Multi-output Deep Learning and Explainable AI methods for Lesion-Symptom Mapping

Andrea García-Tejedor Bilbao-Goyoaga
Imaging and Oncology Division
University Medical Center (UMC) Utrecht
Utrecht, Netherlands
a.garcia-tejedor@students.uu.nl

***Abstract*—Cerebral small vessel disease is a common condition that can lead to stroke and dementia, which cause disability in different cognitive domains. Lesion-Symptom Mapping (LSM) techniques aim to find correlations between the affected locations of the brain and cognition decline. Voxel-based and Support Vector Regression LSM are the current golden standards, but they have limitations in terms of model complexity and interpretability. Deep Learning (DL) methods have the potential of building complex models, which can be useful to tackle this challenge. This study proposes a DL pipeline that can predict multiple artificial cognition scores from lesion MR images and correctly identify the brain locations that are most relevant to make specific score predictions using explainable Artificial Intelligence (xAI) methods. The study analyzes the performance of single and multi-output models and explores different xAI methods (Integrated Gradients, Gradient Shap and Occlusion) to understand the information each can provide. Overall, this project demonstrates that DL techniques can satisfactorily predict multiple regression outputs from segmented lesion MR images and identify the key regions that affect each score, which could be used in the field of LSM to understand the underlying brain mechanisms that contribute to various neurological and psychiatric disorders.**

***Index Terms*—Vascular Cognitive Impairment, Lesion-symptom mapping, Deep Learning, Explainable AI, Neuroimaging, MRI**

1. Introduction

Cerebral small vessel disease (SVD) is the major cause of stroke and dementia, most commonly present in community-dwelling individuals [1]. Stroke is a leading cause of Vascular Cognitive Impairment (VCI), causing long-term disabilities and poor quality of life in over half of the patients [2]. VCI can occur immediately after a stroke, or it can develop over time as a result of ongoing complications with blood flow to the brain. It can bring a range of cognitive and behavioural problems in the domains of memory, attention, language, and executive function [3].

Predicting cognitive decline associated with SVD can be difficult, as it manifests as WMH, infarcts, microbleeds and other lesions in Magnetic Resonance Imaging (MRI) scans [2]. Several studies have shown that cortical involvement, presence of multiple acute infarcts, total infarct volume, and left cerebral hemispheric location (vs right) are associated with VCI [4], [5], [6]. Recent studies on Lesion-Symptom Mapping (LSM) have shown that the locations of these lesions are more correlated to cognition than the total lesion volume, highlighting the importance of studying this relation [7], [8]. LSM is a neuroimaging technique used to identify the lesioned brain regions and determine the relationship between the locations of these regions and specific symptoms or behaviours [9], [3], [10]. This information can be useful not only in developing targeted therapies or rehabilitation strategies to help the patient recover cognitive and functional abilities, but also for understanding the underlying brain mechanisms that contribute to various neurological and psychiatric disorders.

Studies that evaluated lesion location at the voxel level-have shown various correlations between lesion locations and cognitive deficits [11], [12], [13]. However, these studies had an incomplete brain coverage (less than 20% of total brain volume), leaving most of the brain unexplored [13]. With the objective of overcoming this limitation and collect data to increase lesion coverage, the Meta-analyses on Strategic Lesion Locations for Vascular Cognitive Impairment using Lesion-Symptom Mapping (Meta VCI Map) [9] consortium was initiated. Different studies have been carried out using Support Vector Regression with this dataset, which have identified significant associations between total Montreal Cognitive Assessment scores and certain locations of the brain [13][14].

Most studies have focused on tackling LSM challenges using statistical and Machine Learning methods [9], [15], [16], [17], [18], which have limitations in terms of model complexity, interpretability and generalization power. Very little research has been done with a Deep Learning (DL) approach. Convolutional Neural Networks (CNNs) have the potential of building complex models and extracting features from any kind of image data, and have been demonstrated to be a very suitable tool for medical image analysis [19]. First attempts for predicting language disorders from MRI data have outperformed the already existing Machine Learning techniques [20]. However, no study has been carried out for the prediction of multiple outcomes or for mapping these to lesion locations using DL. In actual fact, the research in multi-class regression is still very limited, and there are barely any examples of use in neuroimaging [21].

The drawback of DL-based systems is usually the ”black box” concept since the interpretability of the model is compromised by its complexity. Accessing the procedures that led to a particular output in predictive algorithms is challenging,

so new explainability techniques have emerged to show which features of the input contribute most, and they have proven to have many applications in medical image analysis [22], [23].

This project aims to research how DL techniques can be used to extend the current methods used to research LSM, exploring the advantages and limitations of using CNNs to find lesion-symptom correlations. More specifically, this project proposes a DL model that is not only capable of predicting multiple simulated cognitive scores from lesion MRI images of patients caused by SVD, but can also identify the locations in the brain that have been used to compute the artificial scores using explainability methods. Multi-output models will be analyzed to see if their performance is suitable for this task, and different explainability approaches will be implemented to understand which information each can provide.

2. Materials & methods

This study consists of two main parts. The first part aims to explore the existing explainability methods and apply them to multi-output prediction models with a simple open-source dataset consisting of 2D images of faces and the corresponding labels for age, sex, and ethnicity. The overall objective of this part is to prove that all the explored techniques work for multiple-output models and to have an insight into the expected output for the different explainability methods.
In the second part, a model was trained to predict multiple scores from a single 3D binary image. Several explainability methods were implemented to identify which locations of the image are most relevant for decision-making. The model was trained with artificial cognition scores calculated for images in the TRACE-VCI dataset, to validate that it is capable of identifying which locations of the lesions are responsible for specific outcomes.

All code was written in Python using Pytorch for the DL part.

2.1 UTKFace dataset

For this analysis, UTKFace dataset from GitHub [24] was used. The dataset consists of over 20,000 face images with annotations of biological age, sex, and ethnicity. Images were pre-processed to the same size (200x200) and the faces were aligned and centred, from which a subset of 9000 was selected for the experiments.

2.1.1 Single-output VS multi-output

To predict multiple scores from a single image, different single-output models that make separate predictions can be used for each output. However, this method does not take into account possible intercorrelations between data and is very inefficient. A multiple-output model can learn to predict each output independently, while also considering the interdependence between the outputs. The performances of three single-output models were compared to the one of a multi-output model, using a pre-trained ResNet18 as architecture [25], as it has shown to have an optimal performance in similar tasks [26].

Label	Age	Sex	Ethnicity
Output Type	Regression	Binary classification	Multi-class classification
Values	[0:120]	Male/Female	Black/White/Asian/Indian
Activation Function	Leaky ReLU	Sigmoid	SoftMax
Loss function	Mean Squared Error	Binary CrossEntropy	CrossEntropy

TABLE I: *Specification on optimization problem type, values, last activation function and loss function used for each label*

The specific parameters of the model were established after tuning different combinations of hyperparameters using *wandb.ai* library [27], choosing 9000 samples for training, with a batch size of 16, Adam as optimizer and Mean Squared Error (MSE) as loss function in ResNet18 architecture, and applying linear normalization as pre-processing to set the image values between [0,1]. Table II is a summary of the parameters used to train the models.

Model parameters	
Training size	9000
Pretrain	imagenet
Preprocessing	Image normalization [0:1]
Architecture	ResNet18
Batch Size	16
Optimizer	Adam

TABLE II: *Model parameters for age, sex and ethnicity prediction*

Parameters were the same for all models, except for a modification in the last layer of the architecture. In the multi-output model, the last layer was split into three branches of the same characteristics as the single-output ones, outputting three different values, and each one gets optimized using a different loss function: MSE for age, Binary Cross-Entropy for sex and Categorical Cross-Entropy for ethnicity. Figure 1 shows a representation of the multi-output model and Table I shows the specifications for the last layer for each of the predictions.

2.1.2 Explanations on predictions

It is challenging to understand why such complex models make specific predictions. Explainable AI algorithms can be applied to CNNs to create saliency maps that highlight the critical regions in the image for making a specific prediction, and these maps can be visually inspected to see if they correspond to the real features that contribute to decisions, or if they are incorrect and the good predictions were obtained because of dataset bias. In this part, Gradient SHAP, Integrated Gradients and Occlusion explainability techniques were implemented to analyze the visual saliency maps they provide for single-output and multi-output models [28].

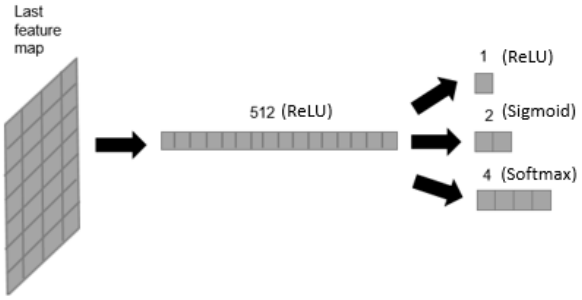

Fig. 1: *Representation of the last linear layers of the ResNet18 architecture with their corresponding activation function. The last feature map is flattened into a linear layer with output size 512 and then branched in 3 outputs.*

The basic idea of Integrated Gradients is to approximate the contribution of each input feature to the model's prediction by integrating the gradients of the model's output with respect to the inputs. The approximation starts from a baseline image and is computed by taking the average of the gradients over a path and multiplying it by the difference between the actual and baseline inputs. The end result is an attribution score for each input feature [29].

GradientSHAP is based on the concept of Shapley values from cooperative game theory. The basic idea is to calculate the contribution of each feature to a prediction by averaging the gradient of the model's output with respect to that feature over all possible combinations of the remaining features. The contribution of each feature is then calculated as the difference between the prediction made using the full set of features and the prediction made using all features except the feature of interest [30].

The Occlusion method is based on selecting windows of interest in the input, replacing them with a baseline value and recalculating the model's prediction, to see how it affects the model output. This process is repeated by sliding the windows through the whole input and the change in the model's prediction is used to compute an attribution score for each region [31].

The three explainability methods were implemented for every output using *captum.ai* library [32]. In order to select which specific output we wanted the explanations for, a wrapped model for each class had to be defined. In the cases of Gradient Shap and Integrated Gradients, Gaussian noise was added with *Noise Tunnel* to each input in the batch, with *smoothgrad_sq* as smoothing type, *n_samples = 10*, and all zeros image as baseline. In the case of Occlusion, the sliding window was of *size = (3,15, 15)* and *stride = (3, 8, 8)*.

2.2 TRACE-VCI dataset

For this analysis, the dataset was obtained from a study on memory clinic patients. The TRACE-VCI dataset is a collection of data from 861 patients with possible VCI that was performed between 2009 and 2013 by three Dutch outpatient clinics at two university hospitals; the outpatient clinic of the VU University Medical Centre (VUMC) registered in the Amsterdam Dementia Cohort (N=665) [33] and the two outpatient memory clinics of the University Medical Centre Utrecht (UMCU) (N=196). The dataset includes every patient that showed cognitive complaints or signs of VCI on MRI scans. Each patient received a standardized extensive 1-day memory clinic evaluation including an interview, physical and neurological examination, laboratory testing, extensive neuropsychological testing, and an MRI scan of the brain [34]. The brain MR images obtained were processed to generate lesion maps using the RegLSM image-processing pipeline (publicly available at www.metavcimap.org) [9]. Then, all lesion maps were transformed into the T1 1-mm MNI-152 (Montreal Neurological Institute) brain template [35] using Elastix toolbox [36] to apply a linear registration followed by nonlinear registration. Then, the images were segmented using the k-nearest neighbour automatic classification with tissue type priors method, as described in [14].

The overall size of the dataset was 822 after excluding patients with dementia, divided into 60% for training, 20% for validation and 20% for testing. Due to GPU limitations, images we downsized to have shape: (109, 131, 109).

Previous studies addressing the weakness of the correlations within the combined lesion locations and scores suggest an initial study on a simulated dataset, where the lesion-symptom mapping can be properly validated [16].

2.2.1 Simulations

The model was trained and evaluated with simulations of lesion-score relations calculated from three cubic Regions of Interest (ROIs) associated with the TRACE-VCI lesion maps.

First, the images were cropped to fit the edges of the brain in the MNI space, obtaining a final lesion-map size of (822, 91, 107, 86). Then, the ROIs were selected with the following procedure to ensure that enough patients cover the area of interest, based on the simulation described in the paper by Zhang et. Al [16]:

- Create a lesion mask by summing the occurrence of lesions in every voxel.
- Define the three cubic ROI of size 8x8x8 mm^3.
- Erode the areas with a Structuring Element of the diameter of the ROI.
- Threshold the areas where at least ten patients have lesions.
- Select a centre and remove the ROI area from the possibilities for the next selection.
- Repeat the last step until getting three ROI centres.

Figure 2 shows two slices of the lesion-matrix, generated by the overlap of the lesioned voxels of all patients and the three ROIs, with the following centres:

- ROI 1: [37, 58, 54]
- ROI 2: [68, 57, 53]
- ROI 3: [35, 78, 44]

Being i the number of patients and *n* the number of ROIs, the scores were calculated with the following procedure: 1

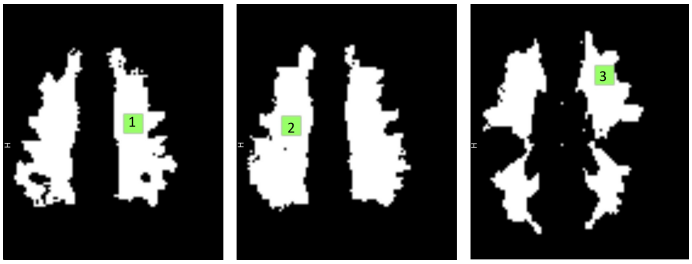

Fig. 2: *Choice of slices 54 (left), 53 (center) and 44 (right) of the coronal view of the lesion matrix (white) overlapping with three ROIs (green)*

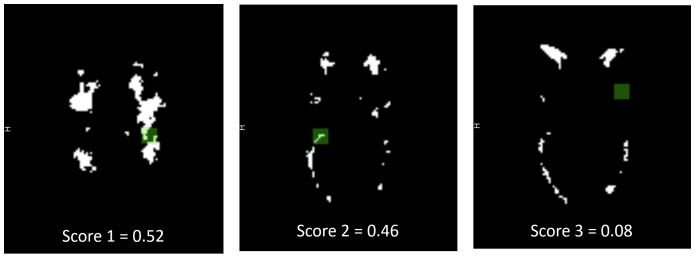

Fig. 4: *Choice of slices 54 (left), 53 (center) and 44 (right) of the coronal view of an example patient with the overlaps of the lesioned voxels with the three ROIs and annotations of the artificial scores*

Algorithm 1 Calculus of scores

for $patient$ **do**
  for ROI **do**
    $overlap_{i,n} \leftarrow LM_i \cap ROI_n$
    $score_{i,n} \leftarrow overlap_{i,n}/sum(ROI_n(:))$
  end for
end for

Obtaining as an output a matrix of size (number of patients, 3) with the score for each ROI for each patient. The computed scores have the following occurrence, displayed in Figure 3.

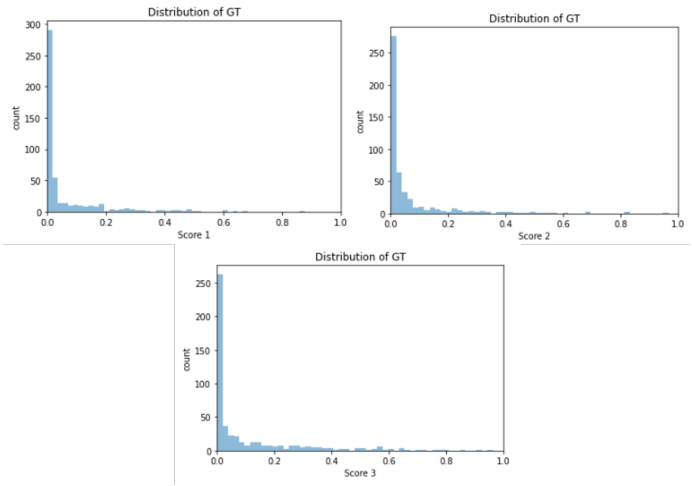

Fig. 3: *Distribution of scores for the simulations*

The overlap between the lesion map of an example patient and the three ROIs is displayed in Fig 4, with the annotations of the corresponding scores.

2.2.2 Model

After running experiments with different combinations of optimizers, learning rates, architectures and loss functions, the parameters to train the final model were chosen for the ones giving the best learning curve and validation metrics. After performing parameter tuning with *wandb.ai*, a 3D adaptation of ResNet10 developed by [37] was trained using Adam optimizer and MSE as loss function, using a batch size of 4. The last layers were modified from the flattening layer to a Linear layer of 512 parameters with ReLU activation, and a last Linear layer with an output of size 1 with Sigmoid activation to clip the output value between 0 and 1. The learning rate showed to work best with 10^{-5} for the first 100 epochs and descending to 10^{-6} after as a regularization technique to avoid noise. MSE was chosen as loss function because of the weight it gives to outliers so that even if most scores have very small values, the biggest ones are correctly predicted. One multi-output and three single-output models were trained on the different scores with this configuration. In the case of the multi-output model, the architecture branches into three equal last linear layers, so that different weights are associated with the different outputs.
The weight configuration was automatically chosen for the epoch with the lowest loss for the validation set. The trained model was then analyzed using MSE metric for the predictions and the ground truth. The square of the Pearson correlation coefficient was also computed.

2.2.3 Explainability

The features of trained models were then analyzed using Gradient Shap, Integrated Gradients and Occlusion techniques, with blank (all zeros) images as baselines. For the occlusion method, the sliding window was a cube of *size = 8* with a sliding window with *stride = 4*. The so called saliency maps were computed for an independent test set, and all the obtained maps were summed pixel-wise to obtain a total map. The only processing applied to the obtained images was filtering only the values indicating a positive contribution in the cases of Gradient Shap and Integrated Gradients, and linear normalization to values between [0,1].
The obtained probabilistic saliency maps were then analyzed by comparing them to a binary image of the corresponding ROI. Precision-Recall (P-R) curves were computed by bina-

rizing the saliency maps with a thousand different thresholds going from the maximum to the minimum image value. At each threshold, Precision and Recall metrics were calculated. The P-R curves were obtained by plotting the Precision against the Recall for all possible thresholds for each saliency map. P-R curve was chosen over the Receiver-Operating Characteristic analysis due to its sensibility for imbalanced data, given the prevalence of negative values over positive ones [38]. A continuous Dice Coefficient (cDC) was also calculated for each method as described in the paper by Shamir et al. [39], a metric that does not threshold the image and is based on the classic Dice Score but for probabilistic images. Precision, Recall and cDC equations are described in Equations 1, 2 and 3, where TP corresponds to True Positives, TN to True Negatives and FN to False Negatives

$$Precision = \frac{TP}{TP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + TN} \tag{2}$$

$$cDc = \frac{2\, sum(A(:) . * (:)B)}{c\, sum(A(:)) + sum(B(:))} \tag{3}$$

where .* denotes elementwise multiplication of matrices and c is described by 4

$$c = \frac{sum(A(:) . * B(:))}{sum(A(:) . * sign(B(:)))} \tag{4}$$

where

$$sign(x) = \begin{cases} -1 & \text{if x<0} \\ 0 & \text{if x=0} \\ 1 & \text{if x>0} \end{cases} \tag{5}$$

Last, the Area Under the Curve (AUC) was computed for every P-R curve.

3. Results

3.1 UTKFace dataset

An example of the Ground Truth and prediction is displayed in Figure 5, in which sex is correctly predicted for all models, but ethnicity is misclassified for the single-output model. The absolute error for age is higher in the multi-output model.

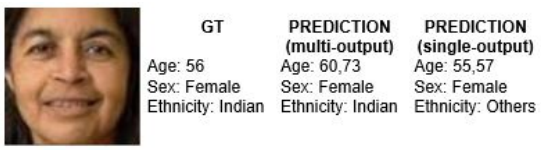

Fig. 5: *Example of a pre-processed image from UTKFace dataset with the corresponding labels for predicted and Ground Truth age, sex, and ethnicity.*

Table III shows the metrics calculated for an independent test set for the models trained with the characteristics described in section 2.1.1.

	Metrics	
	Single-output	Multi-output
Age (Mean squared error)	100.28	236.79
Sex (Accuracy)	78.05%	94.18%
Ethnicity (Balanced accuracy)	56.55%	87.32%

TABLE III: *Metrics for the single-output and multi-output models in terms of MSE, Accuracy, and Balanced Accuracy for UTKFace dataset*

Sex and ethnicity classification show better accuracy for the multi-output model, but age is better predicted for the single-output model.

Figure 6 displays the explanation maps computed for Image 5 for every model with Occlusion, Gradient Shap and Integrated Gradients techniques for each score.

3.2 TRACE-VCI dataset

The learning curves for training and validation sets for all models with the parameters of the models described in subsection 2.2.2 can be found in A.1, and figures in A.2 show the correlation between the prediction and the Ground Truth.

Table IV shows the Mean Squared Error of the predictions and the ground truth calculated for an independent test set. The multi-output model appears to outperform the single-output models in terms of predicting Score 1 and Score 2, while the single-output model has a slightly better performance for Score 3.

	Mean Squared Error	
	Single-output	Multi-output
Score 1	0.0079	0.0061
Score 2	0.0064	0.0034
Score 3	0.0050	0.0066

TABLE IV: *MSE metrics for the single-output and multi-output models for TRACE-VCI dataset predicted scores and ground truth*

Table V shows the square of the Pearson correlation coefficient and the p-value for the predicted scores and the ground truth.

	Squared Pearson correlation (p-value)	
	Single-output	Multi-output
Score 1	0.93 (<0.001)	0.94 (<0.001)
Score 2	0.84 (<0.001)	0.94 (<0.001)
Score 3	0.93 (<0.001)	0.97 (<0.001)

TABLE V: *Square of the Pearson correlation coefficient and p-value of the ground truth and predicted scores for the single-output and multi-output models for the TRACE-VCI dataset.*

Figure 7 shows the ROI origin slice of the explanation maps computed with the methods described in 2.2.3: Gradient Shap, Integrated Gradients and Occlusion for all the scores with the

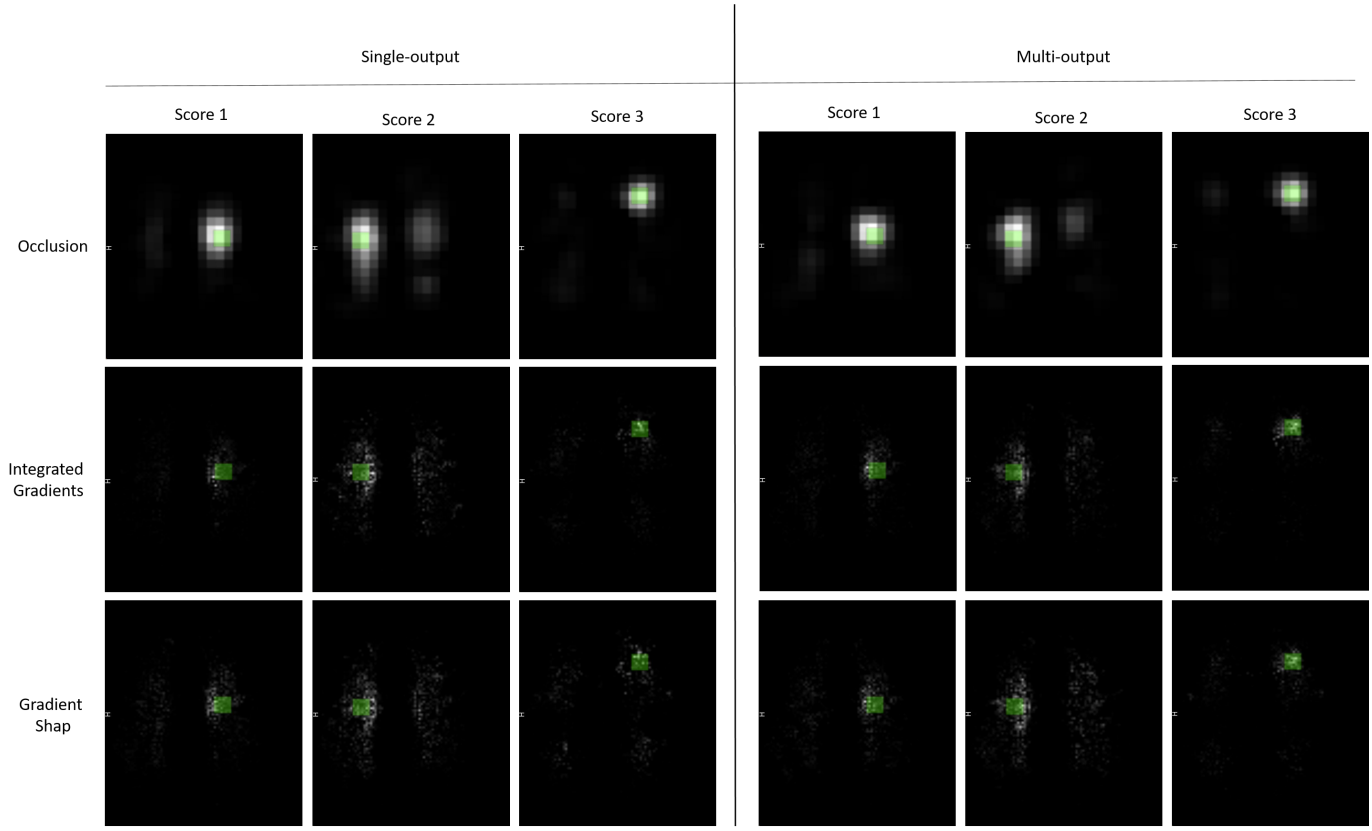

Fig. 7: *Choice of slices 54 (left), 53 (center) and 44 (right) of the coronal view of the overlap between the corresponding ROI and the explanation maps for the single-output and multi-output models*

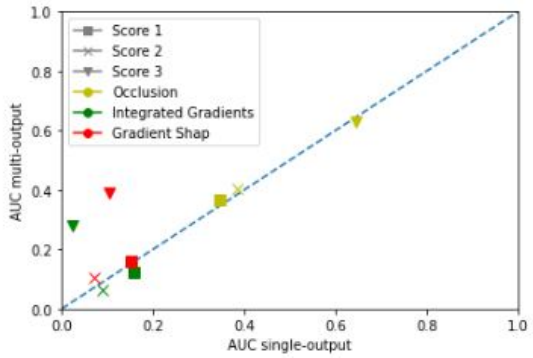

Fig. 8: *Correlation of the AUC of the P-R curves of the single and multi-output models*

an example, so it is not representative of the whole dataset, but several images have been visually analyzed and similar conclusions have been extracted. The features that indicate a person's age, sex, or ethnicity are not well defined and there is not a specific ground truth to be compared with, so from this experiment, we can only obtain qualitative arguments to decide on the methods that work best for this task. In general, this experiment is an example that demonstrates that multi-output models have the potential of performing similarly to multiple single-output models and that explainability methods can be computed for each output. However, the accuracy of the obtained saliency maps could not be evaluated.

As for the second part, a simulation study was performed on WMH segmentations from MR brain images of VCI affected patients by calculating artificial scores associated with specific ROIs. Three single-output and one multi-output 3D ResNet10 architectures were trained, and results from the computed metrics on an independent test set indicate that the single and multi-output models are performing well in predicting scores, as all MSE values are relatively low. However, as indicated in Figure 3, most scores have values close to 0, so low MSE does not necessarily mean good performance. The square Pearson correlation coefficient is independent of the dataset distribution, TableV indicates the following results for scores 1, 2 and 3: 0.94 0.94 and 0.97 for the multi-output model and 0.93 0.84 0.93 for the single-output models. All values are close to 1, which indicates a very high correlation between the predicted and the real scores ($p < 0.001$). However, although these values are in agreement with literature, they can not be directly compared as they use real cognitive scores or simulations that are more complexly dependent on lesion location. To further validate these results and analyze which factors are affecting the predictions, the saliency maps were computed and voxel-wise summed for all images in the test set. Visually, the images in Figure 7 suggest that there is an

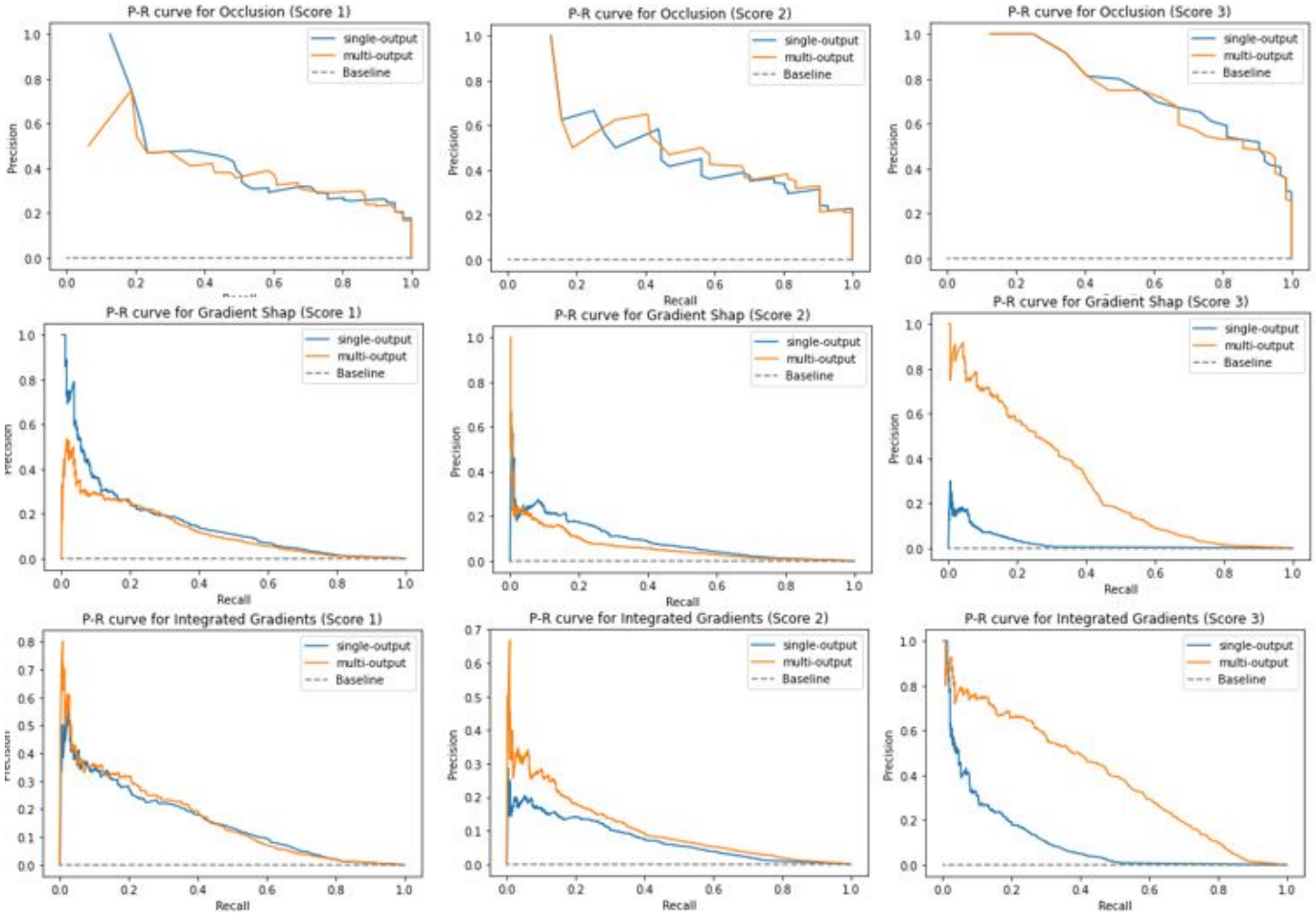

Fig. 9: *P-R curves for single-output, multi-output and baseline models for all scores and explainability methods*

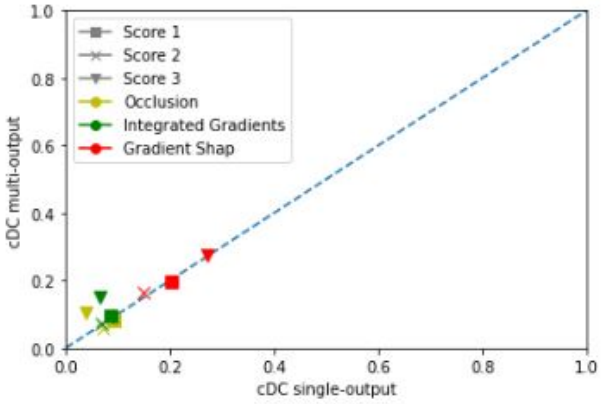

Fig. 10: *Correlation of the cDC of the obtained total explanation maps and the ROIs of the single and multi-output models*

accurate overlap between the model's choices of important regions and the actual ROIs, showing a great ability to spot if a lesion location is directly related to an outcome. It is important to mention the ability of the models to distinguish the locations that affect Scores 1 and 2, as the ROIs are in parallel areas of the brain, for which the lesions are often symmetrical (as it is for example in Fig 4), and the scores are to an extent correlated. From the total maps, we can see that Gradient Shap and Integrated Gradients give more noisy images, as these methods are sensitive to small variations in the input image. While occlusion output is clearer, the computational time was considerably longer.

From the quantitative analysis performed to the explanation maps by comparison to binary images of the ROIs, we can see that P-R curves in Figure 9 show a performance that apparently does not correspond to the overlap observed visually. However, the prevalence (which refers to the ratio of voxels inside the ROI and outside) directly affects the performance of what a baseline classifier would look like. In this case prevalence = **0.0018** is very low, and the baseline performance is indicated by a dashed line in 9, which has an **AUC = 0.0009**, lower than any of the AUC obtained for the P-R curves as observed in Table VI. The performance comparison of the P-R curves is better analyzed by looking at the correlation Figure 8, which shows that all models have similar predictive abilities, except for score 3, in which the multi-output model shows to outperform the single-output one when using Integrated

Gradients and Gradient Shap. The computed cDC presented in Table VII shows values that correspond to what is visually observed, taking into account that we are comparing 3D volumes. The correlation graph displayed in Figure 10 shows consistency between all models, despite a slightly better metric for score 3 for the multi-output model in the cases of Occlusion and IG. Visually analyzing these cases in 3D, the explanation maps show highlights in central slices of the images for the single-output model, whereas, in the multi-output one, explanations are better focused on the ROI. This could indicate that explainability methods have a better ability to identify key regions when applied in multi-output models.

The predicted saliency maps are highly dependent on the chosen parameters since it was observed that changing, for example, baseline image and sliding window size affects considerably the look of the computed image. It was observed that the difference in the calculated metrics for the saliency maps in single and multi-output models is not concluding. This indicates that for this use case, a visual analysis gives enough information to indicate whether the explored methods are suitable to detect ROIs. If a more complex generation of artificial scores were to be implemented and therefore, the performance of single and multiple output models were to be significantly different, then quantitative analyses on the saliency maps would be a more concluding approach for deciding which model is performing better.

Previous studies have only explored simpler models based on statistical and ML analysis to tackle the challenges of LSM, but DL and explainable AI has the potential to yield further discoveries in neuroscience. Overall, this research has demonstrated satisfactorily that DL algorithms and explainability techniques are suitable for the task of identifying key regions for multiple score predictions in 3D lesion MR images of patients affected by VCI. Multi-output models have been shown to have a similar performance to single-output ones in this simulation study, which indicates that, in combination with the explored explainability techniques, they could be useful in the research of LSM. However, more research needs to be done before taking clinical conclusions from real cognitive scores. A more complex simulation study could be performed to test the suitability of multi-output models, in which the ROIs affect the calculus of multiple scores with a different associated weight to resemble more the hypothesis of how brain lesions and cognitive sores are related. In order to apply these methods to real data, the poor correlation between clinical findings and WMH locations discussed in previous studies [40] should also be considered, suggesting the need to develop a very sensitive model, or use images of a different type of lesion segmentation map that is more directly related to a neurological outcome, such as infarcts. Future work should take all the commented limitations into account, especially the need for a robust enough server that can handle the big amount of 3D data and the complex architectures needed to train the CNN, as it was a very limiting issue throughout the whole research process. Lastly, saliency maps can provide valuable insights into the model's behaviour, but they can also be misleading if they are not properly validated.

5. Conclusion

This project presents a 3D multiple-output Deep Learning based pipeline capable of predicting scores associated with lesions in specific regions of the brain from MRI images, and that identifies which regions they are associated to. The first part of the project tests the desired methods on a simpler dataset and proves that multi-output models are suitable for the task of predicting multiple outcomes and that explainability methods can be applied to these models.

The second part proposes a model that has been demonstrated to work satisfactorily for predicting simulated scores and correctly identifying the locations of interest to make such decisions. This method, if used for real cognitive scores of after-stoke patients, has the potential of finding correlations between the location of the lesions in the brain and the different neurological outcomes that VCI can cause, which can help doctors get a better understanding of the underlying brain mechanisms that lead to neurological dysfunction.

This is only a first step towards the use of Deep Learning for Lesion-Symptom Mapping, as many limitations still need to be addressed.

Bibliography

[1] Stéphanie Debette et al. "Clinical significance of magnetic resonance imaging markers of vascular brain injury: a systematic review and meta-analysis". In: *JAMA neurology* 76.1 (2019), pp. 81–94.

[2] Joanna M Wardlaw, Colin Smith, and Martin Dichgans. "Small vessel disease: mechanisms and clinical implications". In: *The Lancet Neurology* 18.7 (2019), pp. 684–696.

[3] Philip B Gorelick et al. "Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association". In: *stroke* 42.9 (2011), pp. 2672–2713.

[4] Nick A Weaver et al. "Strategic infarct locations for post-stroke cognitive impairment: a pooled analysis of individual patient data from 12 acute ischaemic stroke cohorts". In: *The Lancet Neurology* 20.6 (2021), pp. 448–459.

[5] John T O'Brien and Alan Thomas. "Vascular dementia". In: *The Lancet* 386.10004 (2015), pp. 1698–1706.

[6] Perminder S Sachdev et al. "Classifying neurocognitive disorders: the DSM-5 approach". In: *Nature Reviews Neurology* 10.11 (2014), pp. 634–642.

[7] J Matthijs Biesbroek, Nick A Weaver, and Geert Jan Biessels. "Lesion location and cognitive impact of cerebral small vessel disease". In: *Clinical Science* 131.8 (2017), pp. 715–728.

[8] J Matthijs Biesbroek et al. “High white matter hyperintensity burden in strategic white matter tracts relates to worse global cognitive performance in community-dwelling individuals”. In: *Journal of the Neurological Sciences* 414 (2020), p. 116835.

[9] Nick A Weaver et al. “The Meta VCI Map consortium for meta-analyses on strategic lesion locations for vascular cognitive impairment using lesion-symptom mapping: Design and multicenter pilot study”. In: *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 11.1 (2019), pp. 310–326.

[10] Hanna Jokinen et al. “Longitudinal cognitive decline in subcortical ischemic vascular disease–the LADIS Study”. In: *Cerebrovascular diseases* 27.4 (2009), pp. 384–391.

[11] Fanny Munsch et al. “Stroke location is an independent predictor of cognitive outcome”. In: *Stroke* 47.1 (2016), pp. 66–73.

[12] Elizabeth Bates et al. “Voxel-based lesion–symptom mapping”. In: *Nature neuroscience* 6.5 (2003), pp. 448–450.

[13] Lei Zhao et al. “Strategic infarct location for post-stroke cognitive impairment: A multivariate lesion-symptom mapping study”. In: *Journal of Cerebral Blood Flow & Metabolism* 38.8 (2018), pp. 1299–1311.

[14] Nick A Weaver et al. “Cerebral amyloid burden is associated with white matter hyperintensity location in specific posterior white matter regions”. In: *Neurobiology of aging* 84 (2019), pp. 225–234.

[15] Maurizio Corbetta et al. “Common behavioral clusters and subcortical anatomy in stroke”. In: *Neuron* 85.5 (2015), pp. 927–941.

[16] Yongsheng Zhang et al. “Multivariate lesion-symptom mapping using support vector regression”. In: *Human brain mapping* 35.12 (2014), pp. 5861–5876.

[17] David V Smith et al. “Decoding the anatomical network of spatial attention”. In: *Proceedings of the National Academy of Sciences* 110.4 (2013), pp. 1518–1523.

[18] MH Thomas. “Hope Mohamed L Seghier Alex P Leff and Cathy J Price. 2013. Predicting outcome and recovery after stroke with lesions extracted from MRI images”. In: *NeuroImage: clinical* 2 (2013), pp. 424–433.

[19] Dinggang Shen, Guorong Wu, and Heung-Il Suk. “Deep learning in medical image analysis”. In: *Annual review of biomedical engineering* 19 (2017), p. 221.

[20] Sucheta Chauhan et al. “A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images”. In: *Frontiers in neuroinformatics* 13 (2019), p. 53.

[21] Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. “Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications”. In: *Neuroscience & Biobehavioral Reviews* 74 (2017), pp. 58–75.

[22] Mauricio Reyes et al. “On the interpretability of artificial intelligence in radiology: challenges and opportunities”. In: *Radiology: artificial intelligence* 2.3 (2020), e190043.

[23] Bas HM van der Velden et al. “Volumetric breast density estimation on MRI using explainable deep learning regression”. In: *Scientific Reports* 10.1 (2020), p. 18095.

[24] IEEE. *UTKFace*. https://susanqq.github.io/UTKFace/. 2017.

[25] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[26] Modupe Odusami et al. “Analysis of features of alzheimer’s disease: detection of early stage from functional brain changes in magnetic resonance images using a finetuned ResNet18 network”. In: *Diagnostics* 11.6 (2021), p. 1071.

[27] L. Biewald. *Experiment tracking with weights and biases*. https://www.wandb.com/. 2020.

[28] Bas HM Van der Velden et al. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Medical Image Analysis* (2022), p. 102470.

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[30] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[31] Daniel T Huff, Amy J Weisman, and Robert Jeraj. “Interpretation and visualization techniques for deep learning models in medical imaging”. In: *Physics in Medicine & Biology* 66.4 (2021), 04TR01.

[32] Narine Kokhlikyan et al. “Captum: A unified and generic model interpretability library for pytorch”. In: *arXiv preprint arXiv:2009.07896* (2020).

[33] Wiesje M van der Flier et al. “Optimizing patient care and research: the Amsterdam Dementia Cohort”. In: *Journal of Alzheimer’s disease* 41.1 (2014), pp. 313–327.

[34] Jooske Marije Funke Boomsma et al. “Vascular cognitive impairment in a memory clinic population: rationale and design of the “Utrecht-Amsterdam clinical features and prognosis in vascular cognitive impairment”(TRACE-VCI) study”. In: *JMIR Research Protocols* 6.4 (2017), e6864.

[35] Vladimir Fonov et al. “Unbiased average age-appropriate atlases for pediatric studies”. In: *Neuroimage* 54.1 (2011), pp. 313–327.

[36] Stefan Klein et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pp. 196–205.

[37] Tencent. *MedicalNet*. https://github.com/xmuyzz/3D-CNN-PyTorch/blob/

8728325fde8c1980ed300ba8b2cbab0020a302c2 / models/ResNetV2.py. 2019.

[38] Kay Henning Brodersen et al. “The binormal assumption on precision-recall curves”. In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 4263–4266.

[39] Reuben R Shamir et al. “Continuous dice coefficient: a method for evaluating probabilistic segmentations”. In: *arXiv preprint arXiv:1906.11031* (2019).

[40] Elisabeth CW Van Straaten et al. “Impact of white matter hyperintensities scoring method on correlations with clinical data: the LADIS study”. In: *Stroke* 37.3 (2006), pp. 836–840.

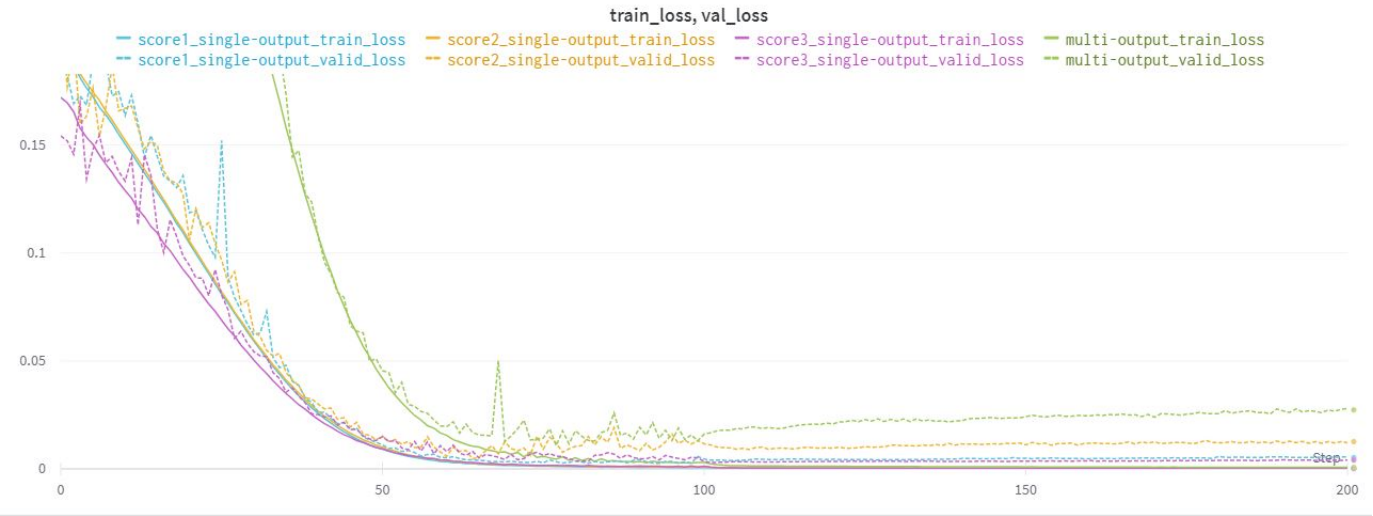

Fig. A.1: *Loss function for the train and validation set for the single and multi-output models for each epoch*

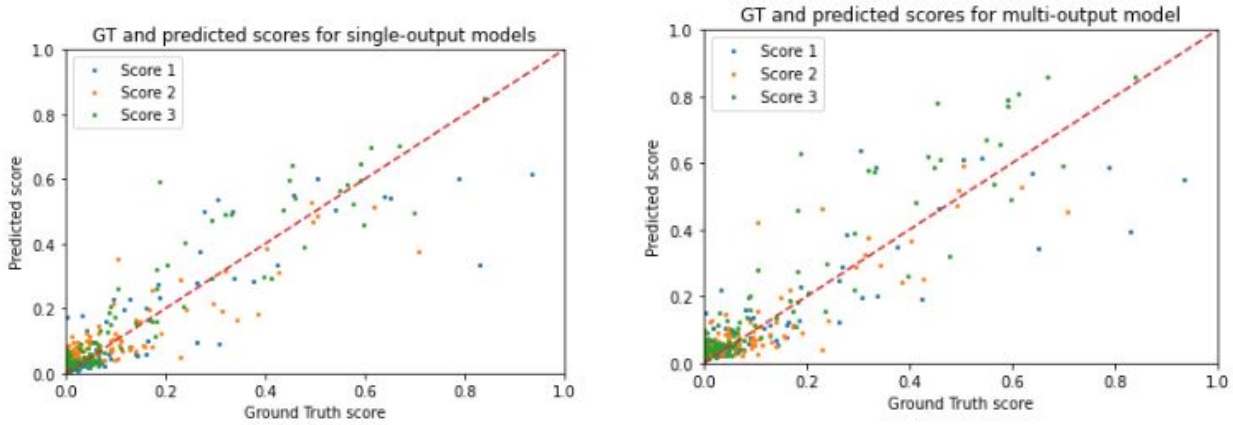

Fig. A.2: *Prediction VS Ground truth for all the scores in the single and multi-output models*

		Area Uunder the Curve		
		Gradient SHAP	Integrated Gadients	Occlusion
Score 1	Single-output	0.1568	0.1526	0.3439
	Multi-output	0.1241	0.1597	0.3651
Score 2	Single-output	0.089	0.0709	0.3854
	Multi-output	0.0620	0.1043	0.4028
Score 3	Single-output	0.0236	0.1042	0.6457
	Multi-output	0.2811	0.3893	0.5835

TABLE VI: *AUC for the Precision-Recall curve of the saliency maps for single and multi-output models for all scores using GS, IG and Occlusion*

		continuous Dice Coefficient		
		Gradient SHAP	Integrated Gradients	Occlusion
Score 1	Single-output	0.0938	0.0857	0.2039
	Multi-output	0.0798	0.0944	0.1984
Score 2	Single-output	0.0712	0.0682	0.1481
	Multi-output	0.0602	0.0727	0.1656
Score 3	Single-output	0.0384	0.0667	0.2724
	Multi-output	0.1059	0.1499	0.2772

TABLE VII: *cDC of the saliency maps and ROIs for single and multi-output models for all scores using GS, IG and Occlusion*