Detection and classification of eye diseases in cattle using image analysis with deep learning

A thesis submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

Sam Tuosheng Xiao

Sydney School of Veterinary Science

Faculty of Science

University of Sydney

Year of submission: 2026

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Sam Xiao

30 June 2025

Acknowledgments

I would like to extend my sincere thanks to my supervisory team for their support and guidance throughout the course of this PhD. I am especially grateful to Dr Mehar Khatkar, who has overseen this project from its inception and provided consistent supervision across all stages. His input has been instrumental in shaping the experimental design, refining the deep learning pipeline, and guiding the analytical approaches taken in this research. He also offered crucial support in troubleshooting technical challenges and helping me maintain clear research objectives over the years.

I would also like to thank Dr Navneet Dhand, who became my main supervisor during the final part of this candidature. His contribution was particularly valuable in reviewing the final thesis drafts, offering constructive and insightful feedback that helped improve the clarity, structure, and interpretation of the findings. His experience in epidemiology and scientific writing brought a helpful external perspective to the project.

I am also very grateful to Dr Peter Thomson for his ongoing support throughout the PhD. His feedback on chapter drafts and discussions around veterinary context helped ensure the research remained grounded in clinical relevance. His encouragement across different phases of the project was important for maintaining momentum and staying focused on the broader goals of the research.

More broadly, I would like to thank my supervisors not only for their academic guidance, but also for their mentorship—helping me grow as an independent researcher, ask the right questions, and develop the skills necessary to carry out interdisciplinary work across computer vision and veterinary science.

Finally, I would like to thank my family for their steady encouragement and support. Their patience, understanding, and belief in me have been an important foundation throughout this journey.

This research reported in this thesis was supported by the award of a Research Training Program scholarship to the PhD Candidate via MLA and University of Sydney partnership.

Authorship Attribution Statement

This thesis contains works that have been published or are under revision for publication. The materials contained in the works are included in Chapter 2. I have played the primary role in this published work. Details of my contribution and publication information are listed below:

- Chapter 2 of this thesis is published as (Xiao, Dhand et al. 2025):

  Xiao S, Dhand NK, Wang Z, Hu K, Thomson PC, House JK and Khatkar MS (2025) Review of applications of deep learning in veterinary diagnostics and animal health. *Front. Vet. Sci*. 12:1511522. doi: 10.3389/fvets.2025.1511522

I conducted the comprehensive search and selection of relevant literature, critically analysed the included studies, and synthesised key findings to identify existing research gaps. I also wrote and edited the manuscript for publication.

Sam Xiao

30 June 2025

Signature:

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Associate Prof. Navneet Dhand

30 June 2025

Signature:

Artificial Intelligence

Generative AI tools were used to support the preparation of this thesis. In particular, OpenAI's ChatGPT was employed to assist with tasks such as refining academic writing and grammar, improving clarity and flow, providing feedback on structure, offering programming support (including Python and deep learning frameworks), and helping to explain technical concepts. All content was written and critically reviewed by the author to ensure accuracy, originality, and alignment with the thesis objectives.

Signature:

Table of Contents

Contents

Table of Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
X-AI	Explainable Artificial Intelligence
CNN	Convolutional Neural Network
Grad-CAM++	Gradient-weighted Class Activation Mapping++
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	SHapley Additive exPlanations
YOLO	You Only Look Once
VGG	Visual Geometry Group
ResNet	Residual Network
DenseNet	Densely Connected Convolutional Networks
EfficientNet	Efficient Convolutional Network
Inception	Inception Neural Network Architecture
ROI	Region of Interest
TPU	Tensor Processing Unit
GPU	Graphics Processing Unit
CSV	Comma-Separated Values

AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
BCE	Binary Cross-Entropy
CCE	Categorical Cross-Entropy
MAE	Mean Absolute Error
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
IBK	Infectious Bovine Keratoconjunctivitis (Pinkeye)
S	Stage
A	Active
R	Resolving
N	Normal

Abstract

Pinkeye, also known as infectious bovine keratoconjunctivitis (IBK), is a highly contagious ocular disease in cattle characterised by inflammation of the cornea and conjunctiva. It remains one of the most common and economically significant health issues in the livestock industry, yet timely diagnosis remains difficult in Australia's vast grazing systems, where veterinary access is limited. This thesis presents a deep learning–based diagnostic framework designed to detect and classify pinkeye from mobile phone–captured images of the eye, offering scalable support for on-farm decision-making.

A systematic review of 422 studies (39 met inclusion criteria) identified gaps in veterinary AI applications, particularly in relation to livestock, ocular diseases, and external photographic modalities. To address this, a four-part research programme was undertaken. Firstly, You Only Look Oncev5 (YOLOv5) was retrained on 2,000 annotated images from Australia and the USA to localise the eye region in images with high precision (mAP = 0.99; 100% specificity). Secondly, a 3,800-image dataset was annotated using a clinician-developed 17-attribute scorecard and used to train multiple Convolutional Neural Network model with varying architectures for analysis. EfficientNetV2B2 consistently outperformed others, achieving high binary classification performance (AUC up to 0.92) and strong ordinal agreement for attributes such as Periocular score (Cohen's kappa = 0.84). Thirdly, full-stage classification (Normal, Active, Resolving, Resolved) and severity grading were modelled using 3,800 images. While binary treatment classification achieved 94% accuracy, performance dropped for multiclass (69%) and ordinal (κ as low as 0.59) tasks due to certain limiting factors such as sample size, class imbalance, overlapping visual features between disease stages, and inconsistencies in field-acquired images.

To enhance transparency, explainable AI tools (Grad-CAM++, LIME, SHAP) were applied to produce heatmaps that reveal the image regions most influential to each prediction, allowing verification that the model's focus aligns with clinically recognised signs of pinkeye. Grad-CAM++ best aligned with clinical intuition, consistently identifying lesion sites and corneal patterns relevant to diagnosis.

This thesis demonstrates that deep learning can deliver accurate and interpretable pinkeye diagnostics from field-acquired images under real-field conditions. The end-to-end pipeline was developed to be robust, transparent, and aligned with clinical reasoning. Designed for mobile deployment, the framework can be embedded into a smartphone application for on-farm testing and decision support, offering a scalable and transferable model for AI-assisted livestock health management in remote and resource-limited settings targeting ocular diseases in cattle.

Chapter 1. General Introduction

1.1 Background and Motivation

The Australian cattle industry is a cornerstone of the nation's agricultural economy, encompassing both beef and dairy production systems that operate across a wide range of climatic and geographic regions. Australia maintains a beef and dairy population exceeding 24 million head, with beef exports alone generating over $10 billion annually (MLA 2025). The industry supports tens of thousands of jobs and contributes significantly to the viability of rural and remote communities (Greenwood, Gardner et al. 2018). However, the extensive nature of Australia's grazing systems presents considerable logistical challenges for the ongoing monitoring and management of animal health. Cattle are often reared across vast pastoral properties in isolated regions, where access to veterinary services, timely health assessments, and treatment interventions can be limited (Kneipp, Green et al. 2021).

Among the diverse health issues affecting cattle, infectious diseases remain one of the most economically impactful types. Within this broad category, ocular diseases pose serious welfare and productivity concerns. Of these, infectious bovine keratoconjunctivitis (IBK), more commonly known as pinkeye, is one of the most prevalent and problematic conditions encountered by cattle producers (Kneipp, Green et al. 2022). Pinkeye is a contagious bacterial disease that causes painful inflammation of the eye, and in severe cases, irreversible vision loss (Alexander 2010). The primary speculated causative agent is *Moraxella bovis*, although other pathogens, such as *Moraxella bovoculi* and *Mycoplasma* spp., may also be involved (Cullen, Engelken et al. 2017). Environmental factors, including UV exposure, dusty pastures, eye irritants like straw heads, and fly activity, play a significant role in disease transmission and exacerbation (Kneipp, Green et al. 2021).

The clinical progression of pinkeye is marked by a series of visible symptoms, including conjunctival redness, lacrimation (tearing), corneal oedema, ulceration, corneal opacity, and in advanced cases, perforation and rupture (Ward and Nielson 1979). These symptoms not only result in substantial discomfort and distress to affected animals but also impair their ability to feed and navigate their environment (Kneipp, Green et al. 2021). As a result, infected cattle frequently exhibit reduced weight gain, diminished milk production, and, in extreme cases, may have to be culled from the herd or sales, leading to significant financial losses for producers. Studies have estimated that pinkeye can cost cattle operations over AUD 200 per case when factoring in decreased growth rates, treatment expenses, and labour requirements (Kneipp, Green et al. 2022). The disease also raises ethical concerns around animal wellbeing, particularly if pain management and early treatment are delayed due to insufficient access to veterinary care.

Accurate and timely diagnosis of pinkeye is therefore essential for both improving animal productivity and limiting the spread of infection within herds. However, in many Australian farming contexts, especially remote inland regions, relying on frequent in-person veterinary visits is neither practical nor economically sustainable. Diagnosis is typically based on visual inspection and clinical signs, requiring experienced veterinary experts to differentiate between early-stage pinkeye and other ocular abnormalities such as trauma, foreign bodies, or congenital defects (Angelos 2015). Inconsistent training or reliance on untrained staff may result in missed or delayed diagnoses, contributing to worsening outcomes (Kneipp, Govendir et al. 2021). Moreover, early symptoms can be subtle without staining or specialty equipment, making visual detection even more challenging.

This creates a compelling need for alternative, scalable tools to support frontline diagnosis and treatment decisions. Recent advances in artificial intelligence (AI), in particular deep learning (DL) approaches, offer promising opportunities to address this gap. DL techniques, such as convolutional neural networks (CNNs), are well-suited to image-based tasks and have already

demonstrated considerable success in domains such as human dermatology, ophthalmology, and radiology (Kshatri and Singh 2023). These models are capable of learning complex visual patterns from annotated datasets, enabling automated classification and localisation of disease-relevant features in images (Pereira, Franco-Gonçalo et al. 2023). In the context of animal health, this opens up the possibility of developing mobile, camera-based diagnostic tools that can take pictures of potentially affected eyes and provide real-time automatic diagnostic feedback to remote farmers isolated from immediate veterinary assistance.

Although interest in AI applications for veterinary diagnostics is growing, current efforts have been largely concentrated on companion animals, such as dogs and cats, and internal imaging modalities like radiographs and ultrasound (Chapter 2;(Xiao, Dhand et al. 2025)). Within livestock, there is a particular lack of research on external, surface-level conditions such as pinkeye. Although the disease presents visible symptoms, it remains challenging to quantify and assess reliably in the absence of consistent diagnostic frameworks or automated tools (Chapter 2;(Xiao, Dhand et al. 2025)). This represents a major research gap, particularly considering the scale and economic relevance of the cattle industry and the widespread occurrence of pinkeye in regions like Australia. Estimates of pinkeye occurrence range widely from 0.6% to 90%, influenced by factors such as age, breed, season, and geographic region. Earlier studies placed the overall prevalence at approximately 4.5% in cattle generally, and up to 10% in calves, which represents a substantial proportion (Kneipp, Govendir et al. 2021) . Moreover, image datasets of livestock eyes tend to be affected by aspects that with impact image quality, such as motion blur, inconsistent angles, poor lighting, and breed-related pigmentation differences, which may present challenges during model development and validation.

Despite these limitations of image quality, Australian cattle owners and farm technicians, equipped with mobile phones, represent an untapped resource for the development and implementation of AI-based diagnostic tools for

livestock health. Their ability to capture and share high-quality images can enable a crowdsourcing approach for building robust training datasets essential for machine learning model development. This same mobile technology infrastructure that exists across rural Australia creates an ideal pathway for deploying AI solutions directly to the farmers themselves. By leveraging existing mobile device ownership, a system for detecting and classifying cattle ocular diseases can be implemented effectively in the field, providing timely diagnostics without requiring specialised equipment or extensive training.

This thesis aims to explore the feasibility of using DL to develop robust, generalisable models for the automated classification of pinkeye in cattle using photographic images of the eye region. It requires the development of a scoring system for ocular diseases and constructing a curated dataset by annotating images representing different stages and severities of the disease, and evaluating a range of CNN architectures to determine their performance under realistic field conditions. By doing so, this work will contribute to the development of practical tools, such as an AI-powered mobile application designed for on-the-go diagnostics. This tool aims to assist farmers and veterinarians in the early detection and classification of pinkeye symptoms directly in the field, facilitating timely treatment decisions, minimising disease transmission, and ultimately enhancing animal welfare and productivity in livestock operations.

Ultimately, the goal is to reduce the diagnostic burden on producers, increase consistency in disease assessment, and enable earlier intervention for one of the most prevalent ocular diseases in cattle. In remote and resource-limited environments, such a mobile app may be essential for overcoming systemic barriers to veterinary access and ensuring better outcomes for animals and producers alike.

1.2 Aims and Objectives

The overarching goal of this research is to develop a deep learning model capable of classifying pinkeye in cattle from digital phone-captured images of the eye. With smartphones now widely accessible across rural and regional Australia, it is feasible for cattle producers and field staff to take and transmit high-quality images without the need for specialised equipment. Such a model has the potential to be integrated into a mobile application, enabling users to obtain real-time assessments and reduce the dependency on immediate in-person veterinary evaluations. By supporting early detection and hence timely treatment, this technology could help mitigate the spread and severity of pinkeye, thereby improving both economic outcomes and animal welfare.

To realise this goal, the thesis addresses the following specific objectives:

- Assemble a comprehensive image dataset of cattle eyes, develop a scorecard with assistance from veterinary experts, annotation of images to capture a range of pinkeye attributes, stage and severity levels.
- Implement cleaning, normalisation, and augmentation techniques to enhance dataset quality and model readiness.
- Employ and compare multiple well-known CNN architectures by including both custom and pre-trained transfer learning models, to identify optimal configurations for classification.
- Use systematic approaches, such as Keras Tuner, to optimise model performance through tuning of learning rates, dropout rates, network depth etc.
- Train models to classify individual eye attributes (e.g., corneal opacity, blood vessels, tear production), disease stages (active, resolving, resolved), and severity levels (1-4) using binary, multiclass, and ordinal strategies.

- Assess model performance through relevant metrics (accuracy, F1-score, AUC, etc.) and interpret outputs using explainable AI (X-AI) methods, such as Grad-CAM++, to visualise the image regions that most influenced the model's decisions, improving transparency and clinical interpretability.

1.3 Structure of the Thesis

This thesis is structured to build progressively from developing training datasets by creating a system of image scoring and annotation through to the implementation and evaluation of deep learning models for pinkeye diagnosis.

- Chapter 1 (Introduction) outlines the background, motivation, aims, and structure of the thesis. It introduces the problem of pinkeye in cattle, the rationale for using deep learning, and the objectives of the research.
- Chapter 2 (Literature Review) provides a detailed synthesis of existing approaches to veterinary diagnostics using deep learning, with a focus on image-based applications. The chapter identifies both the strengths and limitations of current methodologies and highlights the research gaps. A systematic review is included to frame the landscape of related work and justify the study's scope.
- Chapter 3 (Object Detection and Image Preprocessing) outlines the pipeline for removing distracting artefacts from raw images, enhancing model focus on relevant regions of interest. This step ensures higher quality input for classification tasks and mimics a clinical approach by isolating the lesion-bearing area of the eye.
- Chapter 4 (Classification of Eye Attributes) presents the classification framework for several pinkeye-related attributes using binary, multiclass and ordinal approaches. This chapter tests the viability of the dataset and processing pipeline and examines model generalisability across various visual features to determine the optimal DL architecture for these types of analyses.

- Chapter 5 (Classification of Disease Stage and Severity) builds upon earlier chapters to perform more complex classification tasks. It addresses the staging of pinkeye and its severity scoring which are critical for clinical decision-making.
- Chapter 6 (Explainable AI and Model Interpretability) investigates the application of Explainable AI (X-AI) tools to provide transparency into model predictions. Different tools are used to determine which image regions influenced classification decisions, helping to validate model trustworthiness and reveal potential biases or misclassifications.
- Chapter 7 (Results and Discussion) synthesises the findings across all classification tasks, comparing the performance of different architectures and reflecting on real-world deployment considerations. Challenges such as data imbalance, image variability, and interpretability limitations are critically discussed.

Chapter 2. Review of applications of deep learning in veterinary diagnostic and animal health

2.1 Introduction

The field of artificial intelligence (AI) involves the development of computer systems that can emulate human like problem-solving abilities. AI systems are increasingly demonstrating proficiency across a wide range of sectors. These AI methods have been widely studied and applied to improve many aspects of a diverse range of disciplines in human medicine, such as drug development and delivery, patient monitoring, surgery, diagnostic imaging, screening, etc. (Hamet and Tremblay 2017). Numerous studies have consistently demonstrated that many AI models are at least as good as healthcare experts and specialists in performing some of the tasks, which they are designed to do, and even surpass the performance of the experts in some cases (Esteva, Kuprel et al. 2017). This highlights the transformative capabilities of AI in addressing complex healthcare challenges.

2.1.1 Machine learning and deep learning in veterinary medicine

Machine learning (ML) is a core approach in artificial intelligence (AI) that enables computers to learn from data and make predictions without explicit programming. ML encompasses two broad categories: traditional ML and deep learning (DL). Traditional ML methods, such as support vector machines (SVM), k-nearest neighbours (k-NN), and random forests, have been widely applied in livestock health, including oestrus and calving prediction, lameness detection, and disease monitoring (El Naqa and Murphy 2015, Cihan, Gokce et al. 2017, García, Aguilar et al. 2020). While these methods have demonstrated success in precision livestock farming and veterinary diagnostics (Ezanno, Picault et al. 2021), they often require manual feature engineering, which limits their scalability for complex data.

DL, on the other hand, is a subset of AI that is better suited for processing large amounts of complex data compared to ML with the additional costs of requiring higher computation power. The key difference between traditional ML and DL lies in feature engineering and model complexity. Traditional ML models often require manual feature engineering to extract relevant information from the data. In contrast, DL models, based on neural networks, automatically learn relevant features from the data during the training process, reducing the need for extensive manual intervention (Razavi 2021). DL involves the use of various types of neural networks with many layers, hence the term "deep". These models are inspired by the structure and function of the neuron connections in the brain and are capable of learning from complex, high-dimensional data by breaking it down into different layers or representations for analysis. Through DL, machines can now achieve a more nuanced approach to detecting trends in data, leading to accurate predictions and insights across numerous applications. DL technology has been a driving force behind many recent advancements in AI, including speech recognition, image recognition, and natural language processing (Bengio, Goodfellow et al. 2017). Figure 2.1 provides a brief general overview of the key stages involved in the workflow process of the development of a DL model.

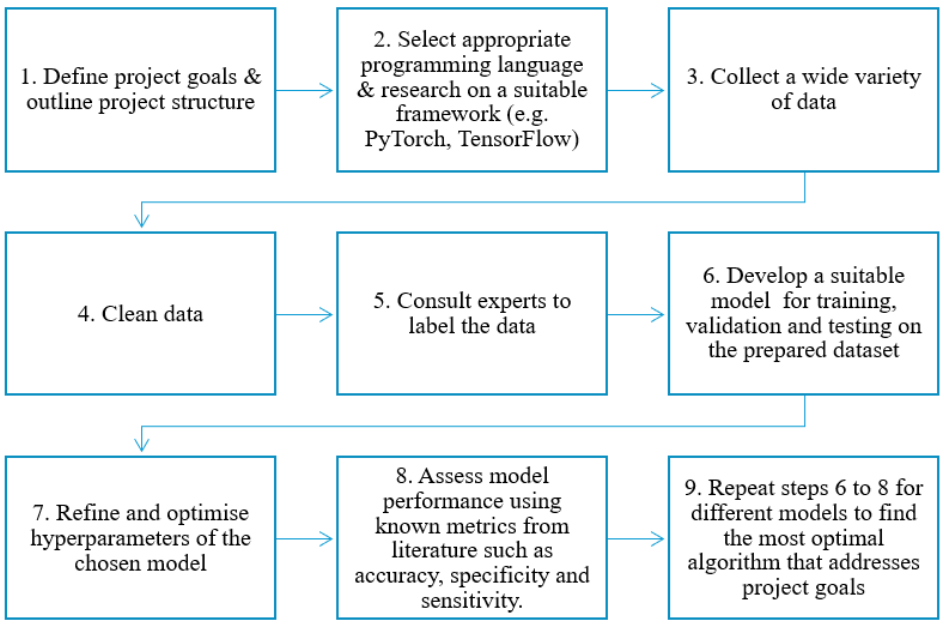


Figure 2.1. Example workflow of the development of a DL system

Among DL architectures, convolutional neural networks (CNNs) are the frequently used for image analysis in veterinary medicine. CNNs excel at processing visual data by learning spatial hierarchies of features, making them highly effective for disease detection, classification, and segmentation in medical imaging. The general CNN architecture consists of:

- Convolutional layers, which apply filters to extract essential patterns such as edges, textures, and structures from input images.
- Pooling layers, which reduce the spatial dimensions of feature maps, enhancing computational efficiency while preserving critical features.
- Fully connected layers, which integrate extracted features for classification or regression tasks (9, 10).

CNNs form the backbone of many advanced DL architectures, such as ResNet, EfficientNet, and Inception, which have been successfully applied in veterinary diagnostics for disease classification and prognosis prediction. These architectures continue to drive progress in AI-assisted veterinary medicine (11).

Deep learning has been extensively applied in human medicine for diagnosis, treatment planning, and disease monitoring, leading to improved patient outcomes and cost reductions (20, 21). However, its application in veterinary medicine remains in its early stages (6). This review examines the current state of DL applications in veterinary diagnostics and animal health, highlighting key advancements, challenges, and potential future directions in the field.

2.2 Methods

We conducted this systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 2.2). The PRISMA 2020 guidelines provide a universal framework that contains a 27-element checklist and a flow diagram to ensure the comprehensive

documentation of the review process, from literature search and study selection to data extraction and synthesis (Page, Moher et al. 2021). The search query included the terms ‘Deep learning’ and ‘Veterinary’ within the PubMed database, aiming to identify literature focused on the application of deep learning (DL) techniques in veterinary medicine. This search yielded 422 relevant articles. The titles and abstracts of these articles were exported as a CSV file for further examination, and only primary research articles were retained. A total of 66 non-primary articles, including those related to veterinary curriculum, review articles, and books or book chapters mentioning AI, were excluded. Abstracts of the remaining articles were carefully reviewed to confirm that they involved the use of deep neural networks for animal disease diagnostics. Articles that mislabelled simple neural networks or ML techniques as DL, used DL only for preprocessing, object detection, or segmentation, or focused on animal models for human medicine were excluded, totalling 142 articles. Additionally, 174 articles related to human medicine and one duplicate were removed. Finally, 39 articles met the inclusion criteria (Table 2.1). Only full-text articles available via open access or the University of Sydney’s institutional access were included in this review.

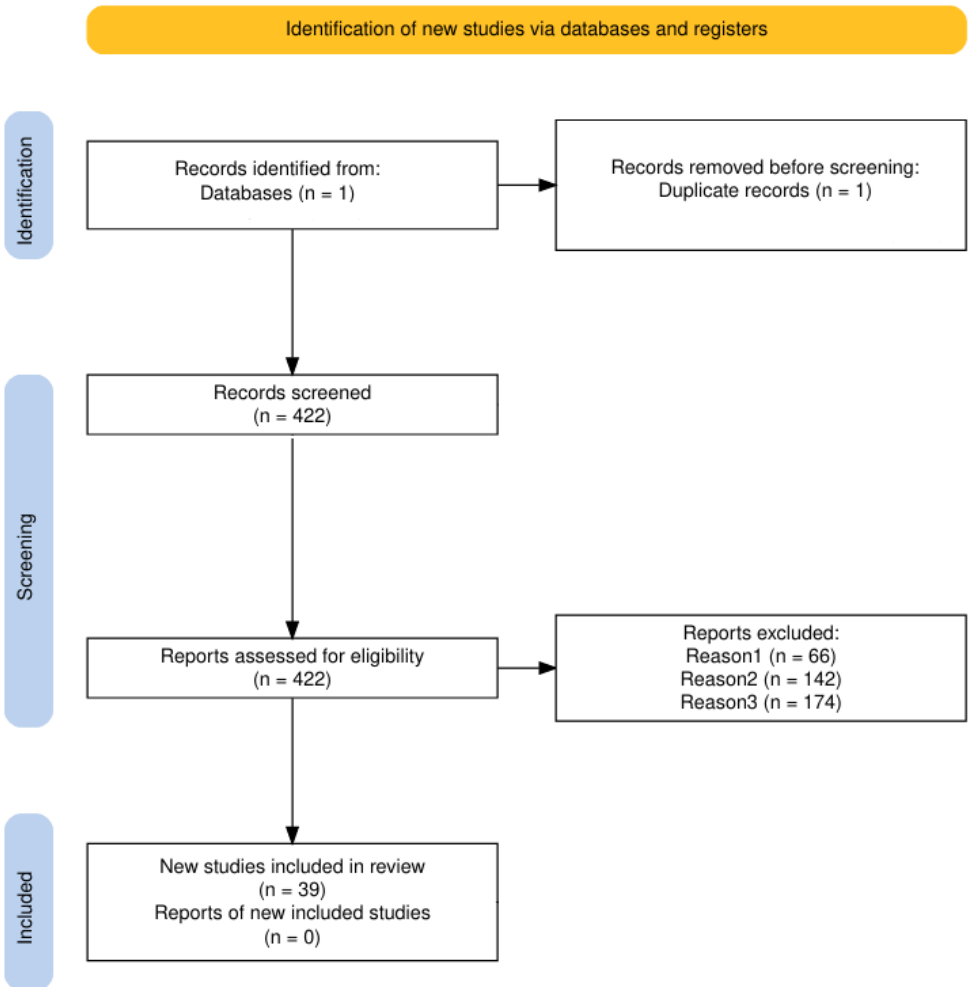


Figure 2.2 PRISMA flow diagram

2.3 Results

The articles that met the inclusion criteria were initially arranged into applications of DL in diagnostics and other domains. Next, in the diagnostics section, articles that were relevant to the diagnosis of diseases were summarised and grouped based on the type of images and/or data they investigated. This helped to synthesise the relevance of these techniques in accordance with the wide variety of information required for accurate diagnosis in the veterinary health context. The remaining section included research on DL applications in areas outside of diagnostics but still consistent with the theme of detecting or predicting diseases.

Most of the diagnostic DL research in veterinary medicine is related to the interpretation of medical images. The proportion of different data types used

in the DL studies is presented in Figure 2.3, and the species-wise studies in Figure 2.4. Most of the DL studies (84%) were on canine (64%) and feline (20%), highlighting the gap in research on other animals, especially those in the livestock industry. The increasing development of DL within the veterinary health/diagnostics context since its inception in 2013 is presented in Figure 2.5.

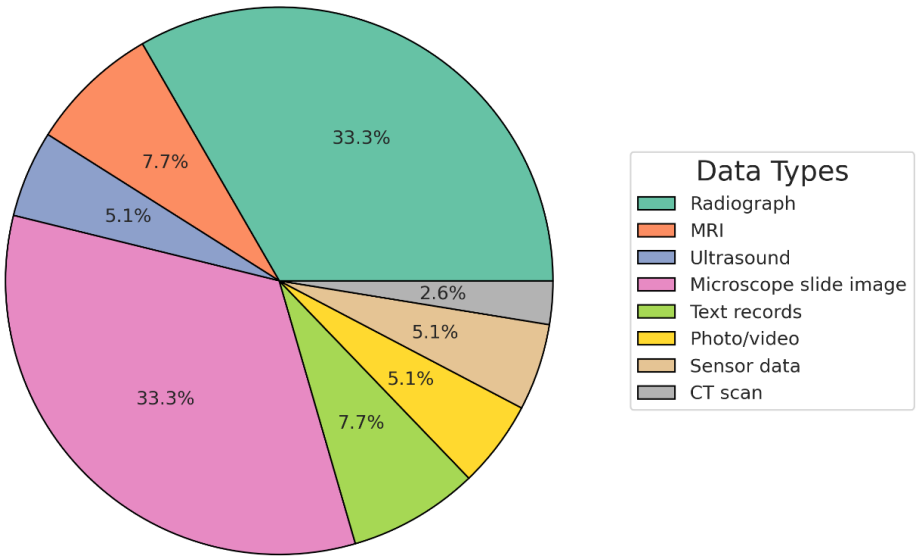


Figure 2.3. The proportion of different data modalities used in the DL studies.

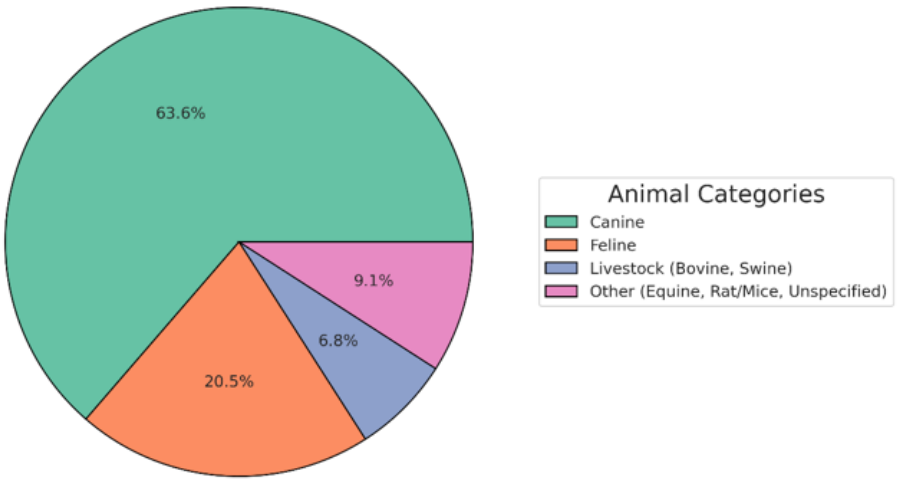


Figure 2.4. The proportion of different species researched in the DL studies.

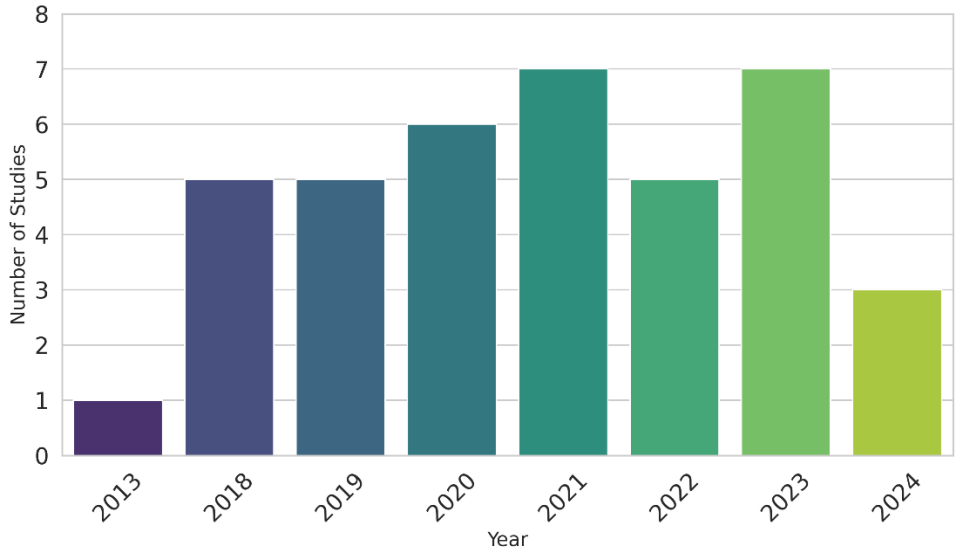


Figure 2.5. The number of Deep Learning studies in veterinary medicine across the years.

Over the years, the quest for improved performance in various studies led to the exploration of different methodologies. When examining the trajectory of accuracy over the years, no clear overarching trend emerged. It appeared model accuracy was dependent on the problems these studies attempted to solve, the animal species, the size of the dataset available, and the complexity of the models utilised. While there was not a consistent upward trend in the size of training data across all studies, variations were notable depending on the domains. Datasets for diagnostics studies involving canine and feline companion animals were larger compared to those involving other animals, which indicates limited research on other animals, such as horses, as DL modelling required larger data sizes for better classification performance. Moreover, there was a discernible disparity in sample sizes, with text/health records and sensor data studies often featuring more substantial datasets compared to studies using radiograph data. Finally, there was not a dramatic shift in the types of models used over the years. CNNs continued to be the predominant model type in various studies, as seen in Table 2.1, which suggested that the research in DL was still at an early stage, as other DL models were rapidly evolving and yet to be investigated in the veterinary health context. However, a noteworthy observation was the incorporation of

transfer learning techniques based on more updated sophisticated models, such as different ResNet versions throughout the years reviewed. The use of transfer learning, a technique which enhances model performance by leveraging pre-existing knowledge from a pre-trained model often on larger datasets of images, and fine tuning it for a new image classification task allows the model to adapt effectively to the specific contexts found in the domain of animal health (Kim, Cosa-Linan et al. 2021). The following sections examine how deep learning have been applied across key domains of veterinary practice.

2.3.1 Deep learning involved in disease diagnosis

2.3.1.1 Veterinary medical imaging

Traditional ML strategies can be used for medical image analysis, such as the bag of features (BoF) strategy that was applied by Yoon et al. (Yoon, Hwang et al. 2018), which aimed to distinguish between normal and abnormal radiographic findings from canine thoracic radiographs across many different regions. The same study included a DL component using a CNN to accomplish the same goal. A direct comparison between the two strategies in this study showed that CNN had higher accuracy and sensitivity measures in performing these tasks when compared to BoF (Yoon, Hwang et al. 2018). Similarly, a comprehensive study trained a CNN model using a large sample of 22,000 veterinary radiographs of cats and dogs combined to predict and identify 15 types of primary thoracic lesions from the radiographs. It showed that classification based on DL produced a significantly lower error rate when compared to the classification performance made by veterinarians (Boissady, de La Comble et al. 2020). Interestingly, the study also asked veterinarians to make predictions with access to the results provided by the DL model, but the experts' prediction results did not improve significantly, which was implied by the authors that experts have a certain level of scepticism about the results of the AI technology. Thoracic radiographs were also analysed to quantify cardiac enlargement to predict cardiac diseases, where DL modelling achieved high concordance between its assessment and human specialists

across both canine and feline patients (Boissady, De La Comble et al. 2021). This success was mirrored in another study, where a DL model's cardiac index calculation outperformed the clinical standard produced by veterinary radiologists in predicting cardiac enlargement (Jeong and Sung 2022). Expanding on these applications in thoracic radiographic analysis, convolutional neural networks (CNNs) have been particularly effective in detecting cardiomegaly (heart enlargement) from these images, achieving high diagnostic accuracy in canines (Burti, Osti et al. 2020, Li, Wang et al. 2020). Further advancements in DL have led to the creation of cardiac scoring models for predicting and diagnosing canine heart diseases (Zhang, Zhang et al. 2021). CNN models have also been applied to feline heart conditions, such as feline hypertrophic cardiomyopathy, achieving diagnostic accuracies exceeding 90% in identifying the disease from radiographic images (Rho, Shin et al. 2023). These findings highlight the adaptability of DL techniques across different species for the diagnosis of heart diseases.

Further illustrating CNN's utility in image analysis, Banzato et al. (Banzato, Bonsembiante et al. 2018) showed that CNNs outperformed non-invasive diagnostic tests, such as serum biochemistry and cytology, in the detection of hepatic diseases from ultrasound images, underscoring DL's potential to enhance diagnostic confidence. However, a separate study on canine chronic kidney disease (CKD) highlighted ongoing challenges, particularly in tackling more complex multi-class classification problems. The CNNs studied struggled to classify five stages of the disease from ultrasound images, achieving a performance accuracy of only 0.46 on average, likely due to the subtle differences between stages and the limitations of the ultrasound imaging. Importantly, ultrasound may not be the most suitable modality for accurate CKD staging, which affects the reliability of the ground truth used to train and evaluate the model. Since DL models are highly dependent on the quality of their training data, a suboptimal ground truth can constrain model performance, regardless of architectural improvements. This underscores the importance of carefully selecting diagnostic modalities and ground truth

definitions in AI studies to ensure meaningful and clinically relevant outcomes. While model selection and hyperparameter tuning remain critical, optimising study design and ensuring high-quality, confirmatory diagnostic data are equally essential for unlocking the full potential of DL applications in veterinary medicine (Yu, Lee et al. 2024).

VGG (visual geometry group) networks have been applied successfully to MRI scans of canine lumbar discs, where they were used to grade intervertebral disc degeneration, achieving over 0.9 accuracy across all five degeneration grades (Niemeyer, Galbusera et al. 2024). Another noteworthy development in DL for veterinary imaging was the use of YOLO (You Only Look Once) a CNN based model. Originally designed for object detection, YOLO v3, v4, and v4 tiny have been adapted to not only detect object but also classify tracheal collapse grades (Normal, grade 1-2, and grade 3-4) from lateral cervicothoracic canine radiographs, offering a versatile tool for both detection and classification tasks (Suksangvoravong, Choisunirachon et al. 2024).

CNN techniques have developed rapidly and branched off into more specialised networks with addition of image segmentation components to better analyse medical images (Patil and Deore 2013). Image segmentation divides the image into multiple significant parts to make the input dataset more informative to analyse (Zaitoun and Aqel 2015). This technique was investigated in a kidney disease diagnosis study where a U-Net CNN with an image segmentation component was tested to successfully estimate kidney volume from CT scans (Ji, Cho et al. 2022). Similarly, inbuilt image segmentation in a CNN was used to detect pulmonary abnormalities in feline radiographs, showcasing its utility (Dumortier, Guépin et al. 2022). This level of diagnostic success was not consistently reproduced in other studies such as in the diagnosis of lung lesions in both dogs and cats from x-ray images, the diagnostic accuracy was only around 70-80% (Arsomngern, Numcharoenpinij et al. 2019). Another study demonstrated low levels of performance for the detection of neoplasms (sensitivity ranged from 0-37.5%) and syringomyelia (sensitivity ranged from 0-10%) on MRI images using custom CNN models.

The poor performance was due to the limited training samples of those cases (Biercher, Meller et al. 2021), demonstrating that in certain contexts, DL remains an evolving technology requiring further refinement. Additional experimentation in both modelling approaches and pipeline design is necessary to optimise the algorithmic efficacy.

In another MRI analysis study, the authors attempted to distinguish between meningioma and glioma conditions in dogs from a selection of MRI images utilizing a process known as transfer learning to develop a CNN model from a pre-trained GoogLeNet CNN (deep neural network consisting of 144 layers) (Banzato, Bernardini et al. 2018). The GoogLeNet has been trained on the ImageNet database with up to 1.2 million images across 1,000 categories to extract the CNN features. It was retrained on a new dataset of MRI images which achieved high accuracy (91% and 94%, respectively) of the model in correctly classifying the condition from both pre-and post-contrast MRI images. The correct differentiation between the two conditions is necessary for choosing the right treatment procedures that could lead to better health outcomes for patients. Retraining pre-trained models reduces the computational resources and potentially the number of images required to apply DL, increasing the accessibility of AI's advantages in various animal health applications. It is still important to note that this improvement in accessibility may come in exchange for accuracy in certain situations.

Artificial neural network (ANN) is another DL technique that is composed of fully connected layers where each neuron is connected to every neuron in directly neighbouring layers, which are more commonly used for general purpose problem solving (Jain, Mao et al. 1996). It was applied to identify canine hip joints on ventrodorsal pelvis radiographs with low classification error, and high sensitivity and specificity measures of 8.9%, 86%, and 100%, respectively (McEvoy and Amigo 2013). ANN has the flexibility to choose different activation functions for nonlinear function learning purposes and change the number of hidden layers and nodes in these hidden layers, to improve its performance to suit various image processing needs (Tracey, Zhu

et al. 2011). An updated study instead used a deep CNN for the detection of the hip joint and extended their aim to the classification of hip dysplasia from pelvis radiographs in two stages. The first stage involved identifying the boundary box of the hip joint from the radiographs using the YOLOv3 object detection algorithm (McEvoy, Proschowsky et al. 2021). These regions were then cropped and put through the second stage of analysis, where a CNN model graded hip dysplasia, which resulted in a high specificity of 0.92 for FCI scores in the "C-E" group. However, the model's sensitivity metric was low at 0.53, suggesting its failure to identify many positive cases (false negative). One contributing factor to this limitation was that the image dataset contained unbalanced annotated images, where certain types of hip dysplasia were under-represented, which impaired the performance of the model in its testing phase (McEvoy, Proschowsky et al. 2021). This outlines the importance of having a large data set to train an effective CNN.

2.3.1.2 Microscope slide images

Microscopic examination of tissues, cells, and blood on microscope slide images is known to be tedious and challenging for disease diagnosis, even for well-trained specialists (Kumar, Singh et al. 2020). DL techniques, particularly CNN, have proven effective in addressing these challenges, such as its utilization in recognizing reticulocytes in cat blood smears to a high accuracy of 98.7% (Vinicki, Ferrari et al. 2018). Another study used CNN modelling to diagnose and classify abnormal cell growth in canine skin samples from cytological images (a subset of microscope slide images), improving cancer detection (Zapata, Chalco et al. 2020). Furthermore, DL methods, particularly CNNs, outperformed veterinary pathologists in grading prognostic elements of canine tumours based on stained canine cutaneous mast cell tumours (Aubreville, Bertram et al. 2020). Along with the results, it was noted that the chosen section of the slide images for analysis by veterinary pathologists was quite varied, thus producing more inconsistent results between experts in their mitotic counts. Also, a more advanced network based on the CNN known as ResNet 50 was able to match with pathologists' grading of

cardiomyopathy severity from microscope slide images in rodents with a Spearman rank-order correlation of 0.82 (Tokarz, Steinbach et al. 2021). In some cases, DL techniques surpass human performance as indicated by a study on horses, where the CNN outperformed human specialists (76% accuracy) in diagnosing exercise-induced pulmonary haemorrhage, achieving a high accuracy of 92% (Bertram, Marzahl et al. 2022). These studies emphasised the utility of DL in diagnosing diseases across different animals, showcasing the potential of these techniques to match or even outperform human specialists in certain contexts within veterinary medicine.

Recent advancements have adopted a more flexible approach to the analysis of microscope slide images, such as using DL models to first preprocess the images by dividing them into grid sections to ensure areas of interest are separated from irrelevant backgrounds. This step is crucial for achieving high classification accuracy, particularly when applying segmentation techniques like U-Net, which enables the identification of relevant regions prior to classification (Haghofer, Fuchs-Baumgartinger et al. 2023), and combining different parts of the algorithm from various known architectures to suit different needs like the aggregate model used in this study, where different CNNs (AlexNet, Inception v3, and ResNet) are combined to form the ARCTA algorithm, which was utilised to accurately classify canine cutaneous round cell tumors (accuracy of 91.7%) and mast cell tumours (accuracy of 100%) (Salvi, Molinari et al. 2021). This approach of utilizing aggregate modelling have also shown success in mitotic figure count, enabling critical early tumour detection (Fitzke, Whitley et al. 2021).

Pretrained models like VGGNet-16, have been applied to classify canine mammary tumours and human breast cancer from histopathological images, achieving improved accuracy when combined with traditional ML classifiers (93%) (Kumar, Singh et al. 2020). The incorporation of transfer learning helps to alleviate the issues of using a small image dataset by only fine-tuning the model parameters based on the knowledge obtained from a large dataset (Greenspan, Van Ginneken et al. 2016). Recent studies have leveraged

advanced CNN architectures such as VGG16, InceptionV3, and EfficientNet as feature extractors for canine tumour histopathology. By removing the final classification layers of these networks and feeding their outputs into traditional machine learning algorithms like support vector machines (SVM), researchers have enhanced differentiation between benign and malignant tumours (Burrai, Gabrieli et al. 2023). In some cases, transfer learning-based approaches have demonstrated particularly high accuracies, such as the use of EfficientNet B5, which achieved approximately 95% accuracy in classifying seven different types of canine skin tumours, though this still fell short of human expert performance levels (Fragoso-Garcia, Wilm et al. 2023). Another study used GoogLeNet transfer learning to classify three classes of canine lymphoma from whole slide images, achieving 99% accuracy in the test set (Hubbard-Perez, Luchian et al. 2024). Similarly, XceptionNet was employed to analyse whole slide images in mice, where it showed a strong correlation ($r = 0.9067$) with pathologists' grading of hepatic fibrosis (Kim, Baek et al. 2023), further illustrating the viability of DL techniques in diagnostic contexts.

Another example where combining different machine learning techniques presents a promising avenue in the veterinary health context. The R-CNN method, an object detection method, was first utilised to extract regions of interest then a ResNet classification model was utilised to detect patterns related to canine stifle joint disease with an accuracy over 80% (Shim, Lee et al. 2023). These results demonstrate the potential of combining multiple CNN networks for improved classification performance in veterinary medicine.

2.3.1.3 RGB images

The widespread availability of RGB images from smartphones and digital cameras has made DL models more accessible, highlighting potential beneficial applications and ease of access to this technology in veterinary medicine (Morikawa, Kobayashi et al. 2021). Hundreds of annotated photo images of canine eyes were trained and evaluated with CNN models: GoogleNet, ResNet, and VGGNet, to determine and predict corneal ulcer severity in dogs (Kim, Lee et al. 2019). It was shown that many of these DL

models achieved accuracies beyond 90% for identifying the different levels of corneal ulcer severity. Another study supports the above claim as its dataset was formed from photo images of equine eyes taken via smartphones (May, Gesell-May et al. 2022). Four different CNN models (MobileNetV2, InceptionV3, VGG16, VGG19) were studied to classify 3 categories of eye conditions with a particular focus on equine uveitis (a particular eye inflammatory disease), the top performing model achieved a validation accuracy of 96% (May, Gesell-May et al. 2022). A part of the imaging data of the eye was collected using a smartphone camera, which further highlights the applicability of DL algorithms in the analysis of photo images gathered via smartphones in the animal health domain. DL techniques have performed well deciphering limited-quality smartphone image data to produce highly accurate results. Images can also be obtained from videos by using object detection DL algorithms to form the image dataset for analysis. In a separate study on canine eye disease, images were isolated from video footages of the face of the animals for application in DL models (Kim, Han et al. 2022).

2.3.1.4 Text analysis

DL can be used to extract textual information to produce diagnostic suggestions. Disease knowledge was first extracted from a dairy cattle disease graph found in literature, as well as information obtained from experts and features found in medical records on a variety of dairy cow diseases such as mastitis, forestomach atony, rumen indigestion, gastroenteritis, rumen acidosis and abomasum dislocation to form the initial dataset. Then a CNN model was pretrained on this initial dataset to obtain the model parameters and weights. Finally, a transfer learning technique was employed utilizing this pretrained model on a more limited separate real-life dataset of textual features of the above outlined dairy cow diseases for training and testing to ascertain the diagnostic performance of the developed DL approach (Gao, Wang et al. 2021). The developed model showed a promising F1 score around 86% in the automation of the diagnosis of dairy cow diseases. The authors compared this model's performance to other standalone ML models,

including support vector machine, random forests, and decision tree, as well as DL methods: recurrent neural network (RNN) and CNN. The results showed the effectiveness of the transfer learning strategy with a pretrained model.

RNN is a DL method which is devised to formulate sequential patterns such as texts and videos (Young, Hazarika et al. 2018). RNN was adopted for detection of chronic cat kidney diseases from historical electronic hospital records of patients and achieved high classification performance of 0.907 and 0.989 for sensitivity and specificity respectively (Bradley, Tagkopoulos et al. 2019). In other text-based veterinary medicine data like necropsy reports as seen in the study performed by Bollig (2020) et al., an RNN with a long short-term memory (LSTM) design successfully classified evidence of gastrointestinal, respiratory and/or urinary diseases, which helps to reveal the features of these diseases such as its epidemiological nature (Bollig, Clarke et al. 2020).

2.3.1.5 Sensors

DL has been used to predict disease occurrence in animals based on historical sensor data. One study utilised a recurrent neural network (RNN) to generate an autoencoder that recognised environmental sensor data, such as CO_2, temperature, and humidity, which were associated with housing environment health for pigs (Cowton, Kyriazakis et al. 2018). The collected sensor data was evaluated by the GRU-autoencoder based on its similarity to normal data. If the data exceeds a specific anomaly threshold, optimised using Particle Swarm Optimisation, the algorithm outputs a prediction warning for respiratory disease. As sensor data is collected in real-time, it allows timely prediction for farmers to act and intervene to prevent the development of respiratory disease. Another type of data which is collected routinely is in the field of dairy, where milk attributes are continuously being collected and recorded by machines at each milking. An important reason for this is to monitor for the occurrence of mastitis in the dairy cows, as contamination will affect the milk quality and cause it to be discarded. Traditionally, statistical methods have been used to predict subclinical mastitis from dairy-related attributes (Gasqui and Barnouin 2003). However, one recent study

utilised a relatively simple multilayer feed-forward deep neural network to forecast subclinical mastitis with a high accuracy of 84% based on multiple milking variables, showcasing the capabilities of DL techniques in this field (Ebrahimi, Mohammadi-Dehcheshmeh et al. 2019).

From the literature, it appears that there is more research being performed on inferring, predicting, and tracking animal behaviour from sensor data, where it found moderate to higher levels of success (Norouzzadeh, Nguyen et al. 2018, Jeantet, Vigon et al. 2021). Compared to this, the analysis of sequential sensor data using DL algorithms appear to be premature at this current stage of development in the diagnosis of diseases in animals. Thus, more research in this area may uncover new and more comprehensive information that can be used to inform disease management and prevention in livestock farming.

Table 2.1. Applications of DL in Veterinary Medicine

Title	Year	DL Technique	Sample	Type of animal	Modality	Evaluation metric	**Reference**
Using machine learning to classify image features from canine pelvic radiographs: evaluation of partial least squares discriminant analysis and artificial neural network models	2013	ANN	256	Canine	Radiograph	Classification error: 0.089 Sensitivity: 0.86 Specificity: 1	McEvoy & Amigo
A methodological approach for deep learning to distinguish between meningiomas and gliomas on canine MR-images	2018	CNN (GoogLeNet)	80	Canine	MRI	Accuracy Post-contrast T1: 0.94 Pre-contrast T1: 0.91 T2 images: 0.90	Banzato et al.
Use of transfer learning to detect diffuse degenerative hepatic diseases from ultrasound images in dogs: A methodological study	2018	CNN (AlexNet)	52	Canine	Ultrasound	AUROC: 0.91 Sensitivity: 1 Specificity: 0.83	Banzato et al.
Using Convolutional Neural Networks for Determining Reticulocyte Percentage in Cats	2018	CNN	1,046	Feline	Microscope slide image	Accuracy: 0.987	Vinicki et al.
Prediction of radiographic abnormalities by the use of bag-of-features and convolutional neural networks	2018	CNN	7,138	Canine	Radiograph	Accuracy: 0.929-0.969 Sensitivity: 0.921-1 Specificity: 0.938-0.96	Yoon et al.

A combined deep learning gru-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors	2018	RNN (Gru)	Pig farms across Europe (unspecified)	Swine	Sensor data	Precision: 0.909 Recall: 0.909	Cowton et al.
Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning	2019	RNN	106,251	Feline	Electric health record	Sensitivity: 0.907 Specificity: 0.989	Bradley et al.
CNN-based diagnosis models for canine ulcerative keratitis	2019	CNN (GoogLeNet, ResNet, VGGNet)	281	Canine	Photograph	Accuracy: > 0.9	Kim et al.
Detection of Cutaneous Tumors in Dogs Using Deep Learning Techniques	2019	CNN	1,500	Canine	Cytological image	Unperformed	Zapata et al.
Computer-Aided Diagnosis for Lung Lesion in Companion Animals from X-ray Images Using Deep Learning Techniques	2019	CNN	2,862	Canine Feline	Radiograph	<u>Abnormal lung classification</u> Accuracy: 0.723 Sensitivity: 0.81 Specificity: 0.637 <u>Lung lesion detection</u> Accuracy: 0.796 Sensitivity: 0.76 Specificity: 0.833	Arsomngern et al.
Comprehensive analysis of machine learning models for	2019	CNN	364,249	Bovine	Dairy attributes/Sensor data	Accuracy: 0.84	Ebrahimi et al.

prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models							
Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs	2020	CNN	792	Canine	Radiograph	Accuracy: 0.827 Sensitivity: 0.684 Specificity: 0.871	Li et al.
Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region	2020	RetinaNet (ResNet18) CNN (U-net, ResNet18, ResNet50)	32	Canine	Microscope slide image	Correlation coefficient: 0.963-0.979	Aubreville et al.
Machine learning for syndromic surveillance using veterinary necropsy reports	2020	RNN (LSTM)	1,000	Unspecified	Necropsy report	F1 scores Gastrointestinal disease: 0.932 Respiratory disease: 0.947 Urinary disease: 0.752	Bollig et al.
Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs	2020	CNN	1,465	Canine	Radiograph	AUROC: > 0.9	Burti et al.
Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer	2020	CNN (VGGNet-16)	352	Canine	Histopathological image	Binary classification of canine mammary tumour accuracy: 0.93	Kumar et al.
	2020	CNN	22,120	Canine	Radiograph	Overall error rate: 0.107	Boissady et al.

Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence				Feline			
Using Deep Learning to Detect Spinal Cord Diseases on Thoracolumbar Magnetic Resonance Images of Dogs	2021	CNN	2,693	Canine	MRI	<u>IVDP detection</u> Sensitivity: 1 Specificity: 0.951 <u>IVDE detection</u> Sensitivity: 0.908 Specificity: 0.989 <u>FCE detection</u> Sensitivity: 0.622 Specificity: 0.979 <u>ANNPE detection</u> Sensitivity: 0.91 Specificity: 0.90	Biercher et al.
Histopathological Classification of Canine Cutaneous Round Cell Tumors Using Deep Learning: A Multi-Center Study	2021	CNN (AlexNet, Inceptionv3, ResNet)	416	Canine	Histopathological image	<u>Accuracy</u> RCT classification: 0.917 Mast cell tumor grading: 1	Salvi et al.
Using Artificial Intelligence to Detect, Classify, and Objectively Score Severity of Rodent Cardiomyopathy	2021	CNN (ResNet50)	300	Rat	Microscope slide image	Spearman rank-order correlation between pathologist median grade and AI grade: 0.82	Tokarz et al.

OncoPetNet: A DL based AI system for mitotic figure counting on H&E stained whole slide digital images in a large veterinary diagnostic lab setting	2021	CNN (ResNet18, U-Net, EfficientNet, SE-Resnext)	3,845	Canine Feline	Hematoxylin and eosin-stained histologic slides	Improved mitotic count compared to human baselines	Fitzke et al.
Computerized assisted evaluation system for canine cardiomegaly via key points detection with deep learning	2021	CNN (HRNet)	2,274	Canine	X-ray	Average performance: 0.864	Zhang et al.
Comparison of a Deep Learning Algorithm vs. Humans for Vertebral Heart Scale Measurements in Cats and Dogs Shows a High Degree of Agreement Among Readers	2021	CNN	60	Canine Feline	Radiograph	Intraclass correlation coefficient for vertebral heart scale between AI and specialists: 0.998 for both canine and feline	Boissady et al.
Disease Diagnosis of Dairy Cow by Deep Learning Based on Knowledge Graph and Transfer Learning	2021	CNN (KGTL_CNN)	21,649	Bovine	Medical records	F1 score for CNN based on knowledge graph: > 0.85	Gao et al.
A deep learning model for CT-based kidney volume determination in dogs and normal reference definition	2022	nnU-Net , UNETR	386	Canine	CT scan	R = 0.96 between manual voxel count and DL model	Ji et al.

Developing a diagnosis model for dry eye disease in dogs using object detection	2022	CNN (YOLOv5)	95	Canine	Eye video image	mAP: 0.995	Kim et al.
DL in veterinary medicine, an approach based on CNN to detect pulmonary abnormalities from lateral thoracic radiographs in cats	2022	CNN (ResNet50V2)	500	Feline	Radiograph	Accuracy: 0.82 F1 score: 0.85 Specificity: 0.75 Positive predictive value: 0.81 Sensitivity: 0.88	Dumortier et al.
Cytologic scoring of equine exercise-induced pulmonary hemorrhage: Performance of human experts and a DL-based algorithm	2022	CNN (RetinaNet)	52	Equine	Microscope slide	Accuracy: 0.923	Bertram et al.
An automated deep learning method and novel cardiac index to detect canine cardiomegaly from simple radiography	2022	CNN (improved attention U-Net)	1,000	Canine	Radiograph	Left atrial & ventricular enlargement F1 score Vertebral heart score: 0.43 Adjusted heart volume index: 0.55	Jeong & Sung
Deep learning-based diagnosis of feline hypertrophic cardiomyopathy	2023	CNN (Resnet50V2, Resnet152, InceptionResnetV2, MobilenetV2, Xception)	273	Feline	Radiograph	Accuracy: > 0.9	Rho et al.

Deep learning-based diagnosis of stifle joint diseases in dogs	2023	CNN (R-CNN, ResNet)	2,382	Canine	Radiograph	Accuracy: > 0.8	Shim et al.
Canine Mammary Tumor Histopathological Image Classification via Computer-Aided Pathology: An Available Dataset for Imaging Analysis	2023	CNN (VGG16, InceptionV3, EfficientNet)	1,056	Canine	Hematoxylin and eosin-stained histologic images	Accuracy: 0.63-0.85	Burrai et al.
Automated diagnosis of 7 canine skin tumors using machine learning on H&E-stained whole slide images	2023	CNN (EfficientNet B5)	350	Canine	Hematoxylin and eosin-stained histologic images	Accuracy: ~0.95	Fragoso-Garcia et al.
Application of convolutional neural network for analyzing hepatic fibrosis in mice	2023	CNN (Xception)	33	Mice	Whole slide images	Correlation with pathologist hepatic fibrosis grade (r = 0.9067)	Kim et al.
Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing	2023	CNN (U-Net++)	116 Canine 38 Feline	Canine Feline	Hematoxylin and eosin-stained histologic images	Accuracy: 0.92 for canine 0.84 for feline	Haghofer et al.

Automatic grading of intervertebral disc degeneration in lumbar dog spines	2023	CNN (VGG16)	5,991	Canine	MRI	Accuracy: > 0.9 Sensitivity: > 0.83 (except 1 class)	Niemeyer et al.
Use of deep learning for the classification of hyperplastic lymph node and common subtypes of canine lymphomas: a preliminary study	2024	CNN (GoogLeNet)	1,530	Canine	Whole slide images	Accuracy: 0.99	Hubbard-Perez et al.
Automatic classification and grading of canine tracheal collapse on thoracic radiographs by using deep learning	2024	CNN (YOLOv3, YOLOv4, YOLOv4 tiny)	600	Canine	Radiograph	Accuracy: 0.989 Sensitivity: 0.983 Specificity: 0.992	Suksangvoravong et al.
Deep learning-based ultrasonographic classification of canine chronic kidney disease	2024	CNN (YOLOv8-n)	883	Canine	Ultrasound	Accuracy: 0.46	Yu et al.

2.4 Issues for consideration when designing and conducting studies

2.4.1 Accountability and ethical considerations in AI assisted veterinary practice.

The technology to replicate human thought processes in AI analysis is still evolving. The tasks assigned to AI algorithms are diverse and complex, requiring tailored infrastructure for each study (Alzubaidi, Zhang et al. 2021). For example, Vinicki's study required a precise image scoring system consistent with expert evaluations to correctly label images and guide AI in accurately classifying reticulocytes (Vinicki, Ferrari et al. 2018). This underscores the importance of expert consultation throughout the design and implementation phases, ensuring results are interpreted with specialist input for meaningful understanding.

Many studies reviewed here indicate that DL models are most effective when used as decision-support tools rather than standalone diagnostic systems. While accurate and reliable AI applications can enhance patient outcomes, unreliable AI use may introduce risks, particularly in veterinary medicine, where training datasets are often limited. Unlike human medical AI, which benefits from vast, standardised, and well-validated datasets, veterinary AI faces challenges due to species-specific variability and context dependent variables (Coghlan and Quinn 2024). Companion animals such as dogs and cats are more commonly studied, while data for exotic and livestock species remain scarce as shown in this review. In addition, animal health data are often lower in quality, unstructured, inconsistently collected across institutions, and subject to less standardisation and scrutiny compared to human medical datasets (Akinsulie, Idris et al. 2024). These factors significantly impact training quality, as DL models rely heavily on well-curated, high-quality datasets to achieve precise and accurate predictions. Consequently, data imbalance and variability affect model generalisability, resulting in an AI system trained on feline or canine images that may not be directly transferable to other species, even when diagnosing the same

disease. This underscores the critical role of veterinarians in validating AI-generated outputs and ensuring that AI is used responsibly as a supporting tool, rather than as an independent diagnostic system (Appleby and Basran 2022). Proper clinical oversight is essential for ethical and effective AI integration in veterinary medicine.

2.4.2 Sample size and data quality

A large volume of training data is crucial for developing and testing reliable DL applications (Kokol, Kokol et al. 2022). Although specific sample size requirements vary by context, certain factors can help estimate the sample size required for effective modelling (Baeza-Delgado, Cerdá Alberich et al. 2022). The size and complexity of the model significantly impact the required dataset, as demonstrated in Krizhevsky's study where training a deep CNN model required over 1 million labelled images (Krizhevsky, Sutskever et al. 2017). Although not directly mentioned, it can be inferred that the number of predicted classes affects the required sample size, given the study's aim to classify 1,000 different classes. In contrast, veterinary medicine often focuses on predicting fewer classes, with currently annotated and labelled data being limited, especially within specific veterinary imaging cases, where small sample sizes and unbalanced classes are common limitations (Dhar, Dey et al. 2023). Veterinary AI studies often have smaller sample sizes due to factors such as having smaller and more dispersed patient (companion and livestock) populations, which makes data collection difficult as veterinary records are often heterogenous (McGreevy, Thomson et al. 2017). In addition, species-specific variability requires separate datasets for different animals, further fragmenting available data for prospective studies. Hence sourcing animal samples that meet specific disease criteria is big challenge, often necessitating retrospective data collection from historical health records, which may contain missing or inaccurate information (Lustgarten, Zehnder et al. 2020). However, increasing sample size alone is not always beneficial if the data quality is compromised. If poorly labelled, inconsistent, or inaccurate ground truth data are used, AI models may exhibit misleading performance

gains while lacking true clinical utility. As veterinary AI research progresses, larger and more diverse datasets will be needed, but they must be carefully curated, standardised, and validated to maintain model reliability (Shahinfar, Meek et al. 2020). In human medicine, one approach to addressing small sample sizes is the use of pre-trained large models, which can be fine-tuned for specific tasks while leveraging existing large scale datasets (Mazurowski, Dong et al. 2023). This is known as transfer learning, which is a promising approach that offers significant potential for further exploration and application in the field of animal health, particularly in settings where data availability is constrained. Data augmentation is another potential solution as it is a method used to enhance the variety of the training dataset by applying various transformations to the existing data, producing altered versions that remain representative of the original dataset (Shorten and Khoshgoftaar 2019). This artificially inflates the dataset and increases the effective sample size for the training of the model, which can help to teach AI models more diverse features aiding in its classification performance.

2.4.3 Evaluation and validation challenges

Ensuring the reliability and generalisability of DL models is highly dependent on the validation strategies employed. While a clean and well-structured dataset is fundamental, models must undergo a rigorous validation process to confirm that they can perform beyond the controlled environment in which they were trained. Without appropriate validation, even highly optimised models may exhibit overfitting, suffer from dataset biases, or fail to generalise to real-world clinical applications, limiting their practical use (Eelbode, Sinonquel et al. 2021). One of the major challenges in veterinary AI research is the lack of universally standardised and fully labelled validation datasets, which complicates fair comparisons between different AI models. Unlike human medical AI, where large-scale benchmark datasets exist for model evaluation, veterinary AI studies often rely on institution specific datasets, making it difficult to directly compare performance across different studies.

This absence of standardisation can lead to inflated performance metrics if validation datasets do not accurately represent real world conditions.

The selection of validation strategies is therefore crucial. Internal validation, where models are assessed using the same dataset on which they were trained, is an essential initial step but does not account for variations in clinical practice, different imaging modalities, or novel cases encountered in real-world settings (Appleby and Basran 2024). External and clinical validation are necessary to test the model's performance across diverse datasets, imaging techniques, and patient populations. Without these steps, a model that appears to perform well under controlled conditions may fail when deployed in practice. Additionally, the quality of validation datasets plays a significant role in model reliability. If datasets are small, unbalanced, or inconsistently annotated, AI models may demonstrate misleading performance gains while lacking true clinical utility. Poor validation design can also lead to circular reasoning, where models inadvertently learn dataset-specific artefacts rather than true disease characteristics.

Given these challenges, AI studies should transparently report their validation methodologies to allow accurate interpretation and comparability across research. Standardising validation frameworks for veterinary AI (Hartung and Kleinstreuer 2025), including establishing common datasets and benchmarking protocols, would be a critical step in improving model assessment and ensuring AI applications are both scientifically rigorous and clinically relevant. Despite the necessity of these validation steps, there remains a lack of regulatory validation strategies that apply to all DL modelling, particularly in veterinary applications. Again, unlike in human medical AI, where regulatory frameworks are more established, veterinary DL models are often developed with less stringent oversight. This gap in regulatory validation could lead to risks for animal patients if veterinary experts are not actively involved throughout the entire model development lifecycle, including post-implementation monitoring and error reporting (Cohen and Gordon 2022).

2.4.4 Data analysis workflow

An effective and smooth analysis pipeline is key to the successful and efficient building of DL networks. General principles to follow include data preprocessing techniques like data augmentation, normalisation, and class imbalance handling to ensure the model robustness (Patterson and Gibson 2017). It is crucial to understand the characteristics of the images and provide accurately annotated data for processing. Lapses in these areas can lead to poor performance or unexpected results. Choosing the most appropriate model is essential, considering factors like the number of classes and problem complexity. Network architecture design and careful selection of hyperparameters, such as the loss function, optimisation algorithm, and learning rate, significantly affect model convergence and performance (Patterson and Gibson 2017).

2.4.5 Black box approach

DL methods are often called “black-box” approaches due to their complex and opaque inner workings, making them difficult to interpret. None of the reviewed veterinary health studies attempted to address this issue in a technical fashion, highlighting the nascent stage of DL research in this field. However, explainable AI (X-AI) approaches, such as GradCAM, which generates gradient maps, and LIME (Local Interpretable Model-agnostic Explanations), which explains how input modifications affect model predictions, are used in human medical image analysis (Dhar, Dey et al. 2023). Researching and implementing X-AI in veterinary medicine could enhance understanding and improve imaging diagnostics. Beyond technical interpretability, transparency is also a critical consideration. Veterinary users need to understand and trust AI driven tools before integrating them into clinical practice (Coghlan and Quinn 2024). A lack of transparency in how these models generate predictions may hinder their adoption, reinforcing the need for X-AI methods to enhance trust and usability in veterinary diagnostics. Researching and implementing X-AI in veterinary medicine could

improve both interpretability and confidence in AI-assisted imaging diagnostics.

2.5 Opportunities for future DL applications in the animal health domain drawing inspiration from human health and other domains

DL techniques in human health can guide and indicate the potential for veterinary medicine. Application of AI image analysis for detecting skin tumours, bone fractures, and lung infections in humans could be similarly developed and adapted for diagnosing animal diseases (Bhatt, Kumar et al. 2021). The application of transfer learning in veterinary medicine for these types of disease classification is a potential avenue for exploration as these types of disease research are currently limiting in the animal health context.

AI can transform healthcare delivery by improving efficacy, accessibility, and personalisation. For example, human healthcare providers utilise monitoring devices and smart phone technology to obtain real-time patient vitals for monitoring purposes (Reddy, Fox et al. 2019). This area could be adapted to track animal health indicators, enhancing veterinary health management. Virtual assistants with NLP capabilities, which provide health and medication information post-hospital visits, have shown improved patient outcomes and could similarly benefit animal healthcare, particularly in rural and underserved areas. NLP techniques that extract information from human clinical records can also be applied to veterinary health records for clinical research, revealing additional information to support veterinarians (Sheikhalishahi, Miotto et al. 2019). Even though this type of research is still relatively nascent in human medicine but as this technology improves, it could potentially be implemented in the analysis of veterinary health records to support better decision making by veterinarians. In another aspect of medicine, CNNs have been used to track and analyse surgical procedures to assess surgeon performance during medical training (Jin, Yeung et al. 2018). This application holds promise in veterinary medicine for training new

surgeons and providing individualised feedback to assist and support veterinarians in training.

Human medicine has successfully utilised specific models such as BERT for natural language processing and ResNet for image classification, achieving high accuracy and robustness (Rasmy, Xiang et al. 2021, Xu, Fu et al. 2023). These successes are partly due to the availability of large, diverse datasets and synthetic data generation techniques, which help augment training datasets and improve model performance. In contrast, veterinary medicine lacks similarly extensive datasets and established models, making it a prime candidate for applying and adapting these advanced techniques. The success of synthetic data in human studies, enhancing model training and performance, suggests a promising avenue for veterinary applications, potentially mitigating the challenges posed by limited real-world data.

2.5.1 Rapid development in AI

AI is advancing at an unprecedented rate, with new models continually emerging in human medical applications. A recent study published by Google introduced Med-PaLM, formerly known as MultiMedQA, a comprehensive benchmark for evaluating the clinical knowledge of large language models (LLMs) across various medical topics (Singhal, Azizi et al. 2023). While Flan-PaLM, a 540-billion parameter LLM, achieves state-of-the-art accuracy on Med-PaLM datasets, human evaluations reveal key shortcomings in areas such as comprehension and reasoning, underscoring the need for improved evaluation frameworks and methodologies to make LLMs safe and useful in clinical settings (Harris 2023). Med-PaLM and its evaluation framework hold significant potential for adaptation in veterinary medicine. By tailoring these benchmarks to veterinary-specific datasets, it is possible to assess and enhance the accuracy and safety of LLMs in diagnosing and treating animal health conditions ensuring that LLMs can effectively support veterinary professionals.

The recent introduction of Med-Gemini, a family of highly capable multimodal models specialised in medicine, highlights the rapid development in AI. Med-Gemini models excel in advanced reasoning, up-to-date medical knowledge access, and complex multimodal data understanding. It achieved state of the art performance on 10 out of 14 medical benchmarks (Saab, Tu et al. 2024). Med-Gemini's capabilities to interpret and analyse complex data could be adapted to diagnose animal health conditions using diverse data sources, including images, health records, and sensor data. Moreover, its capability to surpass human experts in medical text summarisation and video question answering could enhance veterinary training and decision-making processes. These strengths suggest promising potential for the applications of these tools in veterinary medicine, providing more accurate and comprehensive diagnostic tools and improving overall animal healthcare.

The adaptation of transformer models, initially developed for natural language processing, has opened new frontiers in veterinary computer vision applications. The "transformer" architecture, based on a self-attention mechanism, allows the model to weigh the relative importance of different features independently of their order, providing a more flexible and nuanced approach to data interpretation (Vaswani 2017). This approach has demonstrated comparable performance to CNNs in image classification tasks (Han, Wang et al. 2023). In a recent medical imaging study, a Vision Transformer (ViT) was used to analyse PET brain scans, classifying healthy tissue versus Alzheimer's disease, and outperformed the CNN-based VGG19 (Shin, Jeon et al. 2023). In oncology, ViT models have shown superior performance in classifying skin cancer from lesion images (Himel, Islam et al. 2024). Another study highlighted ViT's potential in detecting tuberculosis in chest X-rays, where a hybrid approach incorporating a ViT component with a CNN backbone achieved higher classification performance than a standalone CNN (Duong, Le et al. 2021). These successes suggest that researching ViTs in veterinary health contexts could improve diagnostic accuracy in medical imaging, which is crucial for effective treatment and disease control in

animals. Moreover, generative adversarial networks (GANs) and variational autoencoders (VAEs) showed promise in mitigating the need for extensive labelled datasets in medical image analysis (Frid-Adar, Diamant et al. 2018). By synthesising realistic medical images, these models can augment training data, potentially enhancing the performance and generalizability of DL models even with limited real-world samples (Chen, Lu et al. 2021). Such approaches could be useful in veterinary diagnostic where limited images are currently available. In summary, the rapid development of AI technologies, exemplified by models like Med-Gemini, Med-PaLM, ViTs, GANs, and VAEs, presents significant opportunities for advancing veterinary medicine. By adapting these technologies and benchmarks to veterinary contexts, the field can benefit from improved diagnostic tools, more robust data augmentation techniques, and ultimately, better health outcomes for animals.

2.6 Conclusion

This review has highlighted the application of DL in veterinary medicine. This is a rapidly evolving area of research with increasing attention on its use in the veterinary healthcare industry in recent years. Its advantages are better understood and can be utilised to benefit many aspects of the industry as seen in the examples discussed, particularly in image analysis, which is enabling health specialists and farmers to develop optimal and timely treatment and prevention plans for the best possible health outcomes for affected animals.

The creation of training datasets for veterinary diagnostics is both labour-intensive and costly, presenting a significant bottleneck in the broader application of AI within this field. The limited availability of comprehensive datasets, encompassing various diagnostic modalities, further constrains the successful deployment and optimization of AI-driven tools in veterinary diagnostics.

This review underscores the urgent need to create standardised, high-quality large training datasets that include a wide array of diagnostic modalities and

animal species. Given the inherent diversity of species within veterinary practice, fostering international collaboration is not only advantageous but also essential for the successful implementation and fine-tuning of AI models in veterinary diagnostics. To facilitate this crucial endeavour, we suggest the formation of an international consortium focused on veterinary phenomics for AI. Such a collaborative framework would not only accelerate the assembly of comprehensive and interoperable datasets but also catalyse advancements in AI-driven veterinary diagnostic techniques, thereby elevating the quality and efficacy of animal healthcare globally.

Chapter 3. Object detection for pre-processing Cattle eye images in IBK classification

3.1. Introduction

Infectious bovine keratoconjunctivitis (IBK), more commonly known as pinkeye disease, is an ocular infection found in cattle. It causes a wide list of symptoms that include intermittent or chronic excessive lacrimation, eye discharge, blepharospasm, inflammation, photophobia, epiphora, and ulceration (O'Connor, Shen et al. 2012). Some of these symptoms can lead to severe damage to the eye causing significant health and welfare issues as well as economic and production loss to cattle farmers as severely affected animals cannot be traded. Reports have estimated that IBK is an important eye disease that affects cattle that contributes to a significant loss to cattle farmers in Australia (Kneipp, Govendir et al. 2021), which highlights the importance of detecting this disease early and applying treatments promptly to curb its impact on cattle production in the agriculture industry.

The overall aim of my PhD research is to develop a deep learning pipeline to achieve the goal of scoring and identifying the different stages and severity of IBK from cattle eye images. To develop such deep learning models, a large number of images of cattle eyes are required. These images can be captured by various producers and veterinarians using different electronic devices under a variety of conditions. Due to this, the coverage of the cattle eye in the images taken by potential users may vary, with some images capturing the eye region closely, while others may not. To facilitate the training of deep learning models for pinkeye disease detection, it is crucial to ensure that the proportion of the eye region coverage is broadly consistent across all the images and that other insignificant objects captured are removed. Therefore, it is necessary to develop a model capable of detecting eyes in the images and extracting the eye region of interest for the next step in the modelling

process. This can be accomplished through the use of object detection algorithms to automate this process, to mitigate the manual labour required.

Object detection is a computer vision method used to identify, locate, and classify objects within an image or video (Zhao, Zheng et al. 2019). This includes the drawing of a box known as a bounding box around the identified object in the image or video. The coordinates of these bounding boxes also provide the location information of the identified object. The bounding box can be optionally tagged with the object label. In a generic example presented in Figure 3.1, a bounding box is drawn tightly around each animal, and the algorithm also outputs the correct labels for each bounded dog and cat.

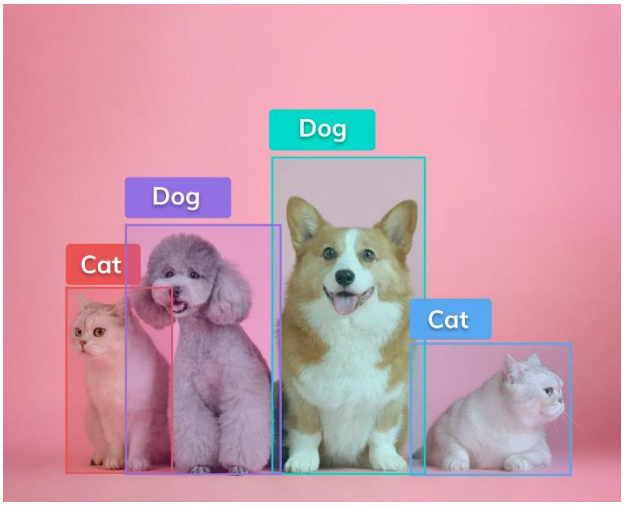


Figure 3.1. The animals in the image are each enclosed by a bounding box and correctly identified and labelled adapted from Rizzoli (2021)

This idea is straightforward and can be performed to a high level manually. However, with the large volume of images generally required for training deep learning models, this manual labelling process is laborious and slow. Thus, we aim to automate this process to speed up the detection and provide users with quicker results. Computer systems have been developed to mimic the successes of humans in object detection in recent years, with the development of deep learning techniques to hasten and automate this process (Krizhevsky, Sutskever et al. 2017). Deep learning-based strategies

use convolutional neural networks (CNN) as their underlying mechanism to accomplish end-to-end object detection (Zou, Chen et al. 2023). The CNN is a type of neural network that is especially useful in parsing image data to perform tasks such as object detection. It achieves this through its network architecture, where its neurons, which are the most fundamental units of processing in the AI algorithm, are arranged in potentially many complex layers to extract, learn, and analyse the different features of the input image.

The dataset we are working with contains images from cattle farms in NSW and Queensland Australia, as well as images obtained from cattle farms in the USA. Example images of the dataset are provided in Figure 3.2, showcasing an image with pinkeye and an image containing a normal eye from the NSW images, and a cattle eye from the USA images.

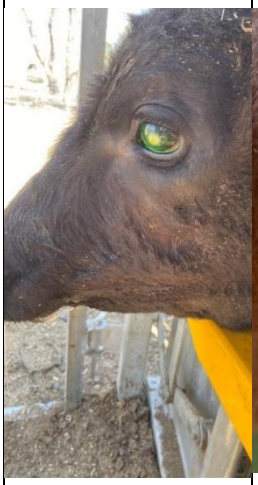
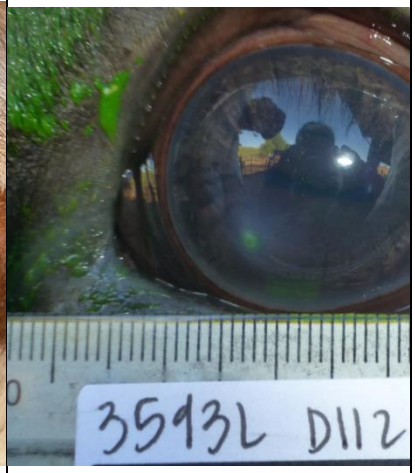
a) Pinkeye from NSW	b) Normal eye from NSW	c) Pinkeye from the USA
		

Figure 3.2. A comparison of three obtained sample images.

Whilst each image in Figure 3.2 is of high quality with the eye of interest being the centre of the image, however, as can be seen on in image A) in Figure 3.2, the picture contains a bright yellow railing, bright sun coloured background which may distract deep learning models from learning the important features that can lead to the incorrect identification of pinkeye. Some other pictures contain a variety of background objects such as fence

posts, hay, clothing, people, tags, among other things. In addition, many images from the USA in our dataset contain a ruler, which although does not obstruct most of the eye, this unique feature is still very distinct from the NSW images, which means it can also impact the performance of the deep learning model in its classification down the pipeline. These background and/or foreground objects can sometimes provide white noise signals, which obscure the deep learning algorithm from correctly analysing the conditions or the image of the eye itself, leading to reduced model performance (Szegedy, Toshev et al. 2013). Literature has shown that removing these background objects or white noise can help to improve the accuracy of the image analysis (Szegedy, Toshev et al. 2013) of the main focus, which in this project is the eye.

Thus, the objectives of this chapter were to:

- train a mainstream deep learning object detection technique, to automatically draw a bounding box to identify the eye in each of the cattle eye images in our datasets (both NSW and USA combined).
- evaluate the performance of this object detection algorithm under different conditions to identify the most effective and efficient approach for this object detection task.

3.2. Background/development of the object detection pipeline involved in the methodology

To address the aim of object detection in this chapter, three standout mainstream object detection approaches – Faster R-CNN, Single Shot Detector (SSD), and You Only Look Once (YOLO)– were considered.

The Faster R-CNN method involves utilising an R-CNN network, which is a region-based convolutional neural network algorithm where analysis takes place in a two-step approach (Srivastava, Divekar et al. 2021). Step 1 involves utilising a selective search method where pixels of similar groupings and

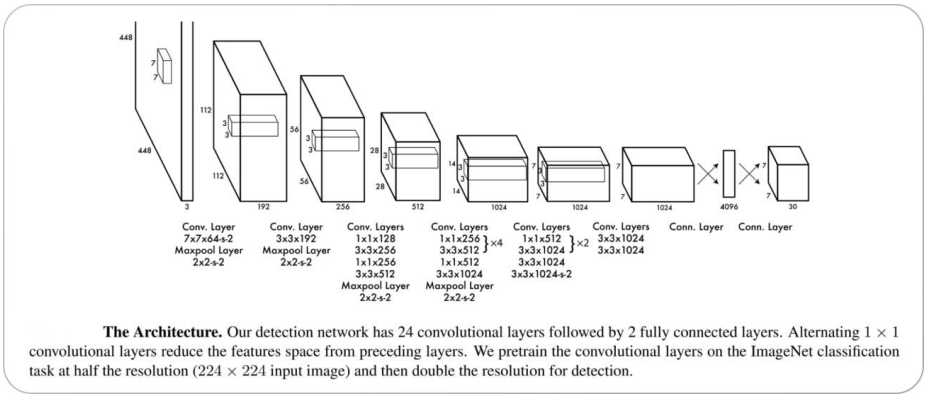
features (e.g. colours, light gradient, texture, etc.) are extracted as groupings or regions known as proposed regions. This creates many proposed bounding boxes, which are cropped and passed through to the CNN network, which constitutes step 2 of the analysis. The Faster R-CNN contains the ROI pooling layer, which unifies all the proposed regions of interest of varying sizes into fixed sizes to be passed to the fully connected layers for class classification and bounding box prediction (Girshick 2015). A fully connected layer uses an objectiveness score for each proposed region to predict whether the region contains the object or not and produces the class label if it contains the object. The other fully connected layer is the regression layer to predict the bounding box and its coordinates. This two-step approach serves as the underpinning fundamental principle behind how more updated R-CNN approaches work (Girshick, Donahue et al. 2014). The main advantage of using the Faster R-CNN is its speed over other algorithms due to its shared convolutional network layers between its region proposal network (RPN) component and its Fast R-CNN component (Ren, He et al. 2016). This layer sharing property allows the Faster R-CNN to act as a unified model in which both components can either be trained as one or each component can be alternatively trained using weights retrieved from the previous convolution step. Research has shown that this process has faster speeds and higher accuracies in achieving object detection compared to its predecessor algorithms, the RCNN and the Fast RCNN (Liu, Ouyang et al. 2020). The Faster R-CNN has shown significant effectiveness in specific applications, such as in the agricultural industry, where it has achieved high accuracy in detecting diseased leaves in tomato plants (Priyadharshini and Dolly 2023). Additionally, in the livestock industry, Faster R-CNN has been utilised to detect features of cattle locomotion, which is crucial for automatic lameness detection (Gardenier, Underwood et al. 2018).

Single-shot detector (SSD) is another one-step approach algorithm, which, as the name suggests, locates and classifies objects and produces the bounding box for each object in a single forward pass from input to output (Liu,

Anguelov et al. 2016). SSD's network architecture consists of two main components. The input image is first fed to a CNN backbone model, which can be a pretrained ResNet model. The convolutional layers extract features from the input image to be processed and passed to the second component of the SSD. The SSD architecture utilises the principle of having more convolutional layers to enlarge receptive fields to predict bounding boxes for objects at different scales. The receptive field measures the association between an input patch of the image and an output feature, which is presented on the feature map (Zeiler and Fergus 2014). As the feature extraction analysis progresses deeper through the CNN layers, the output features develop to represent larger input patches, thus allowing more information from the input to be captured. Features of the same feature map all represent receptive fields of the same size when looking for significant pixels or regions to recognise, but at different locations of the input. It is ideal to have convolutional units in the feature map to have a large receptive field to not miss any significant information, which can be achieved by increasing the layers of the CNN backbone algorithm. Since each convolutional layer of the SSD, representing different receptive field sizes, produces an object detection result, these predictions can be utilised to predict bounding boxes for objects of different sizes. Single-shot Multibox Detection (SSD) networks are known for their accurate object detection capabilities, especially on datasets like VOC2007, where they outperform Faster R-CNN (Liu, Anguelov et al. 2016). However, SSDs' high computational demands limit their usage in real-time applications (Vrbaski, Josic et al. 2023). While real-time processing is not a critical requirement for the current task, which involves identifying a single object class per image, the need for extensive object detection algorithms as required by multiclass challenges like VOC2007 is also unnecessary.

The third object detection algorithm YOLO, is known as a one-step algorithm which only makes one forward propagation pass through its CNN network to make object detection/bounding box predictions (Deng, Xuan et al. 2020). It

works by first dividing an image into specifically sized S × S grids (Jiang, Ergu et al. 2022). For each grid, the CNN predicts a separate output for the number of bounding boxes and the confidence scores for these boxes. YOLO uses the highest 'Intersection over Union' (IOU) to determine the success of the prediction of the bounding boxes as compared to the ground truth. This eventually trains the predictions to improve their ability to ascertain sizes, aspect ratios, and object classes. Further, YOLO uses non-max suppression, a processing step to produce a single bounding box for each object in the image. A big issue in object detection is that objects often get detected multiple times instead of just once. These extra boxes, which often overlap, produce too many white noise signals, which will slow down the processing time for the algorithm. Thus, NMS comes in to discard these irrelevant boxes to produce the one accurate bounding box around the object of interest.



The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

Figure 3.3. YOLO general network architecture illustrating its single-stage convolutional neural network pipeline for real-time object detection using grid-based localisation and classification. Adapted from Redmon, Divvala et al. (2016).

The YOLO network (Figure 3.3) has been extensively researched in diverse fields of computer vision tasks, ranging from detecting pedestrians in autonomous vehicle computer vision tasks to detecting abnormal breast cell masses in mammogram images (Terven, Córdova-Esparza et al. 2023). In the field of agriculture, YOLO has also shown high precision of approximately 94.5% in the detection of pests from images with embedded data augmentation techniques (Lippi, Bonucci et al. 2021). The Faster RCNN

algorithm has also performed to a high accuracy of detecting diseased leaves in tomato plants in the agricultural industry (Priyadharshini and Dolly 2023). In the livestock industry, Faster RCNN was utilised to detect features of cattle locomotion that can point to lameness detection (Gardenier, Underwood et al. 2018). Although the SSD networks have been shown to perform more accurate object detection for datasets such as the VOC2007 compared to the Faster RCNN (Liu, Anguelov et al. 2016), it is also computationally very demanding, which limits their usage in real-time applications (Vrbaski, Josic et al. 2023). While this was not a major limitation during the development and training phases of our research, it becomes more relevant in deployment. Specifically, when the object detection component is integrated into a mobile or field-based application to automatically extract the eye region, the process must operate in near real-time, making inference speed a critical consideration. Additionally, our task differs from VOC2007 in that we are only detecting a single object class per image, rather than performing multiclass object detection, which reduces the need for more complex architectures.

Rapid improvements have been made to the basic YOLO fundamental architecture, and the YOLO algorithm has gone from version 2 to the most state-of-the-art version 8. YOLOv2 utilises batch normalisation and implements anchor boxes to improve model stability and object detection accuracy compared to the initial YOLO algorithm. YOLOv3 utilises the backbone network known as Darknet-53 to better detect objects of different sizes and scales at the cost of increased hardware demand. The advancement of YOLOv3 to YOLOv4 was marked by the change of the backbone network to the CSPDarknet53 architecture, which demonstrated improved accuracy and speed compared to other object detection algorithms available at the time. YOLOv5 is designed to be faster and more accurate for a broader range of object classes (Zhao, Zheng et al. 2019). It utilises a network architecture called EfficientNet, which is a much more complicated and intricate CNN network serving as its backbone (Tan and Le 2019). EfficientNet is an improvement as it is a family of neural network architectures with fewer

parameters than the previous backbone networks. It also introduces a compound scaling method, which scales the neural network in terms of depth, width and resolution simultaneously in a balanced way, which creates a more efficient neural network (Tan and Le 2019). YOLOv5 was also developed on a large and varied dataset called the D5 that contained 600 object classes, which allows the network to learn the diverse features to look for to produce a more accurate output. In addition, YOLOv5 utilises a mechanism known as “dynamic anchor boxes” to cluster all the ground truth boxes to produce centre coordinates as anchor boxes to predict bounding boxes which are more accurate to the object’s dimensions. Moreover, spatial pyramid pooling was added as an improvement to the architecture, which is a pooling layer added on top of the final convolutional layer of the CNN. This layer pools the features at different spatial scales to generate fixed-length outputs (He, Zhang et al. 2015). This means that regardless of the input size, the output generated is still of a fixed length. This spatial pyramid pooling mechanism removes the fixed-size restriction of the CNN network, allowing it to handle varying input sizes. It has been found that this step allows the network to better detect smaller objects as the model is able to see and learn objects at varying scales, contributing to improved overall performance, especially in scenarios where objects have different scales.

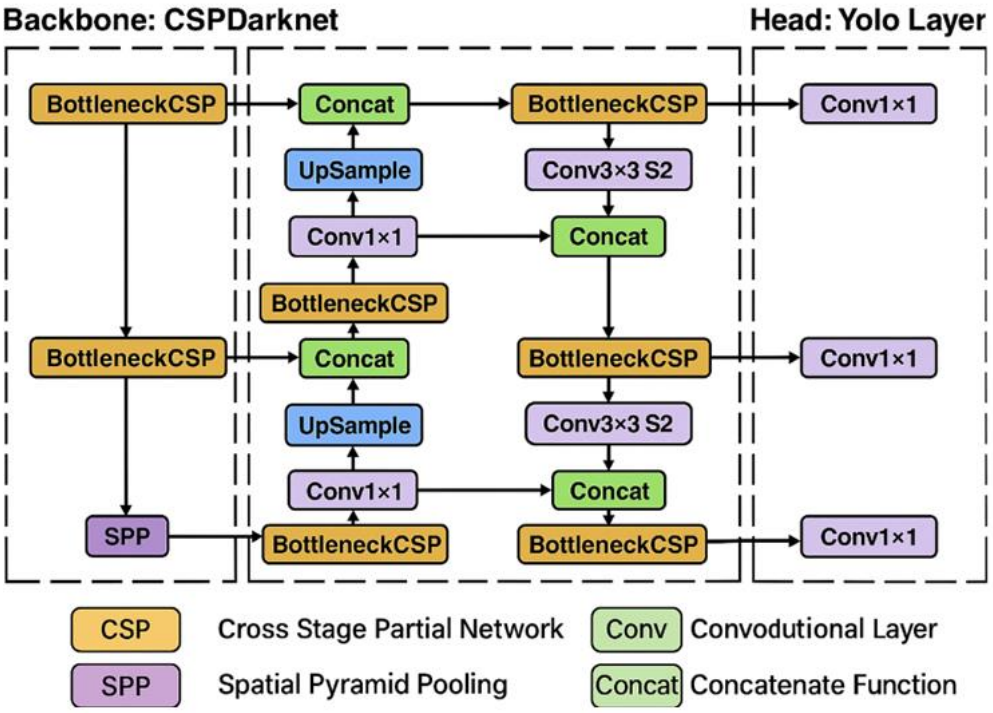


Figure 3.4. YOLOv5 backbone architecture. This figure illustrates the structure of the YOLOv5 backbone, highlighting the sequence of convolutional and CSP (Cross Stage Partial) layers used to extract hierarchical features from input images for object detection. Adapted from Xu, Lin et al. (2021) .

Overall, the developments and improvements introduced in YOLOv5's network architecture (Figure 3.4) make it faster and more accurate for a diverse range of object detection tasks. YOLO has been applied across a diverse range of domains, from detecting pedestrians in autonomous vehicle systems to identifying abnormal cell masses in mammogram images (Terven, Córdova-Esparza et al. 2023). In agriculture, YOLO has demonstrated high precision, achieving approximately 94.5% accuracy in detecting pests from images, especially when combined with embedded data augmentation techniques (Lippi, Bonucci et al. 2021). After an initial exploration of the current state-of-the-art object detection algorithms available, we decided to use YOLO to satisfy the specific objective, which is to predict the bounding box coordinates for the eye area of the cattle images. YOLOv5 was chosen due to a combination of its speed, simplicity, and efficiency in tackling a relatively straightforward object detection task.

3.2.2 Explanation of data augmentation

Data augmentation is a well-practised and documented technique in computer vision and object detection problems. It enhances the variety of the training dataset without the need for additional data collection by applying various transformations to the existing data, producing altered versions that remain representative of the original dataset (Shorten and Khoshgoftaar 2019). This applied transformation to the images also increases the effective sample size of the dataset.

Some of the more common data augmentation transformations which will be applied to our dataset to produce the augmented dataset are outlined below (Perez and Wang 2017):

1. Rotation

Rotation involves rotating the image by a certain angle to assist the model's capabilities to detect the object at different orientations.

2. Scaling

Images can be rescaled to a different size to help the model to detect the objects of interest at a variety of sizes.

3. Flipping

Images can be flipped horizontally or vertically across the middle axis again to improve the model's capabilities to detect the object of interest at different orientations (see Figure 3.5.)

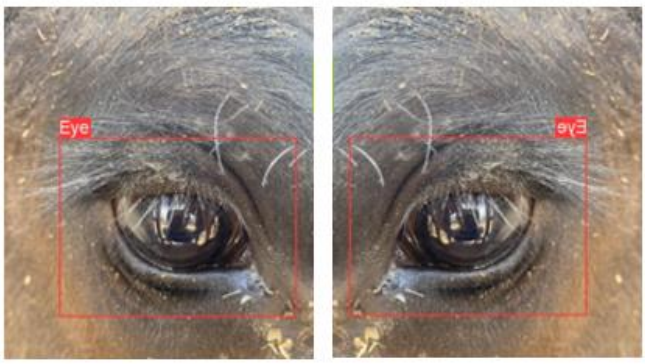


Figure 3.5. Image is flipped along the vertical axis

4. Mosaic transformation

Mosaic transformation involves combining four random training images into a single mosaic image. These images are divided into four quadrants, and then the quadrants are combined to form a mosaic image (Hao and Zhili 2020). The bounding boxes of the objects in the original images are adjusted to reflect their positions in the mosaic, which involves updating the coordinates of the bounding boxes based on their relative positions in the mosaic. The labels associated with the objects are updated to account for the new bounding box coordinates. Mosaic augmentation introduces variations in backgrounds, scales, and object positions, simulating diverse scenes and perspectives within a single training example, enhancing the model's ability to generalise to different scenarios and contexts (Hao and Zhili 2020). It also reduces the risk of overfitting to specific scenes or patterns.

One main purpose of data augmentation is to improve the model's ability to identify and locate objects under a variety of conditions. By exposing the model to a training dataset that includes a more extensive range of variations of the objects of interest, it helps train a model that is more robust to analysing real-world variations of the object in the input data (Mumuni and Mumuni 2022). For example, in future testing images, a submitted input image may contain a cattle eye that is angled or of a different size in a particular picture, and we want the algorithm to still be able to recognise the eye and draw a suitable bounding box around the object.

In addition, data augmentation plays a crucial role in achieving a balanced training dataset by generating additional examples of underrepresented classes or scenarios. This approach mitigates the risk of the model developing a bias towards more prevalent classes, thereby enhancing its performance across all classes (Temraz and Keane 2022). By ensuring that the model is exposed to a more equitable distribution of examples, data augmentation

facilitates a more robust and generalised learning process, enabling the model to perform consistently well across diverse categories.

Another major advantage of implementing data augmentation is helping to mitigate overfitting problems in deep learning models (Shorten and Khoshgoftaar 2019). Overfitting is a significant issue for object detection, as often these algorithms perform highly accurately on the training data but can struggle to extend this high-level performance to unseen testing data. Data augmentation exposes the model to a broader and more varied set of training examples, which helps the model improve its ability to discern patterns and features that remain consistent across diverse transformations. As the model learns from augmented data, it enhances its capacity for generalisation beyond the specifics of the original training dataset.

Moreover, data augmentation can reduce the laborious burden of annotation of a large number of images to form a dataset with a large sample size. The process of manually annotating a large dataset for object detection can be time-consuming, resource-intensive, and mentally exhausting. Data augmentation reduces the need for extensive manual annotation by creating additional training examples through transformations applied to existing annotated images. A large training set is usually required to meet a certain performance threshold for an object detection algorithm, especially if the computer vision task is complex (Montserrat, Lin et al. 2017). The image dataset we are working with in this chapter contains a maximum of 1,000 images to test the required sample size to achieve high-accuracy object detection. The sample size is lower than what was suggested in another study (Montserrat, Lin et al. 2017) which consists of 5,000 images per class for certain object detection algorithms, because of the simplicity of our computer vision problem. We have annotated the images in our dataset, then performed data augmentation of scaling, rotation, flipping and mosaic transformation to increase the effective training sample size to test if this improves model performance.

In summary, data augmentation has proven to be a potent technique, not only elevating a model's capacity for generalisation and resilience but also streamlining the training process by lessening the need for laborious manual annotation. With a meticulously crafted augmentation strategy, models can adeptly navigate an extensive array of conditions and scenarios, ultimately refining their performance in real-world applications. By leveraging data augmentation techniques, it has been consistently shown to notably boost model performance and generalisation (Khosla and Saini 2020), which supports the reasoning behind applying such techniques to improve our object detection model.

Since our computer vision problem of identifying just one object per image is not particularly complex, more complex augmentation techniques such as noise addition and contrast adjustment will not be implemented. These common techniques of rotation (rotate by 90 degrees anticlockwise), scaling, flipping (horizontally and vertically), and mosaic transformations will be implemented instead to produce an augmented dataset during the training phase of the object detection algorithm to create a more robust model. During the testing phase, evaluation metrics discussed were used to compare the performance of the object detection model trained on the augmented dataset to the object detection model trained on the unaugmented dataset, to highlight any differences in the impact of data augmentation on our computer vision problem.

3.2.3 Explanation of the evaluation metrics

3.2.3.1 mAP explained

The YOLOv5's object detection effectiveness and accuracy for the purpose of this chapter are evaluated with the commonly used assessment statistics for object detection computer vision problems, which is the Mean Average Precision (mAP). It operates upon the concept of "Intersection over Union (IoU)", which is an evaluation metric used to determine the accuracy of the predicted bounding boxes for any object detection algorithm (Rezatofighi, Tsoi et al. 2019). It is calculated over a simple equation as follows:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad (3.1)$$

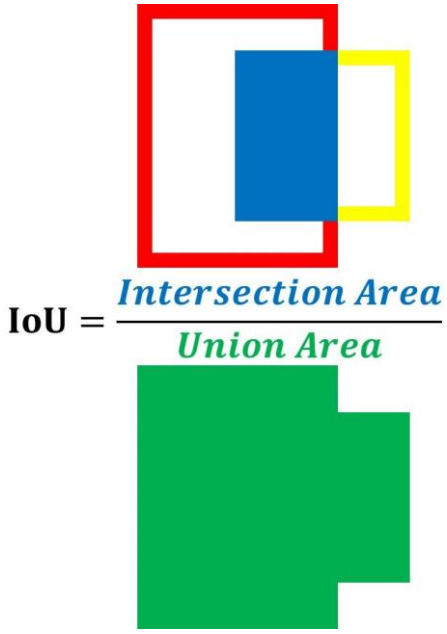


Figure 3.6. Demonstration of the IoU

Once the predicted bounding box is drawn by the YOLOv5 algorithm, the area that overlaps between the ground truth bounding box and the predicted bounding box for a particular object in the input image is determined to form the numerator (Figure 3.6). The denominator consists of the area of the union, which is the area of the image covered by both the ground truth bounding box and the predicted bounding box (Xiao, Tian et al. 2020). Since the area of overlap cannot be greater than the total area covered by both the predicted and the ground truth bounding boxes, the IoU measure outputs a value between 0 and 1. If the predicted bounding box overlaps greatly with the ground truth bounding box, then the IoU score would be closer to 1 because the numerator is large, which means that the prediction was highly accurate. If the predicted bounding box overlaps very little with the ground truth bounding box, then the IoU would be closer to 0. In object detection, it is a common practice to use an IoU score of 0.5 as the binary divider between

what is classified as a good prediction vs. a bad prediction, as illustrated in Figure 3.7.

Figure 3.7. Illustration of the comparison between poor, good and excellent IoU

In object detection, it is rare or unrealistic to expect excellent IoU scores to be consistently outputted due to the many varying factors in the complete image analysis process. Good IoU scores featured in Figure 3.7. can often localise the object of interest to a high degree of accuracy for functionality in day-to-day applications.

IoU is used for the detection of each object for each input image. However, analyses are conducted using many images, which means a combined measure using each image's IoU is used to assess the overall performance of the model over all its predicted bounding boxes over the dataset. The mean average precision (mAP) measure incorporates the IoU in its calculation process to produce a score to assess the performance of the object detection model (Everingham, Van Gool et al. 2009) as a whole for the entire set of images. An IoU score greater or equal to a set threshold, say 0.5, will prompt the model to consider this prediction to be a positive, meaning the model could detect the object of interest correctly. An IoU below the same set threshold is considered to be a negative, which means that the predicted bounding box does not contain the object of interest. This binary classification idea is represented as follows:

$$class(IoU) = \begin{Bmatrix} Positive \rightarrow IoU \geq threshold \\ Negative \rightarrow IoU \leq threshold \end{Bmatrix} \quad (3.2)$$

The predicted positives are compared to the ground truth positives and the false positives of the dataset as follows to find the precision measure.

$$Precision = \left[\frac{True\ Positives}{True\ Positives + False\ Positives}\right] \quad (3.3)$$

For example, if the model detected five eyes in the image and the ground truth of the image contains four eyes, then that means ground truth positive is four, and the false positive is one, which gives a precision of the model four over five which is 80%.

In addition, the predicted positives are compared to the ground truth positive and false negatives according to the following:

$$Recall = \left[\frac{True\ Positives}{True\ Positives + False\ Negatives}\right] \quad (3.4)$$

This means that if a model fails to detect and draw a correct bounding box around an object of interest in the image then the “false negative” portion of the denominator will increase, affecting the recall of the model. For example, if the model detects 9 eyes but there were in fact 10 eyes in the image, then that makes 1 false negative, producing a recall of 9/10, which is 90%. Due to the nature of how precision and recall are calculated, there is a trade-off between these two measures (Figure 3.8), which is characteristically represented by the following graph, where an increase in one often results in the decrease of the other measure (Davis and Goadrich 2006).

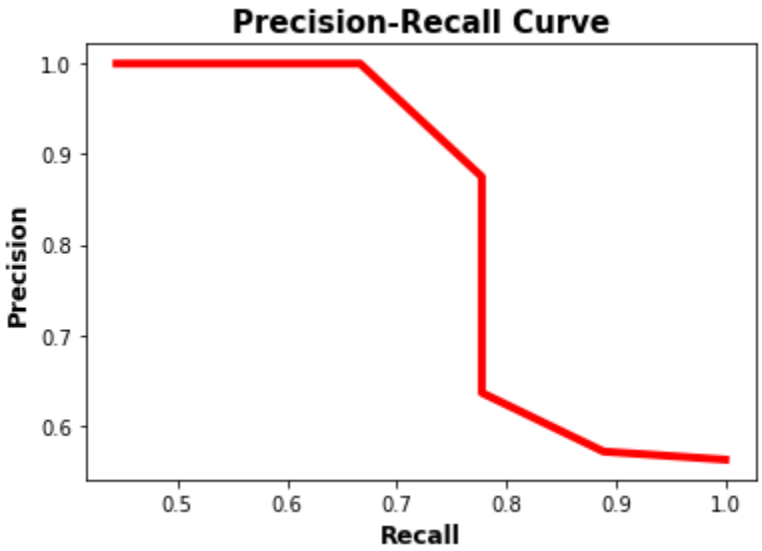


Figure 3.8. Precision-recall curve

An ideal model is to produce object detection results where both measures are relatively high. For the above graph at approximately 85% precision, the recall measure remains relatively high at approximately 78%. Further increasing recall will result in the insufficiency of the precision measure as seen by the sharp dip of precision beyond 78% recall.

These two measures feed into the calculation of the average precision of the model:

$$AP = \sum_{k=0}^{n-1} [Recalls(k) - Recalls(k+1)] * Precisions(k) \quad (3.5)$$

where,

$Recalls(n) = 0$

$Precisions(n) = 1$

$n = number\ of\ thresholds, and\ k = rank\ of\ sorted\ predictions$

The above equation illustrates that average precision is high when both precision and recall are high, which is what an object detection model attempts to accomplish. In many object detection applications, models are often required to detect and localise more than one class of object. The

average precision for each class of object of interest, summed together and then divided by the number of classes, gives the mAP (Henderson and Ferrari 2017). Since our task only requires us to detect one class of object (cattle eye), the mAP then just becomes the average AP across all bounding boxes found in all images contained in the dataset. In our study, the mAP of the validation set is used to tune the hyperparameters of the object detection model. Only models with a mAP of over 0.99 proceed to the testing phase of the research. Finally, the mAP results are outputted for each step of experimentation with the implementation of YOLOv5 in their testing phase are tabulated and presented in the results section and compared and assessed in the discussion section. Using the mean Average Precision (mAP) is important in assessing object detection success because it provides a comprehensive measure of the model's accuracy by considering both precision and recall across different threshold levels, allowing for a more nuanced evaluation of the model's performance in detecting objects correctly and minimising false positives (Everingham, Van Gool et al. 2009).

3.2.3.2 Sensitivity and specificity

A statistical metric known as specificity will also be employed to evaluate the object detection's training and testing performance. It is usually coupled with another metric known as sensitivity, as that represents the ability of the object detection model to correctly draw a bounding box around the region of interest, which is the animal's eye within each image out of all the images in the dataset. Since each image was verified initially during the annotation step for all the good quality images to contain an eye of a cattle, this sensitivity measure also represents the ability to capture the relevant region of the eye in each image, which is also known as "recall" as explained above. This measure is incorporated into the mAP evaluation metric and thus will not be produced in the output.

Specificity for the general purpose of binary classification is calculated as follows:

$$Specificity = \left[\frac{True\ Negatives}{True\ Negatives + False\ Positives}\right] \quad (3.6)$$

This measure is for the purpose of correctly identifying the instances that do not belong to the category of interest. Specificity is slightly different to the above measures as it is not commonly used as a standalone metric for object detection, as it describes the ability of the model to accurately identify the regions that do not contain the specified object, which is anything other than the eye in our example. This is a bit counterintuitive as the purpose of a good object detection model is to successfully identify the object/s of interest, in this case the cattle eye of each image. Since each image was already verified to contain an eye, to output a specificity measure is to ask the model to identify all the regions outside of the eye, which is basically just the reverse of drawing a bounding box around the eye itself. Thus, our approach is to use the optimal object detection model developed on 100 testing images that do not contain the eye at all to validate the algorithm's ability to detect negatives as true negatives on those testing images or simply put, is the YOLOv5 algorithm correctly identifying all these images as the class "background".

In YOLOv5, the process of discerning the background is intricately managed during the training phase, where the model learns to anticipate bounding boxes and class labels based on ground truth annotated training data. Throughout the training process, the model encounters a mix of positive and negative examples. The positive instances represent image regions containing objects of interest, while the negative instances signify areas devoid of any objects of interest. The model becomes adept at distinguishing between these two categories of examples.

In the YOLOv5 algorithm, each predicted bounding box is assigned a confidence score, which reflects the model's certainty that the box contains an object, regardless of its class. This score plays a crucial role during both

training and inference, guiding the model's ability to distinguish between object-containing regions and background.

To optimise this process, YOLOv5 employs a composite loss function comprising three main components: localisation loss, confidence (objectness) loss, and classification loss. The confidence loss specifically penalises the model for assigning high scores to boxes with little or no overlap with ground truth boxes—whether those boxes contain an object (positive) or represent background (negative). By aligning confidence scores with spatial accuracy, the model learns to prioritise more reliable detections.

During the inference phase, a confidence threshold is commonly employed to sift through predictions and discard those with low confidence scores. Such a thresholding mechanism aids in eliminating predictions deemed less reliable. Instances of low-confidence predictions often involve false positives, where the model erroneously identifies the background as an object, and these instances should be suppressed. This post-processing step is applied to filter out low-confidence predictions and improve the precision of the final predictions.

From the 100 testing images, we will work out the correct identification of the true negatives from these images for the numerator of the equation, that is, no bounding boxes are drawn then dividing it by the number of images that contain a bounding box as well as the number of true negatives to output the specificity measure for each YOLOv5 model trained on different sample sizes as well as with or without data-augmentation. This is a significant evaluation step of our object detection model as it serves to make sure that the optimal YOLOv5 model can identify that no object of interest (the cattle eye) is found if incorrect images are uploaded by accident, often mixed in with batches of correct images, and to remove these images before they can contribute to the dataset for the next step of the classification process.

3.3 Methods

3.3.1 Description of the Dataset

The dataset utilised in this study consists of images of cattle eyes collected as part of a research project on pinkeye (MLA funded B.AHE.0319). These images were obtained from two primary sources: NSW cattle farms in Australia and cattle farms in the United States of America. The images typically feature an open cattle eye positioned approximately in the centre of the frame, as illustrated in Figure 3.2. Notably, the USA dataset includes a ruler placed under the eye in many images (Figure 3.2c), providing additional reference for size and scale.

All collected eye images underwent a rigorous manual assessment based on an exclusion criterion. Images were discarded if they met any of the following conditions: the eye was partially or fully closed, the image quality or resolution was poor, the eye was absent, obstructed, or not facing the camera. The final dataset comprises 2,000 high-quality images (1,000 from NSW and 1,000 from the USA), which are used to train and test our object detection algorithm to produce the most optimal model to automatically draw a bounding box around the eye of interest (Figure 3.9).

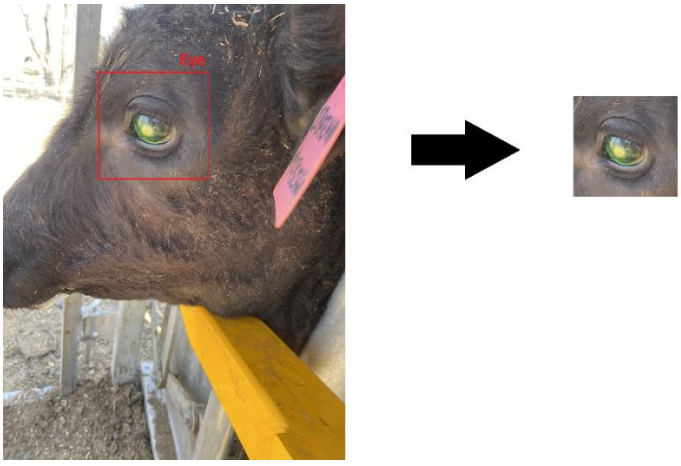


Figure 3.9. Cropping of the eye from the image from the bounding box

3.3.2 Manual Annotation

For the 2,000 high-quality cattle eye images, each eye was manually labelled using Label Studio. A rectangular bounding box was drawn around each eye, serving as the ground truth for object detection (Figure 3.10). A single category, "Eye", was used for classification, represented by the number 1. The annotations were exported in a YOLO-readable format as text files, where the first number indicates the class category, and the bounding box coordinates are represented by normalized values [x_center, y_center, width, height]. This set of label data is linked to the images via the image ID. Any area outside of the region of interest, which is the eye, is classified as the “background” and represented by the number 0, but this is hidden to the user as the background is not of interest anyway and will be eventually cropped out.

Figure 3.10. An example ground truth label of the eye

3.3.3 Data Subsets and Augmentation

The dataset with the 1,000 NSW images was divided into four subsets (Sets 1-4), each subset containing a specific number of images. Likewise, the dataset with 1,000 USA images was divided into 4 subsets (Sets 5-8) in a similar fashion (see Table 3.1).

Table 3.1. Sample sizes of each dataset subset

Sets		Sample size
Set 1	Set 5	400 images
Set 2	Set 6	600 images
Set 3	Set 7	800 images
Set 4	Set 8	1,000 images

Then a third dataset of 1,000 images was created by randomly selecting 500 images from the NSW dataset and 500 images from the USA dataset. This sample size was chosen to ensure a balanced representation across regions while keeping the total number of images same for comparing performance across three sets. Random selection was conducted using Python's random sample function with a fixed random seed (seed = 42) to ensure reproducibility and eliminate sampling bias. The resulting dataset was divided into four subsets (Sets 9–12), as outlined in Table 3.2. This mixed dataset supports the development of an object detection model capable of analysing and detecting eye regions in images from diverse sources and imaging conditions.

Table 3.2. Combined dataset showing the sample size per subset

Sets	Sample size
Set 9	400 images (200 from NSW + 200 from USA)
Set 10	600 images (300 from NSW + 300 from USA)
Set 11	800 images (400 from NSW + 400 from USA)
Set 12	1,000 images (500 from NSW + 500 from USA)

Each subset was further split into training (80%), validation (10%), and testing (10%) sets to evaluate the object detection model's performance. The subsets were used to determine the optimal sample size for achieving reasonably accurate object detection performance.

The same augmentation techniques such as rotation, scaling, flipping, and mosaic transformation were applied to each subset to produce an augmented subset variant, which is denoted by the number followed by "A", e.g. Set 1A is the augmented variant of Set1 and so forth. This setup enabled a thorough comparison of performance between non-augmented and augmented datasets, allowing us to assess the impact of data augmentation techniques on the accuracy of the object detection model when evaluated on the testing

data. The entire experimental set up of each of the subsets are shown in Table 3.3.

Table 3.3. The experiment datasets set up for object detection

Sample size for NSW Images			
Set 1 (400)	Training: 320	Validation: 40	Testing: 40
Set 1A (400 with augmentation)	Training: 320	Validation: 40	Testing: 40
Set 2 (600)	Training: 480	Validation: 60	Testing: 60
Set 2A (600 with augmentation)	Training: 480	Validation: 60	Testing: 60
Set 3 (800)	Training: 640	Validation: 80	Testing: 80
Set 3A (800 with augmentation)	Training: 640	Validation: 80	Testing: 80
Set 4 (1,000)	Training: 800	Validation: 100	Testing: 100
Set 4A (1,000 with augmentation)	Training: 800	Validation: 100	Testing: 100
Sample size for USA Images			
Set 5 (400)	Training: 320	Validation: 40	Testing: 40
Set 5A (400 with augmentation)	Training: 320	Validation: 40	Testing: 40
Set 6 (600)	Training: 480	Validation: 60	Testing: 60
Set 6A (600 with augmentation)	Training: 480	Validation: 60	Testing: 60
Set 7 (800)	Training: 640	Validation: 80	Testing: 80

Set 7A (800 with augmentation)	Training: 640	Validation: 80	Testing: 80
Set 8 (1,000)	Training: 800	Validation: 100	Testing: 100
Set 8A (1,000 with augmentation)	Training: 800	Validation: 100	Testing: 100
Sample size for combined images			
Set 9 (400)	Training: 320	Validation: 40	Testing: 40
Set 9A (400 with augmentation)	Training: 320	Validation: 40	Testing: 40
Set 10 (600)	Training: 480	Validation: 60	Testing: 60
Set 10A (600 with augmentation)	Training: 480	Validation: 60	Testing: 60
Set 11 (800)	Training: 640	Validation: 80	Testing: 80
Set 11A (800 with augmentation)	Training: 640	Validation: 80	Testing: 80
Set 12 (1,000)	Training: 800	Validation: 100	Testing: 100
Set 12A (1,000 with augmentation)	Training: 800	Validation: 100	Testing: 100

3.3.4 YOLOv5 Implementation

In this study, YOLOv5 was selected for its efficiency and high performance in real-time object detection. The training was conducted on Google Colab using the Ultralytics YOLOv5 repository, which provided automated preprocessing and model checkpointing. The YOLOv5 model was trained with the following hyperparameters: an initial learning rate of 0.01, momentum of 0.937, weight decay of 0.0005, batch size of 16, and 300 epochs. Multi-scale training was employed, adjusting the image size every 10 batches to enhance robustness. An early stopping parameter was set to halt training if mAP performance did not improve over the last 10 epochs. The composite loss function used during training included bounding box regression loss, distribution focal loss, and classification cross-entropy loss. These components ensured the model optimised for both localisation and classification tasks. The optimisation algorithm employed was Stochastic Gradient Descent (SGD) with a momentum parameter of 0.937 and a dynamically adjusted learning rate using a cosine annealing scheduler. Each of the trained models underwent

validation on its respective validation set, which produced a working model for evaluation on the testing set. This training process produced four object detection models for Sets 1-4 and four models for Sets 1A-4A, a total of eight models for the NSW dataset. Similarly, eight object detection models were produced for Sets 5-8 (including its respective augmented variants Sets 5A-8A). Finally, eight models were produced for the combined datasets of Sets 9-12 and Sets 9A-12A.

3.3.5 Model evaluation

The performance of the models on the testing sets was primarily assessed using the mean average precision (mAP) measure, comparing predicted bounding boxes to ground truth boxes. These mAP values were used to determine the most optimal model and to evaluate the minimum required sample size for adequately high performance as well as to analyse the impact of data augmentation on model performance.

Additionally, the optimal model of each of the three datasets was tested for specificity using 100 non-related images with no eyes present, sourced ethically and legally from Google. These images were acquired following strict guidelines to ensure compliance with copyright and usage rights. The model's ability to correctly identify non-eye images was used to evaluate its specificity and robustness in real-world applications.

3.4. Results

3.4.1 mAP evaluations

3.4.1.1 mAP evaluation for Dataset subsets 1-4 (NSW images)

The weights obtained after training the object detection model were utilised for testing on the designated testing dataset. The images (shown in Figure 3.11) showcase the comparison between actual labels and predicted labels of the testing set:

Figure 3.11. The ground truth bounding box (red) and the predicted bounding box (green) overlayed on one image to illustrate the object detection results on the testing image.

All experimental sets in the testing phase were able to detect the eye and an appropriate bounding box was formed around the eye of each image, which means the accuracy for labelling and producing the correct bounding box of unseen images is 100%, making this the testing accuracy. The mAP was produced for training, validation and testing for each of the experimental sets and presented in Table 3.4.

Table 3.4. Evaluation metrics for training, validation and testing of Sets 1-4.

Sets	mAP_{50}
1. train	0.963
1. validation	0.984
1. test	0.293
1A. train (augment)	0.963
1A. validation (augment)	0.986
1A. test (augment)	0.420
2. train	0.995
2. validation	0.991
2. test	0.636
2A. train (augment)	0.995
2A. validation (augment)	0.991
2A. test (augment)	0.636
3. train	0.990
3. validation	0.990
3. test	0.636

3A. train (augment)	0.990
3A. validation (augment)	0.990
3A. test (augment)	0.644
4. train	0.993
4. validation	0.993
4. test	0.622
4A. train (augment)	0.993
4A. validation (augment)	0.993
4A. test (augment)	0.620

Where mAP_{50} is is the mean Average Precision calculated using a fixed IoU threshold of 0.50

The mAP calculation for the testing phase was performed as there were ground truth labels manually provided for the testing images to be able to produce this metric for comparison purposes. It can be seen that the mAP for testing sets is generally lower than training and validation but the accuracy is still at 100% for all of the experimental sets, which demonstrates the viability of the YOLOv5 model to automatically crop the eye region of the image to satisfy the objective of the study.

Data augmentation step has shown to improve the mAP of the performance of the model on the testing set. The mAP for set 1 was the lowest at 0.293 and 0.420 with augmentation, whereas the highest mAP for the testing set was achieved by set 3 which used a sample size of 800 (640 training with augmentation).

3.4.1.2 mAP evaluation for Dataset subsets 5-8 (USA images)

The mean Average Precision (mAP) values for the performance of models trained on sets 5-8 are presented in this section (Table 3.5). The results demonstrate a consistent trend of improvement in mAP, with a significant increase of approximately 0.35 observed between set 5 and set 6, corresponding to an increase of just 200 images in the sample size. Data augmentation has shown a notable impact, particularly on the mAP of the testing set for set 5, where the sample size was relatively small at 200 images. However, the enhancement in mAP due to data augmentation becomes less evident for sample sizes exceeding 400 images, suggesting that the benefits of augmentation diminish as the dataset size increases.

Table 3.5. Evaluation metrics for training, validation and testing of Sets 5-8.

Sets	mAP_{50}
5. train	0.968
5. validation	0.973
5. test	0.278
5a. train	0.961
5a. validation	0.988
5a. test	0.414
6. train	0.995
6. validation	0.991
6. test	0.635
6A. train	0.995
6A. validation	0.991
6A. test	0.640
7. train	0.991
7. validation	0.991
7. test	0.633
7A. train	0.991
7A. validation	0.991
7A. test	0.641
8. train	0.993
8. validation	0.993
8. test	0.609
8A. train	0.993
8A. validation	0.993
8A. test	0.613

Where mAP_{50} is is the mean Average Precision calculated using a fixed IoU threshold of 0.50

3.4.1.3 mAP evaluation for Dataset subsets 9-12 (Combined NSW and USA images)

The objective of this analysis is to demonstrate the ability of the object detection algorithm to handle images from diverse sources. Images from the USA often included a ruler, which could partially obstruct the eye and potentially confuse the object detection model trained on images without a ruler, which can be alleviated with a model trained on both types of images equally. The model exhibited a similar trend to that observed earlier, with a significant improvement in mAP at a sample size of 400 images. Data augmentation notably enhanced the mAP for the sample size of 200 images, but its effect was less obvious at larger sample sizes. Overall, there is a general improvement in mAP as the sample size increases, although the improvement becomes marginal at larger sample sizes (Table 3.6).

Table 3.6. Evaluation metrics for Sets 9-12.

Sets	mAP_{50}

9. train	0.965
9. validation	0.985
9. test	0.282
9A. train	0.965
9A. validation	0.985
9A. test	0.405
10. train	0.994
10. validation	0.990
10. test	0.643
10A. train	0.994
10A. validation	0.990
10A. test	0.646
11. train	0.990
11. validation	0.990
11. test	0.655
11A. train	0.990
11A. validation	0.990
11A. test	0.657
12. train	0.993
12. validation	0.993
12. test	0.620
12A. train	0.993
12A. validation	0.993
12A. test	0.621

Where mAP_{50} is is the mean Average Precision calculated using a fixed IoU threshold of 0.50

3.4.2 Specificity evaluation

The YOLO model trained on experiment set 11 with augmentation was deemed to be the most optimal as it required less data than set 12 and yet it has the highest mAP out of all the sets. This model was used for inferential testing on the 100 testing images to test the specificity of the algorithm in its ability to not draw a bounding box in 100 testing images that do not contain the eye.

The results show that there were no eyes labelled and no bounding boxes drawn in any of the images (if an eye is identified it will show “Eye” next to the colon after the .jpg file extension). A sample image is shown in Figure 3.12 without a bounding box drawn after running the object detection algorithm.

Figure 3.12. An example of a testing image with no bounding box drawn after running the object detection algorithm.

3.5. Discussion

The goal of this chapter, which was to identify a single class of image, the cattle eye, in each image was successfully achieved with YOLOv5. The algorithm was able to incorporate a cropping component to automatically crop out the background leaving the region of interest as a preprocessing step for further deep learning model classification of the different attributes of pinkeye. This strategy aligns with the findings in human ophthalmology research, where object detection methods like YOLO have been effectively used to localise eye regions for further disease classification for diseases such as glaucoma and diabetic retinopathy (Gogineni, Pimpalshende et al. 2021). The examination of our object detection problem also revealed several key insights that significantly impact model performance. By leveraging these architectural choices, hyperparameters, training procedures, loss functions, and evaluation metrics, YOLOv5 demonstrated its capability to efficiently and effectively learn from the training data, achieving high performance in object detection tasks. This supports prior research demonstrating YOLO's effectiveness in rapid and accurate localisation across various biomedical applications (Redmon and Farhadi 2018). Our study leverage the full potential of YOLOv5, showcasing its applicability in various real-world scenarios and its ability to generalize well across different datasets.

The investigation into the sample size especially during the training phase has demonstrated a notable positive correlation between an enlarged dataset and improved model performance during testing, which is supported by research that has shown the improvement to the generalisability of object detection models by scaling the size of datasets, particularly in agricultural and veterinary contexts (Song, Kim et al. 2025). The rise in mAP highlights the importance of data volume in training robust object detection models. A larger dataset with augmentation enables the model to grasp diverse representations of the target object, enhancing its ability to generalise effectively to be able to handle unseen images that may contain a variety of different objects along with the object of interest. However, the results showed that the mAP improves to a certain point only even with an increasing sample size to a maximum of 0.657 during the testing phase. This suggests that there may be some homogeneity in the data that was provided for training, validation and testing as these pictures all contained similar attributes such as cattle farm items like fences, hay, tags, and cattle face features like the eye, nose etc. so it does not require the algorithm to learn many pictures to ascertain most of the possible features contained within these images. The results did show that between the increase of a training sample size of 320 images to 480 images, the mAP of the algorithms improved from 0.293 (without augmentation) to 0.636 (without augmentation) for the testing performance, which is a significant improvement. This indicates that even for a 1 class object detection problem, the sample size needs to be at least close to 480 images to achieve a relatively robust object detection model for our dataset of cattle eye images based on our results.

As an additional point, it may seem that the maximum mAP of 0.644 for testing appears to be relatively low compared to the maximum mAP of 0.995 during training, it is to note that the ground truth bounding boxes were manually drawn and annotated by the researcher, which means that there is inconsistency with how the bounding boxes were specifically drawn between

all the images as long as the cattle eye was included as the central focus of the bounding box. This issue of annotation subjectivity is a recognised challenge in medical image detection tasks and has been reported to impact model evaluation metrics (Rajpurkar, Irvin et al. 2017). This means that the model may not be able to conclusively decipher the arbitrary differences between the ground truth boundaries but could learn that the eye must be the focus of the bounding box and thus have been able to produce a bounding box that encircles the eye and captures the essence of the solution to this computer vision problem. The testing images of sets 1-4 with the predicted bounding box by the algorithm were all manually assessed by the annotator to all include the cattle eye as the central focus of the bounding box, thus achieving 100% accuracy, so a maximum mAP of 0.644 for set 3 is a satisfactory evaluation value to demonstrate the success of this YOLOv5 algorithm to achieve the goal of this chapter. The implementation of data augmentation, specifically through flipping, rotation, and mosaic transformations, further enhances the model's robustness as seen by the results. Thus, the results show that it is recommended to incorporate data augmentation in the training process of the data. By introducing artificial diversity into the training set, data augmentation mitigates overfitting and exposes the model to a broader range of spatial configurations, scales, and backgrounds (Shorten and Khoshgoftaar 2019). This contributes to a more resilient and adaptable object detection system.

Upon collecting the USA images, we conducted preliminary testing using the object detection model developed with the NSW images. The initial results were unsatisfactory, as the bounding boxes frequently cropped parts of the eye, yielding low-quality outputs. The presence of a ruler in the foreground of the USA images was identified as a potential factor contributing to this decreased performance. Consequently, we developed a new object detection model based on the USA images to evaluate its performance on this specific dataset.

As demonstrated in Table 3.6, the new model, after learning the features unique to the USA images, achieved results comparable to the initial model developed with the NSW dataset. Notably, the optimal performance, indicated by a maximum mAP of 0.641 during testing, was observed in the model trained on 640 augmented images. This model required the fewest images to achieve the highest mAP. The most significant improvement in mAP was noted between sets 5 and 6, where the number of training images increased from 320 to 480 without augmentation. This suggests that increasing sample size is more effective in enhancing model performance than augmentation. While data augmentation techniques provided slight mAP improvements, the benefits were minimal (set5-set5a: 0.278-0.414; set6-set6a: 0.635-0.640; set7-set7a: 0.633-0.641; set8-set8a: 0.609-0.613). The greatest impact of data augmentation was observed with the smallest training set of 320 images. As the training size increased, the augmentation effect diminished. This finding is crucial for future model development, indicating that performance can be significantly improved through data augmentation if the sample size is below the recommended 400 images. For the final set with 800 training images, a decrease in mAP compared to the 640-image training set was observed, consistent with the model trained on NSW images. This may be attributed to early stopping during model training, as no further performance gains were noted after ten additional epochs. This suggested that the model trained on 640 images was sufficient for the objectives of this chapter. Additionally, an object detection model was developed using a balanced mix of NSW and USA images to effectively crop all available images in our dataset for subsequent classification analysis. This mixed dataset followed the same protocol for increasing sample size and the training, validation, and testing ratio as the previous NSW or USA subsets, resulting in subsets 9-12. Augmentation techniques were applied to each subset to compare the effects of augmentation versus no augmentation on model performance. The results were consistent with previous findings, with the largest mAP improvement occurring between 320 and 480 training images (mAP from 0.282 to 0.643). This further emphasises the importance of

increasing sample size before applying augmentation, which elevated the mAP from 0.282 to 0.405 for the smallest training set of 320 images. When the sample size was sufficient, augmentation only slightly improved mAP for each subset. Thus, the optimal model selected for ongoing analysis to crop each eye from the image dataset is the mixed model developed from 640 images. This model effectively meets the requirements for accurate cropping to enable improved classification in the next phase of analysis. Both combining datasets and varying dataset sizes to assess the efficacy of object detection algorithms represent gaps in the literature, particularly within veterinary and medical imaging analysis.

In the context of identifying a single class in each image, this chapter shows that object detection has the potential to serve as a crucial preprocessing step. By isolating the object of interest and eliminating extraneous background elements, subsequent classification tasks are streamlined. This becomes particularly pertinent when deploying deep learning models for further classification purposes focusing on the attributes of the cattle eye, as the removal of irrelevant information such as fences, colourful objects in the background, sunlight etc. sharpens the model's focus on discriminative features associated with the target class. While the removal of background noise may offer clear advantages, such as improved classification accuracy and enhanced model interpretability, it comes with potential drawbacks. Aggressive background removal may lead to the loss of contextual information, which could be valuable for certain classification tasks. Additionally, the approach is sensitive to object localisation errors, as inaccuracies in this step may propagate to downstream tasks. This was potentially mitigated in this chapter via the specificity evaluation step to make certain that in the case of images that do not contain the cattle eye, the object detection algorithm does not mistakenly draw a bounding box and crop out an incorrect object to feed into the next step in the modelling process. The results show that the specificity of the optimal YOLOv5 algorithm was 1 when the algorithm was used to test upon 100 images that

do not contain the eye, which suggests that this is a robust algorithm that can be used reliably to satisfy the purpose of this chapter.

There are some promising future avenues for refinement and expansion for this object detection technique such as fine-tuning for specialised classes. We can consider fine-tuning the model for specific classes within the single-class detection problem to optimise discrimination within the target class. We only looked at cropping out the eye region for this chapter, however, there are subclasses to the eye such as the normal eye and the diseased eye that can be created for further fine tuning of the model. The creation of these categories and the fine tuning of the model to these categories allows the model to specialise in recognising the characteristics unique to the chosen subclass, potentially leading to improved performance for that specific class. Another aspect we can investigate is the integration with semantic segmentation, to better understand the context surrounding detected objects, potentially mitigating information loss associated with aggressive background removal, which has been shown to be particularly useful in medical AI tasks for delineating lesion boundaries or anatomical structures at pixel-level precision, enhancing downstream classification (Asgari Taghanaki, Abhishek et al. 2021).

The groundbreaking open-source image segmentation method, dubbed "Segment Anything (SA)" or "Segment Anything Model (SAM)," offers another exciting solution to our object detection problem (Kirillov, Mintun et al. 2023). Image segmentation, a sophisticated AI process in computer vision, meticulously partitions an image into regions filled with meaningful data right down to the pixel level, offering a more detailed perspective than traditional object detection methods (Minaee, Boykov et al. 2021). SAM’s new approach bypasses the conventional necessity for model-specific training tailored to distinct object detection tasks. Instead, SAM's innovative framework integrates an image encoder, prompt encoder, and a mask decoder, with Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) serving as its deep learning backbones to achieve high accuracy image

segmentation (Kirillov, Mintun et al. 2023). This enables a versatile approach capable of precise image segmentation without the need for specific preliminary training. Consequently, SAM reduces the initial labour-intensive process of manually annotating images to create ground truths as was employed in this chapter. It also bypasses the training, validation, and testing phases across various datasets required by the traditional approach, swiftly establishing accurate boundaries around cattle eyes at the pixel level. SAM has recently been trialled on animal datasets for poultry monitoring, showing promise in scenarios where traditional object detection pipelines require large, labelled datasets for training (Yang, Dai et al. 2024). Therefore, this new generalised strategy warrants further exploration for its potential to the efficiency in our research.

3.6. Conclusion

This chapter utilised the YOLOv5 object detection model to detect, label and crop out the eye from cattle eye images. The chosen optimal model was a model developed on 640 training images from both NSW and USA datasets, produced a high mAP of 0.99 for both training and validation, and 0.657 for the testing phase. The algorithm also achieved a 100% accuracy rating in correctly identifying and drawing a bounding box around the eye in each of the images. Furthermore, the specificity of the chosen optimal model produced a specificity of 1, which ensures that it does not incorrectly identify the eye in images where no eye is present, which is important in the cases where incorrect images were accidentally uploaded. Overall, these results indicate that the synthesis of an enlarged training dataset and strategic data augmentation the algorithm has successfully achieved the objective of this chapter. This algorithm can be incorporated into the preliminary step to automatically crop out the eye in cattle eye images to be used as inputs for the classification deep learning model to classify and predict the attributes of infectious bovine kerato-conjunctivitis. The trade-offs between background removal and potential information loss underscore the need for a nuanced approach, and our proposed future directions aim to address these

challenges while further refining the model's performance. These results lay a foundation for automated preprocessing in the field, offering potential applications in broader agricultural and veterinary contexts.

Chapter 4. Developing the Deep learning models for the classification of pinkeye attributes in cattle

4.1. Introduction

The detection and classification of medical conditions through image analysis has seen rapid advancement in recent years (Castiglioni, Rundo et al. 2021), particularly by applying deep learning techniques. In this chapter, we describe the application of deep learning algorithms to classify various stages and severities of Infectious Bovine Keratoconjunctivitis (IBK), more commonly known as pinkeye, in cattle eye images. One of the focuses of this chapter is on the development of a Convolutional Neural Network (CNN) model tailored specifically to the task of diagnosing pinkeye in cattle eye images. The CNN architecture, known for its effectiveness in image-based classification tasks, is designed to identify nuanced differences in images, allowing for the classification of the disease based on image features, accurately pinpointing the stages of the disease's progression. Additionally, we explore the technique of transfer learning to leverage popular pre-trained DL networks for comparison and evaluation. This approach, which has shown success in multiclass human ophthalmological disease classification (Glaret subin and Muthukannan 2022), will help us assess its relevance to our study, particularly in the livestock production context. The models with optimal performance encompassing evaluation metrics such as computation time, accuracy and F1 score aim to empower farmers by enabling them to diagnose pinkeye stages and severities in their cattle through simple eye photos, offering accurate diagnoses and actionable follow-up suggestions for the appropriate treatment and management of affected animals.

4.2. Background

Deep learning is a rapidly evolving subset of artificial intelligence that leverages extensive neural networks, often comprising numerous layers and nodes, to discern patterns within large datasets. This technology has garnered significant attention across various domains, including computer vision, language processing, and sound analysis (Shinde and Shah 2018). A particularly impactful application of DL is in the field of medical imaging, where it has outperformed traditional machine learning models in capturing intricate details and patterns within images (Rana and Bhushan 2023). These advancements have enhanced the ability to extract critical health-related information from medical images, thereby improving diagnostic accuracy, treatment planning, and disease management.

The overall simplified process can be summarised as follows: DL algorithms are trained on diverse datasets, including X-rays, MRI scans, CT scans, ultrasound images, and histological slides, which are annotated with labels corresponding to features corresponding to specific diseases or health conditions. Once trained, these DL models can provide diagnostic insights and classify diseases on new images, supporting healthcare professionals in making informed decisions. The purpose of this background section is to outline some key DL models which have found success in the vast research in human medicine studies. This provides insights into the potential direction of DL applications for the classification of pinkeye (bovine keratoconjunctivitis) in cattle, particularly as Chapter 2 of the AI literature review shows a deficiency in research on DL techniques in veterinary ophthalmology, with no specific focus on eye disease diagnosis. Since the volume of DL research in the field of human medicine is too vast and voluminous for the purpose of this background section, which is just to introduce some of the key models that have found success in human medicine for its application for our context of the study, we will focus on some of the examples of DL modelling success in disease diagnostics in ophthalmology to provide insight into attacking our problem. The development of a DL algorithm to detect the stages of pinkeye

can aid cattle farmers and veterinarians in devising and implementing timely treatment plans, thereby improving the health outcomes of affected animals. This chapter draws on the principles and successes of various DL models in ophthalmology medical imaging to inform the development of a model suited to this context.

4.3 Overview of Key Deep Learning Models

In image analysis, the convolutional neural network (CNN) serves as a foundational architecture for learning relationships between input images and their corresponding class labels. The basic structure of a CNN includes an input layer, which processes pixel data, followed by hidden layers that extract features through convolutional and pooling operations (Elngar, Arafa et al. 2021). The output layer typically applies a function like softmax to classify objects within the image (Elngar, Arafa et al. 2021). Over time, numerous advanced deep learning models have been developed, building on the core principles of CNNs. These models, such as ResNet, VGG, and EfficientNet, which will be elaborated upon, are designed to handle increasingly complex image data and improve performance across various domains, including ophthalmology and other medical imaging use cases.

A significant advancement in this area is transfer learning, a useful technique that leverages pre-trained models to address new tasks more efficiently. In the context of computer vision, transfer learning allows us to utilise models that have already been trained on large-scale datasets, such as ImageNet, which contains over 2 million images across thousands of categories. Pre-trained models like ResNet, VGG, and EfficientNet have demonstrated high performance in various image classification tasks due to their extensive training, which enables them to learn a wide array of visual features and patterns (Hussain, Bird et al. 2019). By fine-tuning these models, researchers can adapt them to specific image analysis tasks, such as disease diagnosis in veterinary or human medicine, where labelled datasets are often limited.

In the following sections, we will explore some of the most popular models in the field of ophthalmological medical imaging, including ResNet, DenseNet, VGG, InceptionNet, and EfficientNet, and discuss how their architectures can be effectively utilised for our classification task.

4.3.1 VGGNet:

VGGNet is an older architecture that uses a sequence of small convolutional filters, typically of size 3x3, to systematically extract features from input images (Figure 4.1) (Simonyan and Zisserman 2014). This use of smaller filters allows the network to capture intricate patterns and details, effectively learning a hierarchy of features, from low-level edges and textures to more complex structures, making it adept at capturing fine details essential for tasks such as object recognition and classification. Following the convolutional layers, VGGNet incorporates max pooling layers, which serve to progressively reduce the spatial dimensions of the feature maps (Simonyan and Zisserman 2014). This pooling operation not only decreases the computational load by reducing the number of parameters and activations but also helps in making the network invariant to small translations and distortions, which is particularly useful in handling variations within medical images. The combination of deep stacks of convolutional layers and pooling layers enables VGGNet to learn rich, multi-scale representations of input data.

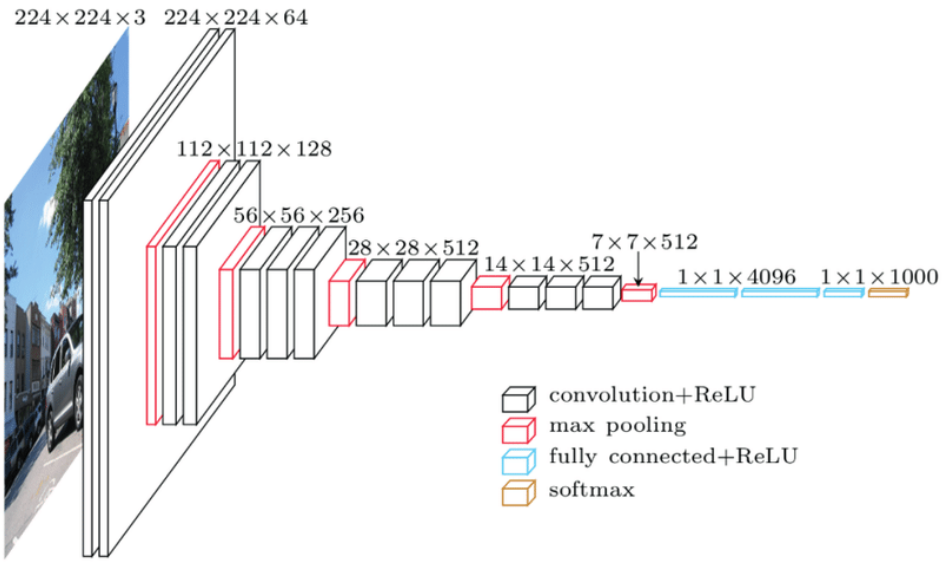


Figure 4.1. VGGNet 19 architecture. Adapted from Bezdan and Bacanin (2019).

Despite its relatively large number of parameters compared to more recent architectures like ResNet and DenseNet, VGGNet remains popular in the field of medical imaging (Kim, Cosa-Linan et al. 2021) due to its simplicity and effectiveness. Its deep architecture, while parameter-heavy, has been shown to perform well in extracting detailed features from medical images, such as MRI scans, X-rays, and histopathology slides (Kora, Ooi et al. 2022).

In the area of classification of eye diseases, VGG-19 achieved high precision (98.19%) and recall (94.7%) in identifying cataracts and diabetic retinopathy (Salem, Negm et al. 2022). In another study, VGGNet was used as a feature extractor in combination with a deeper CNN network to classify eyes with leukocoria, achieving a classification accuracy of 98.5%, outperforming ANN and other machine learning classifiers (Subrahmanyeswara Rao 2020). However, it is important to note that these successes of VGGNet in distinguishing between diseases with visually distinct symptoms may not directly apply to more nuanced tasks, such as differentiating between stages of the same disease, like in our study of classifying stages of pinkeye, where the visual differences are more subtle and harder to separate.

4.3.2 ResNet

Complex information and patterns contained in images typically require a deep network with many layers to parse and analyse effectively. However, this depth introduces a phenomenon known as the degradation problem in deep learning networks. This issue arises when adding more layers to a network does not necessarily lead to better performance. Instead, it can cause the gradients used for updating the weights during backpropagation to either vanish (become too small) or explode (become too large), leading to weight updates that do not contribute meaningfully to improving the network's performance (Borawar and Kaur 2023).

ResNet, whilst inspired by VGGnet, alleviates the degradation problem by introducing skip connections, or shortcuts, that bypass one or more layers, allowing the network to learn residuals rather than directly learning the desired output (Figure 4.2) (He, Zhang et al. 2016). These skip connections enable gradients to flow more easily through the network during backpropagation, reducing the vanishing gradient problem and stabilising training. By focusing on learning the difference between the input and the output, each residual block simplifies the optimisation process, making it easier to train very deep networks effectively (He, Zhang et al. 2016). This approach allows ResNets to capture complex patterns in data, making them particularly useful for tasks like medical imaging, where deep networks are needed to parse and analyse intricate image details.

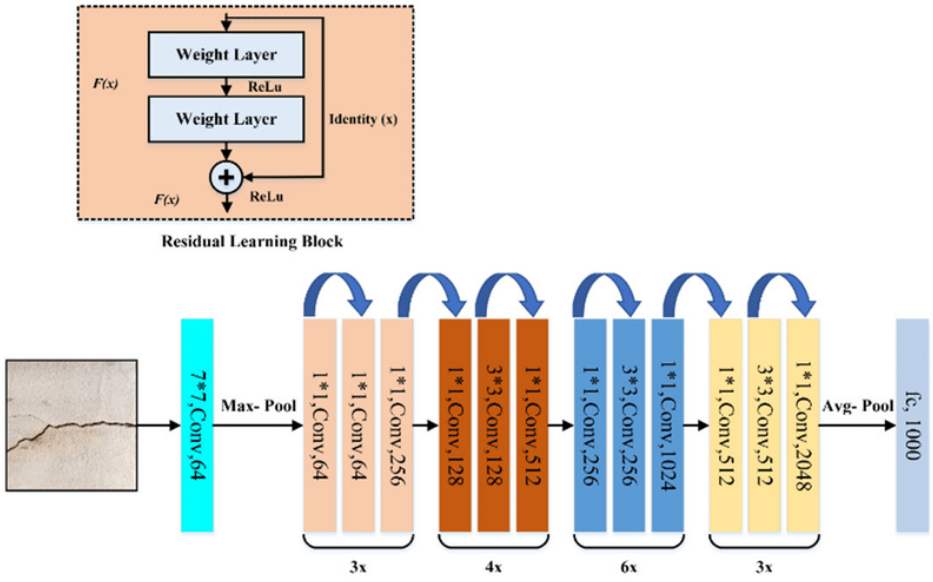


Figure 4.2. ResNet 50 architecture. Adapted from Ali, Alnajjar et al. (2021).

ResNet architectures have been widely adopted in medical imaging, particularly in ophthalmology, where they have demonstrated high accuracy in various diagnostic tasks. For instance, ResNet has achieved an accuracy of over 96% in classifying uveal melanoma (Santos-Bustos, Nguyen et al. 2022). This study utilised transfer learning by initially leveraging ResNet18's pre-trained weights from the large ImageNet database and then fine-tuning these weights on a new eye image dataset to develop a deep learning model

specifically tailored to this task. Similarly, transfer learning with ResNet50 has proven effective in distinguishing between eight categories of ophthalmological conditions—normal (N), diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (AMD), hypertension (H), myopia (M), and other diseases/abnormalities (O)—using fundus images, which are photographs of the interior surface (retina) of the eye. This approach achieved high validation accuracy, with 84.9% using Adam's optimiser and 86.71% using the SGD optimiser (Gour and Khanna 2021), demonstrating the robustness of ResNet models in analysing complex eye images.

4.3.3 DenseNet:

DenseNet enhances the ResNet architecture by introducing the concept of dense connectivity, where each layer in the network is connected to every other layer in a feed-forward manner (Huang, Liu et al. 2017). Unlike traditional architectures, where each layer receives input only from the previous layer, DenseNet ensures that each layer has direct access to the feature maps of all preceding layers. This dense connectivity pattern encourages feature reuse across the network, as each layer can selectively access relevant information from any earlier layer without needing to learn redundant features. This approach mitigates the vanishing gradient problem by providing shorter paths for the gradient to flow during backpropagation (Feng, Yao et al. 2019) , similar to the skip connections in ResNet. However, DenseNet goes a step further by allowing gradients to propagate directly to the initial layers, further stabilising training in very deep networks. In addition, by reusing features across layers, DenseNet significantly reduces the number of parameters required, leading to a more compact and efficient model that performs well, requiring lesser computational power (Zhang, Benz et al. 2021). Overall, DenseNet's dense connectivity structure improves feature propagation throughout the network, allowing for more effective learning. This is especially useful in medical imaging tasks, where subtle

differences in images need to be captured and analysed accurately (Figure 4.3).

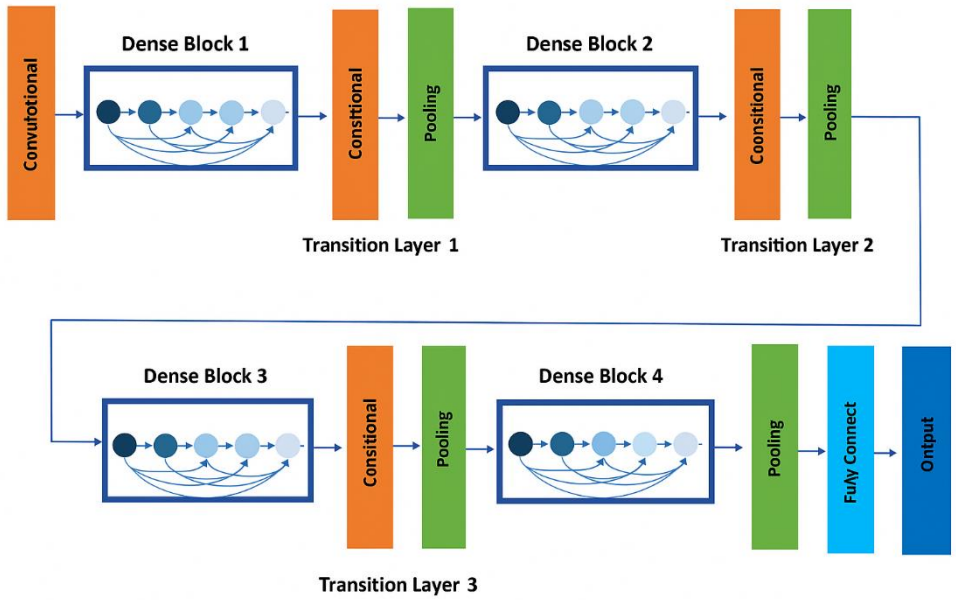


Figure 4.3. An example DenseNet architecture. Adapted from Attallah (2021).

DenseNet has achieved high accuracy 97% in differentiating between normal and glaucoma classifications in fundus images (Ovreiu, Paraschiv et al. 2021). In this relatively simple binary classification task, it has shown its capability. In a more challenging computer vision task, where the detection of 5 recorded stages of diabetic retinopathy, DenseNet was able to reproduce high performance with 98% precision for no diabetic retinopathy only, whereas, for the other four stages, it ranged between 53% to 80% precision (Singh, Dalmia et al. 2024). This again reiterates the difficulty for these models to capture more nuanced differences in images between stages of the same disease. In comparison with some other popular models, DenseNet outperformed ResNet, and VGG in the classification of eye conditions from the iChallenge-GON dataset as it showed DenseNet's advantage in amplifying information between layers, resulting in higher classification accuracy (Mu, Sun et al. 2021).

4.3.4 InceptionV3:

InceptionV3 is a deep convolutional neural network architecture known for its unique ability to capture information at multiple scales using “inception blocks/modules” (Szegedy, Vanhoucke et al. 2016). Unlike traditional convolutional networks that apply a single filter size at each layer, inception blocks process input data using multiple filter sizes (such as 1x1, 3x3, and 5x5) simultaneously (Figure 4.4). This approach allows the model to extract features at various levels of granularity, from fine details to broader contextual information, enhancing its ability to analyse complex patterns in images. Additionally, InceptionV3 incorporates techniques like dimensionality reduction within the inception modules to maintain computational efficiency. By combining convolutional and pooling operations in parallel within these blocks, InceptionV3 significantly reduces the number of parameters compared to a naive, wide architecture, making it more efficient in terms of memory and computational resources (Szegedy, Vanhoucke et al. 2016), contributing to InceptionV3’s effectiveness and versatility in image analysis.

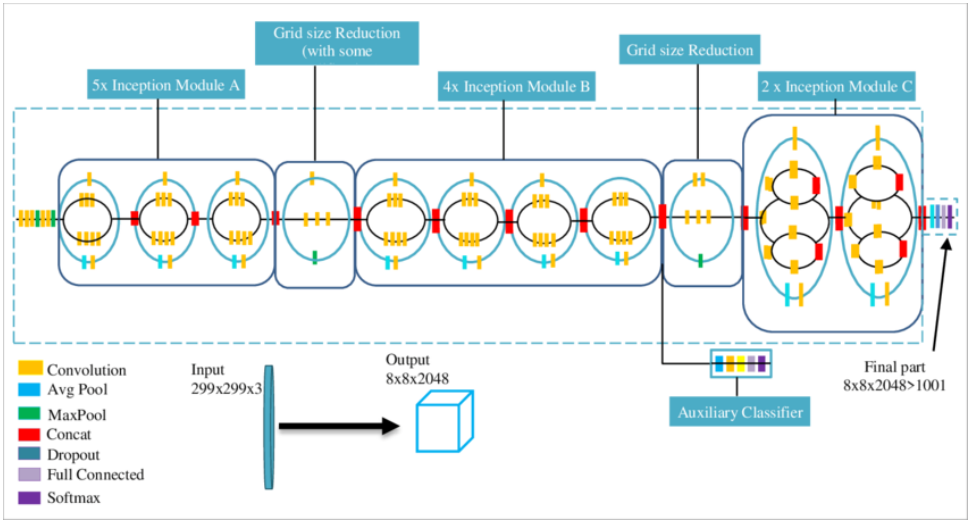


Figure 4.4. InceptionV3 architecture. Adapted from Iparraguirre-Villanueva, Guevara-Ponce et al. (2022).

Across multiple ophthalmology studies, InceptionV3 has demonstrated its effectiveness. For instance, the model achieved an accuracy of 82% in

classifying diabetic retinopathy across five severity levels, illustrating its capability to discern fine-grained differences within the same disease (Kurup, Jothi et al. 2021). Although this performance is respectable, it is not as high as when distinguishing between different diseases such as seen in another study focusing on broader disease categories, InceptionV3 reached an impressive average accuracy of 96.66% for identifying cataracts and glaucoma using fundus images, especially when combined with data augmentation techniques (Raza, Khan et al. 2021). Moreover, recent evaluations have shown that InceptionV3 continues to perform well in classifying eye conditions such as macular degeneration and tessellated fundus images, with accuracy scores exceeding 91% (Pan, Liu et al. 2023). These consistent results across different datasets and disease types suggest that InceptionV3 is a versatile and robust model, especially effective for broader classification tasks in ophthalmology image analysis rather than for fine-grained distinctions.

4.3.5 EfficientNet:

EfficientNet is a deep learning architecture that introduces a novel compound scaling method to balance the model's width (number of channels), depth (number of layers), and resolution (image input size), which is an approach that optimises the network's performance by systematically scaling these dimensions in a coordinated manner, ensuring that the model is both powerful and efficient (Figure 4.5) (Tan 2019). EfficientNetV2, the latest iteration of the model, further improves upon its predecessor by incorporating enhancements that reduce training time and increase accuracy. These improvements include a more refined and flexible compound scaling method that allows for different scaling coefficients for width, depth and resolution (Tan and Le 2021). Also, it incorporated more advanced optimisations in its model architecture to allow for progressive learning with adaptive regularisation, where the model trains on smaller image sizes and resolution to begin with and then gradually increases these qualities as training continues to produce results with better generalisation (Tan and Le 2021). Consequently, EfficientNetV2 offers faster training and better

performance, making it especially suitable for transfer learning applications where rapid adaptation to new datasets is required. This makes EfficientNetV2 an excellent choice for medical imaging tasks, where quick and accurate model deployment is often crucial.

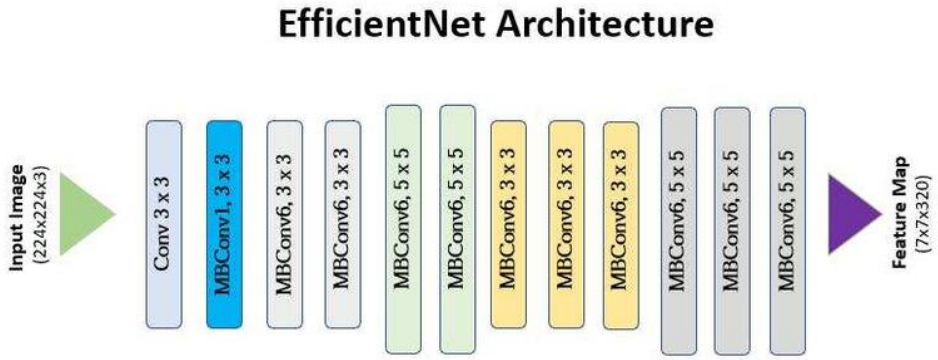


Figure 4.5. An example EfficientNet architecture. Adapted from Hisaria, Sharma et al. (2024).

EfficientNet has consistently demonstrated high performance across various ophthalmology studies, showcasing its versatility and accuracy in medical imaging. It achieved an accuracy above 95% for classifying five categories of common retinal diseases, indicating its effectiveness in handling standard diagnostic tasks (Zhu, Lu et al. 2022). Moreover, an ensemble model using EfficientNet combined with a multi-label classifier successfully classified six categories of eye disorders, further highlighting the model's adaptability to different ophthalmic conditions (Wang, Yang et al. 2020). Additionally, EfficientNet outperformed InceptionResNet in the classification of diabetic retinopathy, suggesting its superior capability for more nuanced classification tasks where distinguishing subtle differences is essential (Ganesh, Dulam et al. 2022). These findings collectively underscore EfficientNet's robustness and suitability for both broad and detailed classifications in eye disease diagnostics.

4.3.6 Segmentation approach

The above are the more popular and more researched DL models found in ophthalmology classifications. This section will cover some additional examples of certain architectures with slightly different approaches to achieving similar goals. A DL architecture, known as U-Net, is a specialised architecture for image segmentation to first identify regions of interest, such as tumours or anatomical structures before sending these features for additional analysis (Du, Cao et al. 2020). A study showcased U-Net's effectiveness when used in combination with other DL models, such as ResNet or DenseNet, for tasks requiring both segmentation and classification, where it was first applied in segmenting retinal fundus images for the diagnosis of glaucoma, followed by classification using models like VGG and ResNet, achieving high performance with accuracy above 96% (Sudhan, Sinthuja et al. 2022). This approach mirrors the method we performed in Chapter 3, where the area of interest (eye region) was first isolated by an object detection algorithm (YOLOv5) and then cropped to pass on the eye without outside noise information for further classification analysis.

4.3.7 Relevance to our study's objective

The background exploration of DL models, including DenseNet and ResNet's classification capabilities and U-Net's specialised segmentation strengths, showcases their potential and versatility for this chapter's classification goals. Each model offers distinct advantages for tackling specific challenges in medical computer vision tasks. Building on this foundation and inspired by the studies discussed in the background section, in this chapter, we applied several deep learning algorithms such as VGG, ResNet, DenseNet, EfficientNet, and ensemble models, using the transfer learning methods previously described to classify various stages and severities of pinkeye. In addition, we have developed a customised CNN model tailored specifically to diagnosing pinkeye in cattle. These models were evaluated based on accuracy, sensitivity, specificity, F1 score, and Area Under the Curve (AUC). The model that demonstrated optimal performance was be selected to aid farmers in

diagnosing the stages and severities of pinkeye in their cattle using simple eye photos. This approach aimed to provide precise diagnoses and actionable follow-up recommendations for the appropriate treatment and management of affected animals.

4.4. Methods

4.4.1 Dataset/data preparation

A total of 3,800 cattle eye photos were obtained by two veterinary specialists with smartphones from cattle farms from NSW and Queensland in Australia forming the initial database. The goal of collating the photographs was to document a spectrum of eyes, from normal to those in different pinkeye stages and severity levels, capturing a diverse range of symptoms and eyes with corneal scars. These pictures primarily focused on the open eyes of the cattle as shown in Figure 4.6. To enhance visibility, most photos were stained with a yellow/green eosin stain by the veterinarians. This staining aimed to reveal any subtle opaque areas that might indicate a pinkeye infection, potentially overlooked during the photo-taking process. In some instances, a blue light was used to accentuate hidden opaque regions that absorbed the dye.

The first research chapter utilised YOLOv5 to identify the eye region automatically and cropped out 3,800 cattle eye images which will form the dataset for this research chapter.

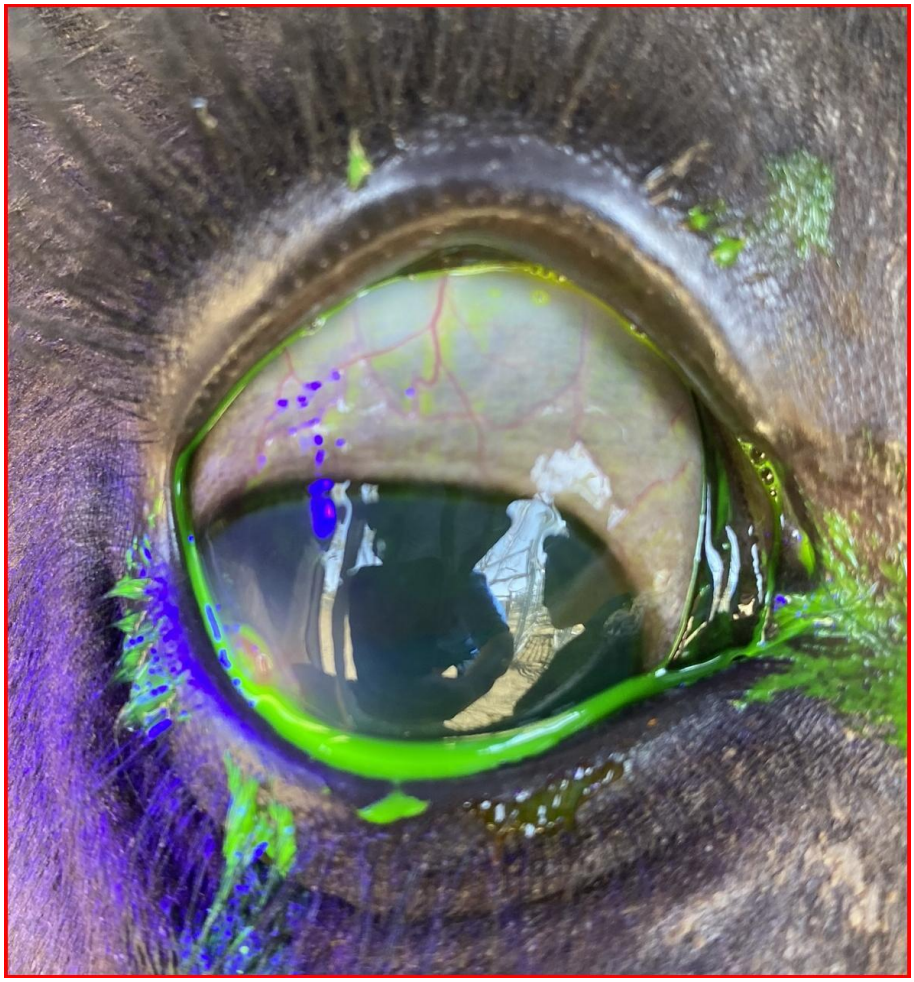


Figure 4.6. An example of a cropped eye to form the new dataset

4.4.1.1 Development of scorecard for Annotation of condition of cattle eye.

To effectively train AI systems to recognise and classify pinkeye in cattle, it is essential to develop a standardised scorecard with detailed descriptors for annotators. This scorecard is crucial to allow for consistent annotations that serve as the "ground truth" for AI training. However, diagnosing and differentiating the stages of pinkeye in cattle is inherently challenging due to the absence of clear boundaries between symptoms and the wide variability in their presentation (Kneipp 2021). Symptoms such as corneal ulceration and eye discharge vary not only in severity but also across stages of the disease, making it difficult to establish a consistent classification system. Existing literature often provides simplistic scoring frameworks (Ward and Nielson 1979), which focus on ulceration severity and eye discharge levels. However, these systems tend not to cover the broad spectrum of symptoms exhibited

by pinkeye infections and do not adequately account for variability influenced by factors such as cattle breed or eyelid pigmentation, both of which significantly affect disease severity as informed by veterinary specialists. Furthermore, the multifactorial nature of pinkeye, primarily caused by *Moraxella bovis* but influenced by other agents, complicates its diagnosis and classification (Kneipp 2021). Results indicate that the severity and progression of pinkeye vary annually and between cattle breeds, further highlighting the need for a scoring system that captures the full range of symptoms as these 3,800 cattle images are captured over a range of seasons and different breeds of animals. This resulted in the selection of 3,301 good-quality images for analysis. This lack of a comprehensive scoring system introduces errors in training AI models, particularly for more obscure cases of pinkeye, further highlighting the limitations of existing literature. To address these gaps, a comprehensive scoring system was developed in collaboration with veterinary specialists from Australia and the USA. Through iterative refinements, the scorecard was designed to categorise pinkeye cases. The score card includes 18 attributes carefully defined and reviewed in collaboration with veterinary experts and informed by both literature and field experience. These attributes provide detailed descriptors that correspond to the stages and severity of pinkeye. Based on these attributes, the eyes were finally classified into three categories—active, resolving, and resolved—each with four severity levels, with a default "normal" category for healthy eyes (Table 4.1). This system was used to annotate 3,301 images.

The annotated labels were exported in CSV format and linked to each image via its unique ID, forming a robust dataset for training DL models. This annotated dataset not only addresses the gaps in existing literature but also ensures reliable and objective training data for AI systems, enabling them to identify patterns in images and classify pinkeye stages in a robust manner.

4.4.1.2 Development of the Excel Tool for Annotation:

During practice annotation sessions, the manual process was found to be laborious and error prone. To streamline this, an Excel spreadsheet with

macros was developed to directly import images into the spreadsheet (Figure 4.7). This design facilitated the simultaneous display of images and attribute labels on a single screen, streamlining the annotation process. Prefilled attribute options reduced data entry errors and ensured consistency among annotators. Labelled data could be exported to a separate Excel tab, consolidating attribute information conveniently with a single click of the "next" button, which advanced to the next image for annotation. The tool also retained user progress, allowing annotators to revisit images and continue from where they left off. This approach enhanced the efficiency, accuracy, and user-friendliness of the annotation process.

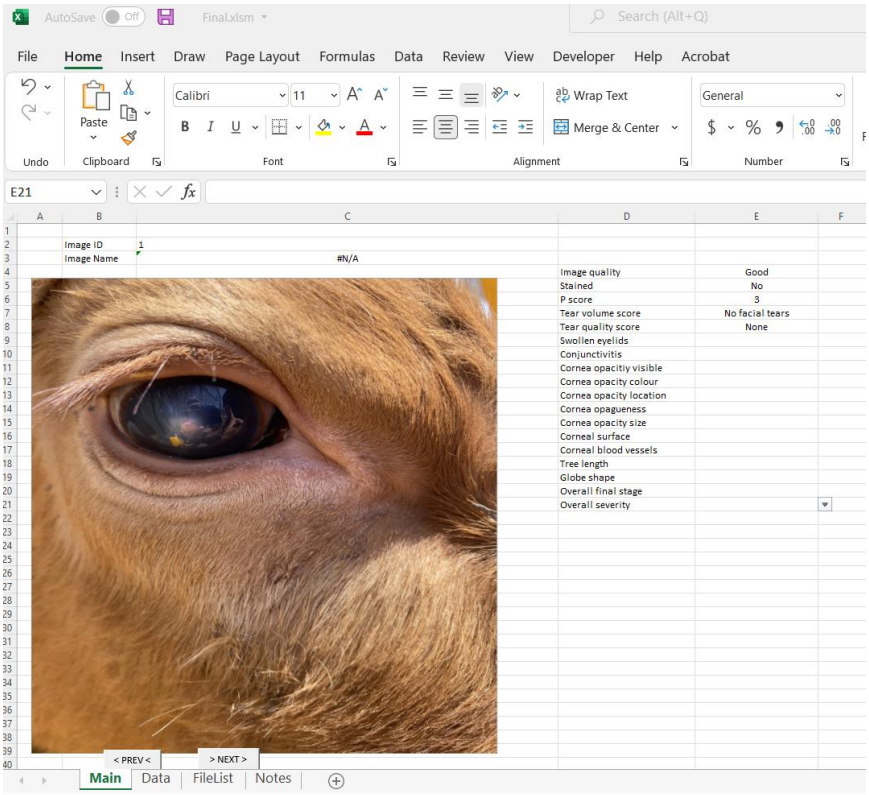


Figure 4.7. Excel annotation tool.

As outlined in Table 4.1, images are first evaluated based on quality in attribute 1; only those of good quality are selected for further annotation. This ensures accurate classification of pinkeye, as poor-quality images may not even contain a clear view of the animal's eye and must be excluded to maintain a clean dataset. After examining the images, those images without a clear opening of an eye to a high quality such as the covering of the eye by

long eyelashes, the eyelid is closed, the eye is too far or too low quality, and images that do not include the eye at all are excluded from the dataset, which resulted in the inclusion of 3,301 images for analysis.

The attributes 2-18 are all descriptors that lead to the determination by the annotator for the final classification of pinkeye stage (attribute 19) and its associated severity level (attribute 20). I conducted a minimum of five practice labelling sessions under the guidance of three veterinary experts to ensure accurate application of the scoring system to the dataset. As the primary annotator, I applied the acquired knowledge and skills to consistently label attributes 2-20 for 3,301 images using the developed scorecard to establish the ground truth labels. These findings serve as a foundation for developing DL approaches to classify attributes 19 and 20 (final pinkeye stage and severity), which will be explored in the subsequent chapter. Because there is so much extensive research on the variety of DL models used in ophthalmology, performing the classification analysis based on attributes 2-18 can narrow down the type of DL models which are more effective at analysing eye images.

Table 4.1. Pinkeye scorecard attributes developed with three veterinary experts

Attributes	Options
Image quality	Good; Poor; Bad
Stained	Yes; No
P score	0; 1; 2; 3; 4
Tear	Yes; No
Tear volume	0; 1; 2
Cornea opacity visible	Yes; No
Cornea opacity colour	Blue; Dull white; Yellow; Red; Black

Cornea opacity touches limbus	Yes; No
Cornea opaqueness	Clear; Mild; Moderate; Completely opaque
Cornea opacity size	Normal eye; <10%; 11-25%; 26-50%; 51-75%; >75%
Cornea melting	Yes; No
Corneal surface	Flat; Raised; Crater/depressed
Corneal blood vessels: hedges	Yes; No
Corneal blood vessels: tree	Yes; No
Corneal blood vessels: across lesion	Yes; No; Unable to determine
Corneal blood vessels: clearning from limbus	Yes; No
Globe shape	Normal; Popeye; Shrunken; Misshapen: include loss of eye
Presence of foreign body	Yes; No
Overall final stage	Normal (1508); A (498); R (547); S (748)
Overall severity	0; 1; 2; 3; 4

Given the distinct nature of each attribute ranging from binary and multi-class to ordinal types, each will be addressed individually in a later more specific section in the methods. A detailed descriptive analysis will be provided for each attribute, followed by a discussion of tailored DL approaches suited to their specific data types. This approach ensures that the DL analysis aligns with the unique requirements of each attribute, enabling accurate and meaningful insights.

4.4.1.3 Data Splitting

For the analyses presented herein, a dataset comprising 3,301 images, all annotated as good quality, was utilised. This dataset underwent stratified splitting, a technique that ensures each subset (training, validation, test) retains the same class distribution as the original dataset, providing a balanced representation of each class across the splits. Where possible, the dataset was divided into three distinct parts: training data (70%), validation data (10%), and test data (20%) for most of these attributes. A fixed random_state parameter was used to ensure consistent splits across different

models and analyses unless other specific tailored data splitting was employed depending on the data's requirements. This 70:10:20 ratio is a slightly unconventional split from the more typical 80:10:10 approach. Here, it was chosen to bolster the testing set, allowing for a more reliable assessment of the model's ability to generalise to unseen images, which is particularly valuable for smaller sample sizes. Although the dataset includes 3,301 images, which is considered a substantial sample size in ophthalmology research, certain attributes contain up to five classes, with some categories having small number of images within a given class. This expanded test set supported evaluation of the model's capability to capture finer, less-represented details in the images, which was crucial for accurate classification across classes.

4.4.2 Deep Learning Models

In this chapter, we address the varying complexities of classifying the 17 attributes associated with the pinkeye scorecard through a two-step DL analysis. First, we develop a custom convolutional neural network (CNN) tailored to this task, establishing a baseline performance. This model is then systematically compared with several well-established pretrained models used in ophthalmology, including ResNet50, VGG19, EfficientNetV2B2, DenseNet121, and InceptionV3. Given the large number of models being compared, we apply a proxy modelling process where all six models are used to analyse each attribute, allowing us to identify the best-performing model for each attribute. For clarity, general testing results for each attribute are summarised in a comprehensive table, comparing performance metrics across all models. In the results section, we presented detailed results, including training, validation, and testing metrics, as well as confusion matrices and other relevant analyses for each attribute's best performing model. This approach ensured a focused and thorough evaluation of each attribute while highlighting the top models suitable for the classification of disease stage and severity in Chapter 5.

4.4.2.1 Custom CNN

A custom CNN was constructed to serve as a baseline for comparison with the transfer learning models (Figure 4.9). The custom CNN architecture developed to analyse the dataset encompassed the following components: an input layer capable of processing images with dimensions of 224×224 pixels and three RGB colour channels, a sequential layer for data augmentation that executed operations such as random flips, rotations, and zooms, four convolutional layers (Conv2D) with filter sizes progressing from 32 to 64, then 64 again, and finally 128, each succeeded by a max pooling layer (MaxPooling2D). Data augmentation was applied to the training set to enhance model generalisation using techniques such as random rotations, width and height shifts, and horizontal flips, implemented dynamically during training with the ImageDataGenerator class from Keras. The feature maps were transformed into a 1D vector using a flattening layer (Flatten) and then passed through a dense layer with 128 units and ReLU activation. Additionally, a dropout layer (Dropout) with a dropout rate of 0.2 and L2 regularisation were incorporated to mitigate overfitting, a dense (fully connected) layer with 64 neurons was introduced, and a final dense output layer with five neurons (number of classes) and softmax activation was established for multi-class classification. To further enhance model performance, a dynamic model building approach was implemented using Keras Tuner. The CNNHyperModel class was defined to construct the model architecture based on hyperparameters, dynamically adding convolutional and dense layers with varying configurations. Hyperparameter tuning was performed using Keras Tuner's RandomSearch algorithm, exploring the number of convolutional layers (1-3), number of filters (32-128), kernel sizes (3, 5), number of dense layers (1-3), number of units in dense layers (64-256), and dropout rates (0.2-0.5). Each trial was evaluated based on validation accuracy, using the RandomSearch tuner to find the optimal combination of architectural and training hyperparameters. Additionally, L2 regularisation was applied to dense layers where necessary. Early stopping was employed during training, monitoring the validation loss with a patience of 5 epochs, to

halt training once the model's performance ceased to improve. The model was compiled using the Adam optimiser and categorical cross-entropy loss and underwent training for 100 epochs with a batch size of 32.

- Activation Function: Rectified Linear Unit (ReLU)
- Learning Rate: An initial learning rate of 0.01
- Batch Size: Initially set to 32
- Number of Epochs: Initially set at 100, with an early stop mechanism
- Dropout Rate: set to 0.2-0.5
- L2 regularisation set to 0.01

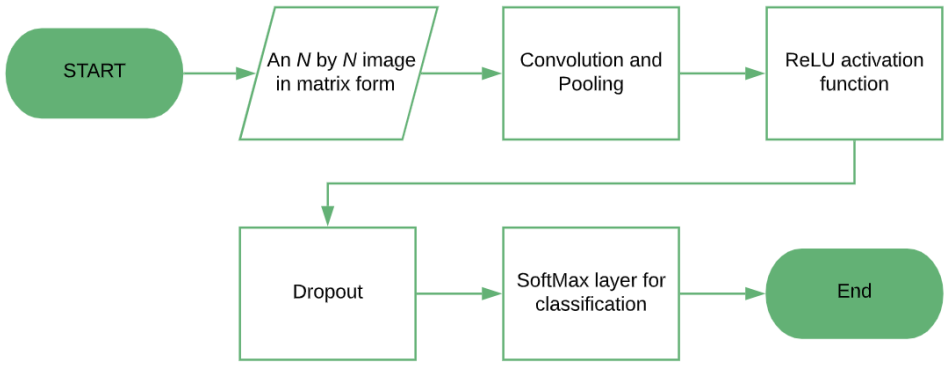


Figure 4.8. A schematic diagram of the custom CNN process

4.4.2.2 Transfer Learning Models

We explored transfer learning with five pre-trained models: InceptionV3, ResNet50V2, EfficientNetV2B2, DenseNet121, and VGG19. These models were selected based on their efficacy in various image classification tasks, particularly in human ophthalmological studies (Suganyadevi, Seethalakshmi et al. 2022). Each model was loaded without its top layers to serve as a feature extractor, retaining the original learned features from the ImageNet dataset by freezing the initial layers. For each model, we added a global average pooling layer, followed by a dense layer with 128 units and ReLU activation. To mitigate overfitting, we included a dropout layer with a rate of

0.5, L2 regularisation at a factor of 0.01, and a final output layer with SoftMax activation for multi-class classification or sigmoid for binary classification, with four output units for the multi-class tasks for an attribute with four categories for example. The Adam optimizer was used with categorical cross-entropy loss for multi-class and binary cross-entropy for binary tasks.

Some architectural adjustments were made to accommodate each model. For example, VGG19's simpler structure required adding a flattening layer before the dense, dropout, and softmax layers. The ResNet50V2, EfficientNetV2B2, InceptionV3, and DenseNet121 models, however, allowed for the consistent inclusion of global average pooling and dense layers. This approach ensured that each model leveraged pre-trained features while adapting effectively to the specific classification tasks.

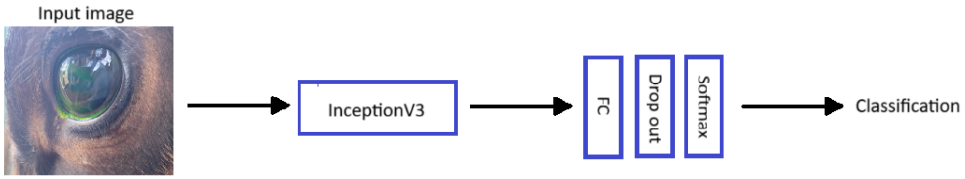


Figure 4.9. Schematic diagram of the transfer learning process for InceptionV3 as an example

4.4.2.3 Adjustments to DL models for analysing different types of variables

Certain eye attributes in this study possess an ordinal nature, where the order of their classes carries meaningful information. Unlike binary or multiclass classification, ordinal classification requires specialised adjustments to the architecture of DL models to ensure the ordinal relationships between classes are preserved and that performance metrics accurately reflect these relationships. Without these adjustments, the model risks treating ordinal classes as independent, leading to suboptimal predictions and less meaningful performance measures.

To address this, a standard approach is to reformulate the problem using the ordinal encoding strategy or cumulative link models, collectively known as ordinal regression in the context of deep learning. This approach converts the

labels (0, 1, 2) into a series of binary classification tasks, where each task predicts whether the sample belongs to a class greater than or equal to a given threshold. The model then combines these probabilities to predict the final ordinal class.

Architecturally, this involves modifying the final layer of the DL model to output probabilities for these thresholds rather than independent class probabilities. This is achieved using sigmoid activation units (one for each threshold) instead of softmax, producing cumulative probabilities that reflect the ordinal structure of the variable. The loss function is also adjusted to accommodate this type of data, with ordinal cross-entropy loss being a common choice specifically designed to handle ordinal data. This loss is computed using binary cross-entropy between the predicted cumulative probabilities and binary-encoded targets, then averaged across all thresholds and batch samples.

4.4.3. Hardware and Software Setup

Both data preparation and analysis procedures were executed using Python in GoogleColab environment and Excel. The development of custom DL models and the application of transfer learning models were undertaken using TensorFlow 2.x (Abadi, Mart et al. 2022), a versatile framework for machine learning, in Python.

4.4.4. Descriptive features, data handling and modelling approach for each attribute

4.4.4.1 Stained

Staining is a binary attribute which refers to the application of a fluorescent dye by a veterinarian to the eyes of cattle before capturing images for analysis (Figure 4.10). This procedure is used to enhance the visibility of corneal opacity by providing a colour contrast that can help identify even the slightest signs of opaqueness.

a) Unstained

b) Stained

Figure 4.10. Images of a) unstained and b) stained cattle eyes in the cropped dataset

Using a binary classification, 1,183 images (approximately 35.8%) were classified as unstained and 2,118 images (approximately 64.2%) as stained out of the 3,301 images in the dataset. Stratified data splitting was employed to ensure that the training, validation, and testing sets closely mirrored the overall distribution of stained (64%) and unstained (36%) images in the full dataset. This partition yielded balanced subsets, maintaining a consistent representation across all three sets (Table 4.2).

4.4.4.2 Tear

Tear refers to whether the eye produces visible tears in the image or not, which is a binary variable (Figure 4.11). Tearing was suggested by veterinary specialists to be a potential indication that the eye is irritated, or if it exhibits a purulent or yellowish quality, it could indicate a possible infection including pinkeye.

	
a) Tear: No	b) Tear: Yes

Figure 4.11. Images of a) Tear: No and b) Tear: Yes cattle eyes.

This dataset consists of 1,224 images (approximately 37.1%) classified as no tears and 2,077 images (approximately 62.9%) classified as tearing out of the 3,301 images in the dataset. Stratified data splitting was employed to maintain the proportional distribution of classes across the training, validation, and testing subsets (Table 4.2).

4.4.4.3 Cornea opacity visible

Cornea opacity refers to whether the cornea of the eye exhibits detectable opaqueness that clouds its clear surface (Figure 4.12). Opacity in the eye is often caused by white blood cells moving to an infected area, which is a common sign of pinkeye. However, other causes, such as corneal ulcers or environmental irritants like hay, can also lead to opaqueness and cannot be definitively ruled out.

a) Cornea opacity visible: No	b) Cornea opacity visible: Yes

Figure 4.12. Images of a) Cornea opacity visible: No and a) Cornea opacity visible: Yes cattle eyes in the cropped dataset

This binary variable consisted of 1,566 images (approximately 47.4%) classified as "cornea opacity visible: no" and 1,735 images (approximately 52.6%) classified as "cornea opacity visible: yes" This represented a fairly balanced distribution for this binary variable. Stratified data splitting was employed to maintain the proportional distribution of cornea opacity classes across the training, validation, and testing sets (Table 4.2).

4.4.4.4 Cornea opacity: touches limbus

Cornea opacity: touches limbus is an attribute that describes whether the ulcer extends to the limbus, the border between the cornea and the sclera of the eye (Figure 4.13). This was considered as a binary variable consisting of 769 images (approximately 23.3%) classified as "yes" and 2,532 images (approximately 76.7%) classified as "no" The "no" category also included cases where no opacity was observed, which contributes to its majority representation, along with opacity that did not reach the limbus.

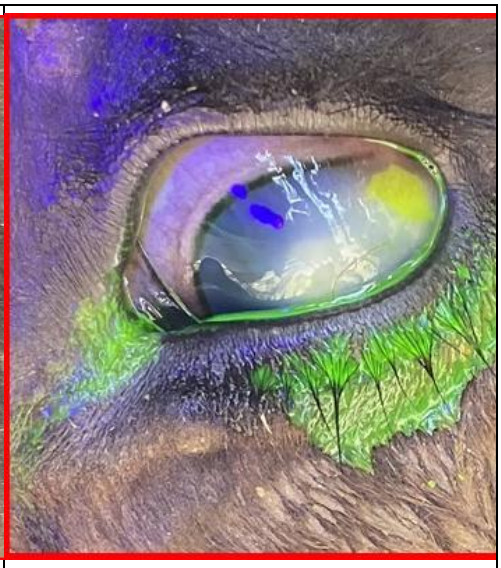
	
a) Cornea opacity: touches limbus: No	b) Cornea opacity: touches limbus: Yes

Figure 4.13. Images of a) Cornea opacity: touches limbus: No and b) Cornea opacity: touches limbus: Yes cattle eyes in the cropped dataset

Stratified data splitting was employed to ensure proportional representation of the two categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.5 Corneal blood vessels (hedges)

Cornea blood vessels (hedges) refers to the bright red vascular patterns that form a circular frame around the perimeter of the opaque area of the cornea (Figure 4.14). These blood vessels deliver increased blood flow to the infected or damaged area, promoting healing through the delivery of immune cells.

	
a) Cornea blood vessels (hedges): No	b) Cornea blood vessels (hedges): Yes

Figure 4.14. Images of a) Cornea blood vessels (hedges): No and b) Cornea blood vessels (hedges): Yes cattle eyes in the cropped dataset

This is a binary variable consisting of 2,906 images (approximately 88.0%) classified as "no" and 395 images (approximately 12.0%) classified as "yes". The dataset is heavily skewed towards the "no" category, so data augmentation may be useful in boosting the representation of the minority class for this variable. Stratified data splitting was employed to maintain the proportional distribution of cornea blood vessels (hedges) categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.6 Corneal blood vessels (trees)

Cornea blood vessels (trees) refers to the blood vessels that surround the opaque area of the cornea in a root- or tree-branch-like pattern. These vessels typically surround the vascular hedges area near the opaque region, although this is not always the case. Tree-like vessels grow from the outer edges towards the centre of the opaque area or the site of infection or damage (Figure 4.15). This vascular growth delivers increased blood flow to the infected or damaged area, promoting healing through immune cell activity. The presence of tree-like vessels can be interpreted in multiple ways. When coupled with hedges, they generally indicate an earlier stage of infection with active inflammation. However, if tree-like vessels are observed

without hedges, it may suggest that the eye is transitioning to a resolving stage where the infection is diminishing, and healing is progressing without additional treatment.

	
a) Cornea blood vessels (trees): No	b) Cornea blood vessels (trees): Yes

Figure 4.15. Images of a) Cornea blood vessels (trees): No and b) Cornea blood vessels (trees): Yes cattle eyes in the cropped dataset

This was considered as a binary variable consisting of 2,791 images (approximately 84.6%) classified as "no" and 499 images (approximately 15.1%) classified as "yes". The dataset was heavily skewed towards the "no" category, so augmentation may help bolster the representation of the minority class and assist deep learning models in learning these features. Stratified data splitting was employed to maintain the proportional distribution of cornea blood vessels (trees) categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.7 Corneal blood vessels (across lesion)

Cornea blood vessels (across lesion) are visible in the eye images as red spots scattered across the opaque ulcer on the cornea. This feature indicates a later stage of healing, where blood vessels have reached the infected area, and healing has progressed into the midway of the resolving stage (Figure 4.16). At this point, the infection is either resolving or no longer active, and scarring may result from the damaged site.

	
a) Cornea blood vessels (across lesion): No	b) Cornea blood vessels (across lesion): Yes

Figure 4.16. Images of a) Cornea blood vessels (across lesion): No and b) Cornea blood vessels (across lesion): Yes cattle eyes in the cropped dataset

This is a binary variable consisting of 2,904 images (approximately 88.0%) classified as "no" and 397 images (approximately 12.0%) classified as "yes". The dataset was heavily skewed towards the "no" category, so data augmentation was implemented to help bolster the representation of the minority class for this variable. Stratified data splitting was employed to maintain the proportional distribution of cornea blood vessels (across lesion) categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.8 Corneal blood vessels (clearing from limbus)

Cornea blood vessels (clearing from limbus) is the final attribute describing blood vessels observed on the cornea. It indicates that the blood vessels have completed delivering white blood cells to heal the eye and are now receding from the infected opaque area. This signifies that the eye is in the late stage of healing, and depending on the severity of the infection, it may result in either a minor or major scar (Figure 4.17).

	
a) Cornea blood vessels (clearing from limbus): No	b) Cornea blood vessels (clearing from limbus): Yes

Figure 4.17. Images of a) Cornea blood vessels (clearing from limbus): No and b) Cornea blood vessels (clearing from limbus): Yes cattle eyes in the cropped dataset

This is a binary variable consisting of 3,219 images (approximately 97.5%) classified as "no" and 82 images (approximately 2.5%) classified as "yes". The dataset was heavily skewed towards the "no" category, making this a challenging attribute for training purposes. However, data augmentation was implemented to reduce the impact of the skewed distribution and improve the generalisation ability of the DL models. Stratified data splitting was employed to maintain the proportional distribution of cornea blood vessels (clearing from limbus) categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.9 Corneal surface

Corneal surface is an attribute that describes the shape of the surface of the cornea. A normal eye typically has a flat surface, but an infected eye can also exhibit a flat surface. The other category is "raised", where the surface of the cornea is elevated, showing slight irregularities. A raised surface is usually associated with more serious infections and always results from ulcer formation. The final category is "crater/depressed", where the raised surface of an ulcer worsens, causing a rupture that forms a crater. Alternatively, an

ulcer may be so deep that it creates a noticeable depression at its epicentre, indicating severe infection and potentially irreversible damage (Figure 4.18). Such damage may lead to the loss of the eye if the centre of the infection ruptures.

a) Flat

b) Raised

c) Crater/Depressed

Figure 4.18. Images of a) Flat, b) Raised, and c) Crater/Depressed of Corneal surface in the cropped dataset

This attribute consists of three categories: 3,089 images (approximately 93.6%) classified as "flat", 139 images (approximately 4.2%) as "raised", and 73 images (approximately 2.2%) as "crater/depressed". Stratified data

splitting was employed to maintain the proportional distribution of these three categories across the training, validation, and testing subsets (Table 4.2).

4.4.4.10 Globe shape

The globe shape of the eye has three classes. The most common class, normal, consists of 3,253 images (approximately 98.5%) and describes a round, spherical eye shape. This classification applies regardless of the severity of the infection, as long as the eyeball maintains its original shape. The next largest class is popeye, with 28 images (approximately 0.8%), where the eye is displaced from the socket but still attached by muscles or nerves. This condition typically results from severe infection or damage. The final class, misshapen, includes cases of complete loss of globe shape, with 20 images (approximately 0.6%). This condition often represents the progression of "popeye", where the eye is no longer viable (Figure 4.19).

One possible approach to address the extreme class imbalance is to combine the two rarer classes, popeye and misshapen, into a single "abnormal globe shape" category. This would convert the variable into a binary format, where normal accounts for 98.5% of the dataset and abnormal globe shape comprises 1.5%. While this strategy may help bulk up the minority category, the imbalance remains significant and likely too extreme for meaningful DL analysis. Given this skewness, this variable was excluded from the further analysis, as training a DL model under such conditions would likely fail to produce reliable or generalisable results.

a) Normal

b) Misshapen: includes loss of globe

c) Popeye

Figure 4.19. Images of unstained and stained cattle eyes in the cropped dataset

4.4.4.11 Presence of foreign body

Presence of foreign body is a binary variable that indicates whether an object is stuck in the eye or abnormal cell growth is protruding in a way that is clearly not part of a normal eye (Figure 4.20). This category consists of only 11 images classified as "yes" which is only 0.33% of the dataset whereas 99.67% were classified as “no” making it heavily imbalanced and unsuitable for training a meaningful DL model. It is assumed that, in most cases, farmers would remove visible foreign objects like hay before taking pictures. Furthermore, the presence of a foreign body, whether caused by abnormal

cell growth or external debris, typically necessitates immediate treatment to improve the animal's welfare, regardless of whether pinkeye is also present. If pinkeye coincides with a foreign body, treatment for both conditions is required. Due to the extreme imbalance of this dataset, this variable was excluded from DL analysis.

a) Presence of foreign body: No	b) Presence of foreign body: Yes

Figure 4.20. Images of Presence of foreign body: No and Presence of foreign body: Yes cattle eyes in the cropped dataset

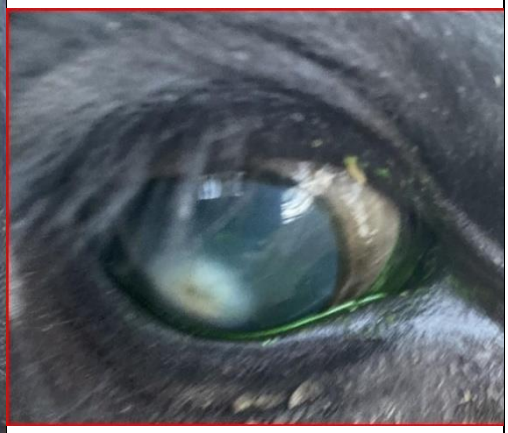
4.4.4.12 Cornea opacity: colour

Cornea opacity colour describes the colours observed on the opaque sections of the cornea due to infection or damage. These colours are distinct from the stain applied to the eye, as the stain's strong colouring (yellow, green, or blue) can sometimes obscure the underlying opacity colours. Red colouring, caused by blood vessels across the lesion, is categorised as "red" and identified as the most common colours observed on damaged corneas as dull white, black, yellow, red, blue, or none. However, this attribute originally allowed multiple options to be selected when multiple colours were observed, leading to an overly complex categorisation with many classes containing very few samples (Figure 4.21). Since there are too many subtle colours observed in the eye and when it's heavily obscured by the fluorescent stain, it was decided that this attribute is not meaningful in its analysis that could contribute to the DL model's performance in the analysis of pinkeye.

a) None

b) Dull white

c) Red



d) Other

Figure 4.21. Images of cornea opacity: colour in cattle eyes in the cropped dataset

4.4.4.13 Cornea melting

Cornea melting is an attribute that describes a condition where the surface of the cornea appears to "melt" forming a thick, lava-like texture due to severe ulceration caused by infection. This condition indicates that the infection has significantly damaged the eye, is deep and severe, and is in an active stage. It is a binary variable with 3,274 images (approximately 99.2%) classified as "no" and only 27 images (approximately 0.8%) classified as "yes". This

highlights the extreme rarity of this level of severe infection within the dataset (Figure 4.22).

Due to the highly skewed nature of this variable, it is unlikely that a DL model could generalise effectively for this condition. This rarity suggests that such severe infections are uncommon in this population of animals, and analysis to predict this rare occurrence may not be effective with limited number of cases.

a) Cornea melting: No	a) Cornea melting: Yes

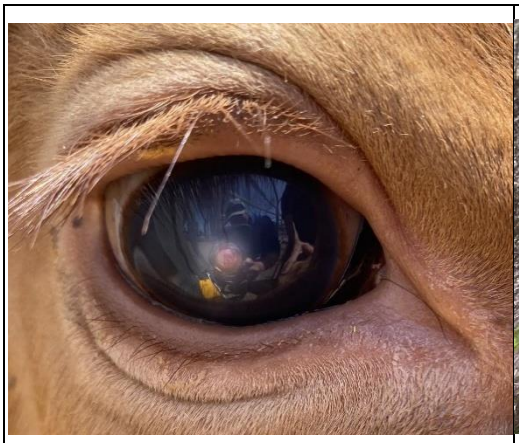
Figure 4.22. Images of a) Cornea melting: No and b) Cornea melting: Yes cattle eyes in the cropped dataset

Table 4.2. Number of images in training, validation and testing set distribution of binary and multiclass attributes included in the analysis

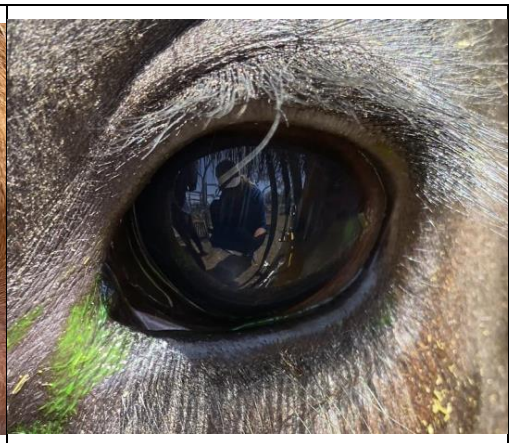
Attribute	Categories	Training	Validation	Testing	Total
Strained	Yes	1482	212	424	2118
	No	828	118	237	1183
Tear	Yes	1453	208	416	2077
	No	857	122	245	1224
Cornea Opacity Visible	Yes	1214	174	347	1735
	No	1096	157	313	1566
Cornea opacity: touches limbus	Yes	538	77	154	769
	No	1772	253	508	2533
Corneal Surface	Flat	2162	309	618	3089
	Raised	97	14	28	139
	Crater/Depressed	51	7	15	73
Cornea blood vessels (hedges)	Yes	276	40	79	395
	No	2034	291	581	2906
Cornea blood vessels (trees)	Yes	349	50	100	499
	No	1954	279	558	2791
Cornea blood vessels (across lesion)	Yes	278	40	79	397
	No	2033	290	581	2904
Cornea blood vessels (clearing from limbus)	Yes	1453	208	416	2077
	No	857	122	245	1224

4.4.4.14 Tear volume

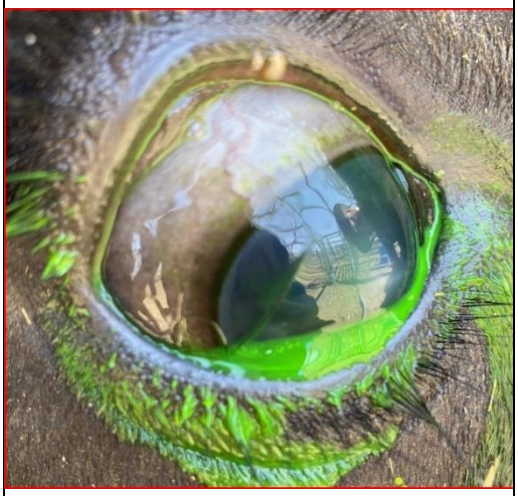
Tear volume is a numeric ordinal variable that describes the amount of tear produced and visible in the eye image. This variable consists of three ordinal classes: 0 represents no tears, 1 indicates a tear streak extending approximately less than 2 cm down the face, and 2 signifies a large quantity of tears and/or tears extending to the facial groove of the maxilla. While the cropping of images impacts the representation of class 2, since most of the face is excluded, it is still possible to identify a large tear streak or volume in these images, as shown in Figure 4.23. This differentiation is particularly clear when comparing tear volume labelled as 2 to that labelled as 1. However, it must be acknowledged that cropping may introduce possible loss of information of the images and may introduce errors or inaccuracies in the modelling process. It was suggested that eyes with larger tear volumes generally indicate a higher level of irritation, which could suggest a possible infection, though this is not always the case as healthy eyes could also produce tears as well. Such patterns are better understood when analysed alongside other variables such as other qualities of tears like texture and colour.



a) Tear volume: 0



b) Tear volume: 1



c) Tear volume: 2

Figure 4.23. Images of a) Tear volume: 0. B) Tear volume: 1 and c) Tear volume: 2 cattle eyes in the cropped dataset

This variable consists of 2,076 images (approximately 62.9%) classified as tear volume 0, 564 images (approximately 17.1%) classified as tear volume 1, and 661 images (approximately 20.0%) classified as tear volume 2. Stratified data splitting was employed to ensure that the training, validation, and testing sets retained proportional distributions of tear volume (Table 4.3).

4.4.4.15 Periocular score

Periocular score refers to the pigmentation (typically brown or black) of the skin and eyelid margins directly surrounding the eye or cornea. It is a numeric ordinal variable with scores of 0, 1, 2, 3, and 4 representing no pigmentation, less than 50% of the eyelid margins pigmented, more than 50% of the eyelid margins pigmented, 100% non-black full eyelid margin pigmentation, and 100% black full eyelid margin pigmentation, respectively. Since this is an ordinal variable, it will be analysed in a manner similar to the aforementioned tear volume variable as ordinal regression with deep learning modelling (Figure 4.24).

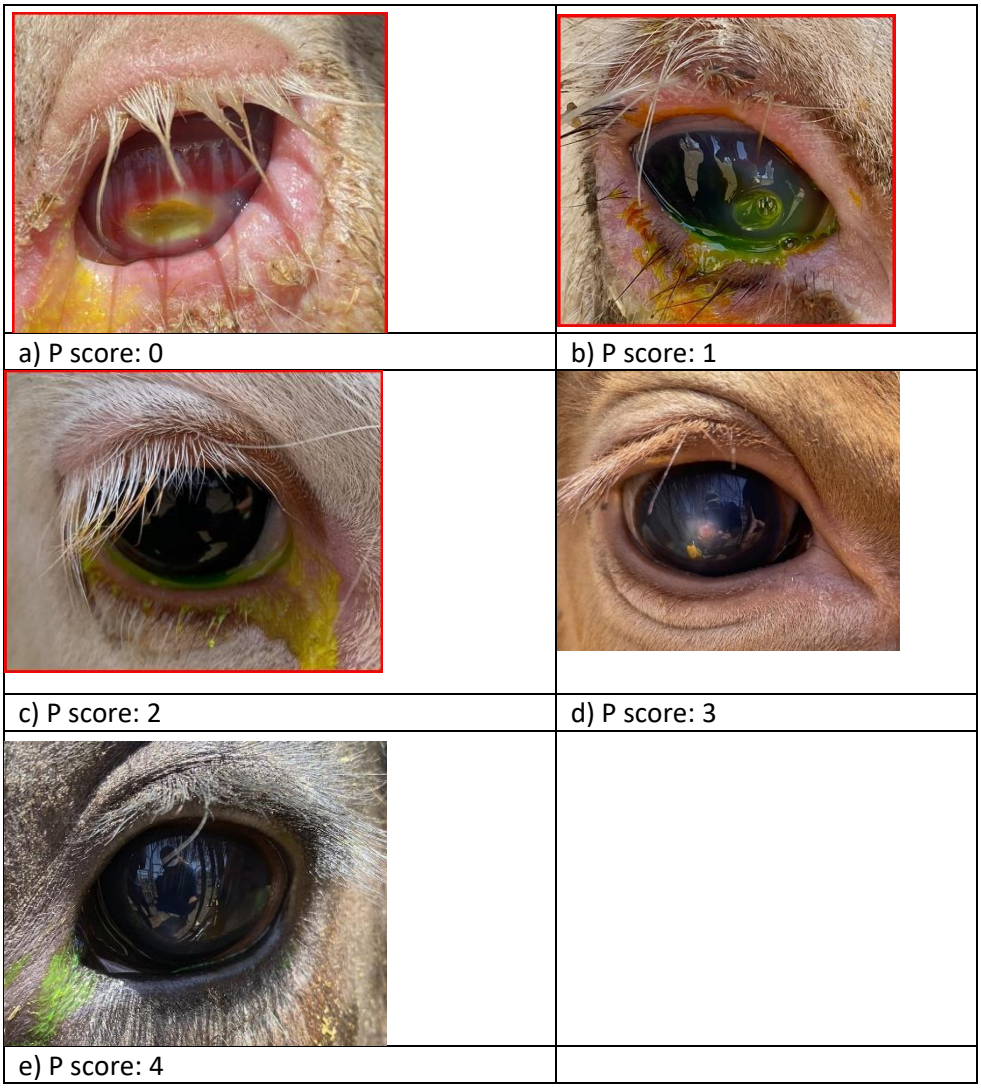


Figure 4.24. Images of a) Periocular score: 0, b) Periocular score: 1, c) Periocular score: 2, d) Periocular score: 3, and e) Periocular score: 4 in the cropped dataset

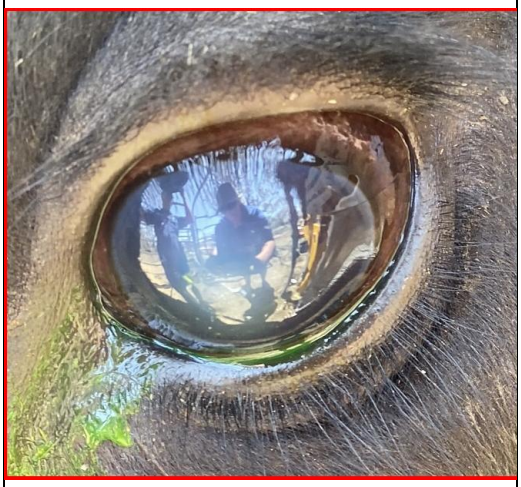
This variable consists of 28 images (approximately 0.8%) classified as score 0, 86 images (approximately 2.6%) classified as score 1, 74 images (approximately 2.2%) classified as score 2, 1026 images (approximately 31.1%) classified as score 3, and 2087 images (approximately 63.3%) classified as score 4. This demonstrates that the majority of images belong to the higher scores of 3 and 4, which make up the bulk of the dataset. Stratified data splitting was employed to maintain the proportional distribution of periocular scores across the training, validation, and testing sets (Table 4.3).

4.4.4.16 Cornea opaqueness

Cornea opaqueness is an attribute that describes the severity of opaqueness in the cornea. For some ulcers, the opaqueness manifests as slight cloudiness, whereas in more severe cases, it appears as a bright white block that obscures the cornea and penetrates deeper into the eye. This attribute consists of four categories: no opaqueness/clear (1,560 images, approximately 47.3%), mild (433 images, approximately 13.1%), moderate (392 images, approximately 11.9%), and complete opaqueness (916 images, approximately 27.7%). To enable ordinal regression analysis, these categories were converted to ordinal ranks, with "clear" as 0, "mild" as 1, "moderate" as 2, and "completely opaque" as 3 (Table 4.3). This variable represents an increasing order of severity in opaqueness, where clearer corneas indicate healthier eyes requiring no treatment. Mild opaqueness may represent a mild infection with superficial ulceration or light scarring, suggesting a relatively successful healing process. On the other hand, moderate or complete opaqueness often corresponds to more severe infections or deeper scarring, which may indicate harsher infection stages or complications. While this general trend provides a useful guideline, each case should be assessed individually to account for nuances in the infection's severity and stage (Figure 4.25).

a) Clear

b) Mild



c) Moderate

d) Completely opaque

Figure 4.25. Images of cornea opaqueness in cattle eyes in the cropped dataset

Since this is an ordinal variable, it will be analysed as ordinal regression with DL models. Stratified data splitting was employed to maintain the proportional distribution of the four categories across the training, validation, and testing subsets (Table 4.3).

4.4.4.17 Cornea opacity size

Cornea opacity size describes the percentage of the cornea covered by an ulcer. These ulcers may represent active infections, resolving infections, or scars. The size of the ulcer generally correlates with the severity of the infection, as more severe infections are typically accompanied by larger ulcer

sizes. Similarly, scarring from more severe infections often results in larger residual scars. Veterinary experts suggested using specific boundaries to categorise ulcer sizes to assist in identifying infection severity levels (Figure 4.26).

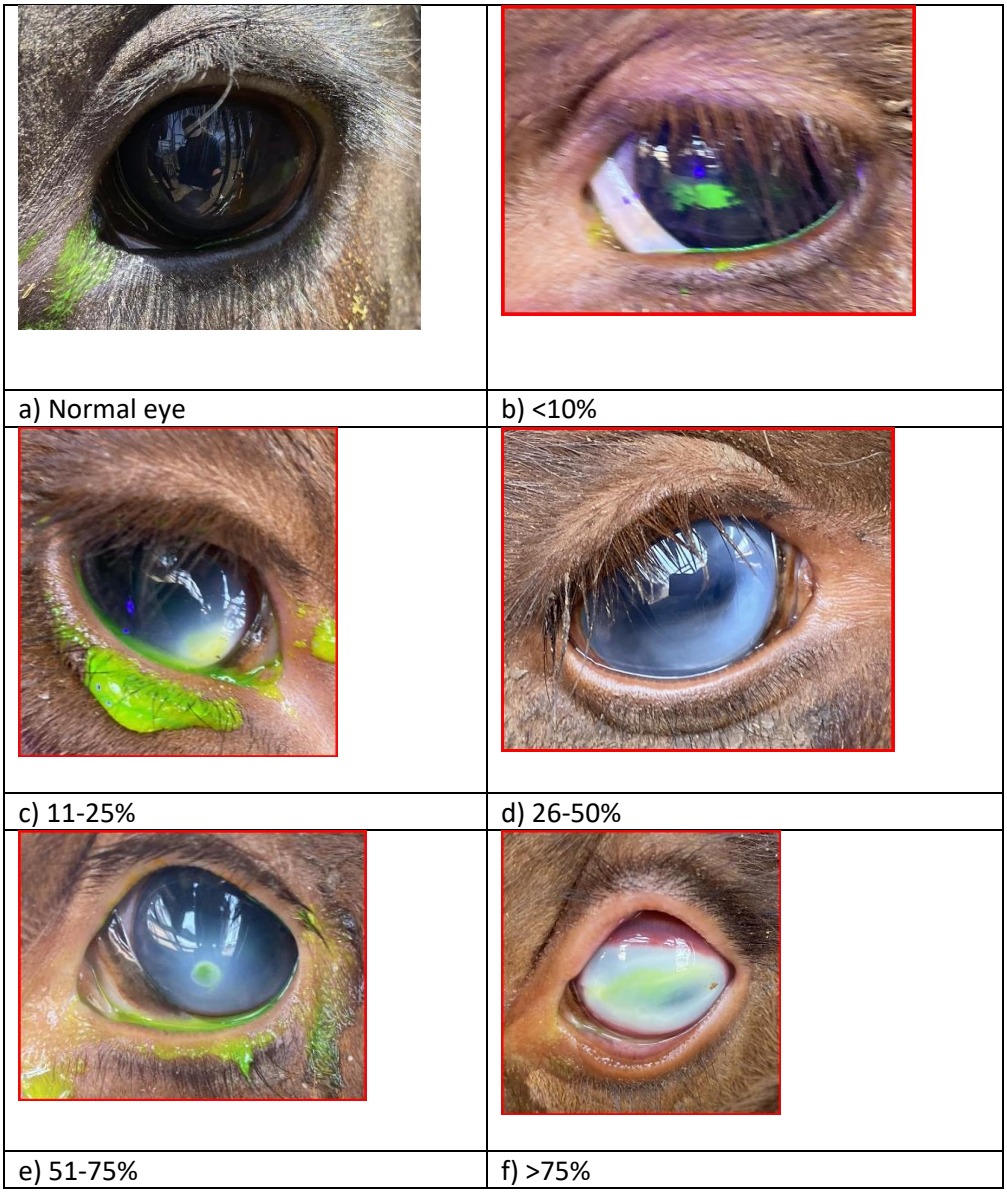


Figure 4.26. Images of cornea opacity size of cattle eyes in the cropped dataset

This variable consists of six categories: normal eyes with no ulcers (1,563 images, approximately 47.4%), ulcers covering less than 10% of the cornea (466 images, approximately 14.1%), ulcers covering 11-25% (290 images, approximately 8.8%), ulcers covering 26-50% (328 images, approximately 9.9%), ulcers covering 51-75% (196 images, approximately 5.9%), and ulcers

covering more than 75% of the cornea (458 images, approximately 13.9%). Since this variable represents an increasing order of ulcer size, it is treated as an ordinal variable. To facilitate ordinal regression analysis, the categories were converted into ranks, with "normal" as 0, "<10%" as 1, "11-25%" as 2, "26-50%" as 3, "51-75%" as 4, and ">75%" as 5. This enables deep learning models to predict and analyse these classes effectively. Stratified data splitting was employed to maintain the proportional representation of all six categories across the training, validation, and testing subsets (Table 4.3).

Table 4.3. Training, validation and testing set distribution of ordinal attributes included in the analysis

Attribute	Categories (ordinal scale)	Training	Validation	Testing	Total
Tear volume	0	1,453	208	416	2,077
	1	208	56	66	330
	2	415	113	132	660
Periocular Score	0	20	3	6	29
	1	60	9	17	86
	2	52	7	15	74
	3	718	103	205	1,026
	4	1,461	209	417	2,087
Cornea opaqueness	Clear (0)	1,092	156	312	1,560
	Mild (1)	303	43	87	433
	Moderate (2)	274	39	92	405
	Completely Opaque (3)	641	92	183	916
Cornea opacity size	Normal eye (0)	1,094	156	313	1,563
	<10% (1)	326	47	93	466
	11-25% (2)	203	29	58	290
	26-50% (3)	230	33	66	329
	51-75% (4)	137	20	39	196
	>75% (5)	321	46	92	459

4.4.5 Evaluation metrics

This study focuses on evaluating DL models used for classifying the developmental stages and severity levels of pinkeye, employing metrics suited to each classification type while considering dataset imbalances. Below, the evaluation metrics are detailed for binary, multiclass, and ordinal classification tasks.

4.4.5.1. Binary classification

For binary classification tasks, the study employed accuracy, AUC (Area Under the Curve), precision, recall, and the F1 score to evaluate model performance. These metrics have commonly been used in the literature.

Accuracy: Represents the proportion of correctly classified samples as one or the other category. While a straightforward metric, it can be misleading in imbalanced datasets, necessitating the use of additional metrics.

Precision and Recall: Precision measures the accuracy of positive predictions, while recall assesses the model's ability to identify all relevant instances. These metrics are especially important when the cost of false positives or false negatives varies.

Precision:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (4.1)$$

Recall:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4.2)$$

F1 Score: The harmonic mean of precision and recall, balancing these two metrics. It is particularly effective for datasets with class imbalances, which has been identified as a significant factor in certain attributes in this study.

F1 equation:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4.3)$$

Area under the curve (AUC): This refers to the area under the receiver operating characteristic (ROC) curve. The AUC is a value ranging from 0 to 1, reflecting the model's overall ability to distinguish between classes in the classification task. An AUC closer to 1 signifies superior performance, indicating the model's capacity to differentiate between positive and negative classes. This makes AUC a more robust metric compared to accuracy when evaluating classification outcomes.

The three key metrics accuracy, AUC and precision and recall incorporated in the F1 score to compare the performance of these DL models align with the metrics commonly utilised in DL modelling (Gour and Khanna 2021). Confusion matrices were also generated to visualise misclassifications between classes.

4.4.5.2 Multiclass classification

For multiclass classification tasks, where the model predicts one of several possible categories, evaluation metrics include accuracy (as described in the binary classification section), weighted-average F1 score, and weighted-average AUC. These weighted metrics account for class imbalance by incorporating the sample size of each class into the calculation.

Weighted-average F1 score is computed by first calculating the F1 score for each class independently, then averaging them using the number of samples in each class as weights. This approach reflects overall model performance in proportion to the prevalence of each class in the dataset. The overall weighted-average F1 score is then calculated as the sum of the F1 scores of each class, weighted by the number of true instances in each class.

The formula is:

$$Weighted\ F1 = \sum_{i=1}^{C} w_i F1_i \quad (4.4)$$

where:

- $C\ is\ the\ total\ no.\ of\ classes$
- $w_i\ is\ the\ weight\ for\ class\ i, calculated\ as$:

$$w_i = \frac{Number\ of\ samples\ in\ class\ i}{Total\ no.of\ samples} \quad (4.5)$$

Weighted average AUC: this extends the concept of AUC for multiclass classification by weighting the AUC for each class using the proportion of samples in that class.

The formula is:

$$Weighted\ AUC = \sum_{i=1}^{C} w_i AUC_i \quad (4.6)$$

where:

- $C\ is\ the\ total\ no.\ of\ classes$
- $AUC_i\ is\ the\ AUC\ for\ class\ i, calculated\ using\ the\ one\ vs\ rest\ method$

  $where\ each\ class\ is\ treated\ as\ the\ positive\ class\ against\ all\ others$
- $w_i\ is\ the\ weight\ for\ class\ i, calculated\ as$:

$$w_i = \frac{Number\ of\ samples\ in\ class\ i}{Total\ no.of\ samples} \quad (4.7)$$

4.4.5.3 Ordinal classification

Ordinal classification requires specialised metrics that account for the ranked nature of the target variable, which differ from those typically used in multi-class or binary classification. The standard evaluation metrics for ordinal regression. Accuracy and F1 score, which are commonly applied in multi-class problems, do not allow for the ordinal structure of the variable. For example, misclassifying a 2 instead of 1 is penalised equally as predicting 2 instead of 0. Intuitively, however, the latter misclassification is more erroneous due to the greater distance between the two categories. Thus, Mean Absolute Error (MAE) and Cohen's Weighted Kappa are more appropriate.

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and true ordinal values:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4.8)$$

where y_i and $\hat{y}_i$ are the true and predicted ordinal values.

Cohen's Weighted Kappa: This metric accounts for the degree of disagreement between predictions and ground truth, assigning higher penalties to larger misclassifications. This metric provides a robust measure of agreement for ordinal data (Yilmaz and Demirhan 2023).

These metrics collectively provide a deeper understanding of model performance in the context of ordinal regression. These modifications will be implemented for the ordinal variables encountered later in the dataset. While ordinal attributes consist of a minority of the overall variables, their significance necessitates these adjustments to ensure accurate and meaningful results.

4.4.5.3 95% Confidence interval

To assess the reliability and variability of model performance, bootstrapped 95% confidence intervals (95% CI) were calculated for the evaluation metrics generated on the independent test set (Bosma, Peeters et al. 2024). For each trained deep learning model, the original test dataset was resampled with replacement to generate 500 bootstrap samples, matching the size of the original test set. At each resampling iteration, the model was evaluated again on the resampled test set, generating new predicted outputs and recalculated performance metrics. This produced a distribution of results from which the 95% CI was derived. For binary and multiclass models, 95% CI were generated for Accuracy, F1 score, AUC (weighted for multiclass). For ordinal regression models, 95% CI were calculated for Accuracy, Mean Absolute Error (MAE), and quadratic-weighted Cohen's κ, consistent with recommendations for ordinal outcomes.

4.5. Results

Based on the above described methodology of carrying out the transfer learning models of InceptionV3, VGG19, EfficientNetV2B2, ResNet50V2, DenseNet121 and the implementation of the custom CNN along with the training (70%), validation (10%) and testing (20%) split of the data as described in the method, the accuracy, F1 score and AUC of each of those models on the testing data set across all 10 binary or multiclass attributes and the accuracy, MAE, and Cohen's weighted kappa produced for 4 ordinal attributes is presented in Table 4.4.

Table 4.4 Evaluation metrics of 6 deep learning models on the test sets of the pinkeye attributes presented in the scorecard.

Attributes	InceptionV3	VGG19	EfficientNetV2B2	ResNet50V2	DenseNet121	Custom CNN	Evaluation Metrics
	0.7895	0.8421	0.9637	0.8772	0.9123	0.8421	Accuracy
	0.5366	0.0000	0.9717	0.4615	0.7368	0.2540	F1
Stained	0.7704	0.3472	0.9912	0.8565	0.9745	0.4792	AUC
	0.8596	0.8947	0.9123	0.8596	0.8596	0.8596	Accuracy
	0.2453	0.3167	0.2891	0.2784	0.3025	0.3502	MAE
P score	0.7012	0.7859	0.8123	0.6751	0.7589	0.6983	Kappa
	0.6842	0.7544	0.7719	0.5789	0.7544	0.6842	Accuracy
	0.5000	0.0000	0.7524	0.5000	0.6316	0.5769	F1
Tear	0.7521	0.3789	0.8661	0.6895	0.7222	0.8077	AUC
	0.6842	0.7193	0.7719	0.5263	0.6667	0.5789	Accuracy
	0.3126	0.3998	0.3567	0.4052	0.3741	0.4295	MAE
Tear volume	0.6452	0.6987	0.7301	0.6712	0.6809	0.6113	Kappa
	0.5614	0.6491	0.7368	0.6140	0.5965	0.5439	Accuracy
	0.6988	0.0000	0.7246	0.7317	0.7089	0.7143	F1
Cornea opacity visible	0.7395	0.7231	0.8908	0.6365	0.7320	0.4404	AUC
	0.8246	0.8421	0.8246	0.7895	0.6667	0.7544	Accuracy
	0.6667	0.6667	0.7828	0.7286	0.6199	0.4300	F1
Cornea opacity touches limbus	0.8704	0.8804	0.8339	0.8073	0.8106	0.5000	AUC
	0.5789	0.7018	0.6140	0.3684	0.5439	0.3860	Accuracy
	0.2841	0.3418	0.3129	0.3798	0.3320	0.3592	MAE
Cornea opaqueness	0.6021	0.6723	0.6907	0.5894	0.6457	0.5638	Kappa
	0.3333	0.5439	0.5614	0.2456	0.3860	0.2456	Accuracy
	0.3994	0.4562	0.4205	0.4697	0.4821	0.4956	MAE
Cornea opacity size	0.5789	0.6134	0.6508	0.5421	0.5894	0.5033	Kappa
	0.8947	0.9474	0.9474	0.8596	0.8596	0.8246	Accuracy
	0.5046	0.7864	0.7864	0.3082	0.3082	0.4706	F1
	0.7531	0.8013	0.8425	0.6398	0.7342	0.6956	AUC
Corneal surface							

	0.9298	0.9123	0.9298	0.8421	0.8947	0.9123	Accuracy
	0.7143	0.7368	0.7500	0.0000	0.5714	0.2500	F1
Corneal blood vessels: hedges	0.8287	0.9583	0.9236	0.7662	0.8056	0.7154	AUC
	0.9123	0.8947	0.9298	0.9123	0.9474	0.9067	Accuracy
	0.2857	0.4000	0.3333	0.0000	0.5714	0.2445	F1
Corneal blood vessels: tree	0.7500	0.7923	0.7385	0.5346	0.7654	0.6450	AUC
	0.9474	0.9123	0.9474	0.9474	0.9474	0.9474	Accuracy
Corneal blood vessels: across	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	F1
lesion	0.4753	0.7160	0.7469	0.5309	0.5247	0.5210	AUC
	0.9001	0.9101	0.9200	0.9101	0.8722	0.8944	Accuracy
Corneal blood vessels: clearing from limbus	0.6742	0.6742	0.6677	0.6501	0.6122	0.6033	F1
	0.8902	0.8902	0.8204	0.7993	0.7540	0.7423	AUC

EfficientNetV2B2 consistently produced the best performance across all metrics (Table 4.4), while also achieving the shortest training times. Therefore, the results of this model for each attribute are presented here.

4.5.1 Stained

The EfficientNetV2B2 model achieved the highest performance on the testing set, reflected by high accuracy, F1 score, and AUC (Table 4.5). This balance between sensitivity and specificity is further supported by the confusion matrix (Figure 4.27), which shows minimal misclassification among the 661 test images. The high AUC value of 0.99 demonstrates the model's strong ability to differentiate between positive and negative classes, indicating excellent discrimination. Additionally, the high F1 score of 0.97 reflects a well-balanced trade-off between precision and recall, confirming that the model effectively identifies positive cases while reducing false positives and false negatives.

Table 4.5. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the stained category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9527	-	0.9858	-	0.9981	-
Validation	0.9424	-	0.9624	-	0.9901	-
Test	0.9637	(0.9470, 0.9773)	0.9717	(0.9589, 0.9827)	0.9912	(0.9905, 0.9983)

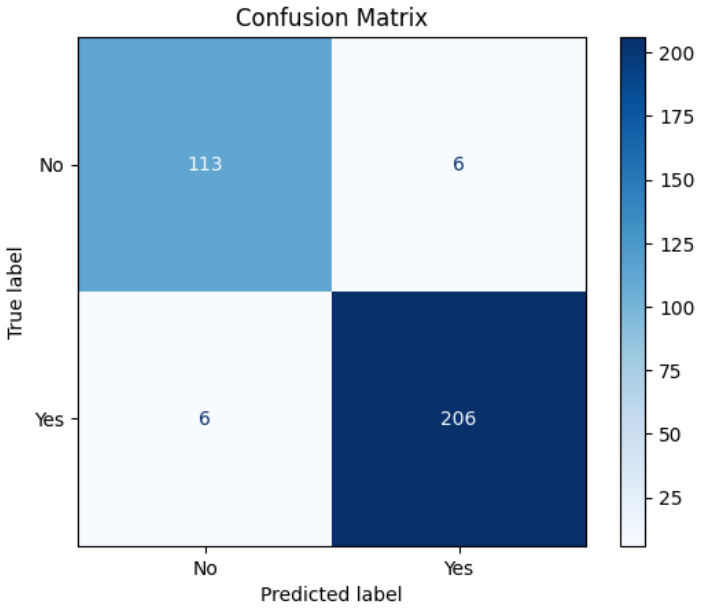


Figure 4.27. Confusion matrix: Class No: unstained eye, Class Yes: stained eye.

4.5.2 Tear

The testing accuracy of 77.19% indicates that the model shows some ability in classifying the 'tear' category and suggests further optimisation to improve will be required for the model's utility in practical applications. The testing AUC of 0.8661and F1 score of 0.7524 suggests moderate discriminative performance between tear and no tear (Table 4.6). Among the 661 testing images, the confusion matrix (Figure 4.28) revealed 103 false positive and 59 false negatives predictions.

Table 4.6. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the tear category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.7931	-	0.7038	-	0.8579	-
Validation	0.7818	-	0.6923	-	0.8408	-
Test	0.7719	(0.7439, 0.7999)	0.7524	(0.7010, 0.8038)	0.8661	(0.8211, 0.9111)

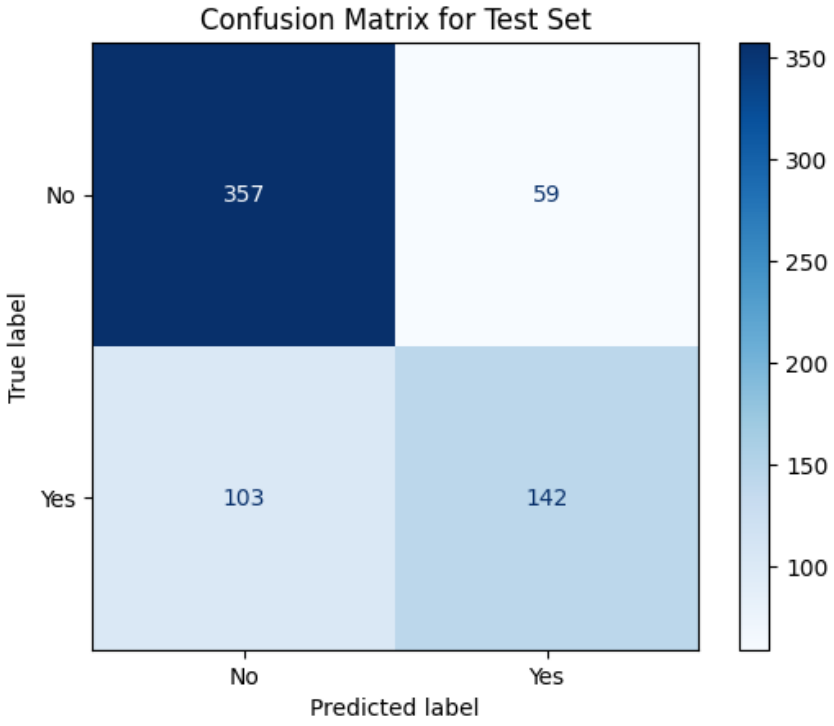


Figure 4.28. Confusion matrix: Class Tear: No; Class Tear: Yes.

4.5.3 Tear volume

The model demonstrated good performance in classifying the ordinal variable tear volume, achieving an accuracy of 77.19%, a MAE of 0.3567, and a Cohen's Weighted Kappa of 0.7301 (Table 4.7). The MAE indicates that most predictions are close to the true class labels, with minimal large deviations, while the Kappa score suggests moderate agreement, reflecting the model's ability to capture the ordinal relationships between classes.

Table 4.7. Accuracy, MAE and Cohen's weighted Kappa metrics for the training, validation and testing sets for the tear volume category using the EfficientNetV2B2 model.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.7965	-	0.4035	-	0.7575	-
Validation	0.8152	-	0.3848	-	0.7843	-
Test	0.7719	(0.7296, 0.8142)	0.3567	(0.3048, 0.4086)	0.7301	(0.6615, 0.7987)

4.5.4 Periocular score

The model demonstrated excellent performance in classifying the ordinal variable periocular score, achieving an accuracy of 91.23%, a low MAE 0.2891, and a high Cohen's Weighted Kappa of 0.8123 (Table 4.8). The lower MAE and higher Kappa demonstrate the model's ability to predict categories close to the true labels and effectively capture the ordinal nature of periocular score.

Table 4.8. Accuracy, MAE and Cohen's weighted Kappa metrics for the training, validation and testing sets for the periocular score category using the EfficientNetV2B2 model.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.9481	-	0.1589	-	0.9260	-
Validation	0.9091	-	0.2091	-	0.8405	-
Test	0.9123	(0.8643, 0.9603)	0.2891	(0.2331, 0.3451)	0.8123	(0.7473, 0.8773)

4.5.5 Cornea opacity visible

The model demonstrated strong performance in classifying the cornea opacity visible attribute, achieving a testing accuracy of 73.68% and an AUC of 0.8908 (Table 4.9). These results indicate that the model has excellent discriminatory power, with the high AUC value highlighting its ability to differentiate between positive and negative cases effectively. The F1 score further confirms the model's balanced performance. Among the testing images, the confusion matrix (Figure 4.29) revealed minimal misclassification, reflecting a good balance between sensitivity and specificity.

Table 4.9. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the cornea opacity visible category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9515	-	0.9526	-	0.9897	-
Validation	0.8606	-	0.8614	-	0.9152	-
Test	0.7368	(0.699, 0.7746)	0.7246	(0.6848, 0.7644)	0.8908	(0.8484, 0.9332)

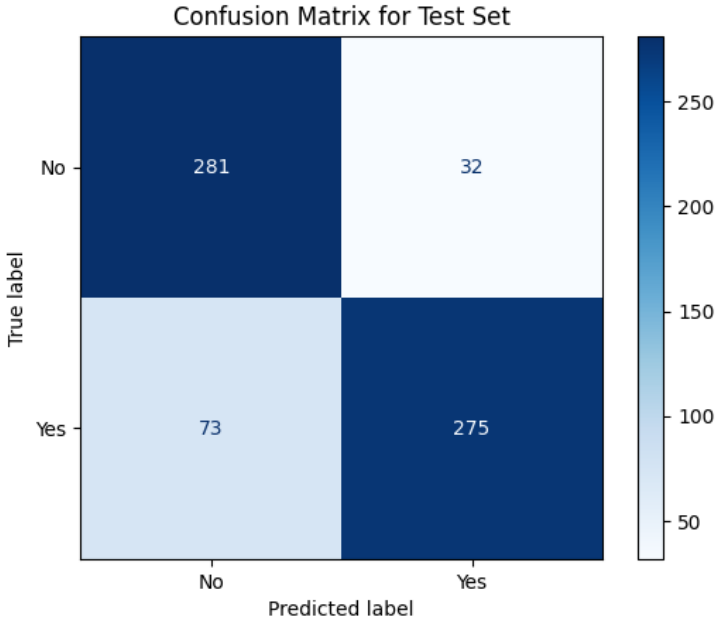


Figure 4.29. Confusion matrix: Class Cornea opacity visible: No; Class Cornea opacity visible: Yes.

4.5.7 Cornea opacity touches limbus

The model achieved robust performance in classifying this attribute, with a testing accuracy of 82.46%, an AUC of 0.8339, and an F1 score of 0.7828 for the testing set (Table 4.10). The AUC value reflects good discriminatory power. The high testing accuracy indicates reliable predictions for most samples, while the F1 score highlights the model's moderate ability to maintain this balance for identifying positive cases. The confusion matrix (Figure 4.30) further supports these findings, showing relatively low misclassification and a good trade-off between sensitivity and specificity.

Table 4.10. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the cornea opacity touches limbus category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.8918	-	0.7465	-	0.9304	-
Validation	0.8727	-	0.6957	-	0.8903	-
Test	0.8246	(0.7934, 0.8558)	0.7828	(0.746, 0.8196)	0.8339	(0.7987, 0.8691)

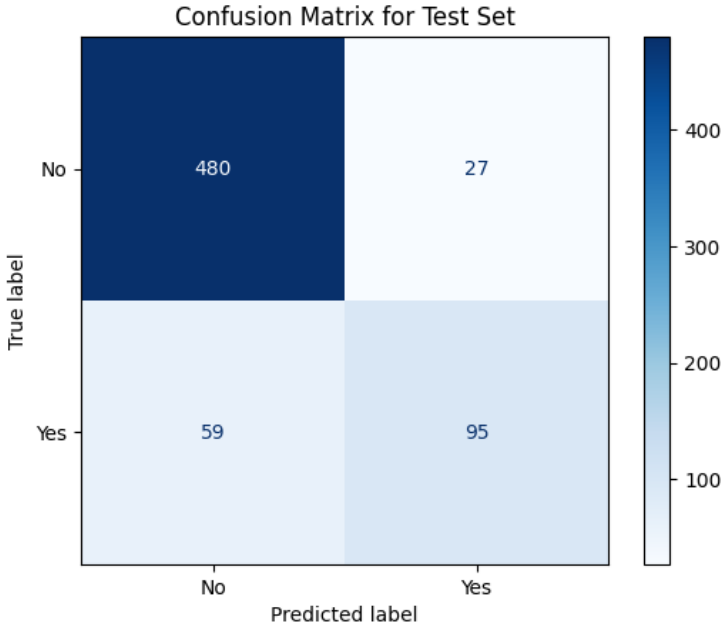


Figure 4.30. Confusion matrix: Class Cornea opacity touches limbus: No; Class Cornea opacity touches limbus: Yes.

4.5.8 Cornea opaqueness

The model showed good performance in classifying the ordinal variable cornea opaqueness, with an accuracy of 61.40%, a MAE of 0.3129, and a Cohen's Weighted Kappa of 0.6907 (Table 4.11). The moderate MAE indicates that the model's predictions are generally close to the true labels, while the high Kappa score highlights strong agreement, suggesting the model effectively captured the ordinal structure of the variable.

Table 4.11. Accuracy, MAE and Cohen's weighted Kappa metrics for the training, validation and testing sets for the cornea opaqueness category using the EfficientNetV2B2 model.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0. 8515	-	0.1805	-	0.8587	-
Validation	0. 7576	-	0.2909	-	0.7760	-
Test	0. 6140	(0.5605, 0.6675)	0.3129	(0.2667, 0.3591)	0.6907	(0.6456, 0.7358)

4.5.9 Cornea opacity size

The model demonstrated moderate performance in classifying the ordinal variable cornea opacity size, with an accuracy of 56.14%, a MAE of 0.4205, and a Cohen's Weighted Kappa of 0.6508 (Table 4.12). Compared to other ordinal variables, the lower accuracy and higher MAE suggest greater difficulty in correctly predicting the class labels, though the relatively high Kappa indicates that the model still captures the ordinal nature of the variable to a reasonable extent.

Table 4.12. Accuracy, MAE and Cohen's weighted Kappa metrics for the training, validation and testing sets for the cornea opacity size category using the EfficientNetV2B2 model.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.7199	-	0.3805	-	0.8768	-
Validation	0.6515	-	0.4848	-	0.8393	-
Test	0.5614	(0.4956, 0.6272)	0.4205	(0.3518, 0.4892)	0.6508	(0.5829, 0.7187)

4.5.10 Corneal surface

The model achieved mixed results in classifying this attribute, with a testing accuracy of 94.74%, a weighted average F1 score of 0.7864, and a weighted average AUC of 0.8425 (Table 4.13). The confusion matrix (Figure 4.31) shows that the model predicted only flat for all inputs, correctly classifying most flat samples but misclassifying all crater/depressed and raised instances. This suggests a strong bias toward Flat, likely due to class imbalance in the data.

Table 4.13. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the Corneal surface category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1 (weighted average)	95% CI	AUC (weighted average)	95% CI
Train	0.9429	-	0.9210	-	0.9527	-
Validation	0.9394	-	0.9127	-	0.9246	-
Test	0.9474	(0.9147, 0.9801)	0.7864	(0.7539, 0.8189)	0.8425	(0.8068, 0.8782)

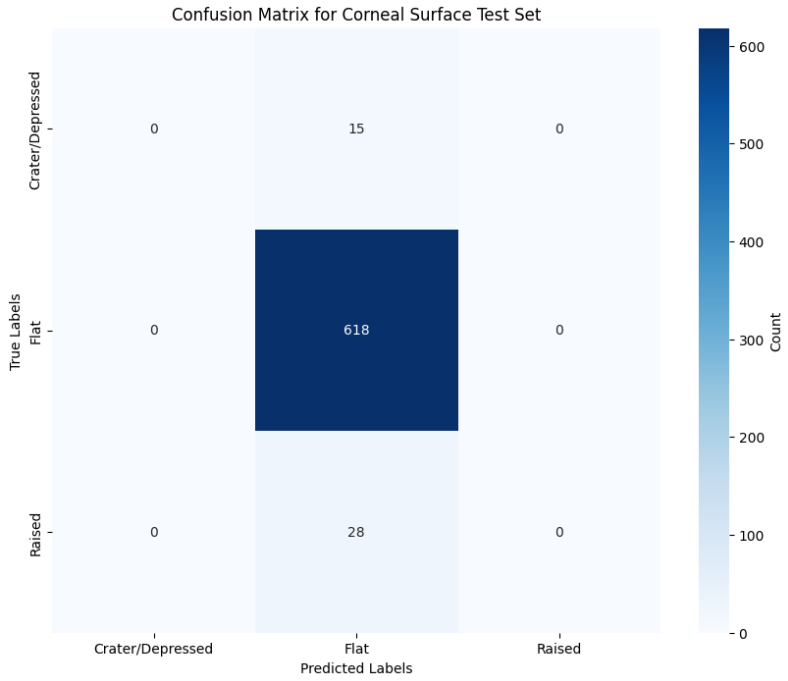


Figure 4.31. Confusion matrix: Class Corneal surface: Crater/Depressed; Class Corneal surface: Flat; Class: Corneal surface: Raised.

4.5.11 Corneal blood vessels (hedges)

The model achieved high performance in classifying the corneal blood vessels (hedges)' attribute, with a testing accuracy of 94.10%, an AUC of 0.9505, and an F1 score of 0.7607 (Table 4.14). The high accuracy and AUC indicate that the model is highly effective at distinguishing between positive and negative cases. However, the F1 score is noticeably lower than the accuracy and AUC, likely due to an imbalance in precision and recall. This suggests that while the model correctly predicts a high proportion of samples overall (reflected in accuracy) and differentiates between classes well (reflected in AUC), it may struggle to achieve a balance between precision and recall for the positive class at the cost of potentially misclassifying negatives as positives as seen in the confusion matrix (Figure 4.32).

Table 4.14. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the Cornea blood vessels (hedges) category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9775	-	0.9068	-	0.9930	-
Validation	0.9545	-	0.8000	-	0.9733	-
Test	0.9289	(0.8994, 0.9584)	0.7500	(0.7227, 0.7773)	0.9236	(0.8955, 0.9517)

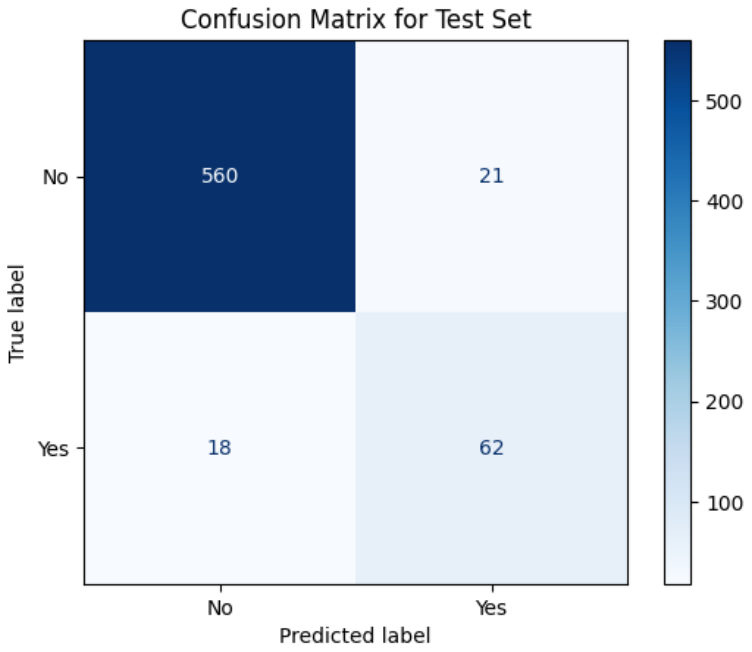


Figure 4.32. Confusion matrix: Class Cornea blood vessels (hedges): No; Class Cornea blood vessels (hedges): Yes.

4.5.12 Corneal blood vessels (trees)

The model demonstrated good performance in classifying this attribute, with a testing accuracy of 92.98%, a weighted average F1 score of 0.33, and a weighted average AUC of 0.74 (Table 4.15). The moderate AUC value indicates fair discrimination between positive and negative cases, suggesting that while the model performs reliably overall, it misclassified the majority of Corneal blood vessels (trees): Yes cases as Corneal blood vessels (trees): No (Figure 4.33).

Table 4.15. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for Corneal blood vessels (trees) category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1 (weighted average)	95% CI	AUC (weighted average)	95% CI
Train	0.8961	-	0.4112	-	0.9225	-
Validation	0.8896	-	0.4322	-	0.8164	-
Test	0.9298	(0.8772, 0.9824)	0.3333	(0.2671, 0.4019)	0.7385	(0.667, 0.81)

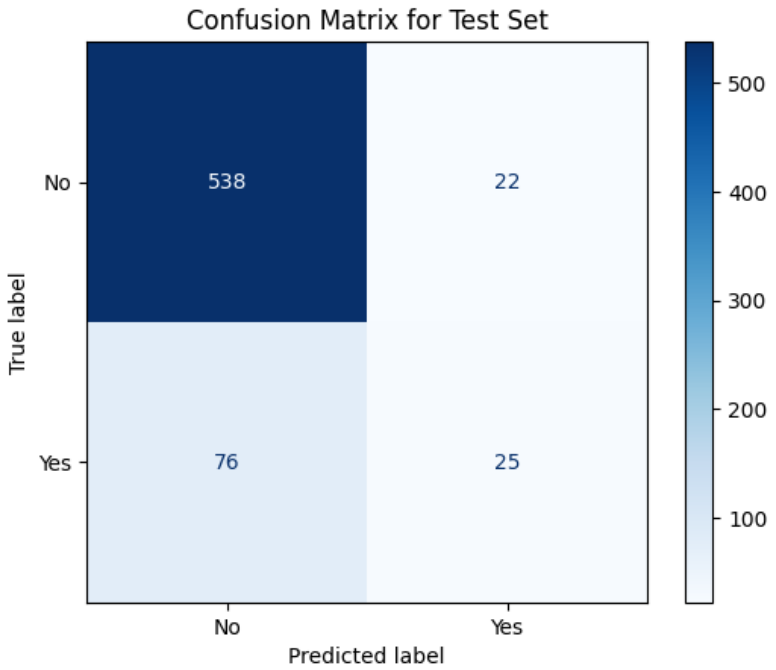


Figure 4.33. Confusion matrix: Class Corneal blood vessels (trees): No; Class Corneal blood vessels (trees): Yes.

4.5.13 Corneal blood vessels (across lesion)

The model achieved mixed performance in classifying this attribute, with a testing accuracy of 94.74%, a weighted average F1 score of 0, and a weighted average AUC of 0.7469 (Table 4.16). The relatively high AUC value highlights the model's excellent ability to distinguish between positive and negative cases, while the weighted F1 score demonstrates a poor balance between precision and recall across all classes. These results along with the confusion matrix (Figure 4.34) indicate that the model is reliable for the classification of Class Corneal blood vessels (across lesion): No, with misclassification issues for Class Corneal blood vessels (across lesion): Yes.

Table 4.16. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the Corneal blood vessels (across lesion) category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1 (weighted average)	95% CI	AUC (weighted average)	95% CI
Train	0.9294	-	0.3333	-	0.9540	-
Validation	0.8818	-	0.3280	-	0.8801	-
Test	0.9474	(0.9147, 0.9801)	0.0000	NA	0.7469	(0.667, 0.8268)

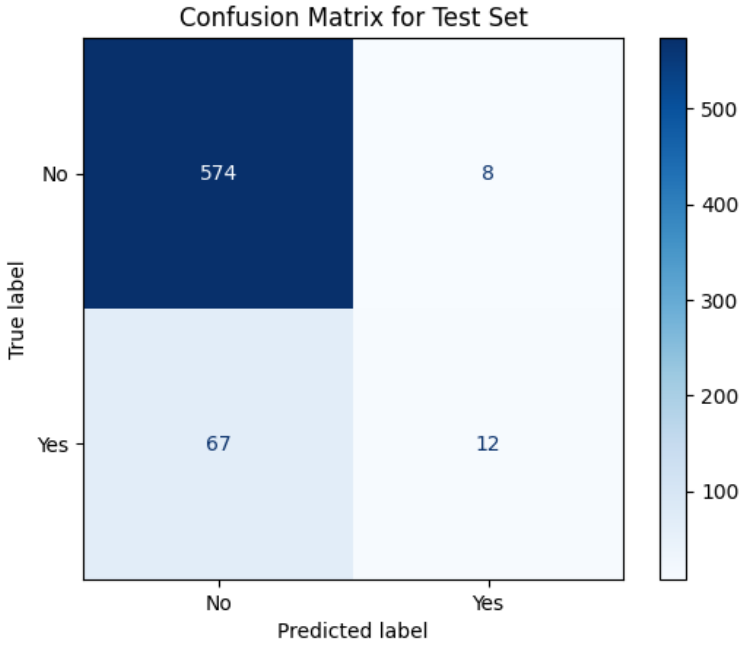


Figure 4.34. Confusion matrix: Class Corneal blood vessels (across lesion): No; Class Corneal blood vessels (across lesion): Yes.

4.5.14 Corneal blood vessels (clearing from limbus)

The model achieved mixed performance in classifying this attribute, with a testing accuracy of 92%, a weighted average F1 score of 0.6677, and an AUC of 0.8204 (Table 4.17). While the AUC indicates that the model has a good ability to distinguish between positive and negative cases, the lower F1 score reflects an imbalance in precision and recall across classes. These results, along with the confusion matrix (Figure 4.35), suggest that the model reliably

identifies Cornea opacity touches limbus: No but fails to correctly classify instances of Cornea opacity touches limbus: Yes.

Table 4.17. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for the cornea opacity touches limbus category using the EfficientNetV2B2 model.

	Accuracy	95% CI	F1 (weighted average)	95% CI	AUC (weighted average)	95% CI
Train	0. 9753	-	0.8655	-	0.9540	-
Validation	0. 9758	-	0.7830	-	0.8801	-
Test	0. 9200	(0.8905, 0.9495)	0.6677	(0.6341, 0.7013)	0.7469	(0.7197, 0.7741)

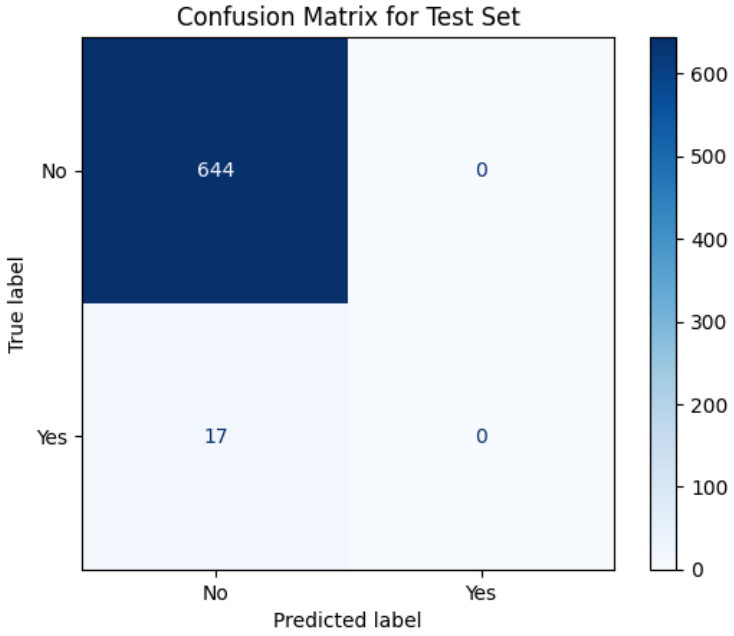


Figure 4.35. Confusion matrix: Class Cornea opacity touches limbus: No; Class Cornea opacity touches limbus: Yes.

4.6 Discussion

4.6.1 Deep Learning Models for pinkeye Analysis

The deep learning models evaluated in this study, EfficientNetV2B2, VGG19, InceptionV3, ResNet50V2, DenseNet121, and a custom CNN demonstrated varying degrees of performance across different attributes of pinkeye,

reflecting their unique strengths and limitations. However, EfficientNetV2B2 consistently emerged as the top performing model, excelling in binary, multiclass classification and ordinal regression classification tasks.

EfficientNetV2B2 achieved notable higher accuracy, F1 score, and AUC across most attributes, suggesting its ability to handle the challenges of pinkeye classification. For binary tasks, it consistently outperformed other models, with results exceeding those of other models for attributes such as 'tear' (77.19% accuracy), 'cornea opacity visible' (73.68% accuracy), and ‘corneal blood vessels (hedges)' (92.98% accuracy). For ordinal attributes, such as PScore (91.23% accuracy) and Tear Volume (77.19% accuracy), EfficientNetV2B2 maintained strong performance despite inherent challenges. Its success is attributed to compound scaling, which balances width, depth, and resolution optimally, alongside integrated regularisation techniques that enhance generalisation, which is reinforced by other ophthalmology DL studies (Fayyad and Mustakim 2024, Mohith, Raja et al. 2024), which suggests that EfficientNet consists of a network architecture that excels in extracting fine-grained details in eye images and parsing that data for analysis.

However, a significant limitation in interpreting these results is the lack of comparable studies on animal eye deep learning tasks, especially involving cattle. Recent studies have mostly focused on canine and feline eyes, applying models like ResNet50 and DenseNet121 to classify different eye diseases in a multiclass setting (Nam and Dong 2023). While these tasks share similarities with our multiclass classification of corneal surface, they differ substantially in context. For example, the accuracy achieved in these studies was lower, at 0.81 for ResNet50 and 0.828 for DenseNet121 for canine eyes, and 0.605 for ResNet50 and 0.609 for DenseNet121 for feline eyes, compared to 93% accuracy in our study. This disparity likely arises from differences in task complexity, dataset characteristics, and preprocessing methods, such as the use of staining in our dataset.

VGG19 also performed well but was slightly outshone by EfficientNetV2B2 in terms of both computational efficiency and accuracy. For instance, while VGG19 produced competitive accuracy for tasks like 'stained' and 'PScore', its longer training times and higher computational costs, due to its more complex architecture and much higher parameter count limited its practicality compared to EfficientNetV2B2. Other models, such as InceptionV3 and ResNet50V2, offered faster training but struggled with more nuanced classifications, such as 'cornea opacity size', where accuracy dipped below 50%. These models performed better in binary tasks with well-defined features but were less effective in multiclass tasks requiring differentiation of subtle visual cues.

4.6.2 Binary and multiclass classification analysis

This study examined eight binary attributes (Tear, Stained, Cornea Opacity Visible, Cornea Opacity Touches Limbus, Corneal Blood Vessels (Hedges), Corneal Blood Vessels (Trees), Corneal Blood Vessels (Across Lesion), and Corneal Blood Vessels (Clearing from Limbus)) and 1 multiclass attribute (Corneal Surface). The binary attributes consistently demonstrated the challenges posed by dataset imbalances, which manifested as disparities between performance metrics such as accuracy, F1-scores, and AUC.

For the binary attributes, models generally achieved high accuracy and AUC scores, reflecting robust discriminatory power overall. However, these metrics often obscured the challenges in correctly identifying instances of minority classes. For example, the Tear attribute achieved a relatively high AUC of 0.8661 on the test set, indicating good overall discriminatory ability. However, the corresponding F1-score was notably lower at 0.7524, highlighting difficulties in balancing sensitivity and specificity. The confusion matrix further revealed a higher false negative rate, with 103 yes samples misclassified as no, compared to 59 false positives. This imbalance in predictions suggests the model struggled to identify the minority yes class effectively, which significantly impacted recall. A possible explanation for this performance could be the trade-offs involved in the image cropping process.

While cropping helped reduce the influence of random objects in the eye images as these potential distractions could lead to classification errors, it may have also removed portions of the broader face that contained important tear streaks as these tear streaks might provide significant contextual information for classifying this attribute. This trade-off between minimising distractions for pinkeye analysis and retaining critical pixel data for tear classification warrants further investigation. Future work could explore optimising cropping strategies to balance these competing needs, maximising classification performance without losing essential information relevant to this attribute. A similar trend was observed for the Cornea Opacity Visible attribute, where the model achieved a strong test AUC of 0.8908 but faced challenges in detecting the yes class. The confusion matrix highlighted a substantial false negative count of 73 compared to 32 false positives. While the model's overall accuracy of 73.68% and F1-score of 0.7246 reflected reasonable generalisation, the imbalance in recall and precision for the minority class remained evident.

The Stained attribute stood out as an exception, with minimal impact from dataset imbalance. This trait's strong predictive accuracy was expected due to its distinct visual characteristics, serving as a positive control to validate the deep learning model's learning capacity and predictive capabilities of the model in this study. The model achieved test accuracy above 98% and an AUC of 0.99. The balanced confusion matrix, with only six false negatives and six false positives, demonstrated robust performance for both classes, which could be due to the obvious highlighted stained/inked pixel area compared to non-inked areas of the eye which can be looked further upon with X-AI techniques to see if these areas are highly significant to the contribution of the DL model’s high performance.

On the other hand, attributes relating to corneal blood vessels and corneal opacity demonstrated significant challenges, particularly with minority class detection since it is too heavily skewed towards the majority class. Weighted F1 scores, which account for class imbalances by assigning proportional

weights to each class based on its size were used to provide a more comprehensive performance evaluation than the raw F1 score. For Cornea Opacity Touches Limbus, the weighted F1-score for the test set was 0.7828, indicating relatively modest overall performance. The confusion matrix revealed 59 false negatives for the minority class (Yes), which represented over 38% of all actual Yes samples. This high false negative rate significantly lowered recall for the minority class, despite a reasonable overall accuracy of 82.46%. The AUC for this attribute was 0.8339 on the test set, reflecting good discriminatory ability but the performance imbalance across classes underscores the difficulty in reliably detecting Yes samples.

A similar trend was observed in the analysis of attributes relating to corneal blood vessels. The attribute Corneal Blood Vessels (Trees) achieved a weighted F1-score of 0.3333 on the test set, reflecting the model's struggles with the minority class (Yes). The model misclassified 76 Yes samples as No, and while the AUC was 0.7385 (a moderate level of discriminatory ability), it could not compensate for the low recall, which significantly hindered the model's effectiveness in detecting Yes samples. Likewise, for the attribute Corneal Blood Vessels (Across Lesion), the test weighted F1-score of 0 indicated reflected the model's difficulty in distinguishing between the two classes when imbalance is a major issue.

The Corneal Blood Vessels (Clearing from Limbus) attribute underscored the limitations of class imbalance even more starkly. The model’s performance was driven entirely by the model’s perfect classification of the majority class (No). The weighted F1 of 0.6677 revealed reduced discriminatory power, and the confusion matrix exposed the model’s complete failure to classify any Yes samples correctly, with all 17 instances misclassified as No.

The use of weighted scores highlights the model's overall ability to balance performance across classes but also reveals the disproportionate contribution of the majority class to these metrics. While the high weighted F1-scores reflect the model’s strength with the dominant class, the poor recall and precision for minority classes emphasise the need for strategies to address

class imbalance, such as oversampling, threshold adjustment, class-weighted loss functions, to ensure more equitable performance across all classes, which can be implemented for the analysis of pinkeye stages and severity levels in Chapter 5, hence showing the importance of these insights obtained from the analyses conducted in Chapter 4.

The multiclass attribute further highlighted the limitations of models when faced with extreme dominance of one class over others. Corneal Surface presented additional challenges due to extreme dominance of the majority class (Flat). While the model achieved over 94% accuracy and a weighted F1-score of 0.7864 on the test set, it failed to predict any instances of the minority classes (Raised and Crater/Depressed). This was evident in the confusion matrix, where no true positives were recorded for either minority class.

Across all attributes, it is clear that class imbalance played a significant role in limiting model performance. High overall metrics often masked poor sensitivity for minority classes, as evidenced by high false negative rates in confusion matrices. Attributes with more balanced data, such as Stained, demonstrated markedly better performance, reinforcing the importance of addressing imbalances.

4.6.3 Ordinal variable analysis

This study analysed four ordinal variables (Tear Volume, Periocular Score, Cornea Opaqueness, and Cornea Opacity Size) to evaluate the model's ability to predict ranked categories effectively. Ordinal data adds complexity to classification tasks due to the need for the model to account for the inherent ordering of the categories. Metrics such as accuracy, MAE, and Cohen’s kappa were used to assess model classification performance.

For the Tear Volume variable, the model achieved moderate performance across datasets. Test accuracy was 77.19%, with an MAE of 0.3567 and a kappa score of 0.7301, reflecting a reasonable level of agreement between predicted and true rankings. In literature, segmentation-based approaches

have also been explored for analysing eye conditions, where a U-Net architecture combined with ResNet and EfficientNet backbones achieved a Jaccard coefficient index of 0.8 for sclera redness but struggled with excessive tearing (0.3) for canine eyes. These challenges align with our findings for tear-related attributes, where nuanced features like tear streaks were difficult for the model to identify accurately (Buric 2024).

While the Periocular Score variable demonstrated strong overall performance, with test accuracy reaching 91.23%, an MAE of 0.2891, and a kappa score of 0.8123, the Cornea Opaqueness variable presented moderate performance, with test accuracy of 75.79%, an MAE of 0.2935, and a kappa score of 0.7704. Upon further breakdown analysis, for the majority classes (Clear and Completely Opaque), the model performed well, achieving 91% and 89% accuracy, respectively. However, intermediate classes such as Mild and Moderate exhibited significantly lower performance. For example, only 32% of Moderate instances were correctly classified, with frequent misclassifications into adjacent classes such as Mild and Completely Opaque. The imbalanced distribution of classes was a significant factor influencing the model's reduced accuracy for underrepresented categories. In addition, these findings show that the performance for minority classes is often overshadowed by the accuracy of dominant ones.

The Cornea Opacity Size variable was the most challenging for the DL models to analyse, with the lowest test accuracy at 56.14%, an MAE of 0.4205, and a kappa score of 0.6508. While the model performed well for the extreme classes (Normal Eye and >75%), achieving 92% and 74% accuracy respectively, intermediate classes such as 26-50% and 51-75% showed significantly higher misclassification rates. For example, only 28% of 51-75% instances were correctly predicted, with notable misclassifications into 26-50% and <10%. These results underscore the model's difficulty in distinguishing between closely ranked and underrepresented classes, which likely stems from the skewed class distribution and the nuanced nature of intermediate categories. It is worth noting that the boundaries for opacity size were established by a

group of experts at the start of the annotation process, and the images were annotated at the eye level by a human annotator. Consequently, human or random errors could occur, particularly when judging whether an opaque area falls between 26–50% and 51–75%. Since these boundaries are not extremely precise, errors are more likely to arise in these overlapping categories.

Across all ordinal variables, the performance was strongest for majority or extreme classes, which benefited from either higher representation in the dataset or more obvious pixelated differences observed in the images. Intermediate and minority classes, on the other hand, exhibited higher misclassification rates and lower predictive accuracy. Metrics such as MAE and kappa provided a balanced evaluation of performance, revealing that even when accuracy was relatively high, the model struggled to achieve precise predictions for underrepresented and intermediate categories.

4.6.4 Limitations and recommendations

This study highlights several challenges inherent in applying deep learning to veterinary diagnostics, particularly for detecting and classifying pinkeye-related attributes. A key limitation observed across multiple attributes was the impact of imbalanced datasets. Attributes such as Corneal Blood Vessels (Trees) and Corneal Blood Vessels (Across Lesion) demonstrated high overall accuracy (92.98% and 94.74%, respectively), yet their performance was heavily skewed toward the majority class. Minority classes were often misclassified, leading to low sensitivity and recall. Similarly, ordinal variables such as Cornea Opacity Size and Cornea Opaqueness revealed significant struggles in distinguishing between underrepresented intermediate classes since there were many classes within each attribute and the visual differences between the intermediate classes are often minute. These challenges highlight the need for more nuanced strategies in the modelling process to handle imbalances effectively such as augmentation and regularisation strategies, particularly as the dataset's natural skew reflects the real-world rarity of certain conditions (Buda, Maki et al. 2018, Johnson and Khoshgoftaar

2019). This insight is helpful for the analyses which will be carried out for the detecting of pinkeye stages and severity levels in Chapter 5 as dataset imbalance will be encountered.

Environmental artefacts such as reflections, obstructions, and varying image quality further complicated classification tasks. These artefacts obscured critical features, reducing model reliability for attributes where fine-grained visual cues were essential, such as Tear Volume and Cornea Opacity Size (Razzak, Naz et al. 2018). In some cases, preprocessing strategies designed to improve classification outcomes for pinkeye inadvertently hampered performance for specific attributes such as Tear volume. For example, the cropping technique employed to remove distracting elements from the background also excluded portions of tear streaks, which are critical for accurately predicting Tear Volume. This trade-off between reducing distractions and retaining informative visual features reflects the broader challenge of balancing preprocessing decisions for multiple classification objectives.

A big avenue for improvement lies in optimising the model's handling of class imbalances as this was one of the biggest and most consistent issue found in all these analyses. Some strategies to address this include refining the loss functions used during training to penalise misclassifications based on their distance within an ordinal scale or their importance in a clinical context (Cao, Mirjalili et al. 2020). Weighted loss functions could ensure that minority classes contribute proportionally to the training process, mitigating the model's bias toward dominant classes (Johnson and Khoshgoftaar 2019). Similarly, advanced augmentation strategies that preserve the integrity of minority class features could bolster model performance especially for underrepresented categories, provided the dataset for these classes is still sufficiently substantial considering the less than ideal results of augmentation implemented for certain attributes. Variables such as Corneal Surface exhibited limited improvements despite augmentation, as some classes lacked sufficient original samples to generate meaningful synthetic data

(Perez and Wang 2017). Moreover, inflated training accuracies relative to testing accuracies suggest that overfitting may have occurred in some cases, emphasising the need for careful implementation of augmentation strategies. This can also increase the risk of overfitting when augmenting underrepresented data. A few targeted strategies can be implemented in Chapter 5 to minimise this issue. First, employing diverse augmentation techniques such as rotations, brightness adjustments, noise addition, and cropping ensures that the augmented dataset contains significant variability, reducing the likelihood of the model memorising repeated patterns (Shorten and Khoshgoftaar 2019). Regularisation techniques, including dropout layers and L2 weight regularisation, can further prevent the model from over-relying on specific augmented features, encouraging it to learn generalisable patterns (Srivastava, Hinton et al. 2014). Additionally, assigning lower weights to augmented samples during training helps the model prioritise learning from the original data while still benefiting from the diversity introduced by augmentation. Finally, incorporating early stopping based on validation loss ensures that training halts before the model begins overfitting, especially to the augmented data.

The insights gained from these attribute-level analyses lay a strong foundation for future work, which will focus on using these findings to refine the detection of pinkeye stages and severity levels. While EfficientNet demonstrated robustness in handling various attributes, its performance highlights the need for further customisation to address the nuances of pinkeye classification. Future iterations will prioritise tailoring the model's preprocessing steps, augmentation techniques, and loss functions to better accommodate the complexities of different attributes. By aligning these refinements with insights from this study, the next chapter will develop a framework for improving diagnostic accuracy across stages and severity levels of pinkeye.

4.7. Conclusion

This chapter has laid the foundation for understanding the performance and limitations of 5 deep learning models (VGG19, ResNet50, DenseNet121, EfficientNetV2B2, and InceptionV3) regularly used in ophthalmology literature in the analysis of cattle eye images, including an inbuilt custom CNN model. In the analysis of pinkeye-related attributes across binary, multiclass, and ordinal classification tasks, EfficientNetV2B2 consistently outperformed other models investigated in accuracy, F1 and AUC (for binary and multiclass classification) and MAE and Kappa (for ordinal classification). By evaluating eight binary attributes, one multiclass attributes, and four ordinal variables, this study uncovered critical insights into the strengths and weaknesses of the top performing DL when faced with real-world challenges such as dataset imbalance, subtle visual distinctions, and environmental artefacts.

The results demonstrated that EfficientNet is robust in handling majority classes, with high accuracy and weighted F1-scores observed for many attributes. However, the analysis also revealed significant challenges in detecting minority classes, particularly for underrepresented categories in imbalanced datasets. For ordinal variables, such as Cornea Opacity Size and Periocular Score, the model struggled to differentiate intermediate classes, highlighting the complexity of subtle distinctions in ranked categories.

Despite these challenges, the chapter provided actionable strategies for addressing the limitations observed as dataset imbalance and subtle category distinctions are expected to persist in the analysis of pinkeye stages in Chapter 5. Techniques such as diverse augmentation pipelines, regularisation, and careful implementation of loss weighting were identified as effective ways to improve performance for underrepresented categories while mitigating overfitting risks. By building on this foundation, the subsequent analysis aims to create practical, reliable AI tools for farmers to promptly diagnose and improve the management of pinkeye in cattle.

Chapter 5. Deep learning modelling for classification of pinkeye disease stage and severity

5.1 Introduction

Pinkeye, or bovine keratoconjunctivitis, is a significant ocular disease affecting cattle, with substantial economic and welfare implications due to reduced productivity, impaired vision, and, in severe cases, permanent blindness. Early detection and accurate classification of disease stages are critical for implementing timely treatment and preventing irreversible damage. However, accurately assessing pinkeye progression remains challenging due to overlapping symptoms across disease stages and variability in presentation influenced by factors such as breed differences, pigmentation, and environmental conditions.

In previous chapters, various DL models were applied to analyse cattle eye images and attributes associated with pinkeye, including tear presence, corneal opacity, and blood vessel patterns. These investigations aimed to identify optimal architectures for accurately classifying and understanding pinkeye disease stages and severity. Classical ophthalmology architectures such as VGG, Inception, DenseNet, EfficientNet, and ResNet were tested, with EfficientNet emerging as the most effective model due to its superior performance across multiple attributes (Chapter 4). This result aligns with findings in related domains, where EfficientNet has demonstrated strong performance in tasks requiring high precision and robustness (Tan and Le 2019, Tan and Le 2021).

Despite these successes, developing a comprehensive and standardised scoring system for pinkeye remains a challenge. Existing scoring frameworks, such as those assessing ulceration severity and eye discharge (Ward and Nielson 1979), are often simplistic, outdated, or insufficiently detailed to

capture the full spectrum of symptoms observed in pinkeye. Furthermore, standardisation is complicated by differences among cattle breeds and pigmentation of the eyelids, which can influence disease severity and complicate visual assessments (Kneipp 2021).

To address these limitations, a new scoring system was developed through an extensive literature review and expert consultations with veterinarians from Australia and the USA. This process involved multiple iterations and revisions, guided by feedback from training sessions conducted by a field veterinarian with expertise in pinkeye, where veterinarians assessed example images from their database. The finalised scorecard, presented in Chapter 4, encompasses a wide range of attributes relevant to pinkeye stages and severity, providing a sound framework for disease assessment. The focus of this chapter is to leverage EfficientNet for the classification of pinkeye disease stages, building upon the foundations established in previous chapters. The objectives are threefold:

1. Apply EfficientNetV2B2 for binary, multi-class, and ordinal classification of pinkeye stages.
2. Evaluate model performance across these frameworks to identify strengths and limitations.
3. Explore potential improvements in classification accuracy through advanced computational methods.

These objectives aim to enhance diagnostic precision and provide quantitative insights into disease progression, demonstrating how automated classification through DL can inform treatment strategies for pinkeye. This chapter contributes to the broader goal of developing a reliable diagnostic tool for pinkeye detection and staging, particularly for application in resource-limited agricultural settings.

5.2 Method

5.2.1 Dataset description

5.2.1.1 Expert review and annotation

To develop a comprehensive dataset for classifying pinkeye stages in cattle, a systematic scoring system was established through a combination of literature review and expert consultations, as detailed in Chapter 4. This scoring system categorised images into four primary classes: Normal, Active (A), Resolving (R), and Resolved (S), each subdivided into specific levels. We used this scoring system to annotate 3,301 images, which formed the ground truth dataset. Annotations were compiled into a CSV file paired with image IDs, serving as the basis for training, validating and evaluating DL algorithms in subsequent analyses.

The general descriptors for each stage are presented in Table 5.1.

Table 5.1. Developed categories and their general descriptors used for classification

Categories	General descriptors
Normal	Healthy; absence of ulcer
Active (A)	Teary discharge; visible opacity; coloured ulcer; hedge and tree blood vessels
Resolving (R)	No discharge; centre is still raised and dense opacity; clearing of blood vessels from the limbus (no hedges); trees (vessels) may be present
Resolved (S)	White scar; no signs of tear or active inflammation; no corneal vascularisation

5.2.2.1 Description of each of the categories

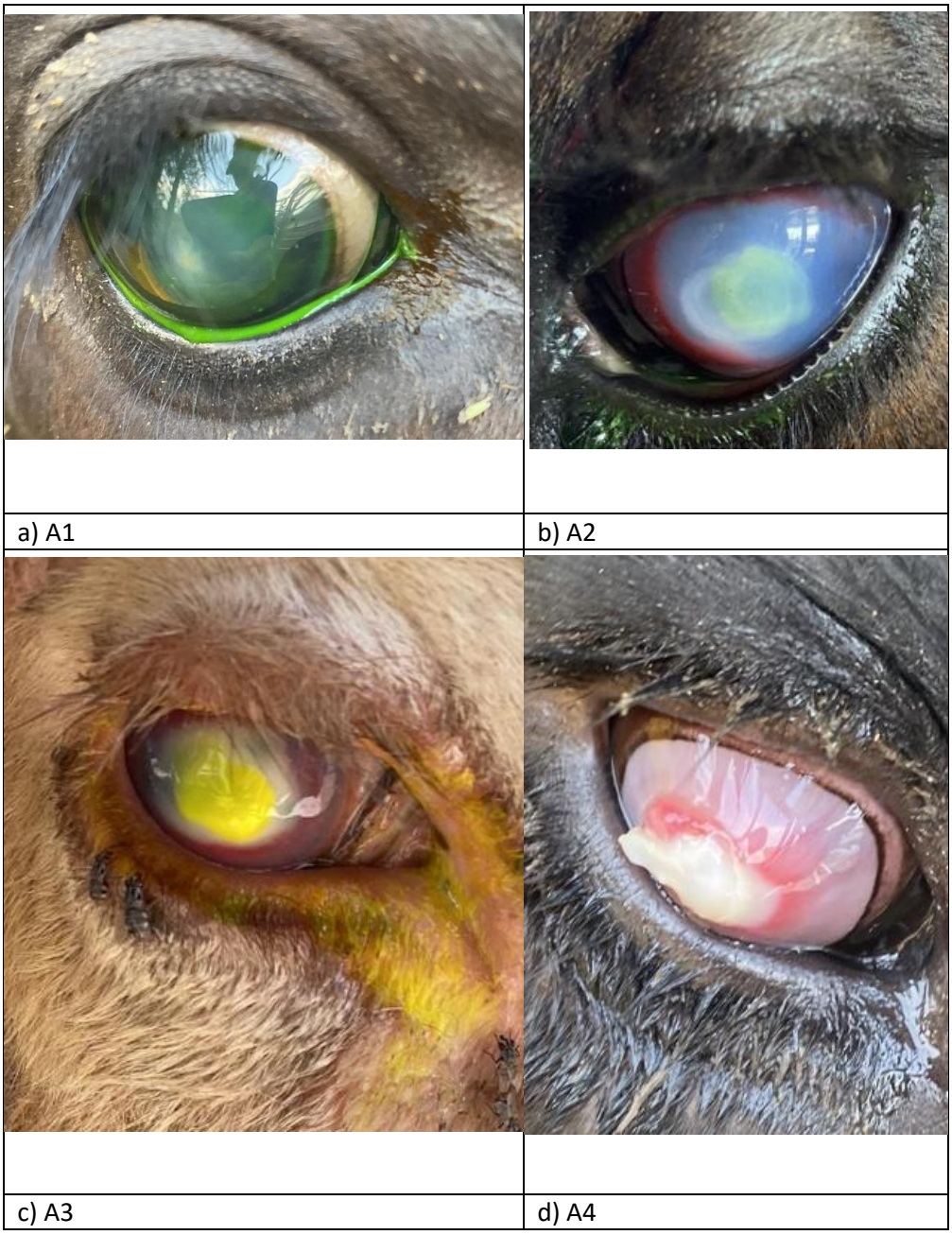
5.2.2.1.1 Normal

Images classified as "Normal" displayed no symptoms of pinkeye, such as ulcers or blood vessel formation. Tears were generally disregarded as a symptom of pinkeye unless they were pus-filled, as serous tears can result from non-specific irritations. This category represented the baseline for comparison against the other stages of pinkeye (Figure 5.1).

Figure 5.1. An example of an eye classified to be normal

5.2.2.1.2 Active

The "Active" category included images classified into four severity levels (A1 to A4). Severity was determined based on ulcer size, the formation of blood vessels, and the colour and texture of the ulcer. White ulcers were classified as A2, while yellow ulcers indicated more advanced severity (A3). A4 represented the most severe cases, characterised by significant ulceration (>75% of the eye), darkened ulcers, malformed or ruptured eyes, or complete eye loss (Figure 5.2).



a) A1 | b) A2

c) A3 | d) A4

Figure 5.2. Example images of a) A1, b) A2, c) A3 and d) A4 eyes classified in the images

5.2.2.1.3 Resolving

The "Resolving" category comprised images exhibiting signs of healing following an active infection. Key features of resolving eyes included tree-like blood vessels approaching the ulcerated area, suggesting healing, and the absence of blood vessel hedges, which differentiated this category from active infection. Resolving eyes were further categorised by ulcer size and/or eye malformation:

- Level 1: Ulcers <10% of the eye
- Level 2: Ulcers between 11–26% of the eye
- Level 3: Ulcers between 26–75% of the eye
- Level 4: Ulcers >75% of the eye

Additional criteria for classification included the presence of blood vessels across the lesion, indicating ongoing healing, and blood vessels clearing from the limbus, which also signified resolution. Eyes in the almost fully recovered stage were classified as Level 1, where the ulcer was either extremely small or no longer visible (Figure 5.3).

a) R1

b) R2

c) R3

d) R4

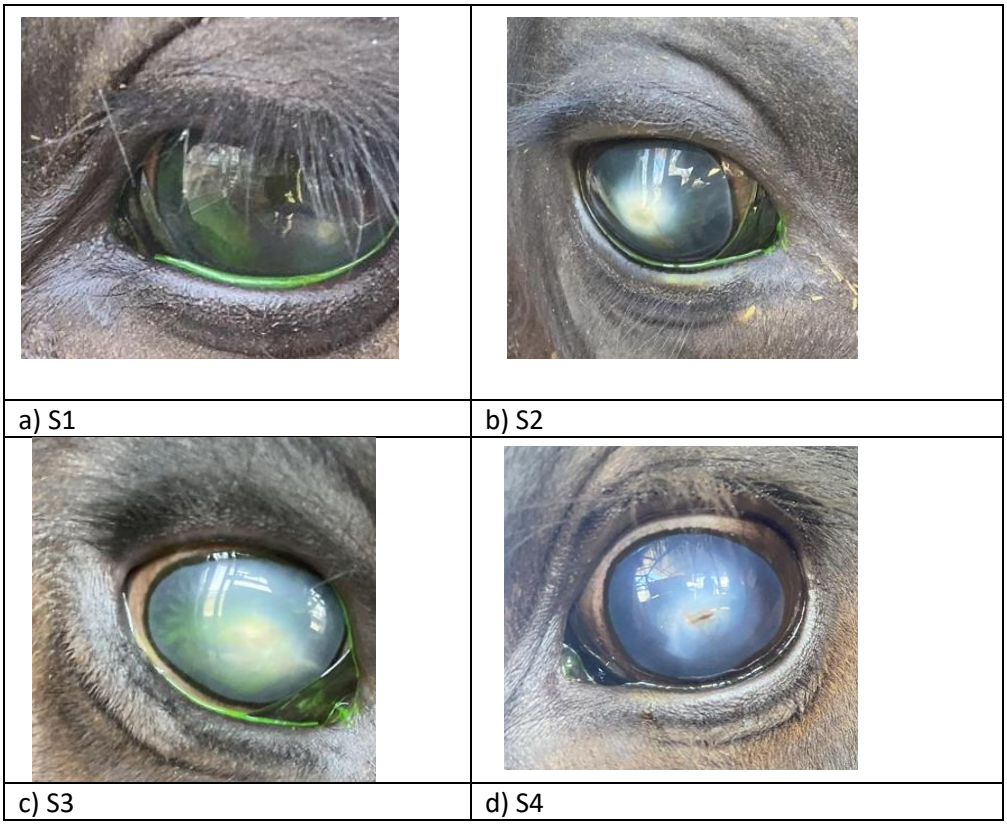
Figure 5.3. Example images of a) R1, b) R2, c) R3 and d) R4 classes of Resolving stage of cattle eyes.

5.2.2.1.4 Resolved

The "Resolved" category included images where pinkeye had healed, leaving dull white scars at the site of previous ulcers. Severity of scarring was determined by the size of the scar relative to the eye:

- Level 1: Scars affecting <25% of the eye
- Level 2: Scars affecting 26–50% of the eye
- Level 3: Scars affecting >50% of the eye
- Level 4: Presence of miscolouring within the ulcer or extensive scarring that significantly damaged the corneal surface

In the most severe cases, scarring resulted in the loss of the eye or substantial damage to the corneal surface. This level of severity was considered beyond treatment and contributed to the overall devaluation of the animal (Figure 5.4).



a) S1 b) S2

c) S3 d) S4

Figure 5.4. Example images of a) S1, b) S2, c) S3 and d) S4 eyes classified in the images

5.2.2 Data preprocessing and modelling approaches

5.2.2.1 Binary classification approach

The dataset's 13 total classes (Normal, Active, Resolving, and Resolved, with respective severity levels) were further regrouped into two classes to simplify analysis and improve practical applicability. For instance, combinations of classes were trialled for the DL model to classify eyes in the active category versus those that are not in the active category requiring immediate attention and treatment. This restructuring aligned with the objective of developing models capable of assisting farmers in early detection and alerting the veterinary specialists, thereby minimising delays in treatment. The composition of these groupings is presented in Table 5.2.

Table 5.2. The combination of the active category for analysis to determine whether the eye requires treatment.

Combination	Require treatment (sample size)	Do not require treatment (sample size)
A	A1 & A2 (201)	All other categories (3,100)
B	A1, A2 & A3 (456)	All other categories (2,845)
C	A1-A4 (498)	All other categories (2,803)

The "Active" category is the primary focus of this analysis, as all eyes in this stage require treatment. In contrast, the "Resolving" and "Resolved" categories represent stages where the condition is already healing or has fully healed, making timely intervention less critical or treatment potentially ineffective for cases that are resolved or scarred. Within the "Active" category, the A1 severity level is significantly imbalanced, with only 61 images out of a total of 3,301, highlighting the rarity of such cases being observed and captured in real-world conditions. This imbalance presents a challenge, as it can disproportionately skew the results, as demonstrated in Chapter 4. Therefore, A1 was not isolated against all other categories to ensure a more robust and overall meaningful evaluation.

From the previous chapter, EfficientNet was the best-performing model and thus was selected to analyse these binary classification combinations. The

hyperparameters for the best performing model include ReLU activation, learning rate at 0.01, batch size at 32, no. of epochs at 100, dropout regularisation at 0.5, L2 regularisation at 0.01, sigmoid activation function, the Adam optimiser with binary cross-entropy. Data augmentation was also implemented with these settings: flip up/down: 0.7, flip left/right: 0.5, mosaic: 0.3 and image rotation: 90.0. The key evaluation metrics used to provide a comprehensive assessment of the performance of this model include accuracy, AUC and sensitivity, specificity and F1 score. A confusion matrix was used to aid in the visualisation of this classification performance.

Stratified data splitting was employed across all analysis combinations, with 70% of the data allocated for training, 10% for validation, and 20% for testing to ensure a balanced representation (Table 5.3)

Table 5.3. Distribution of images in the training, validation, and testing sets for binary analysis included in this chapter.

Combination	Categories	Training	Validation	Testing	Total
A	A1 & A2	140	20	41	201
	All others	2,170	310	620	3,100
B	A1, A2 & A3	319	46	91	456
	All others	1,992	285	569	2,845
C	Active	348	50	100	498
	All others	1,962	280	561	2,803

5.2.2.2 Multiclass approach

The multiclass classification approach involved treating each stage as a distinct category, resulting in four classes: Active, Resolving, Resolved, and Normal. Since this was a multiclass classification task, the EfficientNetV2B2 model was adapted to accommodate the expanded number of categories.

The output layer was configured to use a softmax activation function, enabling the model to assign probabilities across all four classes rather than producing a binary output. The use of softmax ensured that the predicted

probabilities for all classes summed to one, which is appropriate for multiclass classification tasks.

The loss function was changed to categorical cross-entropy, which is specifically designed for multiclass classification tasks. This loss function measured the divergence between the predicted probability distribution and the true distribution across the four classes.

Performance evaluation metrics were adapted to reflect the multi-class nature of the problem. Weighted versions of Accuracy, AUC, sensitivity, specificity, and F1 Score were employed to provide a more comprehensive assessment of model performance which reflects the data distribution. Additionally, to assess the reliability of these performance metrics, 95% confidence intervals were estimated using bootstrapping with 500 resampling iterations, providing a measure of uncertainty around each result of the testing sets. Furthermore, a confusion matrix was generated to assist in visualisation and interpretation of the classification results across the four categories.

The dataset was divided into training, validation, and testing sets using a 70:10:20 split, with the exact proportions of images for each class in the training, validation, and testing sets presented in Table 5.4 below.

Table 5.4. Distribution of images in the training, validation, and testing sets for the multiclass analysis included in this chapter.

Categories	Training	Validation	Testing	Total
Active	348	50	100	498
Resolving	383	55	109	547
Resolved	523	75	150	748

Normal	1,055	151	302	1,508

5.2.2.3 Ordinal analysis

The ordinal analysis in this study was performed to assess severity levels within the Active, Resolving, and Resolved categories after the initial classification stage. This process was structured as a two-stage model, where:

- Stage 1: The EfficientNetV2B2 model was utilised to classify images into one of the four categories: Normal, Active, Resolving, and Resolved. The primary objective of this stage was to accurately identify images belonging to the Active, Resolving, and Resolved categories. Performance for this stage was evaluated using metrics such as accuracy and F1-score to measure the accuracy of the classification task as described in Section 2.3.2.
- Stage 2: After images were classified into their respective categories, a separate ordinal analysis was conducted within each of the Active, Resolving, and Resolved categories. This process involved assessing severity levels based on predefined ordinal scales unique to each category:

The EfficientNetV2B2 model's architecture was modified to accommodate ordinal regression by replacing the softmax activation function in the output layer with a sigmoid activation function. This modification enabled the model to predict cumulative probabilities, which were then interpreted to estimate the appropriate severity level. Additionally, the loss function was adjusted to reflect the ordinal nature of the severity levels, specifically, the ordinal loss function was utilised.

Performance evaluation for the ordinal analysis was conducted using metrics tailored for ordinal variables, including Mean Absolute Error (MAE), Cohen's weighted kappa, and Accuracy (Sakai 2021).

The dataset was divided into training, validation, and testing sets using a 70:10:20 split, respectively. This split was applied uniformly across all four

categories: Normal, Active, Resolving, and Resolved, ensuring that each subset contained representative proportions of images from each category.

5.3. Results

5.3.1 Descriptive results

Following the application of the scorecard criteria, the dataset was divided into four categories: Normal, Active, Resolving, and Resolved. The categorisation process resulted in varying proportions of images across these groups within the dataset after the 70:10:20 preprocessing split (Table 5.8).

Table 5.8. Distribution of images in the training, validation, and testing sets for each category included in this chapter.

Attribute	Categories	Training	Validation	Testing	Total
Active	A1	43	6	12	61
	A2	98	14	28	140
	A3	179	26	51	255
	A4	30	4	8	42
Resolving	R1	132	19	38	189
	R2	98	14	28	140
	R3	115	16	33	164
	R4	38	5	11	54
Resolved	S1	391	56	111	558
	S2	93	13	27	133
	S3	27	4	8	39
	S4	13	2	3	18

5.3.2 Binary classification

5.3.2.1 A1 & A2 vs. all other categories

The binary classification model achieved high accuracy across all datasets, with 94.46% on the training set, 93.64% on the validation set, and 93.80% on the test set (Table 5.9). The AUC values were similarly strong, indicating robust discriminative performance: 0.92 (training), 0.91 (validation), and 0.82 (test). However, the F1 scores were consistently low, with 0.18, 0.09, and 0.09 for training, validation and test, respectively, highlighting challenges in balancing precision and recall, particularly for the “A1 & A2” minority class.

Table 5.9. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for “A1 & A2” vs. “all other categories” using EfficientNetV2B2.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9446	-	0.1795	-	0.9231	-
Validation	0.9364	-	0.0870	-	0.9105	-
Test	0.9380	(0.9228, 0.9532)	0.0889	(0.0765, 0.1013)	0.8235	(0.7966, 0.8504)

The confusion matrix (Figure 5.5) highlights the model's predictions for the binary classification task, showing that out of 661 samples classified as "all other categories," 618 were correctly identified (true negatives), while three were misclassified as "A1 & A2" (false positives). For the "A1 & A2" category, only two samples were correctly classified (true positives), whereas 38 were misclassified as " all other categories " (false negatives).

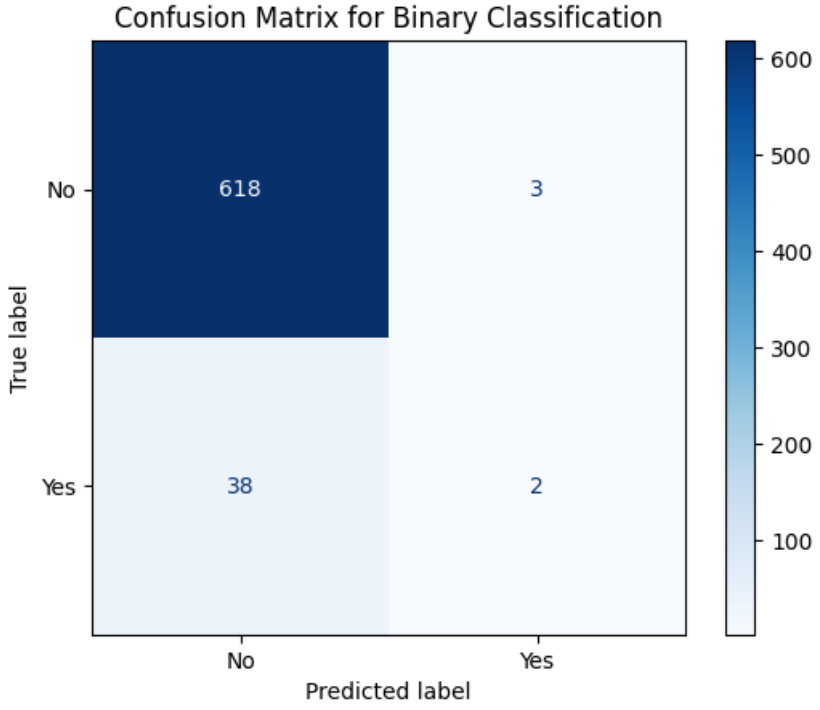


Figure 5.5 Confusion matrix for the binary classification of pinkeye stage, where “Yes” denotes A1 & A2 and “No” denotes all other categories. Of the 661 “No” samples, 618 were correctly classified (true negatives) and 3 were misclassified as “Yes” (false positives). Among the 40 “Yes” samples, only 2

were correctly identified (true positives), while 38 were misclassified as “No” (false negatives).

5.3.2.2 A1, A2 & A3 vs. all other categories

The binary classification model achieved good accuracy across all datasets, with 90.87% on the training set, 89.39% on the validation set, and 89.41% on the test set (Table 5.10). The AUC values were also promosing, demonstrating reliable discriminative performance: 0.93 (training), 0.88 (validation), and 0.92 (test). The F1 scores, while moderate, were noticeably lower than accuracy, with 0.62, 0.56, and 0.53 for the training, validation, and test sets, respectively, reflecting challenges in maintaining a balance between precision and recall for the “A1, A2 & A3” minority class.

Table 5.10. Accuracy, F1 score, and AUC metrics for the training, validation, and testing sets for “A1, A2 & A3” vs. “all other categories” using EfficientNetV2B2.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9087	-	0.6157	-	0.9314	-
Validation	0.8939	-	0.5570	-	0.8803	-
Test	0.8941	(0.8579, 0.9303)	0.5270	(0.4977, 0.5563)	0.9218	(0.8959, 0.9477)

The confusion matrix (Figure 5.6) reveals the model's classification performance in greater detail. For the "all other categories " class, 552 samples were correctly classified (true negatives), while 18 were misclassified as " A1, A2 & A3" (false positives). For the " A1, A2 & A3" class, 39 samples were correctly identified (true positives), but 52 were misclassified as "all other categories" (false negatives).

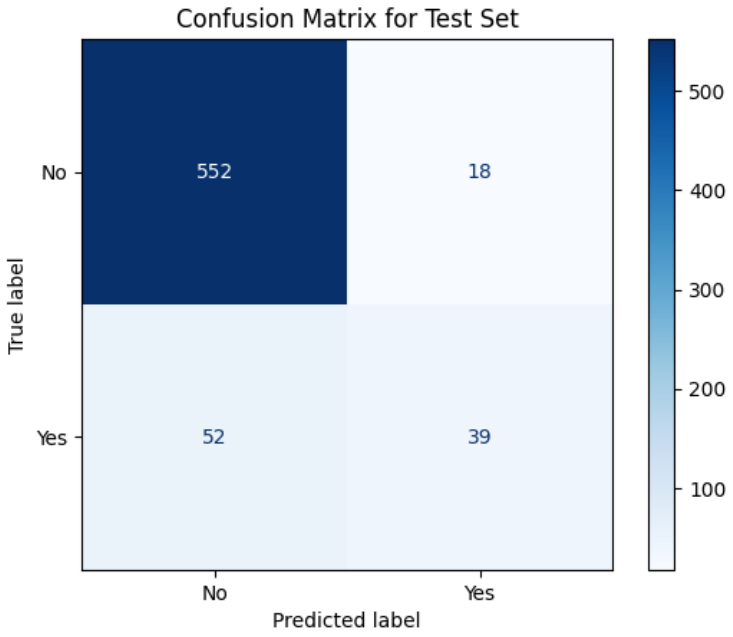


Figure 5.6. Confusion matrix for the binary classification of pinkeye stage, where “Yes” refers to A1, A2 & A3 and “No” refers to all other categories. Of the 570 “No” samples, 552 were correctly classified (true negatives) and 18 were misclassified as “Yes” (false positives). Among the 91 “Yes” samples, 39 were correctly identified (true positives), while 52 were misclassified as “No” (false negatives).

5.3.1.3 Active vs. all other categories

The binary classification model demonstrated good accuracy across all datasets, with 93.07% on the training set, 90.61% on the validation set, and 90.62% on the test set (Table 5.11). The AUC values were similarly robust, indicating some discriminative capability: 0.95 (training), 0.89 (validation), and 0.89 (testing). The F1 scores showed notable improvement compared to previous results, with 0.75 for training, 0.67 for validation, and 0.66 for testing, suggesting the model achieved a better balance between precision and recall, especially for the minority "Active" class.

Table 5.11. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for “Active” vs. “all other categories”.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.9307	-	0.7531	-	0.9524	-
Validation	0.9061	-	0.6737	-	0.8936	-

Test	0.9062	(0.8579, 0.9545)	0.6667	(0.6081, 0.7253)	0.8978	(0.8847, 0.9109)

The confusion matrix (Figure 5.7) further illustrates these improvements. For the "No" class, 537 samples were correctly classified (true negatives), while 24 were misclassified as "Active" (false positives). For the " Active " class, 62 samples were correctly identified (true positives), and 38 were misclassified as "all other categories " (false negatives).

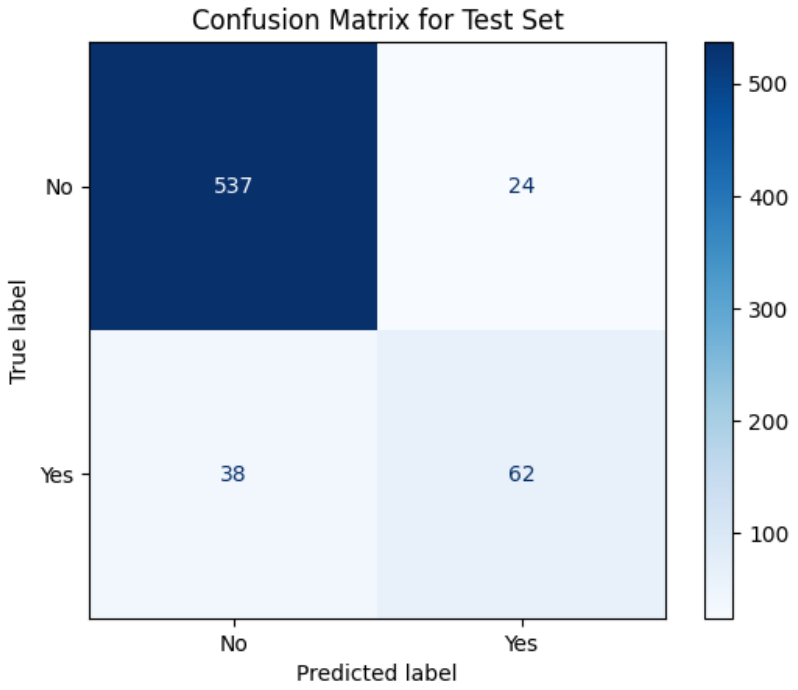


Figure 5.7. Confusion matrix for the binary classification of pinkeye stage, where “Yes” refers to the Active class and “No” refers to all other categories. Of the 561 “No” samples, 537 were correctly classified (true negatives) and 24 were misclassified as “Yes” (false positives). Among the 100 “Yes” samples, 62 were correctly identified (true positives), while 38 were misclassified as “No” (false negatives).

5.3.3 Multiclass classification

The multiclass classification model achieved moderate accuracy across the datasets, with 73% on the training set, 65% on the validation set, and 69% on the test set. The F1 scores were closely aligned with the accuracy, at 0.72

(training), 0.65 (validation), and 0.67 (test), indicating consistent performance in balancing precision and recall across the multiple classes. The AUC values were strong, with 0.91 (training), 0.86 (validation), and 0.87 (test), suggesting the model demonstrated robust discriminative capability in distinguishing among the classes (Table 5.12).

Table 5.12. Accuracy, F1 score and AUC metrics for the training, validation and testing sets for Active, Resolving, Resolved and Normal.

	Accuracy	95% CI	F1	95% CI	AUC	95% CI
Train	0.73	-	0.72	-	0.9056	-
Validation	0.65	-	0.65	-	0.8592	-
Test	0.69	(0.678, 0.702)	0.67	(0.656, 0.684)	0.8707	(0.8487, 0.8927)

The confusion matrix (Figure 5.8) reveals patterns in the model's performance for the multiclass classification task. The "Normal" class was classified with the highest accuracy, as evidenced by the 275 true positives and relatively few misclassifications (1 as "A," 14 as "R," and 12 as "S"). This suggests the model performs best at identifying this majority class.

Conversely, the "S" class exhibited a high number of false negatives, with 64 misclassified as "Normal" and 19 as "R." Similarly, the "R" class had significant misclassifications, with 29 samples labelled as "Normal" and 15 as "A." These patterns highlight challenges in distinguishing among minority or overlapping classes, particularly "R" and "S," which often share features with neighbouring categories.

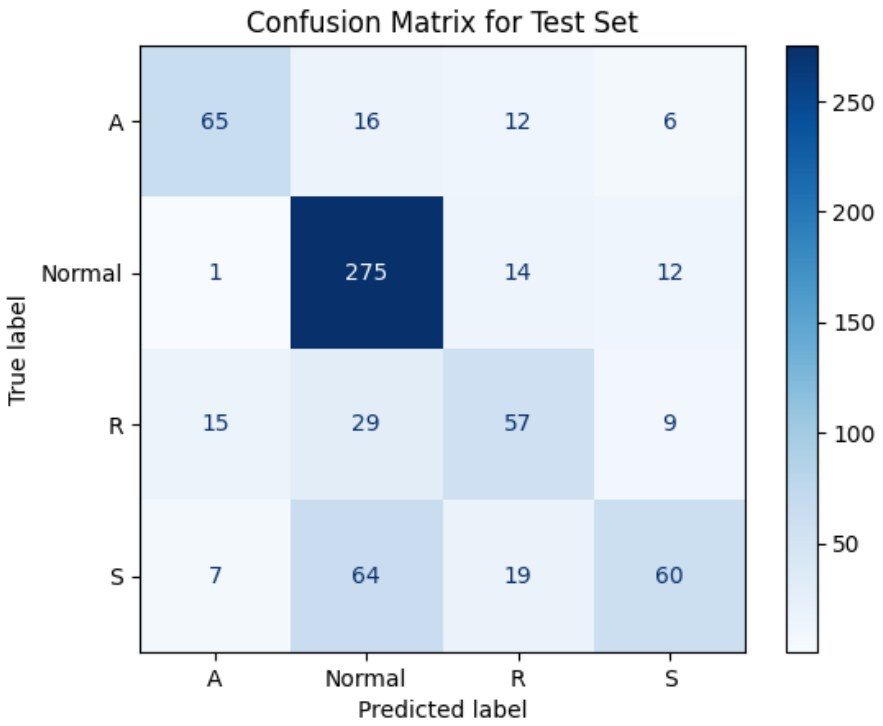


Figure 5.8. Confusion matrix for the multiclass classification of pinkeye stages: A (Active), Normal, R (Resolving), and S (Resolved). For the “A” class, 65 samples were correctly classified, with 16 misclassified as “Normal,” 12 as “R,” and 6 as “S.” The “Normal” class had 275 correct classifications, with 1 misclassified as “A,” 14 as “R,” and 12 as “S.” For the “R” class, 57 samples were correctly classified, while 15 were misclassified as “A,” 29 as “Normal,” and 9 as “S.” The “S” class had 60 correct classifications, with 7 misclassified as “A,” 64 as “Normal,” and 19 as “R.”

5.3.4 Ordinal variables

5.3.4.1 Active

The ordinal classification model for predicting the active classes demonstrated moderate performance across the dataset. The accuracy was 90.80% on the training set, 72.00% on the validation set, and 72.00% on the test set. The MAE values were 0.13 (training), 0.30 (validation), and 0.30 (test), indicating minimal error in predicting ordinal categories. The Cohen's kappa scores were 0.89 for training, 0.62 for validation, and 0.61 for testing, reflecting moderate to strong agreement between predicted and actual ordinal classes (Table 5.13).

Table 5.13. Accuracy, MAE and Cohen's kappa metrics for the training, validation and testing sets for Active using EfficientNetV2B2.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.9080	-	0.1322	-	0.8957	-
Validation	0.7200	-	0.3000	-	0.6229	-
Test	0.7200	(0.704, 0.736)	0.3000	(0.268, 0.332)	0.6147	(0.5786, 0.6508)

5.3.4.2 Resolving

The ordinal classification model for the resolving classes demonstrated varying performance across the dataset. The training set showed strong results, with an accuracy of 81.94%, an MAE of 0.21, and a Cohen's kappa of 0.81, indicating good agreement between predicted and actual classes. However, performance dropped on the validation set, with an accuracy of 52.73%, an MAE of 0.5636, and a kappa score of 0.48, reflecting moderate agreement. On the test set, the model achieved an accuracy of 60.00%, an MAE of 0.46, and a kappa score of 0.59, suggesting improved performance over validation but still showing room for enhancement (Table 5.14).

Table 5.14. Accuracy, MAE and Cohen's Kappa metrics for severity within the Resolving class for the training, validation and testing sets using EfficientNetV2B2.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.8194	-	0.2094	-	0.8144	-
Validation	0.5273	-	0.5636	-	0.4824	-
Test	0.6000	(0.5689, 0.6311)	0.4636	(0.3893, 0.5379)	0.5898	(0.5481, 0.6315)

5.3.4.3 Resolved

The ordinal classification model for the Resolved variable demonstrated overall moderate performance. The training set achieved an accuracy of 79.54% and an MAE of 0.2467, with a Cohen's kappa of 0.38, indicating fair

agreement. Validation and test performances were similar, with an accuracy of 73.33% on both datasets. The MAE was 0.35 for validation and 0.30 for the test set, showing higher prediction errors compared to the training set. Notably, Cohen's Kappa score on the validation set was 0.00, suggesting no meaningful agreement, while the test kappa score improved to 0.30, indicating slight agreement (Table 5.15).

Table 5.15. Accuracy, MAE and Cohen's kappa metrics for the training, validation and testing sets for Resolved using EfficientNetV2B2.

	Accuracy	95% CI	MAE	95% CI	Kappa	95% CI
Train	0.7954	-	0.2467	-	0.3837	-
Validation	0.7333	-	0.3467	-	0.0000	-
Test	0.7333	(0.7043, 0.7623)	0.3000	(0.2523, 0.3477)	0.3005	(0.2148, 0.3862)

5.4. Discussion

This study was conducted to evaluate the effectiveness of deep learning for classifying pinkeye stages and severity in cattle using real-world, phone-captured images. The key finding was that binary classification of active versus non-active stages performed strongly, whereas multiclass and ordinal approaches revealed significant limitations due to class imbalance and overlapping visual features. This is one of the first studies to apply a multi-framework (binary, multiclass, and ordinal) deep learning pipeline specifically for bovine pinkeye, highlighting how explainable AI and ordinal evaluation metrics can reveal performance gaps often hidden by accuracy alone.

The classification frameworks employed in this study encompass binary, multi-class, and ordinal classification approaches, each tailored to address the specific requirements of pinkeye stage classification and severity assessment. The classification tasks were carried out using EfficientNetV2B2, a pretrained model with optimised weights for recognising a wide range of image features, which was adapted for this study by modifying only the final layers for the classification task. This transfer learning approach allowed the model to retain its ability to identify general features while learning specific

characteristics of pinkeye in cattle, such as corneal changes or inflammation. By leveraging pretrained weights, this method reduced computational requirements and training time, making it particularly suitable for medical imaging tasks (Shah, Saeed et al. 2022), where datasets are often small due to privacy concerns and the need for expert annotations. It overall demonstrated strong performance across multiple frameworks, though with varying degrees of success depending on the classification task and data imbalance.

The binary classification approach was conducted through three combinations aimed at distinguishing between eyes requiring treatment and those that do not. Notably, the combination targeting the broadest spectrum of 'Active' cases (A1-A4 vs. all other categories) achieved the highest F1 scores, particularly when compared against the more restrictive combinations. This improvement highlights the model's enhanced sensitivity and ability to generalise when the minority class is more comprehensive, even though challenges remained due to class imbalance (Johnson and Khoshgoftaar 2019), particularly for the most severe cases such as A1, which had the least presenting symptoms, making it especially difficult to train the model effectively for its detection.

A stepwise pattern was observed across the three binary classification tasks. The model showed the greatest difficulty in detecting the "A1 & A2" class, with a high number of false negatives and a low number of true positives. This was reflected in a low F1 score of 0.0870 on the validation set, highlighting its struggle with class imbalance and limited recall. Including A3 in the minority group ("A1, A2 & A3") improved recall, resulting in a higher true positive count and an F1 score of 0.5270, though a substantial number of false negatives remained. When the entire "Active" group (A1 to A4) was classified against all other categories, the model achieved a further increase in performance, with an F1 score of 0.6667. This final combination also showed fewer false negatives and a higher true positive rate, aligning with

strong AUC and accuracy values, suggesting the model could better balance precision and recall when the minority class was broader.

The model achieved strong accuracy across all binary classification tasks, exceeding 89% in each comparison: 93.80% for A1 & A2 vs. all others, 89.41% for A1, A2 & A3 vs. all others, and 90.62% for Active vs. all others. However, the corresponding F1 scores exposed limitations, particularly in the early-stage classification (A1 & A2 vs. all others). This discrepancy between high accuracy and low F1 score suggests that the model predominantly classified the majority category ("all others") while struggling with low recall, as reflected by the confusion matrix where only 2 out of 40 testing images were correctly identified as A1 & A2 for the A1 & A2 vs. all others subset analysis. These results reveal the model's difficulty recognising subtle features of early pinkeye stages. Nonetheless, the inclusion of additional categories (A3 and A4) improved performance. This trend suggests that the model's recall performance can be enhanced with larger datasets, particularly for underrepresented categories.

The multiclass classification approach, which sought to differentiate among all four stages (Normal, Active, Resolving, and Resolved), demonstrated moderate performance with accuracies of 73%, 65%, and 69% for the training, validation, and test sets, respectively. Despite reasonable accuracy, the model struggled to distinguish between classes that share similar visual characteristics, particularly the Resolving and Resolved categories, which likely contributed to the observed misclassifications. The confusion matrix highlighted a tendency for the model to favour majority classes, further emphasising the impact of class imbalance (Ghosh, Bellinger et al. 2024).

The model's overall performance remained moderate, with F1 scores of 0.72, 0.65, and 0.67 for the training, validation, and test datasets, suggesting a consistent yet limited balance between precision and recall. This level of efficacy is echoed in research on human eye disease classification, where multiclass classification yielded similarly moderate results (Guergueb and Akhloufi 2021). Although the accuracy metrics were notably lower than those

achieved in binary classification tasks, F1 statistics remained at a comparable level. This discrepancy in accuracy, but not in F1 score, highlights the model's capability to identify the majority class in the binary scenario. In contrast, the supposed increased complexity of distinguishing between multiple classes, especially when symptoms overlap, did not limit or enhance the model's sensitivity measurement. Robust AUC values (0.9056, 0.8592, and 0.8707) suggest reliable discriminative ability; however, the confusion matrix revealed persistent challenges in identifying minority classes such as "S" and "R." For instance, the "S" class exhibited 64 false negatives, primarily misclassified as "Normal." This issue may be less critical in practice since resolved eyes do not require treatment, but it reflects the broader problem of class imbalance.

The "Normal" category was overrepresented, while critical early-stage categories such as A1 and A2 were significantly underrepresented. Of the 3,301 total images, only 201 belonged to A1 and A2. Diagnosing A2 and A3 is comparatively easier due to distinct features like hedge formations around clearly visible ulcers, whereas A1, which represents the early onset of pinkeye, presents subtle symptoms that are difficult to detect. Early detection of A1 is essential for timely treatment, making accurate classification of this stage pivotal. While data augmentation was applied to address the imbalance, its effectiveness was limited; artificially inflating underrepresented categories often led to overfitting, as replicated patterns lacked sufficient diversity (Huang, Schmelter et al. 2023). This limitation is reflected in the lower F1 scores, indicating ongoing challenges in balancing precision and recall.

From Chapter 3, it was determined that an initial dataset containing at least 200 images per class provides sufficient diversity for effective augmentation. This benchmark suggests that achieving a minimum of 200 images per class for each pinkeye category is necessary for improving classification performance. Since the sample sizes for some classes, such as A1, A4, R2, R4, S3, and S4, fall below this threshold, the effectiveness of data augmentation was constrained by the limited availability of minority class samples. Class

weights were applied during multiclass modelling to mitigate class imbalance (Krawczyk 2016), resulting in the use of weighted accuracy, F1 scores, sensitivity, specificity, and AUC during evaluation. However, despite applying these weighted metrics, the model's performance remained moderate, suggesting that class weighting was insufficient to address the enormous imbalance problem observed in the dataset.

Future efforts should prioritise re-sampling techniques and collecting additional real-world samples, particularly for early-stage categories like A1. Engaging veterinarians in image collection, rather than relying solely on farmers, could enhance dataset quality by capturing subtle symptoms that may otherwise be overlooked. As the dataset grows through broader community usage, the imbalance in class representation may gradually diminish, enhancing the model’s capacity to detect nuanced patterns.

In comparison, binary classification tasks, such as "Active vs. all others," simplified the problem by focusing on broader distinctions and clustering to enlarge grouping sample sizes, resulting in higher accuracy and F1 scores overall. However, this approach sacrificed specificity by including cases that were already apparent to farmers. While multi-class classification aligns more closely with the objective of distinguishing subtle early-stage symptoms like those in A1 and A2, achieving this goal requires addressing dataset imbalance for minority classes to train the model to generalise better. Techniques such as re-sampling to obtain more samples for the minority classes, class weighting mechanisms, improved augmentation strategies, and incorporating attention mechanisms could enhance the model’s ability to capture critical features across multiple classes. Attention mechanisms, however, represent a distinct modelling process as they involve modifying the network architecture to selectively focus on relevant image regions (Gonçalves, Rio-Torto et al. 2022). This approach can improve performance by enhancing feature representation for minority classes, particularly those with subtle or overlapping symptoms.

The ordinal classification approach aimed to assess severity grading within the Active, Resolving, and Resolved categories. Of these, the Active category demonstrated the most robust performance, achieving an accuracy of 90.80% and a high Cohen's kappa score of 0.89 during training, suggesting strong agreement between predicted and actual severity levels. However, this performance decreased significantly on the validation and test sets, with accuracies of 72% and kappa scores of approximately 0.61. The decline in performance across validation and testing indicates that the model was more effective at recognising severity within the training data than generalising to unseen samples. The relatively strong training performance for the Active category may be attributed to clearer, more distinguishable symptoms compared to the Resolving and Resolved stages. However, despite these encouraging results, the drop in performance during testing highlights limitations in the model's generalisability, particularly when faced with subtle or borderline cases within the Active category. The MAE score on the test set reached a maximum of 0.3, which is relatively low for the testing set which shows that the model's predictions are relatively close to the true labels.

In contrast, the Resolving and Resolved categories exhibited weaker ordinal classification performance. For the Resolving category, training accuracy reached 81.94% with a moderate Cohen's kappa score of 0.81, but these metrics sharply declined to 52.73% accuracy and a kappa score of 0.48 on the validation set. Although test accuracy improved to 60.00%, the relatively low kappa score of 0.59 reflects only moderate agreement. The Resolved category demonstrated similarly inconsistent performance, with training accuracy of 79.54% and kappa of 0.38, which dropped to 73.33% accuracy and a kappa of 0.00 during validation. These results suggest that the model struggles particularly with categories like Resolving and Resolved, where symptoms are less visually distinct or present with overlapping features between each of the severity levels. The relatively low Cohen's kappa scores for these categories highlight the model's limited ability to capture ordinal relationships effectively (Yilmaz and Demirhan 2023). This may be due to the complexity of

recognising fine-grained differences between severity levels, particularly for categories that are either progressing toward recovery (Resolving) or already recovered (Resolved). The poorer performance in these categories further underscores the importance of addressing class imbalance and enlarging the sample size to enhance the model's capacity to differentiate between subtle severity gradations. MAE scores were moderate, ranging between 0.3 and 0.5 for both categories, which by comparison shows that the model's predictions here are farther from the true labels compared to distinguishing the severity levels in the active (A) category. This could be attributed to less distinct separation in symptoms between the severity levels found in resolving (R) and recovered (S). Similar moderate level performance have been reported in studies analysing ordinal grading of eye disease progression using other DL techniques (Toledo-Cortés, Useche et al. 2022).

5.4.1 Limitations and Future Directions

These results highlight the strengths of EfficientNetV2B2 in handling complex classification tasks, while also revealing challenges related to class imbalance, overlapping visual characteristics, and the hierarchical nature of disease severity. Despite reasonable performance, the model's difficulty in distinguishing between visually similar classes, particularly those within the Resolving and Resolved categories, underscores the inherent complexity of pinkeye stage classification. Additionally, staining artifacts present in some images may have introduced confounding factors, correlating with class labels and potentially causing the model to learn these artificial visual cues rather than relevant clinical features. Addressing this issue in future work is crucial to improving model generalisability.

As with any diagnostic test, the performance of deep learning classifiers is also subject to potential sources of bias that may influence interpretation and real-world applicability. In this study, some degree of selection bias may have arisen due to the dataset being collected from a limited geographical region and herd population, which may not fully represent the variability in cattle breeds, management environments, or clinical presentations encountered

worldwide. Information bias may also be present because image labels were assigned from expert visual assessment of field-acquired photographs, which inherently involves subjectivity and variation in image quality. These forms of bias are well-recognised in diagnostic epidemiology, and acknowledging their presence reinforces the need for cautious interpretation and future external validation (Thrusfield 2018).

The study also highlights the importance of adapting DL tools for veterinary applications, where challenges differ from human medical imaging. Veterinary imaging presents several unique challenges. Images are often captured in non-clinical settings such as farms or field environments, leading to greater variability in lighting, resolution, angles, and background noise. Unlike human imaging, which benefits from standardised protocols and high-quality clinical equipment, veterinary images are typically taken using mobile devices with inconsistent conditions. Furthermore, annotations are frequently made by non-specialists or without access to confirmed diagnostic outcomes, increasing the risk of label noise and subjective interpretations.

Literature comparisons demonstrate that DL models have achieved remarkable performance in simpler tasks, such as binary classifications of diseased versus healthy eyes in companion animals, with accuracies reaching 99% (Şengöz 2023), which echoes the success of our results in the binary classification of pinkeye vs. normal eyes. This study utilised data augmentation techniques to bolster each category to an effective size of 500 samples, which successfully mitigated class imbalance and enhanced modelling performance. Achieving a sufficiently large and balanced dataset is a crucial factor in DL performance, as demonstrated by studies employing similar strategies. While this approach proved effective for the binary classification tasks in our study, it was less successful for multiclass classification due to the limited number of samples available for certain categories. This limitation is particularly evident in early-stage pinkeye categories such as A1, where even augmented datasets remained insufficient for robust model training. Similar DL techniques were applied to cattle eyes

for cardiovascular disease detection achieved 96% accuracy using ResNet (Cihan, Saygılı et al. 2024). Their study focused on binary classification, which is inherently simpler than the multi-class and ordinal classification tasks undertaken in this study. Binary classification models benefit from clearer distinctions between categories, making them easier to train and optimise. Moreover, their study utilised high-quality, high-resolution retinal fundus images, which are far superior to the phone-captured images used in this study. The latter are more susceptible to external factors and artifacts, such as inconsistent lighting and background noise, which can obstruct the learning process. The simpler binary classification objective also allowed their models to be trained on more balanced datasets, further enhancing performance. For example, Cho, Hwang et al. (2021) reported an accuracy of 85.2% in classifying two stages of glaucoma versus healthy eyes using a dataset of 3,460 high-quality images.. The combination of well-defined classification objectives, superior image quality, and balanced datasets significantly contributed to the higher performance observed in these studies which could be improved upon to address the low performance seen in the challenging multiclass and ordinal classification tasks. In contrast, our dataset, while extensive, remains imbalanced for classes within A, R and S. Moreover, the inherent overlap between stages in pinkeye complicates the classification task, as distinguishing between stages such as "Active" and "Resolving" often requires detecting subtle differences.

In addition, the model's moderate level performance in distinguishing the severity levels within each pinkeye stage reflects the broader scope and challenge of applying DL models to ordinal grading tasks, as most eye disease studies focus on binary classification rather than ordinal scales. Performance in ordinal grading is often moderate at best; for instance, even a study with over 100,000 cleaner image samples utilising ensemble DL models achieved an overall accuracy of only 63.3%, while healthy fundus images were classified with a much higher accuracy of 94.3% demonstrating that

distinguishing healthy images is considerably easier than grading disease severity (Grassmann, Mengelkamp et al. 2018).

Furthermore, the presence of staining artifacts in some images may have introduced confounding factors that correlate with class labels. This issue potentially causes the model to learn these artificial visual cues rather than the relevant clinical features, compromising generalisability and leading to inaccurate predictions. Addressing this limitation is essential for improving model robustness and ensuring that predictions are based on true clinical features rather than unrelated artifacts. Future work should focus on enhancing data quality, potentially through improved preprocessing techniques or by excluding stained images from the training set.

Addressing these challenges requires more advanced model architectures and data handling techniques. Transformer-based models, known for their attention mechanisms, can better capture subtle differences by focusing on the most relevant image regions as a study has shown by utilising this method, the F1 scores of multiclass eye disease classification problems have improved from traditional DL techniques (Gummadi and Ghosh 2023). Additionally, ensemble approaches combining multiple model types could enhance robustness, while graph neural networks might be useful for modelling spatial relationships within lesions. Contrastive learning techniques, which improve feature extraction by leveraging self-supervised methods, also represent a promising avenue, particularly for addressing class imbalance and enhancing minority class representation. Integrating these advanced techniques into the classification pipeline could improve performance, especially when dealing with visually similar or overlapping classes.

Another limitation encountered relates to the application of the scorecard during the annotation process, which was performed solely by a single researcher (myself) without formal veterinary training. Although veterinarians provided guidance during scorecard development, this discrepancy introduces potential bias, particularly when subtle symptoms such as those

associated with early-stage A1 pinkeye are more readily identified by experienced clinicians. The absence of veterinary experts during the annotation process increases the likelihood of inaccurately labelled or overlooked cases, which could propagate errors throughout model training and evaluation. Additionally, the scorecard itself may lack the granularity required for distinguishing closely related categories, such as Active versus Resolving, particularly when slight variations in severity are visually subtle. To enhance the reliability of the scorecard's application, future studies should incorporate expert review during annotation and consider implementing consensus-based approaches to minimise bias. Regular updates to the scorecard, informed by practitioner feedback and ongoing model performance evaluations, could further enhance its utility and accuracy. As DL models become more prominent in veterinary diagnostics, ensuring high-quality annotations through improved scorecard implementation will be essential for achieving robust and clinically useful results.

While the model developed in this study enhances the detection of pinkeye stages, it should be interpreted cautiously, with veterinary consultation remaining essential. The tool is designed as a supplementary resource, particularly useful in settings with limited veterinary access. Future work should prioritise increasing the diversity of the dataset, particularly by engaging veterinarians in data collection to ensure the capture of subtle symptoms that might otherwise be overlooked. As the dataset grows through broader community engagement, the imbalance in class representation may diminish, enhancing the model's ability to detect nuanced patterns across all stages of pinkeye.

5.5 Conclusion

In this study, we trained a model to classify different pinkeye stages in cattle using EfficientNetV2B2, a transfer learning approach, leveraging its pretrained weights to reduce computational demands and improve feature detection.

The model showed good performance in binary classification tasks, achieving accuracies between 89-93% across all different combinations of comparisons, but struggled with class imbalance and low recall for early-stage categories like A1 and A2, as demonstrated by the relatively low to moderate F1 scores across the board. Multiclass classification, such as resolving stages into R1–R4, presented additional challenges, with moderate accuracy (73.33%) and Cohen's kappa (0.3005) on the test set. Misclassification trends in the confusion matrix highlighted the difficulty in distinguishing adjacent stages between R2 and R3, particularly in underrepresented categories, underscoring the limitations of dataset imbalance and feature overlap. Addressing these challenges requires increasing dataset diversity by obtaining more photos, particularly for early pinkeye stages like A1, where timely detection is critical. It was also observed that data augmentation techniques, while useful for addressing class imbalances in theory, had limited impact on model performance due to the insufficient size of the initial sample. In addition, the ground truth annotation sheet can also be shared with all farmers and/or community members as an educational tool to recognise these early symptomatic features. Future efforts should focus on collecting real-world samples through community and veterinary collaboration to utilise the full potential of advanced DL modelling strategies. The study demonstrates the potential of DL for veterinary diagnostics yet highlights the complexity of distinguishing nuanced disease stages compared to broader binary classifications. With continued improvements in dataset quality and model optimisation, DL could serve as a valuable supplementary tool in veterinary ophthalmology.

Chapter 6. Explainable Artificial Intelligence (X-AI) of the DL models developed

6.1. Introduction

Artificial Intelligence (AI) has rapidly advanced in its applications, particularly in image analysis, with significant implications for various fields, including medicine. These advances have led to the development of automated medical image classification and retrieval techniques, which have gained immense significance in clinical and biomedical research. However, the use of AI in veterinary medicine has not kept pace with its applications in human medicine. While recent literature discusses various effective and promising deep learning (DL) models for image classification in veterinary medicine, few have attempted to visualise, understand, or interpret the internal representations and prediction results of their models. This lack of understanding can adversely impact people's confidence in the adoption of automated classification tools. This lack of transparency can lead to mistrust, especially in critical fields like veterinary medicine, where misclassifications can have severe consequences on animal health and wellbeing.

In Chapter 5, DL models were utilised to classify the stages and severity of pinkeye infection in cattle eyes. Understanding which features of cattle eye images contribute to these predictions is crucial, as it can boost confidence in the models' outputs and ensure that veterinary practitioners and other users can trust and act upon these insights. Explainable Artificial Intelligence (X-AI) techniques are designed to address these transparency issues by making the inner workings of DL models more understandable. By applying X-AI techniques, we aim to determine the significant features of cattle eye images that contribute to the correct classifications of pinkeye stages and severity levels. This transparency is particularly important for identifying eyes in the "active" category of pinkeye, as these require immediate treatment to

prevent progression to more severe stages and levels of infection. This chapter explores three mainstream X-AI techniques–Grad-CAM++, LIME, and SHAP– to explain and interpret the DL models developed in the previous chapter, bridging the gap between model predictions and their explainability and interpretability. Techniques like Grad-CAM++ and LIME are used to visualise which regions in the input image are important for the model's predictions. By localising and visualising the regions of interest (ROIs) in the images that are most discriminative for the model's predictions, we aim to enhance the interpretability and reliability of these DL models in the classification of the different stages and severity levels of pinkeye. This approach not only helps in understanding the model's decision-making process but also aids in building trust and confidence among veterinary practitioners, ensuring that the models can be effectively used in clinical settings.

6.2. An overview of X-AI techniques

Despite the success of DL models in image data analysis and pattern recognition, these models are often considered to be "black-box" models due to their complex, opaque nature, making it difficult for users to understand how they work (Murdoch, Singh et al. 2019). Black box or lack of full explainability contributes to the lack of trust in DL systems producing diagnostic results (Shaban-Nejad, Michalowski et al. 2021). Explainable AI, also known as X-AI, is a technique that can provide explanations of how models reach their decisions or classifications, thereby breaking down the opaque nature of black box DL systems (Confalonieri, Coba et al. 2021). This is especially important as it assists users in understanding how to maintain, fix, and improve AI systems, as well as in presenting these systems to others for education and applicability across different domains. It also enhances user trust and accountability, aligning with the ethical boundaries of AI system users. In the extensive literature on X-AI in deep learning image analysis, there tend to be a few general approaches to simplify these black box systems into more understandable ones:

1. The global approach, which encompasses holistic methods of looking at DL models, is to show how each part of the model parses and learns down to the filtering layers, neurons that are activated, etc.

2. The priori approach, where X-AI techniques can be integrated into the design of DL models to train the dataset to improve the explainability of the model. For instance, regularisation methods (Shickel, Tighe et al. 2017), like dropout, data augmentation, and weight decay, were used in Chapters 3 and 4 to enhance model interpretability. These techniques simplify the model by emphasising significant features, thereby, making it easier to understand how the model makes predictions (Ali, Abuhmed et al. 2023).

3. The final approach involves the utilisation of X-AI techniques applied after the model has been trained to understand its decision-making process in reaching a particular output class, known as post-hoc analysis. These methods do not alter the original model but analyse its outputs and internal structures to generate explanations (Minh, Wang et al. 2022). Techniques commonly used in post-hoc explainability include Grad-CAM++, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). By offering a way to understand and trust complex models, post-hoc explainability plays a crucial role in fields where transparency and accountability are essential, such as healthcare, finance, and autonomous systems. Model visualisation strategies are more commonly used in the predominant image analysis literature to understand CNN modelling. In this chapter, we first examine how these techniques were applied in the research and then outline how they can be adapted to explain our AI system for classifying pinkeye.

6.2.1 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique that helps visualise which parts of an image are important for a model's prediction. It uses the gradient information flowing into the final

convolutional layer to produce the activation maps, which are pooled and summed to determine the features of these activation maps that are important for the prediction of a particular class (y^c) (Selvaraju, Cogswell et al. 2020). Mathematically, the summation of all the activation maps weighted by their importance is represented as:

$$L_{Grad-CAM} = ReLU(\sum_k a_k^c A^k) \quad (6.1)$$

where,

- $L_{Grad-CAM}$: The output activation map that highlights important regions in the image for a specific class prediction.
- ReLU: The Rectified Linear Unit activation function.
- $\sum_k$: Summation overall feature maps (*k* channels) in the final convolutional layer.
- a_k^c: The importance weight (gradient) for the *k*-th feature map for class *c*. These weights indicate how important each feature map is for predicting the target class.
- A^k: The *k*-th feature map from the final convolutional layer.

The ReLU functions select only the activation maps that contribute positively to the class prediction score, as those with negative scores most likely mean that part of the image belongs to a different category. This produces an overall heatmap which highlights the regions of the input image that are most influential in the model's decision-making process (Selvaraju, Cogswell et al. 2020).

In recent years, Grad-CAM techniques have been employed in many image analysis studies to elucidate features of images which are important for model decisions. It was used to provide visual explanations of the prediction of COVID-19 on X-ray and CT-scan images as a straightforward approach to highlight the inner workings of the DL models (Panwar, Gupta et al. 2020). A study that compared various GradCAM maps of different groups of brain MRI

scans showed that GradCAM can clearly highlight brain areas important for multiple sclerosis classification in most subjects. Interestingly, the heatmaps also revealed regions that contributed to the DL model's incorrect classifications (Zhang, Hong et al. 2021). Both these results show this technique's usefulness in supporting and understanding DL decisions —even when the predictions are incorrect.

Grad-CAM analysis is not usually conducted independently, but in tandem with DL techniques to verify how DL techniques can be improved to reach a more accurate or desirable result. Priori techniques such as those mentioned in the previous section were incorporated into Grad-CAM modelling to enhance the interpretability of these black box models. A study utilised a prior knowledge of colour and edge information of bronchoscopic images to reduce the hypersensitivity of DL models to high-brightness areas, thereby improving diagnostic outcomes (Yan, Sun et al. 2023). Another study investigated the usefulness of using GradCAM-generated heatmaps in assisting tumour diagnosis in CT images. The consulted medical practitioner confirmed that the GradCAM heatmaps, when combined with the CNN prediction results, aided in determining tumour presence (Chien, Lee et al. 2022). Overall, the Grad-CAM results were used as additional evidence to improve the interpretability of the proposed method's diagnostic outcomes. In recent years, Grad-CAM++ has improved upon Grad-CAM by producing more accurate heatmaps that precisely highlight the image areas that contribute to the model's classification decision (Peng, Jin et al. 2024). Grad-CAM++ will be utilised in this chapter to elucidate the relationship between image features and pinkeye classifications.

6.2.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME is a technique that explains individual predictions by approximating the complex model locally with a simpler, more interpretable linear model. It perturbs the input data, observes the changes in the model's output, and fits a simple model to these changes to understand the local decision boundary between the classes (Ribeiro, Singh et al. 2016). In the context of image

analysis, LIME decomposes the image into superpixels, which are contiguous regions with similar characteristics. By perturbing these superpixels (e.g., by turning them on or off) and observing the effects on the model's predictions, LIME identifies which regions of the image are most influential in the model's decision-making process (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco et al. 2019).

The general formula for LIME:

$$\xi(\mathrm{x}) = \arg\min_{g} \min_{g \in \mathrm{G}} (L(f, g, \pi_x) + \Omega(\mathrm{g})) \qquad (6.2)$$

where:

- $\xi(\mathrm{x})$ is the optimal interpretable model *g* that best explains the prediction for instance *x*
- *g* is the interpretable model
- *G* is the family of possible interpretable models
- *f* is the original DL model
- $L(f, g, \pi_x)$ is the fidelity loss function, measures how well the interpretable model *g* approximates *f* locally around *x*
- π_x is a proximity function, giving higher weights to samples close to the instance *x* being explained$\Omega(\mathrm{g})$ is the complexity penalty, making sure that *g* remains simple and interpretable

The equation represents the minimisation of the explanatory infidelity $L(f,g,\pi_x)$ of a potential explanation *g*, generated by a surrogate model *G*, within a neighbourhood defined by π_x (z) around a given sample of the dataset (*x*). The neighbourhood is obtained by applying perturbations to the dataset *(x)*, enabling the surrogate model to approximate the behaviour of the original model f around the decision boundary.

One study investigated the optimal number of pixels required to form an effective superpixel, ensuring that it encapsulated critical classification information within the image data to meaningfully contribute to the predictions made by DL models (Palatnik de Sousa, Maria Bernardes Rebuzzi

Vellasco et al. 2019). This study also explored the optimal segmentation algorithm for breaking images into superpixels for LIME analysis in the context of identifying tumour tissue in histology slide images, suggesting a range of 20–40 patches per image is suitable (Palatnik de Sousa, Maria Bernardes Rebuzzi Vellasco et al. 2019). Further research in histopathology demonstrated that LIME could generate superpixels that seem to concur with human experts, showcasing its usefulness in explaining the AI system (Sokol and Flach 2020). Additionally, LIME has been shown to improve model performance by identifying the most significant "activation sets" passing through CNN layers before the final classification layer, thereby highlighting the input sections most relevant for each class (Toğaçar, Muzoğlu et al. 2022). In our study, LIME can be applied to visually interpret the classification decisions for each stage and severity of pinkeye. By following the guidelines from the literature, we can ensure that parameters such as the number of superpixels and the choice of algorithm for generating these superpixels are optimised to highlight relevant image regions. This approach will enable us to generate meaningful visualisations that emphasise the areas of the eye image most critical to the model's classification decisions.

6.2.3 SHAP (SHapley Additive exPlanations)

SHAP utilises SHapley values to help explain how CNNs make predictions. They attribute the contribution of each input feature to the model's predicted class output (Štrumbelj and Kononenko 2014) and describe how this prediction is "fairly" distributed amongst these feature inputs (Lundberg and Lee 2017).

$$\phi_i = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \backslash i)] \qquad (6.3)$$

where:

- ϕ_i: The SHAP value for feature i. It represents the contribution of feature *i* to the prediction of the model for a specific input instance
- z': A subset of the simplified input features, which are perturbed feature representations of the input data

- x': the presence or absence of superpixels
- *M*: the total number of simplified input features
- $(M - |z'| - 1)!$: the factorial of the remaining features not in the subset z', excluding *i*. It accounts for the ordering of these excluded features
- $f_x(z')$: the model prediction for the subset of features $z' f_x(z' \backslash i)$: The model prediction for the subset of features z' with feature *i* excluded

Equation 6.3 calculates the marginal contribution of feature *i* by comparing the model's output with and without feature *i* in a given subset of features (z'). It averages this contribution over all possible subsets of the simplified input features, weighted by the size of the subsets and the total number of features (Takeishi 2019).

SHAP has been used to successfully explain deep learning models by identifying specific regions of input images that contribute positively or negatively to a model's predictions (Kirabo, Murindanyi et al. 2024). For classification tasks, SHAP provides detailed visualisation tools that highlight important features influencing the decision-making process (Walia, Kumar et al. 2022). This methodology has been successfully applied in medical imaging, such as identifying critical lung regions in chest x-rays for COVID-19 classification, demonstrating its ability to localise features relevant to different classes while also highlighting regions contributing to misclassifications (Ong, Goh et al. 2021), which also instigated discussions among experts and overall improved the interpretability of the DL models. For our study, SHAP heatmaps were applied to provide additional insights to Grad-CAM++ and LIME into the regions of eye images most critical for classifying the stages and severity of pinkeye.

6.3 Methods

To better understand the deep learning model's decision-making process, three explainability techniques, Grad-CAM++, LIME, and SHAP, were applied to investigate the classification process of images representing each stage of

pinkeye classification. As discussed in the previous section, these methods were selected based on their prevalent use in medical image analysis and their capability to generate interpretable heatmaps highlighting the most influential features for classification (Yang, Wei et al. 2023).

The visualisation methods were applied specifically to the EfficientNetV2B2 model, previously identified in Chapter 4 and utilised in Chapter 5 for the classification of pinkeye stages and severity. Each staging category represented distinct clinical manifestations of pinkeye, ranging from healthy eyes (Normal) through severe ulceration (Active), healing (Resolving), and fully healed but scarred conditions (Resolved). The EfficientNetV2B2 model used in this analysis was previously trained and evaluated in Chapters 4 and 5, employing transfer learning from ImageNet-pretrained weights. Key hyperparameters included an input image size of 224×224 pixels, Adam optimiser with an initial learning rate of 0.001, categorical cross-entropy loss function, and batch size of 32. Additional regularisation strategies included dropout (0.5), and early stopping based on validation loss. Detailed dataset characteristics, annotation processes, and the selection rationale for these specific traits were provided in Chapter 5.

All X-AI methods were implemented in Python using widely used deep learning interpretability libraries, executed on Google Colab with cloud GPU acceleration:

- Grad-CAM++ was implemented using the tf-explain library available at https://github.com/sicara/tf-explain. (Chattopadhay, Sarkar et al. 2018)
- LIME utilised the lime Python package available at https://github.com/marcotcr/lime, employing superpixel segmentation to approximate local feature importance (Ribeiro, Singh et al. 2016).
- SHAP was implemented via the shap Python package (Lundberg & Lee, 2017) available at https://github.com/shap/shap (Lundberg and Lee 2017), specifically using Deep SHAP to compute gradient-based feature attributions.

The X-AI techniques largely utilised default parameter settings provided by their respective libraries, with minimal adjustments:

- Grad-CAM++: Default settings were maintained as per the tf-explain library, involving standard guided backpropagation with no additional tuning.
- LIME: Default parameters included 1,000 perturbations with the quickshift segmentation method (default kernel size of 4 and maximum distance of 200 pixels). No parameter adjustments were made.
- SHAP: Deep SHAP explanations were generated using internal gradient computations of the EfficientNetV2B2 model, with a default background dataset size as recommended by the SHAP documentation.

Explicit hyperparameter tuning was not conducted, as the primary goal was assessing interpretability rather than performance optimisation.

To illustrate the interpretability results, two correctly classified images per pinkeye stage (Normal, Active, Resolving, Resolved) were selected from the test dataset, ensuring accurate classification by EfficientNetV2B2. These particular images were randomly chosen to represent typical presentations and common variations within each stage, guided by clinical consultation and annotation consensus described in Chapter 5. Although selecting a broader set could enhance generalisability, the subset was sufficient to highlight consistent and representative feature-attribution patterns identified by each X-AI method.

Generated heatmaps from Grad-CAM++, LIME, and SHAP were systematically compared to evaluate consistency and clinical relevance of attributed features. Key evaluation aspects included:

- Localisation of clinically important features (e.g., corneal ulcers, blood vessels, scarring).
- Alignment with expert-defined visual features according to the Pinkeye classification Scorecard presented in Chapter 4.

- Identification of potential inconsistencies in the generated heatmaps.

This structured approach provided a comprehensive assessment of how effectively the X-AI methods captured clinically relevant features, thereby supporting a deeper evaluation of model reliability of EfficientNetV2B2.

6.4. Results

The results obtained using Grad-CAM++, LIME, and SHAP are presented below, with visualisations displaying heatmaps overlaid on the original images.

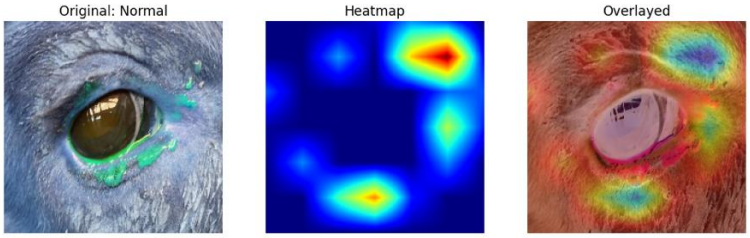
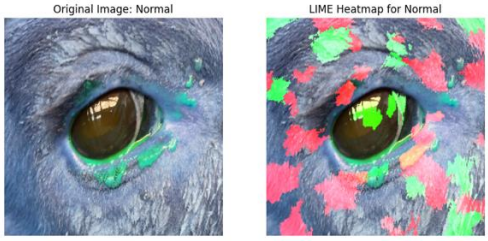
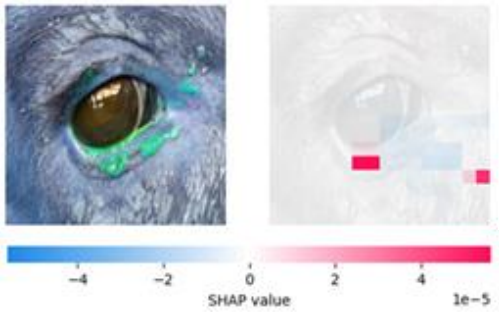
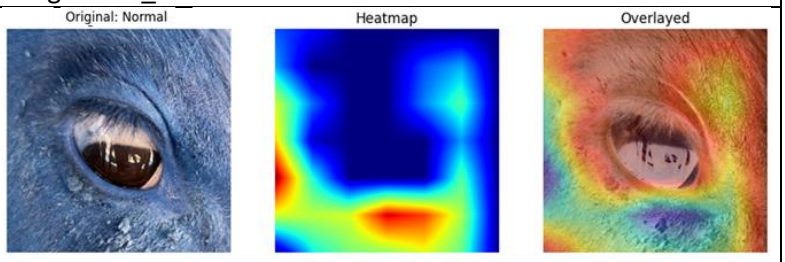
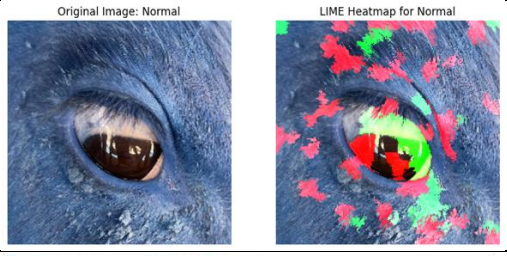
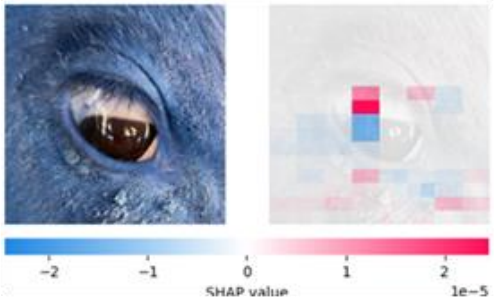
6.4.1 Visualisation results for Normal category

For the Normal stage classification, the Grad-CAM++ heatmap highlighted that the areas around the eye contributed most significantly to the model's decision to classify the image as "normal" for both selected images. Notably, features such as the absence of tearing or visible abnormalities in the medial canthus (front of the eye) appeared to play a crucial role in this classification. This observation was based on visual inspection of the Grad-CAM++ heatmap across the selected images, which consistently showed minimal activation of those areas in cases classified as 'Normal' (Table 6.1). This aligned with what is expected, as a healthy eye without ulcers or other irregularities was consistent with the "normal" category.

In the LIME visualisation, green regions represented features that do not contribute to the classification, while red regions indicated significant superpixels. The heatmap showed that the centre of the eye, particularly the pupil, did not provide significant features for classification for both images. Instead, areas directly underneath the eye at the lower eyelid and at the medial canthus were highlighted as contributing significantly for eye image IMG_5485, aligning with the Grad-CAM++ results. LIME’s heatmap for IMG_5463 showed mixed signals, with the lower eyelid highlighted as significant (Table 6.1).

The SHAP visualisation, which assigned positive values (in red) to regions contributing significantly to the classification and negative values to regions detracting from it (pale blue), offered fewer details compared to Grad-CAM++ and LIME. The SHAP heatmap highlighted two positive regions for IMG_5485: one directly beneath the pupil, which aligned with the findings from the previous heatmaps, and another at the front edge of the image, which did not correspond to any known eye features of significance. Conversely, the negative regions were near the front and below the eye, with slight overlap with areas previously marked as significant by Grad-CAM++ and LIME. For IMG_5463, the significant area was highlighted to be on the superior sclera, and a few spots on the bottom of the lower eyelid were signified to be insignificant (Table 6.1).

Table 6.1. Visualisation heatmap results for Normal classification

Stage: Normal			
Visualisation tools	Image: IMG_5485.JPEG		
Grad-CAM++			
LIME			
SHAP			
	Image: IMG_5463.JPEG		
Grad-CAM++			
LIME			
SHAP			

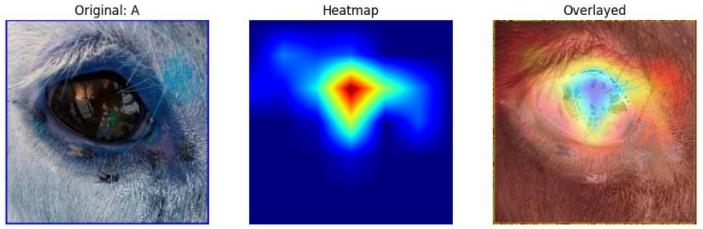
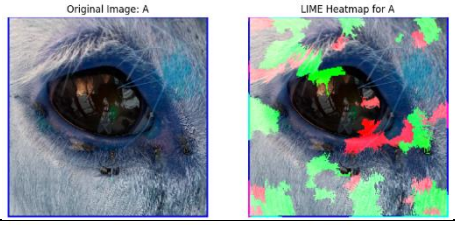
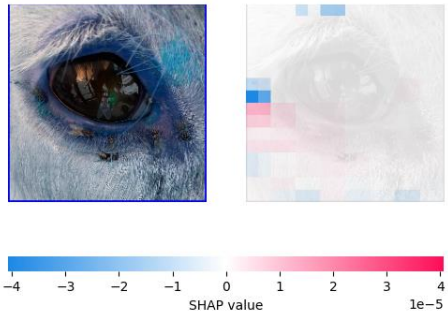
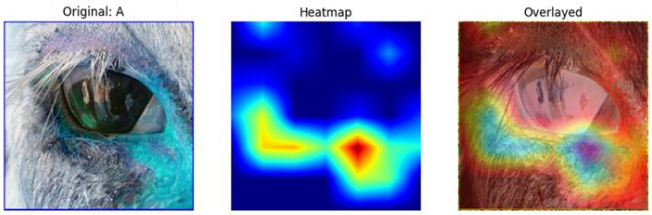
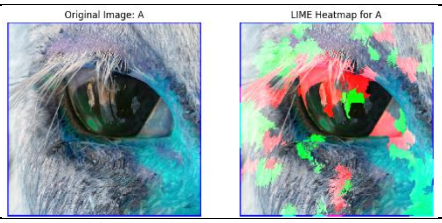
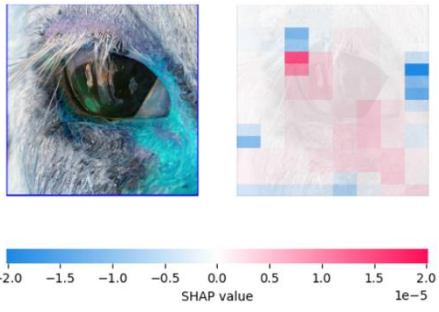
6.4.2 Visualisation results for Active category

For the Active stage classification, the Grad-CAM++ heatmap highlighted a significant region in the centre of the pupil, with a concentrated red area indicating the region of high importance for IMG_1842. When overlaid onto the original image, this heated region aligned with part of the ulcer, which suggested that the model identified this as a key feature contributing to the classification of "Active." For IMG_1862, whilst a portion of the ulcer of the cornea was highlighted, it mostly highlighted the Malar and Nasojugal folds as significant (Table 6.2).

The LIME heatmap also identified he ulcer region as important for IMG_1842. The red areas in the LIME visualisation cover superpixel regions around the eye, although not directly over the ulcer. However, the remaining green and red areas across the image lacked a consistent or interpretable pattern, limiting the additional insights that LIME provides in this instance. On the other hand, LIME did not cover the ulcer region for IMG_1852, yet it signified the upper regions of the sclera as important (Table 6.2).

The SHAP heatmap, on the other hand, highlighted the lateral canthus of IMG_1842, specifically in the area underneath, as significant. This may be linked to the green stain observed in that region, which the model appeared to interpret as important. However, the ulcer itself was not prominently highlighted, indicating that SHAP may have misjudged the staining as a key feature. Similarly, SHAP did not highlight the ulcer region of IMG_1852 as significant instead it highlighted the upper region of the sclera (Table 6.2).

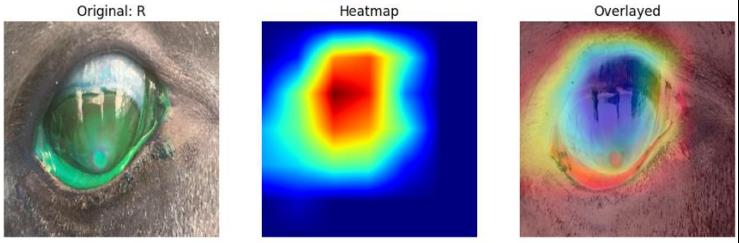
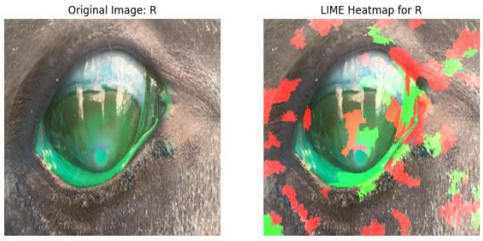
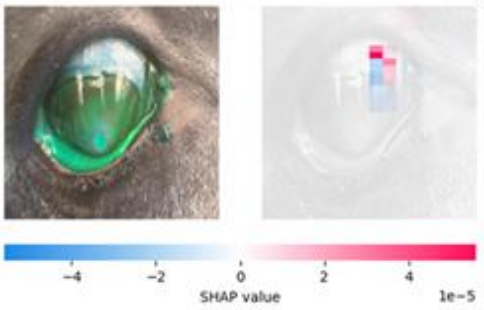
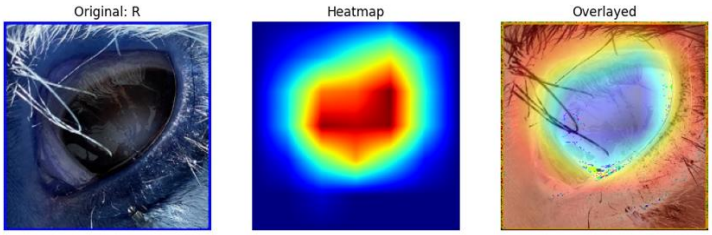
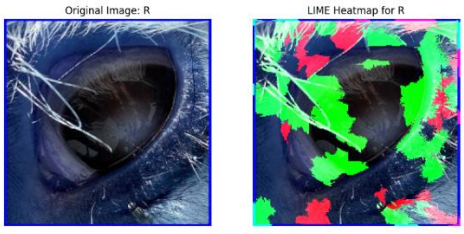
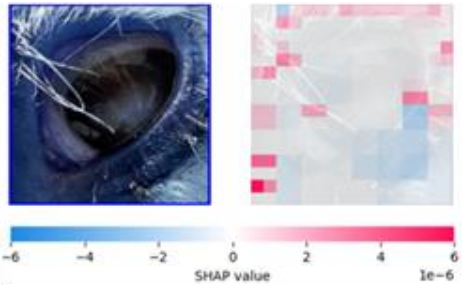
Table 6.2. Visualisation heatmap results for Active stage classification

Stage: Active	
Visualisation tools	Image: IMG_1842.JPEG
Grad-CAM++	
LIME	
SHAP	
Visualisation tools	Image: IMG_1862.JPEG
Grad-CAM++	
LIME	
SHAP	

6.4.3 Visualisation results for Resolving category

For the Resolving stage classification, the Grad-CAM++ heatmap highlighted the entire cornea region as significant for both images (Table 6.3). This region covers both the blood vessels and the ulcer, which aligned well with the eye attributes that defined the resolving stage. The model appeared to have identified the combination of these features, indicative of healing, as critical for this classification. The LIME heatmap similarly showed that the ulcer and blood vessels are important features. Red areas in the LIME visualisation covered certain parts of the ulcer, as well as the blood vessels at the superior sclera for IMG_0466, reinforcing their significance. However, for IMG_3518, the entire sclera was shown to be insignificant in its classification, whereas the superior eyelid was shown to be of importance. The SHAP heatmap highlighted a positive region at the superior sclera near the blood vessels for IMG_0466, indicating that this area contributed significantly to the classification. For IMG_3518, the surrounding regions of the eye were highlighted as significant.

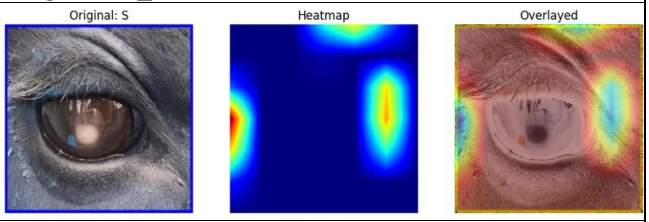
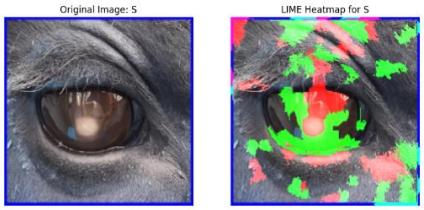
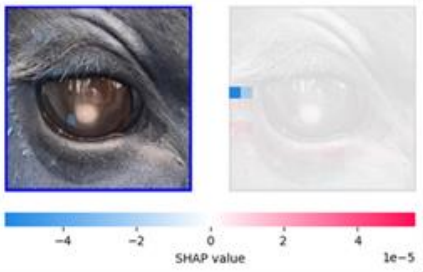
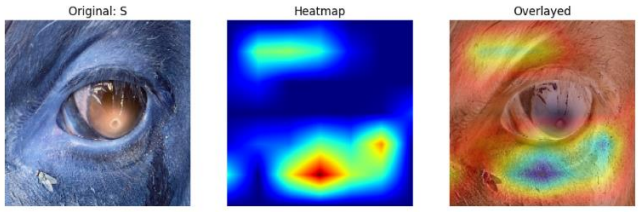
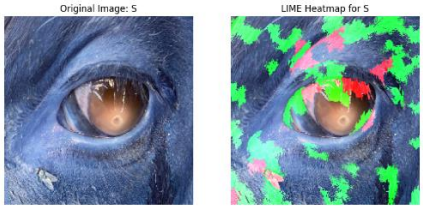
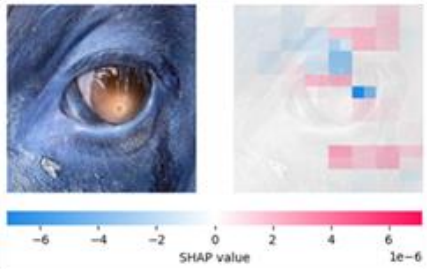
Table 6.3. Visualisation heatmap results for Resolving stage classification

Stage: Resolving	
Visualisation tools	Image: IMG_0466.JPEG
Grad-CAM++	
LIME	
SHAP	
Visualisation tools	Image: IMG_3518.JPEG
Grad-CAM++	
LIME	
SHAP	

6.4.4 Visualisation results for Resolved category

For the "Resolved" category, Grad-CAM++ unexpectedly did not highlight the scarred ulcer as a significant feature for both images (Table 6.4). Instead, the heatmap showed three distinct regions around the eye as contributing to the classification for IMG_2033 and showed the superior and inferior eyelid regions of IMG_4258 to be of significance. LIME, on the other hand, performed exceptionally well in this category by clearly marking the ulcer as the critical feature leading to the "Resolved" classification for IMG_2033. However, LIME did not highlight the ulcer region of IMG_4258 but parts of the surrounding sclera was highlighted. SHAP produced a less detailed explanation for IMG_2033, identifying a minor significant region on the lower eyelid and a highly insignificant region near the lateral canthus. SHAP produced more detail for IMG_4258 where parts adjacent to the ulcer was highlighted to be significant.

Table 6.4. Visualisation heatmap results for Resolved stage classification

Stage: Resolved	
Visualisation tools	Image: IMG_2033.JPEG
Grad-CAM++	
LIME	
SHAP	
Visualisation tools	Image: IMG_4258.JPEG
Grad-CAM++	
LIME	
SHAP	

6.5 Discussion

Understanding the decision-making process of deep learning models is essential, particularly in fields like veterinary medicine, where reliability and explainability are critical for accurate diagnoses and treatment. This chapter employed Grad-CAM++, LIME, and SHAP to visualise the significant features that influenced the DL model's predictions for classifying different stages of pinkeye in cattle. These tools offered complementary perspectives, providing insights into the model's trained weights and learned patterns while highlighting critical regions in the input images.

The XAI results provide valuable insights, though they should be considered alongside the model's performance characteristics discussed in Chapter 5. While the model achieved moderate accuracy, the explainability analysis still offers meaningful understanding of the decision-making patterns and can inform future model improvements.

Grad-CAM++ was most consistent in providing spatial attention maps by highlighting critical regions of the eye, such as ulcers and corneas, as shown by the results in the heatmaps for the "Active" and "Resolving" stages of pinkeye. The heatmaps it generated were clear, focused, and substantially interpretable, making it an effective tool for visualising the model's decision-making process. For the "Normal" stage, Grad-CAM++ identified areas surrounding the eye as key contributors to the classification, reflecting human intuition that a healthy eye lacks significant features in the cornea. In particular, both images showed low activation over the central cornea and superior sclera, while the medial canthus and inferior eyelid were subtly emphasised in one image, suggesting the absence of pathological signs in these regions as a cue for 'Normal' classification.

In the "Active" stage, the presence of an ulcer played a central role in the model's decision-making process, reinforcing the importance of ulcer detection in this classification. In both images, the Grad-CAM++ heatmaps prominently focused on the central cornea; however, one image showed

attention extending to the malar and nasojugal folds. This suggests that certain external periocular structures may also be interpreted as indicators of active inflammation, particularly if ulcer boundaries are diffuse or adjacent staining is present.

In the "Resolving" stage, Grad-CAM++ successfully highlighted both the ulcer and surrounding blood vessels, aligning with the eye attributes described in the Scorecard that reflect clinical understanding. Specifically, attention was concentrated over the corneal blood vessels extending across the superior sclera, and in one case, the superior eyelid was also involved, which may suggest the model interpreted vascularisation and upper eyelid inflammation as signs of healing. However, in the "Resolved" stage, Grad-CAM++ unexpectedly did not highlight the scarred ulcer as a significant feature. Instead, the heatmap identified three distinct regions around the eye as contributing to classification. In IMG_4258, Grad-CAM++ emphasised the superior and inferior eyelid margins, bypassing the central ulcer scar. This repeated pattern of peripheral focus may indicate that the model learned non-lesion-based contextual cues for identifying resolution. Since a scarred ulcer is fully white in appearance, it would have been expected to play a central role, making this result somewhat surprising. This may be of reduced relevance for its precise detection, as resolved cases typically do not require treatment. Except for this anomaly, Grad-CAM++ provided the clearest and most interpretable visualisations overall, making it a reliable tool for explaining model predictions.

Whilst direct comparisons with veterinary ophthalmological studies are currently unavailable in the literature, the effectiveness of Grad-CAM visualisation techniques observed in this study aligns with previous findings in fundus image analysis for human eye diseases. Both Grad-CAM and Grad-CAM++ have been successfully applied to the widely used VGG19 network, producing heatmaps that highlighted key regions of interest in cataract-infected and healthy eyes (Shah, Patel et al. 2023). Similarly, Grad-CAM was applied for diabetic retinopathy classification, effectively localising diseased

areas *(Jiang, Xu et al. 2020, Daanouni, Cherradi et al. 2021).* Beyond ophthalmology, Grad-CAM techniques have been extensively used in radiology, where they have been applied to VGG16, EfficientNet, DenseNet, and ResNet to detect COVID-19 from CT scans and X-rays, successfully highlighting key pulmonary features associated with infection (Palatnik de Sousa, Vellasco et al. 2021, Rajpal, Lakhyani et al. 2021, Shome, Kar et al. 2021). As in the present study, these heatmaps improved model interpretability by localising disease-related regions, thereby enhancing clinical trust in AI predictions. However, unlike lung abnormalities and retinal lesions, which are well-defined in these studies, corneal lesions in pinkeye are more subtle and progressively change with the disease stage, making them more challenging to detect and represent in heatmap outputs, particularly in later stages. Furthermore, many of these studies plainly report that Grad-CAM successfully highlighted relevant regions but do not examine, describe, and explain whether the visualisations precisely match expert labelled disease patterns. Additionally, most of these studies focus on binary classification (disease vs. no disease), whereas the present study involves differentiating between multiple stages of pinkeye. This added complexity increases the difficulty of accurately pinpointing affected regions, potentially explaining some of the inconsistencies observed in the heatmaps produced in our study.

LIME offered feature-level explanations by identifying significant superpixels that contributed to the model's classifications. In the "Active" and "Resolving" stages, LIME corroborated Grad-CAM++ by highlighting the ulcer and surrounding areas as critical features. In the second image of the Resolving category, however, LIME highlighted the superior eyelid as significant while the sclera was deemed non-contributory, suggesting that the eyelid contour may occasionally be weighted more heavily than expected. In contrast, LIME did not clearly segment the ulcer in one of the Active cases and instead highlighted the superior sclera as a contributing region. However, LIME also highlighted adjacent unimportant regions, which could introduce ambiguity in

interpretation. In the Normal category, LIME showed a mix of significant and insignificant areas around the eye, but not within the pupil, failing to reveal a clear or consistent pattern, which could distract users from fully understanding the model's decision-making process. One image showed a clearer pattern of activation around the medial canthus and lower eyelid, partially aligning with Grad-CAM++, while the second image produced more ambiguous patterns.

Similarly, in the Resolving stage, LIME marked some parts of the ulcer as non-contributory, while other regions showed inconsistent red and green areas around the eye, limiting interpretability compared to Grad-CAM++. In contrast, for the "Resolved" stage, LIME outperformed Grad-CAM++ by clearly marking the scarred ulcer as a critical feature for classification. *However, this strength was not consistent across both samples—LIME did not identify the ulcer in one image but highlighted parts of the surrounding sclera, possibly misattributing contextual features to the scar.* Yet, its tendency to highlight adjacent unimportant regions, such as areas near the ulcer, limited trust in its interpretability. While valuable for understanding feature-level contributions, the perturbation-based nature of LIME may have introduced noise in the results during its perturbation processes, making it less consistent and more difficult to interpret than Grad-CAM++ for image data. Yet, its tendency to highlight adjacent unimportant regions, such as areas near the ulcer, limited trust in its interpretability. While valuable for understanding feature-level contributions, the perturbation-based nature of LIME may have introduced noise in the results during its perturbation processes, making it less consistent and more difficult to interpret than Grad-CAM++ for image data.

In retinoblastoma classification, LIME was effective in segmenting regions of the fundus image containing features such as yellow-white masses, calcifications, and retinal detachments which are indicative symptoms of the disease. This corroborates with pinkeye symptoms, as both conditions involve distinct ocular abnormalities leading to disease progression. While LIME successfully highlighted most of the important features, some sections of the

outer eye were omitted, yet the overall segmentation was deemed successful (Aldughayfiq, Ashfaq et al. 2023). However, by visually comparing to our study, this research demonstrated better performance, as LIME segmented clearer regions of interest. Nonetheless, a key limitation was that LIME tended to segment large portions of the image indiscriminately, reducing its precision in identifying specific features essential for DL classification. Similarly, other ophthalmic studies have observed LIME's tendency to highlight excessively large regions, sometimes considering an entire section of the eye as contributing to CNN performance rather than isolating precise pathological features (Shipra and Rahman 2024). However, it was noted that when a pre-trained model such as InceptionV3 was used, LIME's heatmaps became more refined, focusing on smaller, more relevant regions. This trend suggests that while LIME successfully identifies critical features, its lack of spatial specificity may reduce clinical confidence in AI-based predictions. These findings align with the present study, where LIME occasionally failed to isolate key ocular structures, instead highlighting unrelated regions near the eye. The lack of spatial coherence in LIME's superpixel selection may pose a particular challenge for eye disease classification, where precise lesion localisation is crucial for accurate diagnosis.

Finally, SHAP, which relies on Shapley values to attribute importance scores to input features, provided a more global understanding of feature contributions. However, it produced less detailed heatmaps compared to Grad-CAM++ and LIME. In the "Active" and "Resolving" stages, SHAP often failed to highlight ulcers as significant features, occasionally misjudging staining artifacts as important, as shown for the "Normal" and "Active" categories. Additionally, in the Resolving stage, SHAP highlighted a positive region near the blood vessels, reinforcing their importance in classification. However, the presence of a negative region in the same area created inconsistencies, making SHAP's results harder to interpret compared to Grad-CAM++ and LIME. In the Resolved stage, SHAP produced a less detailed explanation, identifying a minor significant region underneath the eye and a

highly insignificant region near the back of the eye. These highlighted areas did not correspond to any known key features for the "Resolved" category, indicating inconsistencies in SHAP's ability to interpret this particular stage. This inconsistency highlights a potential limitation of SHAP in medical imaging, where artifacts may distract the model's attention.

Despite these inconsistencies, SHAP has demonstrated success in human eye disease classification, particularly for diabetic retinopathy and choroidal nevus, where it successfully identified disease-contributing regions that aligned with expert-labelled annotations (Shakeri, Crump et al. 2023). In addition, SHAP was found to be more precise than LIME in pinpointing specific pixel-level information crucial for the classification of retinoblastoma (Aldughayfiq, Ashfaq et al. 2023). In a separate study, however, SHAP heatmaps produced mixed results, as the highlighted pixels contributing to correct classifications appeared vague and were not concentrated in distinct areas (Balaha, Hassan et al. 2025). This study involved multi-class classification (4–5 classes across three datasets), making detailed comparisons of SHAP's performance more complex. However, the diffuse and less focused nature of the heatmaps observed in that study closely resembles the findings in our research, suggesting that SHAP may be more effective for detecting clear, high-contrast visual cues rather than for deep learning tasks requiring fine-grained class differentiation. This trend indicates that while SHAP provides global feature importance, its spatial precision may be limited when applied to complex multi-class classification problems compared to binary classification tasks. This study also utilised a combined CNN and vision transformer model which improved classification accuracy, yet SHAP did not reflect this enhancement, indicating that model performance may not contribute significantly to producing better heatmap results, which is an area that could be further researched.

These visualisation techniques have significant implications for veterinary medicine and X-AI. Grad-CAM++ and LIME enhance trust in AI models by making their decisions transparent and easy to understand. By identifying

clinically significant features such as corneal ulcers and blood vessels, these tools can assist veterinarians in verifying AI predictions. SHAP, while less detailed, offers a broader perspective, which can be useful for explaining the global behaviour of the model. Overall, integrating X-AI techniques into DL workflows can bridge the gap between AI systems and human decision-making, improving transparency, fairness, and accountability.

However, a significant gap remains in the application of X-AI techniques to veterinary ophthalmology, particularly for livestock eye diseases. While Grad-CAM, LIME, and SHAP have been more frequently researched in human ophthalmology in recent years, no comparable studies have explored their use in bovine eye disease classification. Furthermore, existing human studies primarily focus on well-defined retinal pathologies in fundus images, whereas corneal diseases in livestock, such as pinkeye, present unique challenges, including greater lesion variability, progressive disease stages, and lower-quality images with more distracting artifacts. This lack of research presents a critical opportunity for future work in veterinary imaging, particularly in validating heatmap results with expert annotations to improve model interpretability and adapting X-AI methods to better capture the complexities of corneal disease progression in cattle.

6.6 Conclusion

This study was conducted to evaluate the interpretability of deep learning models for classifying different stages of pinkeye in cattle, using three explainable AI (X-AI) techniques: Grad-CAM++, LIME, and SHAP. Grad-CAM++ proved to be the most intuitive and reliable tool, consistently highlighting clinically relevant features such as corneal ulcers and vascularisation. LIME offered detailed, superpixel-level insights but exhibited inconsistencies and sometimes misattributed irrelevant regions. SHAP, though valuable for understanding global feature contributions, often failed to precisely localise key ocular features and occasionally overemphasised artefacts. These findings emphasise that heatmap results can vary significantly even between correctly classified images within the same disease stage, suggesting that no single X-AI

method offers universal reliability across all scenarios. This variability highlights the need for multi-method approaches to visualisation, combining the strengths of region-based (Grad-CAM++) and perturbation-based (LIME, SHAP) techniques for more robust interpretation.

From a practical standpoint, these inconsistencies in heatmap localisation may undermine clinical confidence in AI tools unless properly addressed. Future research should focus on validating X-AI outputs against expert annotations, as emphasised in recent literature [31, 33], to determine whether highlighted regions correspond to known anatomical features or spurious artefacts. Additionally, as demonstrated in studies using retinal fundus and CT imaging, more refined heatmaps are often achieved when paired with advanced architectures, such as hybrid CNN-transformer models or attention-based networks, which may improve spatial focus and reduce noise in attribution maps [34]. Thus, a key recommendation is to explore model architecture enhancements alongside X-AI refinement. This includes investigating whether newer backbone models, improved pretraining on related datasets, or fine-tuning with expert-guided saliency alignment can yield more interpretable outputs in veterinary imaging contexts. As veterinary ophthalmology lacks large-scale annotated datasets, further investment in expert-labelled ground truth for X-AI validation is crucial.

Ultimately, integrating multiple X-AI techniques, guided by domain expertise and supported by refined modelling approaches, will be vital for building trustworthy, explainable, and clinically adoptable AI systems in veterinary diagnostics.

Chapter 7 General discussion

7.1 Introduction

Australia's livestock industry represents a foundational pillar of the national economy, contributing significantly to national food security and economic prosperity, with beef production alone accounting for over $20 billion annually (MLA 2025). However, maintaining this productivity requires consistent attention to animal health and welfare, particularly in extensive pastoral systems where disease surveillance and management are inherently challenging. Diseases in such environments not only compromise animal welfare but also pose substantial economic burdens through decreased productivity, treatment costs, and reduced market value. Among these challenges, infectious bovine keratoconjunctivitis (IBK), commonly known as pinkeye, stands out as a particularly prevalent and economically significant ocular condition in cattle (Alexander 2010, Kneipp, Green et al. 2021)

Pinkeye manifests externally with clinical signs such as corneal opacity, inflammation, increased tearing, and ulceration. Despite its seemingly straightforward clinical presentation, accurately diagnosing and grading pinkeye severity in field conditions remains problematic. Traditional diagnostic methods are heavily reliant on visual inspection, which can be subjective and inconsistent when performed by individuals without specialised veterinary training or under suboptimal environmental conditions. Whilst experts can provide precise diagnoses, many of these farming localities are remote, with veterinary assistance often not readily available, by which time the animal may have already progressed into more harmful stages of the disease. Moreover, the absence of standardised diagnostic criteria or scoring systems further exacerbates variability (Kneipp 2021), undermining both accurate diagnosis and effective disease management.

The veterinary field has increasingly turned to artificial intelligence (AI), particularly deep learning (DL), as a promising means to enhance diagnostic precision, standardisation, and efficiency. While DL technologies have made significant strides in human medical imaging, their adaptation and implementation within veterinary contexts, particularly livestock health management, remain limited and unevenly distributed across disease types and species. Most veterinary DL research has concentrated on companion animals and commonly employed diagnostic modalities such as radiography and microscopy. Consequently, significant gaps exist regarding the application of AI to livestock diseases using easily accessible imaging modalities like mobile phone photography, despite their practical appeal in field conditions (Xiao, Dhand et al. 2025).

Addressing these gaps, this thesis investigates how deep learning techniques can enhance pinkeye diagnostics through the classification of disease attributes, stages, and severity from mobile phone-acquired cattle eye images. By focusing specifically on field-acquired images, this research addresses a critical practical challenge: ensuring diagnostic robustness under realistic, uncontrolled conditions typical of livestock production systems. Through comprehensive exploration and validation, this thesis not only aims to advance understanding of veterinary AI but also seeks to demonstrate its practical viability and applicability within the industry.

7.2 Methodological framework

One of the central contributions of this thesis lies in its comprehensive methodological framework. Unlike prior studies that typically focus primarily on model performance accuracy, this research adopts a holistic approach by integrating multiple stages of AI deployment—image preprocessing, expert-guided attribute annotation, deep learning modelling, and explainable AI (X-AI) interpretation. This end-to-end pipeline was deliberately designed to address not only algorithmic optimisation, but also the practical and diagnostic realities of working with field-acquired livestock images.

Critically, each element of the framework, from object detection to explanation, was developed in response to the practical constraints of remote cattle farming, where inconsistent image quality, limited access to expertise, and delayed diagnosis remain persistent challenges. Rather than treating these as obstacles, this investigation engages with them directly, offering a scalable and transparent AI-assisted solution that meets the unique demands of field-based veterinary medicine. These choices were essential in building reliability, interpretability, and eventual deployability into the modelling process. By combining clinical reasoning with modern AI tools, the investigation offers a replicable framework for disease classification in veterinary contexts, grounded in both methodological rigour and real-world relevance.

7.3 Gaps in veterinary DL diagnostics research

The conceptual and practical choices were informed by Chapter 2, a critical review of the existing literature, which revealed systematic gaps in species focus, modality use, and methodological transparency within current veterinary DL diagnostics research. Recognising that the field remains relatively underdeveloped compared to human medical imaging and lacks the kind of comprehensive meta-analyses found in more mature domains, a dedicated literature review was conducted not merely as background, but as a strategic synthesis to identify where meaningful and novel contributions could be made. In a landscape fragmented by species, imaging modalities, and methodological inconsistency, such a review was essential to map the current state of research and guide the design of a pipeline tailored to the real diagnostic challenges facing livestock health management.

To ensure rigour and transparency, a structured literature review was conducted using PRISMA 2020 guidelines (Page, Moher et al. 2021), focusing on the keywords “deep learning” and “veterinary” in PubMed. This process yielded 39 eligible studies, highlighting both the emerging nature of the field and the need for a systematic appraisal.

What emerged from the review was a striking imbalance in both the type of data used and the species studied. Over 80% of studies focused on companion animals, particularly dogs and cats, often using data collected in well-resourced urban veterinary clinics. These studies also largely utilised radiographs or microscopy images, reflecting the dominant diagnostic modalities available in such settings. Livestock species, particularly cattle, were rarely studied, and where they did appear, the focus tended to be on internal physiological traits such as reproductive status, measured via ultrasound or thermal imaging (Xiao, Dhand et al. 2025). This observation was not just descriptive but foundational: it demonstrated a clear disconnect between the pressing needs of livestock management such as the identification of visible external diseases like pinkeye, and the focus areas of existing veterinary DL research.

Another critical insight from the review concerned imaging modality. Radiographic and microscopic images dominated the landscape, comprising two-thirds of all included studies. These modalities, while valuable, often require specialised equipment and clinic-based settings, making them ill-suited to remote or field-based applications. By contrast, the use of external or photographic images including those captured with standard or mobile phone cameras was almost entirely absent. Only 5.1% of the reviewed studies employed external photography, and even fewer addressed ocular disease or livestock contexts. This gap was central to the thesis rationale: if veterinary diagnostics are to become more accessible, scalable, and relevant to large-scale livestock operations, they must begin to incorporate imaging modalities that are feasible in low-resource and field settings. The literature review thus made a compelling case for using mobile phone-acquired images, not as a convenience, but as a necessity for the practical deployment of AI in cattle health management.

The review also shed light on methodological tendencies across existing veterinary DL studies, particularly in the use of pre-trained convolutional neural networks such as ResNet, VGG, and Inception, models originally

developed for large-scale object recognition tasks like ImageNet classification. While their adoption reflects their proven reliability and ease of integration, most studies offered limited justification for their suitability in veterinary contexts, and few demonstrated efforts to adapt them to domain-specific challenges. In many cases, these networks were applied with little modification, raising concerns about their capacity to detect subtle or anatomically localised disease features in non-standardised animal images.

Validation practices showed similar constraints. Cross-validation was inconsistently applied, external test sets were uncommon, and overall accuracy remained the dominant evaluation metric even in tasks involving multiple classes or imbalanced datasets. More nuanced metrics such as F1-score, area under the curve (AUC), or mean absolute error (MAE) were reported in only a small number of cases, despite their value in reflecting real diagnostic performance, particularly in clinical scenarios where the cost of false positives and false negatives may differ.

These patterns directly informed the methodological design of this thesis. While pre-trained networks were also used here, their selection was based on careful consideration of performance with smaller datasets, efficiency, and suitability for high-variance, real-world imagery which are all characteristics particularly relevant in veterinary fieldwork. More importantly, model evaluation extended beyond accuracy to include AUC, F1-score, Kappa, and MAE where appropriate, reflecting an effort to capture the practical utility of classification in settings marked by class imbalance and diagnostic subtlety. In doing so, this work aimed to build on existing practices without replicating their limitations.

Another key methodological gap identified was the near-total absence of explainable AI (X-AI) tools in veterinary imaging studies. Despite growing attention to model interpretability in human medical AI, very few veterinary applications offered insight into how predictions were made. This presents a barrier to clinical trust and adoption, particularly in contexts where the stakes of misclassification may be high and the integration of AI into practitioner

workflows depends on transparency. This thesis responded to that gap by embedding X-AI techniques such as Grad-CAM++, LIME, and SHAP within the evaluation process, not as auxiliary tools but as essential components for understanding model focus and validating attention against clinical expectations (Shaban-Nejad, Michalowski et al. 2021, Coghlan and Quinn 2024).

Despite the prevalence and economic burden of pinkeye, it remains almost entirely absent from deep learning research in veterinary medicine. Most existing studies have focused on companion animals and relied on imaging modalities such as fundus photography, tools that are impractical in extensive livestock systems. In contrast, pinkeye manifests externally and is readily captured via mobile photography, making it a highly suitable yet overlooked candidate for AI-driven image analysis. This defined a clear gap that this thesis aimed to address. Thus, the focus of research on pinkeye was a deliberate response to a neglected intersection of clinical relevance, technical feasibility, and the potential for training deep learning models on externally acquired, non-specialist images.

7.4 Bridging the gap: Overcoming diagnostic challenges of field-acquired images

The observed shortcomings in current veterinary DL diagnostics research guided the foundational design of the pipeline developed in this thesis. One particularly striking insight was the importance of image standardisation in uncontrolled environments. Veterinary images, especially those captured in the field, are rarely curated or clean; they are affected by lighting conditions, background interference, and inconsistent framing. Any attempt to apply DL in this space, therefore, requires mechanisms to reduce variability at the source. This recognition of the diagnostic challenges posed by field-acquired images—particularly the variability in lighting, background artefacts, and framing—shaped the decision to begin the pipeline with an object detection and preprocessing stage. Rather than treating image cleaning as a preliminary technical task, this stage was conceived as a strategic response to one of the

most persistent limitations in existing veterinary AI literature: the absence of consistent, anatomically focused input data. A model trained on noisy or misaligned images risks learning spurious cues, such as halters or fence lines, rather than pathology-specific features. Consequently, consistent localisation of the eye was essential not only for model accuracy, but for ensuring clinical relevance and interpretability across downstream tasks.

The object detection model implemented for this purpose was based on YOLOv8, chosen for its speed, efficiency, and accuracy. It served two critical functions: isolating the eye region and generating a clean, standardised dataset suitable for classification. On-farm images frequently included artefacts such as ear tags, shadows, or brightly coloured fences, that introduced irrelevant variation and could mislead a classifier. By detecting and cropping around the eye, the pipeline focused subsequent modelling efforts on the region most relevant to clinical diagnosis, while discarding visual noise that had no diagnostic utility.

To develop this model, a manually annotated dataset of 200 images was used as training input, a sample size reflective of realistic annotation constraints in veterinary research. Within this constraint, data augmentation proved especially beneficial. Techniques including random rotations, horizontal flips, brightness shifts, and mosaic transformations were applied to expand the training set, resulting in a notable improvement in mean average precision (mAP) of up to 42%. This finding affirmed one of the thesis's broader methodological commitments: that even in data-scarce environments, thoughtful augmentation strategies can significantly enhance model learning. suggesting that well-targeted augmentation may reduce the need for exhaustive manual labelling while helping to prevent overfitting and minimise computational overhead during early model development.

Once trained, the object detection model achieved 100% sensitivity in identifying exactly one eye per image on a 10% test set. Manual inspection further confirmed that the predicted bounding boxes were anatomically accurate and excluded extraneous elements. This reliability was pivotal for

downstream modelling: with spatial consistency enforced at the outset, classification models could be trained on inputs that mirrored the diagnostic gaze of a veterinarian, reducing the risk of confounding influences and improving generalisability.

Importantly, this step also addresses a structural limitation in veterinary DL research: the lack of high-quality, standardised image repositories. In contrast to human medical imaging, where imaging protocols and institutional databases support uniform data collection, veterinary datasets are often inconsistent, opportunistic, and highly variable, especially when dealing with livestock animals. By introducing a lightweight, scalable object detection solution, this research contributes a practical methodological advance that can be adapted for other field-based diagnostic tasks in veterinary imaging.

The portability of the YOLOv8 architecture further suggests its potential for future use in real-time applications. With minimal computational requirements, the model could be embedded in mobile tools to assist with on-farm triage, alerting producers or technicians to suspected cases of pinkeye that warrant closer inspection. This points to a promising direction for applied veterinary AI: not as a replacement for clinical expertise, but as a first-pass screening tool in contexts where access to diagnostics may be delayed or limited.

Ultimately, this stage laid the groundwork for the classification models that followed. By ensuring each image was cropped to a consistent, clinically relevant region, the pipeline provided a stable foundation for learning fine-grained disease features. This principle of standardising input to enhance both accuracy and clinical validity laid the foundation for the next phase of the thesis, which focused on systematically identifying and modelling the key ocular features used by experts to diagnose pinkeye.

7.5 Seeing what experts see: Closing diagnostic gaps in pinkeye detection through attribute-based modelling

The next challenge lay in translating cropped eye images into clinically interpretable features that mirror how veterinarians assess pinkeye in the field. Instead of simplifying the condition into a binary classification, the approach aimed to replicate expert diagnostic reasoning by modelling the individual ocular signs used to determine disease stage and severity. This involved identifying and modelling a set of specific ocular features that experts use to assess disease presence, severity, and progression. To do so, a custom annotation framework was developed in collaboration with clinical experts, producing a structured scorecard of 17 attributes that collectively describe the morphological presentation of pinkeye.

The decision to pursue attribute-level classification served multiple purposes. From a medical perspective, it allowed the models to engage with the same cues a practitioner would use when inspecting an animal's eye: tearing, corneal opaqueness, blood vessel patterns, lesion size, and so on. From a methodological standpoint, this design enabled the decomposition of a complex and often ambiguous diagnostic process into smaller, well-defined tasks. Each attribute could be treated as a separate modelling problem, with its own distribution, labelling challenges, and evaluation strategy. This modularity not only improved interpretability but also allowed for targeted refinement of complex features without jeopardising the integrity of the broader classification task.

The scorecard system served as the central organising framework for this stage. Attributes were grouped by their relevance to different anatomical structures (e.g. conjunctiva, cornea, periocular tissue), and each was assigned a label format appropriate to its clinical expression. Some features, such as the presence of tearing or whether the corneal opacity touched the limbus, were naturally binary. Others, like corneal surface texture or opacity colour, required multiclass labels to reflect distinct visual categories. The rest, such as opacity size, tear volume, and ocular inflammation, were inherently ordinal,

requiring the model to distinguish graded severity levels. This diversity presented both an opportunity and a challenge i.e. the framework captured the richness of clinical observation, but it also demanded flexibility in model design and evaluation.

Across all attributes, classification models were developed using both a custom-designed CNN and a range of transfer learning architectures, including EfficientNetV2B2, ResNet50V2, VGG19, InceptionV3, and DenseNet121. These CNN-based models were developed in parallel to assess their relative strengths across different attribute types. The decision to incorporate transfer learning was driven by practical and methodological considerations, including the limited availability of annotated veterinary data, the need to accelerate convergence and reduce overfitting, and the demonstrated effectiveness of pre-trained models in related medical imaging tasks. This approach enabled consistent benchmarking across modelling strategies while ensuring that all models were evaluated on the same curated dataset of field-acquired cattle eye images. Binary classification tasks showed the highest overall performance. Several attributes achieved accuracy scores near or above 90%, including 'cornea opacity visible' (93.65%), 'blood vessels across lesion' (88.65%), and 'tear' (91.05%). Area under the curve (AUC) values for these models also remained consistently high, often exceeding 0.90. Sensitivity and specificity were well-balanced for most binary features, with F1-scores ranging from 0.85 to 0.91, reinforcing the models' reliability across both positive and negative cases.

Multiclass tasks presented a greater challenge. Although training accuracy remained moderate, final test performance reflected difficulty in separating closely related categories. For example, in the 'corneal surface' and 'opacity colour' classifiers, misclassifications frequently occurred between visually similar classes. F1-scores were notably lower for these tasks, often falling between 0.60 and 0.72, indicating reduced class-wise precision and recall. Confusion matrices confirmed that errors were not randomly distributed, but

clustered around adjacent categories, suggesting that model uncertainty reflected real-world diagnostic ambiguity rather than random mislabelling.

Ordinal variables proved particularly complex. For attributes like ‘tear volume’, ‘cornea opacity size’, and ‘periocular inflammation’, ordinal regression models were used to preserve the ranking structure of severity levels. While this approach improved overall model fit compared to categorical methods, performance metrics indicated persistent challenges. MAE scores across ordinal tasks ranged from 0.42 to 0.77, with Cohen’s Kappa values between 0.41 and 0.63, highlighting moderate agreement between predictions and ground truth. Importantly, most misclassifications occurred between neighbouring classes, such as mislabelling a grade 2 opacity as a grade 3. This directional accuracy suggests that the models captured the general trend of disease progression, even when fine discrimination between intermediate levels remained difficult. Future work could explore techniques such as label smoothing, attention-based mechanisms to improve sensitivity to subtle differences between adjacent grades, particularly in borderline cases.

Beyond accuracy and ranking metrics, this stage of the pipeline offered important insights into which features were most visually separable, and which may require improved data quality or clinical definition. For instance, attributes related to blood vessel patterns, such as ‘tree’ or ‘hedges’ distribution, showed good performance when the patterns were pronounced, but performance dropped markedly in borderline cases. Similarly, features that relied on relative comparisons (e.g. opacity covering more or less than 50% of the cornea) proved sensitive to framing and annotation consistency.

More broadly, the work presented here contributes a key advance: the operationalisation of veterinary clinical heuristics into structured, machine-readable targets. By formalising expert judgment into labelled features and training models to detect them individually, the research demonstrates a scalable strategy for encoding veterinary expertise in DL systems. Crucially, this approach offers more than just prediction as it provides a framework

through which complex disease states can be interpreted via their component parts, enabling more transparent model evaluation and clearer links to clinical decision-making.

These attribute-level models also laid the groundwork for later components of the thesis, including full disease staging and severity classification. By first establishing how reliably individual features could be detected, the research provided a necessary foundation for asking whether overall pinkeye stage or severity could be inferred from those same images. In this sense, the scorecard did not just scaffold the modelling process, it reflected an intentional effort to structure learning in alignment with the clinical reasoning process it ultimately aims to support.

7.6 From discrete features to integrated diagnosis: Predicting stage and severity of pinkeye from field images

With foundation in place, the thesis progressed to its central diagnostic objective: predicting overall pinkeye stage and severity from single images. Stage and severity classification represent the practical endpoints of the diagnostic pipeline, with direct implications for disease monitoring, treatment decisions, and animal welfare.

Chapter 5 of the research introduced the most clinically complex task: classifying pinkeye into four distinct stages and four severity levels for three of the diseased stages based on field-acquired images. Unlike attribute modelling, this required the model to integrate multiple visual features and patterns to produce a holistic diagnostic outcome. Disease staging and severity grading are central to veterinary decision-making, informing whether treatment is required, how urgently it should be administered, and how disease progression is monitored. As such, the performance and reliability of these models were critical for evaluating the practical diagnostic value of the pipeline.

The four-stage classification task: Normal, Active, Resolving, and Resolved, presented a categorical challenge that required the model to distinguish not

only between healthy and diseased eyes, but between phases of disease expression that often exhibit subtle visual differences. The most robust performance was observed in binary classification tasks, comparing the 'Active' stage against all other categories. Here, the best performing model, EfficientNetV2B2, achieved an accuracy of 94.18%, with an AUC of 0.98, and a strong balance between sensitivity (0.93) and specificity (0.95). These results suggest that the model was particularly effective at identifying active infections, making it a valuable tool for identifying clinically affected animals.

However, as the classification moved beyond binary comparisons into the full four-class task, performance declined. Overall accuracy for the multiclass stage model was 78.35%, with noticeable confusion between intermediate categories, particularly 'Resolving' and 'Resolved'. The F1-scores for these two classes dropped below 0.65, and the confusion matrix revealed consistent misclassification in both directions. This reflects the clinical reality that these stages do not always present with sharply distinct visual features, especially when viewed from variable angles or lighting conditions common in field photography. Notably, the model continued to perform well in distinguishing the 'Normal' and 'Active' categories, indicating that it could reliably detect the presence or absence of overt disease, even if it struggled with finer distinctions within the disease timeline. This difficulty in resolving adjacent categories is not unique to veterinary imaging; similar challenges have been reported in multiclass eye disease classification in human ophthalmology, even with larger and more controlled datasets. Future improvements could involve expanding annotated training sets to include more borderline cases, implementing contrastive learning to better separate visually similar classes, or fine-tuning class-specific thresholds to reduce ambiguity between intermediate stages.

Severity classification introduced additional complexity due to its ordinal nature. The four severity levels, ranging from 1 (minimal signs) to 4 (severe pathology), were treated as an ordinal regression task to preserve the logical structure of increasing severity. Overall, the model achieved an MAE of 0.63

and a Cohen's Kappa of 0.59, indicating moderate agreement with expert labels. Class-wise performance showed the strongest accuracy in classifying grades 1 and 4, with confusion clustering around the middle grades (grades 2 and 3). This suggests the model could identify cases at the extremes of the severity spectrum with reasonable confidence but had difficulty resolving ambiguous or borderline cases, which is an issue that mirrors common diagnostic uncertainty even among clinicians.

These results align with a broader observation made throughout the thesis: that model performance correlates closely with visual distinctiveness and labelling confidence. When disease stages or severity grades are clearly separated by visible clinical signs, such as opacity density, lesion size, or vascular intrusion, model accuracy improves substantially. In contrast, where disease presentation follows a more gradual progression or is subject to annotation variability, performance tends to drop. These patterns reinforce the value of earlier stages in the pipeline: by focusing on consistent image preprocessing and attribute-level modelling, the groundwork was laid for the model to learn higher-level patterns of disease progression.

Despite these efforts, some limitations persisted. The class imbalance inherent in field data affected both stage and severity modelling. 'Normal' and 'Resolved' stages were well represented, while 'Resolving' and especially 'Active' stages had fewer examples. Similarly, extreme severity grades (1 and 4) were more easily distinguishable than middle grades, where inter-annotator disagreement was more likely. While data augmentation and class reweighting strategies were applied during training, their impact was not sufficient to fully compensate for underlying dataset skew. These findings suggest that collecting a more balanced dataset with a particular focus on transitional stages and mid-range severity is essential for improving model granularity and reducing misclassification between clinically adjacent categories.

From a clinical standpoint, the high accuracy and AUC observed in distinguishing Active cases are particularly promising. Early identification of

symptomatic animals is crucial for limiting transmission and reducing welfare impacts. A field-deployable model capable of screening for early active-stage pinkeye could serve as a triage tool, helping producers prioritise animals for treatment or isolation. While full-stage classification and severity prediction remain more challenging, even moderate performance in these tasks may be useful in real-world applications where expert diagnosis is not always immediately available.

Conceptually, this phase of the research marked a shift from modelling isolated visual cues to reasoning about disease holistically. In doing so, it pushed the model to approximate the integrative judgment typically performed by human experts, where multiple features are subconsciously weighed and contextualised. While the models did not achieve perfect classification, the fact that they were able to learn and apply disease progression patterns from raw images illustrates the feasibility of end-to-end pinkeye diagnostics in cattle. Importantly, the findings also underscore that diagnostic AI in veterinary medicine may be best deployed as a decision support tool, highlighting likely disease states and flagging cases for further attention, rather than serving as an absolute replacement for expert input.

This stage of the pipeline also offered a valuable opportunity to reflect on how earlier design choices shaped model behaviour. The consistent eye-centred cropping ensured that training focused on the most diagnostically relevant region, while the attribute-level modelling of corneal and periocular features likely contributed to the model's ability to learn patterns associated with different stages of pinkeye. These upstream decisions supported model generalisation across a range of field conditions, but they could not eliminate all sources of ambiguity, particularly in distinguishing between intermediate or visually subtle categories.

7.7 From prediction to understanding: Closing the gap between AI outputs and clinical reasoning

Given the limitations discussed above, the need for interpretability became increasingly apparent. While the stage and severity models demonstrated promising performance, understanding how and why they arrived at specific predictions was essential for assessing their reliability and clinical trustworthiness. This was especially true for borderline or misclassified cases, where visual cues may have been ambiguous, contradictory, or influenced by noise. As the final component of the thesis, the application of explainable AI techniques provided a crucial layer of insight, allowing for direct visualisation of model attention and enabling a closer alignment between machine reasoning and expert clinical expectations.

To evaluate how the models arrived at their predictions, particularly for disease stage classification, three X-AI methods were employed: Grad-CAM++, LIME, and SHAP. In veterinary contexts where AI tools must support, not replace, expert judgment, such transparency is vital for clinical trust and adoption. Each offered a different interpretability lens. Grad-CAM++ produced heatmaps highlighting the spatial focus of the model's prediction; LIME identified which parts of the image contributed most to the output by perturbing regions; and SHAP provided pixel- or region-level estimates of feature contribution using game-theoretic attribution. These tools were applied to both correctly and incorrectly classified images to assess the reliability and clinical alignment of the model's decision-making.

In correctly classified images, particularly in the 'Active' and 'Normal' stage categories, Grad-CAM++ and LIME consistently focused on relevant anatomical regions, such as the cornea, medial canthus, or areas of visible inflammation. These outputs mirrored the attention patterns used by clinicians and confirmed that the model had learned to prioritise features with diagnostic relevance. SHAP, while less spatially precise, generally supported these findings by attributing high influence to central eye regions.

Together, these methods validated that model predictions were grounded in clinically meaningful cues.

However, in misclassified or borderline cases, especially among ‘Resolving’ and ‘Resolved’ images, attention maps were often more diffuse or inconsistent. In some instances, Grad-CAM++ highlighted irrelevant regions or extended beyond the eye onto the face; LIME sometimes identified fragmented areas of influence, and SHAP occasionally emphasised background features. These inconsistencies reflected the diagnostic ambiguity of these stages and echoed earlier findings from confusion matrices, where the greatest classification errors occurred between visually similar categories.

The X-AI results also served to evaluate upstream design choices. The object detection and cropping step, introduced to standardise input and minimise noise, was supported by saliency maps that largely remained confined to the eye region. This confirmed that the model was not relying on background artefacts or non-clinical signals, an important ethical and practical consideration in field-acquired datasets where contextual confounders are common.

From a deployment perspective, the inclusion of interpretability strengthens the case for AI as a decision support tool. Visual explanations allow end-users (whether veterinarians or producers) to see why a prediction was made, improving confidence and helping to flag uncertain or borderline cases for follow-up. Of the three techniques, Grad-CAM++ proved the most intuitive and clinically interpretable, while LIME offered useful local insights but was less stable in noisy images. SHAP, though valuable in structured data contexts, was less effective for localising image-based features and may be better suited to non-visual veterinary applications.

Overall, the X-AI component reinforced a central aim of the thesis: to develop an AI diagnostic pipeline that is not only accurate but also interpretable, transparent, and usable in real-world settings. By shedding light on the internal reasoning of the models, these methods closed the loop between

prediction and explanation, which is an essential step in making AI outputs actionable, trustworthy, and aligned with veterinary clinical practice.

7.8 Limitations

While this thesis demonstrates the feasibility and potential of deep learning for pinkeye diagnosis in cattle, several limitations remain. First, the dataset was drawn from real-world field conditions, which, while ecologically valid, introduced inherent variability in lighting, angle, and image quality. Although object detection and cropping reduced some of this noise, inconsistencies may still have affected model performance, particularly in multiclass and ordinal classification tasks.

Second, class imbalance was a persistent challenge. Certain stages and mid-range severity levels were underrepresented, limiting the model's ability to learn fine distinctions across the disease spectrum. Data augmentation provided some mitigation, but future work would benefit from collecting larger, more balanced datasets, especially with more examples of resolving and resolved stages, and intermediate severity cases.

A further limitation of this study is the absence of a truly independent external validation dataset drawn from separate farms or geographical regions. All images were sourced from the same population group, and no additional data from external herds were available at the time of this research. Consequently, while the results demonstrate strong internal validity within the sampled population, external generalisability beyond the study environment remains untested.

Third, while multiple pre-trained models and a custom CNN were evaluated, architectural innovation was not a core focus of this work. Future studies may explore novel or veterinary-specific model architectures, as well as ensemble approaches that combine the strengths of different classifiers. In terms of evaluation, this thesis relied primarily on image-level classification. However, pinkeye diagnosis often involves temporal progression, and future research could explore sequential modelling techniques, such as recurrent neural

networks or vision transformers, to better capture disease dynamics over time. Additionally, incorporating metadata, such as animal age, breed, or farm environment, may further improve predictive performance and context relevance as pinkeye disproportionately affects calves compared to adult animals.

Finally, although X-AI tools were employed to interpret predictions, their current use remains exploratory. There is scope to refine veterinary-specific interpretability tools and to conduct user studies with clinicians to assess whether these explanations are clinically meaningful and actionable in practice.

7.9 Recommendations and Future Prospects

This final section presents key recommendations arising from the thesis findings and outlines future prospects for AI-based diagnostic systems in veterinary medicine. Together, they offer practical guidance for researchers, veterinary practitioners, and policymakers, while also reflecting on the long-term direction of the field. The goal is to support the development of transparent, scalable, and context-aware AI tools that address both clinical needs and field-based constraints.

For Future Research

- Expand training datasets to include more images with borderline or ambiguous features, which may improve the model's ability to distinguish between closely related disease stages or severity levels.
- Obtaining multi-site or international datasets to enable validation across broader and more diverse cattle populations, thereby providing a stronger assessment of model robustness and real-world applicability.
- Investigate the use of ensemble learning and attention-based architectures (e.g. Vision Transformers) to enhance model generalisation and class separability.

- Explore hybrid models that incorporate contextual data, such as animal age, breed, or geographic region, alongside image data to improve diagnostic specificity.
- Develop standardised, publicly available veterinary imaging datasets and benchmarks to promote reproducibility and accelerate progress in the field.
- Assess the utility of longitudinal data to track disease progression over time and refine temporal models of pinkeye development.

For Veterinarians and Farmers

- Collaborate with researchers to test the validity of the AI models in the field.
- Adopt image-based diagnostic tools as decision support systems rather than replacements for clinical judgement, particularly in settings where veterinary access is limited.
- Use such tools to aid in disease monitoring, early intervention, and treatment follow-up, especially for herd-level management.
- Provide feedback and annotated images to support the refinement and validation of AI models in diverse production systems and conditions.

For Policymakers

- Allocate funding to support the development and field validation of veterinary AI tools targeting common livestock diseases, including pinkeye.
- Establish regulatory and ethical frameworks to govern the use of AI in veterinary diagnostics, including transparency standards and clinical validation requirements.
- Encourage cross-sector collaboration between academic researchers, industry stakeholders, and animal health agencies to support the adoption of scalable, trustworthy AI solutions in livestock management.

7.10 Conclusions and future prospects

This thesis introduced a complete deep learning framework for the detection and classification of pinkeye in cattle, using field-acquired eye images as the foundation for a scalable, interpretable diagnostic pipeline. By integrating object detection, expert-informed attribute modelling, disease staging, severity grading, and explainable AI, the research demonstrated that clinically relevant and context-aware AI systems can be developed even under field constraints typical of extensive livestock systems.

More than a technical achievement, this work represents a step toward aligning artificial intelligence with veterinary reasoning. It shows how expert heuristics can be formalised into structured, machine-readable targets, enabling AI systems to complement clinical judgement. The resulting models not only support improved pinkeye management but signal broader possibilities for AI in animal health, especially in underserved or remote regions.

Looking forward, the promise of veterinary AI lies not just in higher accuracy, but in developing tools that are adaptable, trustworthy, and easy to deploy in the real world. Achieving this will require overcoming challenges in generalisability across environments and populations, expanding datasets to include edge cases and longitudinal views, and combining image data with other contextual. Incorporating these elements may enable predictive models capable of monitoring health trends over time, rather than offering one-off assessments.

Just as importantly, the future of veterinary AI depends on open, collaborative infrastructure. Shared, standardised datasets and cross-institutional validation will accelerate innovation and ensure transparency. Meanwhile, real-time deployment through computationally-light, interpretable models integrated into mobile applications remains the key to practical impact on farms and in the field.

In summary, this thesis offers both a tested framework and a vision: that AI can be harnessed not just to automate diagnosis, but to support informed, accessible, and ethically grounded decision-making in animal health. Realising this vision will require sustained collaboration between veterinary science, deep learning, and rural practice guided by a shared commitment to clinical relevance and livestock productivity and welfare.

References

Akinsulie, O. C., I. Idris, V. A. Aliyu, S. Shahzad, O. G. Banwo, S. C. Ogunleye, M. Olorunshola, D. O. Okedoyin, C. Ugwu, I. P. Oladapo, J. O. Gbadegoye, Q. A. Akande, P. Babawale, S. Rostami and K. O. Soetan (2024). "The potential application of artificial intelligence in veterinary clinical practice and biomedical research." Frontiers in Veterinary Science **11**.

Aldughayfiq, B., F. Ashfaq, N. Z. Jhanjhi and M. Humayun (2023) "Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP." Diagnostics **13** DOI: 10.3390/diagnostics13111932.

Alexander, D. (2010). "Infectious Bovine Keratoconjunctivitis: A Review of Cases in Clinical Practice." Veterinary Clinics: Food Animal Practice **26**(3): 487-503.

Ali, L., F. Alnajjar, H. Jassmi, M. Gochoo, W. Khan and M. Serhani (2021). "Performance Evaluation of Deep CNN-Based Crack Detection and Localization Techniques for Concrete Structures." Sensors **21**: 1688.

Ali, S., T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez and F. Herrera (2023). "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." Information Fusion **99**: 101805.

Alzubaidi, L., J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan (2021). "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions." Journal of Big Data **8**(1): 53.

Angelos, J. A. (2015). "Infectious Bovine Keratoconjunctivitis (Pinkeye)." Veterinary Clinics: Food Animal Practice **31**(1): 61-79.

Appleby, R. B. and P. S. Basran (2022). "Artificial intelligence in veterinary medicine." Journal of the American Veterinary Medical Association **260**(8): 819-824.

Appleby, R. B. and P. S. Basran (2024). "Artificial Intelligence in Diagnostic Imaging." Advances in Small Animal Care **5**(1): 67-77.

Arsomngern, P., N. Numcharoenpinij, J. Piriyataravet, W. Teerapan, W. Hinthong and P. Phunchongharn (2019). Computer-Aided Diagnosis for Lung Lesion in Companion Animals from X-ray Images Using Deep Learning Techniques. 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST).

Asgari Taghanaki, S., K. Abhishek, J. P. Cohen, J. Cohen-Adad and G. Hamarneh (2021). "Deep semantic segmentation of natural and medical images: a review." Artificial Intelligence Review **54**(1): 137-178.

Attallah, O. (2021). "MB-AI-His: Histopathological Diagnosis of Pediatric Medulloblastoma and Its Subtypes via AI." Diagnostics **11**: 359.

Aubreville, M., C. A. Bertram, C. Marzahl, C. Gurtner, M. Dettwiler, A. Schmidt, F. Bartenschlager, S. Merz, M. Fragoso, O. Kershaw, R. Klopfleisch and A. Maier (2020). "Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region." Scientific Reports **10**(1): 16447.

Baeza-Delgado, C., L. Cerdá Alberich, J. M. Carot-Sierra, D. Veiga-Canuto, B. Martínez de las Heras, B. Raza and L. Martí-Bonmatí (2022). "A practical solution to estimate the sample size required for clinical prediction models generated from observational research on data." European Radiology Experimental **6**(1): 22.

Balaha, H. M., A. E.-S. Hassan, R. A. Ahmed and M. H. Balaha (2025). "Advancing eye disease detection: A comprehensive study on computer-aided diagnosis with vision transformers and SHAP explainability techniques." Biocybernetics and Biomedical Engineering **45**(1): 23-33.
Banzato, T., M. Bernardini, G. B. Cherubini and A. Zotti (2018). "A methodological approach for deep learning to distinguish between meningiomas and gliomas on canine MR-images." BMC Veterinary Research **14**(1): 317.
Banzato, T., F. Bonsembiante, L. Aresu, M. E. Gelain, S. Burti and A. Zotti (2018). "Use of transfer learning to detect diffuse degenerative hepatic diseases from ultrasound images in dogs: A methodological study." The Veterinary Journal **233**: 35-40.
Bengio, Y., I. Goodfellow and A. Courville (2017). Deep learning, MIT press Cambridge, MA, USA.
Bertram, C. A., C. Marzahl, A. Bartel, J. Stayt, F. Bonsembiante, J. Beeler-Marfisi, A. K. Barton, G. Brocca, M. E. Gelain, A. Gläsel, K. d. Preez, K. Weiler, C. Weissenbacher-Lang, K. Breininger, M. Aubreville, A. Maier, R. Klopfleisch and J. Hill (2022). "Cytologic scoring of equine exercise-induced pulmonary hemorrhage: Performance of human experts and a deep learning-based algorithm." Veterinary Pathology **60**(1): 75-85.
Bezdan, T. and N. Bacanin (2019). Convolutional Neural Network Layers and Architectures.
Bhatt, C., I. Kumar, V. Vijayakumar, K. U. Singh and A. Kumar (2021). "The state of the art of deep learning models in medical science and their challenges." Multimedia Systems **27**(4): 599-613.
Biercher, A., S. Meller, J. Wendt, N. Caspari, J. Schmidt-Mosig, S. De Decker and H. A. Volk (2021). "Using Deep Learning to Detect Spinal Cord Diseases on Thoracolumbar Magnetic Resonance Images of Dogs." Frontiers in Veterinary Science **8**.
Boissady, E., A. De La Comble, X. Zhu, J. Abbott and H. Adrien-Maxence (2021). "Comparison of a Deep Learning Algorithm vs. Humans for Vertebral Heart Scale Measurements in Cats and Dogs Shows a High Degree of Agreement Among Readers." Frontiers in Veterinary Science **8**.
Boissady, E., A. de La Comble, X. Zhu and A.-M. Hespel (2020). "Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence." Veterinary Radiology & Ultrasound **61**(6): 619-627.
Bollig, N., L. Clarke, E. Elsmo and M. Craven (2020). "Machine learning for syndromic surveillance using veterinary necropsy reports." PloS one **15**(2): e0228105.
Borawar, L. and R. Kaur (2023). ResNet: Solving Vanishing Gradient in Deep Networks. Proceedings of International Conference on Recent Trends in Computing, Singapore, Springer Nature Singapore.
Bosma, J. S., D. Peeters, N. Alves, A. Saha, Z. Saghir, C. Jacobs and H. Huisman (2024). Reproducibility of training deep learning models for medical image analysis. Medical Imaging with Deep Learning, PMLR.
Bradley, R., I. Tagkopoulos, M. Kim, Y. Kokkinos, T. Panagiotakos, J. Kennedy, G. De Meyer, P. Watson and J. Elliott (2019). "Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning." Journal of veterinary internal medicine **33**(6): 2644-2656.

Buda, M., A. Maki and M. A. Mazurowski (2018). "A systematic study of the class imbalance problem in convolutional neural networks." Neural Networks **106**: 249-259.
Buric, M. (2024). "The Disease of the Canine Eye-From Image to Diagnosis Using AI."
Burrai, G. P., A. Gabrieli, M. Polinas, C. Murgia, M. P. Becchere, P. Demontis and E. Antuofermo (2023). "Canine mammary tumor histopathological image classification via computer-aided pathology: an available dataset for imaging analysis." Animals **13**(9): 1563.
Burti, S., V. L. Osti, A. Zotti and T. Banzato (2020). "Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs." The Veterinary Journal **262**: 105505.
Cao, W., V. Mirjalili and S. Raschka (2020). "Rank consistent ordinal regression for neural networks with application to age estimation." Pattern Recognition Letters **140**: 325-331.
Castiglioni, I., L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico and F. Sardanelli (2021). "AI applications to medical images: From machine learning to deep learning." Physica medica **83**: 9-24.
Chattopadhay, A., A. Sarkar, P. Howlader and V. N. Balasubramanian (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV).
Chen, R. J., M. Y. Lu, T. Y. Chen, D. F. K. Williamson and F. Mahmood (2021). "Synthetic data in machine learning for medicine and healthcare." Nature Biomedical Engineering **5**(6): 493-497.
Chien, J.-C., J.-D. Lee, C.-S. Hu and C.-T. Wu (2022). "The Usefulness of Gradient-Weighted CAM in Assisting Medical Diagnoses." Applied Sciences **12**(15): 7748.
Cho, H., Y. H. Hwang, J. K. Chung, K. B. Lee, J. S. Park, H.-G. Kim and J. H. Jeong (2021). "Deep Learning Ensemble Method for Classifying Glaucoma Stages Using Fundus Photographs and Convolutional Neural Networks." Current Eye Research **46**(10): 1516-1524.
Cihan, P., E. Gokce and O. Kalipsiz (2017). "A review of machine learning applications in veterinary field." Kafkas Universitesi Veteriner Fakultesi Dergisi **23**(4).
Cihan, P., A. Saygılı, C. Şahin Ermutlu, U. Aydın and Ö. Aksoy (2024). "AI-aided cardiovascular disease diagnosis in cattle from retinal images: Machine learning vs. deep learning models." Computers and Electronics in Agriculture **226**: 109391.
Coghlan, S. and T. Quinn (2024). "Ethics of using artificial intelligence (AI) in veterinary medicine." AI & SOCIETY **39**(5): 2337-2348.
Cohen, E. B. and I. K. Gordon (2022). "First, do no harm. Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology." Veterinary Radiology & Ultrasound **63**: 840-850.
Confalonieri, R., L. Coba, B. Wagner and T. R. Besold (2021). "A historical perspective of explainable Artificial Intelligence." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(1): e1391.
Cowton, J., I. Kyriazakis, T. Plötz and J. Bacardit (2018). "A combined deep learning gru-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors." Sensors **18**(8): 2521.
Cullen, J. N., T. J. Engelken, V. Cooper and A. M. O'Connor (2017). "Randomized blinded controlled trial to assess the association between a commercial vaccine against Moraxella bovis and the cumulative incidence of infectious bovine

keratoconjunctivitis in beef calves." Journal of the American Veterinary Medical Association **251**(3): 345-351.
Daanouni, O., B. Cherradi and A. Tmiri (2021). Automatic Detection of Diabetic Retinopathy Using Custom CNN and Grad-CAM. Advances on Smart and Soft Computing, Singapore, Springer Singapore.
Davis, J. and M. Goadrich (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning.
Deng, J., X. Xuan, W. Wang, Z. Li, H. Yao and Z. Wang (2020). A review of research on object detection based on deep learning. Journal of Physics: Conference Series, IOP Publishing.
Dhar, T., N. Dey, S. Borra and R. S. Sherratt (2023). "Challenges of deep learning in medical image analysis—improving explainability and trust." IEEE Transactions on Technology and Society **4**(1): 68-75.
Du, G., X. Cao, J. Liang, X. Chen and Y. Zhan (2020). "Medical Image Segmentation based on U-Net: A Review." Journal of Imaging Science & Technology **64**(2).
Dumortier, L., F. Guépin, M.-L. Delignette-Muller, C. Boulocher and T. Grenier (2022). "Deep learning in veterinary medicine, an approach based on CNN to detect pulmonary abnormalities from lateral thoracic radiographs in cats." Scientific reports **12**(1): 11418.
Duong, L. T., N. H. Le, T. B. Tran, V. M. Ngo and P. T. Nguyen (2021). "Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning." Expert Systems with Applications **184**: 115519.
Ebrahimi, M., M. Mohammadi-Dehcheshmeh, E. Ebrahimie and K. R. Petrovski (2019). "Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models." Computers in biology and medicine **114**: 103456.
Eelbode, T., P. Sinonquel, F. Maes and R. Bisschops (2021). "Pitfalls in training and validation of deep learning systems." Best Practice & Research Clinical Gastroenterology **52-53**: 101712.
El Naqa, I. and M. J. Murphy (2015). What Is Machine Learning? Machine Learning in Radiation Oncology: Theory and Applications. I. El Naqa, R. Li and M. J. Murphy. Cham, Springer International Publishing**:** 3-11.
Elngar, A. A., M. Arafa, A. Fathy, B. Moustafa, O. Mahmoud, M. Shaban and N. Fawzy (2021). "Image classification based on CNN: a survey." Journal of Cybersecurity and Information Management **6**(1): 18-50.
Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature (London) **542**(7639): 115-118.
Everingham, M., L. Van Gool, C. K. Williams, J. Winn and A. Zisserman (2009). "The pascal visual object classes (voc) challenge." International journal of computer vision **88**: 303-308.
Ezanno, P., S. Picault, G. Beaunée, X. Bailly, F. Muñoz, R. Duboz, H. Monod and J.-F. Guégan (2021). "Research perspectives on animal health in the era of artificial intelligence." Veterinary research **52**: 1-15.
Fayyad, M. F. and Mustakim (2024). Application of AlexNet, EfficientNetV2B0, and VGG19 with Explainable AI for Cataract and Glaucoma Image Classification. 2024 International Electronics Symposium (IES).

Feng, X., H. Yao and S. Zhang (2019). "An efficient way to refine DenseNet." Signal, Image and Video Processing **13**: 959-965.
Fitzke, M., D. Whitley, W. Yau, F. Rodrigues Jr, V. Fadeev, C. Bacmeister, C. Carter, J. Edwards, M. P. Lungren and M. Parkinson (2021). "OncoPetNet: A Deep Learning based AI system for mitotic figure counting on H&E stained whole slide digital images in a large veterinary diagnostic lab setting." arXiv preprint arXiv:2108.07856.
Fragoso-Garcia, M., F. Wilm, C. A. Bertram, S. Merz, A. Schmidt, T. Donovan, A. Fuchs-Baumgartinger, A. Bartel, C. Marzahl and L. Diehl (2023). "Automated diagnosis of 7 canine skin tumors using machine learning on H&E-stained whole slide images." Veterinary Pathology **60**(6): 865-875.
Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger and H. Greenspan (2018). "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification." Neurocomputing **321**: 321-331.
Ganesh, M., S. Dulam and P. Venkatasubbu (2022). Diabetic Retinopathy Diagnosis with InceptionResNetV2, Xception, and EfficientNetB3. Artificial Intelligence and Technologies, Singapore, Springer Singapore.
Gao, M., H. Wang, W. Shen, Z. Su, H. Liu, Y. Yin, Y. Zhang and Y. Zhang (2021). "Disease diagnosis of dairy cow by deep learning based on knowledge graph and transfer learning." International Journal Bioautomation **25**(1): 87.
García, R., J. Aguilar, M. Toro, A. Pinto and P. Rodríguez (2020). "A systematic literature review on the use of machine learning in precision livestock farming." Computers and Electronics in Agriculture **179**: 105826.
Gardenier, J., J. Underwood and C. Clark (2018). Object Detection for Cattle Gait Tracking. 2018 IEEE International Conference on Robotics and Automation (ICRA).
Gasqui, P. and J. Barnouin (2003). "Statistical modelling for clinical mastitis in the dairy cow: problems and solutions." Veterinary research **34**(5): 493-505.
Ghosh, K., C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk and N. Japkowicz (2024). "The class imbalance problem in deep learning." Machine Learning **113**(7): 4845-4901.
Girshick, R. (2015). Fast r-cnn. Proceedings of the IEEE international conference on computer vision.
Girshick, R., J. Donahue, T. Darrell and J. Malik (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition.
Glaret subin, P. and P. Muthukannan (2022). "Optimized convolution neural network based multiple eye disease detection." Computers in Biology and Medicine **146**: 105648.
Gogineni, S., A. Pimpalshende and S. Goddumarri (2021). Eye Disease Detection Using YOLO and Ensembled GoogleNet. Evolutionary Computing and Mobile Sustainable Networks, Singapore, Springer Singapore.
Gonçalves, T., I. Rio-Torto, L. F. Teixeira and J. S. Cardoso (2022). "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?" IEEE Access **10**: 98909-98935.
Gour, N. and P. Khanna (2021). "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network." Biomedical Signal Processing and Control **66**: 102329.

Grassmann, F., J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm and B. H. F. Weber (2018). "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography." Ophthalmology **125**(9): 1410-1420.
Greenspan, H., B. Van Ginneken and R. M. Summers (2016). "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique." IEEE transactions on medical imaging **35**(5): 1153-1159.
Greenwood, P. L., G. E. Gardner and D. M. Ferguson (2018). "Current situation and future prospects for the Australian beef industry - A review." Asian-Australas J Anim Sci **31**(7): 992-1006.
Guergueb, T. and M. A. Akhloufi (2021). Ocular Diseases Detection using Recent Deep Learning Techniques. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).
Gummadi, S. D. and A. Ghosh (2023). Classification of Ocular Diseases: A Vision Transformer-Based Approach. Innovations in Computational Intelligence and Computer Vision, Singapore, Springer Nature Singapore.
Haghofer, A., A. Fuchs-Baumgartinger, K. Lipnik, R. Klopfleisch, M. Aubreville, J. Scharinger, H. Weissenböck, S. M. Winkler and C. A. Bertram (2023). "Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing." Scientific Reports **13**(1): 19436.
Hamet, P. and J. Tremblay (2017). "Artificial intelligence in medicine." metabolism **69**: S36-S40.
Han, K., Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang and D. Tao (2023). "A Survey on Vision Transformer." IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1): 87-110.
Hao, W. and S. Zhili (2020). Improved mosaic: Algorithms for more complex images. Journal of Physics: Conference Series, IOP Publishing.
Harris, E. (2023). "Large Language Models Answer Medical Questions Accurately, but Can't Match Clinicians' Knowledge." JAMA **330**(9): 792-794.
Hartung, T. and N. Kleinstreuer (2025). "Challenges and opportunities for validation of AI-based new approach methods." ALTEX - Alternatives to animal experimentation **42**(1): 3-21.
He, K., X. Zhang, S. Ren and J. Sun (2015). "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence **37**(9): 1904-1916.
He, K., X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.
Henderson, P. and V. Ferrari (2017). End-to-end training of object class detectors for mean average precision. Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13, Springer.
Himel, G. M. S., M. M. Islam, K. A. Al-Aff, S. I. Karim and M. K. U. Sikder (2024). "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System." International Journal of Biomedical Imaging **2024**: 3022192.

Hisaria, S., P. Sharma, R. Gupta and S. Konatham (2024). An Analysis of Multi-Criteria Performance in Deep Learning-Based Medical Image Classification: A comprehensive review.
Huang, G., Z. Liu, L. Van Der Maaten and K. Q. Weinberger (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.
Huang, X., F. Schmelter, M. T. Irshad, A. Piet, M. A. Nisar, C. Sina and M. Grzegorzek (2023). "Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and supervised convolutional contrastive learning." Computers in Biology and Medicine **166**: 107501.
Hubbard-Perez, M., A. Luchian, C. Milford and L. Ressel (2024). "Use of deep learning for the classification of hyperplastic lymph node and common subtypes of canine lymphomas: a preliminary study." Frontiers in Veterinary Science **10**: 1309877.
Hussain, M., J. J. Bird and D. R. Faria (2019). A study on CNN transfer learning for image classification. Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK, Springer.
Iparraguirre-Villanueva, O., V. Guevara-Ponce, O. Paredes, F. Sierra-Liñan, J. Zapata-Paulini and M. Cabanillas-Carbonell (2022). "Convolutional Neural Networks with Transfer Learning for Pneumonia Detection." International Journal of Advanced Computer Science and Applications **13**.
Jain, A. K., J. Mao and K. M. Mohiuddin (1996). "Artificial neural networks: A tutorial." Computer **29**(3): 31-44.
Jeantet, L., V. Vigon, S. Geiger and D. Chevallier (2021). "Fully convolutional neural network: A solution to infer animal behaviours from multi-sensor data." Ecological Modelling **450**: 109555.
Jeong, Y. and J. Sung (2022). "An automated deep learning method and novel cardiac index to detect canine cardiomegaly from simple radiography." Scientific Reports **12**(1): 14494.
Ji, Y., H. Cho, S. Seon, K. Lee and H. Yoon (2022). "A deep learning model for CT-based kidney volume determination in dogs and normal reference definition." Frontiers in Veterinary Science **9**: 1011804.
Jiang, H., J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma and W. Qian (2020). A Multi-Label Deep Learning Model with Interpretable Grad-CAM for Diabetic Retinopathy Classification. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).
Jiang, P., D. Ergu, F. Liu, Y. Cai and B. Ma (2022). "A Review of Yolo Algorithm Developments." Procedia Computer Science **199**: 1066-1073.
Jin, A., S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein and L. Fei-Fei (2018). Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV).
Johnson, J. M. and T. M. Khoshgoftaar (2019). "Survey on deep learning with class imbalance." Journal of Big Data **6**(1): 27.
Khosla, C. and B. S. Saini (2020). Enhancing performance of deep learning models with different data augmentation techniques: A survey. 2020 International Conference on Intelligent Engineering and Management (ICIEM), IEEE.

Kim, H.-J., E. B. Baek, J.-H. Hwang, M. Lim, W. H. Jung, M. A. Bae, H.-Y. Son and J.-W. Cho (2023). "Application of convolutional neural network for analyzing hepatic fibrosis in mice." Journal of Toxicologic Pathology **36**(1): 21-30.
Kim, H. E., A. Cosa-Linan, M. E. Maros, N. Santhanam, M. Jannesari and T. Ganslandt (2021). "A review of transfer learning for medical image classification."
Kim, J. Y., M. G. Han, J. H. Chun, E. A. Huh and S. J. Lee (2022). "Developing a diagnosis model for dry eye disease in dogs using object detection." Scientific Reports **12**(1): 21351.
Kim, J. Y., H. E. Lee, Y. H. Choi, S. J. Lee and J. S. Jeon (2019). "CNN-based diagnosis models for canine ulcerative keratitis." Scientific reports **9**(1): 14209.
Kirabo, C., S. Murindanyi, N. P. Kirabo, K. M. Hasib and G. Marvin (2024). SHapley Additive exPlanations for Machine Emotion Intelligence in CNNs. Proceedings of International Conference on Computational Intelligence, Singapore, Springer Nature Singapore.
Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg and W.-Y. Lo (2023). Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision.
Kneipp, M. (2021). "Defining and diagnosing infectious bovine keratoconjunctivitis." Veterinary Clinics: Food Animal Practice **37**(2): 237-252.
Kneipp, M., M. Govendir, M. Laurence and N. K. Dhand (2021). "Current incidence, treatment costs and seasonality of pinkeye in Australian cattle estimated from sales of three popular medications." Preventive Veterinary Medicine **187**: 105232.
Kneipp, M., A. C. Green, M. Govendir, M. Laurence and N. K. Dhand (2021). "Perceptions and practices of Australian cattle farmers for the treatment of pinkeye (infectious bovine keratoconjunctivitis)." Preventive Veterinary Medicine **197**: 105504.
Kneipp, M., A. C. Green, M. Govendir, M. Laurence and N. K. Dhand (2021). "Risk factors associated with pinkeye in Australian cattle." Preventive Veterinary Medicine **194**: 105432.
Kneipp, M., A. C. Green, M. Govendir, M. Laurence and N. K. Dhand (2022). "Perceptions of Australian cattle farmers regarding the impact of pinkeye on farm productivity and animal welfare." Preventive Veterinary Medicine **204**: 105665.
Kokol, P., M. Kokol and S. Zagoranski (2022). "Machine learning on small size samples: A synthetic knowledge synthesis." Science Progress **105**(1): 00368504211029777.
Kora, P., C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak and U. Rajendra Acharya (2022). "Transfer learning techniques for medical image analysis: A review." Biocybernetics and Biomedical Engineering **42**(1): 79-107.
Krawczyk, B. (2016). "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence **5**(4): 221-232.
Krizhevsky, A., I. Sutskever and G. E. Hinton (2017). "Imagenet classification with deep convolutional neural networks." Communications of the ACM **60**(6): 84-90.
Krizhevsky, A., I. Sutskever and G. E. Hinton (2017). "ImageNet classification with deep convolutional neural networks." Commun. ACM **60**(6): 84–90.
Kshatri, S. S. and D. Singh (2023). "Convolutional Neural Network in Medical Image Analysis: A Review." Archives of Computational Methods in Engineering **30**(4): 2793-2810.

Kumar, A., S. K. Singh, S. Saxena, K. Lakshmanan, A. K. Sangaiah, H. Chauhan, S. Shrivastava and R. K. Singh (2020). "Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer." Information Sciences **508**: 405-421.
Kurup, G., J. A. A. Jothi and A. Kanadath (2021). Diabetic Retinopathy Detection and Classification using Pretrained Inception-v3. 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON).
Li, S., Z. Wang, L. C. Visser, E. R. Wisner and H. Cheng (2020). "Pilot study: application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs." Veterinary radiology & ultrasound **61**(6): 611-618.
Lippi, M., N. Bonucci, R. F. Carpio, M. Contarini, S. Speranza and A. Gasparri (2021). A yolo-based pest detection system for precision agriculture. 2021 29th Mediterranean Conference on Control and Automation (MED), IEEE.
Liu, L., W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu and M. Pietikäinen (2020). "Deep learning for generic object detection: A survey." International journal of computer vision **128**: 261-318.
Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg (2016). Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer.
Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions." Advances in neural information processing systems **30**.
Lustgarten, J. L., A. Zehnder, W. Shipman, E. Gancher and T. L. Webb (2020). "Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives—a joint paper by the Association for Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA)." JAMIA open **3**(2): 306-317.
May, A., S. Gesell‐May, T. Müller and W. Ertel (2022). "Artificial intelligence as a tool to aid in the differentiation of equine ophthalmic diseases with an emphasis on equine uveitis." Equine veterinary journal **54**(5): 847-855.
Mazurowski, M. A., H. Dong, H. Gu, J. Yang, N. Konz and Y. Zhang (2023). "Segment anything model for medical image analysis: an experimental study." Medical Image Analysis **89**: 102918.
McEvoy, F. J. and J. M. Amigo (2013). "Using machine learning to classify image features from canine pelvic radiographs: evaluation of partial least squares discriminant analysis and artificial neural network models." Veterinary Radiology & Ultrasound **54**(2): 122-126.
McEvoy, F. J., H. F. Proschowsky, A. V. Müller, L. Moorman, J. Bender-Koch, E. L. Svalastoga, J. Frellsen and D. H. Nielsen (2021). "Deep transfer learning can be used for the detection of hip joints in pelvis radiographs and the classification of their hip dysplasia status." Vet Radiol Ultrasound **62**(4): 387-393.
McGreevy, P., P. Thomson, N. K. Dhand, D. Raubenheimer, S. Masters, C. S. Mansfield, T. Baldwin, R. J. Soares Magalhaes, J. Rand, P. Hill, A. Peaston, J. Gilkerson, M. Combs, S. Raidal, P. Irwin, P. Irons, R. Squires, D. Brodbelt and J. Hammond (2017) "VetCompass Australia: A National Big Data Collection System for Veterinary Science." Animals **7** DOI: 10.3390/ani7100074.

Minaee, S., Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos (2021). "Image segmentation using deep learning: A survey." IEEE transactions on pattern analysis and machine intelligence **44**(7): 3523-3542.
Minh, D., H. X. Wang, Y. F. Li and T. N. Nguyen (2022). "Explainable artificial intelligence: a comprehensive review." Artificial Intelligence Review **55**(5): 3503-3568.
MLA. (2025). "The red meat industry." Retrieved 2025 April, 2025, from https://www.mla.com.au/about-mla/the-red-meat-industry/.
Mohith, V., K. Raja and I. R. Oviya (2024). Elevating Ocular Diagnosis: Harnessing the Power of EfficientNet for Eye Disease Classification. 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT).
Montserrat, D. M., Q. Lin, J. Allebach and E. J. Delp (2017). "Training object detection and recognition CNN models using data augmentation." Electronic Imaging **2017**(10): 27-36.
Morikawa, C., M. Kobayashi, M. Satoh, Y. Kuroda, T. Inomata, H. Matsuo, T. Miura and M. Hilaga (2021). "Image and video processing on mobile devices: a survey." The Visual Computer **37**(12): 2931-2949.
Mu, Y., Y. Sun, T. Hu, H. Gong and T. Tyasi (2021). "Improved model of eye disease recognition based on VGG model." Intell Autom Soft Comput **68**: 729-737.
Mumuni, A. and F. Mumuni (2022). "Data augmentation: A comprehensive survey of modern approaches." Array **16**: 100258.
Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu (2019). "Interpretable machine learning: definitions, methods, and applications." arXiv preprint arXiv:1901.04592.
Nam, M. G. and S. Y. Dong (2023). "Classification of Companion Animals' Ocular Diseases: Domain Adversarial Learning for Imbalanced Data." IEEE Access **11**: 143948-143955.
Niemeyer, F., F. Galbusera, M. Beukers, R. Jonas, Y. Tao, M. Fusellier, M. A. Tryfonidou, C. Neidlinger-Wilke, A. Kienle and H. J. Wilke (2024). "Automatic grading of intervertebral disc degeneration in lumbar dog spines." JOR spine **7**(2): e1326.
Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer and J. Clune (2018). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning." Proceedings of the National Academy of Sciences **115**(25): E5716-E5725.
O'Connor, A., H. Shen, C. Wang and T. Opriessnig (2012). "Descriptive epidemiology of Moraxella bovis, Moraxella bovoculi and Moraxella ovis in beef calves with naturally occurring infectious bovine keratoconjunctivitis (Pinkeye)." Veterinary microbiology **155**(2-4): 374-380.
Ong, J. H., K. M. Goh and L. L. Lim (2021). Comparative Analysis of Explainable Artificial Intelligence for COVID-19 Diagnosis on CXR Image. 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA).
Ovreiu, S., E. A. Paraschiv and E. Ovreiu (2021). Deep Learning & Digital Fundus Images: Glaucoma Detection using DenseNet. 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI).
Page, M. J., D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A.

McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting and J. E. McKenzie (2021). "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews." BMJ **372**: n160.
Palatnik de Sousa, I., M. Maria Bernardes Rebuzzi Vellasco and E. Costa da Silva (2019). "Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases." Sensors **19**(13): 2969.
Palatnik de Sousa, I., M. M. B. R. Vellasco and E. Costa da Silva (2021) "Explainable Artificial Intelligence for Bias Detection in COVID CT-Scan Classifiers." Sensors **21** DOI: 10.3390/s21165657.
Pan, Y., J. Liu, Y. Cai, X. Yang, Z. Zhang, H. Long, K. Zhao, X. Yu, C. Zeng and J. Duan (2023). "Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases." Frontiers in Physiology **14**: 1126780.
Panwar, H., P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj and V. Singh (2020). "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images." Chaos, Solitons & Fractals **140**: 110190.
Patil, D. D. and S. G. Deore (2013). "Medical image segmentation: a review." International Journal of Computer Science and Mobile Computing **2**(1): 22-27.
Patterson, J. and A. Gibson (2017). Deep learning: A practitioner's approach, " O'Reilly Media, Inc.".
Peng, D., C. Jin, J. Wang, Y. Zhai, H. Qi, L. Zhou, J. Peng and C. Zhang (2024). "Defects recognition of pine nuts using hyperspectral imaging and deep learning approaches." Microchemical Journal **201**: 110521.
Pereira, A. I., P. Franco-Gonçalo, P. Leite, A. Ribeiro, M. S. Alves-Pimenta, B. Colaço, C. Loureiro, L. Gonçalves, V. Filipe and M. Ginja (2023) "Artificial Intelligence in Veterinary Imaging: An Overview." Veterinary Sciences **10** DOI: 10.3390/vetsci10050320.
Perez, L. and J. Wang (2017). "The effectiveness of data augmentation in image classification using deep learning." arXiv preprint arXiv:1712.04621.
Priyadharshini, G. and D. R. J. Dolly (2023). Comparative Investigations on Tomato Leaf Disease Detection and Classification Using CNN, R-CNN, Fast R-CNN and Faster R-CNN. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS).
Rajpal, S., N. Lakhyani, A. K. Singh, R. Kohli and N. Kumar (2021). "Using handpicked features in conjunction with ResNet-50 for improved detection of COVID-19 from chest X-ray images." Chaos, Solitons & Fractals **145**: 110749.
Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz and K. Shpanskaya (2017). "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225.
Rana, M. and M. Bhushan (2023). "Machine learning and deep learning approach for medical image analysis: diagnosis to detection." Multimedia Tools and Applications **82**(17): 26731-26769.
Rasmy, L., Y. Xiang, Z. Xie, C. Tao and D. Zhi (2021). "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction." npj Digital Medicine **4**(1): 86.
Raza, A., M. U. Khan, Z. Saeed, S. Samer, A. Mobeen and A. Samer (2021). Classification of Eye Diseases and Detection of Cataract using Digital Fundus Imaging

(DFI) and Inception-V4 Deep Learning Model. 2021 International Conference on Frontiers of Information Technology (FIT).
Razavi, S. (2021). "Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling." Environmental Modelling & Software **144**: 105159.
Razzak, M. I., S. Naz and A. Zaib (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. Classification in BioApps: Automation of Decision Making. N. Dey, A. S. Ashour and S. Borra. Cham, Springer International Publishing**:** 323-350.
Reddy, S., J. Fox and M. P. Purohit (2019). "Artificial intelligence-enabled healthcare delivery." Journal of the Royal Society of Medicine **112**(1): 22-28.
Redmon, J., S. Divvala, R. Girshick and A. Farhadi (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition.
Redmon, J. and A. Farhadi (2018). "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767.
Ren, S., K. He, R. Girshick and J. Sun (2016). "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence **39**(6): 1137-1149.
Rezatofighi, H., N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese (2019). Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
Rho, J., S.-M. Shin, K. Jhang, G. Lee, K.-H. Song, H. Shin, K. Na, H.-J. Kwon and H.-Y. Son (2023). "Deep learning-based diagnosis of feline hypertrophic cardiomyopathy." Plos one **18**(2): e0280438.
Ribeiro, M. T., S. Singh and C. Guestrin (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.
Rizzoli, A. (2021). "The Ultimate Guide to Object Detection." Retrieved May 2023, from https://www.v7labs.com/blog/object-detection-guide.
Saab, K., T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park and E. Vedadi (2024). "Capabilities of gemini models in medicine." arXiv preprint arXiv:2404.18416.
Sakai, T. (2021). Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification, Online, Association for Computational Linguistics.
Salem, H., K. R. Negm, M. Y. Shams and O. M. Elzeki (2022). Recognition of Ocular Disease Based Optimized VGG-Net Models. Medical Informatics and Bioimaging Using Artificial Intelligence : Challenges, Issues, Innovations and Recent Developments. A. E. Hassanien, R. Bhatnagar, V. Snášel and M. Yasin Shams. Cham, Springer International Publishing**:** 93-111.
Salvi, M., F. Molinari, S. Iussich, L. V. Muscatello, L. Pazzini, S. Benali, B. Banco, F. Abramo, R. De Maria and L. Aresu (2021). "Histopathological Classification of Canine Cutaneous Round Cell Tumors Using Deep Learning: A Multi-Center Study." Frontiers in Veterinary Science **8**.
Santos-Bustos, D. F., B. M. Nguyen and H. E. Espitia (2022). "Towards automated eye cancer classification via VGG and ResNet networks using transfer learning." Engineering Science and Technology, an International Journal **35**: 101214.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra (2020). "Grad-CAM: visual explanations from deep networks via gradient-based localization." International journal of computer vision **128**: 336-359.
Şengöz, N. (2023). "UTILIZING DEEP LEARNING AND DATA AUGMENTATION FOR EARLY DETECTION OF EYE DISEASES IN PETS." International Journal of Engineering and Innovative Research **5**(2): 112-122.
Shaban-Nejad, A., M. Michalowski, J. S. Brownstein and D. L. Buckeridge (2021). "Guest editorial explainable AI: towards fairness, accountability, transparency and trust in healthcare." IEEE Journal of Biomedical and Health Informatics **25**(7): 2374-2375.
Shah, H., R. Patel, S. Hegde and H. Dalvi (2023). XAI Meets Ophthalmology: An Explainable Approach to Cataract Detection Using VGG-19 and Grad-CAM. 2023 IEEE Pune Section International Conference (PuneCon).
Shah, H. A., F. Saeed, S. Yun, J. H. Park, A. Paul and J. M. Kang (2022). "A Robust Approach for Brain Tumor Detection in Magnetic Resonance Images Using Finetuned EfficientNet." IEEE Access **10**: 65426-65438.
Shahinfar, S., P. Meek and G. Falzon (2020). ""“How many images do I need?” Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring." Ecological Informatics **57**: 101085.
Shakeri, E., T. Crump, E. Weis, E. Mohammed, R. Souza and B. Far (2023). "Explaining Eye Diseases Detected by Machine Learning Using SHAP: A Case Study of Diabetic Retinopathy and Choroidal Nevus." SN Computer Science **4**(5): 433.
Sheikhalishahi, S., R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi and V. Osmani (2019). "Natural language processing of clinical notes on chronic diseases: systematic review." JMIR medical informatics **7**(2): e12239.
Shickel, B., P. J. Tighe, A. Bihorac and P. Rashidi (2017). "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." IEEE journal of biomedical and health informatics **22**(5): 1589-1604.
Shim, H., J. Lee, S. Choi, J. Kim, J. Jeong, C. Cho, H. Kim, J. i. Kim, J. Kim and K. Eom (2023). "Deep learning-based diagnosis of stifle joint diseases in dogs." Veterinary Radiology & Ultrasound **64**(1): 113-122.
Shin, H., S. Jeon, Y. Seol, S. Kim and D. Kang (2023) "Vision Transformer Approach for Classification of Alzheimer’s Disease Using 18F-Florbetaben Brain Images." Applied Sciences **13** DOI: 10.3390/app13063453.
Shinde, P. P. and S. Shah (2018). A review of machine learning and deep learning applications. 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE.
Shipra, E. H. and M. S. Rahman (2024). An Explainable Artificial Intelligence Strategy for Transparent Deep Learning in the Classification of Eye Diseases. 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS).
Shome, D., T. Kar, S. N. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang and A. K. Saudagar (2021) "COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare." International Journal of Environmental Research and Public Health **18** DOI: 10.3390/ijerph182111086.
Shorten, C. and T. M. Khoshgoftaar (2019). "A survey on Image Data Augmentation for Deep Learning." Journal of Big Data **6**(1): 60.

Shorten, C. and T. M. Khoshgoftaar (2019). "A survey on image data augmentation for deep learning." Journal of big data **6**(1): 1-48.
Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.
Singh, M., S. Dalmia and R. K. Ranjan (2024). "Detection of diabetic retinopathy and age-related macular degeneration using DenseNet based neural networks." Multimedia Tools and Applications.
Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan (2023). "Large language models encode clinical knowledge." Nature **620**(7972): 172-180.
Sokol, K. and P. Flach (2020). "One Explanation Does Not Fit All." KI - Künstliche Intelligenz **34**(2): 235-250.
Song, J., D. Kim, E. Jeong and J. Park (2025) "Determination of Optimal Dataset Characteristics for Improving YOLO Performance in Agricultural Object Detection." Agriculture **15** DOI: 10.3390/agriculture15070731.
Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting." J. Mach. Learn. Res. **15**(1): 1929–1958.
Srivastava, S., A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni and V. Pattabiraman (2021). "Comparative analysis of deep learning image detection algorithms." Journal of Big data **8**(1): 66.
Štrumbelj, E. and I. Kononenko (2014). "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems **41**: 647-665.
Subrahmanyeswara Rao, B. (2020). "Accurate leukocoria predictor based on deep VGG‐net CNN technique." IET Image Processing **14**(10): 2241-2248.
Sudhan, M., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran and Y. A. Waji (2022). "Segmentation and Classification of Glaucoma Using U‐Net with Deep Learning Model." Journal of Healthcare Engineering **2022**(1): 1601354.
Suganyadevi, S., V. Seethalakshmi and K. Balasamy (2022). "A review on deep learning in medical image analysis." International Journal of Multimedia Information Retrieval **11**(1): 19-38.
Suksangvoravong, H., N. Choisunirachon, T. Tongloy, S. Chuwongin, S. Boonsang, V. Kittichai and C. Thanaboonnipat (2024). "Automatic classification and grading of canine tracheal collapse on thoracic radiographs by using deep learning." Veterinary Radiology & Ultrasound.
Szegedy, C., A. Toshev and D. Erhan (2013). "Deep neural networks for object detection." Advances in neural information processing systems **26**.
Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition.

Takeishi, N. (2019). Shapley values of reconstruction errors of pca for explaining anomaly detection. 2019 international conference on data mining workshops (icdmw), IEEE.
Tan, M. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946.
Tan, M. and Q. Le (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, PMLR.
Tan, M. and Q. Le (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning. C. Kamalika and S. Ruslan. Proceedings of Machine Learning Research, PMLR. **97:** 6105--6114.
Tan, M. and Q. Le (2021). Efficientnetv2: Smaller models and faster training. International conference on machine learning, PMLR.
Temraz, M. and M. T. Keane (2022). "Solving the class imbalance problem using a counterfactual method for data augmentation." Machine Learning with Applications **9**: 100375.
Terven, J., D.-M. Córdova-Esparza and J.-A. Romero-González (2023) "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS." Machine Learning and Knowledge Extraction **5**, 1680-1716 DOI: 10.3390/make5040083.
Thrusfield, M. (2018). Veterinary epidemiology, John Wiley & Sons.
Toğaçar, M., N. Muzoğlu, B. Ergen, B. S. B. Yarman and A. M. Halefoğlu (2022). "Detection of COVID-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from CNNs." Biomedical Signal Processing and Control **71**: 103128.
Tokarz, D. A., T. J. Steinbach, A. Lokhande, G. Srivastava, R. Ugalmugle, C. A. Co, K. R. Shockley, E. Singletary, M. F. Cesta and H. C. Thomas (2021). "Using artificial intelligence to detect, classify, and objectively score severity of rodent cardiomyopathy." Toxicologic pathology **49**(4): 888-896.
Toledo-Cortés, S., D. H. Useche, H. Müller and F. A. González (2022). "Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression." Computers in Biology and Medicine **145**: 105472.
Tracey, J. A., J. Zhu and K. R. Crooks (2011). "Modeling and inference of animal movement using artificial neural networks." Environmental and ecological statistics **18**: 393-410.
Vaswani, A. (2017). "Attention is all you need." arXiv preprint arXiv:1706.03762.
Vinicki, K., P. Ferrari, M. Belic and R. Turk (2018). "Using convolutional neural networks for determining reticulocyte percentage in cats." arXiv preprint arXiv:1803.04873.
Vrbaski, V., S. Josic, V. Vranjkovic, P. Teodorovic and R. Struharik (2023) "Puppis: Hardware Accelerator of Single-Shot Multibox Detectors for Edge-Based Applications." Electronics **12** DOI: 10.3390/electronics12224557.
Walia, S., K. Kumar, S. Agarwal and H. Kim (2022) "Using XAI for Deep Learning-Based Image Manipulation Detection with Shapley Additive Explanation." Symmetry **14** DOI: 10.3390/sym14081611.
Wang, J., L. Yang, Z. Huo, W. He and J. Luo (2020). "Multi-label classification of fundus images with efficientnet." IEEE access **8**: 212499-212508.

Ward, J. K. and M. K. Nielson (1979). "Pinkeye (Bovine Infectious Keratoconjunctivitis) in Beef Cattle." Journal of Animal Science **49**(2): 361-366.
Xiao, S., N. K. Dhand, Z. Wang, K. Hu, P. C. Thomson, J. K. House and M. S. Khatkar (2025). "Review of applications of deep learning in veterinary diagnostics and animal health." Frontiers in Veterinary Science **12**: 1511522.
Xiao, Y., Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du and X. Lan (2020). "A review of object detection based on deep learning." Multimedia Tools and Applications **79**: 23729-23791.
Xu, R., H. Lin, K. Lu, L. Cao and Y. Liu (2021). "A Forest Fire Detection System Based on Ensemble Learning." Forests **12**: 217.
Xu, W., Y.-L. Fu and D. Zhu (2023). "ResNet and its application to medical image processing: Research progress and challenges." Computer Methods and Programs in Biomedicine **240**: 107660.
Yan, P., W. Sun, X. Li, M. Li, Y. Jiang and H. Luo (2023). "PKDN: Prior Knowledge Distillation Network for bronchoscopy diagnosis." Computers in Biology and Medicine **166**: 107486.
Yang, W., Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin and B. Kang (2023). "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects." Human-Centric Intelligent Systems **3**(3): 161-188.
Yang, X., H. Dai, Z. Wu, R. Bahadur Bist, S. Subedi, J. Sun, G. Lu, C. Li, T. Liu and L. Chai (2024). "An innovative segment anything model for precision poultry monitoring." Computers and Electronics in Agriculture **222**: 109045.
Yilmaz, A. E. and H. Demirhan (2023). "Weighted kappa measures for ordinal multi-class classification performance." Applied Soft Computing **134**: 110020.
Yoon, Y., T. Hwang and H. Lee (2018). "Prediction of radiographic abnormalities by the use of bag-of-features and convolutional neural networks." The Veterinary Journal **237**: 43-48.
Young, T., D. Hazarika, S. Poria and E. Cambria (2018). "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine **13**(3): 55-75.
Yu, H., I.-G. Lee, J.-Y. Oh, J. Kim, J.-H. Jeong and K. Eom (2024). "Deep learning-based ultrasonographic classification of canine chronic kidney disease." Frontiers in Veterinary Science **11**: 1443234.
Zaitoun, N. M. and M. J. Aqel (2015). "Survey on image segmentation techniques." Procedia Computer Science **65**: 797-806.
Zapata, L., L. Chalco, L. Aguilar, E. Pimbosa, I. Ramírez-Morales, J. Hidalgo, M. Yandún, H. Arias-Flores and C. Guevara (2020). Detection of Cutaneous Tumors in Dogs Using Deep Learning Techniques. Advances in Artificial Intelligence, Software and Systems Engineering, Cham, Springer International Publishing.
Zeiler, M. D. and R. Fergus (2014). Visualizing and understanding convolutional networks. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer.
Zhang, C., P. Benz, D. M. Argaw, S. Lee, J. Kim, F. Rameau, J.-C. Bazin and I. S. Kweon (2021). Resnet or densenet? introducing dense shortcuts to resnet. Proceedings of the IEEE/CVF winter conference on applications of computer vision.

Zhang, M., K. Zhang, D. Yu, Q. Xie, B. Liu, D. Chen, D. Xv, Z. Li and C. Liu (2021). "Computerized assisted evaluation system for canine cardiomegaly via key points detection with deep learning." Preventive Veterinary Medicine **193**: 105399.
Zhang, Y., D. Hong, D. McClement, O. Oladosu, G. Pridham and G. Slaney (2021). "Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging." Journal of Neuroscience Methods **353**: 109098.
Zhao, Z.-Q., P. Zheng, S.-t. Xu and X. Wu (2019). "Object detection with deep learning: A review." IEEE transactions on neural networks and learning systems **30**(11): 3212-3232.
Zhu, S., B. Lu, C. Wang, M. Wu, B. Zheng, Q. Jiang, R. Wei, Q. Cao and W. Yang (2022). "Screening of common retinal diseases using six-category models based on EfficientNet." Frontiers in medicine **9**: 808402.
Zou, Z., K. Chen, Z. Shi, Y. Guo and J. Ye (2023). "Object detection in 20 years: A survey." Proceedings of the IEEE.

Xiao, S (2025). Detection and classification of eye diseases in cattle using image analysis with deep learning: code and resources. Zenodo: 10.5281/zenodo.17735797