An Explainable Convolutional Neural Network for the Detection of Drug Abuse

Giulia Tufo[1,*,†], Meriam Zribi[1,†], Paolo Pagliuca[2,†] and Francesca Pitolli[1]

[1] *Department of Basic and Applied Sciences for Engineering, Università degli Studi Roma La Sapienza, Via Antonio Scarpa 14, Roma*

[2] *Institute of Cognitive Sciences and Technologies, National Research Council (CNR), Via Gian Domenico Romagnosi 18/A, Roma*

Abstract

The spread of Artificial Intelligence methods in many contexts is undeniable. Different models have been proposed and applied to real-world applications in sectors like economy, industry, medicine, healthcare and sports. Nevertheless, the reasons of why such techniques work are not investigated in depth, thus posing questions about explainability, transparency and trust. In this work, we introduce a novel Deep Learning approach for the problem of drug abuse detection. Specifically, we design a Convolutional Neural Network model analyzing lateral-flow tests and discriminating between normal and abnormal assays. Moreover, we provide evidence regarding the attributes that enable our model to address the considered task, aiming to identify which parts of the input exert a significant influence on the network's output. This understanding is crucial for applying our methodology in real-world scenarios. The results obtained demonstrate the validity of our approach. In particular, the proposed model achieves an excellent accuracy in the classification of the lateral-flow tests and outperforms two state-of-the-art deep networks. Additionally, we provide supporting data for the model's explainability, ensuring a precise understanding of the relationship between attributes and output, a key factor in comprehending the internal workings of the neural network.

Keywords

Drug abuse detection, Lateral-flow tests, Explainability, Convolutional Neural Networks

1. Introduction

Artificial Intelligence (AI) is a field in considerable and continuous expansion that is part of our lives and has spread to many sectors, like economy [1], industry [2], sports [3], medicine [4, 5] and healthcare [6]. Focusing on these two latter fields, AI provides a valid support for helping doctors and other professionals to make diagnosis [7] and predictions [8], explain and analyze medical data [9, 10]. Moreover, the use of assistive robots in rehabilitation and elderly monitoring is widespread nowadays [11, 12].

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

[*] Corresponding author.

[†] These authors contributed equally.

giulia.tufo@uniroma1.it (G. Tufo); meriam.zribi@uniroma1.it (M. Zribi); paolo.pagliuca@istc.cnr.it (P. Pagliuca); francesca.pitolli@uniroma1.it (F. Pitolli)

0009-0003-8187-274X (G. Tufo); 0009-0003-6232-3745 (M. Zribi); 0000-0002-3780-3347 (P. Pagliuca); 0000-0002-7159-0533 (F. Pitolli)

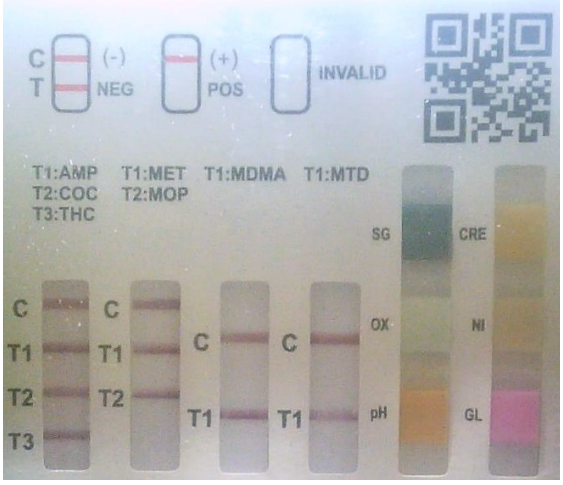

Figure 1: Example of a lateral flow test for drug abuse detection.

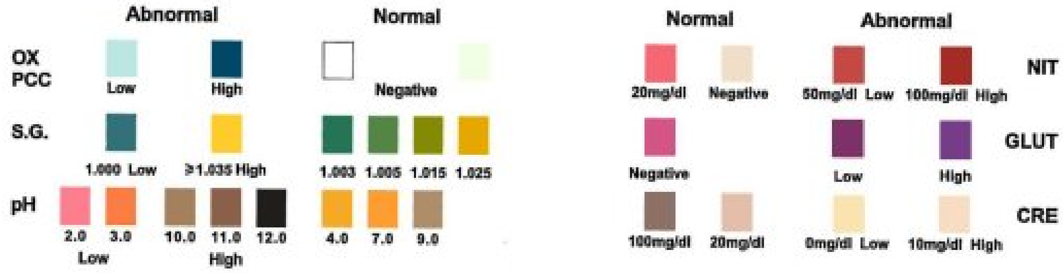

Figure 2: Adulterant guide chart. For each adulterant, the list of colors identifying normality and abnormality of the test is provided. Image taken from [13].

In particular, AI is a unique tool for analyzing huge amounts of data effectively and on time [14]. Human evaluation, constrained by various factors such as subjectivity, limited computational capacity, past and personal experiences, fatigue, stress, and data quality (such as image resolution and/or lighting conditions), may be prone to generate inaccurate predictions and/or faults. Especially in medicine and healthcare, the error minimization is paramount, since it might affect the diagnosis of potential diseases, prompt interventions, therapies for rehabilitation and other aspects. A largely applied approach in the medical data analysis and classification relies on the use of Convolutional Neural Networks (CNNs) [15–17]. CNNs enable the analysis of broad datasets containing thousands of data faster than human operators. Notwithstanding the abundance of the examples, a major concern is the lack of explainability of some models proposed in the literature, posing a challenge that even involves the developers themselves. This turns out to be critical in the fields of medicine and healthcare, where the use of explainable AI approaches is pivotal [18–20].

In this work, we analyzed the issue of detecting the presence of substances/drugs in rapid lateral-flow tests (Fig. 1) [21]. Similar works investigating this topic are those reported in

[22–24]. In particular, in [23] the authors propose an image processing algorithm combined with a Least Squares Support Vector Machine (LS-SVM) to investigate pH indicator paper assays. Their approach achieve excellent performances in terms of accuracy.

The analysis of the test results is generally made by human operators. As we stated above, the interpretation of the test is affected by factors like the subjectivity of the person and/or her/his physical and mental conditions, with consequent possible errors. Instead, we propose a novel Computer Vision (CV) approach based on the use of a deep CNN. Specifically, we employed the model introduced in [25, 26] with the addition of pooling layers [27] in the convolutional part of the network. The use of pooling allows us to reduce the complexity of the problem without losing accuracy. Indeed, pooling helps the model to become invariant to small translations of the input [27]. The model must distinguish between normal and abnormal results in lateral-flow tests analyzed for the detection of drug abuse. The primary goal of the model is to verify the suitability of the sample to ensure it has not been compromised in any way. Once the sample's suitability is confirmed, the analysis can proceed to investigate the presence of narcotic substances. The cartridge undergoes a color change upon contact with the human biological sample (urine). Based on the detected color gradation, it can be concluded whether the sample has been adulterated or not. A test is considered as "abnormal" if any of the adulterants is not compliant with the corresponding guide (Fig. 2).

While the detection of strips in rapid check tests with Deep Learning techniques has already been addressed in the literature [22, 28–30], to our knowledge the use of a CNN model to verify that the biological sample is indeed urine and has not been tampered with in the adulterant section of lateral-flow tests has not been investigated. Collected results indicate the validity of our approach: the proposed model manages to discriminate between normal and abnormal tests. Moreover, we discuss the reasons of why such model is effective, thus providing evidence of its explainability, which represents a paramount property to successfully apply the methodology in real-world scenarios. The main contributions of our work can be summarized as follows:

- we propose a novel Deep Learning (DL) approach to address the issues related to the visual inspection of lateral-flow tests, which are generally examined by human operators, with a focus on the adulterant section of the assay;
- we apply the recently proposed ConvNet3_4 model [25, 26] to discriminate between normal and abnormal tests;
- we achieve an excellent classification capability proving the validity of the model;
- we compare the model with two state-of-the-art deep networks and we demonstrate the superiority of our approach;
- we provide a thorough analysis of the relevant features extracted by the model in order to associate the proper output class to each image.

The remainder of the article is structured as follows: section 2 contains a description of the methodology we applied with respect to the considered problem. Results of our experiments are provided in section 3. Finally, our conclusions and final remarks are reported in section 4.

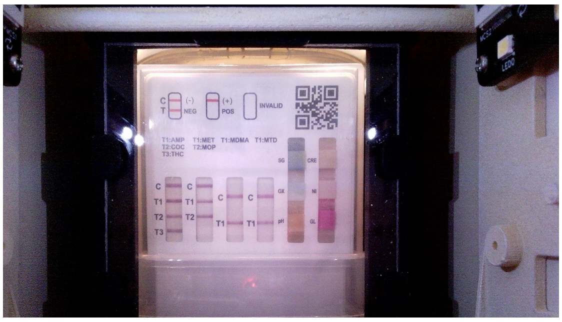

Figure 3: Image showing the entire setup: the lateral flow test is inserted into the device. Acquisition is made through a camera placed at the front.

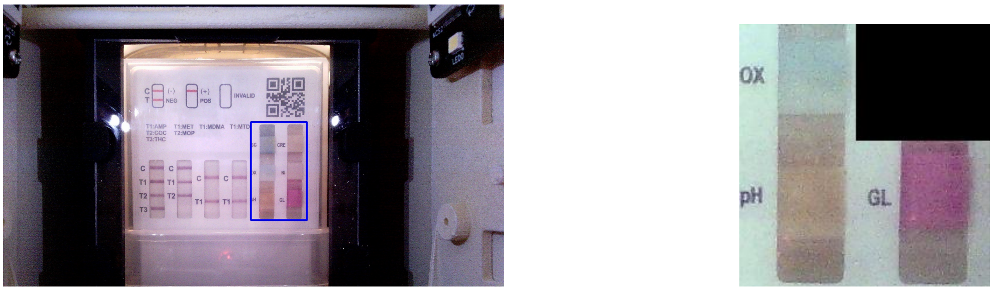

Figure 4: Left: Adulterant section of the lateral flow test (area inside the blue rectangle). **Right:** Example of image used for model training. It highlights the portion of the adulterant section considered. The adulterants taken into account are: pH, OX and GL. We filled the image with a black box in order to minimize the impact of meaningless pixels on the model's prediction.

2. Materials and methods

As stated in section 1, our study focuses on assessing the suitability of samples through the analysis of lateral-flow tests for the detection of substance abuse (Fig. 1). All images used were obtained from a specialized medical devices company and were labeled by professional laboratory technicians. An example of images obtained with this setup is shown in Fig. 3.

Our analysis focuses on the suitability of the sample by examining the portion of the test image containing the six different adulterants (see Fig. 4 left, highlighted area in the blue rectangle): Specific Gravity (SG), pH, Oxidant (OX), Creatinine (CRE), Nitrite (NI), and Glutaraldehyde (GL). Due to the presence of elements belonging to a specific class only (i.e., samples being either always normal or always abnormal), we concentrated our analysis on only three adulterants - pH, OX, and GL - for which we managed to collect data belonging to both classes. Consequently, we created a dataset consisting of 181 images. Fig. 4 right provides an example of the input image where we cropped the specific portion of interest, filling the remaining part with black pixels. The size of pictures is 215×225 pixels. Our model was then trained exclusively on images considering these components.

Given the small size of our dataset, we employed the data augmentation technique [31], which is crucial to avoid overfitting when the amount of available data is limited [32]. Furthermore, due to the difficulty to collect normal assays, the original dataset is unbalanced between the two classes, with 133 images of abnormal tests and only 48 pictures of normal assays (the ratio is around 2.77). The class imbalance problem is a major concern in Machine Learning (ML) and Deep Learning (DL) [33]. In fact, training models on unbalanced data may result in learning most from the larger class, with consequent sub-optimal performances and poor generalization capabilities (for instance, a model could associate one class to all the input data regardless of the image features). Aiming to mitigate such issue, we first apply transformations so as to balance the two types of data (see Table 1), thus obtaining 300 images equally split between the two classes, 80% of which constitute the training set and the remaining 20% are the test set. The balancing operation has been performed in order to ensure that each image and its variation(s) cannot be in both training and test sets, hence excluding the possibility of overfitting. Then, we use the *RandomAdjustSharpness* transformation [34] (parameters: $factor \in [0, 0.25, 0.5, 0.75, 1.25, 1.5, 2, 2.5, 3]; prob = 1.0$) to widen the set of input images in both training and test sets. The type of transformations employed to increase the number of data has been chosen carefully by taking into account the specific nature of the problem and the criticality of modifying image colors (see Fig. 2). Overall, the final training set consists of 2400 images, while the final test set contains 600 images.

Our model has been trained 10 times, each one starting with a different network weight initialization. The use of multiple replications minimize the risk of overestimating the model's performance due to lucky conditions. Training lasts 50 epochs, the learning rate is set to 10^{-4} and the batch size is set to 16. The model's optimizer is Adam [35] with weight decay, whose value is 10^{-2}. The size of pooling filters is 2×2. The experimental parameters have been derived from [26] and are empirically determined. Before training our model, we applied the k-fold cross-validation technique [36] to verify whether our deep network is suitable for the considered problem and mitigate the data overfitting issue [37]. We set $k = 5$ and measured the average accuracy of the model by computing the Cross-Entropy (CE) loss metric. The average CE obtained during cross-validation phase is 92.917%, that is the proposed model achieves sufficiently good categorization performances and correctly discriminates between normal and abnormal tests. Therefore, we can state that our model is suitable for the considered problem.

Aiming to demonstrate the novelty and efficacy of the proposed model, we perform a comparison with the DenseNet121 [38] and ResNet18 [39] pre-trained networks, which represent two state-of-the-art models. We choose these networks since they are characterized by a number of trainable parameters of similar orders of magnitude compared to our approach, as we will illustrate in the next section. This allows us to perform a fair evaluation of the presented model.

3. Results

In this section we provide the outcomes of our experiments. As for the k-fold cross-validation phase, we employed the CE loss as a performance measure.

Fig. 5, left shows the CE loss of the model during training. As it can be observed, the training error quickly decreases and stabilizes from the epoch 20 (see Fig. 5 left, blue curve). Conversely, the test error increases in the first 10 epochs (see Fig. 5 left, red curve), then it starts decreasing and almost stabilizes from epoch 20-25. The peak at epoch 40 (see Fig. 5 left, red curve) is due

Table 1
List of transformations applied to the original images in order to balance pictures between the two output classes (i.e., “normal” and “abnormal”). Specifically, we generate 17 images of abnormal tests and 102 pictures of normal assays. The resulting set contains 300 images equally distributed among the two classes. For further details about the transformations, the reader is referred to [34].

Output class	Type	# of transforms	Parameter
Abnormal	Center crop + pad	17	$size : (215, 205)$ $padding : 5$ $fill = 0$
Normal	Center crop + pad	48	$size : (215, 205)$ $padding : 5$ $fill = 0$
Normal	Center crop + pad	48	$size : (205, 195)$ $padding : 10$ $fill = 0$
Normal	Center crop + pad	6	$size : (205, 195)$ $padding : 15$ $fill = 0$

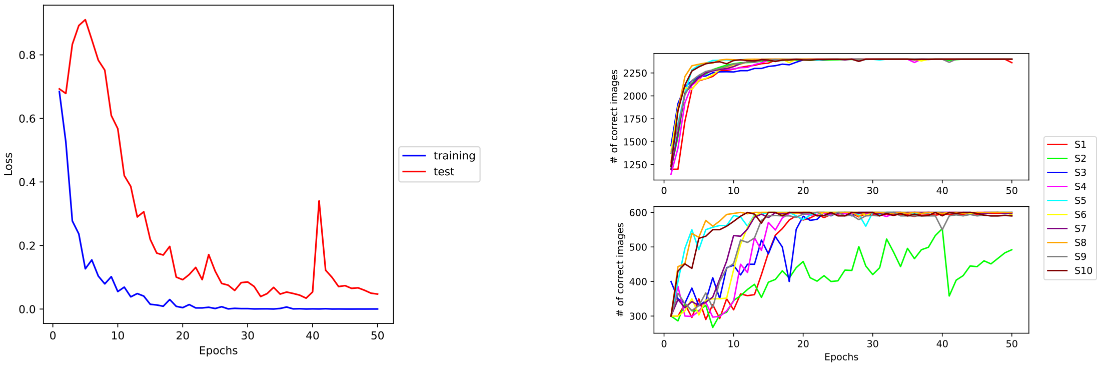

Figure 5: Model classification capability during training. **Left:** Error curve during training. Data are obtained by averaging 10 replications of the experiment. **Right:** Number of images correctly classified during training. Curves show how many images are correctly categorized as “normal” in each replication (labeled as S1, S2, ..., S10). **Top:** data referring to training set. **Bottom:** data collected on the images belonging to test set.

to a sudden increase of the error observed in one of the 10 replications (see Fig. A.1), probably related to a particular batch of images. The model achieves an average accuracy of 97.683%, i.e. a very good classification capability. Specifically, the proposed approach succeeds in correctly categorizing all the lateral-flow assays in both the training set and the test set in 4 out of 10 replications and reaches a test accuracy over 98% in 9 replications (see Fig. A.2). Overall, these results imply that our outcomes have been obtained systematically and are not due to chance or lucky initialization of the network’s weights. Instead, the proposed model is able to extract

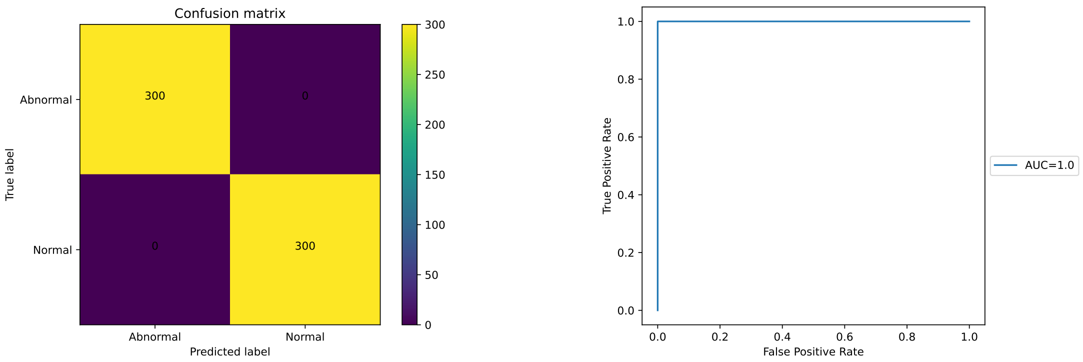

Figure 6: Analysis of the best model's categorization. **Left:** Classification results of best model found. The confusion matrix indicates the number of correctly/wrongly categorized images in the test set. Data in the main diagonal represent correct model predictions, while data outside the diagonal indicate classification errors. **Right:** ROC curve of the best model. The AUC score is indicated in the legend.

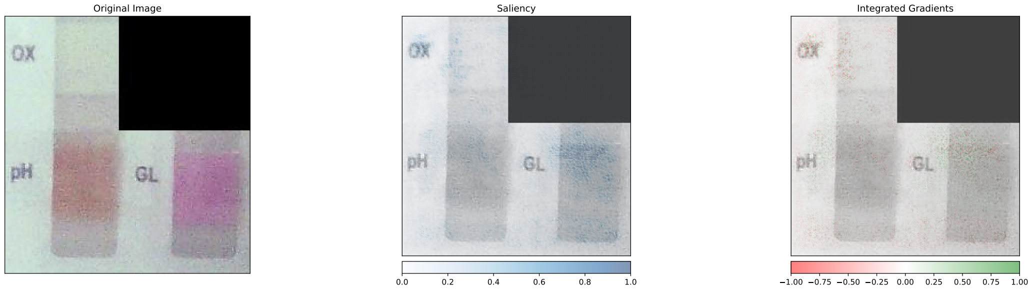

Figure 7: Feature maps of a lateral-flow test labeled as "normal". **Left:** original image of the assay. **Middle:** *Saliency* feature map. Colors in the bar range from white (absence of saliency) to blue (positive saliency). **Right:** *Integrated Gradients* feature map. Colors in the bar range from red (negative attribution) to green (positive attribution).

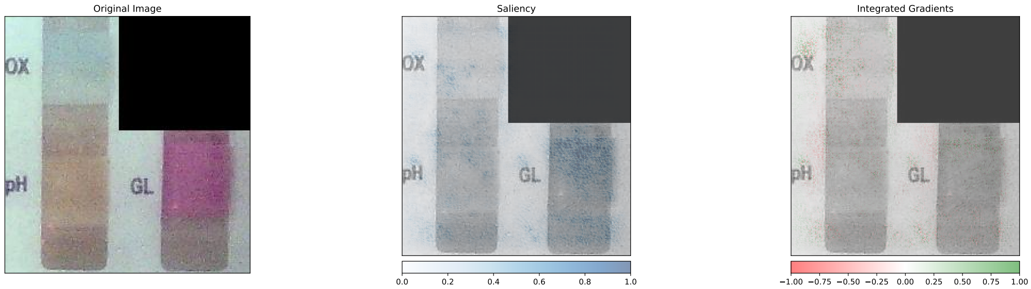

Figure 8: Feature maps of a lateral-flow test labeled as "abnormal". The abnormality of the assay is due to the OX and GL adulterants. **Left:** original image of the assay. **Middle:** *Saliency* feature map. Colors in the bar range from white (absence of saliency) to blue (positive saliency). **Right:** *Integrated Gradients* feature map. Colors in the bar range from red (negative attribution) to green (positive attribution).

Table 2
Analysis of the accuracy and the efficiency of the ConvNet3_4, DenseNet121 and ResNet18 models. With regard to ConvNet3_4, we considered the best model for the comparison. Bold values denote the best outcomes (concerning both *Model parameters* and *Training time*, the lower the better). Interestingly, the ConvNet3_4 model requires remarkably less time than the other two networks during training.

	ConvNet3_4	DenseNet121	ResNet18
Accuracy	**100%**	52.5%	50.667%
Model parameters	**3464527**	6955906	11177538
Training time (s)	**513**	11668	3790

the relevant features from the input images and predict the corresponding output class. The oscillations of the CE during training may be due to the randomization of the order of input images across epochs. If we analyze the number of lateral-flow assays correctly categorized in each replication (see Fig. 5 right), we can observe how all the networks manage to properly associate each test in the training set to the right class in around 20 epochs (Fig. 5 right, top figure). The classification of images in the test set is subject to oscillations and stabilizes after around 20 epochs except for one replication (Fig. 5 right, bottom figure). This outcome is not surprising since the latter set is used as a tool for validating the model. By considering the best model only, Fig. 6 left explains whether and how the model categorizes the lateral-flow assays in the test set. The latter consists of 300 images of abnormal tests and 300 pictures of normal assays (see Fig. 6 left). As it can be seen, the model manages to correctly classify all the 600 images in the test set. Fig. 6 right illustrates the Receiver Operating Characteristic (ROC) curve of the best model, which plots the true positive rate against the false positive rate. Because the Area Under Curve (AUC) is 1.0, our best model corresponds to a perfect classifier.

As we mentioned above, we compared the outcomes of our model with those achieved with the DenseNet121 and ResNet18 networks. The analysis is illustrated in Table 2 and reveals that the ConvNet3_4 model is notably superior to both DenseNet121 and ResNet18 with respect to the accuracy: pre-trained networks manage to correctly classify only around half of the images, i.e. they are not able to discriminate between the two possible output classes (see also Fig. B.1). The result is in line with those reported in [26]. Moreover, our ConvNet3_4 model is also remarkably better than DenseNet121 and ResNet18 in terms of efficiency (see the significant discrepancy concerning the training time in Table 2), which represents a pivotal property for a model applicability in real scenarios.

The results we presented so far demonstrate that our model is suitable to address the considered problem and achieve excellent performances in terms of classification capability. Nonetheless, as we stated in section 1, a worthwhile aspect in the fields of medicine and healthcare is the explainability of the proposed models, which is necessary to practically employ them in real-application cases. To this end, we performed a feature analysis by using two different techniques: *Saliency* [40] and *Integrated Gradients (IG)* [41]. Both methods are widely employed to interpret the outcomes of a model's classification [42–44]. The former method allows the identification of the parts of the image contributing more to the output prediction. An example of the feature map extracted through the *Saliency* method is shown in Fig. 7 middle. Saliency values range from 0 (absence of saliency) to 1 (positive saliency), as indicated in Fig. 7 middle. For a more detailed description of the *Saliency* algorithm, the reader is referred to [40]. Conversely, the IG method identifies the regions of the image that most influenced the model's classification decision by considering the entire input-output trajectory and the reference input distribution (baselines) used in the attribution calculation. As pointed out in [41], the IG method first considers the input image x and a baseline x', which is an input characterized by absence of features. Then, a straightline path from x' to x is taken into account. IG computes the gradients at all points along such path. The integrated gradients are obtained as the cumulative sum of these gradients. Put more formally, if we denote our CNN model with $F\colon \mathbb{R}^n \to [0,1]$, the integrated gradients along the i^{th} dimension is calculated as:

$$IG_i(x) = (x_i - {x'}_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{1}$$

where $\frac{\partial F(x)}{\partial x_i}$ indicates the gradient of $F(x)$ along the i^{th} dimension. Overall, the IG method enables to detect the portions of the picture providing positive (parts in green, see Fig. 7 right) and negative (parts in red, see Fig. 7 right) contributions to the output prediction.

Fig. 7 shows the image of a lateral-flow test categorized as "normal" (Fig. 7 left), the *Saliency* feature map (Fig. 7 middle) and the *Integrated Gradients* heat map (Fig. 7 right). Fig. 8 contains the same data for an "abnormal" assay. The colorbars below the heat maps specify respectively the intensity of the saliency and the magnitude of the importance attribution of each region of the image to the model's prediction.

By examining the feature maps associated to a normal lateral-flow test (Fig. 7), we can observe that the *Saliency* method returns a positive saliency for the pH and GL adulterants and a slightly positive saliency for the OX adulterant (Fig. 7 middle). Concerning the *Integrated Gradients* technique, the heat map displays a positive attribution for the pH and GL adulterants and a slightly positive attribution for the OX adulterant (see Fig. 7 right). Therefore, in the case of a normal assay, the model assigns the same importance to the portions of the image containing the considered adulterants. This outcome is in line with the actual test result.

If we look at the relevant features highlighted by the *Integrated Gradients* technique with regard to an abnormal assay, we can see that a slightly positive attribution is conferred to the pH, OX and GL adulterants (Fig. 8 right). As far as the *Saliency* algorithm is concerned, it returns a positive saliency for the GL adulterant and a slightly positive attribution for the pH and OX adulterants (Fig. 8 middle). Also in this case, the results are coherent with the actual test outcome, which indicates non-compliance for the OX and GL adulterants. Overall, our findings suggest that the proposed model primarily relies on the portions of the image containing the pH, OX and GL adulterants. However, the amount of contribution strongly depends on the test result (e.g., normal or abnormal) and the specific color gradation of the adulterants. Indeed, except for one case, the OX adulterant is characterized by soft nuances tending to be as similar as the background color of the image. Similarly, the normality and abnormality of the pH adulterant are defined based on subtle color gradations. Therefore, distinguishing between the two cases might be challenging even for laboratory operators. Finally, it is worth noting that cropping the picture does not affect the classification capability of the model. Indeed, the use of black pixels providing no information allows the model to focus only on the remaining parts of the image, which contains the relevant data.

To summarize, our outcomes demonstrate the capability of the ConvNet3_4 model to extract the most relevant features of the input image in order to generate a precise prediction of the output class. In particular, the identification of the portions containing the adulterants as the key elements of the input image implies that the model is capable of making assumptions from a medical point of view. Indeed, discriminating between the color nuances of the considered adulterants (see Fig. 2) is far from trivial even for experienced and well-trained operators.

4. Discussion and conclusions

The spread of AI methods poses questions about the explainability of such models, particularly when they are applied in real-world contexts. Especially in medicine and healthcare, using explainable and trustworthy approaches is paramount in order to help doctors and other professionals to make diagnoses of possible diseases, design adequate therapies for prevention or rehabilitation, explain and collect historical data. Generally, analyzing huge amount of medical data is addressed through DL methods and CNNs represent a widespread tool, although they are often tailored to specific applications. This represents a major concern in the possibility to develop cross-cutting tools. In this work, we proposed a novel approach for the problem of automatically classify lateral-flow tests for drug abuse detection. Specifically, we considered the adulterant section of an assay and we trained a CNN model for the ability to categorize tests (i.e., normal or abnormal) by analyzing the pH, OX and GL adulterants only. We used the network introduced in [25, 26] with the addition of pooling layers in the convolutional part of the model. The use of pooling enables the development of slim networks that can be used in real-world scenarios, particularly when dealing with limited hardware resources. We verified the suitability of the model through a 5-fold cross-validation and we ran the training 10 times. We collected promising results on the chosen task, with an excellent average accuracy. The proposed approach is also notably superior to two state-of-the-art deep networks. Moreover, we provided an explainability of our model by performing a feature analysis. Our outcomes reveal the importance of some portions of the input image (those containing the adulterants), while other parts affect the final prediction only partially.

In spite of the good results we achieved, further research is needed to generalize our approach. First, we are collecting samples so as to broaden our analyses to the SG, CRE and NI adulterants, which are out of the scope of this work, and increase the size of our dataset. Owning a significant high number of images is pivotal to apply the model in a real case, since medical data usually include hundreds or thousands of pictures. However, extending the analysis to all adulterants implies considering input images of different sizes, with an unavoidable effect on the model's performance. Furthermore, the ConvNet3_4 model proves effective at dealing with the considered problem, with a very good classification capability, and outperforms the DenseNet121 and ResNet18 pre-trained networks. Nonetheless, we observe the presence of oscillations during training due to the sensitivity to specific batches of input images. Aiming to address such undesired behavior, in the future we will consider the possibility to adapt the learning rate and/or the weight decay during training. In addition, with respect to the explainability issue, we provide evidence of the reasons behind the success of our model. Nonetheless, because the visualization of feature maps revealed the model identifying sometimes as regions of interest those that should not influence it (e.g., the portions of the cartridge container surrounding the adulterants, see Figs. 7 - 8), in future developments we plan to create a mask. This mask, overlaid on the original image, will eliminate areas of low interest, allowing the model to focus solely on relevant regions. Finally, we are investigating the applicability of our model to other datasets in the medical and health care fields with the aim to generalize the validity of the approach.

References

[1] M. Jrad, A role of artificial intelligence in the context of economy: Bibliometric analysis and systematic literature review, International Journal of Membrane Science and Technology 10 (2023) 1563–86.

[2] Z. Jan, F. Ahamed, W. Mayer, N. Patel, G. Grossmann, M. Stumptner, A. Kuusk, Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities, Expert Systems with Applications 216 (2023) 119456.

[3] D. Araújo, M. Couceiro, L. Seifert, H. Sarmento, K. Davids, Artificial intelligence in sport performance analysis, Routledge, 2021.

[4] C. J. Haug, J. M. Drazen, Artificial intelligence and machine learning in clinical medicine, 2023, New England Journal of Medicine 388 (2023) 1201–1208.

[5] O. Marques, Artificial intelligence and medicine: The big picture, in: AI for Radiology, CRC Press, 2024, pp. 1–17.

[6] D. Houfani, S. Slatnia, O. Kazar, H. Saouli, A. Merizig, Artificial intelligence in healthcare: a review on predicting clinical needs, International Journal of Healthcare Management 15 (2022) 267–275.

[7] J. G. Richens, A. Buchard, Artificial intelligence for medical diagnosis, in: Artificial Intelligence in Medicine, Springer, 2022, pp. 181–201.

[8] R. G. Nadakinamani, A. Reyana, S. Kautish, A. Vibith, Y. Gupta, S. F. Abdelwahab, A. W. Mohamed, et al., Clinical data analysis for prediction of cardiovascular disease using machine learning techniques, Computational intelligence and neuroscience 2022 (2022).

[9] F. Khader, T. Han, G. Müller-Franzes, L. Huck, P. Schad, S. Keil, E. Barzakova, M. Schulze-Hagen, F. Pedersoli, V. Schulz, et al., Artificial intelligence for clinical interpretation of bedside chest radiographs, Radiology 307 (2022) e220510.

[10] J. Tveit, H. Aurlien, S. Plis, V. D. Calhoun, W. O. Tatum, D. L. Schomer, V. Arntsen, F. Cox, F. Fahoum, W. B. Gallentine, et al., Automated interpretation of clinical electroencephalograms using artificial intelligence, JAMA neurology (2023).

[11] S. Coşar, M. Fernandez-Carmona, R. Agrigoroaie, J. Pages, F. Ferland, F. Zhao, S. Yue, N. Bellotto, A. Tapus, Enrichme: Perception and interaction of an assistive robot for the elderly at home, International Journal of Social Robotics 12 (2020) 779–805.

[12] G. D'Onofrio, D. Sancarlo, M. Raciti, D. Reforgiato, A. Mangiacotti, A. Russo, F. Ricciardi, A. Vitanza, F. Cantucci, V. Presutti, et al., Mario project: experimentation in the hospital setting, in: Ambient Assisted Living: Italian Forum 2017 8, Springer, 2019, pp. 289–303.

[13] Craig Medical: Adulterant Validity Chart Interpretation, Rapidcheck pro 10 dsc with adulterant check, https://www.craigmedical.com/Drug_5Panel_DSC-Adulterant.htm, 2024.

[14] H.-J. Jang, K.-O. Cho, Applications of deep learning for the analysis of medical data, Archives of pharmacal research 42 (2019) 492–504.

[15] D. Sarvamangala, R. V. Kulkarni, Convolutional neural networks in medical image understanding: a survey, Evolutionary intelligence 15 (2022) 1–22.

[16] X. Yao, X. Wang, S.-H. Wang, Y.-D. Zhang, A comprehensive survey on convolutional neural network in medical image analysis, Multimedia Tools and Applications 81 (2022) 41361–41405.

[17] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, M. J. Deen, Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives, Neurocomputing 444 (2021) 92–110.

[18] S. Reddy, Explainability and artificial intelligence in medicine, The Lancet Digital Health 4 (2022) e214–e215.

[19] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 5–22.

[20] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Medical Image Analysis 79 (2022) 102470.

[21] G. Tufo, M. Zribi, F. Pitolli, P. Pagliuca, Advanced computer vision techniques for drug abuse detection, in: 21st edition of the IMACS world congress (IMACS2023), 2023, p. 226.

[22] A. Carrio, C. Sampedro, J. L. Sanchez-Lopez, M. Pimienta, P. Campoy, Automated low-cost smartphone-based lateral flow saliva test reader for drugs-of-abuse detection, Sensors 15 (2015) 29569–29593.

[23] M. H. Tania, K. T. Lwin, A. M. Shabut, M. Najlah, J. Chin, M. A. Hossain, Intelligent image-based colourimetric tests using machine learning framework for lateral flow assays, Expert Systems with Applications 139 (2020) 112843.

[24] V. Turbé, C. Herbst, T. Mngomezulu, S. Meshkinfamfard, N. Dlamini, T. Mhlongo, T. Smit, V. Cherepanova, K. Shimada, J. Budd, et al., Deep learning of hiv field-based rapid tests, Nature medicine 27 (2021) 1165–1170.

[25] M. Zribi, P. Pagliuca, F. Pitolli, Convolutional neural networks for the automatic control of consumables for analytical laboratories, in: BUILD-IT2023 worskhop, 2023, pp. 95–97.

[26] M. Zribi, P. Pagliuca, F. Pitolli, A computer vision-based quality assessment technique for the automatic control of consumables for analytical laboratories, Expert Systems with Applications (2024, in press).

[27] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[28] H. J. Min, H. A. Mina, A. J. Deering, E. Bae, Development of a smartphone-based lateral-flow imaging system using machine-learning classifiers for detection of salmonella spp., Journal of Microbiological Methods 188 (2021) 106288.

[29] S. Yan, C. Liu, S. Fang, J. Ma, J. Qiu, D. Xu, L. Li, J. Yu, D. Li, Q. Liu, Sers-based lateral flow assay combined with machine learning for highly sensitive quantitative analysis of escherichia coli o157: H7, Analytical and Bioanalytical Chemistry 412 (2020) 7881–7890.

[30] Y. Zha, Y. Li, J. Zhou, X. Liu, K. S. Park, Y. Zhou, Dual-mode fluorescent/intelligent lateral flow immunoassay based on machine learning algorithm for ultrasensitive analysis of chloroacetamide herbicides, Analytical Chemistry (2024).

[31] M. A. Tanner, W. H. Wong, The calculation of posterior distributions by data augmentation,

Journal of the American statistical Association 82 (1987) 528–540.

[32] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.

[33] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (2002) 429–449.

[34] PyTorch Illustration of transforms, https://pytorch.org/vision/main/auto_examples/plot_transforms.html, 2024.

[35] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[36] C. Schaffer, Selecting a classification method by cross-validation, Machine learning 13 (1993) 135–143.

[37] L. A. Yates, Z. Aandahl, S. A. Richards, B. W. Brook, Cross validation for model selection: a review with examples from ecology, Ecological Monographs 93 (2023) e1557.

[38] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4700–4708.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[40] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[41] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[42] I. Čík, A. D. Rasamoelina, M. Mach, P. Sinčák, Explaining deep neural network using layer-wise relevance propagation and integrated gradients, in: 2021 IEEE 19th world symposium on applied machine intelligence and informatics (SAMI), IEEE, 2021, pp. 000381–000386.

[43] G. Li, Y. Yu, Visual saliency detection based on multiscale deep cnn features, IEEE transactions on image processing 25 (2016) 5012–5024.

[44] M. Schwegler, C. Müller, A. Reiterer, Integrated gradients for feature assessment in point cloud-based data sets, Algorithms 16 (2023) 316.

Appendix

A. ConvNet3_4

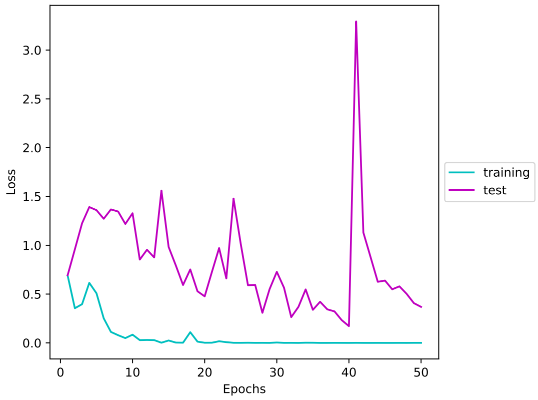

Figure A.1: Error curve of the worst ConvNet3_4 model out of 10 replications. The training error decreases and goes to 0 after few epochs. Instead, the test error increases at the beginning of training and oscillates; the sudden increase at epoch 40 prevents this replication from achieving a good accuracy on images in the test set.

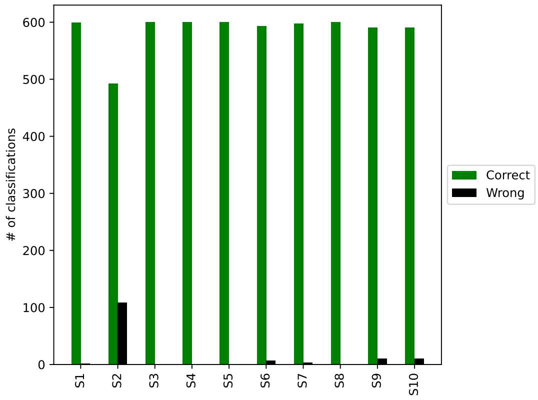

Figure A.2: Number of correct and wrong classifications with respect to the images in the test set. Data refer to 10 replications of the experiment.

B. Pre-trained models

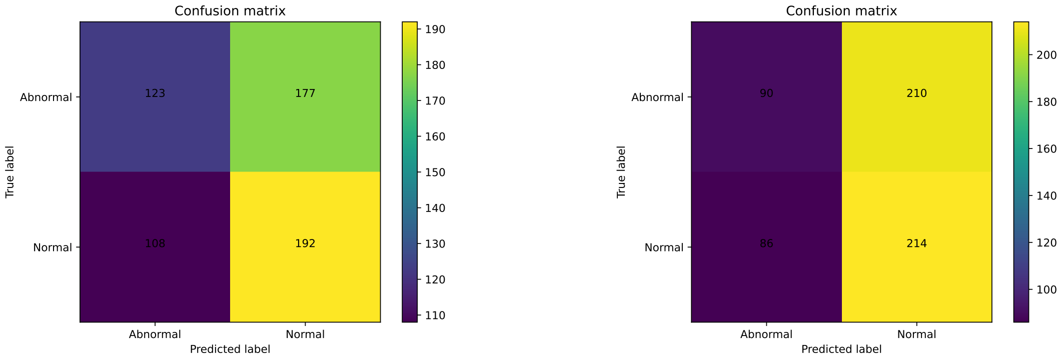

Figure B.1: Analysis of the classification capability of pre-trained networks on the images belonging to the test set. Data in the main diagonal denote correct predictions, while values outside the diagonal represent faulty categorizations. **Left:** DenseNet121 model. **Right:** ResNet18 model.