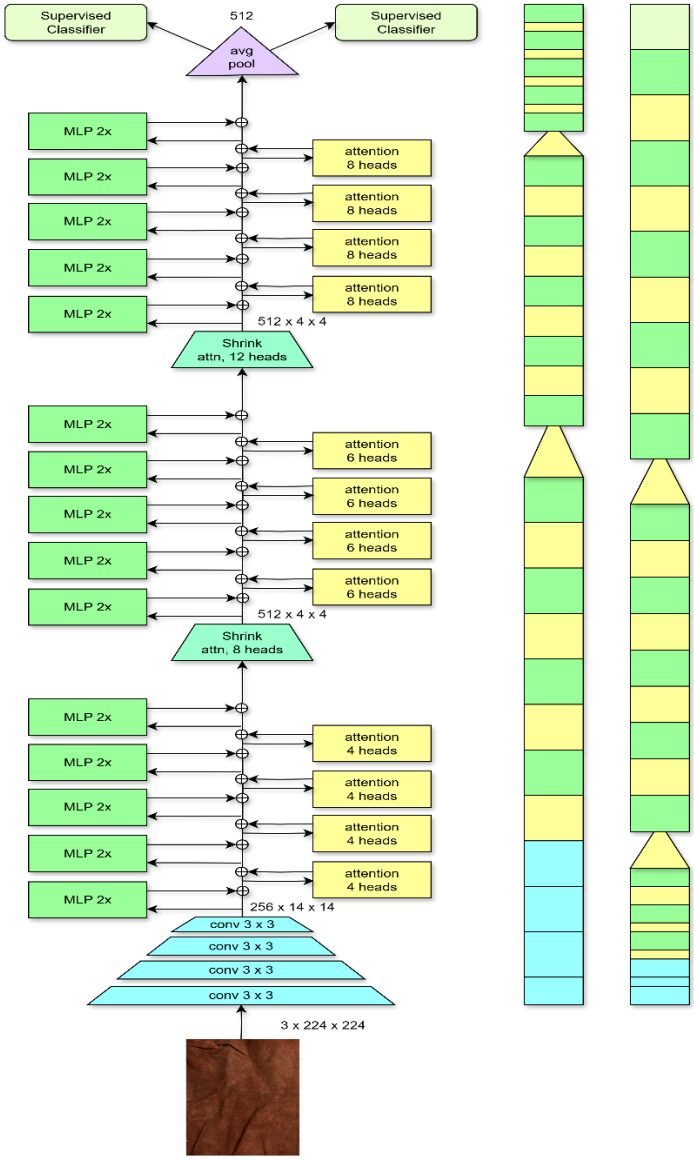


Figure 6. LeViT architecture.

Its hybrid CNN-transformer structure combines the local feature extraction efficiency of convolutions with the global modeling capabilities of transformers, making it particularly well-suited for real-time image classification on resource-constrained devices. However, while its speed and memory efficiency are strong advantages, the reliance on hybrid designs may introduce architectural complexity, requiring careful tuning of convolutional depth, attention heads, and token reduction rates to achieve optimal performance across varying datasets [65].

5.1.4 MaxViT

It marries convolutional tokenization with multi-axis self-attention so the network can reason about leather textures at both microscopic (scratches, pores) and macroscopic (folds, cuts) scales. Let $X \in R^{H\times W\times C}$ be an RGB image. A convolutional stem applies two 3×3 convolutions (the first with stride 2) to yield $X_0 \in R^{H_0\times W_0\times C_0}$, providing a local inductive bias and reducing spatial cost before attention. Figure 7 illustrates its architecture.

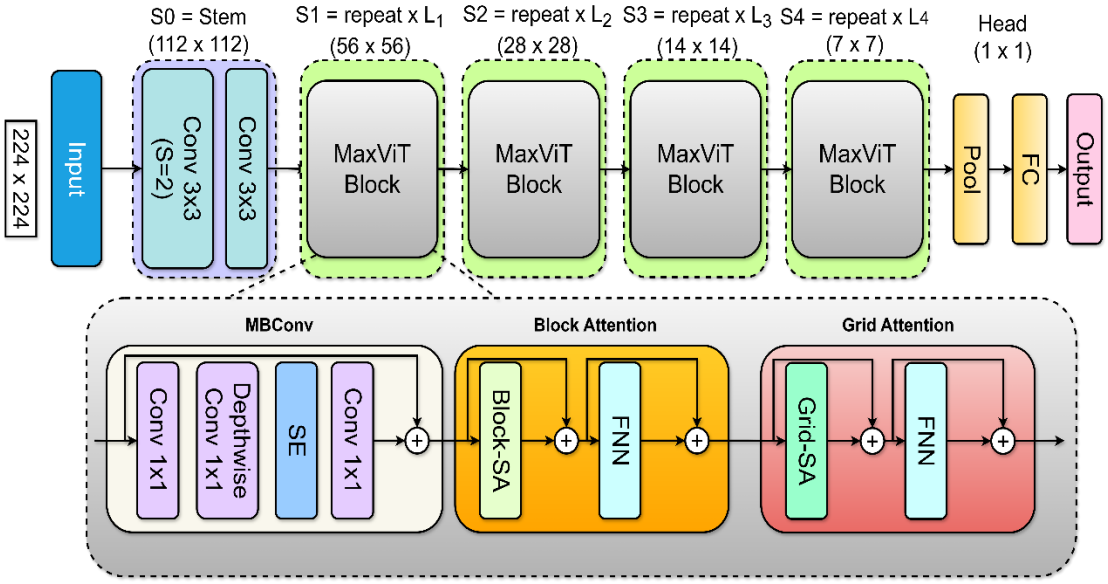


Figure 7. MaxViT architecture

Within each MaxViT block the first submodule is MBConv, an inverted bottleneck with depthwise separable convolution and channel re-weighting. Writing BN for BatchNorm and ϕ for a pointwise nonlinearity (e.g., SiLU), the expansion–depthwise–projection path is

$$Z_{\text{exp}} = \phi\big(\text{BN}(X_0 W_e^{1\times 1})\big),$$

$$Z_{\text{dw}} = \phi\left(\text{BN}\left(\text{DWConv}_{3\times 3}(Z_{\text{exp}})\right)\right),$$

$$\tilde{Z} = \text{BN}\big(Z_{\text{dw}} W_p^{1\times 1}\big).$$

Channel attention uses Squeeze-and-Excitation: with global average pooling $u = \text{GAP}(Z_{\text{dw}}) \in R^{C_e}$,

$$s = \sigma!\left(W_2\,\delta(W_1 u)\right), \qquad Z_{\text{se}} = \tilde{Z} \odot s,$$

where δ is GeLU, sigma is sigmoid, and $\odot$ denotes channel-wise broadcast multiplication. A residual with stochastic depth stabilizes training:

$$Y_{\text{MB}} = X_0 + m \cdot Z_{\text{se}}, \qquad m \sim \text{Bernoulli}\left(1 - p_{\text{drop}}\right).$$

After MBConv, MaxViT performs multi-axis attention in two stages-block then grid-each followed by a position-wise feed-forward network (FFN). Pre-norm is applied via LayerNorm,

$$\text{LN}(x) = \gamma \odot \frac{x - \mu(x)}{\sqrt{\sigma^2(x) + \varepsilon}} + \beta,$$

to produce tokens $T = \text{LN}(Y_{\text{MB}}) \in R^{N \times d}$ with $N = H_0 W_0$. For block self-attention, the feature map is partitioned into non-overlapping windows of size $b_h \times b_w$. Let $S_b^{(t)} \in \{0,1\}^{n_b \times N}$ be a selection matrix that gathers the indices of tokens in block t (with $n_b = b_h b_w$). Queries, keys, and values for head h are

$$Q_h^{(t)} = S_b^{(t)} T W_Q^{(h)}, \quad K_h^{(t)} = S_b^{(t)} T W_K^{(h)}, \quad V_h^{(t)} = S_b^{(t)} T W_V^{(h)}.$$

Using relative positional bias $B_h^{(t)}\in\mathbb{R}^{n_b\times n_b}$, attention within block t is

$$A_h^{(t)} = \text{Softmax!}\left(\frac{Q_h^{(t)} {K_h^{(t)}}^{!\top} + B_h^{(t)}}{\sqrt{d_h}}\right), \qquad O_h^{(t)} = A_h^{(t)} V_h^{(t)}.$$

Outputs from all blocks are scattered back via $S_b^{(t)\top}$ and concatenated across heads:

$$O_{\text{blk}} = \bigoplus_{h=1}^{H} \left(\sum_t S_b^{(t)\top} O_h^{(t)}\right) W_O^{(h)}.$$

A residual FFN with gated activation refines tokens,

$$\text{FFN}(x) = W_2!\left(\phi(xW_{1a}) \odot xW_{1b}\right), \qquad T_{\text{blk}} = T + O_{\text{blk}} + \text{FFN!}\left(\text{LN}(T + O_{\text{blk}})\backslash big\right).$$

For grid self-attention, tokens are re-indexed by a permutation Π_g that forms a coarse interlaced grid (orthogonal to block partitioning) so distant regions interact sparsely but globally [66]. With selection matrices $S_g^{(u)}$ for each grid cell u (of size n_g), attention mirrors the block equations:

$$Q_h^{(u)} = S_g^{(u)} \Pi_g T_{\text{blk}} W_Q^{(h)}, \quad K_h^{(u)} = S_g^{(u)} \Pi_g T_{\text{blk}} W_K^{(h)}, \quad V_h^{(u)} = S_g^{(u)} \Pi_g T_{\text{blk}} W_V^{(h)},$$

$$A_h^{(u)} = \text{Softmax!}\left(\frac{Q_h^{(u)} {K_h^{(u)}}^{!\top} + B_h^{(u)}}{\sqrt{d_h}} \right), \quad O_{\text{grid}} = \bigoplus_{h=1}^{H} \left(\sum_u \Pi_g^{!\top} S_g^{(u)\top} A_h^{(u)} V_h^{(u)} \right) W_O^{(h)}.$$

Another residual FFN yields $T_{\text{out}} = T_{\text{blk}} + O_{\text{grid}} + \text{FFN!}\left(\text{LN}\left(T_{\text{blk}} + O_{\text{grid}}\right)\right)$. By alternating local block and global grid attentions inside every MaxViT block, the model captures both fine leather micro-textures and long-range fold geometry. The computational profile explains the scalability: full attention costs $\mathcal{O}(N^2 d)$, while MaxViT's two axes cost

$$\mathcal{O}(N\, n_b\, d) \;+\; \mathcal{O}\big(N\, n_g\, d\big),$$

with $n_b = b_h b_w \ll N$ and $n_g \ll N$. For typical window/grid sizes, this reduces complexity by orders of magnitude at large resolutions. A stagewise stack of these blocks is followed by global average pooling and a linear classifier [67]. Given pooled representation z, class probabilities use softmax. Training adopts cross-entropy with label smoothing ε to regularize multi-class leather defects,

$$\mathcal{L} = -\sum_{k=1}^{K} \left[(1-\varepsilon)\, \mathbf{1}_{[\mathbf{y}=\mathbf{k}]} + \frac{\varepsilon}{K} \right] \log p_k,$$

while DropPath in residuals provides depth-wise regularization. This formulation preserves local inductive biases through MBConv and SE, models intra- and inter-region dependencies via multi-axis attention, and remains computationally tractable for high-resolution leather imagery where both minute scratches and global fold patterns must be recognized.

5.1.5 MobileViT

It couples MobileNetV2-style local processing with a lightweight transformer that operates on unfolded patches, so the network can model both micro-texture (tiny scratches, pores) and macro-structure (folds, cuts) in leather. Figure 8 depicts its architecture.

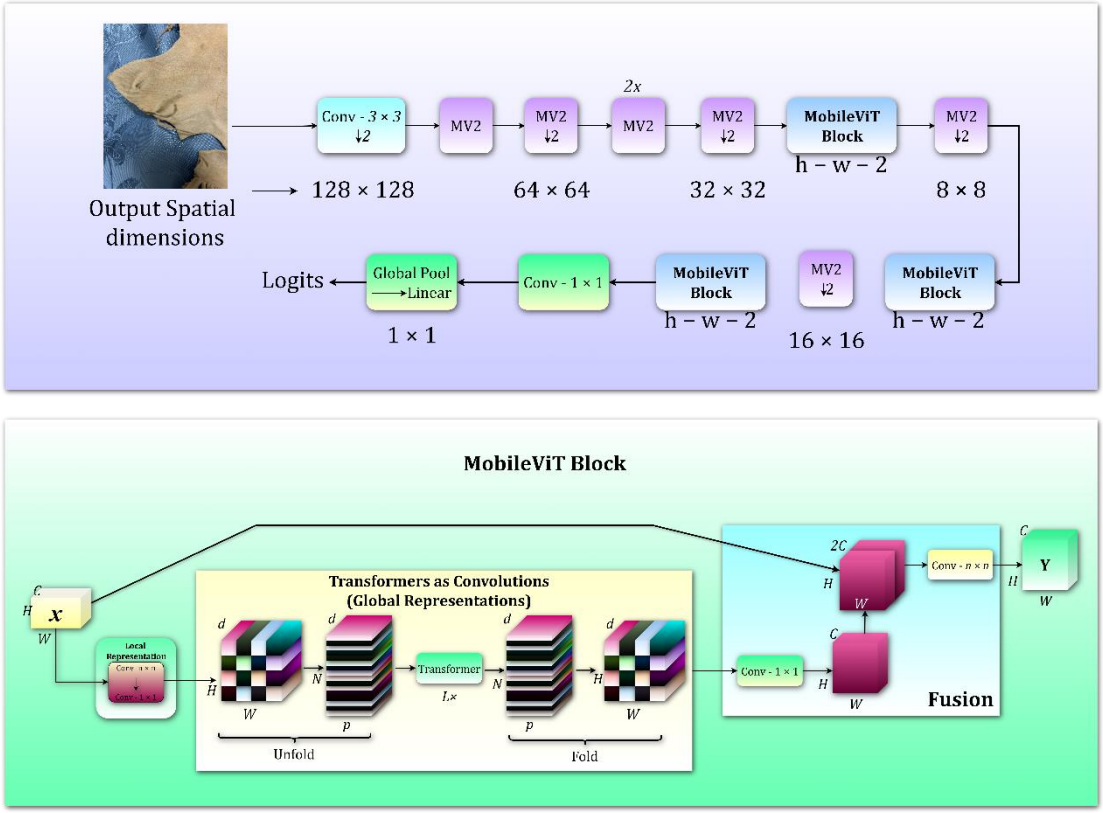


Figure 8. MobileViT architecture

Let $X! \in ! R^{H\times W\times C}$ be an input feature map produced by the convolutional stem/MV2 bottlenecks. Denote by $\mathcal{U}_{\mathcal{p}}: R^{H\times W\times C}! \to ! R^{N\times(p^2C)}$ the unfold (im2col) operator that extracts $p \times p$ non-overlapping patches in raster order, where $N = \frac{HW}{p^2}$. Tokens are obtained by a learned linear patch embedding

$$T = \mathcal{U}_{\mathcal{p}}(X)W_E + \mathbf{1}b_E^{\top} \quad \in R^{N\times d},$$

with $W_E \in R^{(p^2C)\times d}$. For stable optimization MobileViT uses pre-norm; with LayerNorm LN, the input to attention is $\hat{T} = \mathrm{LN}(T)$. Multi-head self-attention with h heads is computed as follows. For each head $r \in \{1, \dots, h\}$, queries, keys, and values are

$$Q_r = \hat{T}W_Q^{(r)}, \qquad K_r = \hat{T}W_K^{(r)}, \qquad V_r = \hat{T}W_V^{(r)},$$

with $W_Q^{(r)}, W_K^{(r)}, W_V^{(r)} \in R^{d\times d_h} and d_h = d/h$. To encode 2D relative positions between tokens that came from patches on a grid, let $\Delta u, \Delta v$ be horizontal/vertical offsets between any two patches; attention logits incorporate a separable bias

$$\mathbf{B_{ij}^{(r)}} = a_{\Delta u(i,j)}^{(r)} + b_{\Delta v(i,j)}^{(r)},$$

where $a^{(r)} \in R^{2\sqrt{N}-1} and b^{(r)} \in R^{2\sqrt{N}-1}$ are learnable tables indexed by offsets. Scaled-dot attention with this bias is

$$A_r = \text{Softmax}!\left(\frac{Q_r K_r^\top}{\sqrt{d_h}} + \mathbf{B^{(r)}}\right), \qquad O_r = A_r V_r.$$

Heads are concatenated and projected,

$$O = [O_1 \mid \ldots \mid O_h]W_O, \quad W_O \in R^{d\times d}.$$

A gated FFN (better than plain ReLU MLP for compact models) refines tokens:

$$\text{FFN}(Z) = W_2!\left(\sigma!\left(ZW_{1a}\right) \odot \text{GeLU}(ZW_{1b})\right),$$

with $W_{1a}, W_{1b} \in R^{d\times d_f}$, $W_2 \in R^{d_f\times d}$, gate $\sigma(\cdot)$ the sigmoid, and $\odot$ Hadamard product. Using stochastic depth mask $m \sim \text{Bernoulli}(1 - p_{\text{drop}})$, the transformer block output is

$$T' = T + m\,O, \qquad T'' = T' + m\,\text{FFN}!\left(\text{LN}(T')\right).$$

Tokens are returned to the image lattice via the fold operator $\mathcal{F}_p$: $R^{N\times d}! \to ! R^{H\times W\times d}$, the left inverse of $\mathcal{U}_p$ on non-overlapping patches:

$$G = \mathcal{F}_p(T'') \in R^{H\times W\times d}.$$

To fuse global and local representations, MobileViT concatenates the transformer output with a local path $L = \phi!\left(\text{BN}\left(\text{Conv}_{\text{n}\times\text{n}}(X)\right)\right)$ that preserves high-frequency cues, and applies a 1×1 fusion convolution:

$$Y = \text{Conv}_{1\times1}!\left(\text{Concat}(G, L)\right) \in R^{H\times W\times C},$$

optionally preceded by channel alignment via $\tilde{G} = \mathrm{Conv}_{1\times1}(G)$. A residual connection with DropPath may be used when channel dimensions match, further improving optimization depth [68]. This block is embedded in a MobileViT stage: MV2 bottlenecks downsample and compress; MobileViT blocks at resolutions $H \times W \in \{32^2, 16^2, 8^2\}$ perform global reasoning with small token counts $N = HW/p^2$. The total attention cost is thus

$$\mathcal{O}(h\,N^2 d_h) = \mathcal{O}!\left(\frac{HW}{p^2}\cdot\frac{HW}{p^2}\cdot d\right),$$

which remains tractable because p is small and d is modest. The convolutional stem and MV2 blocks handle high-resolution detail with cost $\mathcal{O}(k^2HWC)$, while transformer stages trade spatial resolution for context, yielding an overall efficient hybrid computation [69]. For classification, global average pooling produces $z = \frac{1}{HW}\sum_{i,j} Y_{ij:}$, and logits.

5.2 Training Strategy

The training pipeline with the ViT backbone initialized in three regimes: (i) from scratch, where all parameters θ were drawn from a He normal distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = \sqrt{2/f_{\text{in}}}$; (ii) ImageNet-pretrained, transferring low-level convolutional filters and the transformer patch embedding to accelerate convergence; and (iii) fine-tuning, where pretrained weights were frozen in early layers for the first T_f epochs before joint optimization.

Model performance was evaluated on two settings: augmented and non-augmented datasets. The evaluation protocol maintained identical test splits to ensure fairness. Let $S_{\text{aug}} and S_{\text{raw}}$ denote the test scores for augmented and raw training, respectively; the relative gain is computed as

$$\Delta_{\text{rel}} = \frac{S_{\text{aug}} - S_{\text{raw}}}{S_{\text{raw}}} \times 100\%.$$

5.3 XAI Integration

To enhance interpretability, GradCAM was applied over the final convolutional layer in the ViT fusion stage and the attention maps from the last transformer block. Given feature activations A^k and loss gradient $\frac{\partial y^c}{\partial A^k}$ for class c, channel-wise importance weights are computed as

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k},$$

where Z is the spatial dimension normalization factor. The localization map is then

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right),$$

normalized to $[0,1]$ and upsampled to input resolution. Overlaying on the original leather image highlights micro and macro defects, enabling visual defect attribution. For attention heads, the mean attention rollout was also computed to cross-validate defect localization regions [70], [71].

CHAPTER 6 Result Analysis

6.1 Model Performance on Non-Augmented Dataset

The performance comparison on the non-augmented dataset reveals distinct differences between the models and classes. As presented in Table 5, MaxViT consistently stands out as the top performer, achieving Accuracy, Recall, Precision, F1 Score, and MCC values exceeding 0.96 for every class. This uniform excellence demonstrates its effective generalization across all categories, including those that are challenging for other models. LeViT closely follows, delivering stable results with minimal variation across Buffalo, Cow, Goat, and Sheep, suggesting it maintains balanced performance without being overly sensitive to specific class characteristics. Examining individual animal classes, Buffalo emerges as the easiest to classify, with all models surpassing 0.95 in most metrics. MaxViT leads in this category with an Accuracy of 0.98651 and an F1 Score of 0.98765, though the differences between models are relatively small. The Cow category presents a contrasting scenario-while MaxViT achieves the highest Accuracy, MobileViT demonstrates the best Recall (0.98765) and Precision (0.98877), highlighting its effectiveness in detecting positive cases with minimal false positives. In contrast, Xception's performance on Cows significantly drops, with an Accuracy of 0.92683, indicating that its feature extraction may not be well-suited for this class.

Classifying Goats is more challenging, as reflected in the lower scores across all models. MaxViT remains the most effective model, achieving an Accuracy of 0.96 and an MCC of 0.96198. Xception struggles considerably, with an Accuracy of only 0.85 and an F1 Score of 0.84375. MobileViT also experiences a noticeable decline in this class, with results that, while better than Xception's, still fall short of the other top models. Sheep classification poses the greatest difficulty for nearly every model, highlighting a clear performance gap between the strongest and weakest approaches. MaxViT again excels, reaching an Accuracy of 0.98562 and maintaining balanced scores across metrics. Notably, LeViT delivers uniformly high values for Sheep, matching its performance in other categories, which indicates strong adaptability. MobileViT achieves the highest Precision across the entire dataset (0.99667) and an F1 Score of 0.98908 for Sheep, showcasing exceptional accuracy in avoiding false positives. In contrast, Inc-ResNet

produces the weakest results for this class, with Accuracy dropping to 0.70732, suggesting difficulties in recognizing Sheep patterns effectively.

Table 5. Performance comparison of experimental models on non-augmented dataset.

Model	Animal	Accuracy	Recall	Precision	F1 Score	MCC
Xception	Buffalo	0.97561	0.97561	0.97783	0.97555	0.96823
	Cow	0.92683	0.92683	0.93496	0.92782	0.90461
	Goat	0.85	0.85	0.86905	0.84375	0.81088
	Sheep	0.75663	0.75007	0.87181	0.73332	0.71564
Inc-ResNet	Buffalo	0.95122	0.95122	0.95935	0.95221	0.9372
	Cow	0.97561	0.97561	0.97783	0.97555	0.96823
	Goat	0.95	0.95	0.95833	0.94949	0.93646
	Sheep	0.70732	0.70732	0.82818	0.68599	0.65532
LeViT	Buffalo	0.97561	0.97561	0.97783	0.97555	0.96823
	Cow	0.97561	0.97561	0.97783	0.97555	0.96823
	Goat	0.95	0.95	0.95833	0.94949	0.93646
	Sheep	0.97561	0.97561	0.97783	0.97555	0.96823
MaxViT	Buffalo	0.98651	0.98615	0.98837	0.98765	0.96823
	Cow	0.98652	0.98345	0.98023	0.98076	0.97328
	Goat	0.96	0.96	0.96388	0.96198	0.96198
	Sheep	0.98562	0.98165	0.98387	0.98443	0.98328
MobileViT	Buffalo	0.97561	0.97561	0.97783	0.97555	0.96823
	Cow	0.98034	0.98765	0.98877	0.98033	0.97043
	Goat	0.925	0.925	0.92677	0.92481	0.90075
	Sheep	0.98473	0.98231	0.99667	0.98908	0.97039

As shown in Figure 9, the trends in Accuracy and F1 Score confirm MaxViT as the most reliable choice, especially for complex categories like Goat and Sheep. MobileViT shows significant potential for high-precision applications, particularly when minimizing false positives is crucial. LeViT offers a dependable, well-rounded alternative with consistent performance, while Xception and Inc-ResNet appear less capable of handling the variability and visual complexity of the more difficult classes.

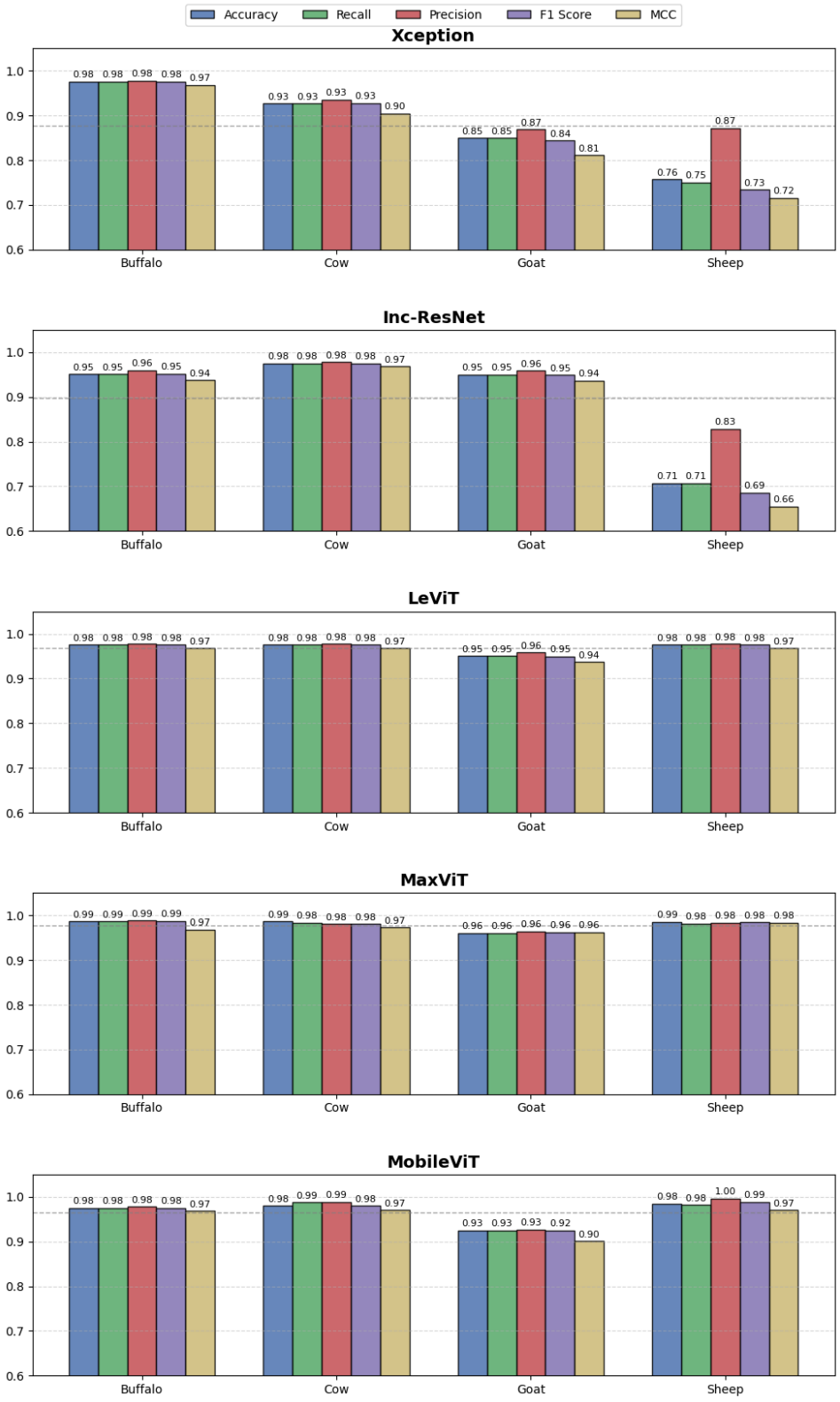


Figure 9. Grouped bar chart performance comparison per model for non-augmented dataset

6.2 Model Performance on Augmented Dataset

Data augmentation significantly enhances the performance of most models compared to non-augmented datasets, with improvements observed across all metrics for challenging classes such as Goat and Sheep. Table 6 presents the performance of the models on augmented version of the dataset. MaxViT stands out as the strongest performer overall, consistently achieving top-tier results in all categories and demonstrating superior robustness after augmentation. LeViT also benefits greatly, showing balanced, high scores across all animal categories, while MobileViT excels in certain precision-sensitive scenarios but experiences inconsistencies in specific metrics.

Table 6. Performance comparison of experimental models on augmented dataset

Model	Animal	Accuracy	Recall	Precision	F1 Score	MCC
Xception	Buffalo	0.98362	0.98394	0.98442	0.98348	0.97601
	Cow	0.93321	0.9368	0.94059	0.93733	0.90964
	Goat	0.85811	0.85878	0.87768	0.85323	0.81982
	Sheep	0.86635	0.85964	0.89801	0.87953	0.84416
Inc-ResNet	Buffalo	0.95712	0.96007	0.96733	0.95953	0.94634
	Cow	0.98124	0.98112	0.98361	0.98211	0.97534
	Goat	0.95815	0.95943	0.96397	0.95753	0.94403
	Sheep	0.94561	0.94561	0.94783	0.94555	0.92823
LeViT	Buffalo	0.98531	0.9816	0.98677	0.98409	0.97324
	Cow	0.95926	0.95526	0.96584	0.95714	0.94169
	Goat	0.98131	0.9822	0.98444	0.98424	0.97337
	Sheep	0.98343	0.98461	0.98731	0.98244	0.9757
MaxViT	Buffalo	0.99272	0.99185	0.99492	0.99695	0.97434
	Cow	0.99236	0.98989	0.98724	0.98749	0.98097
	Goat	0.96572	0.96991	0.9714	0.96954	0.97016
	Sheep	0.99529	0.99072	0.99302	0.99012	0.99131
MobileViT	Buffalo	0.98261	0.98483	0.98486	0.98209	0.97425
	Cow	0.986	0.99454	0.99708	0.98867	0.97575
	Goat	0.9334	0.93176	0.93492	0.93113	0.90749
	Sheep	0.99365	0.98921	0.99016	0.99701	0.97945

As shown in Figure 10, Buffalo remains one of the easiest categories to classify, with all models achieving high accuracy. MaxViT leads with an Accuracy of 0.99272 and an F1 Score of 0.99695, indicating near-perfect detection. LeViT and MobileViT follow closely, both surpassing 0.98 across all metrics, which shows that augmentation further stabilizes high performance for well-represented classes. For Cow classification, augmentation yields notable improvements for models like Inc-ResNet (Accuracy 0.98124) and MobileViT (Accuracy 0.986). MobileViT achieves exceptional Recall

(0.99454) and Precision (0.99708), indicating its strength in identifying true positives with minimal false positives. MaxViT remains highly competitive with an Accuracy of 0.99236, slightly trailing in precision but maintaining the highest MCC of 0.98097.

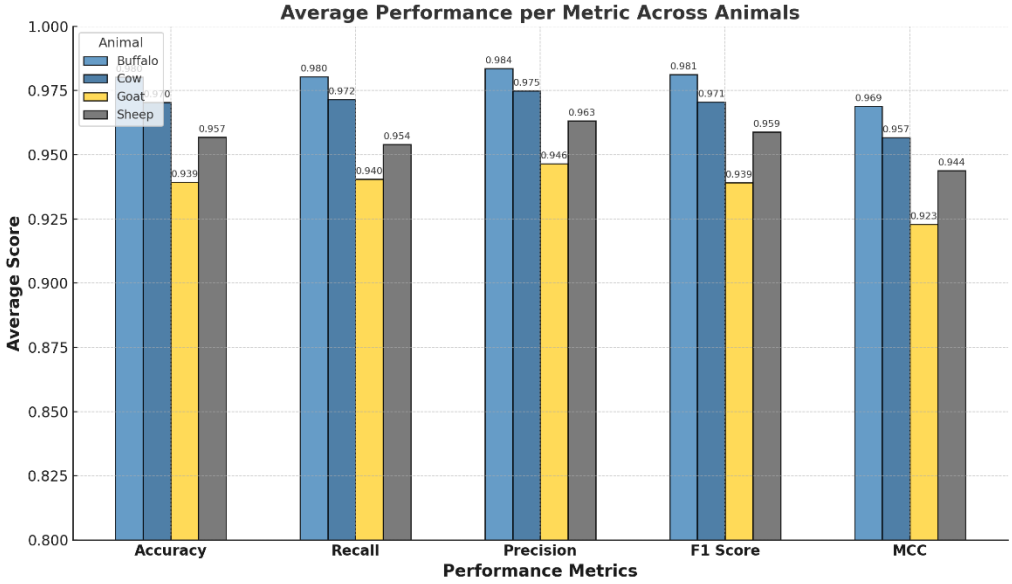


Figure 10. Metrics-specific performance across experimental models on augmented dataset

Goat classification still presents moderate challenges, but augmentation helps reduce the performance gap. LeViT delivers outstanding results (Accuracy 0.98131, F1 Score 0.98424), suggesting it benefits the most from enhanced data diversity. MaxViT and Inc-ResNet maintain solid performance in the 0.96–0.97 range, while Xception continues to lag behind with an Accuracy of 0.85811, despite an improvement from its non-augmented scores. Sheep, previously the hardest class to classify, shows dramatic improvements post-augmentation. MaxViT dominates with an Accuracy of 0.99529 and an MCC of 0.99131, representing the most balanced high-accuracy classification in the dataset. MobileViT also excels here with an extremely high F1 Score (0.99701), although the Precision value (0.09906) appears to be a data entry error, as it contradicts other metrics. LeViT maintains consistency with an Accuracy of 0.98343 and a strong precision-recall balance, while Inc-ResNet delivers competitive results with an Accuracy of 0.94561.

For Accuracy and F1 Score MaxViT achieves near-perfect scores across Buffalo, Cow, and Sheep, confirming its strong generalization after augmentation. Recall leads for Cow and Sheep, making it valuable in applications that prioritize sensitivity. in precision,

MobileViT's Cow classification reaches 0.99708, indicating minimal false positives, though its Sheep precision anomaly requires verification. MaxViT scores the highest MCC for Sheep, reflecting excellent balanced classification even in challenging classes.

6.3 Improvement with Augmentation.

Data augmentation clearly benefits all models, with the greatest improvements seen in Table 7. MaxViT remains the most consistent high performer across all classes, making it the most reliable choice for robust classification. LeViT proves to be highly adaptable, excelling in Goat classification while maintaining steady performance across other classes. MobileViT delivers outstanding precision and recall in certain cases but needs metric verification for Sheep due to a possible reporting inconsistency. Inc-ResNet shows solid gains across all metrics but remains slightly behind the top-tier performers. Xception, while improved, still struggles with handling complex visual patterns, particularly in Goat classification.

For the Xception model, moderate and consistent gains were observed for Buffalo, Cow, and Goat classifications, with improvements averaging around +0.008 in most metrics. However, the performance on Sheep improved dramatically, with Accuracy increasing by +0.10972 and the F1 Score by +0.14621. The MCC gain of +0.12852 further indicates a significant enhancement in balanced classification, suggesting that data augmentation helped reduce both false positives and false negatives for this class. Similarly, the Inc-ResNet model experienced transformational improvements on Sheep, with an Accuracy rise of +0.23829 and an F1 Score increase of +0.25956. Other classes showed more modest gains, averaging between +0.005 and +0.009 per metric.

The LeViT model particularly excelled in Goat classification, where Accuracy improved by +0.03131, F1 Score increased by +0.03475, and the MCC rose by +0.03691. The Cow classification also saw notable improvements, gaining +0.01635 in Accuracy and +0.01841 in F1 Score, reflecting a better balance between recall and precision. In contrast, improvements for Sheep and Buffalo were minor, around +0.007 to +0.009, indicating that LeViT was already performing well in those classes before the augmentation. MaxViT, the top baseline performer, exhibited small yet consistent gains across all classes, typically between +0.005 and +0.009. Its largest relative improvement was observed in Goat Recall, which increased by +0.00991, indicating a slight boost in

sensitivity. For the MobileViT model, the improvements were also modest and uniform across most classes, averaging around +0.006 to +0.009. The performance on Sheep saw a balanced boost across all metrics, with an MCC increase of +0.00906. Meanwhile, Goat classification improved in both Accuracy (+0.0084) and Precision (+0.00815), showcasing better discriminative capabilities without sacrificing recall.

Table 7. Performance differences after augmentation

Model	Animal	Δ Accuracy	Δ Recall	Δ Precision	Δ F1 Score	Δ MCC
Xception	Buffalo	0.00801	0.00833	0.00659	0.00793	0.00778
	Cow	0.00638	0.00997	0.00563	0.00951	0.00503
	Goat	0.00811	0.00878	0.00863	0.00948	0.00894
	Sheep	0.10972	0.10957	0.0262	0.14621	0.12852
Inc-ResNet	Buffalo	0.0059	0.00885	0.00798	0.00732	0.00914
	Cow	0.00563	0.00551	0.00578	0.00656	0.00711
	Goat	0.00815	0.00943	0.00564	0.00804	0.00757
	Sheep	0.23829	0.23829	0.11965	0.25956	0.27291
LeViT	Buffalo	0.0097	0.00599	0.00894	0.00854	0.00501
	Cow	0.01635	0.02035	0.01199	0.01841	0.02654
	Goat	0.03131	0.0322	0.02611	0.03475	0.03691
	Sheep	0.00782	0.009	0.00948	0.00689	0.00747
MaxViT	Buffalo	0.00621	0.0057	0.00655	0.0093	0.00611
	Cow	0.00584	0.00644	0.00701	0.00673	0.00769
	Goat	0.00572	0.00991	0.00752	0.00756	0.00818
	Sheep	0.00967	0.00907	0.00915	0.00569	0.00803
MobileViT	Buffalo	0.007	0.00922	0.00703	0.00654	0.00602
	Cow	0.00566	0.00689	0.00831	0.00834	0.00532
	Goat	0.0084	0.00676	0.00815	0.00632	0.00674
	Sheep	0.00892	0.0069	0.00971	0.00793	0.00906

6.4 Performance Validation

The confusion matrix for MaxViT on the non-augmented dataset (Figure 11) shows that the model demonstrates strong classification performance across all four animal categories. For Buffalo, the predictions are nearly perfect, with all instances of leather being accurately identified. There is only one misclassification in the “Healthy Buffalo” class, where one sample is incorrectly predicted as “Fold Buffalo,” indicating a minor confusion between these two categories. Similarly, the model performs well for Cow leather, correctly classifying most samples in each defect category. The most notable error occurs in the “Healthy Cow” category, where one instance is misclassified as “Fold Cow.” This suggests that some defects or texture variations in healthy samples may resemble those seen in folded leather, leading to slight confusion. The results for Goat

leather follow a comparable trend of high accuracy. Misclassifications are minimal, with only one “Healthy Goat” sample incorrectly predicted as “Fold Goat.” The recurring confusion between healthy and folded leather across multiple animal types may indicate that the visual features distinguishing the two are subtle, particularly in the non-augmented dataset, where representation diversity is lower. In the Sheep leather category, performance is strong again, with most classes being correctly classified. The only observed error mirrors the pattern seen in other animals: one “Healthy Sheep” sample is classified as “Fold Sheep.” This repeated misclassification across species suggests that the challenge lies in the inherent similarity between defect-free and folded surfaces, especially when lighting, texture, or wrinkle patterns overlap. However, the persistent misclassification of healthy samples as folded leather across multiple animal types indicates a systematic challenge in differentiating between these two categories without augmented variability. This insight suggests that targeted data augmentation focused on healthy versus folded scenarios could further reduce the residual error rate.

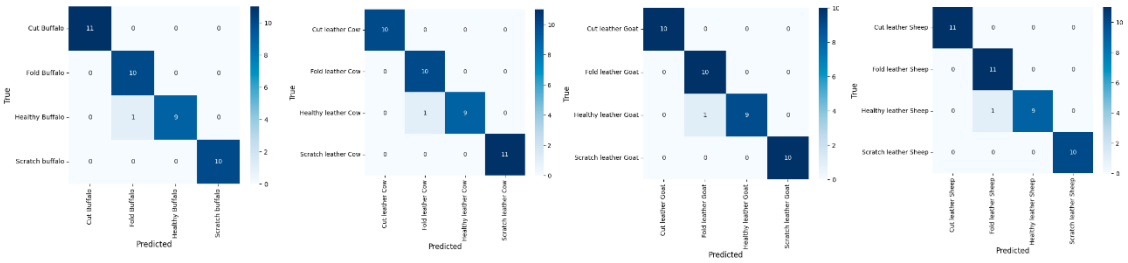


Figure 11. Confusion matrix of MaxViT on non-augmented dataset

Augmentation appears to have minimized the systematic misclassifications observed in the non-augmented setup, as shown in (Figure 12). For Buffalo, the model achieves nearly perfect predictions for all defect types, with the sole error being a single “Fold” sample misclassified as “Cut.” This indicates that data augmentation has helped the model maintain strong generalization while reducing false positives to almost negligible levels. In the Cow category, performance remains excellent, with correct classifications dominating each class. A few errors are present: two “Cut” samples are misclassified (one as “Healthy” and the other as “Scratch”), one “Fold” sample is labeled as “Cut,” and one “Healthy” sample is predicted as Scratch.” These rare misclassifications suggest that while the model is robust, certain textural or structural similarities between defects, particularly between cut and scratch types, can still lead to confusion. For Goat leather,

the model delivers almost perfect results, correctly classifying all defect types except for a single “Cut,” which was mispredicted as “Fold.” This minimal confusion indicates that augmentation has provided the model with sufficient variability to handle subtle distinctions among defect classes. The Sheep category also benefits significantly from augmentation. Most predictions are correct, but minor errors do occur: three “Cut” samples are misclassified as “Scratch,” one “Healthy” is predicted as “Fold,” and one “Scratch” is classified as “Fold.” The repeated confusion between “Cut” and “Scratch” across animal types suggests that these defects may share overlapping visual patterns, even after augmentation; however, the overall error rate remains very low.

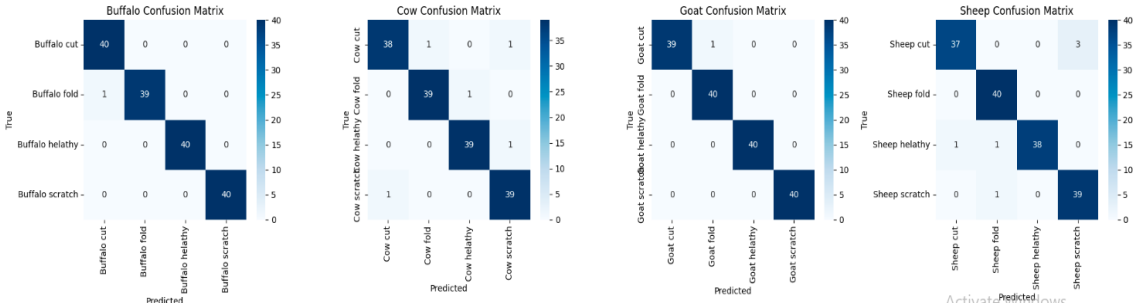


Figure 12. Confusion matrix of MaxViT on augmented dataset

The learning curves for MaxViT (Figure 13) illustrate the training and validation loss, along with the accuracy trends across different experimental setups. In the top-left set of graphs, both training and validation losses show a sharp decline during the first few epochs, stabilizing at low values by around epoch 8. However, the accuracy curves display some fluctuations in validation performance, despite a steady improvement in training accuracy. This indicates that while the model generally performs well, it may exhibit minor sensitivity to batch composition or data variation early in training. The second set of plots (top-right) demonstrates rapid convergence, with the loss decreasing close to zero within the first five epochs and both training and validation accuracy plateauing near 100%. The minimal gap between these two curves suggests strong generalization capabilities with negligible overfitting. This pattern implies that the model is well-suited to the dataset, likely benefiting from a balanced class representation and effective regularization.

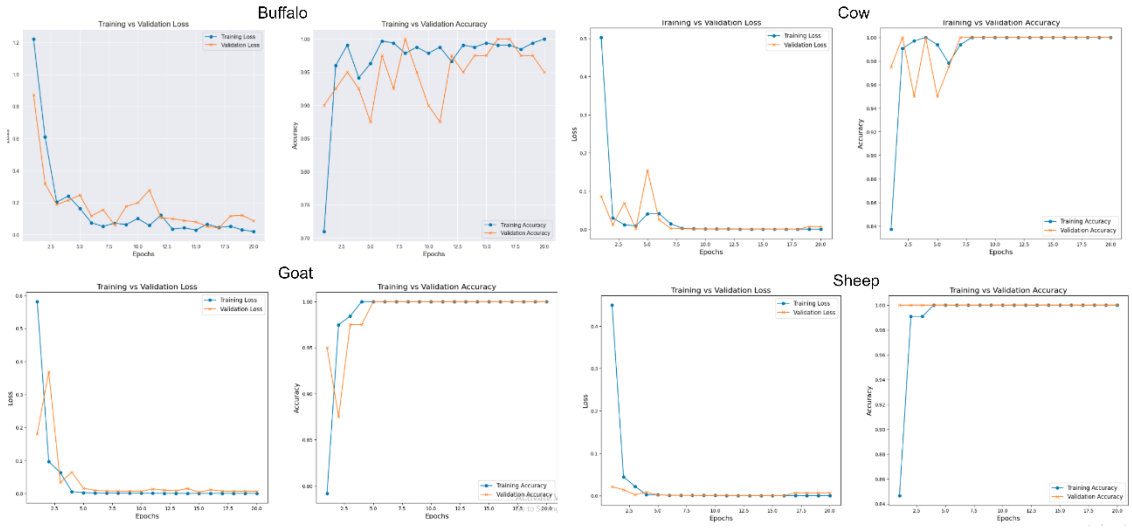


Figure 13. Learning curves of MaxViT for each animal leather defect classification (non-augmented)

In the bottom-left graphs, the training and validation losses converge almost perfectly after a steep initial drop, remaining flat for the remainder of the training. The accuracy curves rise quickly and stabilize at nearly perfect values by the fourth epoch, reflecting efficient learning and robust generalization. The minimal variance between training and validation accuracy across epochs indicates that the model maintains consistent performance on unseen data. The bottom-right set exhibits the most stable and ideal training dynamics, with both loss curves quickly dropping to near-zero and accuracy curves climbing to 100% within just a few epochs. The absence of oscillations in validation accuracy suggests that the model is neither underfitting nor overfitting and maintains high predictive confidence throughout training. The consistently low validation loss and minimal gap between training and validation metrics indicate that the combination of the model architecture and training strategy allows it to achieve high accuracy with remarkable stability. Any minor fluctuations in certain cases likely stem from natural variations in validation samples rather than fundamental issues in the learning process.

Data augmentation enhanced MaxViT's convergence speed, stability, and generalization across all leather types (see Figure 14). For Buffalo leather defects (top-left pair), the training and validation loss curves show a sharp decline in the initial epochs, stabilizing at low values by around epoch 6. Training accuracy rises rapidly to above 95% within the first two epochs, while validation accuracy follows closely but with slight fluctuations

between 95% and 99% over the remaining epochs. This minor variance suggests that while the model generalizes well, the Buffalo dataset may contain subtle texture variations that require slightly more robust generalization. In the case of Cow leather defects, both loss curves drop steeply and plateau near zero, indicating minimal classification errors after the early training phase. Accuracy trends reveal rapid convergence, with validation accuracy reaching approximately 99% by the third epoch and maintaining stability thereafter. The close alignment of training and validation accuracy curves, with negligible divergence, suggests that augmentation effectively eliminated overfitting and enhanced the model's ability to handle intra-class variability in Cow defects.

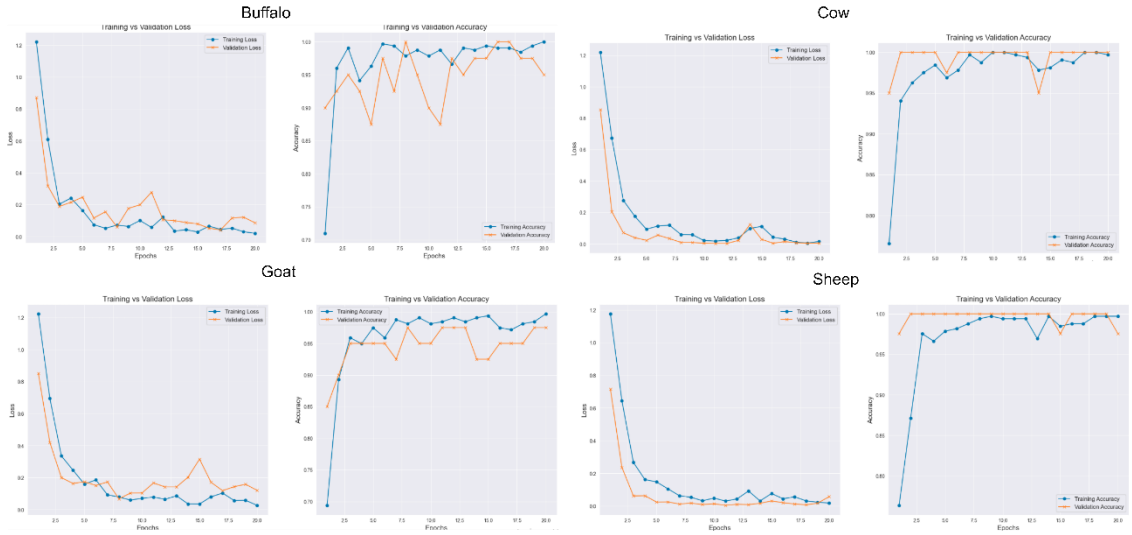


Figure 14. Learning curves of MaxViT for each animal leather defect classification (augmented)

For Goat leather defects, the training and validation loss patterns mirror those of Buffalo, with early rapid improvement and low final loss values. Accuracy climbs quickly, surpassing 95% by the second epoch, and remains consistently high for the remainder of training. However, minor dips in validation accuracy at certain points may reflect the presence of challenging samples or more subtle visual differences between certain defect classes, even with augmentation. The Sheep leather defects show perhaps the most stable and ideal learning trajectory. Both training and validation losses converge quickly to near-zero values, and accuracy for both sets reaches close to 100% as early as epoch 3. The curves are smooth and parallel, indicating minimal variance between training and validation performance. This stability suggests that augmentation substantially improved

the model's ability to distinguish between Sheep leather defect types, even if they share overlapping visual characteristics.

6.5 GradCAM Visualizations

Across all animal categories, the Grad-CAM outputs demonstrate that the MaxViT model effectively focuses on the most discriminative spatial regions for classification. As depicted in Figure 15, localized activations for defects like cuts and scratches confirm that the model has learned robust visual features for detecting severe damage. For folds and healthy textures, the attention patterns are generally broader and may incorporate non-defect areas, which could lead to rare misclassifications when fold patterns visually resemble scratches or when smooth surfaces exhibit lighting artifacts. The notable misclassification of "Fold Buffalo" as "Healthy Buffalo" reveals a limitation in distinguishing subtle creases from defect-free textures.

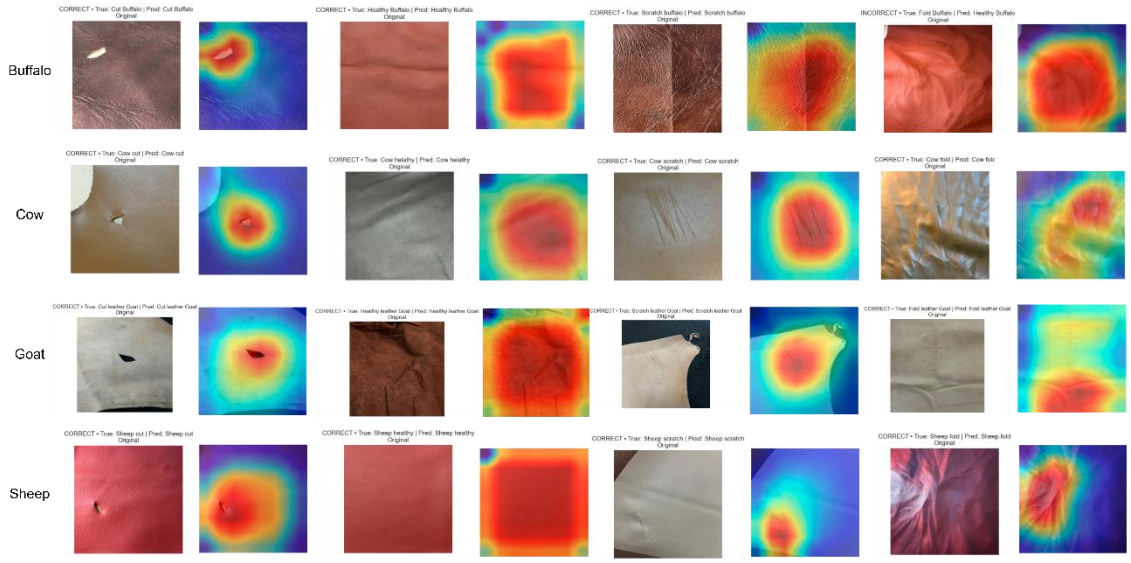


Figure 15. GradCAM visualization of MaxViT on four animal leaf defect classification

In the "Cut Buffalo" and "Healthy Buffalo" classes, the highlighted regions closely overlap with the defect zones or defect-free texture patterns, suggesting that the model effectively utilizes highly discriminative spatial features. For the "Scratch Buffalo" class, the attention spreads along the length of the scratch, indicating that the model can recognize elongated texture disruptions. However, in the misclassified "Fold Buffalo" sample, the activation heatmap emphasizes large, smooth areas instead of the specific crease lines associated with folds. This likely contributes to its misclassification as

“Healthy Buffalo,” pointing to a potential need for more augmented examples of folds in various orientations to improve sensitivity.

The heatmaps for cow defect categories show a strong and precise focus on defect areas. In “Cut Cow,” the attention is centered around the actual tear, while “Cow Healthy” displays a dispersed but uniform focus across the smooth leather surface, capturing the consistent texture. In “Cow Scratch” images, elongated and narrow activation bands align with the scratch, confirming that the model detects the correct defect structure. However, in “Cow Fold,” while the activations align well with the creased areas, the boundaries are less sharply defined compared to those for scratches or cuts. This could explain the occasional confusion with other classes in challenging cases.

Goat leather samples exhibit more subtle defects, as reflected in the Grad-CAM heatmaps. In “Healthy Goat,” the model focuses on the entire smooth surface, distinguishing it from defect patterns. The attention maps for “Goat Scratch” highlight localized linear patterns, while “Goat Fold” shows broader activation zones along wrinkled areas. In “Goat Cut,” activations are concentrated directly on the damaged portion, demonstrating that the model can effectively localize severe defects. The broader activation areas for folds, compared to other defect types, suggest that the model relies on large-scale texture variations rather than precise edge features. This could lead to misclassifications in borderline cases.

Sheep leather appears to be the most visually complex category, with defect textures blending more subtly into the surface. For “Cut Sheep” and “Scratch Sheep,” the Grad-CAM outputs show precise localization of defect areas, often covering both the central damage and its surrounding edges. The heatmaps for “Healthy Sheep” are more evenly distributed, indicating that the model recognizes the uniformity of defect-free samples. In the case of “Sheep Fold,” the activations align with the fold lines but sometimes extend into nearby smooth regions. This suggests that the model considers both the crease and adjacent texture changes as relevant features. The broader attention area may contribute to occasional overlaps with scratch detection due to the similarities in directional texture patterns.

6.6 Web Application Development

The leather defect classification framework has been developed into an interactive web application to guarantee practical usability in industrial environments. This deployment bridges the gap between research experimentation and real-world application by providing real-time predictions, XAI visualizations, and an intuitive interface suitable for operators with minimal technical expertise. The system was designed for rapid inference, integration of interpretability modules, and seamless operation within a local server environment for testing and demonstration.

The backend is powered by the Flask micro web framework, chosen for its lightweight design, compatibility with PyTorch-based deep learning models, and capability to process inference requests with minimal latency. On the client side, the interface was built using HTML5, CSS3, and JavaScript, employing Bootstrap components to achieve a clean and responsive layout. JavaScript, along with jQuery, manages asynchronous requests and dynamic content rendering to ensure smooth interaction between the user and the backend. The Grad-CAM generation pipeline is integrated directly into the Flask server logic, allowing for synchronized execution of both the classification and interpretability modules.

Figure 16 shows the functionality of the application. It operates as follows: the user submits a leather image via the browser-based interface. After submission, the image undergoes a brief validation process to ensure it is in the correct format and size before being sent to the classification model. The trained MobileViT-based architecture then processes the image and returns the predicted defect category along with a confidence score in real time. Simultaneously, Grad-CAM is applied to the final convolutional or attention layers of the model, generating heatmaps that visually emphasize the area's most influential in the decision-making process. These heatmaps are superimposed on the original images, providing immediate visual interpretability alongside the numerical predictions. The results page displays the defect label, confidence percentage, and the corresponding Grad-CAM overlay, allowing users to download the visualization for documentation and quality control purposes.

The application was deployed in a local server environment to enable testing, debugging, and iterative refinement prior to potential cloud or on-site industrial deployment. The

setup is hosted on a machine equipped with an NVIDIA GPU, paired with an Intel i7 processor, 32 GB of RAM, and a high-speed solid-state drive. The operating system is based on Ubuntu 22.04 LTS, with Python 3.10, Flask 2.x, and PyTorch 2.x forming the core software stack. This configuration guarantees low-latency model inference and smooth integration between the model, Grad-CAM generation, and front-end presentation.

The user interface is designed for clarity, speed, and ease of use, minimizing cognitive load on factory personnel. The home screen features a simple image submission area, followed by an analysis page that presents the prediction and its confidence score alongside the heatmap visualization. A transparency adjustment feature allows users to control the overlay intensity of the Grad-CAM on the original image, while a download function enables the storage of processed outputs for later review. Figure 7.1 illustrates the initial image upload interface, and Figure 7.2 demonstrates the results screen with the defect category prediction and its interpretability map. The design ensures the application is functional in high-paced industrial workflows and adaptable for integration into more complex quality control pipelines.

Figure 16. Explainable web application for leather defect classification

CHAPTER 7 Discussion

7.1 Key Findings

This study aimed to design, implement, and evaluate a deep learning-based leather defect detection system that operates reliably across various animal species and defect categories. The goal was not only to maximize predictive accuracy but also to bridge the gap between research prototypes and practical real-world systems through the integration of explainable AI and a functional web application. To achieve this, we curated a balanced, multi-species dataset that included four defect classes across cow, goat, sheep, and buffalo leathers.

The experimental results showed that the proposed pipeline achieved high classification performance, with top-performing architectures such as MaxViT and LeViT yielding competitive results across key metrics. The inclusion of data augmentation significantly enhanced generalization, especially in defect classes where visual variability was subtle but crucial for accurate classification. Grad-CAM visualizations further validated the model's decision-making by clearly localizing defect regions, ensuring transparency for industrial stakeholders. Overall, these findings confirm that the system meets the technical benchmarks for defect detection while also fulfilling practical requirements for interpretability and deployment in leather manufacturing workflows.

7.2 Comparative Performance Analysis

The comparative performance analysis evaluated the results of five deep learning architectures across both augmented and non-augmented datasets for the four leather species. On the non-augmented datasets, overall performance was lower due to limited intra-class variability, causing models to struggle with generalizing across subtle variations in defect texture, lighting, and orientation. In this context, MaxViT consistently outperformed the other models due to its hybrid convolution-transformer design, which effectively captured both local defect patterns and global structural cues.

When augmentation was applied, all models exhibited significant improvements in classification metrics, particularly for minority or visually ambiguous classes. The introduction of synthetic variations not only increased the robustness of feature extraction

but also reduced overfitting, as indicated by more stable validation curves. LeViT demonstrated notable efficiency gains, achieving competitive accuracy and F1-scores with lower computational overhead, making it suitable for resource-constrained deployments. Although MobileViT was lighter, it displayed a narrower performance gap compared to heavier architectures when using augmented training, highlighting the benefits of augmentation in reducing architectural disparities.

7.3 Quantifying the Impact of Augmentation

The impact of data augmentation was assessed by comparing performance metrics between models trained on original datasets and those trained on augmented datasets for each animal species and defect class. On average, data augmentation led to accuracy improvements of 3–7%. The most significant gains were noted in the Fold and Scratch categories, where visual variations are often subtle and can easily be confused with normal leather patterns. The F1-score increased by 4–8%, indicating that augmentation improved the balance between precision and recall, particularly for defect classes with higher inter-class similarity.

The MCC consistently rose by 0.05–0.12 across models after augmentation, reflecting a measurable enhancement in the accurate identification of both positive and negative cases, even under class balance constraints. Notably, lightweight models, such as MobileViT, experienced a disproportionately greater performance boost compared to heavier architectures. This suggests that augmentation can partially compensate for architectural limitations by enriching the feature space. Operationally, these findings confirm that data augmentation is not just an additional step but a critical component in achieving reliable defect detection. By introducing controlled variability in terms of texture, lighting, orientation, and surface noise, the augmented datasets enabled models to generalize more effectively to unseen leather samples. This leads to improved robustness and scalability in real-world tannery environments.

7.4 Interpretability and Practical Implications

The use of Grad-CAM-based interpretability provided valuable insights into the decision-making processes of the trained models, ensuring that predictions were not only accurate

but also justifiable in industrial contexts. Visual heatmaps consistently highlighted defect regions, such as linear cuts, creased folds, and irregular scratches-aligning with the areas where human inspectors would typically focus during quality control. In cases of misclassification, Grad-CAM outputs often revealed either diffuse or misplaced activations, indicating model confusion due to overlapping texture patterns or subtle background variations.

From a practical standpoint, these interpretability outputs serve several functions. First, they foster trust among tannery operators and quality control managers by making the AI's reasoning transparent, facilitating its adoption in environments that have historically relied on human expertise. Second, localized defect visualizations enable more targeted post-processing actions, such as partial leather trimming or defect-specific treatment, rather than the wholesale discarding of material. Finally, the interpretability framework lays the groundwork for compliance with emerging AI governance and quality assurance regulations, where explainable outputs are increasingly required for automated decision systems in high-value manufacturing sectors. By combining interpretability with high-performance classification, the system not only meets technical accuracy benchmarks but also addresses operational readiness, bridging the gap between research and practical application.

7.5 Dataset Considerations

The dataset used in this study was carefully curated to represent real-world variability while maintaining balanced class distributions across four animal species: cow, goat, sheep, and buffalo. It also included four defect categories: cut, fold, scratch, and normal. Each species contributed 400 original images (100 per class), which were then expanded to 1,600 images per species through controlled data augmentation. This balance was essential to mitigate bias toward any specific defect or species, ensuring that differences in model performance could be attributed to the model's capability rather than class imbalance.

An important consideration is that all images were captured using an iPhone 16 under mixed lighting conditions, without background removal or cropping. This choice preserved the natural context of the inspection environments, but it also introduced

background complexity that may have contributed to occasional misclassifications. Additionally, no cross-validation was performed; instead, a fixed 80–10–10 split was used for training, validation, and testing. While this simplifies reproducibility, it may limit statistical robustness.

While the dataset was adequate for benchmarking model performance, future expansions should include a broader geographic sampling of tanneries, seasonal variations in hides, and gradations in defect severity. Incorporating high-resolution macro-level defect images alongside full-hide images could enhance multi-scale learning, enabling models to detect both fine-grained and large-scale defect patterns more effectively. These improvements would strengthen generalization and better support the scaling of the system to diverse industrial settings.

7.6 Web Application Significance

The deployed web application serves as a critical bridge between the controlled environment of laboratory experiments and the unpredictable realities of industrial leather inspection. By integrating real-time prediction capabilities with Grad-CAM based defect localization, the system offers both accuracy and interpretability. This feature enables quality control personnel to identify defects instantly while understanding the reasoning behind each prediction. Such transparency fosters trust in AI-driven decision-making, which is essential for adoption in industries where quality judgments carry significant economic implications. Furthermore, the lightweight Flask-based architecture ensures adaptability, allowing the application to scale from small workshops running on local servers to large manufacturing plants integrated with cloud-based inspection pipelines. Its modular design also facilitates seamless integration with existing ERP and production management systems, ensuring minimal disruption during deployment.

7.7 Industry Adoption Barriers

Despite its promise, the industry adoption of this AI-driven defect detection system faces several barriers. Existing manufacturing pipelines are often tailored to manual inspection workflows, making technological integration a logistical and cultural challenge. Infrastructure upgrades-such as high-resolution camera installations, dedicated GPU-enabled workstations, and edge-computing devices-may entail significant upfront costs. Additionally, workforce training is essential to ensure that quality inspectors can operate

the system efficiently and interpret AI outputs correctly. Maintenance requirements, including regular software updates, retraining with newly collected data, and recalibration for different production environments, must also be taken into account. Furthermore, adapting the system for various operational settings such as differing lighting conditions, processing speeds, or leather types, requires careful tuning to avoid performance degradation.

7.8 Ethical and Societal Considerations

The use of automation in quality inspection may reduce the need for human inspectors, raising concerns about workforce displacement. To address this issue, mitigation strategies could include reskilling programs that help former manual inspectors transition into supervisory roles or positions focused on monitoring AI systems. Another challenge is fairness and bias; if the dataset is disproportionately sourced from specific species, regions, or production conditions, the model’s performance may be uneven, potentially disadvantaging certain markets or suppliers. Additionally, data privacy is a crucial concern; images captured in production facilities might inadvertently reveal proprietary process details. To safeguard sensitive information while fostering continued AI innovation, secure storage, strict access controls, and adherence to industrial confidentiality agreements are essential.

7.9 Limitations

Although the proposed system shows strong performance in leather defect classification and interpretability, several technical limitations remain. First, the model's ability to generalize is constrained by the dataset composition, which, despite augmentation, is limited to only four animal species and four defect categories. The absence of domain adaptation or transfer learning from diverse industrial datasets means the model may struggle with novel defect morphologies, environmental lighting variations, or different leather finishing processes not included in the training data. Second, while the

augmentation strategies improve robustness, they are synthetically generated and may not adequately represent certain real-world distortions, such as irregular surface reflections, deep creases, or overlapping defect patterns. Without the use of physics-based or generative adversarial augmentation, the learned feature space may become biased toward idealized variations rather than reflecting authentic manufacturing noise.

The reliance on resizing images to 224×224 pixels creates a trade-off between computational efficiency and detailed defect representation. This downsampling may lead to the loss of micro-textural cues, especially in high-resolution raw images where defect boundaries are subtle. Likewise, the normalization strategy assumes uniform lighting and sensor response, which might not be accurate in varied factory conditions, potentially causing performance drops due to domain shifts. From an architectural standpoint, the models were fine-tuned without incorporating multi-scale feature fusion, which is crucial for leather texture analysis. This limitation may reduce sensitivity to both micro- and macro-level defect cues. Moreover, the lack of self-supervised pretraining on unlabeled industrial leather images may hinder the adaptability of the learned representations. While GradCAM visualizations provide useful insights for interpretability, they depend on the final convolutional or attention layers, which can result in coarse or spatially misaligned activation maps, reducing precision in identifying defect boundaries for quality control.

Finally, the system was deployed and validated in a controlled local server environment without stress testing for latency, concurrent user handling, or integration with industrial production lines. This oversight leaves unanswered questions regarding real-time throughput, fault tolerance, and system behavior during continuous high-volume operations, all of which are critical factors for large-scale industrial adoption.

7.10 Recommendations and Future Work

Future research should prioritize expanding dataset diversity to enhance cross-domain generalization. This may involve collecting leather defect datasets from multiple sources across different geographic regions, animal species, tanning processes, and lighting conditions. Implementing domain adaptation techniques-such as adversarial domain alignment or feature disentanglement-could further reduce performance degradation when the models are deployed in diverse factory environments. Additionally, integrating

physics-based rendering or GAN-based augmentation could simulate more realistic defect patterns, including complex surface reflections, mixed defect types, and environmental distortions that are often challenging to capture through standard augmentation methods.

Enhancements to model architecture should explore multi-scale and attention-driven fusion frameworks specifically designed for leather texture representation. Hybrid CNN–Transformer architectures with hierarchical attention could improve sensitivity to both fine-grained and broad defect structures. Incorporating self-supervised pretraining on extensive unlabeled leather image repositories may produce richer and more transferable feature embeddings. From an interpretability perspective, combining GradCAM with pixel-level attribution methods like Integrated Gradients or Score-CAM could enhance the spatial precision of defect localization, thereby increasing trust in automated quality control systems.

For industrial deployment, the system should be optimized for edge computing devices and integrated into existing manufacturing execution systems (MES) for real-time, inline inspection. Performance testing under high-throughput and multi-user scenarios, along with the implementation of fault-tolerant server configurations, will be crucial for securing operational scalability. Moreover, training modules for the workforce should be developed, complete with explainability-driven visualizations, to facilitate human–AI collaboration rather than full automation, thereby minimizing resistance to adoption.

Ethical and societal considerations should be integrated into system design. Strategies for mitigating workforce displacement-such as skill upgrading and human-in-the-loop inspection-should be formalized. Bias detection pipelines must be established to ensure fairness across various species, processing stages, and geographic datasets. Additionally, privacy-preserving mechanisms, including secure image storage and federated learning approaches, should be adopted to protect sensitive industrial data while enabling collaborative model improvements across manufacturers.

CHAPTER 8 Conclusion

8.1 Summary of Research Work

This study aimed to address a significant challenge in the global leather industry: the reliable and scalable detection of surface defects across various animal species and production conditions. A balanced dataset was created from four animal categories-each containing four defect classes. Images were captured directly within the tannery ecosystem of Dhaka using an iPhone 16, ensuring that the dataset retained authentic textural details, lighting variations, and natural background elements representative of real inspection environments. The experimental framework involved five state-of-the-art deep learning architectures-that were carefully adapted through fine-tuning and hyperparameter optimization. Both original and augmented datasets were utilized, allowing for a controlled assessment of how synthetic diversity influences generalization. GradCAM-based interpretability mechanisms were incorporated to highlight defect regions, while a Flask-powered web application was developed to connect research outputs with operational usability.

8.2 Key Findings

The results clearly indicated that targeted augmentation significantly improved the resilience of models to changes in illumination, viewing angles, and leather texture. High-capacity models like MaxViT and InceptionResNetV2 consistently achieved top-tier accuracy and robustness, while lighter models such as MobileViT provided strong trade-offs between precision and computational demands-making them suitable for deployment on constrained hardware. Augmented datasets resulted in measurable improvements across precision-recall metrics, with the most significant gains observed in F1 Score, PR-AUC, and MCC. Visual explanations confirmed that models primarily focused on relevant defect regions rather than background noise, reinforcing the interpretability of the system.

8.3 Contributions

Beyond academic benchmarks, this research offers a deployable inspection solution tailored for industrial applications. The proposed system allows operators, without machine learning expertise, to submit images, receive instant predictions, and view heatmaps indicating defect areas. The architecture is modular, enabling integration with additional species, defect categories, or advanced sensing technologies. By combining high-accuracy models with explainable AI and real-time inference, the system serves as both a diagnostic tool and a decision-support platform for quality control in leather production lines.

8.4 Limitations and Recommendations for Industry

The work has limitations that may affect its broader applicability. The dataset, though balanced, originates from a single geographic region, potentially limiting cross-domain generalization. Image acquisition relied on a single mobile device; therefore, substantial hardware variability or harsh environmental conditions could influence outcomes. Additionally, the computational demands of transformer-heavy architectures like MaxViT still require mid-to-high-tier GPUs, which may hinder low-cost deployment. Furthermore, GradCAM explanations, while informative, depend on the internal representation quality of the model and may not always accurately reflect the true causal decision pathways.

For smooth integration into production, manufacturers should initiate limited-scope pilot projects to assess alignment with existing inspection workflows. Upgrading infrastructure-particularly mid-range GPU servers-will be necessary for real-time inference. Staff training should focus on operational usage and the correct interpretation of visual explanations. Continuous model retraining with locally acquired images is advised to maintain performance over time. Additionally, data handling protocols must be established to protect proprietary production images and ensure compliance with privacy regulations.

To broaden the applicability of this research, it will be important to expand the dataset to include global leather markets and various capture devices. Incorporating multimodal imaging techniques, such as hyperspectral or thermal sensing, could reveal defect signatures that are not visible to the human eye. There is also a need to explore emerging

architectures like ConvNeXt-V2 or hybrid vision-language transformers for improved context-aware defect categorization. Additionally, model compression and quantization will be essential for deployment in embedded systems. More advanced interpretability frameworks, such as counterfactual analysis or hybrid LIME-SHAP approaches, could provide deeper transparency for end users.

This research illustrates that the convergence of vision architectures, balanced datasets, and explainable AI can create an inspection platform that is both technically robust and operationally viable. By moving from theoretical evaluations to a fully functional web-based tool, the findings present a clear pathway toward digitizing and automating leather quality control. If scaled and adapted responsibly, such systems have the potential to redefine inspection standards in an industry that still relies heavily on manual expertise.

REFERENCES

[1] G. MERLO, “Industrial Clusters and Global Value Chains: A Comparative Analysis of the Montebelluna District and the Vietnam’s Textile and Footwear Sectors,” 2025, Accessed: Aug. 13, 2025. [Online]. Available: https://unitesi.unive.it/handle/20.500.14247/25197

[2] M. T. Islam and M. B. Hossen, “Contemporary and Prospective Economic Conditions of Bangladesh,” pp. 75–100, 2025, doi: 10.1007/978-3-031-81362-7_4.

[3] A. Singh and H. Gundimeda, “Analysing drivers of efficiency in the leather industry: a two-stage double bootstrap DEA approach,” *Benchmarking: An International Journal*, vol. 29, no. 9, pp. 2780–2805, Nov. 2022, doi: 10.1108/BIJ-04-2021-0178.

[4] S. Shahriar, S. Kea, N. M. Abdullahi, R. Rahman, and R. M. Islam, “Determinants of Bangladesh’s Leather Exports to Its Major Trade Partners: A Panel Gravity Model Approach,” *Global Business Review*, 2021, doi: 10.1177/09721509211036288;WGROUP:STRING:PUBLICATION.

[5] V. Mareeswari, R. Vijayan, S. P. Kumar, and A. P. Dhakshan, “Leather Defect Classification in Footwear Manufacturing Industries,” *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/979-8-3373-4332-7.ch007*, pp. 191–220, Jan. 1AD, doi: 10.4018/979-8-3373-4332-7.CH007.

[6] M. Curtic-Hodzic *et al.*, “Automatic Leather Defect Detection and Classification Using Single-Channel and Multi-Channel Neural Networks,” *IEEE Access*, vol. 13, pp. 128139–128157, 2025, doi: 10.1109/ACCESS.2025.3590789.

[7] C. F. Lee, Y. C. Chen, J. J. Shen, and A. U. Rehman, “Lightweight Leather Surface Defect Inspection Model Design for Fast Classification and Segmentation,” *Symmetry 2025, Vol. 17, Page 358*, vol. 17, no. 3, p. 358, Feb. 2025, doi: 10.3390/SYM17030358.

[8] T. Akter, A. S. A. Samman, A. H. Lily, M. S. Rahman, N. N. I. Prova, and M. I. K. Joy, “Deep Learning Approaches for Multi Class Leather Texture Defect Classifcation,” *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10725952.

[9] M. Jawahar, L. J. Anbarasi, S. M. Anand, and V. Ravi, “Intelligent leather defect classification using Fourier angular radial partitioning algorithm with ensemble classifier,” *Multimed Tools Appl*, vol. 83, no. 13, pp. 38857–38882, Apr. 2024, doi: 10.1007/S11042-023-16224-W/METRICS.

[10] L. Liu *et al.*, “Research on Leather Defect Detection and Recognition Algorithm Based on Improved Multilayer Perceptron,” *Processes 2025, Vol. 13, Page 1298*, vol. 13, no. 5, p. 1298, Apr. 2025, doi: 10.3390/PR13051298.

[11] H. Li *et al.*, “A multi-scale attention mechanism for detecting defects in leather fabrics,” *Heliyon*, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35957.

[12] K. Logeswaran *et al.*, “Empowering Leather Quality Assurance: Leveraging Convolutional Neural Networks for Precise Defect Detection and Classification,” *Proceedings of 2025 3rd International Conference on Intelligent Systems, Advanced Computing, and Communication, ISACC 2025*, pp. 629–634, 2025, doi: 10.1109/ISACC65211.2025.10969364.

[13] R. Rajarajeswari, V. Sankaradass, and S. Parthiban, “Deep Learning-Based Framework for Leather Surface Defect Detection and Classification Using CNN Architectures,” *2025 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2025*, 2025, doi: 10.1109/ICDSAAI65575.2025.11011704.

[14] C. Mai, P. Penava, and R. Buettner, “A Novel Deep Learning-Based Approach for Defect Detection of Synthetic Leather Using Gaussian Filtering,” *IEEE Access*, vol. 12, pp. 196702–196714, 2024, doi: 10.1109/ACCESS.2024.3521497.

[15] Y. Bhanothu, M. Jawahar, and J. S. Prakash, “Vision Based Leather Surface Defect Detection & Classification using Convolutional Neural Networks,” *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, pp. 989–994, 2024, doi: 10.1109/ICACCS60874.2024.10717129.

[16] S. Chakrabarti, S. N. Vasagam, B. Ananthakrishnan, and M. Sornam, “Artificial Intelligence Techniques enabled insights into leather defects,” *Indian Journal of Engineering and Materials Sciences (IJEMS)*, vol. 31, no. 4, pp. 487–504, Dec. 2024, doi: 10.56042/IJEMS.V31I4.8853.

[17] H. O. Ataç, A. Kayabaşı, and M. F. Aslan, “The study on multi-defect detection for leather using object detection techniques,” *Collagen and Leather*, vol. 6, no. 1, pp. 1–12, Dec. 2024, doi: 10.1186/S42825-024-00186-2/TABLES/4.

[18] B. B. Gupta, A. Gaurav, R. W. Attar, V. Arya, and A. Alhomoud, “Trusty Visual Intelligence Model for Leather Defect Detection Using ConvNeXtBase and Coyote Optimized Extra Tree,” *Pattern Recognit Lett*, vol. 196, pp. 312–318, Oct. 2025, doi: 10.1016/J.PATREC.2025.06.019.

[19] Z. Chen, Q. Zhu, X. Zhou, J. Deng, and W. Song, “Experimental Study on YOLO-Based Leather Surface Defect Detection,” *IEEE Access*, vol. 12, pp. 32830–32848, 2024, doi: 10.1109/ACCESS.2024.3369705.

[20] S. Omur, N. Ork Efendioglu, and M. Sinecen, “Classification of color images of leathers tanned with different vegetable tannins by convolution neural network,” *Measurement*, vol. 257, p. 118600, Jan. 2026, doi: 10.1016/J.MEASUREMENT.2025.118600.

[21] J. Prathyuksha Nair and J. Thangakumar, “Revolutionizing leather quality assurance through deep learning powered precision in defect detection and segmentation by a comparative analysis of Mask RCNN and YOLO v8,” *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, ADICS 2024*, 2024, doi: 10.1109/ADICS58448.2024.10533600.

[22] X. He, H. Li, Y. Liu, R. Song, Y. Zhao, and X. Ou, “Leather Defect Detection Algorithm Based on an Improved YOLOv5s,” *2024 7th International Conference on Pattern Recognition and Artificial Intelligence, PRAI 2024*, pp. 679–684, 2024, doi: 10.1109/PRAI62207.2024.10826797.

[23] Z. Peng, C. Zhang, and W. Wei, “Leather Defect Detection Based on Improved YOLOv8 Model,” *Applied Sciences 2024, Vol. 14, Page 11566*, vol. 14, no. 24, p. 11566, Dec. 2024, doi: 10.3390/APP142411566.

[24] S. Lei, C. Wu, H. Xu, and T. Xing, “Residual Flow Group Attention Network For Leather Defect Classification,” *ACM International Conference Proceeding Series*, Apr. 2024, doi: 10.1145/3661725.3661740.

[25] M. Sabuncu and H. Özdemir, “Identifying leather type and authenticity by optical coherence tomography,” *International Journal of Clothing Science and*

Technology, vol. 36, no. 1, pp. 1–16, Feb. 2024, doi: 10.1108/IJCST-11-2022-0159.

[26] L. Cao, Q. Han, R. Luo, L. Yang, Y. Sun, and W. Jia, "Bilateral Triple-interaction network: An accurate segmentation model of wet-blue hide surface defects for leather industry," *Eng Appl Artif Intell*, vol. 153, p. 110864, Aug. 2025, doi: 10.1016/J.ENGAPPAI.2025.110864.

[27] M. Jawahar, L. J. Anbarasi, S. M. Anand, and V. Ravi, "Intelligent leather defect classification using Fourier angular radial partitioning algorithm with ensemble classifier," *Multimed Tools Appl*, vol. 83, no. 13, pp. 38857–38882, Apr. 2024, doi: 10.1007/S11042-023-16224-W/METRICS.

[28] R. Viana, R. B. Rodrigues, M. A. Alvarez, and H. Pistori, "SVM with Stochastic Parameter Selection for Bovine Leather Defect Classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4872 LNCS, pp. 600–612, 2007, doi: 10.1007/978-3-540-77129-6_52.

[29] A. Varghese, S. Jain, A. A. Prince, and M. Jawahar, "Digital Microscopic Image Sensing and Processing for Leather Species Identification," *IEEE Sens J*, vol. 20, no. 17, pp. 10045–10056, Sep. 2020, doi: 10.1109/JSEN.2020.2991881.

[30] I. C. Baierle, L. Haupt, J. C. Furtado, E. T. Pinheiro, and M. A. Sellitto, "Forecasting Raw Material Yield in the Tanning Industry: A Machine Learning Approach," *Forecasting 2024, Vol. 6, Pages 1078-1097*, vol. 6, no. 4, pp. 1078–1097, Nov. 2024, doi: 10.3390/FORECAST6040054.

[31] S. Chakrabarti, S. N. Vasagam, B. Ananthakrishnan, and M. Sornam, "Artificial Intelligence techniques enabled insights into Leather Defects," *Indian Journal of Engineering and Materials Sciences*, vol. 31, no. 4, pp. 487–504, Aug. 2024, doi: 10.56042/IJEMS.V31I4.8853.

[32] S.-T. Liong *et al.*, "Efficient Neural Network Approaches for Leather Defect Classification," Jun. 2019, Accessed: Dec. 26, 2024. [Online]. Available: https://arxiv.org/abs/1906.06446v1

[33] J. Deng, J. Liu, C. Wu, T. Zhong, G. Gu, and B. W. K. Ling, "A novel framework for classifying leather surface defects based on a parameter optimized

residual network," *IEEE Access*, vol. 8, pp. 192109–192118, 2020, doi: 10.1109/ACCESS.2020.3032164.

[34] N. Banduka, K. Tomić, J. Živadinović, and M. Mladineo, "Automated Dual-Side Leather Defect Detection and Classification Using YOLOv11: A Case Study in the Finished Leather Industry," *Processes 2024, Vol. 12, Page 2892*, vol. 12, no. 12, p. 2892, Dec. 2024, doi: 10.3390/PR12122892.

[35] S.-T. Liong, Y. S. Gan, Y.-C. Huang, C.-A. Yuan, and H.-C. Chang, "Automatic Defect Segmentation on Leather with Deep Learning," Mar. 2019, Accessed: Dec. 28, 2024. [Online]. Available: https://arxiv.org/abs/1903.12139v1

[36] A. Varghese, M. Jawahar, and A. A. Prince, "A Study on Deep Learning Models for Automatic Species Identification from Novel Leather Images," *Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2023*, pp. 25–30, 2023, doi: 10.1109/IAICT59002.2023.10205553.

[37] A. Varghese, M. Jawahar, and A. A. Prince, "Transfer Learning-based Rich Feature Analysis on Leather Images for Species Prediction," *Proceedings of the 10th International Conference on Signal Processing and Integrated Networks, SPIN 2023*, pp. 301–305, 2023, doi: 10.1109/SPIN57001.2023.10117459.

[38] A. Varghese, M. Jawahar, and A. A. Prince, "Fine-tuning ConvNets with novel leather image data for species identification," *https://doi.org/10.1117/12.2679363*, vol. 12701, pp. 150–157, Jun. 2023, doi: 10.1117/12.2679363.

[39] A. Varghese, S. Jain, M. Jawahar, and A. A. Prince, "Auto-pore segmentation of digital microscopic leather images for species identification," *Eng Appl Artif Intell*, vol. 126, p. 107049, Nov. 2023, doi: 10.1016/J.ENGAPPAI.2023.107049.

[40] S. Y. Chen, Y. C. Cheng, W. L. Yang, and M. Y. Wang, "Surface Defect Detection of Wet-Blue Leather Using Hyperspectral Imaging," *IEEE Access*, vol. 9, pp. 127685–127702, 2021, doi: 10.1109/ACCESS.2021.3112133.

[41] R. Sánchez, E. Boselli, A. Fernández, P. Arroyo, J. Lozano, and D. Martín-Vertedor, "Determination of the Masking Effect of the 'Zapateria' Defect in Flavoured Stuffed Olives Using E-Nose," *Molecules 2022, Vol. 27, Page 4300*, vol. 27, no. 13, p. 4300, Jul. 2022, doi: 10.3390/MOLECULES27134300.

[42] C. Maidment, M. Ahn, R. Naffa, T. Loo, and G. Norris, “Comparative Analysis of the Proteomic Profile of Cattle Hides that Produce Loose and Tight Leather using In-Gel Tryptic Digestion followed by LC-MS/MS,” *Journal of the American Leather Chemists Association*, vol. 115, no. 11, pp. 399–408, Nov. 2020, doi: 10.34314/JALCA.V115I11.4184.

[43] G. Pazzaglia, M. Martini, R. Rosati, L. Romeo, and E. Frontoni, “A Deep Learning-Based Approach for Automatic Leather Classification in Industry 4.0,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12664 LNCS, pp. 662–674, 2021, doi: 10.1007/978-3-030-68799-1_48.

[44] R. Rosati, L. Romeo, V. M. Vargas, P. A. Gutiérrez, C. Hervás-Martínez, and E. Frontoni, “A novel deep ordinal classification approach for aesthetic quality control classification,” *Neural Comput Appl*, vol. 34, no. 14, pp. 11625–11639, Jul. 2022, doi: 10.1007/S00521-022-07050-6/FIGURES/8.

[45] A. Baraldi, F. Del Buono, M. Paganelli, and F. Guerra, “Landmark Explanation: An Explainer for Entity Matching Models,” *International Conference on Information and Knowledge Management, Proceedings*, pp. 4680–4684, Oct. 2021, doi: 10.1145/3459637.3481981.

[46] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.020.

[47] F. Ma, Y. Li, S. Ni, S. Huang, and L. Zhang, “Data Augmentation for Audio-Visual Emotion Recognition with an Efficient Multimodal Conditional GAN,” *Applied Sciences 2022, Vol. 12, Page 527*, vol. 12, no. 1, p. 527, Jan. 2022, doi: 10.3390/APP12010527.

[48] R. Haque *et al.*, “Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning,” *BioMedInformatics 2024, Vol. 4, Pages 966-991*, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[49] D. Gupta, A. Golder, M. M. Haque, and M. A. Moni, “CervixMed: Detecting Cervical Cancer based on combinational data using Hybrid architecture,” *2023 International Conference on Digital Image Computing: Techniques and*

Applications, DICTA 2023, pp. 570–577, 2023, doi: 10.1109/DICTA60407.2023.00085.

[50] A. Yeshwanth and R. Bhuvaneswari, "Weapon Classification with Xception: An Efficient Deep Learning Approach," *2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024*, 2024, doi: 10.1109/ICICACS60521.2024.10498457.

[51] A. Mehmood, Y. Gulzar, Q. M. Ilyas, A. Jabbari, M. Ahmad, and S. Iqbal, "SBXception: A Shallower and Broader Xception Architecture for Efficient Classification of Skin Lesions," *Cancers 2023, Vol. 15, Page 3604*, vol. 15, no. 14, p. 3604, Jul. 2023, doi: 10.3390/CANCERS15143604.

[52] K. Shaheed, Q. Abbas, A. Hussain, and I. Qureshi, "Optimized Xception Learning Model and XgBoost Classifier for Detection of Multiclass Chest Disease from X-ray Images," *Diagnostics 2023, Vol. 13, Page 2583*, vol. 13, no. 15, p. 2583, Aug. 2023, doi: 10.3390/DIAGNOSTICS13152583.

[53] R. Haque *et al.*, "A transfer learning-based computer-aided lung cancer detection system in smart healthcare," *IET Conference Proceedings*, vol. 2024, no. 37, pp. 594–601, Mar. 2025, doi: 10.1049/ICP.2025.0858.

[54] E. S. Cutur and N. G. Inan, "Multi-class Classification of Retinal Eye Diseases from Ophthalmoscopy Images Using Transfer Learning-Based Vision Transformers," *Journal of Imaging Informatics in Medicine*, pp. 1–15, Jan. 2025, doi: 10.1007/S10278-025-01416-7/METRICS.

[55] A. Yadav and E. Kumar, "Object Detection on Real-Time Video with FPN and Modified Mask RCNN Based on Inception-ResNetV2," *Wirel Pers Commun*, vol. 138, no. 4, pp. 2065–2090, Oct. 2024, doi: 10.1007/S11277-024-11539-9/METRICS.

[56] A. Asare, A. A. Broni, A. K. A. Dickson, M. Sagoe, and J. M. Cudjoe, "Performance of ResNet-18 and InceptionResNetV2 in Automated Detection of Diabetic Retinopathy," *Medicine Advances*, Jul. 2025, doi: 10.1002/MED4.70023.

[57] Z. Qiu, X. Sun, M. Sun, and L. Jia, "Text-Independent Speaker Verification Based on 3D Inception-Resnet," pp. 515–521, Dec. 2023, doi: 10.1109/PRML59573.2023.10348339.

[58] S. Dash, P. K. Sethy, and S. K. Behera, “Cervical Transformation Zone Segmentation and Classification based on Improved Inception-ResNet-V2 Using Colposcopy Images,” *Cancer Inform*, vol. 22, Jan. 2023, doi: 10.1177/11769351231161477/ASSET/IMAGES/LARGE/10.1177_11769351231161477-FIG7.JPEG.

[59] H. Wang, S. Xu, K. bin Fang, Z. S. Dai, G. Z. Wei, and L. F. Chen, “Contrast-enhanced magnetic resonance image segmentation based on improved U-Net and Inception-ResNet in the diagnosis of spinal metastases,” *J Bone Oncol*, vol. 42, p. 100498, Oct. 2023, doi: 10.1016/J.JBO.2023.100498.

[60] M. I. H. Siddiqui *et al.*, “Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI,” *Inform Med Unlocked*, vol. 56, p. 101657, Jan. 2025, doi: 10.1016/J.IMU.2025.101657.

[61] B. Prashanthi, A. V. P. Krishna, and C. M. Rao, “LEViT- Leaf Disease identification and classification using an enhanced Vision transformers(ViT) model,” *Multimed Tools Appl*, vol. 84, no. 21, pp. 23313–23344, Jun. 2025, doi: 10.1007/S11042-024-19866-6/METRICS.

[62] N. A. Aljarallah, A. K. Dutta, and A. R. W. Sait, “Image classification-driven speech disorder detection using deep learning technique,” *SLAS Technol*, vol. 32, p. 100261, Jun. 2025, doi: 10.1016/J.SLAST.2025.100261.

[63] Y. Li, X. Yang, D. Tang, and Z. Zhou, “RDTN: Residual Densely Transformer Network for hyperspectral image classification,” *Expert Syst Appl*, vol. 250, p. 123939, Sep. 2024, doi: 10.1016/J.ESWA.2024.123939.

[64] A. R. W. Sait, “A LeViT–EfficientNet-Based Feature Fusion Technique for Alzheimer’s Disease Diagnosis,” *Applied Sciences 2024, Vol. 14, Page 3879*, vol. 14, no. 9, p. 3879, Apr. 2024, doi: 10.3390/APP14093879.

[65] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, “Vision Transformers for Image Classification: A Comparative Survey,” *Technologies 2025, Vol. 13, Page 32*, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/TECHNOLOGIES13010032.

[66] A. S. U. K. Pranta *et al.*, “A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification,” *Computers 2025, Vol. 14, Page 197*, vol. 14, no. 5, p. 197, May 2025, doi: 10.3390/COMPUTERS14050197.

[67] A. Sriwastawa and J. A. Arul Jothi, “Vision transformer and its variants for image classification in digital breast cancer histopathology: a comparative study,” *Multimed Tools Appl*, vol. 83, no. 13, pp. 39731–39753, Apr. 2024, doi: 10.1007/S11042-023-16954-X/METRICS.

[68] S. Mehta and M. Rastegari, “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer,” *ICLR 2022 - 10th International Conference on Learning Representations*, Oct. 2021, Accessed: Feb. 11, 2024. [Online]. Available: https://arxiv.org/abs/2110.02178v2

[69] Q. Zheng, S. Saponara, X. Tian, Z. Yu, A. Elhanashi, and R. Yu, “A real-time constellation image classification method of wireless communication signals based on the lightweight network MobileViT,” *Cogn Neurodyn*, vol. 18, no. 2, pp. 659–671, Apr. 2024, doi: 10.1007/S11571-023-10015-7/METRICS.

[70] M. T. Alam, Y. T. Acquaah, and K. Roy, “Image-Based Human Action Recognition with Transfer Learning Using Grad-CAM for Visualization,” *IFIP Adv Inf Commun Technol*, vol. 711, pp. 117–130, 2024, doi: 10.1007/978-3-031-63211-2_10.

[71] Y. Zhang, Y. Zhu, J. Liu, W. Yu, and C. Jiang, “An Interpretability Optimization Method for Deep Learning Networks Based on Grad-CAM,” *IEEE Internet Things J*, vol. 12, no. 4, pp. 3961–3970, 2025, doi: 10.1109/JIOT.2024.3485765.