Journal of Business, Social and Technology
Volume 7, Issue 4, 1971-1983
e_ISSN: 2807-6362
https://bustechno.polteksci.ac.id/index.php/jbt/index
DOI: doi.org/10.59261/bustechno.v7i4.761

Tarsius Image-Source Classification Using VGG Feature Extraction and Machine Learning Classifiers

Caecilia Eva Martina Korompis*
Universitas Airlangga, Indonesia

Imam Yuadi
Universitas Airlangga, Indonesia

***Corresponding author:**
Caecilia Eva Martina Korompis, Universitas Airlangga, Indonesia.
✉caecilia.eva.martina-2025@pasca.unair.ac.id

Article Info:

Article history:
Received: July 06, 2026
Revised: September 09, 2026
Accepted: September 15, 2026
Available Online: September 21, 2026

Keywords:
AI-generated Images; Image Classification; InceptionV3; Tarsius; VGG.

Abstract

Background: Reliable classification of heterogeneous Tarsius imagery is relevant to digital wildlife-data management, yet real photographs and AI-generated images may share visual characteristics that complicate automated separation.
Objective: This study compares VGG-16, VGG-19, and InceptionV3 feature extraction combined with Logistic Regression and Neural Network classifiers for real, cartoon-style, and AI-generated Tarsius images.
Methods: A balanced dataset of 300 web-sourced images (100 per category) was processed in Orange Data Mining. Performance was evaluated using AUC, classification accuracy (CA), F1-score, precision, recall, Matthews Correlation Coefficient (MCC), confusion matrices, Multidimensional Scaling (MDS), and silhouette plots.
Results: Logistic Regression with VGG-19 produced the strongest reported combination (AUC = 0.997, CA = 0.967, F1 = 0.967, precision = 0.968, recall = 0.967, and MCC = 0.950), closely followed by VGG-16 with Logistic Regression. Cartoon-style images were generally the most distinguishable, whereas real and AI-generated images showed the greatest overlap. InceptionV3 produced clearer two-dimensional MDS separation but lower aggregate predictive performance than the VGG-based Logistic Regression configurations.
Conclusion: VGG-19 with Logistic Regression provides the most balanced performance for the available dataset. The findings demonstrate that predictive metrics and feature-space visualizations should be interpreted jointly and remain limited by the small, web-sourced dataset and incomplete provenance records.

To cite this article: Korompis, C. E. M., & Yuadi, I. (2026). Tarsius image classification using VGG feature extraction and machine learning classifiers. *Journal of Business, Social and Technology, 7*(4), 1971–1983. https://doi.org/10.59261/bustechno.v7i4.761

INTRODUCTION

Tarsiers are small nocturnal and crepuscular primates whose ecological and conservation importance makes reliable visual documentation necessary. Their activity under low-light conditions, small body size, and habitat use can complicate field observation and image acquisition (Gursky-Doyen, 2010; Hidayatik et al., 2025). In Indonesia, pressures associated with hunting, habitat disturbance, and incomplete distribution records further increase the need for reproducible identification support. Within this context, image-based computational tools can complement, not replace, expert verification by organizing large collections of visual records and

identifying patterns that warrant closer examination.

Tarsiers receive legal protection under Law No. 5/1990 and Government Regulation No. 7/1999, while conservation assessments have also identified threatened tarsier taxa (IUCN, 2025; Shekelle & Salim, 2008). Conventional identification depends strongly on observer expertise and can be affected by subjectivity, limited observation time, image quality, and morphological similarity. Automated image classification is therefore relevant as a decision-support approach, particularly when digital image collections contain heterogeneous sources and visual styles. However, a classifier trained on web-sourced images must be interpreted cautiously because online images may contain unknown editing histories, repeated content, uneven resolution, and labels that have not been confirmed through field observation.

Advances in deep learning have broadened the use of image recognition for objects and animals (He et al., 2020; Zhu et al., 2022). Transfer learning is especially useful when the available dataset is limited because pretrained convolutional neural networks can provide feature representations without training a full architecture from the beginning. VGG-16 and VGG-19 employ deep stacks of convolutional layers, whereas InceptionV3 uses multi-scale operations to represent visual information. Extracted features may then be supplied to classifiers such as Logistic Regression or a Neural Network. Comparable hybrid strategies have produced useful classification performance in other image domains Amjoud and Amrouch (2022) and Pearline et al. (2019), although performance in one domain cannot be assumed to transfer directly to tarsier imagery.

Previous studies establish the methodological basis but leave a specific gap. Samudra et al. (2023) compared conventional classifiers using pretrained VGG-16 features for animal-type classification, but did not examine whether real, cartoon-style, and AI-generated images of the same animal concept occupy overlapping feature spaces. Pearline et al. (2019) compared conventional image processing and deep learning for plant recognition, demonstrating the value of learned representations in a different biological domain. Amjoud and Amrouch (2022) combined transfer learning with Logistic Regression for image-orientation detection, but the task did not involve visually similar real and synthetic categories. Thus, the available literature supports transfer learning and hybrid classification, yet provides limited evidence about source-style classification for Tarsius images and about the representational similarity between real and AI-generated images.

The novelty of this study lies in comparing three pretrained feature extractors (VGG-16, VGG-19, and InceptionV3) combined with Logistic Regression and Neural Network classifiers for three image-source categories: real Tarsius images, cartoon-style Tarsius images, and AI-generated Tarsius images. In addition to predictive metrics, Multidimensional Scaling (MDS) and silhouette plots are used to inspect the geometry and cohesion of the extracted feature space. This combination distinguishes classification performance from cluster structure: a model may achieve a strong predictive score while its two-dimensional projection still reveals overlap between categories.

The expected contribution is both methodological and applied. Methodologically, the comparison shows how architecture choice affects source-style discrimination in a small, balanced Tarsius image dataset. Applied implications concern the careful use of computer vision for sorting heterogeneous wildlife-related image collections. The analysis also identifies a practical limitation: AI-generated images can reproduce textures, lighting, and facial configurations that resemble photographs, whereas cartoon-style images usually contain more distinctive contours and colors. Consequently, the findings should be interpreted as evidence for classifying the three defined image-source categories, not as taxonomic identification of Tarsius species or as validation of an operational conservation system.

The study therefore aims to evaluate and compare the classification performance of VGG-16, VGG-19, and InceptionV3 feature extraction when paired with Logistic Regression and Neural Network classifiers. It also examines how the three image categories are distributed in the extracted feature space and identifies which architecture provides the most balanced evidence across AUC, classification accuracy (CA), F1-score, precision, recall, Matthews Correlation Coefficient (MCC), MDS, and silhouette analysis. The best-performing configuration is determined from the reported predictive metrics, while MDS and silhouette plots are treated as complementary descriptive evidence rather than substitutes for classification performance.

METHOD

The dataset comprised 300 Tarsius images obtained through Google image-search results and organized into three balanced image-source categories: 100 real Tarsius images, 100 cartoon-style Tarsius images, and 100 AI-generated Tarsius images. Images were selected when the depicted subject was visually recognizable as a tarsier and the file could be opened by Orange Data Mining; unreadable files and visibly irrelevant results were excluded. Because the original source URLs, retrieval date, licensing metadata, duplicate-screening log, and independent label-verification record were not retained in the study materials, these details are not reconstructed here. This limitation restricts provenance auditing and external replication and is addressed in the discussion.

The workflow was designed to compare image embeddings and downstream classifiers under a consistent evaluation framework. It comprised dataset organization, automated preprocessing in Orange Data Mining, feature extraction, classification, metric-based evaluation, and descriptive analysis of feature-space structure. Figure 2 summarizes this sequence without changing the original experimental scope.

Real Tarsius Images AI-Generated Tarsius Images Cartoon-Style Tarsius Images

Figure 1. Examples of Tarsius Images from Different Image-Source Categories

The experimental setup is described through combinations of various feature extractors and classifiers tested consistently to ensure fair comparison of results. Model performance was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, AUC, and Matthews Correlation Coefficient (MCC), which formed the basis for assessing and comparing model effectiveness. Further explanation is presented in the following research methodology diagram.

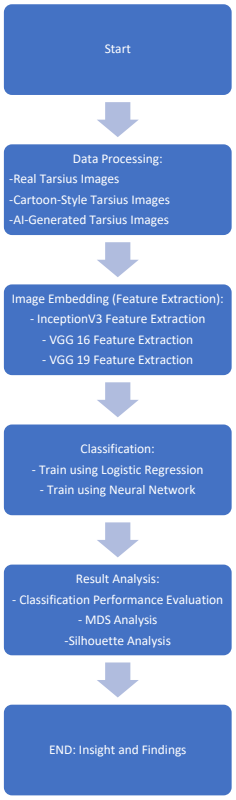


Figure 2. Research Methodology Flowchart

Research objective and dataset preparation

The experiment was defined as a three-category image-source classification task using the 300-image dataset. The labels represented real Tarsius images, cartoon-style Tarsius images, and AI-generated Tarsius images; they did not represent different Tarsius species.

Data processing

The classification experiment was conducted using Orange Data Mining and the Image Analytics add-on. The 300 images were placed in three category folders, with folder names used as class labels by the Import Images widget in Orange Data Mining. File readability and label-folder correspondence were checked before embedding. Three parallel workflows were developed using VGG-16, VGG-19, and InceptionV3 as pretrained feature extractors. The Image Embedding widget then applied the preprocessing required by each pretrained architecture, including conversion to RGB where required, architecture-specific input resizing, and normalization. These operations were executed by Orange Data Mining, using Logistic Regression and Neural Network classifiers. Test and Score was used to assess predictive performance, while confusion matrices, multidimensional scaling, and silhouette plots were used to examine classification errors and feature-space structure. Figure 3 presents the complete experimental workflow.

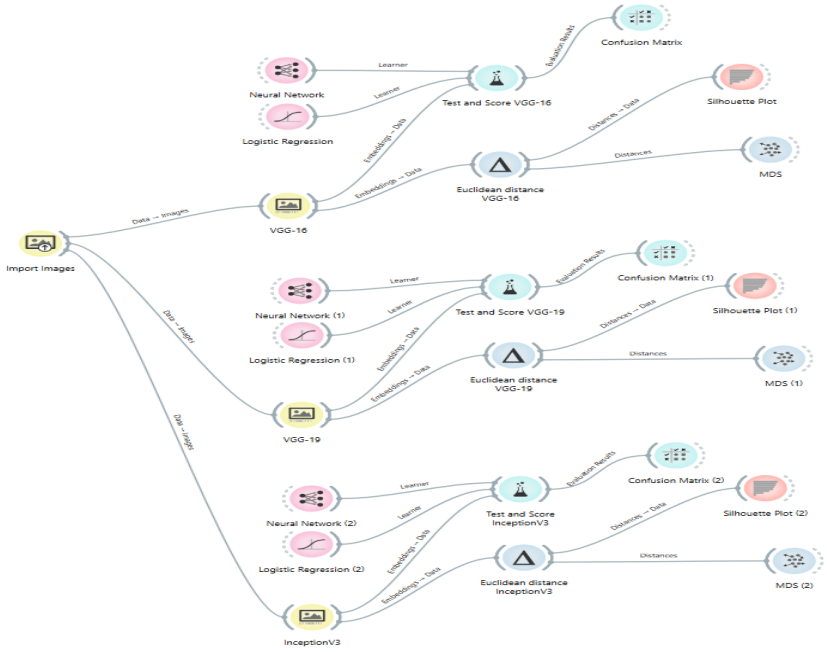


Figure 3. Orange Data Mining workflow for Tarsius image classification using VGG-16, VGG-19, and InceptionV3 feature extraction

Figure 3 provides a visual representation of the Orange Data Mining workflow. As shown in Figure 3, the same image dataset and analytical procedure were applied to all three feature extractors. This parallel design supported a consistent comparison of the model configurations. To reduce leakage risk, evaluation should keep each original image in only one validation fold; however, the original executable workflow file was not retained.

Image embedding

VGG-16, VGG-19, and InceptionV3 were used as pretrained feature extractors. Each architecture transformed an image into a numerical feature vector representing visual properties such as shape, texture, and color distribution. The vectors, rather than raw pixels, were supplied to the classifiers. The experiment relied on the architecture implementations available in the Orange Image Analytics workflow; the original software version, layer-output selection, and exported configuration were not preserved, so undocumented parameter values are not asserted. This reporting choice avoids introducing settings that cannot be verified from the original study records (Crowley, 2023; Simonyan & Zisserman, 2015; Szegedy et al., 2016).

Logistic Regression and Neural Network

The embeddings generated by VGG-16, VGG-19, and InceptionV3 were classified using Logistic Regression and Neural Network models in Orange Data Mining. Each feature extractor was evaluated with the same classifier settings to ensure a consistent comparison across the six architecture–classifier combinations. Logistic Regression estimated class membership through a linear probabilistic decision function, whereas the Neural Network captured nonlinear relationships within the extracted feature representations (Goodfellow et al., 2016; Hosmer et al., 2013; Yousef & Allmer, 2023)

The Logistic Regression classifier used ridge regularization with an L2 penalty. The inverse regularization strength was set to $C = 1$. Class-distribution balancing was disabled because the dataset was balanced, with 100 images in each category: AI-generated, real Tarsius, and cartoon-style images. The same Logistic Regression configuration was applied separately to the embeddings produced by VGG-16, VGG-19, and InceptionV3.

The Neural Network classifier consisted of one hidden layer containing 100 neurons. It used the rectified linear unit function as its activation function and Adam as the optimization solver. The regularization coefficient was set to $\alpha = 0.0001$, while the maximum number of training iterations was limited to 200. The reproducible training option was enabled to reduce variation caused by model initialization. These settings were maintained across the three feature-extraction architectures.

Both classifiers were evaluated using stratified 10-fold cross-validation in the Test and Score widget. The stratified procedure maintained the relative distribution of the three categories in each fold. Model performance was compared using AUC, classification accuracy, F1-score, precision, recall, and Matthews Correlation Coefficient. Confusion matrices were subsequently generated from the cross-validation predictions to examine category-specific classification errors.

Result analysis

Performance was compared using AUC, classification accuracy (CA), F1-score, precision, recall, and MCC. CA summarizes the proportion of correct predictions; precision reflects the reliability of positive predictions; recall measures recovery of actual class members; F1-score balances precision and recall; and MCC provides a correlation-based summary that is informative beyond accuracy alone (Chicco & Jurman, 2020, 2023; Goutte & Gaussier, 2005). MDS projected pairwise relationships among embeddings into two dimensions for visual inspection (Kruskal and Wish ,1978) and (Hagele et al., 2023), while silhouette values described within-category cohesion relative to between-category separation (Bartlett & Dorribo Camba, 2023; Nevill et al., 2023; Rousseeuw, 1987). The best predictive model was selected primarily from the tabulated CA, F1-score, precision, recall, MCC, and AUC values. MDS and silhouette plots were used as complementary diagnostics; consequently, visually clearer MDS separation alone did not override stronger predictive metrics. Pairwise distances between image embeddings were calculated using normalized Euclidean distance. The resulting distance matrices were used for MDS visualization and silhouette analysis. Silhouette values were grouped according to the category variable to evaluate the cohesion and separation of AI-generated, real Tarsius, and cartoon-style images.

Interpretation

Findings were synthesized across predictive metrics, confusion matrices, MDS, and silhouette plots. Conclusions are limited to the three image-source categories and the available 300-image dataset; they do not establish taxonomic species identification or field-level generalizability.

RESULTS AND DISCUSSION

Results

Table 1 reports six combinations of two classifiers and three pretrained feature extractors. Performance is summarized using AUC, CA, F1-score, precision, recall, and MCC.

Table 1
Performance of Classification Models Based on Feature Extraction Architectures

Model	Feature Extraction	AUC	CA	F1	PREC	RECALL	MCC
Logistic	VGG-16	0.996	0.967	0.967	0.967	0.967	0.950
Regression	VGG-19	0.997	0.967	0.967	0.968	0.967	0.950
	InceptionV3	0.997	0.957	0.957	0.957	0.957	0.935
Neural Network	VGG-16	0.983	0.947	0.946	0.947	0.947	0.920
	VGG-19	0.997	0.950	0.950	0.951	0.950	0.925
	InceptionV3	0.996	0.957	0.957	0.957	0.957	0.935

For Logistic Regression, VGG-19 produced AUC = 0.997, CA = 0.967, F1 = 0.967, precision = 0.968, recall = 0.967, and MCC = 0.950. VGG-16 reached the same CA, F1, recall, and MCC, with AUC = 0.996 and precision = 0.967. InceptionV3 produced AUC = 0.997, CA = 0.957, F1 = 0.957, precision = 0.957, recall = 0.957, and MCC = 0.935.

For the Neural Network, InceptionV3 produced the highest CA, F1-score, precision, recall, and MCC within that classifier (0.957, 0.957, 0.957, 0.957, and 0.935, respectively), with AUC = 0.996. VGG-19 produced AUC = 0.997, CA = 0.950, F1 = 0.950, precision = 0.951, recall = 0.950, and MCC = 0.925; VGG-16 yielded the lowest Neural Network scores (AUC = 0.983; CA = 0.947; MCC = 0.920).

On the tabulated predictive metrics, Logistic Regression with VGG-19 was selected as the best-performing configuration because it tied for the highest CA (0.967), F1-score (0.967), recall (0.967), and MCC (0.950), while also attaining AUC = 0.997 and the highest reported precision (0.968). Logistic Regression with VGG-16 was nearly equivalent but had slightly lower AUC and precision.

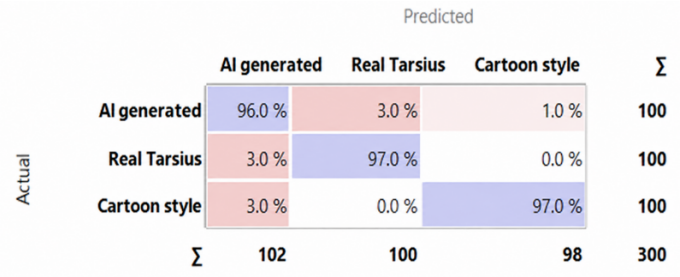


Figure 4. Confusion Matrix Results (VGG-16)

Figure 4 presents the confusion matrix for Logistic Regression with VGG-16 across the three balanced image-source categories.

For AI-generated Tarsius images, 96% were classified correctly, while 3% were assigned to the real category and 1% to the cartoon-style category. For real Tarsius images, 97% were classified correctly and 3% were assigned to the AI-generated category. The reciprocal errors are consistent with visual similarity between photographic and synthetic representations.

For cartoon-style Tarsius images, 97% were classified correctly and 3% were assigned to the AI-generated category. The limited confusion with the synthetic category may reflect shared stylization, whereas no reported cartoon-style samples were assigned to the real category.

The VGG-16 confusion matrix therefore supports the aggregate CA of 0.967 in Table 1 and identifies the real-versus-AI-generated boundary as the principal source of error.

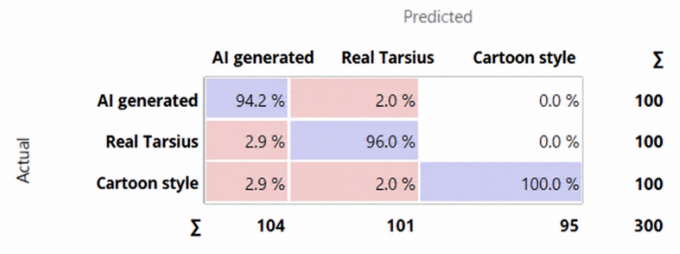
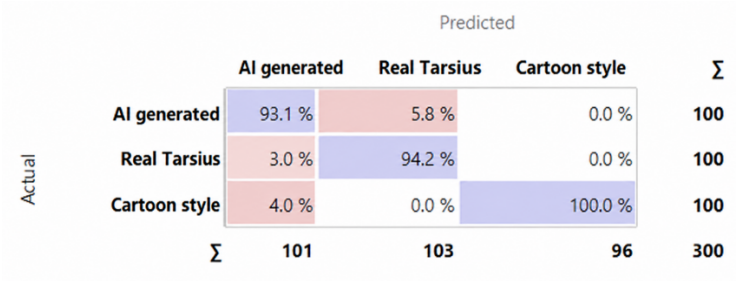


Figure 5. Confusion Matrix Results (VGG-19)

Figure 5 presents the confusion matrix for Logistic Regression with VGG-19. Its interpretation is based on the displayed matrix and the aggregate metrics in Table 1; with the reported CA of 0.967.

The VGG-19 results show limited cross-category error and support its selection as the strongest tabulated configuration. The main remaining confusion occurs between real and AI-generated Tarsius images, while cartoon-style images are more distinctly represented. This pattern is interpreted comparatively rather than described as perfect classification.



Actual \ Predicted	AI generated	Real Tarsius	Cartoon style	Σ
AI generated	93.1 %	5.8 %	0.0 %	100
Real Tarsius	3.0 %	94.2 %	0.0 %	100
Cartoon style	4.0 %	0.0 %	100.0 %	100
Σ	101	103	96	300

Figure 6. Confusion Matrix Results (InceptionV3)

Figure 6 presents the confusion matrix for Logistic Regression with InceptionV3 across the three image-source categories.

For AI-generated Tarsius images, the displayed matrix reports 93.1% correct classification and 5.8% assignment to the real category. For real Tarsius images, 94.2% were classified correctly, with a smaller proportion assigned to the AI-generated category. These errors again locate the main decision boundary between realistic synthetic images and photographs.

Cartoon-style Tarsius images showed the clearest separation in the InceptionV3 confusion matrix. Nevertheless, model selection was based on the complete metric set rather than performance in one category, and InceptionV3–Logistic Regression had a lower aggregate CA and MCC than the two VGG–Logistic Regression combinations.

Across confusion matrices, cartoon-style images were generally easier to distinguish, whereas real and AI-generated Tarsius images showed recurring cross-category errors. These observations motivate closer inspection of feature-space geometry through MDS and silhouette analysis.

Figure 7. MDS Results (VGG-16)

Figure 7 projects the VGG-16 embeddings into two dimensions using MDS. Points close together represent images with relatively similar extracted features; the projection is descriptive and may not preserve every relationship in the original high-dimensional space.

The cartoon-style category forms the most compact and visibly distinct region, whereas real and AI-generated images overlap more strongly. This structure is compatible with the confusion-matrix errors and suggests that VGG-16 captures stylization cues more readily than subtle source differences between realistic synthetic images and photographs.

Because MDS is a two-dimensional projection, overlap should not be treated as a direct error count or as proof that the full feature vectors are inseparable. It instead provides qualitative evidence about the proximity of the three categories.

Figure 8. MDS Results (VGG-19)

Figure 8 presents the two-dimensional MDS projection of VGG-19 embeddings. The distribution is denser and more complex than the VGG-16 projection, indicating heterogeneous feature relationships within the available dataset.

Real and AI-generated Tarsius images remain substantially intermingled, while cartoon-style points are more distinguishable but not entirely isolated in the projection. The overlap may reflect shared visual cues such as eye shape, fur-like texture, lighting, and composition.

The VGG-19 projection therefore does not, by itself, show the clearest visual separation. Its selection as the best predictive configuration instead follows the tabulated Logistic Regression metrics and the comparatively stable silhouette pattern.

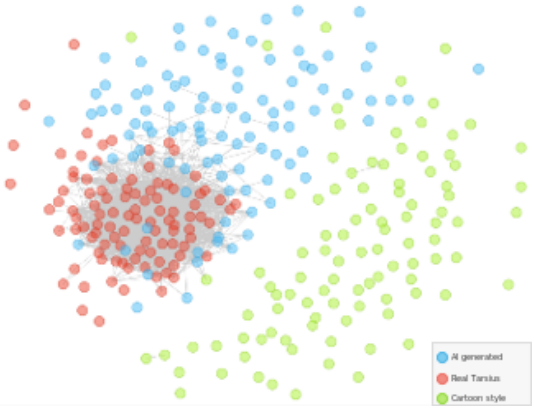


Figure 9. MDS Results (InceptionV3)

Figure 9 presents the InceptionV3 MDS projection. Among the three projections, it shows the clearest separation of cartoon-style images from the other categories.

The pattern suggests that InceptionV3 represents broad differences in contour, color distribution, and illustrative style effectively. This is a descriptive interpretation of the projection rather than evidence that InceptionV3 is the best classifier overall.

Real and AI-generated images remain close and partially mixed in the central region. Thus, clearer global separation in the MDS plot coexists with lower Logistic Regression CA and MCC than the VGG-based configurations.

InceptionV3 provides the clearest two-dimensional category structure, particularly for cartoon-style images, but MDS is not the sole model-selection criterion. This distinction resolves the apparent conflict between the visual projection and the predictive metrics.

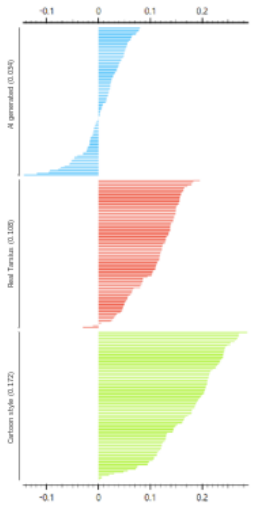


Figure 10. Silhouette Plot Results (VGG-16)

Figure 10 reports silhouette values for the VGG-16 embeddings. Values closer to +1 indicate stronger within-category cohesion and separation; values near zero indicate boundary cases; negative values indicate closer association with another category.

Most VGG-16 silhouette values are positive. The cartoon-style category is comparatively cohesive, while some real and AI-generated samples approach zero, consistent with their overlap in the MDS projection.

The dispersion of real-image values indicates that several samples lie close to the real–AI-generated boundary. The result supports cautious interpretation of the strong aggregate classifier scores because cluster structure is not uniformly separated for every sample.

Taken together, the VGG-16 MDS and silhouette plots show a distinct cartoon-style category and weaker separation between real and AI-generated images.

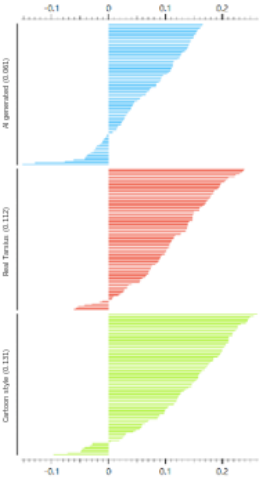


Figure 11. Silhouette Plot Results (VGG-19)

Figure 11 shows that the VGG-19 silhouette values are generally positive and comparatively stable across the three categories. AI-generated images are concentrated in a positive range, although the MDS projection still indicates proximity to real images.

The real-image category shows improved cohesion relative to the VGG-16 plot, with fewer values near the category boundary. This stability complements the strong Logistic Regression metrics reported for VGG-19.

The cartoon-style category retains positive and comparatively stable silhouette values. Accordingly, VGG-19 offers the most balanced combination of strong predictive metrics and stable category cohesion, even though InceptionV3 gives a visually clearer two-dimensional MDS projection.

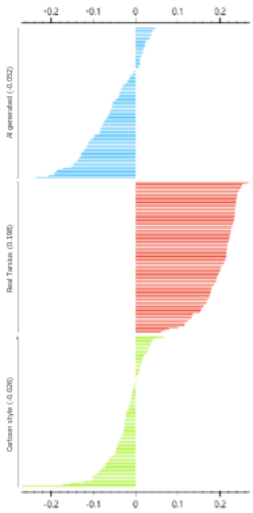


Figure 12. Silhouette Plot Results (InceptionV3)

Figure 12 shows a mixed InceptionV3 silhouette pattern. Several AI-generated samples lie near zero or below it, while the reported category averages are small. The cartoon-style category remains broadly recognizable, but an average silhouette value of approximately 0.026 indicates weak separation for some samples. The reported value of approximately 0.055 for another category is also modest. These small averages explain why clearer visual spacing in an MDS projection does not necessarily translate into the strongest cohesion or predictive metrics.

Discussion

Across the six combinations, Logistic Regression with VGG-19 provides the strongest overall predictive evidence. Its CA of 0.967 and MCC of 0.950 equal the highest values in the table, while its precision of 0.968 is the highest reported. VGG-16 with Logistic Regression is practically close, differing only slightly in AUC and precision. The Neural Network does not improve the aggregate scores, suggesting that the extracted embeddings are already sufficiently structured for a linear probabilistic decision boundary in this dataset. This finding aligns with the broader rationale for combining pretrained deep features with classical classifiers on limited datasets (Amjoud & Amrouch, 2022; Samudra et al., 2023).

Within these boundaries, the study provides a useful methodological finding: strong aggregate classification can coexist with visible feature-space overlap. Combining predictive metrics with MDS and silhouette diagnostics makes that tension explicit and discourages model selection from a single graph or score. For the available data, Logistic Regression with VGG-19 is the most balanced configuration, while the recurring real-versus-AI-generated confusion identifies the principal direction for subsequent feature engineering and validation.

Future validation should use a larger, provenance-controlled dataset with source URLs or collection metadata, explicit permission or licensing records, perceptual duplicate detection, and independent expert review. Field-camera photographs from multiple locations, seasons, devices, and lighting conditions should be separated by source or capture event before cross-validation to prevent leakage. AI-generated images should be stratified by generator and prompt family, with external test sets from unseen generators. Reporting the Orange workflow, software versions, random seeds, preprocessing choices, classifier hyperparameters, and fold assignments would also make comparisons reproducible.

The image-source labels also define a narrow task. The 'real' category refers to photographs presented as real Tarsius images in the collected dataset, not field-verified observations linked to species, location, sex, age, or camera metadata. Similarly, the AI-generated category may vary substantially across generation systems and prompts, none of which are documented here. A model could therefore learn compression artifacts, backgrounds, watermarks, or style conventions instead of biologically meaningful characteristics. The reported metrics should not be interpreted as evidence of species identification performance or readiness for conservation deployment.

Several limitations constrain generalization. First, the dataset contains only 300 images and was assembled from Google search results rather than a controlled field protocol. Search-

engine ranking, web availability, editing, compression, and uncertain provenance may shape the sample. Second, the retained materials do not include source URLs, retrieval dates, licensing metadata, a duplicate-image audit, or expert confirmation of every label. Third, the classifier configurations and stratified 10-fold cross-validation procedure were documented. However, the Orange Data Mining version, Image Analytics add-on version, exact validation-fold assignments, and duplicate-screening records were unavailable. These omissions prevent exact replication and leave open the possibility of near-duplicate leakage or source-specific shortcuts.

MCC is important in this interpretation because it summarizes agreement between predictions and labels and is less easily overstated than accuracy alone (Chicco & Jurman, 2020, 2023). The balanced category sizes reduce one common source of misleading accuracy, yet they do not eliminate sampling bias. The near-equivalence of VGG-16 and VGG-19 with Logistic Regression also indicates that the reported advantage of VGG-19 is modest rather than absolute. Claims of superiority should therefore be limited to this dataset and evaluation output.

The comparison with previous research highlights the study's specific contribution. Samudra et al. (2023) demonstrated that pretrained VGG-16 features can support conventional animal classification, whereas the present results compare two VGG depths and InceptionV3 across both Logistic Regression and Neural Network classifiers. Pearline et al. (2019), Tan et al. (2021) and Shoaib et al. (2022) likewise show that learned image representations can support biological image classification, but their domains and category definitions differ. The current contribution is not a new taxonomic identification model; it is evidence that feature-extractor choice changes the discrimination of real, cartoon-style, and AI-generated Tarsius image sources.

Silhouette analysis adds a second complementary perspective. Positive and stable silhouette values indicate that samples are closer to their assigned category than to alternatives (Rousseeuw, 1987; Saito et al., 2023). VGG-19 shows the most balanced cohesion pattern across categories, which supports, but does not independently prove, its selection from the classification metrics. InceptionV3 includes values near zero or negative for some samples and small reported category averages, indicating boundary ambiguity. These findings demonstrate that predictive accuracy and unsupervised category geometry can agree in broad pattern while differing in detail.

The results also clarify why InceptionV3 should not be selected solely because its MDS projection appears more separated. MDS compresses high-dimensional relationships into two dimensions and is primarily a visualization of relative proximity (Hagele et al., 2023; Kruskal & Wish, 1978). A visually clean projection can emphasize global structure while discarding local relationships relevant to classification. InceptionV3 separates the cartoon-style region clearly, but its Logistic Regression CA (0.957) and MCC (0.935) remain below the corresponding VGG-19 values. Model selection therefore prioritizes the complete predictive metric set, with MDS used to explain feature geometry rather than to rank classifiers independently.

The most persistent error pattern occurs between real and AI-generated Tarsius images. This is scientifically plausible because contemporary synthetic images can reproduce photographic cues (including fur-like texture, lighting gradients, facial symmetry, and background depth) while both categories depict the same animal concept. By contrast, cartoon-style images often use simplified contours, flatter color regions, and non-natural proportions, producing a more distinctive representation. The confusion matrices, MDS projections, and silhouette plots converge on this interpretation, although their numerical and visual outputs describe different aspects of performance.

CONCLUSION

This study met its objective of comparing VGG-16, VGG-19, and InceptionV3 embeddings with Logistic Regression and Neural Network classifiers for real, cartoon-style, and AI-generated Tarsius images. Logistic Regression with VGG-19 was the best-performing reported configuration because it combined AUC = 0.997, CA = 0.967, F1 = 0.967, precision = 0.968, recall = 0.967, and MCC = 0.950. Logistic Regression with VGG-16 was nearly equivalent. Cartoon-style images were generally the most distinguishable, whereas real and AI-generated images showed recurring confusion and overlapping feature representations. InceptionV3 produced the clearest two-dimensional MDS separation, particularly for cartoon-style images, but its lower aggregate predictive metrics and mixed silhouette pattern did not support selecting it as the best classifier.

The principal contribution is a multi-perspective comparison showing that predictive

metrics, MDS, and silhouette analysis answer related but non-identical questions. The findings are limited by the small web-sourced dataset, incomplete provenance and configuration records, possible duplicate or source bias, and the absence of field-based external validation. Future work should expand the dataset, document all sources and hyperparameters, remove near-duplicates before data partitioning, use source-grouped validation, and test field images and images from unseen AI generators. Accordingly, the present results support comparative image-source classification within this dataset rather than taxonomic species identification or operational conservation deployment.

ACKNOWLEDGEMENT

The authors acknowledge the institutional and technical support that facilitated completion of this study. No specific external funder or sponsor was reported in the available manuscript records.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed to the study conception, methodology, data processing, analysis, manuscript preparation, and critical revision, and approved the final version.

REFERENCES

Amjoud, A. B., & Amrouch, M. (2022). Transfer learning for automatic image orientation detection using deep learning and logistic regression. *IEEE Access*, *10*, 128543–128553. https://doi.org/10.1109/ACCESS.2022.3225455

Bartlett, K. A., & Dorribo Camba, J. (2023). The role of a graphical interpretation factor in the assessment of Spatial Visualization: A critical analysis. *Spatial Cognition and Computation*, *23*(1). https://doi.org/10.1080/13875868.2021.2019260

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., & Jurman, G. (2023). The Matthews Correlation Coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1). https://doi.org/10.1186/s13040-023-00322-4

Crowley, J. L. (2023). Convolutional Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13500 LNAI*. https://doi.org/10.1007/978-3-031-24349-3_5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Goutte, C., & Gaussier, É. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *Advances in Information Retrieval: 27th European Conference on IR Research (ECIR 2005)* (Vol. 3408, pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Gursky-Doyen, S. (2010). The function of scentmarking in spectral tarsiers. In S. Gursky-Doyen & J. Supriatna (Eds.), *Indonesian Primates* (pp. 359–369). Springer. https://doi.org/10.1007/978-1-4419-1560-3_20

Hagele, D., Krake, T., & Weiskopf, D. (2023). Uncertainty-Aware Multidimensional Scaling. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1). https://doi.org/10.1109/TVCG.2022.3209420

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175

Hidayatik, N., Agil, M., Iskandar, E., Farajallah, D., Khairullah, A., Saputro, S., & Maulana, V. (2025). Behavior of female Tarsius spectrumgurskyae at the primate research center breeding facility. *Open Veterinary Journal*, *15*(5), 2059. https://doi.org/10.5455/OVJ.2025.v15.i5.22

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

IUCN. (2025). The IUCN Red List of Threatened Species (Versión 2025-1). *Iucn*.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE Publications.

https://us.sagepub.com/en-us/nam/multidimensional-scaling/book4259

Nevill, C. R., Cooper, N. J., & Sutton, A. J. (2023). A multifaceted graphical display, including treatment ranking, was developed to aid interpretation of network meta-analysis. *Journal of Clinical Epidemiology*, *157*. https://doi.org/10.1016/j.jclinepi.2023.02.016

Pearline, S. A., Kumar, V. S., & Harini, S. (2019). A study on plant recognition using conventional image processing and deep learning approaches. *Journal of Intelligent & Fuzzy Systems*, *36*(3), 1997–2004. https://doi.org/10.3233/JIFS-169911

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saito, Y., Omae, Y., Nagashima, K., Miyauchi, K., Nishizaki, Y., Miyazaki, S., Hayashi, H., Nojiri, S., Daida, H., Minamino, T., & Okumura, Y. (2023). Phenotyping of atrial fibrillation with cluster analysis and external validation. *Heart*, *109*(23). https://doi.org/10.1136/heartjnl-2023-322447

Samudra, J. T., Rosnelly, R., Situmorang, Z., & Ramadhan, P. S. (2023). Model klasifikasi jenis hewan dengan SVM, KNN, Logistic Regression menggunakan pre-trained VGG 16. *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer)*, *22*(2), 225–231. https://doi.org/10.53513/jis.v22i2.8314

Shekelle, M., & Salim. A. (2008). Tarsius tarsier. The IUCN Red List of Threatened Species 2008. In *IUCN Red List of Threatened Species*. https://doi.org/10.2305/IUCN.UK.2008.RLTS.T21491A9288932.en

Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S. M., Ali, F., & Park, S. H. (2022). Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1031748

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations. *International Conference on Learning Representations (ICLR)*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tan, L., Lu, J., & Jiang, H. (2021). Tomato leaf diseases classification based on leaf images: A comparison between classical machine learning and deep learning methods. *AgriEngineering*, *3*(3), 542–558. https://doi.org/10.3390/agriengineering3030035

Yousef, M., & Allmer, J. (2023). Deep learning in bioinformatics. *Turkish Journal of Biology*, *47*(6). https://doi.org/10.55730/1300-0152.2671

Zhu, H., Wang, Y., & Fan, J. (2022). IA-Mask R-CNN: Improved Anchor Design Mask R-CNN for Surface Defect Detection of Automotive Engine Parts. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136633