Fish Erythrocytes Nuclear Abnormalities Classification using Machine Learning

Newton Loebens[1], Bruno do Amaral Crispim[2], Nathalya Alice de Lima[3], Everton Tetila[4], Celso Costa,
[5] Willian Paraguassu Amorim[6], Alexeia Barufatti[7], Pedro Henrique Neves da Silva[8],
Gabriel Toshio Hirokawa Higa[9], and Hemerson Pistori[10]
[1,9,10]Universidade Católica Dom Bosco, Campo Grande, Brazil
[2,3,4,6,7] Universidade Federal da Grande Dourados, Dourados, Brazil
[5] Instituto Federal de Mato Grosso do Sul, Campo Grande, Brazil
[8,10] Universidade Federal de Mato Grosso do Sul, Campo Grande, Brazil
Email: [1]newtonloebens@gmail.com

***Abstract*—The creation of automated systems capable of detecting anomalies in fish erythrocytes is an important concern in the area of marine biology. We investigate the possibility of using machine learning to classify images of abnormal and normal nuclei of fish erythrocytes, considering three abnormalities: nuclear bud, notched nuclei, and vacuole nuclei, among others. Random Forests were shown to have the highest AUC median in both sets, reaching AUC values of 0.896 and 0.959 for all sets of classes and the vacuole set, respectively, being able to correctly classify a high percentage of the bud and notched cells. However, when all classes are considered, the outcome is impressively better.**

I. Introduction

A significant challenge in the field of marine biology involves the development of automated systems capable of identifying abnormalities in fish erythrocytes. Through machine learning techniques, we investigate this issue with a focus on shallow learning algorithms, specifically Random Forests.

The selection of shallow learning algorithms is justified by the nature of the problem at hand. The realm of microscopic image analysis for classification has found value in these approaches, as they are particularly well-suited for problems with datasets of moderate size and features of low complexity [1] [2]. Moreover, this kind of algorithm has shown effectiveness in numerous classification applications [3] [4].

The chosen feature extractors and parameters for this study were based on prior outcomes attained in analogous problems involving the classification of microscopic images [5]. This underscores the significance of drawing upon accumulated knowledge and established best practices to ensure the efficiency of the proposed system. Employing these features and parameter extractors will contribute to a robust and well-founded strategy for addressing the problem in question.

A noteworthy aspect of the addressed problem lies in the highly imbalanced nature of the dataset. This imbalance arises from the scarcity of nuclear abnormalities relative to the overall number of erythrocytes analyzed. Consequently, it is expected that the correct classification rate (CCR) will surpass the Area Under the ROC Curve (AUC) within the framework of this study, since the CCR considers both the total count of accurate and inaccurate classifications, while the AUC evaluates the model's ability to differentiate between classes, potentially influenced by the infrequent occurrence of nuclear abnormalities.

The dataset used in this work and the experiments carried out play a significant role in advancing research in the area of classification of nuclear abnormalities in fish erythrocytes. Until now, the accurate and automated detection of nuclear abnormalities in fish erythrocyte cells has been a complex challenge. However, by introducing a well-curated and comprehensive dataset, we offer the scientific community a valuable tool for the development and evaluation of this operation. Furthermore, we diversify the exploration of machine learning approaches and image processing techniques, contributing to the accurate and effective identification of nuclear abnormalities in fish erythrocytes. The results of these experiments have the potential to improve the understanding of fish health conditions, providing essential information for the conservation of aquatic ecosystems and the monitoring of these environments.

Also, this approach of combining nuclear abnormality data with the latest image analysis and machine learning techniques paves the way for substantial advances in the field. The ability to accurately and efficiently identify nuclear abnormalities in fish erythrocytes is crucial for scientific assessment research in the aquatic environment [6], [7]. In addition to computer vision practitioners and scholars, the scientific community and aquatic biology professionals can greatly benefit from the insights and methodologies resulting from our experiments, paving the way for a deeper understanding of nuclear abnormalities in fish erythrocytes and their biological implications.

In 2021, Phillip et al. [8] introduced a protocol and software to analyze the morphology of cells and nuclei from fluorescence or brightfield images. Analysis of cell morphology distributions in automatically identified shape modes allows relating cell shapes to cell subtypes based on endogenous and exogenous conditions. The VAMPIRE algorithm was used to profile and classify cells in shape modes based on equidistant points along their contours, being highly automated and fast, allowing the quantification of morphologies in 2D projections of cells on 2D substrates or in 3D microenvironments, such as hydrogels and fabrics.

Reliable automated tools for cell cycle classification at the individual cell level using in situ imaging are still limited, so it is necessary to establish precise strategies that combine bioimaging with high-content image analysis for reliable classification. Narotamo et al. [9] developed a supervised machine learning method for cell cycle phase classification in individual adherent cells using in situ fluorescence imaging of DAPI-stained nuclei. A Support Vector Machine (SVM) classifier operated on normalized core features using over 3500 DAPI-stained nuclei. True molecular labels were obtained by automatic image processing using fluorescent ubiquitin-based cell cycle indicator technology (Fucci).

Talapatra et al. [10] employed a two-part approach to analyze fish peripheral erythrocytes by firstly detecting cell numbers and measuring the shape of cells, cytoplasm, and nuclei in Giemsa-stained images of fish peripheral erythrocytes. This was achieved through the utilization of CellProfiler, an image analysis tool. Then various machine learning algorithm models, including BayesNet, NaiveBayes, logistic regression, Lazy.KStar, decision tree J48, Random Forest, and Random Tree, were evaluated using the WEKA tool. The aim was to predict the accuracy of the dataset generated from the images. The CellProfiler provided primary, secondary, and tertiary object data, including cell numbers and individual cellular area shape, for cells, cytoplasm, and nuclei. Both CellProfiler and WEKA proved effective in extracting rich information from the dataset and yielded promising results for classifier accuracy, contributing to computational biological research, facilitating the extraction of valuable dataset information through ML modeling and the potential for future analysis of biological big data using WEKA.

II. Materials and Methods

For the genotoxicity test, juveniles of Oreochromis niloticus were exposed for 96 hours to a concentration of 40 mg / kg of cyclophosphamide, used as a positive control for inducing genotoxicity. At the end of the exposure period, the fish's caudal vein was punctured and blood smeared on slides. The slides were hydrolyzed in HCl2mol at 60°C for 10 min and stained with the reagents Schiff and Fast Green, respectively. Subsequently, the slides were observed in a Nikon optical microscope with a connected camera at 100x magnification. The photomicrographs were recorded, cut, and separated into folders according to each change.

Figure 1 shows how the 3099 annotated images are divided into the 4 classes, three corresponding to each kind of abnormality along with another one for normal cells. As expected, most of the images are from normal cells (n=2695). The least common abnormality is nuclear bud, with only 36 samples. Vacuole nuclei cells have a reasonable number of samples (n=309) and 59 examples are from cells with notched nuclei. Three examples for each class are presented in Figure 2.

A set of 395 features has been extracted from each image using several color, shape, and texture descriptors. Thirty-six features are related to the mean, standard deviation, minimum

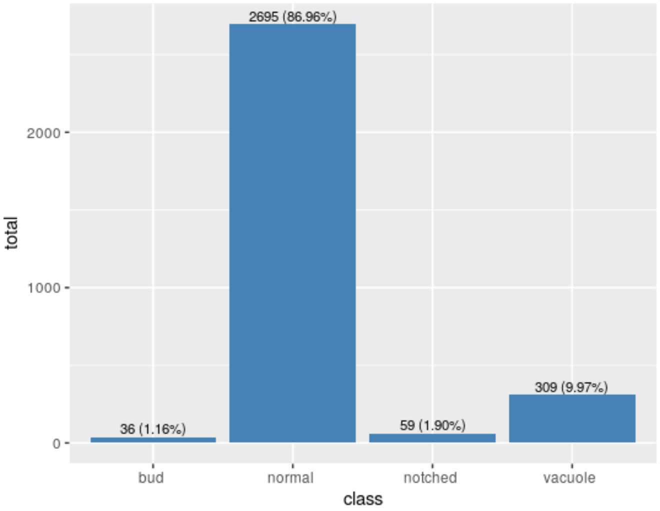
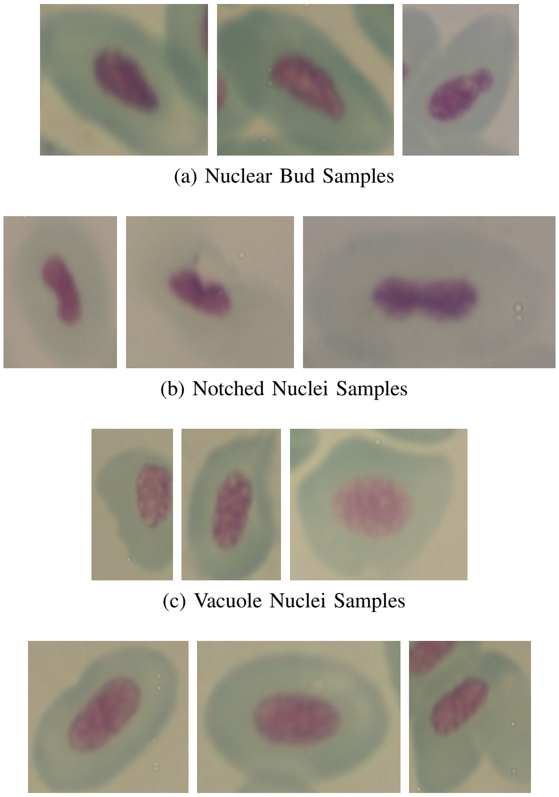


Fig. 1: Number of images per abnormal and normal nuclei. Three abnormalities have been considered: nuclear bud (bud), notched nuclei (notched), and vacuole nuclei (vacuole).



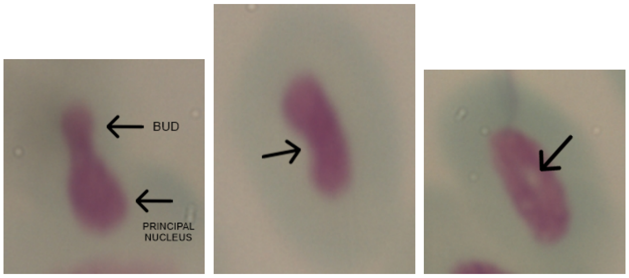
(a) Nuclear Bud Samples

(b) Notched Nuclei Samples

(c) Vacuole Nuclei Samples

(d) Normal Cells

Fig. 2: Sample images for each of the classes considered: (a) nuclear bud, (b) notched nuclei, (c) vacuole nuclei, and (d) normal cells



(a) Nuclear Buddings (b) Notched Nuclei (c) Vacuolated Nuclei

Fig. 3: The black arrows in these images are pointing to important characteristics that define the 3 alterations studied in this paper: (a) nuclear bud has a protuberance on one of the nucleus extremities, (b) notched nuclei are characterized by a concavity in the nucleus center and (c) the vacuole nucleus has a space inside.

and maximum values for each band of the RGB, HSV, and CIELab color spaces. Contrast, dissimilarity, homogeneity, angular second moment, energy, and correlation values have been extracted from gray-scale co-occurrence matrices (GLCM) in the 0°, 45° and 90° directions and at 1 and 2 pixels distances, giving 36 more features. Seventeen image moments were also used: raw, central and Hu, together with 18 Local Binary Patterns (LBP) and 160 Gabor Filter Banks from 8 directions: 0°, 45°, 90°, ..., 315°; 6 sine wave frequencies: 0.01, 0.10, 0.25, 0.50 and 0.90; 2 Gaussian envelop standard deviations: 1 and 3. Finally, a histogram of oriented gradients (HOG) was used to extract the final 160 features. Table I summarizes all the features extracted to serve as input to the shallow machine learning algorithms. These feature extractors and parameters were chosen based on previous results over other classification problems with similar images [5].

TABLE I:
Features Extracted from each image

Feature Group	Number of Features	Reference
Color Space Statistics	36	[11]
Co-occurrence Matrices	36	[12]
Image Moments	17	[13]
Local Binary Patterns	18	[14]
Gabor Filter Banks	160	[15]
Histogram of Oriented Gradients	128	[16]
TOTAL	395	

Two decision tree based approaches have been used as the machine learning models: Random Forest (RF) and a more traditional decision tree inducer based on C4.5 (DT). Two support vector machines (SVM) have also been used, one with a polynomial kernel (SVMP) and the other with a radial basis function (RBF) kernel (SVMR). Finally, a k-Nearest Neighbour (kNN) approach was tested using k equal to 1, 5 and 10 (1NN, 5NN and 10NN respectively). All the other hyper-parameters have been set according to the default values of Weka 3.9.4. Table II presents the shallow machine learning models used and some references.

TABLE II:
Shallow learning models used in the experiments

Acronym	Model	Reference
RF	Random Forest	[17]
DT	Decision Tree	[18]
SVMP	Support Vector Machine with Polynomial Kernel	[19]
SVMR	Support Vector Machine with RBF Kernel	[19]
1NN	k-Nearest Neighbours with k=1	[20]
5NN	k-Nearest Neighbours with k=5	[20]
10NN	k-Nearest Neighbours with k=10	[20]

A 5-fold stratified cross-validation strategy, with 10 repetitions, has been used to run the machine learning techniques over 2 configurations of the dataset: one involving all the 4 classes (**all**) and the other using only the 2 most frequent classes: normal and vacuole nuclei (**vacuole**). Two main metrics were calculated: the Correct Classification Rate (CCR) and the Area Under the Receiver Operating Characteristic Curve (AUC). Nonetheless, we focus on the AUC, since it can be considered less biased than the CCR for heavily imbalanced datasets. For the analysis of the results, boxplots and confusion matrices were used. An Analysis of Variance (ANOVA) was also conducted, followed by the Scott-Knott clustering test, at a 5% significance level. Confusion matrices are presented for the technique with the highest AUC mean. A test has also been made using a balanced version of the data where each class has been randomly under-sampled to have only 36 images. Other metrics, such as precision, recall and f-score were calculated when relevant to the discussion.

III. Results and Discussion

Figure 4 shows the boxplots for AUC and CCR performance metrics, both using all the classes (all) and only the vacuole and normal classes (vacuole). The median values are higher for almost all configurations when only the most frequent class, vacuole, is considered against normal cells. The Random Forest has the highest AUC median in both sets, but its result is significantly higher when all classes are considered (averaging 0.896, against 0.656 in second place). As expected, in general, CCR is higher than AUC, which is likely due to the extremely imbalanced nature of the problem, as one can argue from Figure 5.

The mean values for AUC and CCR are shown in Table III, grouped using the Scott-Knott test. For both, all classes set and vacuole set the Random Forest model achieved the highest means, 0.896 and 0.959 respectively, and the Scott-Knott test confirms that its performance is statistically better than

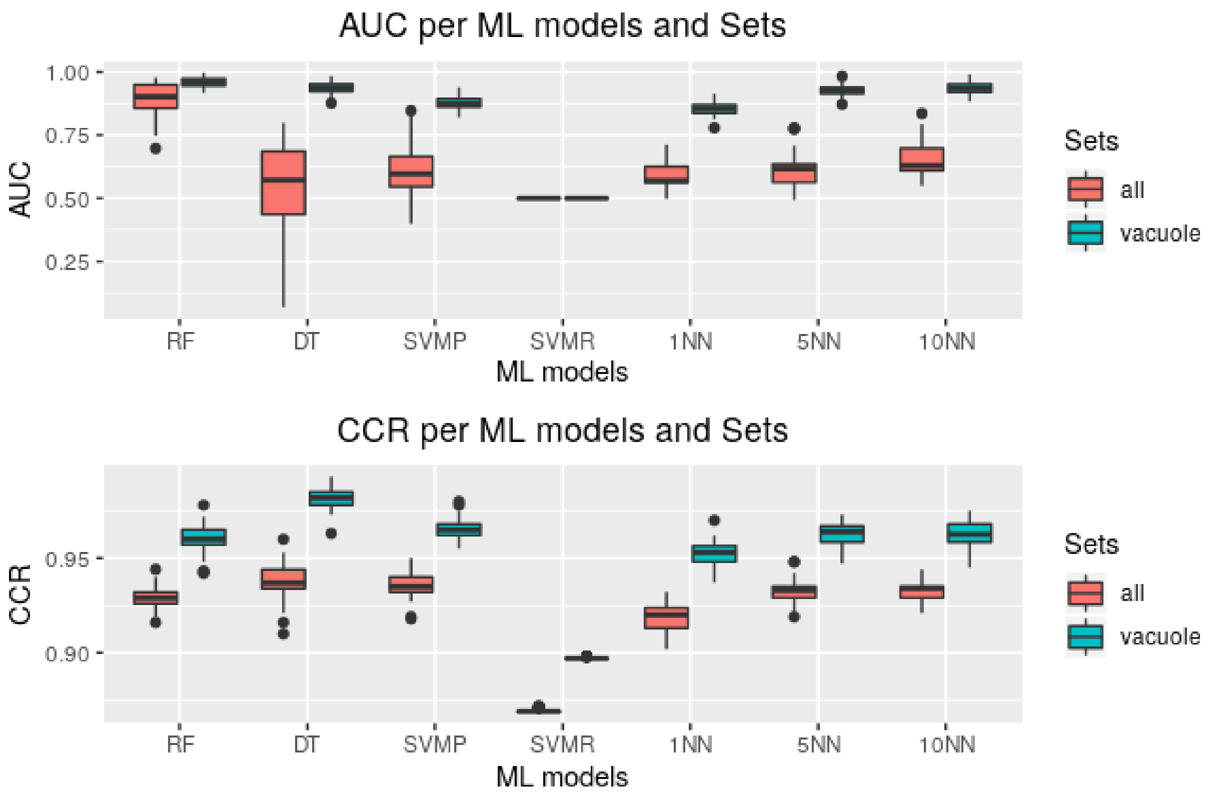


Fig. 4: Boxplot for AUC and CCR considering all the ML models in the complete dataset (all) and the datasets with normal and vacuole nuclei cells (vacuole).

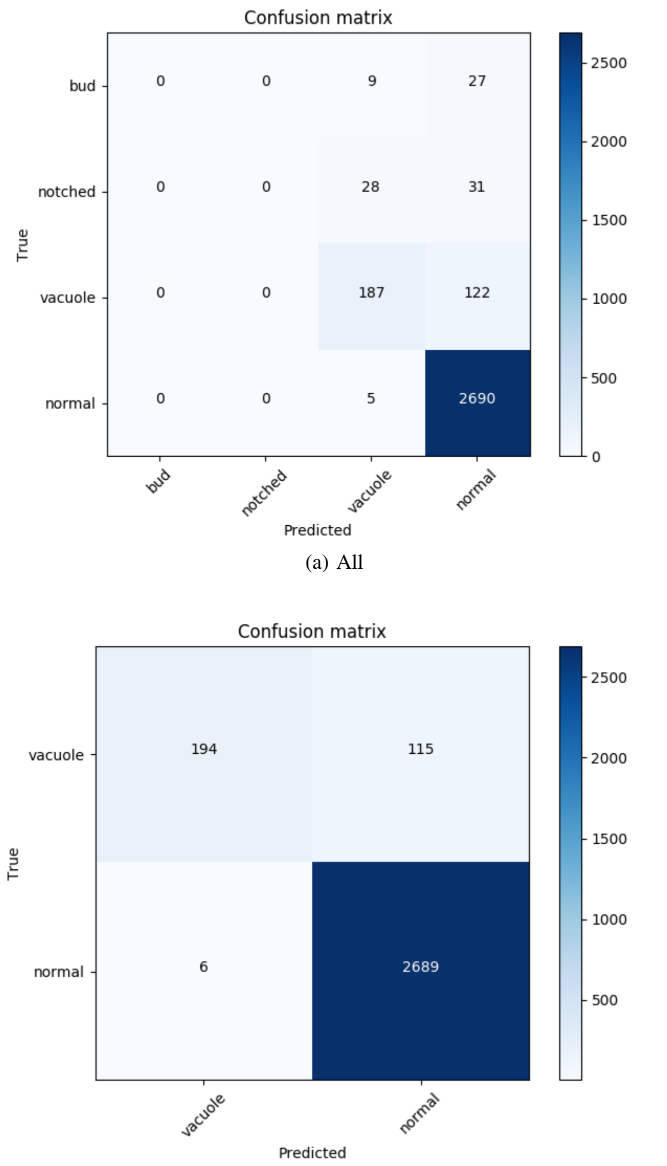
that of other algorithms. As for the CCR, decision trees and support vector machines with a polynomial kernel were ranked higher, but this result may be due to the normal class having a very high number of examples and the algorithms being biased toward this class. The low mean AUC that these two algorithms have over all classes set, 0.547 and 0.619 respectively, suggests that when the nuclear bud and notched nuclei are used, the performance gets much worse.

TABLE III: Machine learning models grouped by means using AUC and CCR metrics. Means followed by different letters in the same column differ by the Scott-Knott test at 5% significance level.

	AUC		CCR	
ML Models	all	vacuole	all	vacuole
RF	0.896 a	0.959 a	0.929 c	0.960 d
DT	0.548 d	0.936 b	0.938 a	0.981 a
SVMP	0.619 c	0.877 d	0.936 a	0.965 b
SVMR	0.500 e	0.500 f	0.869 e	0.897 f
1NN	0.576 d	0.855 e	0.919 d	0.952 e
5NN	0.615 c	0.926 c	0.933 b	0.963 c
10NN	0.656 b	0.936 b	0.932 b	0.963 c

The confusion matrices shown in Figure 5 indicate that the model has learned to ignore the bud and notched classes, so that the AUC and CCR values that RF achieved are due to the correct classifications of vacuole nuclei and normal cells. Even the classification of vacuole nuclei is hard, as 155 of the samples have been misclassified as normal cells, but a higher number, 194, has been correctly classified. Figure 6 suggests that the poor performance displayed by RF in Figure 5a is not due to a lack of modeling capacity, but rather, as stated, to the imbalance of the dataset. Machine learning algorithms often struggle in challenging scenarios due to the dearth of samples from the minority classes. That can result in biased learning towards major classes, leading to poor performance in detecting rare classes such as the aforementioned nuclear classification [21] [22] [23].

The resulting confusion matrix for the balanced dataset can be seen in Figure 6. It can be seen that RF was, actually, able to correctly classify a high percentage of the bud and notched cells, albeit still not as well as the normal cells. The vacuole class appears as the one that was confused the most, being almost equally confused with bud ($n = 8$), notched ($n = 7$) and normal ($n = 7$) classes. When all classes and an imbalanced dataset are considered, a global precision of 81.66% and a global recall equal to 60.52% is achieved. When only two classes are considered, RF achieved up to 97% precision and 62.78% recall. These values drop to 43.75% (global precision) and 38.89% (global recall) when a balanced dataset is used. Some of this difference may be due to the smaller dataset. However, one could argue that the results on the balanced version of the dataset are actually the ones within



(a) All

(b) Vacuole

Fig. 5: Confusion matrix using Random Forest for all classes (a) and for vacuole against normal (b)

expected. It shows that vacuole nuclei are the hardest to model among the four classes. By an inspection of Figure 3, it is possible to hypothesize that cells with vacuolated nuclei have more subtle visual differences.

IV. Conclusion

In this work, the classification of nuclear abnormalities in fish erythrocytes through machine learning was studied. The AUC and CCR performance metrics were used. The results showed that the models achieved higher median values for the identification of vacuole versus normal cells, while the classification performance in the complete dataset was subpar. Also, the Random Forest method consistently performed the best means, reaching notable AUC values of 0.896 and 0.959 for all sets of classes and the vacuole set, respectively.

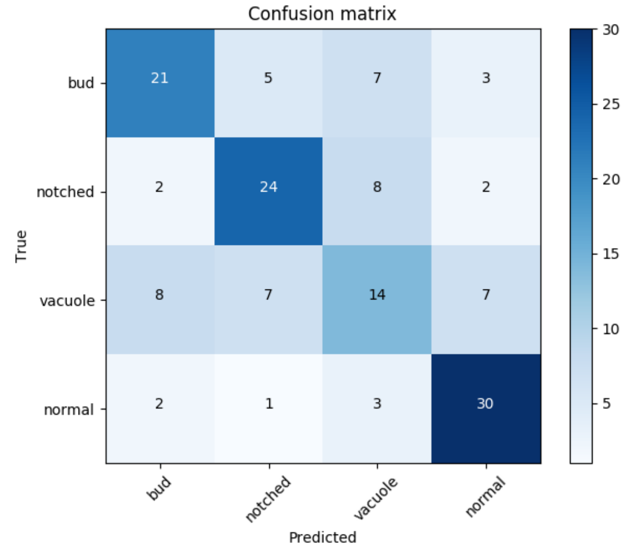


Fig. 6: Confusion matrix after balancing the data using 36 random samples per class

The AUC metrics also indicate that the performance of the nuclear bud and notched nuclei classes decreased considerably, with averages of 0.547 and 0.619, respectively, on the set of all classes. This suggests that these classes are not being learned well, and the high AUC and CCR achieved by Random Forest are due to correct classifications of vacuole nuclei and normal cells. Even the classification of vacuole nuclei presents difficulties since some samples were incorrectly classified as normal cells. On the other hand, when the dataset is balanced to contain 36 samples per class, there is a notable improvement in the capacity of the Random Forest in classifying all three abnormalities.

For further research, possible improvements are: the expansion of the dataset, by gathering more images of the least represented classes; the extraction of other features, such as variations on the chord length function [24] and on the color statistics; and finally, the use of deep learning, either in a stand-alone neural network or in a larger strategy with other techniques.

Acknowledgment

We thank Dom Bosco Catholic University (UCDB), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Process: 423256/2021-1, Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do estado de Mato Grosso do Sul (FUNDECT) – Process: 83/007.898/2023 TO: 11/2023 SIAFEM: 32834, NVIDIA Corporation for the donation of the graphics card, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Process: 2023/03870-8 and Embrapa Agricultura Digital/CNPTIA/EMBRAPA – Process: 2022/09319-9.

REFERENCES

[1] S. K. Vohra and D. Prodanov, "The active segmentation platform for microscopic image classification and segmentation," *Brain Sciences*, vol. 11, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2076-3425/11/12/1645

[2] D. Yoshida, K. Akita, and T. Higaki, "Machine learning and feature analysis of the cortical microtubule organization of arabidopsis cotyledon pavement cells," *Protoplasma*, vol. 260, pp. 1–12, 10 2022.

[3] K. Mimura, S. Minabe, K. Nakamura, K. Yasukawa, J. Ohta, and Y. Kato, "Automated detection of microfossil fish teeth from slide images using combined deep learning models," *Applied Computing and Geosciences*, vol. 16, p. 100092, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590197422000143

[4] K. Bijari, G. Valera, H. López-Schier, and G. A. Ascoli, "Quantitative neuronal morphometry by supervised and unsupervised learning," *STAR Protocols*, vol. 2, no. 4, p. 100867, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666166721005736

[5] C. S. Costa, E. C. Tetila, G. Astolfi, D. A. Sant'Ana, M. C. Brito Pache, A. B. Gonçalves, V. A. Garcia Zanoni, H. H. Picoli Nucci, O. Diemer, and H. Pistori, "A computer vision system for oocyte counting using images captured by smartphone," *Aquacultural Engineering*, vol. 87, 2019.

[6] J. d. S. Azevedo, E. d. S. Braga, and C. A. O. Ribeiro, "Nuclear abnormalities in erythrocytes and morphometric indexes in the catfish cathorops spixii (ariidae) from different sites on the southeastern brazilian coast," *Brazilian Journal of Oceanography*, vol. 60, pp. 323–330, 2012.

[7] M. Stankevičiūtė, T. Gomes, and J. A. C. González, "Nuclear abnormalities in mussel haemocytes and fish erythrocytes," 2022.

[8] J. M. Phillip, K.-S. Han, W.-C. Chen, D. Wirtz, and P.-H. Wu, "A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei," *Nature protocols*, vol. 16, no. 2, pp. 754–774, 2021.

[9] H. Narotamo, M. S. Fernandes, A. M. Moreira, S. Melo, R. Seruca, M. Silveira, and J. M. Sanches, "A machine learning approach for single cell interphase cell cycle staging," *Scientific Reports*, vol. 11, no. 1, p. 19278, 2021.

[10] S. N. Talapatra, R. Chaudhuri, and S. Ghosh, "Cellprofiler and weka tools: image analysis for fish erythrocytes shape and machine learning model algorithm accuracy prediction of dataset," *World Scientific News*, vol. 154, pp. 101–116, 2021.

[11] L. Tao and G. Xu, "Color in machine vision and its application," *CHINESE SCIENCE BULLETIN*, vol. 46, no. 17, pp. 1411–1421, SEP 2001.

[12] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[13] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[15] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, pp. 429–441(12), November 1946.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[17] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," *ACM Comput. Surv.*, vol. 51, no. 3, May 2018.

[18] B. Chakradhar, I. S. Rao, V. J. Archana, and C. V. K. Hari, "Detection of malignancy on dermis using j48 and random forest classifiers," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. IEEE, 2020, pp. 1–6.

[19] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*, vol. 52, pp. 857–900, 2019.

[20] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Applied Sciences*, vol. 1, no. 1559, 2019.

[21] P. Grimaldi, M. Lorenzati, M. Ribodino, E. Signorino, A. Buffo, and P. Berchialla, "Predicting astrocytic nuclear morphology with machine learning: A tree ensemble classifier study," *Applied Sciences*, vol. 13, no. 7, p. 4289, 2023.

[22] F. J. Moreno-Barea, L. Franco, D. Elizondo, and M. Grootveld, "Application of data augmentation techniques towards metabolomics," *Computers in Biology and Medicine*, vol. 148, p. 105916, 2022.

[23] Y. Li, J. Gong, X. Shen, M. Li, H. Zhang, F. Feng, and T. Tong, "Assessment of primary colorectal cancer ct radiomics to predict metachronous liver metastasis," *Frontiers in Oncology*, vol. 12, p. 861892, 2022.

[24] B. Wang and C. Shi, "Shape matching using chord-length function," in *Intelligent Data Engineering and Automated Learning – IDEAL 2006*, E. Corchado, H. Yin, V. Botti, and C. Fyfe, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 746–753.