scientific reports

Check for updates

OPEN

Bulldogs stenosis degree classification using synthetic images created by generative artificial intelligence

Gustavo da Silva Andrade[2]✉, Gabriel Toshio Hirokawa Higa[1], Jarbas Felipe da Silva Ribeiro[1], Joyce Katiuccia Medeiros Ramos Carvalho[1,3], Wesley Nunes Gonçalves[2], Marco Hiroshi Naka[1,4] & Hemerson Pistori[1,2]

Nasal stenosis in bulldogs significantly impacts their quality of life, making early diagnosis crucial for effective treatment. This study developed an automated deep learning model to classify the severity of nasal stenosis using 1020 images of bulldog nostrils, including both real and AI-generated samples. Five neural network architectures were tested across three experiments, with DenseNet201 achieving the highest median F-score of 54.04%. The model's performance was directly compared to trained human evaluators specializing in veterinary anatomy, achieving comparable levels of accuracy and reliability. These results demonstrate the potential of advanced neural networks to match human-level performance in diagnosis, paving the way for enhanced treatment planning and overall animal welfare.

Keywords Brachycephalic dogs, Stenotic nares, Airway obstruction, Brachycephalic obstructive airway syndrome, Computer vision, Deep learning

Brachycephalic obstructive airway syndrome (BOAS) is a complex respiratory disease that affects dogs with flattened facial and cranial features, such as pugs, French bulldogs and English bulldogs[1,2]. The conformation of these breeds, with shortened skulls and flattened snouts, leads to deformations in the upper respiratory tract, resulting in obstruction and a series of clinical problems, ranging from noisy breathing and difficulty to exercise, up to syncope, collapse, and other potentially fatal events[3].

The occurrence of BOAS has increased significantly over the last few decades, partly due to the growing popularity of these breeds. Recent studies have attempted to identify specific conformational characteristics associated with BOAS, using non-invasive measures that can be applied in a practical way to help breeders select animals with a lower risk of developing the syndrome[1,4].

In addition, the severity of BOAS is often assessed through exercise tests, which measure the dog's ability to exercise without showing signs of respiratory distress. Such tests can be particularly useful for assessing the effectiveness of surgical interventions aimed at relieving the symptoms of BOAS[1].

Nasal stenosis can be considered one of the symptoms of BOAS. It refers to the narrowing of the nostrils, which restricts the flow of air through the nasal airways[2]. When present, this narrowing can vary between mild, moderate and severe as shown in Fig. 1. In dogs with BOAS, the narrowed nostrils significantly compromise their ability to inhale enough air through the nasal passages, often forcing them to breathe through their mouths, which is neither the natural nor the most efficient method for canine breathing[5].

Stenosis prevents adequate airflow, increasing breathing effort and exacerbating other components of BOAS, such as stretching of the soft palate, which may already be partially obstructing the laryngeal inlet[6]. This increased effort to breathe can lead to inflammation and additional edema in the airways, making respiratory symptoms even worse. In addition, forced and labored breathing contributes to the development of secondary changes in the airways, such as laryngeal collapse and tracheal changes[2,6].

The management of nasal stenosis in dogs with BOAS often involves surgical approaches to widen the nostrils, a technique known as stenotic nostril widening, which can significantly improve the quality of life of

[1]Universidade Católica Dom Bosco, Campo Grande, Brazil. [2]Universidade Federal de Mato Grosso do Sul, Campo Grande, Brazil. [3]Clínica de Odontologia Veterinária - OdontoPet, Campo Grande, Brazil. [4]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul, Campo Grande, Brazil. ✉email: gustavo.s.andrade@ufms.br

Fig. 1. Three different degrees of stenosis from our dataset of real images, ranging from mild (**a**), moderate (**b**) and severe (**c**).

affected dogs. In addition, careful lifestyle management, such as avoiding excessive heat, controlling weight and limiting exercise, is essential to minimize the risks associated with the condition[5].

This connection between nasal stenosis and BOAS illustrates how anatomical features inherited through specific breeding practices can predispose certain breeds to complex and challenging conditions, emphasizing the need for awareness and careful considerations in the management and breeding of these popular breeds[2,4].

The use of DL and AI, together with computer vision, has revolutionized many areas, including animal health[7]. In recent years, the growth of this area has been remarkable, with applications ranging from the automatic identification of physiological characteristics to the behavioral and welfare assessment of animals[8], it also offers new possibilities for monitoring, diagnosing and treating diseases in animals through the advanced analysis of images and behavioral data. In addition, it allows the automation of complex diagnostic tasks that traditionally rely on human interpretation, offering comparable or even superior accuracy in some cases[9].

Several studies have demonstrated the potential of artificial intelligence in transforming animal health management. For example, automatic animal recognition techniques and the extraction of specific physiological characteristics, such as breathing rate and heart rate from images, show that it is possible to obtain vital data non-invasively and continuously, which was previously only possible through specialized equipment and physical contact[8].

In recent years, the use of text-to-image models has become a great ally for data augmentation. For example, in Bahani[10], the use of this type of tool was able to improve Recall by 2.1%, Specificity by 1.9% and smooth out overfitting during dataset training. In addition, a major advantage of using generative models is the possibility to generate data for experiments and storage that does not endanger individual safety[11].

In an earlier paper by Higa et al.[12], it was argued that the use of neural networks to classify the degree of stenosis is complex but possible. The main indication for this was the maximum median f-score of 53.77% obtained using the MobileNetV3 architecture, when treated as a multi-class problem. In this sense, this paper proposes a new approach to improve these results.

Using an innovative combination of synthetic and real images, the work aims to improve the performance of neural networks for classifying and diagnosing the level of stenosis in bulldogs, a critical step given the prevalence of respiratory problems in this breed due to its brachycephalic nature. The inclusion of a diverse data set, comprising a set of synthetic images and a mixed set, with real and synthetic images, addresses the significant challenge of data scarcity and increases the model's ability to generalize across different clinical presentations. This methodological choice not only strengthens the training process, but also tests the model's performance in a variety of conditions, more closely emulating real-world scenarios.

Leveraging sophisticated neural network architectures such as ResNet50, MobileNetV3, DenseNet201, SwinV2, and MaxViT, this research stands at the forefront of applying deep learning in veterinary science. Each of these models brings unique strengths in handling image data, which is crucial for accurately capturing and learning from the nuanced differences in bulldogs's nasal anatomies.

In addition to the aforementioned neural networks, we also employed GPT-4o, a model developed by OpenAI, to classify the images. Although GPT-4o is primarily a language model, its multimodal capabilities allowed us to utilize it for image classification tasks. We further enlisted veterinary students to classify the same images. Regardless of the classification approach, all tests were carried out using the same experimental protocol.

We emphasize that our model, despite its F-score of 54.04%, performs comparably to human evaluators and surpasses large multimodal AI models while using significantly fewer computational resources. This suggests that our method provides a strong starting point for future research. The primary contribution of this study is to introduce a novel problem for the scientific community to explore, particularly in the field of computer science, paving the way for advancements in AI-driven veterinary diagnostics.

Despite the promising approach, the work faces challenges such as ensuring the synthetic data's realism, balancing the influence of synthetic versus real images to prevent model bias, and managing the computational demands of processing extensive datasets with complex models. Overcoming these challenges is essential for the successful deployment of these technologies in clinical settings, where they can potentially transform the diagnostics landscape by providing quick, reliable, and non-invasive diagnosis tools. This work not only highlights the potential of artificial intelligence in enhancing animal welfare but also sets a precedent for future research in the domain of veterinary medical imaging, promising substantial improvements in the accuracy and efficiency of diagnosing and treating animal diseases.

In summary, the main contributions of this paper are threefold: a new dataset specifically designed for experiments on nasal stenosis diagnosis in bulldogs, addressing the scarcity of available data in veterinary medicine for this condition; we present the first comparative analysis between human performance and

deep neural network performance in classifying the degrees of stenosis, providing valuable insights into the capabilities and limitations of both human experts and AI models; and report initial findings using state-of-the-art foundational models, such as GPT-4o and DALL-E, which demonstrate the complexity of the problem and highlight the challenges and potentials of applying advanced AI technologies to veterinary diagnostics.

Results

Classifying the degree of nasal stenosis in bulldogs presents significant challenges, not only for computational models but also for trained human evaluators. The subtle morphological variations between different levels of stenosis often make consistent evaluation difficult, even among individuals with specialized training. This complexity underscores the importance of developing robust models that can assist in this nuanced task.

Tables 1, 2 and 3 show precision, recall and f-score results for each neural network architecture trained with different dataset configurations. It also shows the results of the Scott–Knott clustering test. The results are further illustrated by boxplots in Fig. 4. Despite the inherent difficulties, our models achieved convincing performance levels. Notably, the combined dataset led to the highest average results across all three metrics, with DenseNet201 achieving a precision of 61%, a recall of 58%, and an F-score of 56%. For a better understanding of this architecture's performance, accuracy and loss were calculated using the complete history of 10 training and validation runs of the DenseNet201 model. The loss and accuracy curves are shown in Fig. 2. These results indicate that incorporating both real and synthetic images enhances the model's ability to generalize and accurately classify the degree of nasal stenosis.

Interestingly, while MobileNetV3 trained solely on real images achieved the highest median precision and recall, the highest average results were generally achieved with the combined dataset. This suggests that synthetic images contribute valuable variability to the training process, helping the model better capture the subtle differences between stenosis levels. In some cases, using the synthetic dataset alone led to better results than using the real dataset, such as with MaxViT across all three metrics and ResNet50 in recall and F-score. However, the combined dataset consistently outperformed the others, highlighting the benefits of a diversified training set.

Figure 3 shows the confusion matrix for DenseNet201 trained with the combined image set. The dataset is not heavily imbalanced, allowing for a straightforward interpretation of the normalized matrix. The highest percentage of correct predictions is 28%, corresponding to severe stenosis examples correctly classified. Among the incorrect predictions, 12% of the images are examples of moderate stenosis misclassified as severe. This highlights the difficulty in distinguishing between adjacent classes due to subtle morphological differences.

To enhance the interpretability of the confusion matrix, we included both absolute values (number of nostrils) and normalized percentages within a single figure. The absolute values represent the exact number of analyzed nostrils, reinforcing that the total corresponds to 190 samples (from real images). The normalized values ensure an intuitive understanding of classification performance relative to the dataset distribution.

The ANOVA results for precision indicated that there was a difference both for architectures ($p = 2.0 \times 10^{-9}$) and for training sets ($p = 1.2 \times 10^{-5}$). It also indicated a significant interaction ($p = 4.1 \times 10^{-5}$). Regarding

Precision			
Mean SK (SD)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.390 Ba (±0.144)	0.397 Ba (±0.170)	0.386 Aa (±0.243)
DenseNet201	0.512 Aa (±0.120)	**0.612 Aa** (±0.084)	0.325 Ab (±0.060)
MobileNetV3	0.565 Aa (±0.192)	0.358 Bb (±0.178)	0.313 Ab (±0.062)
SwinV2	0.401 Ba (±0.233)	0.311 Ba (±0.092)	0.209 Ab (±0.143)
MaxViT	0.183 Ca (±0.117)	0.303 Ba (±0.016)	0.284 Aa (±0.042)
ResNet50	0.438 Bb (±0.162)	0.548 Aa (±0.045)	0.342 Ab (±0.106)
Humans	**0.572 Aa** (±0.131)	–	–
Median (IQR)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.359 (0.122)	0.408 (0.119)	0.366 (0.297)
DenseNet201	0.514 (0.181)	0.579 (0.077)	0.331 (0.093)
MobileNetV3	**0.629 (0.293)**	0.319 (0.193)	0.320 (0.079)
SwinV2	0.441 (0.417)	0.318 (0.028)	0.115 (0.225)
MaxViT	0.118 (0.103)	0.307 (0.010)	0.273 (0.046)
ResNet50	0.458 (0.171)	0.536 (0.050)	0.316 (0.133)
Humans	**0.590 (0.135)**	–	–

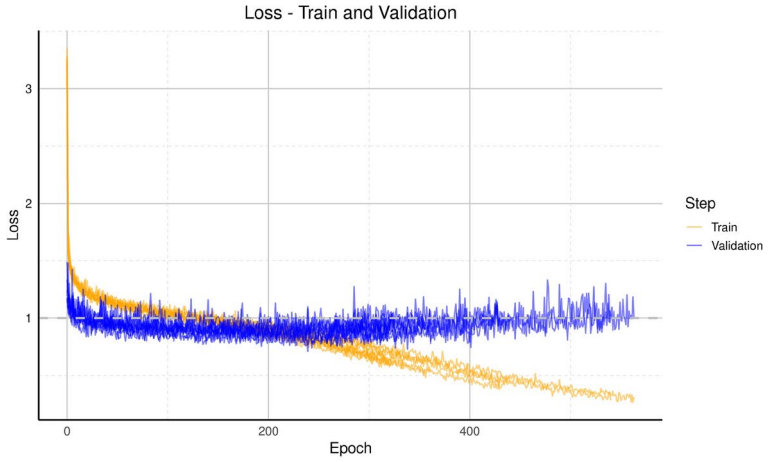
Table 1. Precision statistics. The results of the Scott–Knott test are shown next to the mean values. In each column, mean values indicated by the same capital letters did not differ according to the 5% significance threshold. In each row, mean values indicated by the same lowercase letters did not differ according to the same threshold. Significant values are in bold.

Recall			
Mean SK (SD)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.407 Ba (±0.174)	0.400 Ba (±0.219)	0.386 Ba (±0.271)
DenseNet201	0.512 Aa (±0.075)	**0.581 Aa** (±0.088)	0.489 Aa (±0.079)
MobileNetV3	0.556 Aa (±0.134)	0.453 Aa (±0.066)	0.472 Aa (±0.080)
SwinV2	0.451 Ba (±0.134)	0.380 Ba (±0.053)	0.391 Ba (±0.106)
MaxViT	0.350 Ba (±0.057)	0.459 Ba (±0.024)	0.441 Ba (±0.061)
ResNet50	0.436 Aa (±0.111)	0.547 Aa (±0.029)	0.499 Aa (±0.116)
Humans	**0.580 Aa** (±0.125)	–	–
Median (IQR)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.402 (0.261)	0.431 (0.129)	0.352 (0.329)
DenseNet201	0.533 (0.064)	0.554 (0.119)	0.489 (0.116)
MobileNetV3	**0.572 (0.144)**	0.486 (0.047)	0.461 (0.118)
SwinV2	0.439 (0.231)	0.392 (0.037)	0.333 (0.150)
MaxViT	0.333 (0.017)	0.463 (0.027)	0.439 (0.085)
ResNet50	0.420 (0.078)	0.538 (0.030)	0.494 (0.185)
Humans	**0.600 (0.140)**	–	–

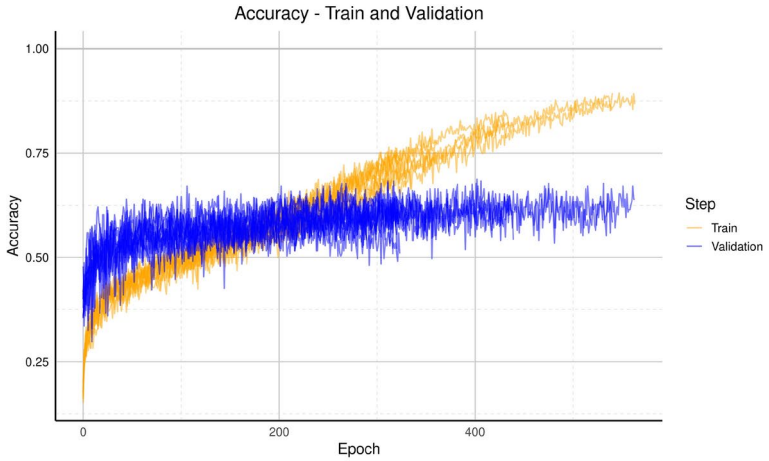
Table 2. Recall statistics. The results of the Scott–Knott test are shown next to the mean values. In each column, mean values indicated by the same capital letters did not differ according to the 5% significance threshold. In each row, mean values indicated by the same lowercase letters did not differ according to the same threshold. Signficant values are in bold.

F-score			
Mean SK (SD)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.365 Ba (±0.146)	0.362 Ba (±0.184)	0.376 Aa (±0.251)
DenseNet201	0.474 Aa (±0.068)	**0.561 Aa** (±0.102)	0.386 Ab (±0.068)
MobileNetV3	0.529 Aa (±0.152)	0.345 Bb (±0.098)	0.367 Ab (±0.068)
SwinV2	0.380 Ba (±0.185)	0.294 Ba (±0.076)	0.255 Aa (±0.132)
MaxViT	0.221 Cb (±0.086)	0.364 Ba (±0.019)	0.338 Aa (±0.045)
ResNet50	0.385 Ba (±0.138)	0.505 Aa (±0.039)	0.397 Aa (±0.108)
Humans	**0.575 Aa** (±0.130)	–	–
Median (IQR)			
	Training set		
Architecture	**Real**	**Combined**	**Synthetic**
GPT-4o	0.333 (0.113)	0.377 (0.151)	0.354 (0.278)
DenseNet201	0.482 (0.088)	**0.540 (0.148)**	0.390 (0.103)
MobileNetV3	0.538 (0.188)	0.385 (0.057)	0.360 (0.084)
SwinV2	0.362 (0.338)	0.297 (0.034)	0.167 (0.230)
MaxViT	0.174 (0.086)	0.367 (0.020)	0.326 (0.053)
ResNet50	0.363 (0.177)	0.492 (0.033)	0.379 (0.152)
Humans	**0.590 (0.138)**	–	–

Table 3. F-score statistics. The results of the Scott–Knott test are shown next to the mean values. In each column, mean values indicated by the same capital letters did not differ according to the 5% significance threshold. In each row, mean values indicated by the same lowercase letters did not differ according to the same threshold. Signficant values are in bold.



(a) Loss Curve



(b) Accuracy Curve

Fig. 2. Loss and accuracy curves calculated during training and validation steps. Since a tenfold cross validation was used, ten lines are shown for each procedure, for both loss and accuracy.

recall, there is no evidence either for difference between training sets ($p = 0.55$) or for a significant interaction ($p = 0.09$). On the other hand, the ANOVA result was significant for architectures ($p = 3.0 \times 10^{-5}$). Finally, f-score results were significant for architectures ($p = 1.8 \times 10^{-7}$) and for the interaction ($p = 2.3 \times 10^{-4}$), but not for the training sets ($p = 0.06$). The specification of the differences by the SK test are shown in Tables 1, 2 and 3.

Discussion

Classifying nasal stenosis in bulldogs presents considerable challenges, even for trained human evaluators and specialists. The subtle morphological features that differentiate the degrees of stenosis make consistent evaluation difficult, emphasizing the need for advanced tools to assist in diagnostics. Our study demonstrates that advanced neural networks can achieve convincing performance levels in this task, offering a promising tool to support veterinary diagnostics.

While it is true that the dataset used by us, composed of 190 nostril images from 95 animals, does not have the usual size expected for training a neural network, it is worth noticing that data collection within this field is a difficult task. Schmid et al.[13], for instance, managed to sample data from 84 french bulldogs for their study. Although their work had different goals and different methods, this allows us to argue that our sample is within the possible working range for the problem. Within deep learning, models trained with small datasets may not generalize well. In order to address this issue, we have used data augmentation with synthetic images generated by DALL-E. Other promising options that could be evaluated in the next steps are the use of few-shot learning and fine grained image classification techniques[14], as well as those of multi-task learning[15,16].

By incorporating both real and synthetic images, our models reached commendable performance metrics. Specifically, our hybrid dataset approach allowed DenseNet201 to achieve a precision of 61%, a recall of 58%, and an F-score of 56%, outperforming configurations that relied solely on one type of data. This blended approach

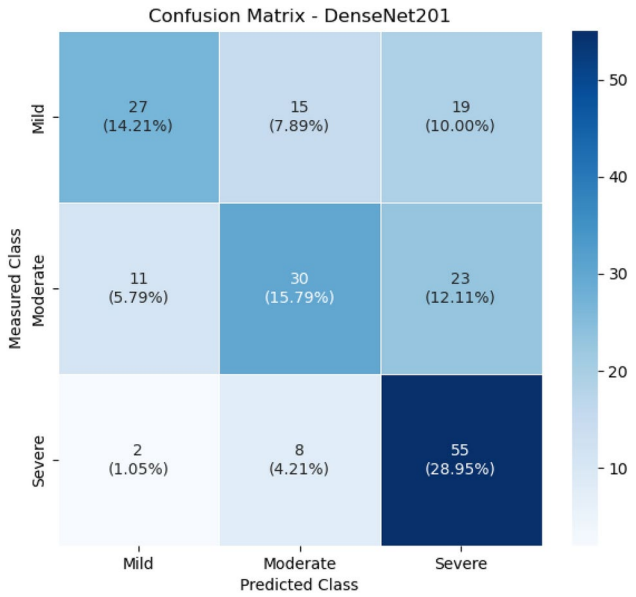


Fig. 3. Confusion matrix for DenseNet201 trained with the combined set. The matrix presents both absolute values (number of nostrils) and normalized percentages based on the total number of real images (190).

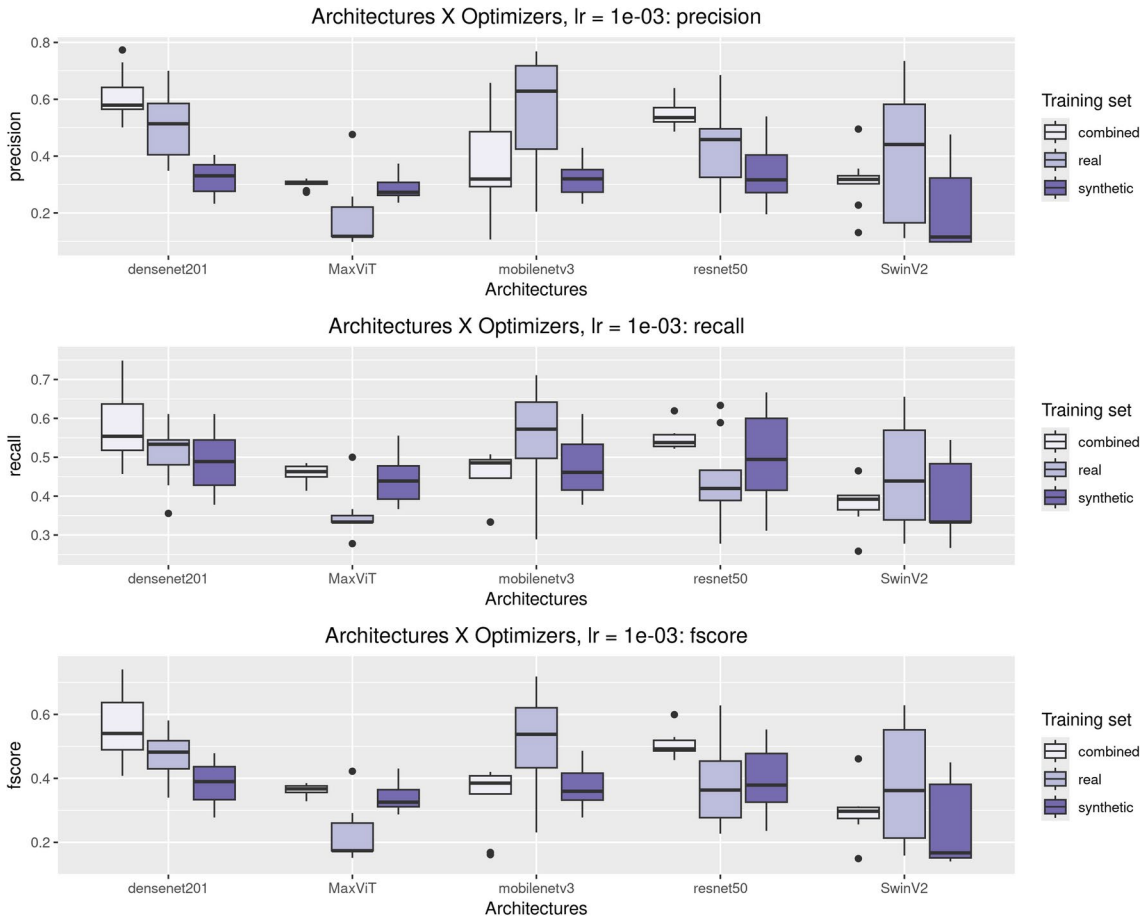


Fig. 4. Boxplots for each metric, for each achitecture evaluated in the different training regimens investigated in this study.

provided a balance between generalization and specificity, showcasing the benefits of a diversified dataset for complex classifications.

The curves in Fig. 2 indicate that training and validation loss and accuracy values followed similar patterns across the folds. In general, training loss and accuracy continued to improve whereas validation did not, which indicates overfitting. This suggests that regularization techniques, such as L1 and L2 regularization, can help improve performance. A promising next step would be to carry out a systematic evaluation of these techniques. Finally, Fig. 2 also shows that both loss function and accuracy stabilize within the first epochs. This suggests that a reduction on the number of epochs, or on the patience, can lead to a more efficient evaluation pipeline for this problem.

Our findings align with those of Higa et al.[12], where predicting severe stenosis was easier for certain neural networks. In their work, 24% of the images were examples of severe stenosis correctly classified, whereas we observed 28% (Fig. 3). This consistency suggests that using a combination of synthetic and real images can be fruitful for binary classification tasks, such as recognizing severe stenosis. In addition, by looking at Fig. 3, it is possible to argue that the model's predictions were, in a sense, pessimistic, since more images were misclassified on the upper right of the main diagonal, where the labeled stenosis degrees are lower than the predicted ones.

Our analysis reveals that while real images anchor the model in clinically applicable contexts, synthetic images generated through DALL-E enrich the dataset by introducing a variety of scenarios. This combination enhances model generalization without sacrificing clinical relevance. The success of this blended approach underscores the advantages of using both real and synthetic data, especially in veterinary applications where data scarcity is often an obstacle.

By comparing the synthetic samples in Fig. 6 with the real ones in Fig. 5, one can see that, while the synthetic images do resemble the real ones, they are still clearly taken from a different distribution. Important differences arguably include textures and lighting. Promising possibilities for future works are to carry out a systematic evaluation of the quality of synthetic images, to test other generative models, and to look for strategies to improve image generation. By doing this, it is possible that using generative models for data augmentation will lead to better-performing models.

Despite the complexities, human evaluators with specialized training achieved notable F-scores and recall rates, underscoring the inherent difficulty of the task and highlighting the potential for human expertise to complement automated systems. However, neural networks like MobileNetV3 still had an edge in precision, likely due to their ability to minimize false positives more effectively.

While GPT-4o did not outperform the other models, its use as a stenosis degree classification tool presents certain pitfalls. As a multimodal foundation model used within commercial software like ChatGPT, it operates as a black box with opaque procedures. This lack of transparency makes it challenging to ascertain how the model processes the data and whether it effectively utilizes the images for classification.

One of the key findings of our study is that our approach, despite using significantly fewer computational resources, was able to outperform GPT-4o in the classification of stenosis severity. This highlights the potential of our method as a baseline for future studies, showing that specialized deep learning models trained with domain-specific data can achieve better results than general-purpose multimodal AI systems. The results reinforce the importance of domain adaptation and the role of dataset quality in improving classification accuracy.

Furthermore, while the F-score of 54.04% may seem low in absolute terms, it is crucial to note that our model performs comparably to trained human evaluators. In clinical applicability, this result suggests that AI-based approaches can support veterinarians in stenosis classification, especially considering that human evaluators also struggle with consistent classifications in complex medical imaging tasks. This reinforces the importance

(a) L and R moderate **(b)** L and R mild **(c)** L open, R mild

Fig. 5. Different examples of images from the real image set. Degree of stenosis in both nostrils of each bulldog: left (L) and right (R).

of continued development in this area to improve diagnostic reliability and practical usability in real-world settings.

According to a recent work by Schmid et al.[13], exercise tests are the most reliable and satisfactory method to diagnose BOAS in french bulldogs. However, their results also reasserted the relationship between nasal stenosis and BOAS. More specifically, moderate and severe stenosis were found to have a significant positive effect on BOAS degree. While exercise tests were more reliable, it is clear that they present risks to the animal, thus requiring safety and comfort protocols. Furthermore, other highly effective methods, such as computer tomography and endoscopy, may be even more risky, since they are more invasive and require anesthesia.

Oren et al.[17] evaluated machine learning models for the diagnosis of BOAS with respiratory sounds. In their work, animals of different brachycephalic races were used. While their approach achieved a maximum f-score of 85% in the determination of BOAS degree with pugs, their procedure involved an exercise test and required multiple audio samplings. Similarly to highly reliable diagnostic methods, their procedure can also be stressful to the animals.

One should notice that these works targeted the condition itself (BOAS). Meanwhile, classifying stenosis degree targets an important symptom of that condition. However, the method proposed by us has the advantage of being neither invasive nor stressful, since ultimately it only requires a picture of the animal's face. While the performance of the model is still not ideal for diagnostics, further work could investigate model performance when classes are separated differently. For instance, a separation of degrees between (i) open-mild and (ii) moderate-severe could lead to acceptable results, while still being potentially useful, given that mild stenosis, according to the findings of Schmid et al.[13], does not have a significant effect on the degree of BOAS.

The findings highlight the feasibility and potential of integrating deep learning into veterinary diagnostics. Advanced neural networks can complement traditional diagnostic methods, making the diagnosis and management of conditions like nasal stenosis in bulldogs faster and more precise. Future research should focus on expanding synthetic image diversity and exploring alternative neural network architectures to further optimize performance metrics. By doing so, we can push veterinary diagnostic tools to new levels of accuracy and accessibility, ultimately improving animal health outcomes.

Methods

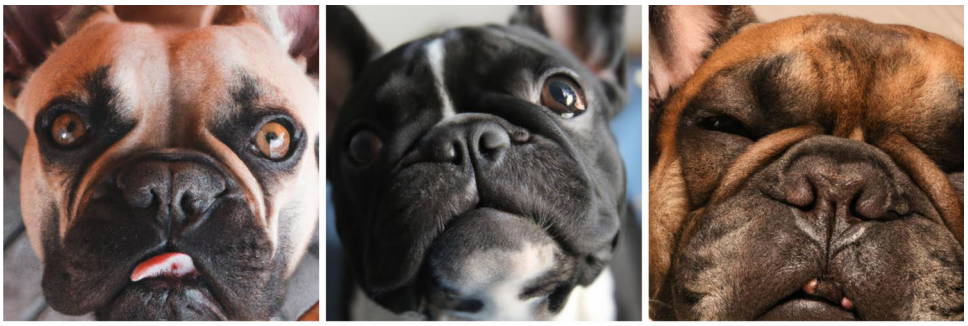
Dataset

For the collection and use of images, we have approval from the Ethics Committee for the Use of Animals of the Dom Bosco Catholic University under protocol number 010/2022. In this work, we utilized the dataset of real images established in our previous project by Higa et al.[12], which comprises 95 images collected between July 2021 and April 2023 in partnership with the OdontoPet clinic in Campo Grande-MS.

In total, 95 images were collected and each nostril was labeled by an experienced veterinarian according to the four degrees of stenosis: open (non-stenotic), moderate, mild and severe. Due to the limited number of non-stenotic nostrils (only three images), these were incorporated into the mild stenosis category for classification purposes. The images were then cropped using the LabelMe tool, isolating the nostril area and assigning the respective stenosis degree to each sample.

Additionally, we created a set of synthetic images categorized alongside the real ones. The synthetic set consists of 415 images generated using OpenAI's Generative Artificial Intelligence, known as DALL-E. These synthetic images were also manually validated to ensure they accurately represented the varying degrees of nasal stenosis.

The 415 images of French Bulldogs were generated using the Generative Artificial Intelligence technology known as DALL-E. DALL-E is a cutting-edge model developed by OpenAI. Using advanced deep learning techniques and neural networks, DALL-E has the ability to transform text descriptions into highly realistic images. By applying a multi-level attention mechanism and reinforcement learning, DALL-E is able to generate a wide variety of high-quality images, including different degrees of severity of stenosis in bulldogs. To generate the images, we gave DALL-E the following command in Portuguese: "Narinas de um Bulldog Francês vistas de frente" ("French Bulldog's nostrils seen from the front"), in its literal translation, as can be seen in Fig. 6. This



(a) L severe, R moderate **(b)** L and R mild **(c)** L and R severe

Fig. 6. Different examples of synthetic images generated for the dataset. Degree of stenosis in both nostrils of each bulldog: left (L) and right (R).

Group	Class			
	Open (A)	Mild stenosis (B)	Moderate stenosis (C)	Severe stenosis (D)
Synthetic	170 (20.48%)	324 (39.03%)	167 (20.12%)	169 (20.36%)
Real	3 (1.58%)	56 (29.47%)	66 (34.74%)	65 (34.21%)

Table 4. Set of images created for the article, with their respective classes annotated.

approach allowed for the creation of a set of images that complements the actual images collected, providing a variety of visual representations for analysis and study[18].

Each image was carefully assessed to determine the degree of stenosis of the nostrils, following criteria established by experts in the field. The images were then annotated using the RoboFlow tool, and according to Liu et al.[2] four classes were classified as open (*i.e.*, non stenotic) (A), mild stenosis (B), moderate stenosis (C) and severe stenosis (D), as can be seen in Table 4. It is important to note that each image contains two nostrils, which are classified individually. Therefore, the total of 510 images (415 synthetic and 95 real) results in 1020 nostril images.

Deep learning

In this work, we used convolutional neural networks (CNN) and transformers, advanced technologies that, together with DL, are becoming increasingly important for study and experimentation in solving complex problems[19,20]. With these approaches, we have obtained good results. CNNs are effective at extracting complex visual features from images, while transformers excel at capturing long-range dependencies and contextual relationships within data[21].

We evaluated five neural network architectures known for their effectiveness in image classification, particularly in medical imaging. ResNet50, MobileNetV3, and DenseNet201, which are based on CNNs, were selected for their strong performance in extracting complex visual features. DenseNet201, in particular, was chosen for its efficiency in feature reuse and gradient propagation, which are crucial for distinguishing the nuanced differences in nasal anatomy. Additionally, SwinV2 and MaxViT, transformer-based architectures, were included to leverage their ability to capture long-range dependencies and contextual relationships within the data.

- *ResNet50* proposed in 2015 by He et al.[22], brought significant innovation with the introduction of jump connections, which help to avoid the problem of the gradient disappearing. This allows networks to become deeper and more effective. ResNet50, in particular, is made up of 50 layers and is widely used not only for image classification, but also as the basis for other computer vision tasks, such as object detection. In this work, we chose ResNet50 both because of its widespread acceptance in the scientific community, which enables our findings to be relevant in a broader context, and to serve as a benchmark against which to evaluate the performance of the other architectures we investigated. ResNet50 is known for its excellent performance and efficiency, facilitating the construction of robust and accurate models for a variety of computer vision applications.
- *MobileNetV3* is a neural network architecture designed for mobile devices and resource-limited environments. It combines depthwise separable convolutions and attention blocks to achieve high efficiency. The first version was released in 2017[23], followed by two updates: MobileNetV2 in 2018[24] and MobileNetV3 in 2019[25], which is the version used in this work. MobileNetV3 is especially suitable for applications where processing capacity and memory are restricted, while maintaining good performance in extracting visual features from images. The choice of MobileNetV3 in this work is due to its efficiency and its ability to operate in environments with limited resources. In addition, it serves as a useful comparison for evaluating the performance of other architectures. Its innovations, such as the use of Squeeze-and-Excitation (SE) attention blocks and optimization through the use of the Neural Architecture Search (NAS) algorithm, contribute to its high efficiency and accuracy in computer vision tasks.
- *DenseNet201* is a neural network architecture that connects each layer to all subsequent layers, promoting feature reuse and improving gradient propagation, proposed in 2017 by Huang et al.[26]. This approach allows DenseNet201 to learn complex representations with a relatively smaller number of parameters, due to its dense structure that facilitates the transmission of information through the network. We chose DenseNet201 in this work because of its ability to improve network efficiency and reduce data redundancy. This feature is particularly useful for applications that require detailed and accurate analysis of visual characteristics. DenseNet201 also stands out for its ability to mitigate common problems in deep networks, such as gradient disappearance, resulting in a more robust and effective network for extracting complex visual features.

The SwinV2 and MaxViT networks are based on transformers, a technology originally introduced for Natural Language Processing (NLP) in 2017[27]. Since the launch of the Vision Transformer (ViT) in 2021[20], several variants and improvements have been developed to overcome the initial limitations of these architectures, with a particular focus on training efficiency and adaptability to datasets of different sizes.

The networks used based on transformers are:

- *SwinV2* Swin Transformer V2, or SwinV2[28], improves on the original design of Swin Transformers by using a hierarchical structure and sliding windows for attention. This approach allows efficient scalability for high-resolution images. SwinV2 is notable for its ability to capture long-range relationships in images, of-

fering a flexible and powerful solution to complex challenges in computer vision. It excels at problems that require the analysis of large, detailed images, making it an important benchmark for comparing the performance of other transform-based architectures. The choice of SwinV2 in this work is justified by its efficiency and scalability when dealing with high-resolution images, characteristics that make it exceptionally suitable for computer vision applications that require precision and detail. The architecture also demonstrates enhanced parallel processing capabilities and adaptability to different dataset sizes, making it a robust choice for a variety of application scenarios.

- *MaxViT*: MaxViT, a recently proposed architecture, effectively combines the techniques of Convolutional Neural Networks (CNNs) and Transformers for computer vision[29]. This architecture is designed to efficiently capture the local and global characteristics of images, facilitating more detailed and contextualized processing. MaxViT is particularly efficient at handling computer vision tasks that require deep understanding of images at varying levels of granularity. MaxViT achieves this through an innovative strategy that interleaves convolution layers with transform modules, optimizing the flow of information and the effectiveness of the model at different resolutions and scales. This allows the architecture to adjust more dynamically to the complexities of visual data, making it ideal for applications such as image recognition, object detection and semantic segmentation.

These innovations not only address some of the initial limitations of transformers in computer vision, but also broaden their applications, showing promising capabilities in various computer vision tasks, from basic image classification to more complex applications such as real-time detection and video analysis. These advanced architectures were selected for our study due to their relevance and potential for dealing with the proposed dataset, demonstrating how technological advances can significantly impact the field of computer vision.

Experimental design

To evaluate the effectiveness of synthetic data for training the five neural networks, and to provide GPT-4o with prior knowledge, we used three distinct training regimens. Importantly, we always used a set of real images for testing, reflecting real-life applications where models are expected to handle real, unseen data.

First, we relied on the results from Higa et al.[12], where neural networks were trained exclusively on real images using a tenfold stratified cross-validation strategy. Secondly, we employed a combination of real and synthetic images for training, again using the tenfold stratified approach. For this setup, the synthetic and real images were kept in different datasets and separately split into ten folds, yielding a total of 20 fold-subsets. Then, for any run i, $1 \leq i \leq 10$, the training and validation sets were sampled from the union between all synthetic and real fold-subsets, except for the i-th ones. The test set for this run i is the i-th fold-subset of real images. The i-th fold-subset of synthetic images was not used in run i. The purpose of this exclusion was to increase data variability in the evaluation. Finally, a last configuration was setup, similarly to the second one, but the training and validation sets were sampled only from the synthetic images. The test set, of course, remained being a given fold-subset i of real images. These three setups will be referred to as the real, combined, and synthetic training sets, respectively.

To assess GPT-4o's performance, we applied a similar approach. For each fold, nine were inputted as prior knowledge to the model, while the tenth (test fold) comprised only real images, which were provided in a single batch for classification. This process was repeated in separate instances of GPT-4o to ensure prior knowledge input without revealing test image labels. Each instance of GPT-4o was labeled with a title and description formatted as: *dataset - fold_X*, where *dataset* was either real, synthetic or combined. During the evaluation, it had access neither to the internet nor to image generation with DALL-E; it did have, however, access to the code interpreter, which seems to be necessary for it to manipulate files. Then, it was instructed with the following text:

> Behave like a stenosis degree classifier. I'll send you images that have already been classified so that you can understand and learn what each class is. You will receive input images of bulldog nostrils, which you should return as: mild, moderate or severe. Inside the zip file that I am attaching to the knowledge, there is the following structure:

mild;
moderate;
severe.

> Note that the folder names are: mild, moderate, severe. These names classify the images inside each folder, so that you can learn each class. You will always return a csv with two columns one being the file name of the image and the other being the degree you have classified, order the rows by the names of the images. You will always analyze the set of images uploaded here in the knowledge to classify the input images. Do not create a second model. Look at each image, analyze the image and classify it by yourself. Limit yourself to using code only where necessary to unzip and organize the image files, and to load the images themselves so that you can analyze them. You must not use code to classify these images, that is, do noticreate a second model and do not resort to strategies such as calculating the distance between the images. The classification must be intended to correspond to the real stenosis degree. Each classification performed should follow the template: show the name of the image file, and show the image itself with your classification. Furthermore, describe the image and your reasoning, so that we know you are not just randomly classifying the image.

For neural network training, we followed the same setup used by Higa et al.[12] to ensure a fair comparison. Since using Sharpness-Aware Minimization (SAM)[30] with Stochastic Gradient Descent (SGD) yielded the best median

results for all the metrics in that work, we selected this optimization strategy for our neural networks in this experiment. In each cycle, 20% of the training images were used for validation. The images were standardized to a size of 256 × 256 pixels to match the input requirements of the networks, and pixel values were normalized to the range [0, 1]. Training with real images only was performed on a 12 GiB Nvidia RTX 3060 GPU, which limited the batch size to 8. For the other configurations, we used a 16 GiB Nvidia RTX A4000 GPU, but the batch size remained the same. For optimization, the categorical cross entropy was used as loss function. The learning rate was set at 0.001, and validation loss values were monitored with a patience of 300 epochs and tolerance of 0.01. Various data augmentation techniques-including color adjustments and geometric transformations-were applied to the training set (but not the test set) to improve the model's ability to generalize. This does not apply to the evaluation of GPT-4o, which followed the procedure described above but received images without prior transformations, either in size or data augmentation.

Data augmentation was extensively applied to the training set (excluding the test set) to improve generalization and reduce overfitting; these techniques included color jittering, random grayscale conversion, inversion, solarization (with a threshold of 0.75), auto-contrast adjustment, random cropping, horizontal and vertical flips (with a 50% probability), 90° rotations, perspective distortions, and sharpness adjustments (factor 2, 50% probability).

In addition to the machine learning models, we conducted an experiment with 52 trained human evaluators specializing in veterinary anatomy to assess their ability to classify nasal stenosis in bulldogs based on the same set of real images used in neural network training. Each evaluator received a subset of the images and was asked to classify them as mild, moderate, or severe stenosis. The results were evaluated based on the accuracy, precision, recall, and F-score of their classifications. These metrics were calculated using the same methods applied to the neural network and GPT-4o results, allowing for a direct comparison between human and machine classification accuracy.

Regarding the neural networks utilized, except for GPT-4o, training and validation accuracy values were calculated, in addition to loss values, in order to further assess model training. Finally, precision, recall, and F-score metrics were calculated in each fold to evaluate the neural networks. For the final evaluation on the test set, these classification metrics were preferred over accuracy because the properties measured by them were considered more important within the context. Statistical analysis included means, standard deviations, medians, and interquartile ranges. A two-way Analysis of Variance (ANOVA) was performed, and the Scott-Knott clustering test (SK) was used post-hoc to explore significant differences between the models. Boxplots and confusion matrices were also generated and used in the analysis and discussion of the results.

The experiments with the mixed and synthetic dataset were conducted on a machine running Linux Ubuntu 22.04.4 LTS, equipped with 32GiB of RAM, a 13th-generation Intel Core i5-13500 processor, and a NVIDIA RTX A4000 GPU with 16 GiB memory. For the implementation of deep learning techniques, the 2.4.0 version of the PyTorch library was used. The Scikit-learn, version 1.4.2, was used for calculating the metrics evaluated on this experiment. The matplotlib library, version 3.9.0, was used for generating the confusion matrices during the testing on each fold. The Torchvision, 0.18.0 version, and TIMM, 1.0.3 version, were used to import the models for our experiment. Finally, the R language, version 4.1.2, with the package ggplot2, version 3.5.0, was used to generate boxplots, accuracy and loss graphics and confusion matrices for the best results obtained at the end of all 10 runs.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to privacy and ethical considerations but are available from the corresponding author, Gustavo da Silva Andrade (gustavo.s.andrade@ufms.br), upon reasonable request.

Received: 14 November 2024; Accepted: 3 March 2025
Published online: 21 March 2025

References

1. Akinsulie, O. C. et al. The potential application of artificial intelligence in veterinary clinical practice and biomedical research. *Front. Vet. Sci.* **11**, 550. https://doi.org/10.3389/fvets.2024.1347550 (2024).
2. Liu, N.-C. et al. Conformational risk factors of brachycephalic obstructive airway syndrome (boas) in pugs, french bulldogs, and bulldogs. *PLoS ONE* **12**, 181928. https://doi.org/10.1371/journal.pone.0181928 (2017).
3. Ezanno, P. et al. Research perspectives on animal health in the era of artificial intelligence. *Vet. Res.* **52**, 4. https://doi.org/10.1186/s13567-021-00902-4 (2021).
4. Åsbjer, E., Hedhammar, Å. & Engdahl, K. Awareness, experiences, and opinions by owners, breeders, show judges, and veterinarians on canine brachycephalic obstructive airway syndrome (boas). *Canine Med. Genet.* **11**, 3. https://doi.org/10.1186/s40575-024-00137-4 (2024).
5. Dupré, G. & Heidenreich, D. Brachycephalic airway syndrome. *Vet. Clin. N. Am. Small Anim. Pract.* **46**, 691–707. https://doi.org/10.1016/j.cvsm.2016.04.003 (2016).
6. Liu, N.-C., Adams, V. J., Kalmar, L., Ladlow, J. F. & Sargan, D. R. Whole-body barometric plethysmography characterizes upper airway obstructions in three brachycephalic breeds of dogs. *J. Vet. Intern. Med.* **30**, 853–865. https://doi.org/10.1111/jvim.13953 (2016).
7. Linhoss, J. & Zhao, Y. Image analysis and computer vision applications in animal sciences: An overview. *Front. Anim. Sci.* **21**, 1492. https://doi.org/10.3390/s21041492 (2021).
8. Barbedo, J., Koenigkan, L., Santos, T. & Santos, A. The livestock farming digital transformation: Implementation of new and emerging technologies using artificial intelligence. *Anim. Health Res. Rev.* **22**, 75–85. https://doi.org/10.1017/S1466252320000166 (2021).
9. Linhoss, J. & Zhao, Y. Practices and applications of convolutional neural network-based computer vision systems in animal farming: A review. *Sensors* **21**, 1492. https://doi.org/10.3390/s21041492 (2021).

10. Bahani, M., El Ouaazizi, A., Avram, R. & Maalmi, K. Enhancing chest x-ray diagnosis with text-to-image generation: A data augmentation case study. *Displays* **83**, 735. https://doi.org/10.1016/j.displa.2024.102735 (2024).
11. Rossi, S., Rossi, M., Mukkamala, R. R., Thatcher, J. B. & Dwivedi, Y. K. Augmenting research methods with foundation models and generative AI. *Int. J. Inf. Manag.* **77**, 102749. https://doi.org/10.1016/j.ijinfomgt.2023.102749 (2024).
12. Higa, G. T. H., Carvalho, J. K. M. R., Zanoni, P. B. P., de Andrade, G. B. & Pistori, H. A new machine learning dataset of bulldog nostril images for stenosis degree classification. http://arxiv.org/abs/2403.07132 (2024).
13. Schmid, C. et al. Anatomical, functional, and blood-born predictors of severity of brachycephalic obstructive airway syndrome severity in french bulldogs. *Front. Vet. Sci.* **11**, 440. https://doi.org/10.3389/fvets.2024.1486440 (2025).
14. Ren, J., Li, C., An, Y., Zhang, W. & Sun, C. Few-shot fine-grained image classification: A comprehensive review. *AI* **5**, 405–425. https://doi.org/10.3390/ai5010020 (2024).
15. Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **34**, 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203 (2022).
16. Thung, K.-H. & Wee, C.-Y. A brief review on multi-task learning. *Multimedia Tools Appl.* **77**, 29705–29725 (2018).
17. Oren, A. et al. Brachysound: Machine learning based assessment of respiratory sounds in dogs. *Sci. Rep.* **13**, 20300 (2023).
18. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. http://arxiv.org/abs/2204.06125 (2022).
19. Pan, S. et al. Land-cover classification of multispectral lidar data using cnn with optimized hyper-parameters. *ISPRS J. Photogramm. Remote Sens.* **166**, 241–254. https://doi.org/10.1016/j.isprsjprs.2020.05.022 (2020).
20. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. http://arxiv.org/abs/2010.11929 (2011).
21. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (eds Pereira, F. et al.), vol. 25 (Curran Associates, Inc., 2012).
22. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778. https://doi.org/10.1109/CVPR.2016.90 (2016).
23. Howard, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*. http://arxiv.org/abs/1704.04861 (2017).
24. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4510–4520. https://doi.org/10.1109/CVPR.2018.00474 (IEEE Computer Society, 2018).
25. Howard, A. et al. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 1314–1324. https://doi.org/10.1109/ICCV.2019.00140 (IEEE Computer Society, 2019).
26. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269. https://doi.org/10.1109/CVPR.2017.243 (IEEE Computer Society, 2017).
27. Vaswani, A. et al. Attention is all you need. http://arxiv.org/abs/1706.03762 (2017).
28. Liu, Z. et al. Swin transformer v2: Scaling up capacity and resolution. Preprint at http://arxiv.org/abs/2106.13230 (2021).
29. Tu, Z. et al. Maxvit: Multi-axis vision transformer. Preprint at http://arxiv.org/abs/2204.01697 (2022).
30. Foret, P., Kleiner, A., Mobahi, H. & Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the International Conference on Learning Representations* (2021).

Acknowledgements

This work has received financial support from the Dom Bosco Catholic University and the Foundation for the Support and Development of Education, Science and Technology from the State of Mato Grosso do Sul, FUNDECT. Some of the authors have been awarded with Scholarships from the the Brazilian National Council of Technological and Scientific Development, CNPq and the Coordination for the Improvement of Higher Education Personnel, CAPES.

Author contributions

G.S.A.: Methodology, Software, Validation, Formal analysis, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization. G.T.H.H.: Methodology, Software, Formal analysis, Writing—Original Draft, Writing—Review and Editing. J.F.S.R.: Data-Curation and Writing—Original Draft, Writing—Review and Editing. J.K.M.R.C.: Data-Curation and Resources. W.N.G.: Investigation, Writing—review. M.H.N.: Investigation, Writing—review. H.P.: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing—review, Supervision, Project administration, Funding acquisition.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.S.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.