

A New Method for Binary Classification of Proteins with Machine Learning

Damiano Perri[1,3(✉)], Marco Simonetti[1,3], Andrea Lombardi[2], Noelia Faginas-Lago[2], and Osvaldo Gervasi[3]

[1] Department of Mathematics and Computer Science, University of Florence, Florence, Italy
damiano.perri@unifi.it
[2] Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy
[3] Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy

Abstract. In this work we set out to find a method to classify protein structures using a Deep Learning methodology. Our Artificial Intelligence has been trained to recognize complex biomolecule structures extrapolated from the Protein Data Bank (PDB) database and reprocessed as images; for this purpose various tests have been conducted with pre-trained Convolutional Neural Networks, such as InceptionResNetV2 or InceptionV3, in order to extract significant features from these images and correctly classify the molecule. A comparative analysis of the performances of the various networks will therefore be produced.

Keywords: Machine Learning · Computational chemistry · Protein Data Bank · Convolutional Neural Network · Image processing · Orthogonal axonometry

1 Introduction

The classification of the geometric structures of proteins and the individuation of possible simple criteria to base their discrimination are complex tasks and a longstanding issue in chemical sciences. To investigate the relationships between structure and activity and for a satisfactory theoretical understanding of the protein folding process, the ability to assess the "correctness" and similarity of possible spatial arrangements of such macromolecules is a prerequisite.

The recent increased practicability of computational approaches based on Deep Learning and Neural Networks further motivates renewed efforts in such direction, since it permits one to resort to approaches based on the search for hidden patterns and regularities across large set of experimentally resolved protein structures.

In a recent paper [1], we developed an approach to the basic problem of classifying as "real" a protein given its amino acid sequence, using a Deep Learning

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12958, pp. 388–397, 2021.
https://doi.org/10.1007/978-3-030-87016-4_29

approach, based upon a Convolutional Neural Network (CNN) trained on a large set of data.

In the present paper, which is intended as a continuation of such previous work, a Convolutional Neural Network is again aimed at classifying as "true" or "false" a given structure, but the CNN has been developed after new significant improvements to the original approach for the recognition of the geometric structures. The idea was not to lose valuable spatial information regarding the shape of the protein to be examined, so we moved beyond the molecule model as a simple sequence of amino acids, to get to a more effective and realistic description preserving spatial information. To this purpose we exploited the well known suitability of Convolutional Neural Networks for image analysis, where they are particularly appreciated in the recognition of images and their characteristics, both on two-dimensional or three-dimensional objects, through the extraction of particular features from images so that different kind of objects, like people or things, can be correctly classified.

In this article we illustrate our approach to the problem using a two-dimensional representation based on 2D Convolutional Neural Networks, where any given protein is mapped into a two-dimensional grid of coloured pixels and then processed by the CNN in order to extract the relevant features and the characteristic properties to carry out the protein classification. In order to train the neural network, similarly to the previous work, the set of protein structures was obtained from the Protein Data Bank (PDB) [2] an open access repository containing data about proteins and nucleic acids' structures.

The paper is organized as follows. In Sect. 2 we briefly point out some key points about data extraction for protein classification. Section 3 illustrates the characteristics of our CNN, reports details about data extraction and processing. Preliminary applications, for training and validation, are also reported. Conclusions and perspectives are in Sect. 4.

2 Strategies for Classification of Complex Molecular Structures

For some years now, the use of Machine Learning techniques has rapidly become more and more pervasive in the world of Biology [3,4] and Chemistry [5–8], especially in the field of classification of macro-molecules that are generally found in the modeling of protein and bio-molecular structures [9]. The identification and correct assignment of the protein attribute to a generic bio-molecule is of considerable importance both for the purposes of genomic mapping and for the preparation of new and more specific groups of drugs [10–12].

Various methodologies are continuously suggested that take into consideration different aspects [13], such as chemical-physical properties [14] or geometric structure [15,16], to reach the goal. In the first case, two different approaches are possible:

1. attention is focused on a chemical-physical feature of interest and the biomolecules are labelled, also using a deep learning or SVM (Support Vector Machine) technique to obtain an accurate and automated classifier;
2. an n-dimensional vector with the descriptors of the chemical-physical properties to be examined is produced with an attached class label [17].

Several studies have confirmed that the combination of protein characteristics is preferable in order to obtain better predictive information than the use of single protein characteristics [18].

In the second case, the spatial arrangement of the various amino acid chains is evaluated for classification, both with SVM [19] and Machine Learning/Deep Learning [20,21] techniques. Today, convolutional neural networks are widely used in image analysis. They allow the extraction of features thanks to which object recognition or image classification can be performed [22–24]. Our research fits into this last channel with the aim of testing a way to correctly classify protein groups. Most of the work in this area has mainly focused on the study of the sequence and position of amino acids in the protein chain (primary or secondary sequence); our research has tried to maintain the information related to the sequences, simultaneously capturing all the geometric characteristics using 2D axonometric maps.

3 Methodology: The Architecture of the System

The system is a classic binary classifier whose task is to correctly subdivide the biomolecules given in input as "protein" or "non-protein". The data relating to the molecules to be examined are passed to the neural network as two-dimensional image processing, which faithfully reconstructs the geometric structure of the molecule itself.

3.1 Data Extraction and Processing

In order to validate our methodology, we have extracted from the PDB a sufficient number of records useful to effectively train the network; after several tests, it was found that good results were obtained with a number of samples around 3,000 units. This allowed us to select a group of proteins (equal to 2,911 molecules, with 16,924,350 amino acids) with the best images of their structure and focus our attention on the results. The whole process, from data extraction to image evaluation by the neural network, was performed in Python3, with the help of well-known libraries useful for scientific computing and data processing, such as Numpy and Pandas.

The various steps required for the generation and management of the dataset are listed below:

1) Data extraction in XML format from the PDB database, identification of the necessary information we wish to keep and its subsequent transformation into a single object managed by the Pandas library (Fig. 1)

	id	Cartn_x	Cartn_y	Cartn_z	auth_comp_id	amin_code	group_PDB	type_symbol	file
0	1	44.080	-3.124	8.654	MET	8421504	ATOM	N	190l.txt
1	2	43.646	-1.771	8.714	MET	8421504	ATOM	C	190l.txt
2	3	42.260	-1.784	9.335	MET	8421504	ATOM	C	190l.txt
3	4	41.490	-2.712	9.135	MET	8421504	ATOM	O	190l.txt
4	5	43.668	-1.224	7.269	MET	8421504	ATOM	C	190l.txt
...	...	...	...	...	...	...	...	...	...
16924345	2309	4.709	50.805	38.087	GLY	65535	ATOM	O	9abp.txt
16924346	2310	2.795	50.913	36.873	LYS	8421376	ATOM	N	9abp.txt
16924347	2311	2.496	52.321	37.153	LYS	8421376	ATOM	C	9abp.txt
16924348	2312	1.126	52.680	36.577	LYS	8421376	ATOM	C	9abp.txt
16924349	2313	0.882	52.522	35.363	LYS	8421376	ATOM	O	9abp.txt

16924350 rows × 9 columns

Fig. 1. Information extracted: spatial position, name and colour code of amino acids

2) Data cleaning, with the elimination of any duplicate records, possibly generated by the two different methods of measuring the crystal lattice structure for the molecule
3) Association of a unique RGB color code to each amino acid present in proteins (e.g. Alanine 128, Glycine 65280, Lysine 8421376), in order to visualize the structure of the molecule as an image, on which every single amino acid is coloured in a different way

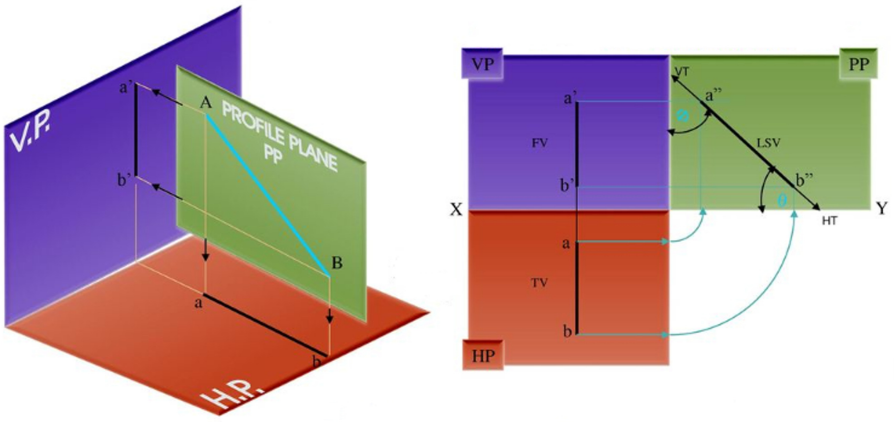


Fig. 2. Orthogonal axonometry

- Visualization of the molecule according to a multiview orthographic axonometric map with orthogonal projections on three planes (horizontal plane, vertical plane and lateral/profile plane - an example of this type of

axonometry is shown in Fig. 2), in order to split and project the whole 3D image on a flat surface, without losing the isometricity and symmetries on the x, y and z components for the individual amino acids. From the analysis of the coordinates of all the amino acids present in the dataset, it was possible through an appropriate translation for the origin of the Cartesian reference system and an integer mapping for the numerical values of the coordinates themselves to represent each single protein in the domain $D = [0, 3200]^3 \subset \mathbb{N}^3$; this allowed us to refer to each point (x, y, z) belonging to the cube D as a 3-indices tuple for the tensor with dimensions $3200 \times 3200 \times 3200$, capable of containing the entire biomolecule

- Each image has been processed to fit within the 299×299 pixel dimensions, necessary as input dimensions for a 2D convolutional neural network; in order to avoid distorting the original axonometric proportions, the figures have been carefully cut out at the edges and in the central areas, to reduce unnecessary black padding.

Therefore, for each single protein, the various amino acids were projected, as colored dots, on the three main planes: in Fig. 3 four images obtained with our method for four different proteins are shown.

4) Generation of false samples, necessary for the learning of the neural network. It was decided to proceed starting from the original images; for the single amino acids belonging to each protein we applied a mutation probability, established at the beginning of the process (in our case experimental tests made us lean towards a fixed value of 5%), which induced a colour change on the coloured points representing the amino acid: 2911 images of false proteins were thus produced. In Fig. 4 a portion of a true protein is reproduced (above) with its false analogue (below): it is possible to notice the differences due to the mutation process (pixel that occupies the same position in the two figures, but has different colors).

3.2 Training and Validation

The performances of two neural networks were analyzed: InceptionV3 and InceptionResNetV2. These networks were trained through the **transfer learning** technique which allowed us to obtain high accuracy values with a reduced number of learning epochs.

Transfer learning is the process by which a neural network learn a new task through knowledge's transfer a related task previously learned. Using transfer learning, the weight values of the convolutional layers are imported (they are generally used for feature extraction and already present as initial parameters of the network itself), and only the final layers of the neural network that are generally used for classification are trained [25,26]. Both networks are pre-trained on ImageNet which is an image dataset used for object recognition consisting of 14 million photographs.

The model used for the InceptionV3 neural network is shown in the Table 1, while the model used for the InceptionResNetV2 neural network is shown in the Table 2.

Table 1. InceptionV3 model

Layer (type)	Output Shape	Param #
inception_v3	(None, 8, 8, 2048)	21802784
flatten (Flatten)	(None, 131072)	0
dense (Dense)	(None, 64)	8388672
dense_1 (Dense)	(None, 64)	4160
dense_2 (Dense)	(None, 1)	65
Total Params:		30,195,681
Trainable Params:		30,161,249

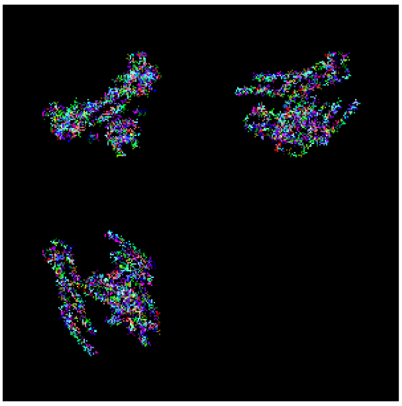
Table 2. InceptionResNetV2 model

Layer (type)	Output Shape	Param #
inception_resnet_v2	(None, 8, 8, 1536)	54336736
flatten (Flatten)	(None, 98304)	0
dense (Dense)	(None, 64)	6291520
dense_1 (Dense)	(None, 64)	4160
dense_2 (Dense)	(None, 1)	65
Total Params:		60,632,481
Trainable Params:		60,571,937

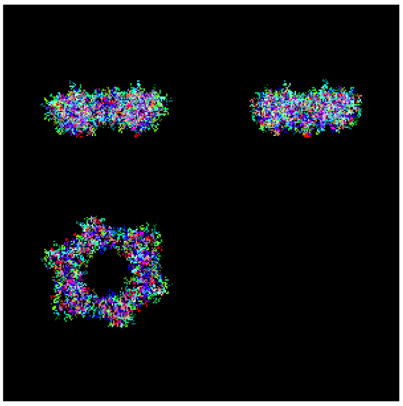
The results of the training are summarized in the Table 3, where it can be seen that the InceptionV3 network has a higher accuracy's percentage in recognizing false proteins from true ones compared to the results obtained by InceptionResNetV2. The Confusion Matrices [27] of the two models indicate that they are both able to classify proteins with good results, also we believe that a correct parameters' modulation, through the realization of ad-hoc models, can further increase the degree of accuracy. The matrices are shown in Fig. 5.

Table 3. Final results

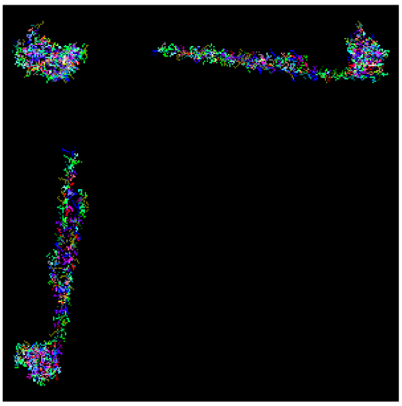
	InceptionV3	InceptionResNetV2
Accuracy Train Set	99.06%	98.36%
Accuracy Test Set	94.57%	92.14%
Model Size	461 MB	465 MB



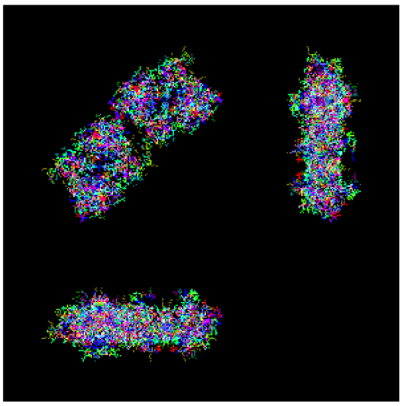
(a) 5AFR, N-terminal fragment of dynein heavy chain



(b) 5AGU, sliding clamp of Mycobacterium tuberculosis



(c) 6ABO, human XRCC4 and IFFO1 complex



(d) 6AGX, cocrystal structure of FGFR2

Fig. 3. Images of 4 proteins obtained with our representation method

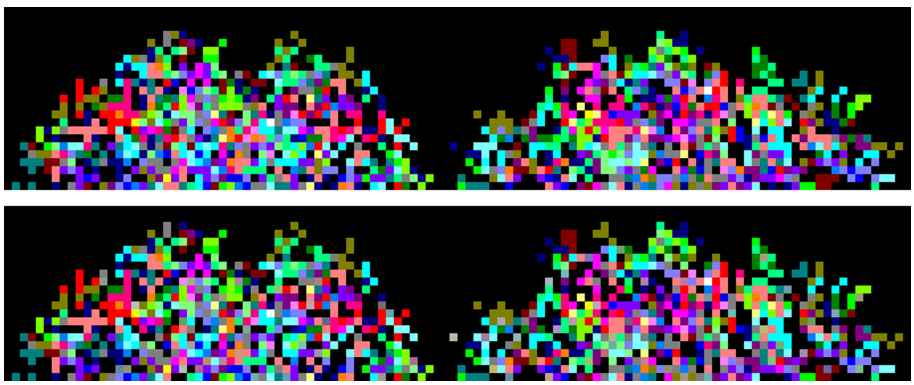
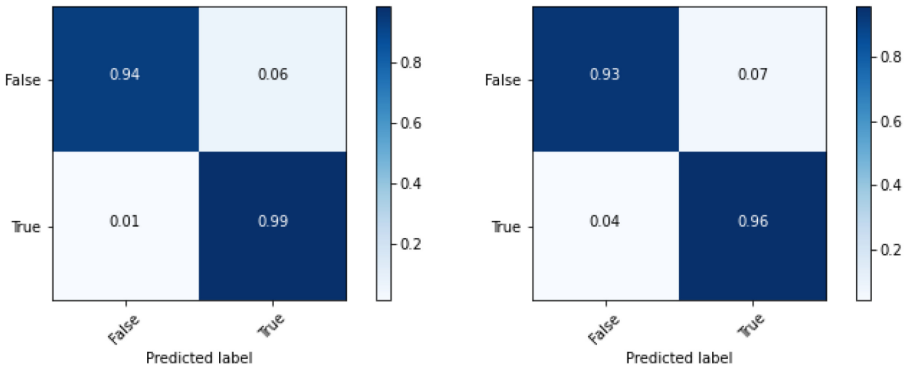
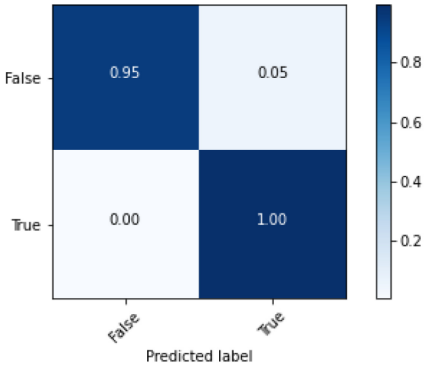


Fig. 4. Differences between the same portion of a real molecule (above) and its false analogue (below)

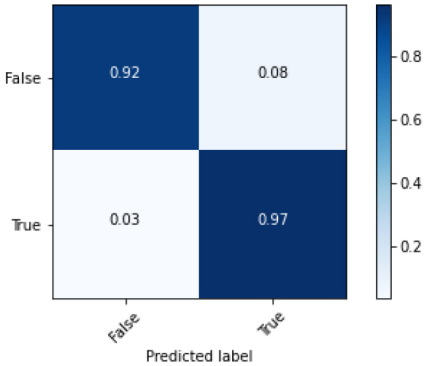


(a) InceptionV3: Training set

(b) InceptionV3: Validation set



(c) InceptionResNetV2: Training set



(d) InceptionResNetV2: Validation set

Fig. 5. Confusion matrices

4 Conclusions and Future Works

In this work we presented an improved Convolutional Neural Network for the classification of protein structure. Preliminary results appear encouraging, showing that the methodology used for the representation of the protein geometry, based on a 2D projected image associated to each molecule, retains most of the spatial information and is suitable for recognition. Also, the computational performances seems quite good, being calculations extremely fast as we provide the neural network with an image for each molecule.

However, considerations on the general nature of the CNNs used lead us to think that specifically designed neural networks could significantly improve the results, or even outperform them. A further research path worth being followed is to train the neural network using a greater number of samples, to better analyze the link between the samples' structural complexity and the classification capacity of the neural network itself. We believe that even better performance might be achieved if we developed a neural network customised for the graphical representation we proposed. Our representation is in fact very particular and

within our images the areas with a high information content are located in very specific sectors of the images. Furthermore, a personalised neural network could also reduce the size in MegaBytes of the model obtained in output.

Acknowledgments. AL and NFL thank the Dipartimento di Chimica, Biologia e Biotecnologie dell'Università di Perugia (FRB, Fondo per la Ricerca di Base 2019 and 2020) and the Italian MIUR and the University of Perugia for the financial support of the AMIS project through the program "Dipartimenti di Eccellenza". AL thanks the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma, for allocated computing time.

References

1. Perri, D., Simonetti, M., Lombardi, A., Faginas-Lago, N., Gervasi, O.: Binary classification of proteins by a machine learning approach. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12255, pp. 549–558. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58820-5_41
2. Berman, H.M., et al.: The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000). http://www.rcsb.org/
3. Cartwright, H.M.: Artificial neural networks in biology and chemistry—the evolution of a new analytical tool. In: Artificial Neural Networks, pp. 1–13 (2008)
4. Krishnan, V.G., Westhead, D.R.: A comparative study of machine-learning methods to predict the effects of single nucleotide poly- morphisms on protein function. Bioinformatics **19**(17), 2199–2209 (2003)
5. Dral, P.O.: Quantum chemistry in the age of machine learning. J. Phys. Chem. Lett. **11**(6), 2336–2347 (2020)
6. Goldman, B.B., Walters, W.P.: Machine learning in computational chemistry. Annu. Rep. Comput. Chem. **2**, 127–140 (2006)
7. Panteleev, J., Gao, H., Jia, L.: Recent applications of machine learning in medicinal chemistry. Bioorg. Med. Chem. Lett. **28**(17), 2807–2815 (2018)
8. Mater, A.C., Coote, M.L.: Deep learning in chemistry. J. Chem. Inf. Model. **59**(6), 2545–2559 (2019)
9. Bakhtiarizadeh, M.R., Moradi-Shahrbabak, M., Ebrahimi, M., Ebrahimie, E.: Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. J. Theoret. Biol. **356**, 213–222 (2014)
10. Ou-Yang, S., Jun-yan, L., Kong, X., Liang, Z., Luo, C., Jiang, H.: Computational drug discovery. Acta Pharmacologica Sinica **33**(9), 1131–1140 (2012)
11. Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W.: Computational methods in drug discovery. Pharmacol. Rev. **66**(1), 334–395 (2014)
12. Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., Ebrahimie, E.: DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discovery Today **21**(5), 718–724 (2016)
13. Rishi Das Roy and Debasis Dash: Selection of relevant features from amino acids enables development of robust classifiers. Amino Acids **46**(5), 1343–1351 (2014). https://doi.org/10.1007/s00726-014-1697-z
14. Taniguchi, M.: Combination of single-molecule electrical measurements and machine learning for the identification of single biomolecules. ACS Omega **5**(2), 959–964 (2020)

15. Brandt, S., Sittel, F., Ernst, M., Stock, G.: Machine learning of biomolecular reaction coordinates. J. Phys. Chem. Lett. **9**(9), 2144–2150 (2018)
16. Cang, Z., Mu, L., Wei, G.-W.: Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Comput. Biol. **14**(1), e1005929 (2018)
17. Michael Gromiha, M., Ahmad, S., Suwa, M.: Neural network based prediction of protein structure and function: comparison with other machine learning methods. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1739–1744. IEEE (2008)
18. Ong, S.A.K., Lin, H.H., Chen, Y.Z., Li, Z.R., Cao, Z.: Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinform. **8**(1), 1–14 (2007)
19. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z.: SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. **31**(13), 3692–3697 (2003)
20. Cui, J., et al.: Advances in exploration of machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology. Curr. Bioinform. **2**(2), 95–112 (2007)
21. Ding, C.H.Q., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics **17**(4), 349–358 (2001)
22. Perri, D., Sylos Labini, P., Gervasi, O., Tasso, S., Vella, F.: Towards a learning-based performance modeling for accelerating deep neural networks. In: Misra, S., et al. (eds.) ICCSA 2019. LNCS, vol. 11619, pp. 665–676. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24289-3_49 ISBN 978-3- 030-24289-3
23. Biondi, G., Franzoni, V., Gervasi, O., Perri, D.: An approach for improving automatic mouth emotion recognition. In: Misra, S., et al. (eds.) ICCSA 2019. LNCS, vol. 11619, pp. 649–664. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24289-3_48 ISBN 978-3-030-24289-3
24. Labini, P.S., et al.: On the anatomy of predictive models for accelerating GPU convolution kernels and beyond. ACM Trans. Archit. Code Optim. **18**(1) (2021). https://doi.org/10.1145/3434402, ISSN 1544-3566
25. Franzoni, V., Biondi, G., Perri, D., Gervasi, O.: Enhancing mouth-based emotion recognition using transfer learning. Sensors **20**(18) (2020). https://doi.org/10.3390/s20185222, https://www.mdpi.com/1424-8220/20/18/5222, ISSN 1424-8220
26. Benedetti, P., Perri, D., Simonetti, M., Gervasi, O., Reali, G., Femminella, M.: Skin cancer classification using inception network and transfer learning. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12249, pp. 536–545. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58799-4_39
27. Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E.: Confusion matrix-based feature selection. MAICS **710**, 120–127 (2011)