Beef Carcass Grading using Deep Convolutional Networks

Geazy Vilharva Menezes[1], Everton Castelão Tetila[2], Diogo Nunes Gonçalves[3],
Vanessa Aparecida de Moraes Weber[4], Gabriel Toshio Hirokawa Higa[5], Marcelo Fontes Pereira[6],
Marina de Nadai Bonin Gomes[7], Rodrigo da Costa Gomes[8], Hemerson Pistori[9]
[1, 3, 7, 9]*Universidade Federal de MS*, Campo Grande, Brazil
[2, 4, 5, 6, 9]*Universidade Católica Dom Bosco*, Campo Grande, Brazil
[2]*Universidade Federal da Grande Dourados*, Dourados, Brazil
[4]*Universidade Estadual de MS*, Campo Grande, Brazil
[4]*KeroW Soluções de precisão*, Campo Grande, Brazil
[8]*EMBRAPA Gado de Corte*, Campo Grande, Brazil
gabrieltoshio03@gmail.com[5]

***Abstract*—Beef carcass grading is an invaluable tool to ensure meat quality. In most of the Brazilian abattoirs, carcasses are graded through visual analysis by trained graders. In order to automate this process, we evaluate seven image deep learning models. For this purpose, a new dataset was created containing images of 670 bovine half-carcasses taken during regular operation in an abattoir. The images were graded by three professionals. All three experts agreed in only 9.9% of the cases, and two out of three graders agreed in 58.82%. The graders disagreed on 31.28% of the images. These results indicate the complexity of the problem. Nonetheless, an overall accuracy of 53% was achieved using convolutional neural networks, which is close to human performance, when the agreement between the graders is considered. Furthermore, an accuracy of around 91% can be achieved if the cases of disagreement are disregarded.**

I. Introduction

The demand for high quality beef as a source for protein and other important nutrients for humans has been growing constantly worldwide during the last decades [1], [2]. Evaluation of carcass fat is an important step in quality control for meat production and can be made using different kinds of sensors and visual inspection strategies [3]–[5]. Currently, there are some systems that assist in carcass grading, some of which already present some level of automation, using equipment like the VIAscan [6]. However, their usage is usually limited to specific and sometimes subjective criteria that may differ from region to region, specially when different grading protocols are considered. For instance, VIAscan is not adjusted to any Brazilian beef grading system. Nevertheless, even when works across different grading protocols are considered, the complete automation of the process using only whole carcass images remains an open problem.

Recently, many approaches based on computer vision and machine learning have been proposed to automate meat quality evaluation for different important livestock animals [7]. A system to evaluate chicken meat freshness has been proposed by Taheri-Garavand et al. [8], using a combination of genetic algorithms and artificial neural networks. Poultry carcasses has also been studied in Chmiel et al. [9], who used computer vision to estimate fat content. In Alcayde et al. [10] a custom-built platform was used to identify pork meat age using image analysis and three different regression algorithms. The quality analysis of lamb carcasses using videos and a computer vision system has been investigated by Araújo et al. [11], with positive results.

When bovines are considered, some studies have been conducted on the usage of computer vision systems to evaluate not only carcass quality (specially carcass conformation and fat cover), but also bovine meat quality. Common prediction targets in these cases are the percentage of intramuscular fat and the marbling score. For instance, Pinto et al. [12] proposed the application of the local binary pattern method to extract color and texture features, and tree-based machine learning algorithms to classify beef according to their marbling score, and Pannier et al. [13] evaluated the capacity of a RGB scanner mounted above a conveyer belt along with a machine learning system to predict intramuscular fat percentage and marbling scores.

Research regarding assistive systems for bovine carcass grading has been conducted in different regions, such as the US and Europe. Negretti et al. [14], for instance, proposed one such high-performance system for the evaluation of carcass conformation and fat cover in accordance to the European legislation (specifically, the system was developed in Italy). However, the proposed system is actually semi-automatic, requiring that the user indicates some reference points in the image to be used in the process of evaluation. The usage of semi-automatic systems that require some user input seems to be a common strategy in systems of this kind. Another example of a very similar system developed in Norway is the one by Heggli et al. [15], which also addressed the problem of objectivity. In their case, a model was fitted that required not only information such as age, but also manual length measurements of the carcasses (albeit with manually operated lasers).

In this work, a new automatic beef carcass evaluation system, based on deep learning techniques and using images from carcasses freely passing in front of a standard RGB camera is presented. The system predicts a grade that follows one of the grading protocols used in Brazil, based on the Normative Instruction n. 9 of May 4th, 2009, of the Brazilian Ministry of Agriculture, Livestock and Provision. Its performance was compared to trained human graders using a new dataset of images captured in real conditions and not in a controlled laboratory situation.

The carcass grading protocol used in this paper defines nine carcass finishing classes related to visual aspects of the bovine half-carcass. As some of these classes rarely happen in the Brazilian state of Mato Grosso do Sul, where the images were collected, only 6 grades were used, due to the lack of a minimum amount of samples for the other classes. Seven state-of-the-art deep learning architectures were tested and evaluated using four performance metrics.

The main contributions of this paper are 2-fold: (1) an evaluation of deep learning techniques for beef carcass grading in accordance to a Brazilian system; and (2) the quantification of the complexity of the problem related to human visual grading. Detailed information regarding materials, methods and results are presented in the following sections.

II. Materials and Methods

A GoPro Hero 3+ camera was attached to a tripod 4 meters away from an overhead rail where beef half-carcasses move on, inside an abattoir in Campo Grande city, Brazil, during the months of March and April 2018. The recordings happened during regular operation in three different days, from 4am to 1pm, at 60fps and 1080p spatial resolution. Each frame from the videos was visually analysed and, from each half-carcass, frames with enough visual quality were selected. As the recordings happened during regular operation, some of the frames portrayed workers passing in front of the camera and were removed. Another reason for discard is related to the movement of the carcasses as they turned alongside the overhead rail, resulting on incomplete views or non-frontal angles of the carcass.

The selected frames were also cropped and corrected for the radial distortion from the GoPro fisheye lens. In this way, each frame has a clear and large view of the half-carcass, as seen in Figure 2. A total of 687 images were then graded independently by three human graders. The adopted grading system is based on the Normative Instruction (NI) n. 9 of May 4th, 2009, of the brazillian Ministry of Agriculture, Livestock and Provision. In itself, this NI defines five possible grades related to carcass conformation and fat cover: Absent, Scarce, Median, Uniform and Excessive. Two of these have been subdivided for better evaluation of carcass finishing[1], yielding nine possible grades. These are: Absent (A), Scarce^{-} (S^{-}), Scarce^{o} (S^{o}), Scarce^{+} (S^{+}), Median^{-} (M^{-}), Median^{o} (M^{o}), Median^{+} (M^{+}), Uniform (U) and Excessive (E). Images with the grades M^{+} (n=12), U (n=5) and E (n=0) were removed from the final dataset due to the small number of samples, so that only six different grades were kept for the experiment. The dataset ground-truth grade was established by the majority vote among the three graders, or a random grade when all of the graders used a different grade. The distribution of grades for the 670 images of the final dataset is shown in Figure 1. One exemplar image for each of the six carcass grades are shown in Figure 2.

[1]The subdivision and grading criteria are explained here: https://www.frivasa.com.br/imagens/pecuaristas/classificacaodecarcacafrivasa.pdf

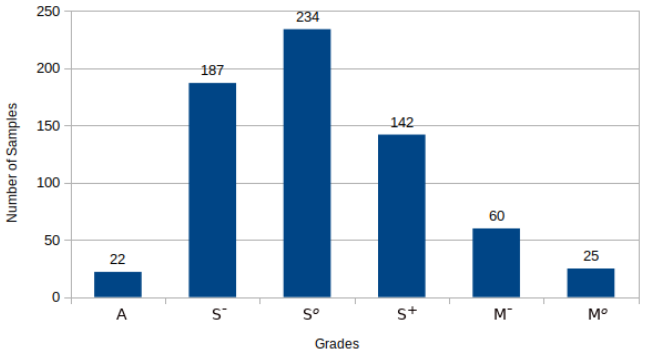
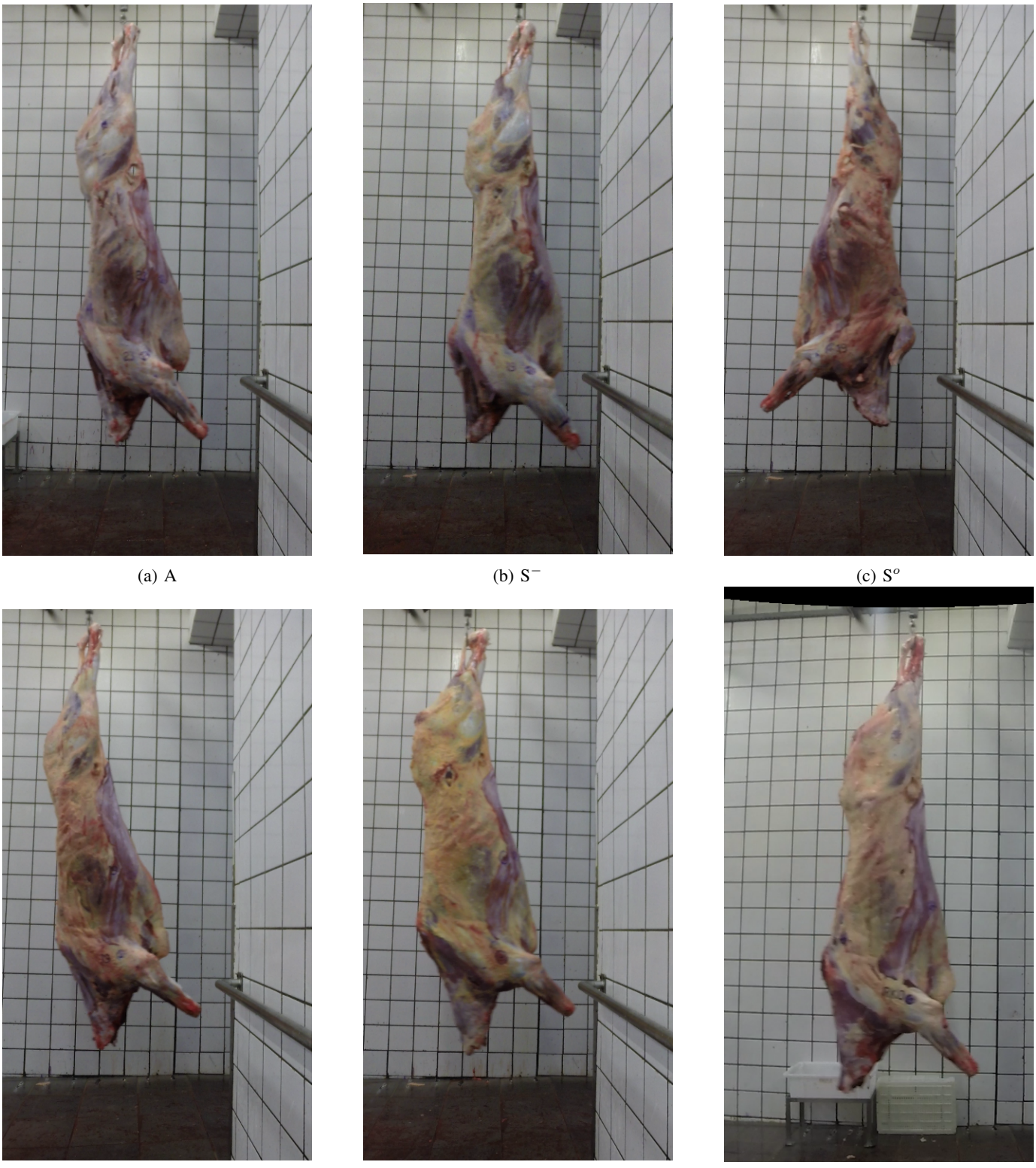


Fig. 1: Number of images per grade that compose the dataset used in this work

By means of a 5-fold stratified cross-validation, seven state-of-the-art deep learning architectures were compared on the task of automatically grading beef carcass: DenseNet201 [16], InceptionResNetV2 [17], InceptionV3 [18], ResNet50 [19], VGG16 [20], VGG19 [20] and Xception [21]. All model hyperparameters were set to the same values used in Tetila et al. [22]. The neural networks were optimized with Stochastic Gradient Descent (SGD) in minibatches of size 8. The training was performed in 100 epochs, with a learning rate of 0.001 and momentum of 0.9. Cross-entropy was used as the loss function. All the images were rescaled to 256x256 pixels. Data-augmentation was performed using rotations and translations in a factor of $\pm 30\%$ of $2\pi rad$, for rotations, and of the image height and width, for translations, resulting in 100 new images for each original one. Transfer learning using weights pretrained in the Imagenet dataset [23] with fine-tuning was also applied. Accuracy, precision, recall and f-score were calculated from the cross-validation results over the test set. The f-scores were subjected to a Friedman's non-parametric hypothesis test.

III. Results and Discussion

The three graders agreed on 9.9% images. In 58.82% two graders agreed but one of them used a different grade. The remaining 31.28% had 3 different grades. Table I shows the accuracy, precision, recall and f-score values achieved by each of the 7 deep learning models. InceptionV3 presented the highest values for precision, recall and f-score but DenseNet201 resulted on a higher accuracy of 53.71%. The non-parametric Friedman hypothesis test over the f-score metric resulted on



(a) A (b) S^- (c) S^o

(d) S^+ (e) M^- (f) M^o

Fig. 2: One sample for each of the classes used in the automatic carcass grading experiment: (a) Absent A, (b) Scarce$^-$ S^-, (c) Scarce^o S^o, (d) Scarce$^+$ S^+, (e) Median$^-$ M^-, (f) Median^o M^o

Architec.	Accuracy	Precision	Recall	F-score
DenseNet201	**53.71**	46,60	48,00	46,00
IncResNetV2	49.45	44,40	45,40	44,20
InceptionV3	53.06	**47,00**	**48,00**	**46,40**
ResNet50	53.40	43,60	45,00	41,6
VGG16	48.71	43,00	43,00	38,40
VGG19	47.74	41,40	43,80	39,80
Xception	50.16	42,80	45,00	41,20

TABLE I: Accuracy, Precision, recall and f-score for each of the architectures, in percentage, with the highest values in bold

a p-value of 0.097, indicating no evidence of a statistically significant difference between the architectures even at a 5% significance level.

DenseNet201 and InceptionV3 have been chosen for a further analysis using the confusion matrix shown in Figures 3 and 4. Both matrices show a clear pattern of error concentration along the main diagonal, suggesting that when the learned models misclassify a carcass, the error is not usually far from the correct grade. It is also clear that DenseNet201 and InceptionV3 present a different pattern of errors regarding mostly the confusion of S^+ with S^o that happened in 69 images with DenseNet201 but with just 47 images with InceptionV3. The reverse also happened, with the learned models differently confusing S^o with S^+ 24 times for DenseNet201 but 52 times for InceptionV3.

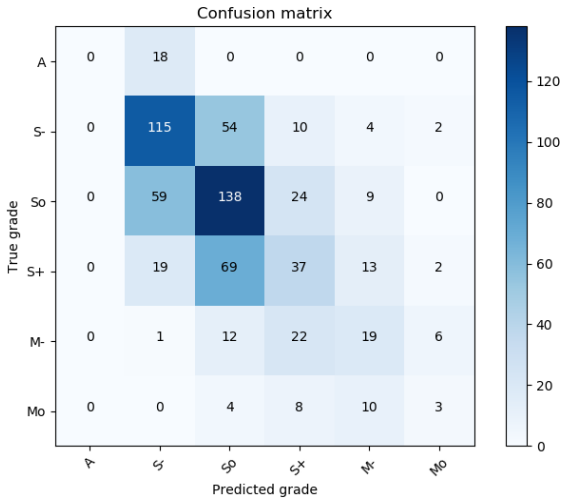


Fig. 3: Confusion matrix for DenseNet201

If a less fine-grained grading system was used, without the subdivisions inside the scarce and median grade, the resulting confusion matrices would be those shown in Figures 5 and 6, which were obtained by regrouping the existing results. The confusion would be significantly reduced and the accuracy for DenseNet201 would increase to 85.56% and InceptionV3 to 87.04%.

Figure 7 shows 3 examples of images pointed out by some of the graders as more difficult to evaluate due to wrong carcass angles and low quality related to the blurring effect. It is possible to see in these examples that some of the images

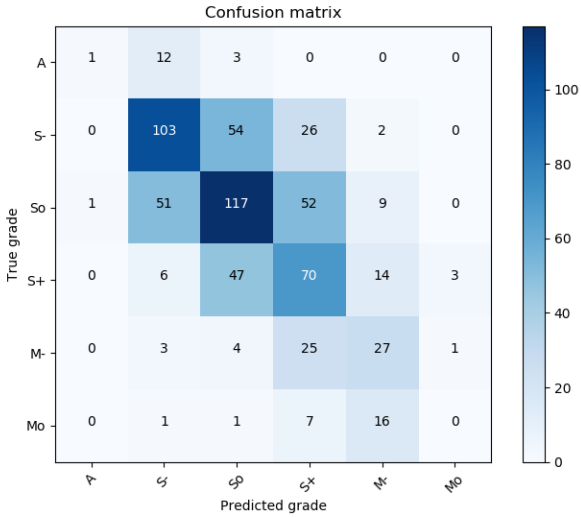


Fig. 4: Confusion matrix for InceptionV3

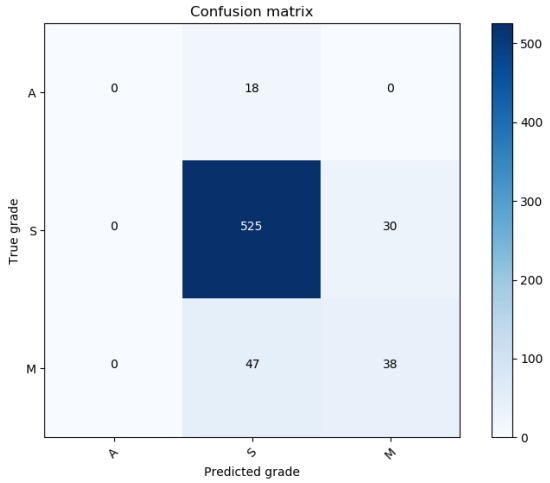


Fig. 5: Confusion matrix for DenseNet201 without the subgrades for scarse and median grades

also have workers [2] helping to stabilize the carcass for the video recordings. Figure 8 presents the amount of grades used by each human grader. It is possible to see a bias from the third grader toward grade S^-, which has been used a lot. On the other side, this grader marked much less carcasses as a S^+. This may indicate some confusion using the labeling software or some conceptual difference on how he/she interprets the grading protocol. Further studies using more graders would be an interesting path for the future.

IV. Conclusion

In this work, the automatic carcass grading using deep learning has been evaluated and the results are encouraging. The proposed approach has been tested on a operational

[2]Workers faces have been blurred for the paper due to privacy issues but in the dataset the faces are not blurred

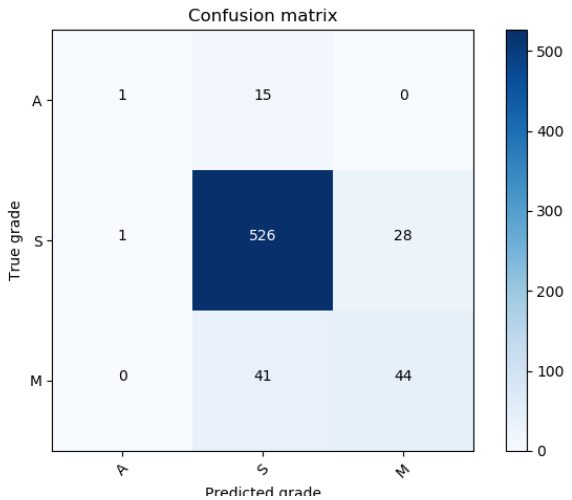


Fig. 6: Confusion matrix for InceptionV3 without the sub-grades for scarse and median grades

environment using a very simple camera capture arrangement. Results indicate that the problem is very hard, even for humans, but machine learning algorithms were able to produce grades that almost compare to human grading. The highest accuracy of 53% was achieved using InceptionV2 but with errors concentrated around the true grade. When only 3 grades where used, absent, scarce and median, the accuracy reached a much higher value of 87.04%. For future work we suggest the use of balancing techniques, as the dataset is highly unbalanced among different grades. Increasing the number of human graders in order to decrease the variability is also recommended based on our results. It would also be important to compare the grading from experts using just the images, as we did here, and those from live evaluation inside the abattoir.

ACKNOWLEDGMENT

This work has received financial support from the Dom Bosco Catholic University and from the Foundation for the Support and Development of Education, Science and Technology from the State of Mato Grosso do Sul, FUNDECT. The work has also received financial support from the São Paulo Research Foundation, FAPESP - Process: 2023/03870-8, and from the Embrapa Digital Agriculture/CNPTIA/EMBRAPA - Process: 2022/09319-9. 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. We would also like to thank NVIDIA for providing the Titan X GPU used in the experiments with deep learning.

REFERENCES

[1] C. Gajaweera, K. Y. Chung, S. H. Lee, H. I. Wijayananda, E. G. Kwon, H. J. Kim, S. H. Cho, and S. H. Lee, "Assessment of carcass and meat quality of longissimus thoracis and semimembranosus muscles of hanwoo with korean beef grading standards," *Meat Science*, vol. 160, p. 107944, 2020.

[2] C. Bonnet, Z. Bouamra-Mechemache, V. Réquillart, and N. Treich, "Viewpoint: Regulating meat consumption to improve health, the environment and animal welfare," *Food Policy*, p. 101847, 2020.

[3] S. Piao, T. Okura, and M. Irie, "On-site evaluation of wagyu beef carcasses based on the monounsaturated, oleic, and saturated fatty acid composition using a handheld fiber-optic near-infrared spectrometer," *Meat Science*, vol. 137, pp. 258 – 264, 2018.

[4] G. K. Naganathan], K. Cluff, A. Samal, C. R. Calkins, D. D. Jones, C. L. Lorenzen, and J. Subbiah, "Hyperspectral imaging of ribeye muscle on hanging beef carcasses for tenderness assessment," *Computers and Electronics in Agriculture*, vol. 116, pp. 55 – 64, 2015.

[5] E. Fiore, G. Fabbri, L. Gallo, M. Morgante, M. Muraro, M. Boso, and M. Gianesella, "Application of texture analysis of b-mode ultrasound images for the quantification and prediction of intramuscular fat in living beef cattle: A methodological study," *Research in Veterinary Science*, vol. 131, pp. 254 – 258, 2020.

[6] Y. Ye, N. Schreurs, P. Johnson, R. Corner-Thomas, M. Agnew, P. Silcock, G. Eyres, G. Maclennan, and C. Realini, "Carcass characteristics and meat quality of commercial lambs reared in different forage systems," *Livestock Science*, vol. 232, p. 103908, 2020.

[7] A. Taheri-Garavand, S. Fatahi, M. Omid, and Y. Makino, "Meat quality evaluation based on computer vision technique: A review," *Meat Science*, vol. 156, pp. 183 – 195, 2019.

[8] A. Taheri-Garavand, S. Fatahi, F. Shahbazi, and M. de la Guardia, "A nondestructive intelligent approach to real-time evaluation of chicken meat freshness based on computer vision technique," *Food Process Engineering*, vol. 42, no. 4, 2019.

[9] M. Chmiel, M. Słowiński, and K. Dasiewicz, "Application of computer vision systems for estimation of fat content in poultry meat," *Food Control*, vol. 22, no. 8, pp. 1424 – 1427, 2011.

[10] M. Alcayde, F. Elijorde, and Y. Byun, "Quality monitoring system for pork meat using computer vision," in *2019 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, 2019, pp. 1–7.

[11] J. Araújo, A. Lima, M. Nunes, M. Sousa, G. Serrão, E. Morais, L. Daher, and A. Silva, "Relationships among carcass shape, tissue composition, primal cuts and meat quality traits in lambs: A pls path modeling approach," *Small Ruminant Research*, vol. 182, pp. 52 – 66, 2020.

[12] D. L. Pinto, A. Selli, D. Tulpan, L. T. Andrietta, P. L. M. Garbossa, G. V. Voort, J. Munro, M. McMorris, A. A. C. Alves, R. Carvalheiro, M. D. Poleti, J. C. de Carvalho Balieiro, and R. V. Ventura, "Image feature extraction via local binary patterns for marbling score classification in beef cattle using tree-based algorithms," *Livestock Science*, vol. 267, p. 105152, 2023.

[13] L. Pannier, T. van de Weijer, F. van der Steen, R. Kranenbarg, and G. Gardner, "Prediction of chemical intramuscular fat and visual marbling scores with a conveyor vision scanner system on beef portion steaks," *Meat Science*, vol. 199, p. 109141, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0309174023000475

[14] P. Negretti, G. Bianconi, G. Cannata, G. Catillo, R. Steri, R. Barrasso, and G. Bozzo, "Visual image analysis for a new classification method of bovine carcasses according to eu legislation criteria," *Meat Science*, vol. 183, p. 108654, 2022.

[15] A. Heggli, L. E. Gangsei, M. Røe, O. Alvseike, and H. Vinje, "Objective carcass grading for bovine animals based on carcass length," *Acta Agriculturae Scandinavica, Section A — Animal Science*, vol. 70, no. 2, pp. 113–121, 2021. [Online]. Available: https://doi.org/10.1080/09064702.2021.1906940

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 4278–4284.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

Fig. 7: Examples of images that the graders pointed out as more difficult due to wrong carcasses angles and blurring effect

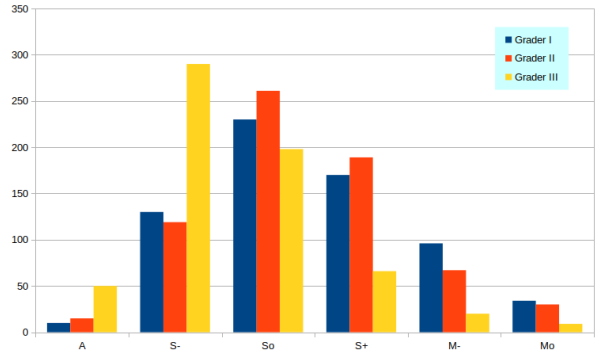


Fig. 8: Distribution of the grades for each of the 6 possible classes by the 3 human graders

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[22] E. C. Tetila, "A deep learning approach for automatic counting of soybean insect pests," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, 2015.