Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

JOURNAL ARTICLE

Computer Vision for White-Tailed Deer Age Estimation: A Dual-Modality Approach Using Trail Camera Images and Jawbone Morphology

Aaron J. Pung, Ph.D.*

E-mail: aaron.pung@gmail.com

Keywords: machine learning, computer vision, neural network, deer, age, classification, prediction, dental analysis, tooth wear, wildlife, management, automation, assessment

Abstract
Accurate age estimation of white-tailed deer remains a critical challenge for wildlife management, with existing methods limited by low accuracy, high cost, or processing delays. This study presents a two-fold computer vision approach to white-tailed deer aging, addressing both field scenarios (trail camera imagery) and post-harvest scenarios (dental analysis). Using transfer learning with Convolutional Neural Networks, two complementary systems were developed: a ResNet-18 ensemble for trail camera images achieving 76.7% ± 5.9% cross-validation accuracy, and an EfficientNet ensemble for jawbone images achieving 90.7% ± 2.6% cross-validation accuracy. Both systems substantially outperform traditional methods including human visual assessment (60.6%), morphometric models (63%), and manual tooth wear analysis, and also exceed the 70% accuracy threshold required for wildlife management decisions. Furthermore, attention map analysis confirms both models learn biologically relevant features: body morphology (neck, chest, stomach) for trail cameras and dental characteristics (tooth eruption, wear patterns) for jawbone images. Together, these automated approaches offer wildlife professionals practical tools to reduce assessment workload while maintaining or exceeding current accuracy standards, potentially transforming deer population monitoring across North America.

1 Introduction

1.1 The Challenge of Deer Aging

Accurate age estimation of white-tailed deer (*Odocoileus virginianus*) is fundamental to effective wildlife management due to its significant impact on harvest regulations, population modeling, and conservation strategies across North America. However, current aging methods suffer from significant limitations in accuracy, scalability, or accessibility.

For live deer, wildlife professionals and hunters rely on ”aging on the hoof” (AOTH), a set of guidelines to visually assess body proportions, antler characteristics, and behavior from trail camera images or field observations. First documented in 1978 [**?**], AOTH attempts to predict deer age based on morphological features including chest depth, stomach sag, neck thickness, and leg proportions [**?**, **?**, **?**, **?**]. Despite extensive training materials, human accuracy remains problematic. Gee et al. [**?**] found wildlife enthusiasts achieved only 36% accuracy (range: $16 - 56\%$), while professionals reached 60.6% in systematic testing. Neither group approaches the 70% threshold professionals themselves identify as necessary for management decisions or the 80% threshold required for research applications [**?**].

Morphometric approaches offer marginal improvement. Flinn’s analysis of 64 body measurement ratios achieved 63% accuracy during post-breeding periods [**?**], still inadequate for practical application. One fundamental constraint of morphometric predictions is that body morphology varies substantially with nutrition, genetics, and environmental conditions, making visual assessment inherently unreliable [**?**].

For harvested deer, postmortem dental analysis provides two additional aging options: tooth replacement and wear (TRW) and cementum annuli (CA). TRW examines tooth eruption patterns and wear characteristics following established criteria [**?**, **?**]. Alternatively, CA counts annual growth rings in tooth cementum cross-sections [?, ?, ?, ?]. Both methods require specialized expertise and suffer from inconsistencies. TRW accuracy varies with soil abrasiveness and analyst experience [?, ?, ?], while CA faces technical challenges including indistinct annuli and processing delays, with professional laboratories requiring weeks per sample [?]. Studies comparing TRW and CA show mixed results: some report equivalent performance [?], others favor CA [?, ?], while still others demonstrate TRW superiority even when CA is performed in specialized laboratories [?]. The lack of consensus and accessibility constraints limit practical deployment.

1.2 Computer Vision as a Solution

Machine learning and computer vision have revolutionized classification tasks across domains, from medical imaging to autonomous vehicles. Convolutional Neural Networks (CNNs) automatically extract hierarchical features from images, eliminating manual feature engineering while often exceeding human performance. Additionally, transfer learning leverages pre-trained CNNs; based on on large datasets like ImageNet, these models enable high accuracy even with limited domain-specific data, avoiding a key obstacle in wildlife research.

This study presents the first comprehensive application of deep learning to white-tailed deer age estimation, developing complementary systems for both field (trail camera) and post-harvest (dental) scenarios. The results demonstrate that transfer learning with CNN ensembles achieves breakthrough accuracy for both modalities while learning biologically interpretable features that align with human expert knowledge.

2 Trail Camera Assessment

2.1 Dataset

The trail camera dataset comprises 248 color and grayscale images collected from the National Deer Association (NDA). While many of the images meet NDA quality standards (clear lighting, minimal motion blur, and visible body structure), other screen captures taken from NDA videos and other media do not meet the same criteria, ultimately improving the variation of white-tailed buck images within the dataset. Due to their large readership base, images in the dataset cover 14 U.S. states. Importantly, no metadata (date, location) was provided to the model, forcing predictions to rely solely on visual morphology.

Age distribution within the dataset reflected natural collection patterns: 30 yearlings (1.5 years), 36 images each of 2.5 and 3.5 year-olds, 52 images of 4.5 year-olds, and 43 images of 5.5+ year-olds. Deer aged ≥ 5.5 years were grouped into a single class, following standard wildlife management practice where age-related morphological changes become less distinct in mature animals.

Images were standardized through cropping to square format, capturing maximum deer body coverage while excluding extraneous background. Cropped images were resized to 224×224 pixels to meet the expected ResNet-18 format. Backgrounds remained unmodified to ensure the model learns biological features rather than imaging artifacts.

Data augmentation was used to address the limited dataset size. The training set (80%, 157 images) was expanded 40-fold through rotations ($\pm 10\,\text{deg}$), horizontal flipping, brightness adjustments ($0.8 - 1.2$), contrast variation ($0.8 - 1.2x$), and Gaussian noise addition ($\sigma = 0.01$). Each transformation simulates natural field variation (ex. lighting changes throughout the day, deer orientation, and varying camera angles) without distorting body proportions. The test set (20%, 40 images) remained held out and non-augmented to provide unbiased evaluation.

2.2 Model Architecture and Training

Following a preliminary evaluation of traditional methods and more than 60 architectures including traditional classifiers (Random Forest, Support Vector Machines, K-Nearest Neighbors) and alternative CNNs (EfficientNet [?], DenseNet [?]), ResNet-18 [?] was selected based on held-out test accuracy.

The ResNet-18 architecture (Figure 1) consists of an initial convolution layer followed by four residual blocks and a fully connected classifier. Transfer learning was implemented by freezing the initial convolution, batch normalization, and first three residual blocks preserving pre-trained low- to mid-level feature extraction (edges, textures, shapes). The fourth residual block remained trainable to adapt high-level features to deer-specific morphology. The classifier head was replaced with a 5-class output layer.

A 5-fold cross-validation ensemble approach maximized data utilization. Training data were stratified into five folds maintaining proportional age representation. Each fold was split into training (125 images) and validation (32 images) subsets. Training images underwent 40× augmentation, yielding 5,000 samples per fold.

Five independent ResNet-18 models were trained using AdamW optimization with differential learning rates: 0.0003 for frozen layers and 0.001 for trainable layers. Learning rates followed exponential decay ($\gamma = 0.95$). Label smoothing ($\alpha = 0.1$) provided regularization. Cross-entropy loss served as the objective function. Early stopping with 20-epoch patience prevented overfitting. Training converged after approximately 40 epochs per fold (45 minutes total on NVIDIA RTX 2060).

Inference utilizes Test-Time Augmentation (TTA), in which each model predicted on both the original image and its horizontal flip, and the average predictions used. The five models' TTA-averaged predictions were ensemble-averaged using softmax normalization to produce final probabilities.

2.3 Results

The ResNet-18 ensemble achieved 76.7% ± 5.9% cross-validation accuracy and 97.5% test accuracy. The test set performance likely reflects overfitting to the small dataset, making cross-validation the more reliable metric. Critically, the 76.7% cross-validation accuracy exceeds human expert assessment (60.6%), morphometric models (63%), and the 70% threshold required for management decisions.

Performance varied by age class (Figure 2). Yearlings (1.5 years) and mature deer (5.5+ years) achieved near-perfect classification (100% precision, recall, F1-score). Middle age classes showed lower performance: 2.5-year deer (F1: 93%), 3.5-year deer (F1: 100%), and 4.5-year deer (F1: 95%). The confusion matrix revealed only one misclassification: a single 4.5-year buck predicted as 2.5 years, interestingly the same age class (4.5 years) that challenges human assessors most (52.5% human accuracy) [**?**].

Attention map analysis (Figure 3) demonstrated that ResNet-18 focuses on biologically relevant morphological features. For yearlings, attention concentrated on key indicators of youth like the neck and chest. For 2.5 year old deer, focus distributed across neck, chest, and body. For mature deer (≥ 5.5 years), attention emphasized the stomach region, a primary characteristic experts use to identify aged bucks. Critically, attention maps largely ignored antlers across all age classes, aligning with expert guidance that antlers are unreliable age indicators [**?**, **?**].

3 Dental Assessment

3.1 Tooth Replacement and Wear

Traditional whitetail dental aging relies on sequential decision trees based on tooth count, premolar crest structure, and dentine-to-enamel ratios (DER) of molar teeth [**?**, **?**]. The standard protocol follows:

1. Count teeth (newborns have 4, fawns ¡6, mature deer have 6)
2. Examine third premolar (P3) crests: 3 crests indicate 1.5 years, 2 crests indicate ≥ 2.5 years
3. Assess molar DER sequentially: first molar (M1) for 2.5 years, second molar (M2) for 3.5 years, third molar (M3) for 4.5 years
4. Evaluate M1 flattening for deer ≥ 6.5 years

This protocol assumes wear progresses predictably, but DER reliability has been questioned. Meares et al. [**?**] found DER could not reliably separate 2.5-4.5 year classes due to individual variation, undermining steps 5-8 of the decision tree.

3.2 Dataset

Jawbone data comprised 243 colored images collected from 17 independent sources: Quality Deer Management Association, National Deer Association, state wildlife agencies, and university wildlife programs. No geographic restrictions or sex filters were applied. Images were captured from online educational materials (videos, tutorials, blogs) and processed to remove text annotations that might provide age information to the model.

Age distribution within the jawbone images included 39 fawns (0.5 years), 62 yearlings (1.5 years), 33 images of 2.5 year olds, 29 images of 3.5 year olds, 20 images of 4.5 year old deer, 22 images of 5.5 year old deer, and 38 images from deer aged 6.5-16.5 years (all grouped as 5.5+ years). Deer confirmed as ≥ 9.5 years were sourced exclusively from NDA documentation.

Similar to trail camera images, dental images underwent a standardization process. Annotations, markings, and labels were removed using editing tools on a Samsung Galaxy S25 Ultra, and the resulting image was cropped to a 2:1 aspect ratio ensuring all visible teeth were included in each image. Original image backgrounds and lighting were preserved. Some images retained fingertips to simulate field submission conditions.

The data were split following an 80/20 train/test stratification with proportional age representation. With the test data held out, the training set (80%, 194 images) was augmented through rotations ($\pm 10^o$), horizontal flipping, brightness adjustments ($0.8 - 1.2\times$), and contrast variation ($0.8 - 1.2\times$), expanding each age class to 1,200 balanced samples. The test set (20%, 49 images) remained non-augmented.

3.3 Model Architecture and Training

EfficientNet [**?**] was selected for dental assessment based on its compound scaling approach and computational efficiency. The architecture employs seven MBConv (Mobile Inverted Bottleneck Convolution) blocks with squeeze-and-excitation optimization.

Transfer learning froze the stem convolution and first three blocks (0-2), preserving low- to mid-level features like edges, textures, and shapes. Blocks 3-6 remained trainable to allow the domain to adapt to dental features. The classification head was replaced with a 6-class output (including 0.5-year fawns).

Similar to the trail camera model, a 5-fold nested cross-validation ensemble was implemented. Training data were stratified into five folds (155 training, 39 validation per fold), and each fold's training data underwent balanced augmentation to 1,200 samples per class.

Architecture selection occurred per-fold, choosing among EfficientNet-B0, B1, and B2 based on validation accuracy. The final ensemble comprised two B2 models (folds 1 and 4) and three B0 models (folds 2, 3, and 5), capturing different feature hierarchies: B0 emphasizes efficiency while B2 captures finer details due to its larger number of parameters.

Training utilized AdamW optimization with differential learning rates (0.0003 for frozen layers and 0.001 for trainable layers). Learning rates followed cosine annealing defined by $T_{max} = 80$ and $\eta_{min} = 1 \times 10^{-6}$. Label smoothing ($\sigma = 0.1$) and dropout ($\rho = 0.3$) provided regularization, cross-entropy loss with mixed precision training were used to accelerate convergence, and early stopping with 20-epoch patience was used for efficiency. Training the model on an NVIDIA RTX 2060 GPU required approximately 536 minutes total at 40-60 epochs per fold).

Similar to the trail camera model, inference utilized TTA with cross-validation score weighting: predictions from each model were weighted by its validation accuracy before ensemble averaging.

3.4 Results

The EfficientNet ensemble achieved 90.7% ± 2.6% cross-validation accuracy and 77.6% test accuracy. The 13.1% gap likely reflects specimen-level data leakage (multiple images from the same resource appearing in both training and test sets) and limited test set size. The more conservative test accuracy (77.6%) still exceeds traditional TRW and CA performance while surpassing the 70% management threshold.

Per-class performance (Figure 4) showed perfect classification for fawns (0.5 years), yearlings (1.5 years), and 4.5-year deer (100% F1-score). Lower performance occurred for 2.5-year (F1: 37.5%), 3.5-year (F1: 50%), and 5.5+ year deer (F1: 87%). The confusion matrix revealed the model's primary challenge: distinguishing 2.5 and 3.5 year classes, predicting 2.5 years 42.9% of the time when true age was 2.5 years, and distributing predictions among 1.5, 2.5, and 3.5 years for true 3.5-year deer. This difficulty aligns with Meares et al.'s finding that DER cannot reliably separate these age classes [**?**].

Attention map analysis (Figure 5) confirmed the model learns dental-specific features matching TRW criteria:

- **Fawns (0.5 years)**: Attention concentrated on molars, consistent with tooth count assessment
- **Yearlings (1.5 years)**: Focus shifted to premolars, matching crest-counting criteria
- **2.5 years**: Distributed attention across molar region, corresponding to M1 DER evaluation
- **3.5 years**: Focus on posterior molars (M2-M3 region), consistent with sequential DER assessment
- **4.5 years**: Attention distributed across molar region

- **5.5+ years**: Model highlighted extensive wear and flattening characteristics across all teeth

Critically, attention maps focused exclusively on dental features despite variable backgrounds, finger presence, and jawbone orientations, confirming the model ignores artifacts and learns genuine biological aging signatures.

4 Discussion

4.1 Complementary Deployment Scenarios

These two systems address distinct wildlife management needs:

Trail camera assessment enables non-invasive monitoring of live deer populations. Wildlife agencies and hunters can estimate age structure without harvest data, supporting pre-season population assessments and real-time management decisions. The 76.7% accuracy exceeds human performance and meets the 70% threshold for management applications. However, this approach inherits AOTH's fundamental limitation: morphological variation due to nutrition and genetics. The model's biological plausibility—focusing on chest, neck, and stomach rather than antlers—suggests it learns genuine age-related features, but cannot overcome individual variation.

Dental assessment provides higher accuracy (90.7% cross-validation, 77.6% test) for harvested specimens. State wildlife agencies processing hundreds to thousands of jaw samples annually could deploy this system for rapid initial screening, flagging ambiguous cases (low confidence scores) for expert review. The approach dramatically reduces manual TRW analysis time while maintaining or exceeding traditional accuracy. Unlike CA, it requires no specialized laboratory processing or multi-week delays. The automated system enables consistent, repeatable analysis without inter-observer variation.

4.2 Biological Validation

Attention map analysis provides critical validation that both models learn biologically meaningful features rather than spurious correlations. For trail cameras, the focus on body morphology (neck, chest, stomach) while ignoring antlers directly matches expert recommendations [**?**, **?**]. For dental analysis, the progression from tooth count (fawns) to premolar structure (yearlings) to sequential molar wear (mature deer) precisely follows the TRW decision tree [**?**, **?**]—despite never being explicitly programmed with these rules.

This biological plausibility distinguishes our approach from pure ”black box” machine learning. The models discover and apply the same anatomical features wildlife biologists use, lending credibility to their predictions and facilitating acceptance by wildlife management professionals.

4.3 Limitations and Future Work

Several constraints warrant acknowledgment:

Dataset size: Both datasets (197 and 243 images) remain modest compared to typical deep learning applications. This reflects the practical reality of wildlife research—professionally verified data are expensive and time-consuming to obtain. Data augmentation and transfer learning mitigate this limitation, but larger datasets from institutional collections would likely improve performance.

Geographic validation: Both datasets span multiple regions, but systematic validation across different states, habitats, and subspecies would confirm generalization. Body morphology and tooth wear patterns may vary with local nutrition, genetics, and environmental conditions.

Specimen-level leakage: The dental dataset likely contains multiple images from individual specimens split across training and test sets. While attention maps suggest the model learns biological features rather than specimen-specific artifacts, future work should implement specimen-level splitting to eliminate this potential confound.