AI and Ethics (2024) 4:151–156
https://doi.org/10.1007/s43681-023-00399-x

ORIGINAL RESEARCH

Using ScrutinAI for visual inspection of DNN performance in a medical use case

Rebekka Görge[1] · **Elena Haedecke**[1,2] · **Michael Mock**[1]

Published online: 20 December 2023

Abstract
Our Visual Analytics (VA) tool ScrutinAI supports human analysts to investigate interactively model performance and data sets. Model performance depends on labeling quality to a large extent. In particular in medical settings, generation of high quality labels requires in depth expert knowledge and is very costly. Often, data sets are labeled by collecting opinions of groups of experts. We use our VA tool to analyze the influence of label variations between different experts on the model performance. ScrutinAI facilitates to perform a root cause analysis that distinguishes weaknesses of deep neural network (DNN) models caused by varying or missing labeling quality from true weaknesses. We scrutinize the overall detection of intracranial hemorrhages and the more subtle differentiation between subtypes in a publicly available data set.

Keywords Visual analytics · Medical image analysis · Trustworthy AI · Human-centered machine learning · Machine learning · Artifical intelligence

1 Introduction

Machine Learning (ML) models, especially based on Deep Learning (DL), promise a high potential for medical applications. To use machine based decisions in medical practice, models must be reliable and transparent. Additionally to local explanations of single model decisions, a detailed understanding of the deep neural network (DNN) is needed to uncover hidden patterns within the inner structure of the algorithm [11]. The performance and reliability of a supervised model is influenced by aleatoric uncertainty resulting, e.g., from label noise. This concerns especially the medical domain [7]: Labeling medical data sets, such as image data, requires resource intensive domain knowledge which is biased due to subjective expert judgement, annotation habits as well as annotator's errors. This results in a lack of consistency among different observers, which is defined as inter-observer variability [8]. Final annotations are often a consensus from multiple experts, which challenges appropriate aggregation of these labels. As a consequence of the negative affection of label noise on model performance, the trustworthiness of quality metrics based on these labels is questionable. As handling label noise is still largely unnoticed in medical domain [7]. we use Visual Analytics (VA) to analyze the aleatoric uncertainty of labels and base the detection of true weaknesses on this.

Visual Analytics is a multidisciplinary field where interactive systems and tools are being developed that enable the human analyst to engage in a structured reasoning process by providing appropriate visualizations and representations of the data. An important aspect is to explicitly benefit from the humans' tacit and expert knowledge by supporting the analyst with specific workflows to gain insights and knowledge into the problem domain [10]. Integrating the human into the analysis process is especially important in case of complex DNN models [1]. To further facilitate DNN interpretability, transparency methods, e.g. Grad-CAM++ [2], are often employed. For this task, the data and its representations must be prepared in such a way that no relevant information is lost or hidden, and at the same time the workflow and integrated widgets provide enough flexibility for deep dive analysis. For example, CheXplain [12] focuses at

✉ Rebekka Görge
rebekka.goerge@iais.fraunhofer.de

Elena Haedecke
elena.haedecke@iais.fraunhofer.de

Michael Mock
michael.mock@iais.fraunhofer.de

[1] Fraunhofer IAIS, 53757 St. Augustin, Germany

[2] University of Bonn, 53113 Bonn, Germany

addressing the specific needs in the healthcare domain for exploring and understanding a detection model for radiographic chest images.

Our own focus lies in assessing potential model vulnerabilities by leveraging the semantic context of the data, specifically object-level description of entities, through the use of detailed descriptive meta data. For this task, we developed the VA tool ScrutinAI [6], whose functionalities and workflow promote the generation of semantic hypotheses during the exploration in iterative **analysis cycles**. The meta data, as well as precomputed model predictions of the model, are loaded into the tool via an easily exchangeable CSV file. Additionally, image data are displayed with options for zooming and overlaying. Various widgets provide options for filtering and querying the data, such as textual queries, an interactive selection of interesting data points, data distribution along categories, or correlation plots. Although being originally developed for uses cases in the automotive domain, the modular design of ScrutinAI allows a simple adaption to other domains by the integration of customized widgets.

2 Visual analysis of the medical use case

In this paper, we demonstrate the visual analysis of the influence of label noise on model performance within ScrutinAI for the use case of DNN detection of intracranial hemorrhages. We give details on the medical use case, model and data sets and the modularity and extensibility of ScrutinAI. We describe our analysis process within ScrutinAI and our findings, in which we uncover inter-observer variability in the general detection of hemorrhages and among all classes in the data sets, in particular with respect to very similar classes. In consecutive analysis cycles, we reveal a negative influence of this label noise on our model's performance.

2.1 Background of the medical use case

Intracranial hemorrhage is an urgent and life-threatening emergency requiring rapid medical treatment. To determine region and size of a hemorrhage, imaging techniques such as computed tomography (CT) can be used, consisting of individual slices giving a three-dimensional impression of the head. Automatic image recognition using DL can assist doctors with quick detection and characterization. Therefore, we use a DNN of our prior work [5] trained on the biggest publicly available multinational and multi-institutional data set of intracranial CT scans provided by the Radiology Society of North America (RSNA) [4]. For each sample, one of 60 experienced radiologists annotated on slice-level the region of the hemorrhage with the corresponding subtype *any*, *epidural*, *intraparenchymal*, *interaventricular*, *subarachnoid* or *subdural*. We use the labeled data set part with 80% train and 20% test split and window setting[1] as preprocessing method. If the model predicts a hemorrhage in a region, we generate with Grad-CAM++ a heatmap as local explanation.

We evaluate our model in addition to the RSNA test split on the commonly used public CQ500 data set [3]. After data selection and cleaning, we obtain 490 CT scans, using only those slice series with the lowest sampling rate. The data set is annotated on CT-level by three independent senior radiologists with eight, twelve, and 20 years of experience, using the same six subtypes of hemorrhages as in the RSNA data set. We use the individual annotations as well as the original ground truth derived from the majority vote of the three radiologists.

In addition to the CT-level annotations, we use labels and bounding boxes on slice-level of [9]. The labels are an aggregation of annotations from three different neuroradiologists with six, four, and less than one year of practice. The single annotations of the radiologists have not been published. The slice-level labels by [9] are uniquely matched to the CT-level annotations by the *SOP-Instance UID* and the *Study-Instance UID*. In the following, we treat the annotations of [3] as radiologists number one, two and three, and of [9] as fourth radiologist. To make the annotations of [9] as well as the model's output comparable to the others, we generalize them to a CT-level by using the maximum value per region and CT. Besides generating more knowledge by combining all various annotations, we enrich the meta data, e.g., by specifying for each CT the proportion of radiologists detecting a specific hemorrhage.

2.2 Analysis in ScrutinAI and results

Since labeling noise is a challenging and well-known problem in medical image analysis, we investigate the application of our model to the CQ500 data set in this regard. Using ScrutinAI, we aim to expose dependencies and relationships between model performance and inter-observer variability in labeling. Therefore, we load the CQ500 data set, the annotations, and the precomputed model's predictions in ScrutinAI. As visible in Fig. 1a, all structured data is accessible over the meta data overview Ⓑ. The unstructured image data can be displayed in use case specific views Ⓗ, enabling an analysis in different window settings or an inspection of annotated bounding boxes and local explanations generated with GradCAM++.

[1] Typically used by radiologists. Corresponds to gray-value mapping, where a specific interval of the CT range is selected to highlight different intensity ranges.

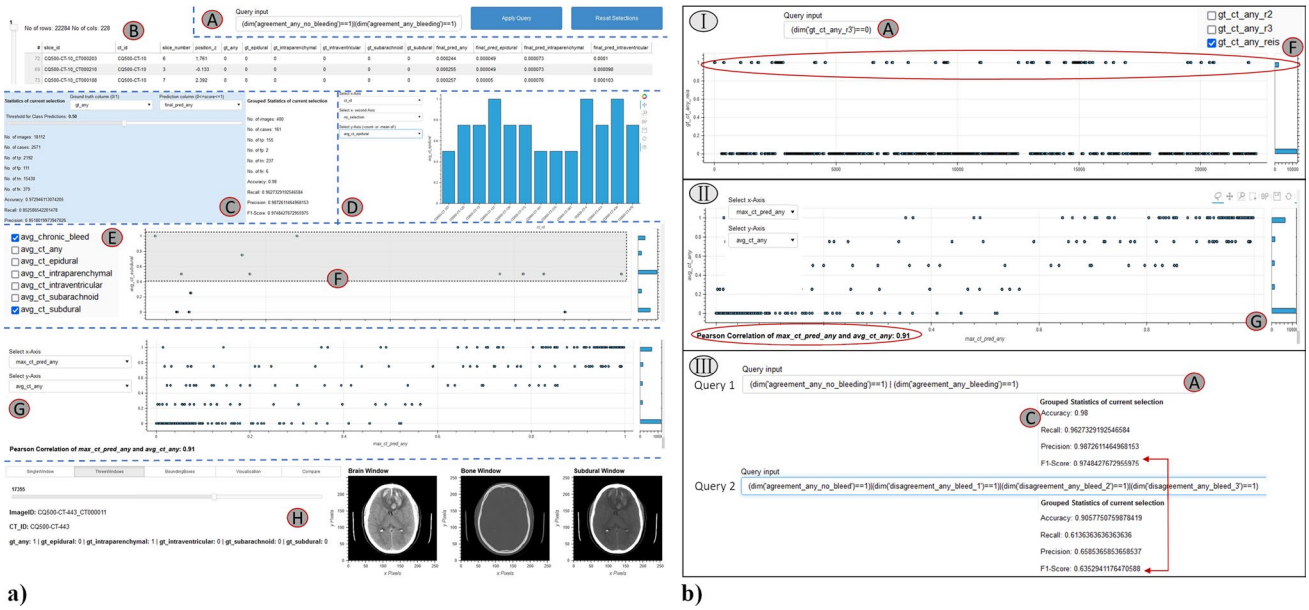

Fig. 1 (a) Overview of the interactive functionalities in ScrutinAI applied to the medical use case. (b) Detailed screenshots of the findings in analysis cycles I-III. (The individual widgets are shown in a compressed form to give an overall view)

Following, we present the walk-throughs of four consecutive **analysis cycles** (I-IV) within ScrutinAI, where we gained six *(i-iv)* findings.

In **cycle I**, we want to compare the model's performance against the opinion of all four radiologists as ground truth. For this task, we iteratively use the metric overview widget (as illustrated in Fig. 1a Ⓒ) and interactively change which annotation column to use as ground truth. All other dependent performance metrics are re-computed automatically. Based on this, we find that among radiologists, a varying number of hemorrhages and corresponding subtypes is detected. The model shows the best performance compared to the annotations of the fourth radiologist (Acc: 92,4%, F1-Score: 91%, occasions class any: 217), and the worst performance in comparison to the third radiologist (Acc: 88% and F1-Score 87,4%, occasions class any: 193). This can also be explored visually in ScrutinAI using the widgets textual query Ⓐ and scatter plot Ⓔ. We select with the query all hemorrhages labeled as negative by one radiologist, and create three plots for the annotations of the other radiologists to visually compare patterns. Exemplary, one of these plots in Fig. 1b - cycle I shows that several annotations of the fourth radiologist differ to the third expert and are labeled as true (value 1). We gain the first two findings: *(i)* the test data set shows a high inter-observer variability with respect to the general occurrence of hemorrhage as well as among different classes, and *(ii)* the model's performance varies greatly depending on which radiologist it is compared to.

We deepen the analysis in **cycle II** and calculate for each CT how many radiologists agreed on the presence of a hemorrhage. In the drop-down menu of the Pearson correlation plot Ⓖ, we choose the agreement of the radiologists on the incidence of a hemorrhage and the prediction of the model as independent categories and as result we get a positive correlation of 0.91 (see Fig. 1b - cycle II). The plot visualizes, that in cases where radiologists agree on the occurrence of a hemorrhage, the prediction score of the model is higher, whereas the less radiologists agree on a hemorrhage, the lower the prediction score of the model. We conclude as third finding *(iii)* a correlation between the model's performance and the inter-observer variability.

Using again the textual query, we further explore this observation in **cycle III**. Based on the data selection, ScrutinAI provides the number of cases detected by one, two, three or four radiologists, respectively, and let us easily compare how many of those potential cases are detected by our model. The results (see Table 1) support finding *(iii)*. Most cases are detected by all four radiologists and the detection rate of the model for those cases is the highest. Detection overlap between model and radiologists decreases, if only

Table 1 Number of radiologists agreeing on a hemorrhage and model prediction for these potential cases

		Four	Three	Two	One
Cases		161	37	28	25
Model	True	155	25	11	3
	False	6	12	17	22
Detection overlap		97%	68%	61%	44%

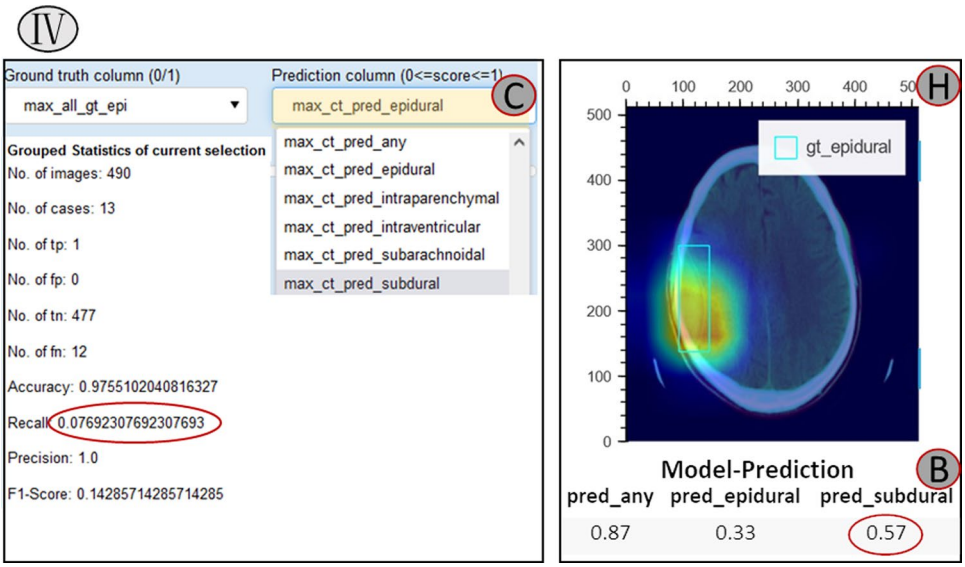

Fig. 2 The findings of analysis cycle IV show in the performance overview and image view that the model incorrectly classifies epidural hemorrhages as subdural

three or two radiologists agree on the occurrence. For those cases in which only one radiologist detects a hemorrhage, the model classifies a hemorrhage even less frequently.

To compare the cases in which the radiologists agree with those in which they disagree, we query in ScrutinAI two data subsets (see Fig. 1b - cycle III). The first subset of data is selected by textual query 1, consisting of all cases on which all radiologists either agreed on "no hemorrhage" or "hemorrhage". The second subset is filtered analogously (query 2), containing the data points where all radiologists agree on "no hemorrhage", plus, all cases for which only some of the radiologists (one, two or three) have detected a hemorrhage. We read off the model performance for both subsets by selecting the ground truth of the original annotations. Using the same procedure, we split the data according to the concordance of all radiologists within each class. We reveal as finding *(iv)*, that accuracy and F1-score decrease in all classes for cases in which the radiologists do not agree on the label, while both metrics increase for all cases in which all radiologists agree on the label.

As the performance of the model for epidural hemorrhages is extremely low, we scrutinize those cases in **cycle IV**. By examining the meta data overview, we observe an increased prediction score in the subdural class for many cases that have been annotated as epidural. As visible in Fig. 2, we explore in the image view Ⓗ an example case in which the model classifies a hemorrhage as subdural (see increased prediction score for subdural), while the ground truth of the slice and the bounding box are epidural. Still, the local explanation with Grad-CAM++ shows that the model has detected the region within the bounding box but it has classified it as wrong subtype. As we are interested if the model systematically detects epidural hemorrhages as subdural, we select in the metric overview as visible in Fig. 2Ⓒ as ground truth column "epidural". Accordingly, we compare it firstly to the prediction column of epidural. As the grouped statistics show, there are 13 cases out of the 490 CT-Scans with an epidural hemorrhage, but only one case is detected by the model (true positive), while 12 epidural cases are not detected as epidural (false positive). In a second step, we select in the drop down menu as prediction column "subdural", still comparing it to the epidural ground truth. We observe that now 11 out of the 13 epidural cases are detected and therefore classified as subdural. The recall increases to 85%. This result leads to finding *(v)*, that the model did not learn to distinguish between the classes subdural and epidural. A major reason for the bad performance might be due to the fact that the epidural cases were undersampled in the RSNA training data set. Still, we want to assess the radiologists' agreement on the label epidural, as distinguishing between both (spatially very close) regions requires a lot of expert domain knowledge. Filtering the epidural cases once more in ScrutinAI, we find that in 6 cases all radiologists, in 13 cases three, in 4 cases two, and in 9 cases only one radiologist agreed on an epidural hemorrhage. We select with a textual query all cases labeled in the original ground truth as epidural but not as subdural aiming to exclude images with hemorrhages in both regions. We obtain only 7 cases. We visualize the radiologists' assessment for subdural in a scatter plot and detect that in 4 out of the 7 cases at least one radiologist still labeled the case as subdural. Even if the number of cases is not representative, it indicates as finding *(vi)* that a clear distinction between similar classes, as subdural and epidural, is even non-trivial for experienced radiologists and similarly leads to higher inter-observer variation affecting model performance negatively.

3 Conclusion

We have shown that ScrutinAI can be easily adapted to a new domain such as healthcare. New data sets and models can be loaded and analyzed with the existing structure and features. Based on the modular structure, the tool can be easily extended by use case specific features. With the exemplary analysis of DNN detection of intracranial hemorrhages, we demonstrate that ScrutinAI can be used to interactively explore the dependencies between model performance and label noise in data sets. ScrutinAI's workflow and interactive functionalities were shown to efficiently support the analyst by means of VA principles. Using linked brushing, data could be easily filtered across the different widgets and data representations (structural and visual), enabling deep-dive analysis without the need to deal with individual scripts or tools to get the same functionality.

In summary, the analysis of the use case in ScrutinAI reveals the aleatoric uncertainty in the CQ500 data set, which interrelates with the inter-observer variability. Based on finding *(iii)*, we assume that in the RSNA data set, used for the training of our model, a similar label noise consists. We face the challenge of discerning whether the label noise arises from hard to detect hemorrhages being detected only by individual experts or from radiologists misclassifying artifacts as hemorrhages. Moreover, analysis cycles III and IV revealed the difficulty for a clear and consistent distinction between specific and, in particular, similar regions among observers. The negative effect of label noise on model performance detected in finding *(iv)* and *(v)*, confirms that learning patterns for the model is more difficult due to inter-observer variability. The question raises, whether a model trained either on the annotations of only one expert, or on individual annotations of several consistent experts, would more easily learn a stable behavior and perform better on hard to detect samples. To answer this question, we would need to compare our model to a model trained only on annotations of a single radiologist, which we leave open for future work. For a deeper analysis of the actual correspondence of the detected hemorrhages, we plan to compare the location of the bounding boxes of [9] to the rough location regions annotated in [3] as well as to a location's approximation of the occurrences detected by the model and extracted from the heatmap generated through Grad-CAM++.

Acknowledgements The development of this publication was supported by the Ministry of Economic Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia as part of the flagship project ZERTIFIZIERTE KI. The authors would like to thank the consortium for the successful cooperation.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The CQ500 dataset analyzed as part of the study is available at http://headctstudy.qure.ai/dataset. Additional annotations are used from https://physionet.org/content/bhx-brain-boundingbox/1.1/. The RSNA Intracranial Hemorrhage Detection dataset used for training the model is accessible at https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/overview/hemorrhage-types.

References

1. Andrienko, N., Andrienko, G., Adilova, L., Wrobel, S.: Visual analytics for human-centered machine learning. IEEE Comput. Graph. Appl. **42**(1), 123–133 (2022)
2. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V. N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV), 839–847. IEEE (2018)
3. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V., Rao, P., Warier, P.: Development and validation of deep learning algorithms for detection of critical findings in head CT scans. arXiv preprint arXiv:1803.05854 (2018)
4. Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artif. Intell. **2**(3), e190211 (2020)
5. Görge, R.: Verwendung von Deep Learning zur Erkennung von hämorrhagischen Schlaganfällen in CT-Scans des Schädels. Master's thesis, University Koblenz-Landau (2021)
6. Haedecke, E., Mock, M., Akila, M.: ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions. In Bernard, J.; and Angelini, M., eds., EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association. ISBN 978-3-03868-183-0 (2022)
7. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)
8. Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P.: On modelling label uncertainty in deep neural networks: automatic estimation of intra-observer variability in 2d echocardiography quality assessment. IEEE Trans. Med. Imaging **39**(6), 1868–1883 (2019)
9. Reis, E. P., Nascimento, F., Aranha, M., Secol, F. M., Machado, B., Felix, M., Stein, A., Amaro, E.: Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images. https://physionet.org/content/bhx-brain-bounding-box/1.1/. Accessed: 2023-01-13 (2020)
10. Sacha, D., Stoffel, A., Stoffel, F., Kwon, B.C., Ellis, G.P., Keim, D.: Knowledge generation model for visual analytics. IEEE Trans. Visual Comput. Graph. **20**, 1604–1613 (2014)

11. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Trans. Neural Netw. Learn. Syst. **32**(11), 4793–4813 (2020)
12. Xie, Y., Chen, M., Kao, D., Gao, G., Chen, X.: CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13 (2020)