WARSAW UNIVERSITY OF TECHNOLOGY

Ph.D. Thesis

Discipline of Science:
Information and Communication Technology

Field of Science:
Engineering and Technology

M. Sc.
Paulina Anna Tomaszewska

Role of spatial and temporal context in classification using deep neural networks

Supervisor
Prof. **Przemysław Biecek**, Ph.D., D.Sc.

Warsaw 2026

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor Prof. Przemysław Biecek. Thank you for your guidance and scientific support.

I gratefully acknowledge the financial support from the following projects: 'HOMER: Human-Oriented autoMated machinE leaRning', funded by the National Science Centre under the SONATA BIS-9 programme (grant no. 2019/34/E/ST6/00052); 'xLungs: Responsible Artificial Intelligence for Lung Diseases', funded by the National Centre for Research and Development (grant no. INFOSTRATEG-I/0022/2021-00); and YOUNG PW, funded by Warsaw University of Technology within the Excellence Initiative – Research University (IDUB) programme (grant no. 1820/97/Z01/2023). The work was carried out with the support of the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science, Warsaw University of Technology.

I am grateful for the exceptional opportunities to do research internships at Institute of Science and Technology Austria (Machine Learning and Computer Vision Lab) and École polytechnique fédérale de Lausanne (Visual Intelligence and Learning Lab), as well as to participate in enriching research visits at the University of Oxford (Reasoning with Machines AI Lab) and Carnegie Mellon University (Auton Lab). Each of these experiences has broadened my perspective, challenged my thinking, and contributed significantly to my growth in the field.

I am also grateful to my research collaborators from my alma mater (Dr Elżbieta Sienkiewicz among others) and other research institutes in Poland and abroad. I would like to mention here Professor of pathology Mai P. Hoang from Harvard Medical School and Professor of ophthalmology Gustav Stålhammar from Karolinska Institutet. I have learnt a lot about the medical perspective while working together on projects.

Last but not least I would like to say thank you to my family, members of the MI^2.AI research group and friends for their support.

Abstract

Deep Learning (DL) models are perceived as the so-called black boxes that are difficult to interpret due to a large number of trainable parameters. However, due to legal requirements and stakeholders' expectations, various techniques were introduced to shed more light on the model decision-making process. Information on the model inner workings can be valuable insight for model developers. This is the goal of the so-called RED XAI (Explainable Artificial Intelligence), which prioritizes explanations for model validation and debugging, in contrast to BLUE XAI, which focuses on aligning explanations with human understanding and values.

When making decisions, it is important to have as complete information as possible. This often includes contextual information, which supports a holistic judgment and leads to more informed decisions. This should also apply to Deep Learning models. The contextual information is especially important when processing data from street surveillance systems, autonomous cars, robots, healthcare, agriculture, etc.

There are various types of contextual information, including spatial and temporal, which are the two main types investigated in this thesis. In the dissertation, I examine the perception of spatial contextual information by Deep Learning models, with digital pathology and visual IQ tests serving as illustrative use cases. Additionally, temporal context is analysed in a model continuously classifying tweets. These investigations align with the principles of Red XAI.

Moreover, I showed that popular XAI techniques do not account for spatial relationships between objects in the input images, limiting themselves to investigations of the role of single objects in model decision-making process. The lack of dedicated methods for analyzing such relationships is identified as a research gap to be addressed in the thesis.

Several contributions have been made to investigate the role of spatial context at various scales (from local to global) in deep neural networks used for tissue classification. To inspect the role of local context in convolutional and transformer-based models in classifying tissue patches, an occlusion-based method was proposed. The study examined whether the models are sensitive to the lack of contextual information in the form of tissue surrounding the patch. Moreover, the role of semiglobal and global context (such as image compositionality and long-range neighborhood) was evaluated in Multiple Instance Learning (MIL) models which are often used to process high resolution images such as

Whole Slide Images in digital pathology. With such an aim, the set of three datasets (SyntheticSMIL), along with the ReasonSMIL evaluation protocol and Deep Spatial Context method were introduced. The Deep Spatial Context is a mathematically grounded approach for quantifying spatial reasoning, broadening, together with the occlusion-based method, the arsenal of *spatial XAI* techniques. Lastly, comparative analysis was conducted between MIL models trained on Whole Slide Images and those trained on subsets of their constituent patches. The empirical results show that extra effort in the selection of regions of interest (ROIs) by physicians does not improve model performance compared with using raw data. It was also observed that in inter-evaluations performed on WSIs, the model trained on random patches achieves the highest performance among models trained on subsets of WSIs defined as ROIs or top-attended patches extracted from the model trained on WSIs.

Moreover, the spatial context was also addressed when designing the Duel-IQ model to solve visual IQ tests, which require the understanding of spatial relationships.

The role of temporal context is explored in the LIMES method, where it is used in the adaptation mechanism to handle class-prior shifts over time.

In conclusion, this dissertation demonstrates that contextual information (both spatial and temporal) plays an important role in deep neural network classification across its scales (from local to global) and domains.

Keywords: spatial context, digital pathology, Multiple Instance Learning, temporal context, explainable Artificial Intelligence.

Abstrakt

Modele uczenia głębokiego (oparte na głębokich sieciach neuronowych) są postrzegane jako tzw. czarne skrzynki, które są trudne do interpretacji ze względu na dużą liczbę trenowalnych parametrów. Jednak z uwagi na wymogi prawne oraz oczekiwania interesariuszy, opracowano różne techniki mające na celu lepsze wyjaśnienie sposobu analizowania danych i podejmowania decyzji przez modele. Informacja o tym jak działają różne komponenty sieci neuronowej jest ważna dla osób tworzących modele. To jest celem tzw. czerwonej wyjaśnialnej sztucznej inteligencji (ang. Explainable AI, XAI), która stawia na pierwszym miejscu wyjaśnienia służące technicznej walidacji modelu, w przeciwieństwie do niebieskiej XAI, koncentrującej się na dopasowywaniu wyjaśnień do sposobu myślenia ludzi i ich wartości.

Podczas podejmowania decyzji istotne jest, aby dysponować możliwie jak najbardziej kompletnymi informacjami. Często obejmuje to informacje kontekstowe, które pozwalają na holistyczną ocenę sytuacji i w efekcie bardziej świadome decyzje. Zasada ta powinna również odnosić się do modeli uczenia głębokiego. Informacje kontekstowe mają szczególne znaczenie w przetwarzaniu danych pochodzących z systemów monitoringu ulicznego, samochodów autonomicznych, robotów, opieki zdrowotnej, rolnictwa itp.

Istnieją różne rodzaje informacji kontekstowej, w tym przestrzenna i temporalna, które są głównym przedmiotem analiz w niniejszej rozprawie. Postrzeganie kontekstu przestrzennego przez modele oparte na sieciach neuronowych jest badane na przykładach cyfrowej histopatologii oraz wizualnych testów IQ. Z kolei kontekst temporalny analizowany jest w modelu służącym do klasyfikacji tweetów w trybie ciągłym. Prowadzone badania są zgodne z założeniami czerwonej XAI.

Ponadto pokazano, że popularne techniki XAI nie uwzględniają relacji przestrzennych pomiędzy obiektami w obrazach wejściowych, ograniczając się jedynie do analizy roli pojedynczych obiektów w procesie podejmowania decyzji przez model. Oznacza to, że modele głębokie nie mogą być właściwie ewaluowane, gdy istotna jest informacja kontekstowa. Brak dedykowanych metod do analizy takich relacji został wskazany jako luka badawcza, którą niniejsza praca doktorska podejmuje.

Opracowano szereg rozwiązań umożliwiających badanie roli kontekstu przestrzennego o różnym rozmiarze (od lokalnego do globalnego) w sieciach neuronowych stosowanych do klasyfikacji tkanek. Aby zbadać rolę kontekstu lokalnego w modelach konwolucyjnych i

opartych na transformerach do klasyfikacji łatek (ang. *patches*) przedstawiających tkankę, zaproponowano metodę opartą na zasłonięciu kontekstu. Badano, czy modele są wrażliwe na brak informacji kontekstowej w postaci tkanki otaczającej analizowaną łatkę. Ponadto oceniono rolę kontekstu semiglobalnego i globalnego (kompozycyjności obrazu i relacji sąsiedztwa dalekiego zasięgu) w modelach opartych na wielu instancjach (ang. *Multiple Instance Learning*), które są często wykorzystywane do analizy obrazów wysokiej rozdzielczości, takich jak obraz całej tkanki (ang. *Whole Slide Images*) w histopatologii cyfrowej. W tym celu opracowano trzy zestawy danych (*SyntheticSMIL*), protokół ewaluacyjny (*ReasonSMIL*) oraz metodę *Deep Spatial Context*. Metoda *Deep Spatial Context* jest matematycznie ugruntowanym podejściem do ilościowej oceny rozumowania w oparciu o informację przestrzenną. Metoda ta razem z tą opartą na zasłanianiu kontekstu poszerzają arsenał technik w ramach "przestrzennej wyjaśnialnej sztucznej inteligencji". Przeprowadzono również analizę porównawczą skuteczności modeli trenowanych na zdjęciach całych tkanek a tych trenowanych na podzbiorze łatek (regionów) z każdego oryginalnego zdjęcia. Wyniki eksperymentalne wskazują, że dodatkowy wysiłek włożony w wybór kluczowych regionów przez lekarza nie przekłada się na lepszą skuteczność modeli w porównaniu z wariantem, gdy użyte są surowe dane z całymi tkankami. Zaobserwowano, że podczas testowania modeli na surowych danych w ramach inter-ewaluacji, najlepszą skuteczność, spośród modeli trenowanych na podzbiorach regionów, osiągnął model trenowany na losowo wybranych regionach. Tym samym był to lepszy model niż te trenowane na regionach wybranych przez lekarza oraz regionach wyselekcjonowanych na podstawie najwyższych wartości atencji z modelu wytrenowanego na surowych danych.

Ponadto, kontekst przestrzenny został uwzględniony w projektowaniu algorytmu *DuelIQ* do rozwiązywania wizualnych testów IQ, które wymagają zrozumienia relacji przestrzennych.

Rola kontekstu temporalnego była eksplorowana w ramach zaproponowanej metody *LIMES*, w której kontekst temporalny jest wykorzystany w mechanizmie adaptacji do zmiennego rozkładu klas w danych.

Podsumowując, w ramach rozprawy doktorskiej pokazano, że informacja kontekstowa (zarówno przestrzenna, jak i temporalna) odgrywa kluczową rolę w klasyfikacji realizowanej przez głębokie sieci neuronowe. Jest to obserwowane przy różnych rozmiarach kontekstu (od lokalnego do globalnego) i w różnych obszarach.

Słowa kluczowe: kontekst przestrzenny, cyfrowa histopatologia, modele oparte na wielu instancjach, kontekst przestrzenny, wyjaśnialna sztuczna inteligencja.

Contents

1 Introduction **13**
1.1 Research hypothesis . . . 14
1.2 Research questions . . . 15
1.3 List of papers included in the thesis . . . 17
1.4 List of papers not included in the thesis . . . 19
1.5 Structure of the dissertation . . . 20

2 Background **23**
2.1 Model evaluation . . . 23
2.2 Context . . . 31

3 P1: Lightweight Conditional Model Extrapolation for Streaming Data under Class-Prior Shift **39**
3.1 Preface . . . 39
3.2 Introduction . . . 40
3.3 Method . . . 41
3.4 Experimental Evaluation . . . 47
3.5 Summary and Conclusions . . . 51

4 P2: Position: Do not explain vision models without context **53**
4.1 Preface . . . 53
4.2 Introduction . . . 54
4.3 Contextual Information within Images . . . 56
4.4 Application domains . . . 59
4.5 Deep Learning model vs. spatial reasoning – datasets . . . 61
4.6 Spatial context in DL models . . . 64
4.7 XAI Failures . . . 67
4.8 Spatial context in XAI . . . 69
4.9 Needed Research Directions . . . 71
4.10 Spatial XAI within Pipeline . . . 72
4.11 Conclusions . . . 73

5 P3: Duel-based Deep Learning system for solving IQ tests 75
5.1 Preface . . . 75
5.2 Introduction . . . 76
5.3 Problem description . . . 77
5.4 Related work . . . 78
5.5 Proposed solution . . . 79
5.6 Experimental setup . . . 82
5.7 Experimental results . . . 84
5.8 Human performance . . . 91
5.9 Conclusions . . . 92

6 Digital pathology 93
6.1 Multiple Instance Learning . . . 94

7 P4: Does context matter in digital pathology? 101
7.1 Preface . . . 101
7.2 Introduction . . . 102
7.3 Motivation . . . 103
7.4 Method . . . 104
7.5 Experiments . . . 105
7.6 Results . . . 105
7.7 Conclusions . . . 112

8 P5: From Input to Insight: Probing the Reasoning of Attention-based MIL Models 115
8.1 Preface . . . 115
8.2 Introduction . . . 118
8.3 Related work . . . 120
8.4 SyntheticSMIL datasets . . . 121
8.5 ReasonSMIL protocol . . . 126
8.6 Experiments . . . 127
8.7 Results . . . 128
8.8 Conclusions . . . 138

9 P6: Information Injection to Deep Learning Solutions in Knowledge Transfer 141
9.1 Preface . . . 141
9.2 Introduction . . . 142
9.3 Proposed Contribution . . . 144

10 P7: Deep Spatial Context: When Attention-based Models Meet Spatial Regression **147**
10.1 Preface . . . 147
10.2 Introduction . . . 150
10.3 Related work . . . 152
10.4 Preliminaries . . . 152
10.5 Deep Spatial Context . . . 153
10.6 Deep Learning models . . . 155
10.7 Application of *Deep Spatial Context* . . . 157
10.8 Results . . . 158
10.9 Conclusion . . . 168

11 P8: Role of patch diversity while predicting time of death due to uveal melanoma metastasis **169**
11.1 Preface . . . 169
11.2 Introduction . . . 171
11.3 Dataset and Task . . . 172
11.4 Model trained on WSI . . . 173
11.5 Model trained on subsets of patches from WSIs . . . 174
11.6 Patch inspection . . . 176
11.7 Conclusions . . . 179

12 Summary **181**

13 Discussion and Future Work **185**
13.1 Other types of model architectures . . . 185
13.2 Conflicting information . . . 185

Chapter 1

Introduction

Deep Learning (DL) is a rapidly evolving field of Artificial Intelligence (AI), built upon artificial deep neural networks (hereinafter referred to simply as neural networks for brevity). Deep Learning models are perceived as *black boxes* due to their complex architectures with a large number of trainable parameters (Guidotti et al., 2018; Hassija et al., 2024). This makes it challenging to decipher why a model made a particular prediction. This lack of transparency poses significant concerns, particularly in high-stakes domains such as healthcare or autonomous driving, where decisions must be correct, responsible, and explainable (Longo et al., 2024). There is an expectation that the reasoning of AI systems will align with human knowledge and values. It means that it is expected that models will base their predictions on features that a human expert would also consider relevant.[1] However, systematic evaluation and interpretation of the behavior of Deep Learning models remains a formidable challenge.

Various approaches are taken to inspect the key decision-making factors from the perspective of Deep Learning models. Today this area of research is called eXplainable AI (XAI) (Holzinger et al., 2020; Ali et al., 2023). There are a number of methods (Simonyan et al., 2014; Ribeiro et al., 2016; Sundararajan et al., 2017; Ancona et al., 2018; Chattopadhay et al., 2018) that were introduced to investigate Deep Learning models for image classification. The outputs of these methods are colorful heatmaps that highlight the key pixels (regions) in the input image from the model decision-making perspective. Moreover, there are attempts to name the concepts that a model bases its prediction on (Achtibat et al., 2022; Hernandez et al., 2022; Oikarinen and Weng, 2023; Bai et al., 2025). The concepts are often atomic, simple objects (e.g. beak) or features (e.g. striped). **Although XAI is a rapidly evolving field, there is still little attention paid to the relationships between objects (which can be translated as a form of context), even though relationships such as distance, rotation, inside/outside and order, may play a significant role in applications such as street surveillance,**

[1] At this stage, we omit a discussion on the field known as Explainable Machine Learning for Scientific Insights and Discoveries (Roscher et al., 2020), which will be outlined in Chapter 11.

autonomous driving, land analysis, archaeology and healthcare. This is the research gap that I want to address in my PhD dissertation. My aim is to focus on the investigation of an abstract concept within Deep Learning models which is context (spatial and temporal).

My investigation is inspired by the fact that human decision-making in everyday life relies on integration of multiple sensory cues and contextual information. Such an approach helps in accurate and holistic judgments. If context plays such a central role in human reasoning, it is natural to ask whether Deep Learning models operate in a similar manner.

1.1 Research hypothesis

The research hypothesis of this dissertation is formulated as follows: **Contextual information plays an important role in a decision-making process of Deep Learning classifiers.** The topic is analysed along several axes (Table 1.1):

- type of context (temporal — spatial context)

  There are several types of context: textual (provided by surrounding text), temporal (where the time axis plays a key role) and spatial (defined by the relationships between objects or regions within images). In the thesis, I investigate the last two types, with the main focus on spatial context.

- context scale (global — semiglobal — local context)

  Regardless of its type, context can be perceived over shorter or longer spans. In this thesis, I focus on investigation of context, defined at varying scales ranging from local to global.

- use case (histopathology — IQ tests — tweets — other)
  The context plays an important role in various applications. In the thesis, the main focus is on the investigation of the role of context in models suitable for the analysis of histopathological images. There are also other use cases described, like the classification of the country of origin of the tweets and solving visual IQ tests by choosing an answer from a given set of possibilities.

- stage of model development (Informed Machine Learning — explainable AI)

  This axis concerns at which stage of the development process of Deep Learning models, the role of the context is investigated. In Informed Machine Learning (von Rueden et al., 2023), the prior knowledge (in the form of contextual information, inductive biases etc.) is incorporated into the model design a priori (before the model training begins). It is expected that it will guide a model towards more transparent

and reliable decision-making. In the case of explainable AI, the already trained model is analysed post-hoc to understand whether it benefits from the contextual information within input data. The analysis along this axis distinguishes between approaches that are inspired by the paradigm of models explainable by design and post-hoc analyses of black-box models.

Table 1.1: Axes regarding context investigated in the thesis.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

1.2 Research questions

The research hypothesis was investigated by decomposing it into the following research questions. In the square brackets below, the references to the papers are indicated (more details are given in Section 1.3).

RQ1. Is incorporation of class distribution from the historical timestamp beneficial for the model when predicting classes at the next timestamp in the continuous training setup with class-prior shift? [P1]

Selecting the most accurate historical record to serve as temporal context for a model is an interesting research question. The choice of temporal contextual information should account for changes in the time series data, including periodicity, trends, and other relevant patterns.

RQ2. Can widely adopted explainable AI (XAI) approaches to inspect Deep Learning Computer Vision models account for spatial context in data? [P2]

Ensuring that Deep Learning models are responsible and aligned is crucial. The models should be able to benefit from contextual information as humans do. This brings up the question of whether popular XAI methods are capable of analysing the model decisions through the lens of contextual understanding. A thorough analysis of this issue may lead to a shift in the current paradigm within the explainable AI field.

RQ3. Can a model improve its decision-making by comparing one option with an alternative in terms of how well each fits various spatial relationships? [P3]

There are various ways of assessing the ability of abstract reasoning. One of them is via performance analysis in solving visual IQ tests in the form of Raven Progressive

Matrices (RPMs). In such a test, it may happen that more than one answer fits the spatial relationships in a given RPM but the correct answer is the one that has an easier explanation. Due to such a relativity, it may be valuable to create a Deep Learning system where various potential RPM answers are compared against each other before the model returns the final answer.

RQ4. Does the limitation of surrounding region (context) impact the classification performance of patch-based image models? [P4]

The inspection of the role of surrounding tissue (local context) when predicting if the sample is healthy or with tumor can bring valuable insights for the developers of the Deep Learning models for digital pathology.

RQ5. Are Deep Learning models used for an analysis of histopathological images capable of encoding and leveraging spatial information? [P5], [P6]

A special class of Deep Learning models is used to process the high resolution images of tissue (Whole Slide Images). These are called Multiple Instance Learning (MIL) models. It would be important to investigate these models from the perspective of spatial context to check if they are in line with histopathologists who analyse samples at tissue (global) and cell (local) levels in their daily practice.

RQ6. Is it advantageous for the widely used CLAM model, as a representative approach in the Multiple Instance Learning domain, to rely on physician-selected regions of interest as input, or is it preferable to use raw Whole Slide Images directly providing full context? [P8]

To mitigate the risk of model failure and improve its performance, often the irrelevant information is excluded from the input. One can imagine a similar approach applied in the case of models analyzing histopathological images where instead of using Whole Slide Images (WSIs) as input, the model will be provided with only regions of interest (ROIs) selected by a physician. The selection of regions is an extra burden but if it results in significant performance boost, it may be worth it.

RQ7. Does analysis of similarity between patch features allow one to reconstruct the tissue structure? [P7]

Such similarity measures can provide insight on the features extracted from a pre-trained network. Moreover, such measures can potentially be injected into the model in the Informed Machine Learning paradigm.

1.3 List of papers included in the thesis

[P1] **Paulina Tomaszewska** & Christoph H. Lampert. (2022). "Lightweight Conditional Model Extrapolation for Streaming Data under Class-Prior Shift". Proceedings of the International Conference on Pattern Recognition (ICPR).

[P2] **Paulina Tomaszewska** & Przemysław Biecek. (2024). "Position: Do Not Explain Vision Models Without Context". Proceedings of the International Conference on Machine Learning (ICML).

[P3] **Paulina Tomaszewska**, Adam Żychowski & Jacek Mańdziuk. (2022). "Duel-based Deep Learning system for solving IQ tests". Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS).

[P4] **Paulina Tomaszewska**, Mateusz Sperkowski & Przemysław Biecek. (2023). "Does context matter in digital pathology?". Workshop on Modelling and Representing Context (MRC) at European Conference on Artificial Intelligence (ECAI).

[P5] **Paulina Tomaszewska**, Michał Gozdera, Elżbieta Sienkiewicz & Przemysław Biecek. (2025). "From Input to Insight: Probing the Reasoning of Attention-based MIL Models". Proceedings of the World Conference on eXplainable Artificial Intelligence.

[P6] **Paulina Tomaszewska**, Elżbieta Sienkiewicz, Mai P. Hoang & Przemysław Biecek. (2025). "Deep Spatial Context: When Attention-based Models Meet Spatial Regression (Digital Pathology)". Proceedings of the International Conference on Artificial Intelligence in Medicine (AIME).

[P7] **Paulina Tomaszewska**. (2022). "Information Injection to Deep Learning Solutions in Knowledge Transfer". Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Doctoral Consortium.

[P8] **Paulina Tomaszewska**, Krzysztof Kotowski, Piotr Donizy, Gustav Stålhammar & Przemysław Biecek. (—). "Role of patch diversity while predicting time of death due to uveal melanoma metastasis". version submitted to the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2026.

The relationships between the papers and the categories regarding context to which they belong are depicted in Figure 1.1. The main outcomes of the papers constituting the dissertation are outlined in Figure 1.2. In P1, P3, P4, P5, P6 and P8, the code was released on GitHub repositories to facilitate other researchers reproduction of the results and application of the same methodology in their use cases.

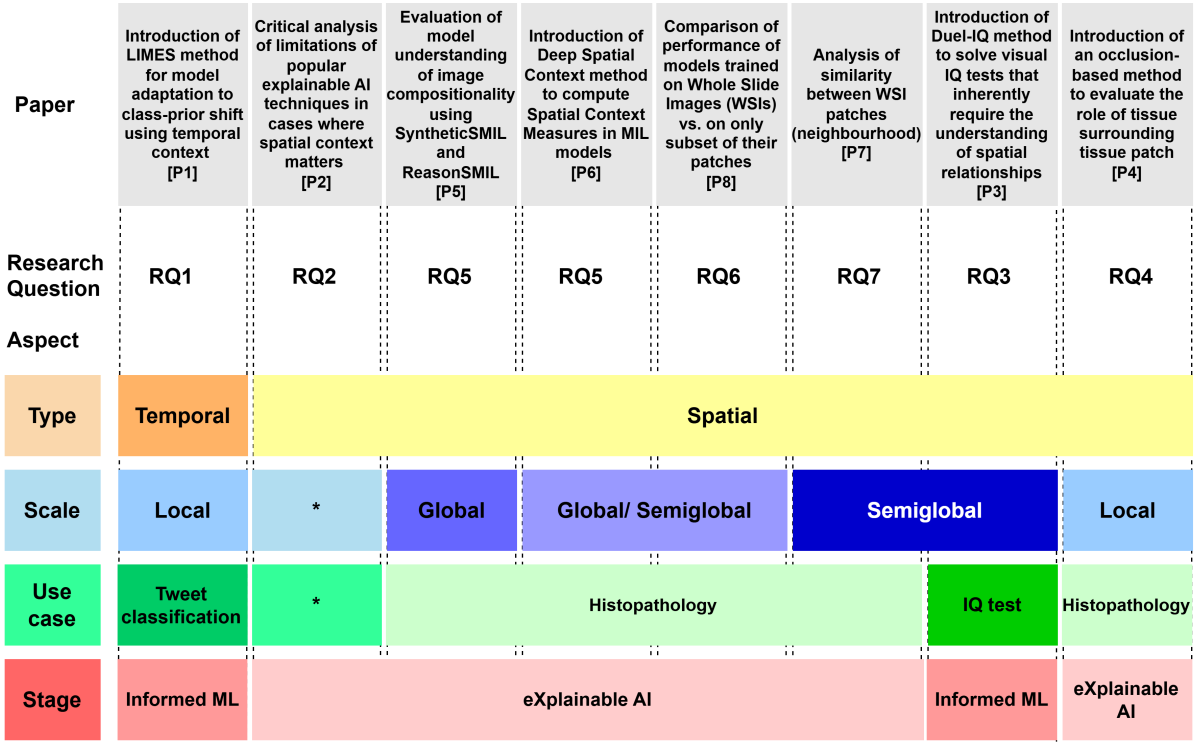


Figure 1.1: Mapping of papers and the corresponding research questions to axes regarding context (type, scale, use case, stage). The sign '*' means the work addresses all the values defined within the axis. The information in square brackets directs to the details on the papers provided in Section 1.3.

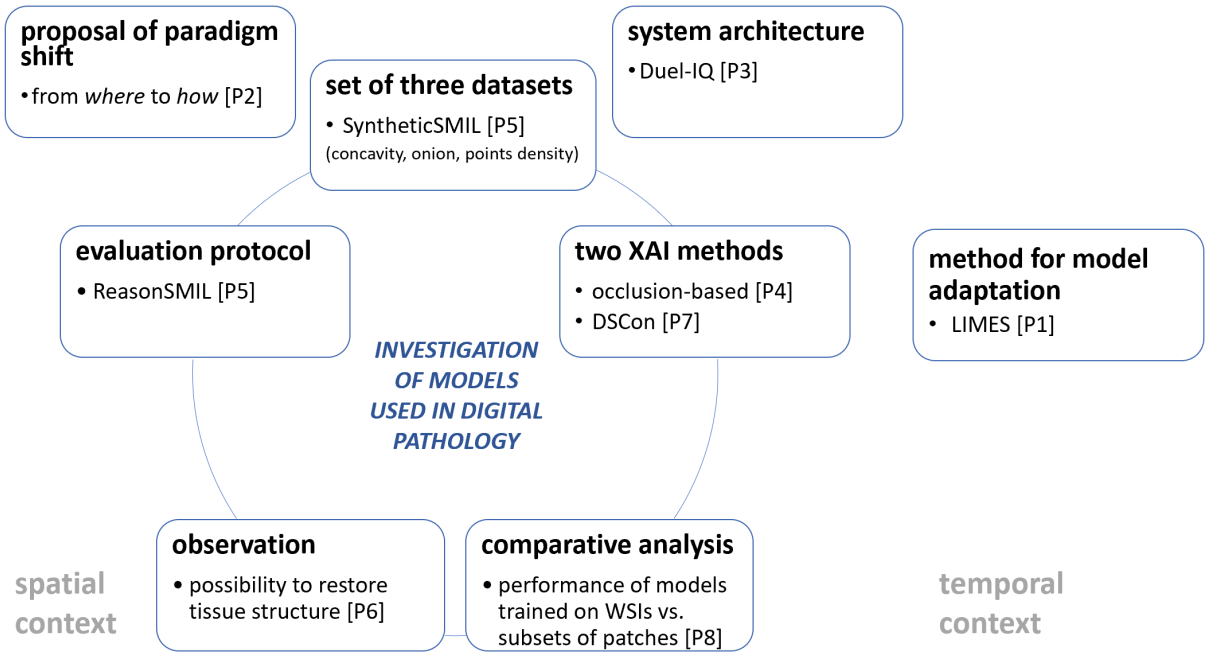


Figure 1.2: The key outcomes developed in this dissertation.

1.4 List of papers not included in the thesis

During PhD studies, I was involved in collaborations with researchers from various external institutes, including foreign ones. Some of the resulting papers are not included in the thesis, as they are of limited relevance to its main topic.

[A9] Weronika Hryniewska, Przemysław Bombiński, Patryk Szatkowski, **Paulina Tomaszewska**, Artur Przelaskowski & Przemysław Biecek. (2021). "Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies". Pattern Recognition. 118. 1–17.

[A10] Maciej Tyszko, Anna Lemańska-Perek, Jakub Śmiechowicz, **Paulina Tomaszewska**, Przemysław Biecek, Waldemar Goździk & Barbara Adamik. (2023). "Citrulline, Intestinal Fatty Acid-Binding Protein and the Acute Gastrointestinal Injury Score as Predictors of Gastrointestinal Failure in Patients with Sepsis and Septic Shock". Nutrients. 15. Article 9.

[A11] Georg Starke, Felix Gille, Alberto Termine, Yves Saint James Aquino, Ricardo Chavarriaga, Andrea Ferrario, Janna Hastings, Karin Jongsma, Philipp Kellmeyer, Bogdan Kulynych, Emily Postan, Elise Racine, Derya Sahin, **Paulina Tomaszewska**, Karina Vold, Jamie Webb, Alessandro Facchini & Marcello Ienca. (2025). "Finding Consensus on Trust in AI in Health Care: Recommendations From a Panel of International Experts". Journal of Medical Internet Research.

[A12] Travis Mitchell, **Paulina Tomaszewska**, Łukasz Łaniewski-Wołłk & Christopher Leonardi. (2022). "Automatic construction of open-boundary conditions for lattice Boltzmann methods". Proceedings of the Australasian Fluid Mechanics Conference.

[A13] Bertrand Kerautret, Konstantin Kirchheim, Daniel Lopresti, Phuong Ngo & **Paulina Tomaszewska**. (2023). "Promoting Reproducibility of Research Results in International Events (Report from the 4th RRPR)". Proceedings of the Reproducible Research in Pattern Recognition Workshop at the International Conference on Pattern Recognition (ICPR).

[A14] **Paulina Tomaszewska** & Christoph H. Lampert.(2023). "On the Implementation of Baselines and Lightweight Conditional Model Extrapolation (LIMES) Under Class-Prior Shift". Proceedings of the Reproducible Research in Pattern Recognition Workshop at the International Conference on Pattern Recognition (ICPR).

[A15] **Paulina Tomaszewska**, Marlena Broniszewska, Mariusz Kaleta. (2024). "Initialization of an Evolutionary Strategy Using the Weibull Distribution for Non-Pharmaceutical Epidemic Policy". Prace naukowe Wydziału Elektroniki i Technik

Informacyjnych Politechniki Warszawskiej. vol. 3. Oficyna Wydawnicza Politechniki Warszawskiej.

The first three papers [A9], [A10], [A11] deal with the use of AI in healthcare. The following paper [A12] proposes the use of neural networks in computational fluid dynamics to allow more accurate simulations in a shorter domain that results in a decrease in computational costs. It is achieved by the proposed automatic way of constructing open-boundary conditions. The papers [A13], [A14] deal with the issue of reproducibility in AI research. The paper [A13] provides guidelines for the conference organizers on possible actions to enhance reproducibility whereas the paper [A14] provides implementation details and measures taken to facilitate reproducibility and replicability of the method proposed in paper [P1]. The paper [A15] proposes the initialization scheme for the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm (Hansen and Ostermeier, 2001) to increase convergence when modelling the non-pharmaceutical policy in the pandemic situation.

1.5 Structure of the dissertation

The dissertation starts with Chapter 2 that provides background information on model evaluation, including detection of concepts within neural networks and the literature analysis of different approaches to context incorporation in models processing different data modalities. The chapters dedicated to my papers start with a preface where the paper is put into a broader context.

Chapter 3 is the first chapter about my publications. It refers to the topic of temporal context. I propose the LIMES method, where temporal context is incorporated into the model adaptation mechanism to address the class-prior shift in a continuous training scheme. The work was a result of my research internship at the Institute of Science and Technology Austria.

Later, the focus in the thesis is on spatial context, its importance and the fact that it is an overlooked topic when investigating Computer Vision models (Chapter 4). Examples are provided where popular techniques from the field of Explainable AI fail, particularly when the ground truth label depends on spatial relationships. These cases serve as a motivation for the analysis and development of new techniques presented in the thesis.

Chapter 5 provides the description of the new neural-based algorithm called Duel-IQ suitable for solving visual IQ tests that require the understanding of spatial relationships between the figures in the 3×3 grid of panels constituting Raven's Progressive Matrices.

Subsequently, in Chapter 6, there is an introduction to the Deep Learning models used for the processing of histopathological images, including Multiple Instance Learning models. This is a starting point for the following five chapters, where I analyse the role

of context in models dealing with histopathological images. In this regard, the set of synthetic datasets, two evaluation protocols and the XAI method are proposed. Also, the comparative analysis involving intra- and inter-evaluations is performed. In the thesis, the analyses are undertaken at various context scales (from local to global).

The narrative begins with the work where I propose the occlusion-based evaluation protocol to inspect the role of local context defined as a small area of neighbouring tissue around the patch (Chapter 7).

In the following chapters, the focus is on semiglobal and global context which is present within bigger images (Whole Slide Images, WSIs). Analysis of such images requires the use of a dedicated class of Deep Learning models called Multiple Instance Learning models. First, I propose the set of synthetic datasets (SyntheticSMIL) and the evaluation protocol (ReasonSMIL) to examine the properties and abilities of MIL models to understand image compositionality (Chapter 8).

In subsequent chapters, real high-resolution histopathological images are used. In Chapter 9, it is shown that by the analysis of the similarity of features of neighboring patches, the structure of the tissue can be restored. Later, the Deep Spatial Context method is proposed where by the introduction of three Spatial Context Measures, it is possible to analyse if spatial context is preserved within features, targets or residuals extracted from Multiple Instance Learning models. This property is examined under scenarios involving different input data subsets, as well as across various model variants and families (Chapter 10). Subsequently, in Chapter 11, there is a comparative analysis of model intra- and inter-performance when the input to models are WSIs or only subsets of the WSI patches (selected randomly, manually as regions of interest by a physician or based on the attention scores extracted from the trained model). The conclusions drawn from the evaluation results are further strengthened by an analysis of diversity measures at different levels of granularity. This provides insight into the underlying factors that drive the observed outcomes. The results are supported by the patch inspection made by a physician. The dissertation ends with chapters with a summary (Chapter 12) and discussion of plans for future work (Chapter 13).

Chapter 2

Background

2.1 Model evaluation

Deep Learning (DL) models are particularly effective at processing complex data such as images, speech, and text. Unlike traditional Machine Learning, which often relies on ante-hoc feature engineering, Deep Learning models learn features directly from raw input. It is believed that the early layers typically capture low-level patterns (e.g., edges in images), while deeper layers learn more abstract representations (e.g., complex objects) (Zeiler and Fergus, 2014).

Recent Deep Learning models achieve similar or superior performance to that of humans in various tasks (Figure 2.1). For example, in the classification of images from the Imagenet dataset, best Deep Learning models achieve a lower error rate than humans (Alzubaidi et al., 2021). In addition, the ensemble of Deep Learning models outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images (Zhou et al., 2021b). The question is, however, whether DL models are reliable and whether they take into account similar aspects as humans when solving their tasks. Moreover, the question is whether the datasets are challenging enough (more on the topic of benchmarking in the next section).

One of the key analyses preceding the decision whether the model can be deployed is a verification of model performance on data that was not used during model training. Models are also evaluated for stability, robustness to input perturbations, and resilience against adversarial attacks. Verification is perceived as one of the pillars of Model Science (Biecek and Samek, 2025).

There are various challenges when evaluating models. It may happen that models overfit to the public test sets. Recht et al. (2019) introduced a new test set for image classification with objects from the Imagenet dataset and found that many models, that performed well on the original test set, exhibited a noticeable drop in accuracy on the

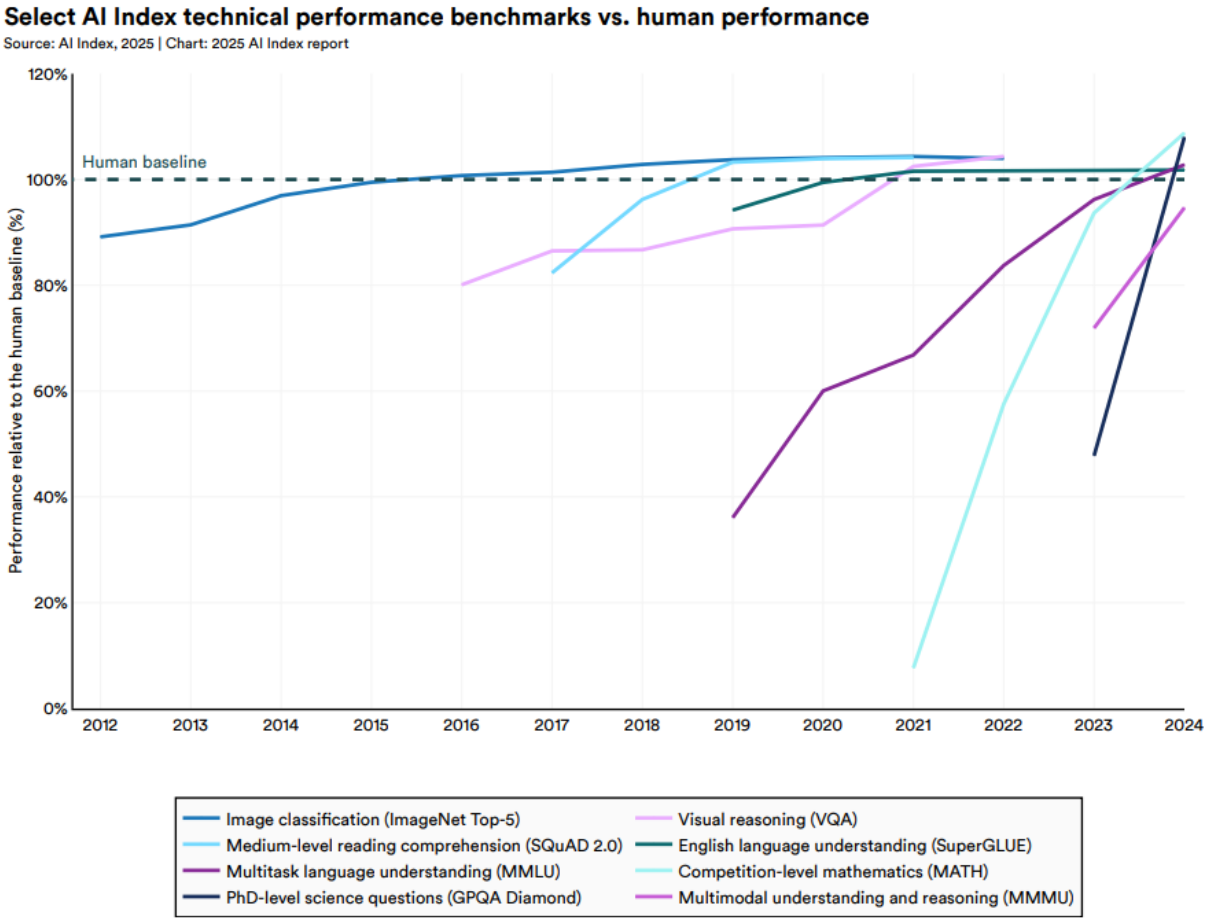


Figure 2.1: AI performance vs. human performance in benchmarks. The values are scaled to establish a standard metric for comparison of different benchmarks. The scaling function is calibrated such that the performance of the best model for each year is measured as a percentage of the human baseline in a given task. A value of 105% indicates, for example, that a model performs 5% better than the human baseline. Source: Maslej et al. (2025).

new test set. It suggested that some form of overfitting to the original test set could have occurred. To mitigate this problem, more complex benchmarks were introduced to assess model performance. The potential solution could be the use of dynamic benchmarks in which the test data would undergo some changes to mitigate the saturation of model performance observed in the case of static benchmarks (Kiela et al., 2021). Another challenge is shortcut learning, which may not be caught during model evaluation if the test set is generated in the same way as the training and validation sets that feature some spurious correlations. This problem was highlighted by Dubanowska et al. (2025), who showed that hallucination detectors based on internal values of Large Language Models were not reliable. Their performance is strongly correlated with the specific task on which they were evaluated (data-to-text, summarization, or question answering) instead of the general hallucination patterns. As a result, the detectors' performance on a dataset without clearly defined tasks (claimed as out of distribution) was close to random.

The problem of evaluation is even more difficult in the case of generative models where there is no fixed ground truth as opposed to scenarios where discriminative models are used. In the text generation tasks, often the *LLM-as-a-judge* approach (Gu et al., 2026) is applied where one language model (usually a larger one) evaluates the output generated by another model (typically a smaller one). The key advantage is scalability but there are concerns about reliability of such evaluations. Another method of assessing the quality of the generated text is via the public preference leaderboards (Chiang et al., 2024).

The other challenging area is an analysis of the reasoning capabilities of Large Reasoning Models (LRMs). Shojaee et al. (2025) claimed that these skills are limited on puzzles beyond certain complexity thresholds. However, a week later Lawsen (2025) wrote a comment paper claiming that the conclusions in the paper by Shojaee were the result of mistakes in the experimental design. It was noted that the Tower of Hanoi task assigned to the LRM required producing more tokens than the experiment allowed. Moreover, it turned out that the model output was scored as failure even though the task (River Crossing benchmarks with $N > 5$) was unsolvable itself. Consequently, the automated evaluation framework was unable to differentiate between failures in model reasoning and the inherent unsolvability of certain tasks, leading to wrong conclusions about low model capabilities. Lawsen showed in preliminary experiments that after removing the methodological errors in the model evaluation protocol, the LRM solved the tasks with high accuracy. This highlights the need for well-suited and carefully designed model evaluations.

In this paragraph, the term reasoning was used in the context of Large Language Models where it refers to the ability to solve complex problems by decomposing them into smaller subproblems and addressing them sequentially through logically coherent steps to reach a conclusion. Importantly, evaluating reasoning skills may involve tasks that the model has not encountered during training. Reasoning can take various forms, including deductive, inductive, abductive, informal, and formal reasoning. As claimed by Huang and Chang (2023), the commonly investigated form of reasoning in LLMs is informal deductive reasoning. One prominent method used to facilitate such reasoning is Chain-of-Thought (Wei et al., 2022). Nonetheless, the definition of reasoning in LLMs remains abstract. **In the following parts of the dissertation, the term 'reasoning' will refer to the decision-making process in discriminative models.**

2.1.1 Need for an understanding of model behavior

Despite the model evaluation procedures undertaken to inspect the Deep Learning models, the understanding of their in-depth behavior remains limited. The difficulty stems from the black-box nature of such models, which arises from their complexity.

Transparency It is often anticipated that AI models will make decisions consistent with human understanding of a given phenomenon. Similarly, there is an expectation that the outputs of Large Language Models will reflect human values, for instance, by avoiding the generation of harmful content. Understanding a model's decision logic involves identifying which features, regions, or patterns drive its predictions. Further analysis can then determine whether the model relies on meaningful evidence or on spurious correlations (Geirhos et al., 2020; Dogra et al., 2024). However, the fact that model decision making process does not follow the human logic, does not have to mean that the model is wrong. It may be that the model identifies predictors that are not yet known by humans as important while solving a particular problem. For instance, not all histopathological factors influencing tumor metastases are yet well understood from the medical point of view (Riggi et al., 2018; Rigamonti et al., 2021; Tyagi, 2022). Therefore, it would be difficult to assess with high certainty the correctness of features used by the DL model to predict metastases. Even in Mathematics it is sometimes not known what input features or structures are crucial in a particular study. This was the case within knot theory, where mathematicians had some initial conjecture and hypothesis but were lacking the evidence (Davies et al., 2021). It turned out that the analysis of the attribution maps extracted from the model served as a guidance to ultimately develop a theorem.

Debugging and improvements There are, however, the cases where models clearly follow spurious correlations or irrelevant features. For instance, model reasons based on regions outside of the lungs to predict COVID-19 (DeGrave et al., 2021; Hryniewska et al., 2021); presence of a ruler to decide if a tissue is benign or malignant (Narla et al., 2018) or decides that there is a horse in the image only based on the text tag in the image corner (Lapuschkin et al., 2019). Models tend to be biased, e.g when segmenting lesions putting high importance on skin color (Benčević et al., 2024), or classifying an object in the image based on background (Ribeiro et al., 2016; Beery et al., 2018). Moreover, a model may over-rely on textures rather than focus on object shapes (Geirhos et al., 2019). It was also observed that models follow gender bias i.e. the person in the image is classified as a nurse only based on a female face (Selvaraju et al., 2017). All of these aspects can impact the ability of model to generalize to unseen data. Problems related to such shortcut learning, which lead to biased and unreliable models, can often be mitigated through improved data curation. However, identification of aspects of the dataset that require enhancement frequently relies on explainable AI methods.

Legal aspects The use of AI is regulated by law as it relates to the responsibility behind the autonomous model decisions, fair treatment of everyone, ethics and privacy. It is especially important in high-stake domains, where an AI system can be deployed in real-world scenarios only if these key aspects are ensured. This is especially a challenge

in healthcare (Starke et al., 2025) where physicians and patients are key stakeholders. They are afraid that AI systems will make wrong decisions i.e. based on some irrelevant features. Moreover, there are many risks for patients. For instance, due to false positive model decision, a patient could be exposed to unnecessary invasive treatment. In another scenario, due to false negative decision the disease could be overlooked which would delay the proper treatment leading to lower chances for full recovery. To mitigate risks, the model should be well tested before the deployment. It is imposed by law that proper analyses of the model's decision-making process will be performed to verify model correctness and fairness. In Article 13 of the EU Artificial Intelligence Act (European Commission, 2024), there is a claim that "high-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." Moreover, the Executive Order 14110 (The White House, 2023) issued in the United States was dedicated to "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence".

2.1.2 Explainable AI vs. Mechanistic Interpretability

The growing demand from various stakeholders for a better understanding of AI model decision-making has led to the emergence of the field of eXplainable AI (XAI) (Samek et al., 2019). This area offers insights into how decisions are made and identifies the input features that most influence the model's output. Often, the model reasoning is compared with the common knowledge or domain expertise. If there is alignment, users perceive the model as more trustworthy. This is the focus of the so-called Blue XAI (Biecek and Samek, 2024) which is human- and value-centered. The aim is to provide the model explanations that are understandable to end-users who do not have extensive knowledge about AI. The other approach is undertaken in the so-called Red XAI which is dedicated to technical model inspection, auditing, validation and debugging. The provided model explanations are addressed to AI researchers and developers who have a deep understanding of the technical aspects of AI. Therefore, the explanations in Red XAI are much more complex than in Blue XAI where user-friendliness and readability are crucial. It can be concluded that in Blue XAI the focus is on 'what' the model does, whereas Red XAI focus on 'why' model behaves in a particular way.

A field closely related to Red XAI is Mechanistic Interpretability (Rauker et al., 2023; Bereska and Gavves, 2024). It focuses on the automated fine-grained analysis of model internals (specific neurons, layers or circuits) in producing particular outputs, especially in Large Language Models (LLMs) and Large Vision Language Models. Often, analyses are performed on the activations extracted at different stages of the model. The gained insights can be used for changing model behaviour at inference time e.g. via activation steering (Turner et al., 2023; Rimsky et al., 2024; Cho and Hockenmaier, 2025; Zhao et al.,

2026).

Nowadays, the boundaries between the explainable AI and Mechanistic Interpretability are becoming increasingly blurred. In my dissertation, I draw inspiration from both fields.

2.1.3 Concepts in neural networks

In the early stages of explainable AI research on vision models, the primary focus was on highlighting key regions of the input image from the model's perspective ('where'). Later, the focus shifted towards 'what', therefore the aim was to name the most important input regions from the model's point of view. Various XAI techniques were introduced in line with this notion: Testing with Concept Activation Vectors (TCAV) (Kim et al., 2017), Automatic concept-based explanations (ACE) (Ghorbani et al., 2019), ConceptSHAP (Yeh et al., 2020), Concept Relevance Propagation (CRP) (Achtibat et al., 2022) etc.

In parallel, techniques to identify concepts at the neuron level of the vision model were introduced. In the nomenclature in this field, 'neuron' describes a channel in convolutional neural networks. The first work in this area was Network dissection (Bau et al., 2017). The authors proposed a Broaden dataset where each image is accompanied with fine-grained binary concept masks annotated by human. The particular neuron is considered sensitive to the particular concept c if the intersection over union score (IoU) between the ground-truth binary masks with respect to c and the binarized mask from the activations of the concerned neuron is greater than a threshold. Bau et al. (2020) replace the human-annotated labels with the outputs of a segmentation model. Note that this method is limited to the predefined set of concepts used to train the segmentation model[1].

This limitation is overcome in the MILAN (Hernandez et al., 2022) method where a Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997) is used to generate annotations (captions) for top activating images. Consequently, MILAN method requires a dataset preparation and a training of the LSTM model from scratch which are its drawbacks.

However, there are also training-free methods like CLIP-Dissect (Oikarinen and Weng, 2023) and Describe-and-Dissect (DnD) (Bai et al., 2025) to name concepts in neurons.

In CLIP-Dissect (Figure 2.2), the first step is to extract the embeddings (I_i) for images from the probing dataset and the text embeddings (T_j) for the predefined set of concept names using the CLIP model's encoder. Then, the concept-activation matrix P is computed where $P_{ij} = I_i * T_j$. Next, the activations of the target neuron are extracted for all images from the probing dataset. The activation for a given image refers here to a mean over spatial dimensions of the neuron output (activation map). Therefore, it is a scalar value. The last step is to compute a similarity score in the form of Soft Weighted Pointwise Mutual Information (which is an extension of Weighted Pointwise Mutual In-

[1] At that time, the Segment Anything Model (Kirillov et al., 2023) was not yet known.

formation (WPMI) (Schneider, 2005)) between columns of P (corresponding to concepts) and the vector of activations for all images from the probing dataset. The final label of the concept represented in the neuron is the one with the highest similarity score.

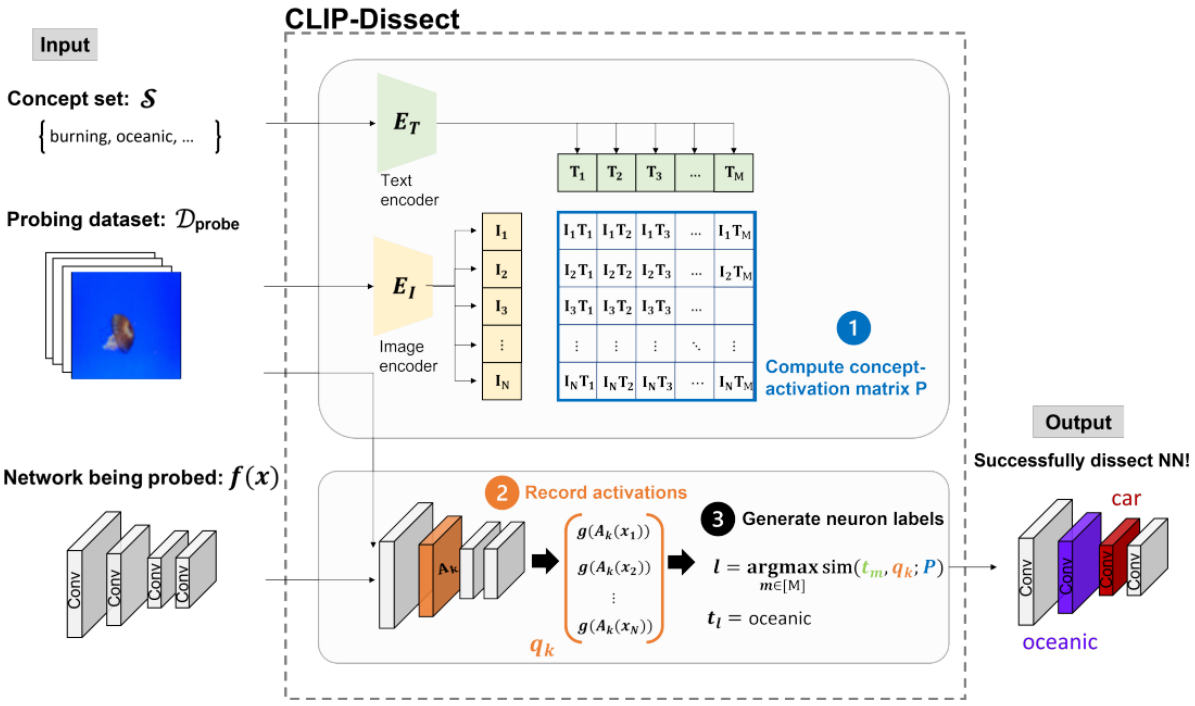


Figure 2.2: The CLIP-Dissect method. Source: Oikarinen and Weng (2023).

The Describe-and-Dissect method (Figure 2.3) benefits from already trained large models. First, the probing dataset is enriched with augmentations such as cropping and resizing. Later, BLIP model (Li et al., 2022) is used to generate captions for top activating images. The captions are summarized using GPT-3.5 Turbo model, which returns a set of candidate concepts. Subsequently, the process of selection of the most accurate candidate is performed. Here, the Stable Diffusion model (Rombach et al., 2021) is used to generate an image for each candidate concept. These generated images are then fed into the model under investigation, and during inference, the activations of the specified neuron are recorded. Lastly, all the extracted activations are analysed within the concept scoring module to select the best concept label for a given neuron.

To enable a proper comparison between various methods for naming concepts in neurons, the CoSy (Concept Synthesis) framework was proposed (Kopf et al., 2024). The key element of the framework is the generation of synthetic images depicting concepts named in a particular neuron using methods like MILAN, CLIP-Dissect, Describe-and-Dissect etc. In addition, a control set of images is created. The synthetic and control sets are then used for inference in the investigated Deep Learning model and the activation scores at a given neuron are collected. Later, the AUC score and Mean Activation Difference are computed to check if the activation scores for the synthetic set are significantly higher

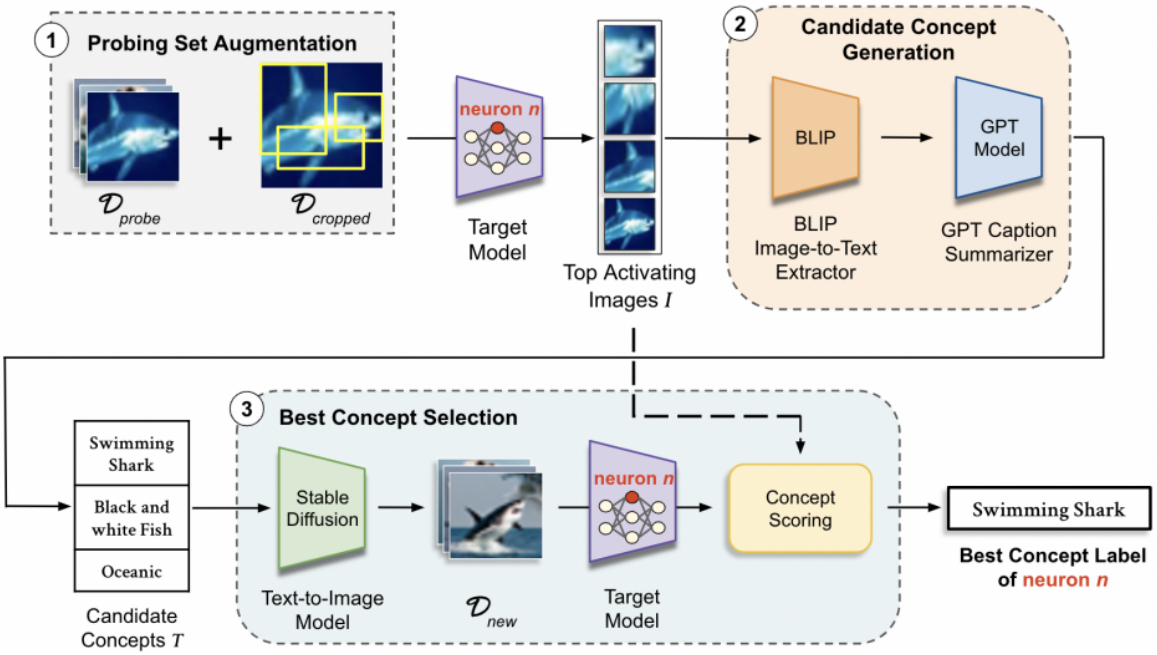


Figure 2.3: The Describe-and-Dissect method. Source: Bai et al. (2025).

than those for the control set. The high metrics suggest that the concept in a neuron was correctly assigned.

Discussion Note that in existing methods, the concepts examined in Deep Learning models are typically those with clear visual identifiers, making it possible to construct image sets representing a particular concept in order to test the model's sensitivity to it. For example, consider the concept of a cat: although there are many types of cats, a strong visual signal would be an image with a cat prominently positioned in the center with no distracting elements.

The question arises, however, whether it is equally straightforward to create clear signals for more complex concepts, such as spatial relationships (e.g., distance). One would expect such a dataset to contain images with at least two objects (A, B) on a plain background, arranged at varying distances.

However, if objects A and B in images in the probing dataset are always from the same two separate classes, then it is not clear whether the particular neuron fires up due to object A, object B, or the distance between them. In this case, additional image sets would be required: (1) with only object A in the center, and (2) with only object B in the center. If, during inference with these single-object sets, the neuron in question remains inactive, despite being active when both objects were present, we may then suspect that the neuron is sensitive to distance.

Alternatively, the solution would be to create images with objects A and B of various classes in different images. In this case, if the neuron activates, it can be assumed that it

is due to the distance concept.

This hypothetical scenario illustrates that the investigation of more complex concepts is inherently more challenging and more aspects have to be taken into account. **It may suggest that, due to the abstract nature of concepts like spatial relationships, dedicated XAI techniques must be developed.**

In all the aforementioned methods for naming concepts within neural networks, there was a strong assumption that only one concept is represented in a neuron. However, it may not be the case as the polysemantic hypothesis (superposition) is becoming more and more popular (Bereska and Gavves, 2024). Oikarinen and Weng (2024) show that the explanation of a neuron is a linear combination of concepts. This entanglement is especially significant in the case of transformer neural networks. Huben et al. (2024) propose to apply sparse autoencoders (SAEs) to perform the disentanglement of model activations. It is expected that the concepts in the SAE latent space are monosemantic. The concepts are named using the autointerpretability score (Paulo et al., 2025). The approach with sparse autoencoders comes also with challenges such as dead neurons (largely inactive neurons in the latent space) and feature absorption (Chanin et al., 2024). The latter can be observed, for instance, when in the SAE latent space one neuron activates for words starting with the letter 'L', another neuron responds specifically to the token 'laser', and a third neuron attends to the token 'lion'. Similarly, one neuron might activate for tokens related to mathematics, while another focuses specifically on the term 'algebra'. In such cases, we observe overlapping concepts in the latent space, which should ideally be avoided.

The other approaches to the problem of feature entanglement in neural networks are Concept Bottleneck Models (Koh et al., 2020) and Prototypical Neural Networks (Chen et al., 2019) which are largely interpretable-by-design.

2.2 Context

The contextual information plays a significant role in daily situations. As pointed out by Mehrabian and Ferris (1967), words, body language, and tone of voice account for 7%, 55%, and 38% of effective communication in particular situations, respectively. It means that the content constitutes only 7% of communication and the rest is the form of context. This translates into the challenges in voice-only or text-only conversations where some of the senses are switched off. This can be especially problematic when writing text messages that contain sarcasm, which is difficult to spot without knowing the tone of voice. The problem can be mitigated with the use of emoticons (emojis) serving as contextualization cues (Skovholt et al., 2014; Garcia et al., 2022).

The other type of problem is observed when different sources of information (senses)

give us different information about a particular situation. For instance, there is a well-known McGurk effect (McGurk and MacDonald, 1976) where due to auditory and visual information conflict, the person hears a different fused sound from what was actually said. For example, if the move of the lips suggests that the sound /Ga/ is produced, the actual sound spoken is /Ba/, the sound that is often perceived by a person is actually /Da/. The issue of conflicting sources of information is also analysed in Large Language Models. The conflict is often observed between the contextual information provided in the LLM prompt and the parametric knowledge encoded in the LLM weights. Such a conflict can lead to hallucinations (Sun et al., 2025) and can be used to create successful adversarial attacks (Zhu et al., 2025).

In this dissertation, the issue of conflicting contextual information from various sources will not be widely investigated. Instead, I focus on the incorporation of contextual information for more holistic overall information.

When the main information is presented without its surrounding context, it results in what is known as being 'out of context'. According to the Cambridge Dictionary, this term refers to a situation where 'only a small separate part of what was originally said or written is reported, with the result that their meaning is not clear or is not understood'[2]. **In the dissertation, context will be analysed in a broader sense as it will not only refer to text and speech but also to visual cues.**

2.2.1 Context in Deep Learning models

Context can refer to the amount of information provided alongside a particular sample – essentially, how much background or surrounding detail is given. This is often analysed with respect to text, times series and videos. In the case of text, it is about the number of surrounding sentences/ words/ tokens, in the case of time series – the number of surrounding data points (from other timestamps or other signals referred to as temporal context), videos – number of neighboring frames. The context is handled differently depending on the data modality.

Text

Context is important when analysing text because the word can have different meanings depending on the surrounding words. For instance, the word 'bank' can refer to a financial institution or a border of a river. The other example is the sentence 'Can you pass me a bottle of water?'. Here, the verb 'can' does not refer to a person's skill of passing a bottle

[2]https://dictionary.cambridge.org/dictionary/english/out-of-context

but is a polite way of asking for a favor. Sentences with double negations or sarcasm pose another sort of challenge.

These semantic intricacies are reflected in Deep Learning advancements in the field of Natural Language Processing (NLP). Initial approaches were focused on manual extraction of features from texts. The features could be later analysed using classical Machine Learning models e.g. tree-based to predict the text sentiment. Later, the process of feature extraction was automated using e.g. Term Frequency-Inverse Document Frequency (TF-IDF) algorithm (Sparck Jones, 1972) which does not account for context. This changed with the advent of such methods as word2vec (Mikolov et al., 2013) which take into account surrounding words (context window). It is performed in two variants: Continuous Bag of Words (CBOW) and Skip gram. In the CBOW setup, the current word is predicted based on the context (few preceding and following words), whereas in the Skip-gram, surrounding words are predicted based on the current word. The training of an embedding model in such a manner allows for preservation of semantic relationships i.e. distance between the embeddings corresponding to words 'king' and 'queen' is the same as a distance between the embeddings of the words 'man' and 'woman'. The other task in NLP that requires a deep understanding of the context is machine translation. Here, the recurrent neural networks such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014) networks find application, as their built-in memory enables processing entire sentences to capture context. This, in turn, leads to translations of higher quality and correctness. The next breakthrough model architecture in the field of text processing was Transformer (Vaswani et al., 2017) which serves as a building block of Large Language Models (LLMs) (Zhao et al., 2025).

In LLMs, the term context refers to the amount of information provided in the prompt. It relates to in-context learning (Dong et al., 2024), where a user provides in a prompt one or few examples of the task to solve with the correct answers (one- or few-shot). Thanks to these demonstrations, the LLM is able to solve the tasks that it was not trained on without the need for the update of the model weights in the fine-tuning process like LoRA (Hu et al., 2022). Several methods were introduced to allow LLMs to understand long context. For instance, fine-tuning using the Focused Transformer method (Tworkowski et al., 2023) and memory with gating system (analogous to the one in LSTM) in Large Memory Models (Kang et al., 2025). LLMs can also benefit from a large reservoir of documents which is the idea behind Retrieval-Augmented Generation (Fan et al., 2024). It is one of elements of efficient context engineering (Mei et al., 2025).

Speech

The importance of context in text relates directly to its importance in speech processing as it is a recording of a text. Here, context is also about the amount of information

provided together with the main message. However, the methods of speech processing are different than in the case of text. Speech can be analysed as a sound wave or more often using the melspectrogram which can be treated as a form of an 'image' with local and global patterns.

Time series

Context also plays an important role in time series analysis. In this case, the temporal context can be in the form of past records of the signal of interest and other related time series (called channels) accounting for interactions and showing the broader aspect of the phenomenon. Information on trends and seasonality is useful when predicting values at the following timestamps. The temporal context is important e.g. when analysing disease progression and effectiveness of treatment (Wijeratne et al., 2023). The usefulness of historical values of a given asset and the additional indicators (e.g. about global market) is observed at stock exchange markets although the impact of the noise factor can be significant (Olorunnimbe and Viktor, 2023). The correlations across channels are also significant in weather forecasting, e.g. the current high humidity (one variate) can lead to a very plausible high amount of rain (another variate) in the near future.

The time series analysis can be performed using Multilayer Perceptron (MLP) networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers. CNNs primarily account for local patterns, whereas Transformers allow for processing of longer dependencies. The main drawback of the transformer-based methods, despite their state-of-the-art performance, is the high complexity of the attention mechanism. To address it, Wang et al. (2023) introduce the convolutional model enriched with the mechanism to not only account for local but also global context. The resulting hybrid model is of smaller complexity than transformer-based approaches. By combining the local features and global correlations, it is possible to capture a more holistic view of time series including fluctuations and trends.

There are channel-independent solutions such as LTSF-Linear (long-term time series forecasting) (Zeng et al., 2023), PatchTS (Nie et al., 2023) where it is assumed that values at a particular timestamp depend only on a historical data from the same channel. The alternative approach is to account for channels' interactions (channel mixing). It is used in solutions such as Informer (Zhou et al., 2021a), Conformer (Li et al., 2023), TSMixer (Chen et al., 2023), Crossformer (Zhang and Yan, 2023), TimeMixer (Wang et al., 2024a). Abdelmalak et al. (2025) show that in the case of usually considered datasets, the channel-independent solutions are sufficient. When the lookback-window is properly hyperparameter-tuned, the channel-independent PatchTS (Nie et al., 2023) model is still the state-of-the-art solution and it is observed that dropping mixing layers in the channel-dependent models (TimeMixer, TSMixer, Crossformer) does not result in

a performance drop.

Most often in modelling, the lookback-window of size 96 is used (Abdelmalak et al., 2025). However, there are attempts to incorporate longer context (Żukowska et al., 2024). Recently, the in-context learning was proposed for use in foundation models for time series data to allow for a generalization outside of the training data distribution (Faw et al., 2025; Lu et al., 2025).

Video

Temporal context in videos refers to the way models for video processing consider the relationships between frames over time to interpret actions, events, and overall video content. This is observed when solving such tasks as long-term action anticipation and spatio-temporal video grounding. In the latter, the combination of spatial and temporal contexts plays an important role.[3]

In the case of long-term action planning where the goal is to predict the next action and its duration in the video, Maté and Dimiccoli (2024) introduce two components to enrich an existing Deep Learning solution with contextual information. The first component is an end-to-end transformer-based network TCCA (Temporal Context Consistency above All), which is supposed to leverage global and local interactions. The second new element is a Bidirectional Action Context Regularizer to amplify local temporal context in parallel decoding.

In the spatio-temporal video grounding (STVG) task, the goal is to detect the object specified in the free-form textual query. The result should be a bounding box around the proper object in video frames. The fact that the solutions rely only on the textual prompt to identify the object can be limiting in the case of complex scenes with distractors or significant appearance changes in the video. Gu et al. (2024) pointed out that providing more fine-grained text in a query to increase model performance may not always be feasible. Hence, they propose another approach. They leverage contextual information inspired by the saying 'a picture is worth a thousand words' and claim that visual cues can provide richer information about the target object. As it is not allowed in the STVG task to have any external source of information apart from the video and the query, the authors propose in their solution, called context-guided STVG, to use contextual information from the video itself. The key modules of the new solution are instance context generation and instance context refinement. The first discovers the potential features of the foreground (so-called instance context), including both appearance and motion features. The second module is responsible for the spatial and temporal refinement by removing irrelevant context features. The instance context learned from one decoding stage is fed to the next

[3]Spatio-temporal context is also important in weather forecasting (Yu et al., 2025; Alves et al., 2025).

stage and used as supplementary guidance to enhance overall model performance.

Recently, in-context learning has also been applied to video generation (Zhang et al., 2025), where a model generates content-aware results guided by demonstration clips, ensuring semantic consistency. There are also attempts to enable long-context video understanding via memory consolidation (Balazevic et al., 2024) and Recurrent Memory Bridges (Wang et al., 2024c).

Image

The combination of spatial and temporal context is important not only when analysing videos but also in the case of a set of medical images taken at different moments in time over a longer period (Konwer et al., 2022; Yang et al., 2024; Karaman et al., 2024). It is crucial for analysis of disease progression.

Context also plays a significant role when processing images. In this case, the in-context learning for scene understanding (Balazevic et al., 2023) is an interesting field of research. For instance, Pariza et al. (2025) introduce a new post-pretraining adaptation method applicable to any pretrained Vision Transformer. The core of the method is a new loss function that enforces patch-level nearest neighbor consistency.

The context (in-)consistency was also an important factor in adversarial attacks and defenses.

Li et al. (2020) introduce a method to identify adversarial perturbation attacks by examining potential inconsistencies in contextual information within images. For instance, if there is a zebra crossing and a traffic sign nearby that an AI system recognizes as a traffic sign of speed limit to 100 km/h, it seems suspicious, as it is expected that there will be a stop sign in front of a crossing. Such context inconsistency may suggest that an adversarial attack was applied to the stop sign. The approach draws inspiration from extensive prior research on context-aware object recognition (Dvornik et al., 2018; Liu et al., 2018; Barnea and Ben-Shahar, 2019). In the solutions, it is assumed that the detection is performed using Faster R-CNN with Region Proposal. Such an architecture for detection first 'proposes' many regions that potentially contain objects and then classifies them. When designing the defence methods, Li et al. (2020) account for four types of relationships among the regions, all of which together form the context for each proposed region: a) regions corresponding to the same object; b) regions corresponding to other objects likely to co-exist within a scene; c) the regions likely to co-exist with the background; and d) the consistency of the regions within the holistic scene. Using the provided information, a fully connected graph is constructed where the nodes represent regions along with a super-region node that captures the overall scene context. Each node in this graph is assigned a context profile, which includes both node features (i.e., the original features used for classification) and edge features (i.e., contextual relationships). The node fea-

tures describe the region of interest (RoI), while the edge features capture how this region relates to others in the feature space.

An autoencoder is then trained on this graph to model the distribution of context profiles for each object category. During testing, the autoencoder assesses whether the predicted class of objects aligns with the test-time context profile. Specifically, if a region originally belonging to class A is perturbed adversarially and misclassified as class B, attempting to reconstruct its context profile using the autoencoder trained on class B will yield a high reconstruction error (meaning an anomaly). This discrepancy serves as an indicator that the classification result may be adversarial.

Later, Yin et al. (2022) propose the ADC method (short for Adversarial attacks against object Detection that evade Context consistency checks). The method circumvents a context-aware defense. Consider an intersection with a zebra crossing and a stop sign. If adversarial perturbations are applied only to the traffic sign, it may be misclassified as a speed limit sign. However, this inconsistency would be easily detected by the context-aware defense. Therefore, additional manipulations are required to bypass the defense mechanism. Consequently, Yin et al. (2022) postulate to perturb not only the object of interest (here a stop sign), but also other parts of the image so that the contextual consistency (object co-occurrence) is preserved. For instance, in the discussed example, it would also require the application of adversarial perturbation to the zebra crossing in a way such that it will be classified as cars. Such a modification would result in a context consistency between cars (originally zebra crossing) and the speed limit sign (originally stop sign). In such a case, the adversarial attack on the stop sign will be successful as context consistency defense will not detect it.

Moreover, the spatial context is important in human-object interaction (HOI) detection which involves the identification of ⟨human, object, interaction⟩ triplets from a given image (Kim et al., 2023). Jia et al. (2025) introduce ContextHOI where they apply spatially contrastive constraints together with semantic-guided context exploration and context aggregator to enhance model performance. This serves as a novel spatial context learning paradigm tailored for transformer-based HOI detectors.

The spatial context considered as the understanding of location of objects within the scene was investigated with regard to Vision Transformers (ViTs) that do not have spatial inductive bias unlike convolutional neural networks. To provide ViTs with information on the structure of the input (patch locations), positional encoding is used. Its role was investigated in various works (Jelassi et al., 2022; Chu et al., 2023; Lee et al., 2025). In addition, it was shown that by extending the training process with an auxiliary task (classification of position of image patches), it is possible to train ViT to implicitly add positional information to the patches themselves without directly injecting positional encodings (Zhang and Gong, 2023).

In this dissertation, the main focus is on the spatial context in images, which is described in more detail in Chapter 4. However, some inspiration and analogies were drawn from the context present and addressed in other data modalities.

The following chapters (3, 4, 5, 7, 8, 9, 10, 11) are based on my publications. The title of each such chapter starts with the letter 'P' and the index of the paper as it was specified in Section 1.3. Such chapters begin with a preface with a question and a short answer and subsequent broader context. After such a background, the text of the paper is provided.

Chapter 3

P1: Lightweight Conditional Model Extrapolation for Streaming Data under Class-Prior Shift

3.1 Preface

In Deep Learning, it is often assumed for simplicity that data is stationary. However, the deployed models often require some periodic fine-tuning to adjust to changes in data that occur over time. This procedure can be time-consuming. Moreover, there are scenarios in which there is no distinct training phase and inference phase; instead, the model is trained and used for inference continuously. The question arises: ***Is it possible to perform continuous model adaptation to the data changes in the form of class-prior shift in a lightweight manner by benefiting from temporal context?***
The short answer is 'yes'.

In the work below, I introduce the method for LIghtweight Conditional Model Extrapolation for Streaming Data under Class-Prior Shift (in short, *LIMES*) which benefits from information on temporal context. The solution is in line with the Informed Machine Learning paradigm, where external knowledge (here temporal context) is injected into Deep Learning models (von Rueden et al., 2023; Schulz and Bauckhage, 2025). In general, the prior knowledge can be incorporated through the choice of a model architecture and its inherent inductive bias such as in the case of Convolutional Neural Networks (Wang and Wu, 2023; HaoChen and Ma, 2023), Physics-Informed Neural Networks (Raissi et al., 2019) etc. Another direction is to provide prior knowledge in the form of context or human expertise as additional model input or using neurosymbolic methods. In the work, the information is incorporated as an additional component (bias term) in a fully connected layer. The aforementioned techniques are supposed to increase

model reliability, trustworthiness and performance.

Note that at the time of the project, little attention was paid to the bias weights in neural networks. In our work, we showed that their well-crafted modification can bring improvements in model performance. In parallel and later works, researchers increasingly investigated the role of the bias term in neural networks, particularly in parameter-efficient fine-tuning (Ben Zaken et al., 2022; Bu et al., 2024; Huang et al., 2025).

The proposed *LIMES* algorithm to perform model adaptation by the injection of temporal context is general and can be applied to any classifier trained in a continuous manner dealing with data undergoing class-prior shift. The use case with tweet classification is just for an illustration. The brief overview of how temporal context is addressed in Deep Learning (DL) models in the literature is provided in Section 2.2.

The paper is categorized as a study of the role of local context because the adaptation relies on the injection of a single historical data sample (temporal context) chosen in a so-called extrapolation step. However, since this data sample can be selected from the entire available history, the approach also exhibits a slightly global character. To summarize, the work is positioned along the axes defined in Chapter 1 as shown in Table 3.1.

Table 3.1: Mapping of the paper by Tomaszewska and Lampert (2022) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

The work is the result of my research internship in the Machine Learning and Computer Vision Group at the Institute of Science and Technology Austria.

Paulina Tomaszewska & Christoph H. Lampert. (2022). "Lightweight Conditional Model Extrapolation for Streaming Data under Class-Prior Shift". Proceedings of the International Conference on Pattern Recognition (ICPR).

3.2 Introduction

In our current era of connected devices, data is often generated in a continuous manner, e.g by mobile devices, intelligent sensors or surveillance equipment. This creates a challenge for machine learning systems, especially when the available computational and memory resources are limited, such as when learning *at the edge*. On the one hand, systems cannot store all prior data as the necessary amount of storage would grow linearly over the runtime of the system. On the other hand, the systems have to make predictions

continuously, interleaved with learning, rather than following the traditional *offline* setting with separate training and prediction phases. Furthermore, continuously generated data is typically non-stationary: its data distribution varies over time, e.g. due to changing environmental conditions or changing user behavior.

For a Machine Learning system, a time-varying data distribution is problematic, because it means that previously observed data might not be representative of future data. Consequently, simply learning a single fixed model, as is usually done in the standard setting of independent and identically distributed (i.i.d) data, leads to suboptimal results. Alternatively, it is possible to learn multiple models, essentially starting a new one whenever the data distribution changes. This, however, can also lead to suboptimal results when each model achieves acceptable prediction quality only after having seen enough data. Furthermore, learning multiple models leads to increased memory requirements.

In this work, we introduce *LIMES* (short for Lightweight Model Extrapolation for Streaming data) that offers a new way for learning classifiers from non-stationary streaming data in resource-constrained settings. Inspired by recent progress in *meta-learning*, *LIMES*'s core idea is to combine the resource efficiency of having only a single set of parameters with the flexibility of learning different models for different data distributions. We construct a base model that can be adapted efficiently to different target distributions together with a procedure for forecasting the parameters necessary for adapting the base model to be used on future data, even before any examples of that data have been observed.

Specifically, we adopt the scenario of multi-class classification under *class-prior shift*, in which the class frequencies change over time, while the class-conditional data densities remain approximately constant. One of our contributions is to show how in this setting it is possible to learn a model from multiple data distributions, and how make use of it for yet another data distribution, with hardly any computational or memory overhead. Furthermore, we introduce an *extrapolation* step that forecasts the class distribution of future data based on the observed previous data.

We show experimentally on exemplary tasks using Twitter data that *LIMES* achieves higher accuracy than either learning a single model or an ensemble of multiple models.

3.3 Method

In this section, we formalize the learning scenario we study and introduce the proposed *LIMES* method.

3.3.1 Learning Scenario

We assume a time-varying data distribution $p_t(x, y)$ for $t = 1, 2, \ldots$, where $x \in \mathcal{X}$ are the designated inputs to the learning system and $y \in \mathcal{Y} = \{1, \ldots, L\}$ are the output labels (Bartlett, 1992). The goal of a learning system is to create a predictive model, $f_t : \mathcal{X} \to \mathcal{Y}$, for any time step t, with as small as possible *risk*, $R_t(f) = \mathbb{E}_{(x,y)\sim p_t}[\ell(y, f_t(x))]$, where ℓ is the 0/1-loss function, $\ell(y, y') = y \neq y'$[1] . For this task, the learning system can make use of observed data from prior time steps, S^1, ..., S^{t-1}, where each $S^\tau = \{(x_1^\tau, y_1^\tau), \ldots, (x_{n_\tau}^\tau, y_{n_\tau}^\tau)\} \sim p_\tau(x, y)$, for $\tau = 1, \ldots, t-1$. In line with practical resource-constrained applications, we assume a continuously running learning system that is kept up-to-date by *periodic retraining* (Graepel et al., 2010; McMahan et al., 2013; Li et al., 2015). That means, only at the end of each time step it is possible to update the model parameters. The model obtained is then used to make predictions for all data during the next time step.

A typical example of this setting is the real-time classification of social media streams. For example, for our experimental evaluation in Section 8.6, the inputs are tweets and the outputs are the countries from which the tweets originated. The time index corresponds to time intervals over which the data distribution can be assumed relatively stable, in our case *hours*. Within a few time steps, the data distribution can change quite rapidly. Intuitively, users from different countries tweet predominantly at different times because of their respective time zones. This causes a time-varying data distribution in which class proportions fluctuate substantially over time.

3.3.2 Learning Methods

A number of possible approaches can be used to tackle the learning problem introduced above. The most classical is standard *empirical risk minimization* (Vapnik, 2013), in which a model, f, from a class of possible ones, $\mathcal{F}$, is sought that minimizes the training error across all data observed so far,

$$f_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{\tau=1}^{t} \hat{R}_\tau(f), \tag{3.1}$$

where

$$\hat{R}_\tau(f) = \sum_{(x,y)\in S^\tau} \ell(y, f(x)) \tag{3.2}$$

is the training error (empirical risk) of a model f on the data from time step τ. Optionally, a regularizer or other terms can be added to improve generalization or to make the

[1] The *Iverson* brackets P for a predicate P take the value 1 if P is true and 0 otherwise.

optimization problem better behaved.

Empirical risk minimization is widely used when the data is sampled i.i.d from a stationary distribution. However, for learning from non-stationary data, it has a number of shortcomings. First, Equation (3.1) will not actually learn a model adapted for the coming time step $t+1$, but rather one that works well on average across all data distributions in the past. Second, minimizing Equation (3.1) is intractable in a practical streaming setting. It would require having stored all training sets, $S^1, \ldots, S^t$. which is impractical for systems that are meant to run continuously over a long time with limited resources, such as *at the edge* (Murshed et al., 2021).

Alternatively, one can find a model by training only on the most recently available data,

$$f_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_t(f). \tag{3.3}$$

This *restart* approach overcomes the second problem and mitigates the first one. However, the procedure completely ignores non-recent data. Therefore, the models it produces are usually limited in the accuracy they can achieve.

Incremental training (Giraud-Carrier, 2000) avoids the problem of having to store all data, but still allows models to benefit from all available data. Each f_{t+1} is obtained by running a minimization routine on the most recent empirical risk, $\hat{R}_t$, as in Equation (3.3). However, instead of running the optimization routine until convergence, one initializes the process at the parameters of the previous model, f_t, and performs early stopping, e.g after one pass through the data. Models learned by incremental training depend on all available data. At the same time, they tend to be biased towards recently seen data. The latter can be a benefit if the data distribution changes slowly, but it can also be a disadvantage, if data from more distant time steps is valuable, such as for periodic data distributions.

A way to preserve information from more distant time step is to combine incremental training with *model ensembling* (Street and Kim, 2001; Muhlbaier and Polikar, 2007; Minku and Yao, 2011; Krawczyk et al., 2017). Instead of training a single classifier, one trains multiple prototypical models, $g_1, \ldots, g_K$, on different subsets of the available data. The desired f_{t+1} is then constructed out of this ensemble based on contextual information or, if available, data from the target distribution (Cruz et al., 2018; Almeida et al., 2018). The cost of the gained flexibility is a K-fold increase in memory requirements in order to store all models. At the same time, the amount of training data necessary to achieve satisfactory accuracy tends to be larger as well, when each model is trained only on a subset of all observed data.

Recently, the problems of *class-incremental learning* (Rebuffi et al., 2017) and *task-incremental learning* (Li and Hoiem, 2017) have attracted a lot of attention in the Machine Learning community, see e.g. (Delange et al., 2021) for a survey. These problems are

special cases of the problem of learning with non-stationary data, based on additional assumptions which data appears when and to which tasks the resulting functions will be applied. Specialized algorithms have been derived that aim at giving deep neural networks the ability to learn continually without suffering from catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). However, those recent methods are not applicable in the situation in which we are interested. On the one hand, that is because they solve different problems, and on the other hand, because they require additional resources either for storing part of the observed data (Rebuffi et al., 2017; He et al., 2020a) or growing the model architecture over time (Rusu et al., 2016; Yoon et al., 2018).

3.3.3 Classifier Adaptation

The *LIMES* method we propose combines the best aspects of the different approaches discussed before. Like in incremental training, only a single model is trained, for which all available data is used. Like in ensemble learning, the predictions are made with a model specifically adapted to the target distribution, thereby allowing for higher prediction accuracy. The core concept to achieve both properties simultaneously is a form of *classifier adaptation* (Saerens et al., 2002), that we discuss in the following.

We adopt a setting of probabilistic multi-class classification, such as implemented by logistic regression or any neural networks with softmax output layer. Such a model represent conditional distributions over the label set, $\mathcal{Y} = \{1, \dots, L\}$, in log-linear form

$$\hat{p}(y|x; W, b) \propto \exp(w_y^\top \phi(x) + b_y). \tag{3.4}$$

The model parameters are the weight vectors $W = (w_1, \dots, w_L) \in \mathbb{R}^{d \times L}$, and per-class bias terms $b = (w_1, \dots, b_L) \in \mathbb{R}^L$. The feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ can be fixed (in the case of logistic regression) or learned together with the classifier parameters (in the case of a neural network).

The proportionality constant for the right hand side of Equation (3.4) can be computed efficiently for any $x \in \mathcal{X}$ as $\frac{1}{Z(x)}$ with $Z(x) = \sum_{z \in \mathcal{Y}} \exp(w_z^\top \phi(x) + b_z)$. We generally suppress this step from our notation for the sake of conciseness.

The probabilistic model allows one to make predictions for new inputs $x \in \mathcal{X}$ by

$$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \hat{p}(y|x; W, b) = \operatorname*{argmax}_{y \in \mathcal{Y}} [w_y^\top \phi(x) + b_y]. \tag{3.5}$$

The parameters W, b (and potentially ϕ) are typically trained to make the conditional model distribution, $\hat{p}(y|x; W, b)$, as close as possible to the conditional distribution of the training data, $p(y|x)$. In the case that they are identical, Equation (3.5) yields the *Bayes-optimal* prediction rule, i.e. the classifier has minimal expected error on future data from the same distribution. However, if the data distribution at prediction time, $q(x, y)$, differs

from $p(x, y)$, then the classifier f can become suboptimal, i.e. have higher error rate than necessary. Unfortunately, the latter situation is the rule rather than the exception in the setting we study. For a time-varying data distribution, the data available at training time practically never is distributed identically to the data at prediction time.

There is no general solution for this problem of *distribution mismatch*, unless we assume that $p(x, y)$ and $q(x, y)$ are in some way related to each other. One such relatedness assumption, which is often justified in practice, is that the two distributions differ only by a *class-prior shift*, that is, $q(y) \neq p(y)$ but $q(x|y) = p(x|y)$ (Quiñonero-Candela et al., 2008). In that case, it is possible to *adapt* f to a classifier that yields optimal predictions for the new situation. To see this, we first write

$$q(y|x) = \frac{q(x|y)q(y)}{q(x)} = \frac{p(x|y)q(y)}{q(x)} = p(y|x)\frac{q(y)}{p(y)}\frac{p(x)}{q(x)}, \tag{3.6}$$

where the first and third equality are Bayes' rule and the second equality follows from the class-prior shift assumption. Consequently, the optimal decision for the distribution $q(x, y)$ is

$$\operatorname*{argmax}_{y \in \mathcal{Y}} q(y|x) = \operatorname*{argmax}_{y \in \mathcal{Y}} p(y|x)\frac{q(y)}{p(y)}, \tag{3.7}$$

where we have used the fact that the argmax is not affected by multiplication with factors that do not depend on y.

For a log-linear model parameterized as in Equation (3.4) that approximates $p(y|x)$, we obtain the *adapted model*

$$\hat{q}(y|x; W', b') \propto \exp({w'_y}^\top \phi(x) + b'_y) \tag{3.8}$$

that approximates $q(y|x)$ by setting, for all $y \in \mathcal{Y}$,

$$w'_y = w_y, \quad b'_y = b_y + \log\frac{q(y)}{p(y)}, \tag{3.9}$$

as long as $q(y)$ and $p(y)$ are known, or can at least be estimated sufficiently well. We denote the classifier associated with the adapted model (3.8) as $f_{p \to q}$.

3.3.4 Lightweight Model Extrapolation for Streaming Data

The ability to adapt a learned model from one data distribution to another has been used previously, for example for *domain adaptation* (Quiñonero-Candela et al., 2008) or to exploit changing class proportions at prediction time (Royer and Lampert, 2015).

In contrast to prior work, we use it at training time to learn a single set of parameters

that will work well – after respective adaptation – under many different conditions. Conceptually, the step resembles meta-learning (Finn et al., 2017, 2019), where also a base model is learned in a way such that it can be adapted easily to new situations. Specifically, we form a single log-linear model f_u with modeled conditional distribution

$$\hat{p}(y|x; W, b) \propto \exp(w_y^\top \phi(x) + b_y). \tag{3.10}$$

Instead of training this model to minimize the average loss across all time steps, we aim for the adapted models to result in minimal loss,

$$\min_{W,b} \quad \sum_{\tau=1}^{t} \hat{R}_\tau(f_{u \to p_\tau}), \tag{3.11}$$

where u denotes the distribution in which all classes have equal probability. Despite the similar formulation, learning a model using Equation (3.11) differs fundamentally from learning a model with Equation (3.1). For the latter, the terms, $\hat{R}_1(f), \ldots, \hat{R}_t(f)$, in the summation compete with each other: the parameters that optimally minimize one of them are not optimal for any of the others, when their respective training data is distributed differently. As a consequence, the optimization has to settle for finding parameters that perform well on average, but are never truly optimally. This effect is purely due to the time-varying nature of the data, so neither observing larger datasets nor more of them will overcome it.

In the newly proposed formulation (3.11) the terms, $\hat{R}_1(f_{u \to p_1}), \ldots, \hat{R}_t(f_{u \to p_t})$, support each other: up to finite sample effects, the parameters that minimize any of the terms $\hat{R}_\tau(f_{u \to p_\tau})$ for $\tau = 1, \ldots, t$, are identical, and they are also the same ones that would minimize the risk of the modeled f_u over a dataset sampled from u, i.e. with uniform class distribution. The reason is that the differences in class distributions are compensated by the corresponding adaptation step, so the model parameters do not have to change. As a consequence, the more datasets we observe or the larger they are, the higher the precision with which we can recover the optimal parameters for all time steps.

As in the case of empirical risk minimization, solving Equation (3.11) exactly is intractable because it would require us to store all past data. Instead, we again use incremental training: we update the parameters by always minimizing the most recent empirical risk

$$\min_{W,b} \quad \hat{R}_t(f_{u \to p_t}), \tag{3.12}$$

starting at the previous parameter values and using early stopping after a single pass through the data.

Forming $f_{u \to p_\tau}$ for any $\tau = 1, \ldots, t$, requires knowledge of the class probabilities $p_\tau(y)$.

These are not known exactly but after having observed the training set S^τ we can estimate them easily by the observed class frequencies, $\hat{p}_\tau(y) = \frac{1}{n_\tau} \sum_{i=1}^{n_\tau} y_i^\tau = y$. Those are just L-dimensional vectors, so they can be stored, even continuously, with only a very small memory footprint.

The model f_u resulting from the above procedure can be expected to work well for data with a uniform class distribution, and also for any other class distribution if we adapt it suitably. This property is particularly useful for constructing a model for the next time step, $t+1$: instead of having to train on a data set $S^{t+1} \sim p_{t+1}$, which is not available yet, we can simply perform the very efficient adaptation step (3.8) to form $f_{u \to p_{t+1}}$. This, however, requires an estimate of the class probabilities at the following time step, $p_{t+1}(y)$, which are a priori unknown. To overcome this problem, we propose a history-based *forecasting step* that extrapolates the class distribution. First, we find the time step, t^*, among all earlier ones whose class distribution is most similar to the one at the current time step, t,

$$t^* = \operatorname*{argmin}_{\tau=1,\dots,t-1} \operatorname{dist}(\hat{p}_\tau(y), \hat{p}_t(y)), \tag{3.13}$$

where dist denotes the L^1-distance. We then use $\hat{p}_{t^*+1}$ as a proxy for $\hat{p}_{t+1}$. In other words, we forecast that the situation one time step after the current, $t+1$, will be similar to the situation that had emerged one time step after a similar situation had occurred in the past, t^*+1.

The combination of both steps, adaptation at training time and forecasting of the next class distribution, we call *LIMES* (Lightweight Model Extrapolation for Streaming Data). It has a number of desirable properties. First, it is easy to implement, as it requires neither a new model architecture nor a special form of optimization. Instead, one simply trains a model continuously on the incoming data, and only adjusts the model's bias terms at each new time step. *LIMES*'s computational and memory overhead are negligible, as only the vectors of prior class distributions have to be stored and searched. For very long-running systems, the forecasting step could be made even more efficient (time and memory complexity $O(1)$ instead of $O(t)$) by either using a rolling history buffer, or by selecting a finite number of prototypical histories in a data-dependent way. Such steps are orthogonal to our main contribution in this work, so we leave them to future work.

3.4 Experimental Evaluation

We evaluate *LIMES* on the `geo-tweets` dataset[2]. It consists of a large number of tweets with geographic information that were collected using Twitter's free Streaming API[3] from

[2] `https://cvml.ist.ac.at/geo-tweets`

[3] `https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data`

Table 3.2: Experimental Results: classifier accuracy of the proposed *LIMES* method and baselines as mean and standard deviation across 10 data splits for six experimental settings (see main text for details). Reported are the average per-day accuracy (*avg-of-avg*) and the minimum-across-the-day accuracy (*avg-of-min*).

(a) features: *tweet*, subset: *early*

method	**avg-of-avg**	**avg-of-min**
LIMES	$52.47_{\pm 0.04}$	$43.15_{\pm 0.21}$
incremental	$51.94_{\pm 0.05}$	$40.51_{\pm 0.25}$
ensemble	$39.04_{\pm 0.02}$	$30.73_{\pm 0.16}$
restart	$34.36_{\pm 0.05}$	$24.62_{\pm 0.27}$

(b) features: *tweet*, subset: *late*

method	**avg-of-avg**	**avg-of-min**
LIMES	$52.28_{\pm 0.04}$	$42.90_{\pm 0.22}$
incremental	$51.63_{\pm 0.03}$	$39.90_{\pm 0.22}$
ensemble	$38.85_{\pm 0.04}$	$30.71_{\pm 0.16}$
restart	$33.89_{\pm 0.05}$	$24.21_{\pm 0.19}$

(c) features: *location*, subset: *early*

method	**avg-of-avg**	**avg-of-min**
LIMES	$81.40_{\pm 0.03}$	$75.20_{\pm 0.07}$
incremental	$81.27_{\pm 0.03}$	$74.49_{\pm 0.14}$
ensemble	$68.65_{\pm 0.05}$	$64.44_{\pm 0.13}$
restart	$54.45_{\pm 0.05}$	$44.28_{\pm 0.17}$

(d) features: *location*, subset: *late*

method	**avg-of-avg**	**avg-of-min**
LIMES	$81.35_{\pm 0.04}$	$75.07_{\pm 0.16}$
incremental	$81.16_{\pm 0.04}$	$74.24_{\pm 0.17}$
ensemble	$68.35_{\pm 0.04}$	$63.79_{\pm 0.21}$
restart	$54.07_{\pm 0.09}$	$43.34_{\pm 0.39}$

(e) features: *tweet+location*, subset: *early*

method	**avg-of-avg**	**avg-of-min**
LIMES	$85.73_{\pm 0.03}$	$79.93_{\pm 0.11}$
incremental	$85.66_{\pm 0.03}$	$79.40_{\pm 0.14}$
ensemble	$72.54_{\pm 0.04}$	$68.64_{\pm 0.17}$
restart	$58.49_{\pm 0.04}$	$48.08_{\pm 0.18}$

(f) features: *tweet+location*, subset: *late*

method	**avg-of-avg**	**avg-of-min**
LIMES	$85.67_{\pm 0.03}$	$79.75_{\pm 0.11}$
incremental	$85.55_{\pm 0.02}$	$79.15_{\pm 0.11}$
ensemble	$72.27_{\pm 0.05}$	$68.30_{\pm 0.23}$
restart	$57.86_{\pm 0.06}$	$46.54_{\pm 0.26}$

April to July 2020. The associated classification task is to predict from which of 250 countries a tweet was sent based on its textual content and/or the user metadata. Because users in different timezones are active at different times of the day, the country distribution fluctuates strongly across time. Figure 3.1 illustrates this effect: the x-axis shows the hours of a day in GMT time zone. Between hours 14 and 23 as well as 0 and 6, the most active countries on Twitter are the USA and Brazil. At their peak (hours 22-4) they together are responsible for more than half of all tweets in the dataset. In contrast, between hours 7-14, tweets from the USA and Brazil are less frequent, whereas Japan, Great Britain and the Philippines are responsible for larger shares.

To reflect the time-varying nature of the data, we split the data into hourly chunks, which on average contain approximately 127000 tweets (minimum: 61000, maximum: 149000). We use random subsets of 80% of each chunk for training and 20% for evaluation. In order to assess statistical variations, we subdivide the data further. First, we create two subsets, one containing tweets of the 1st to 5th day of each month (*'early'*) and one for the 11th to 15th (*'late'*). From each subset, we create ten realizations of the time series by subsampling the data with a stride of 10 from different starting indices.

From each tweet we extract two 512-dimensional feature vectors using the pretrained

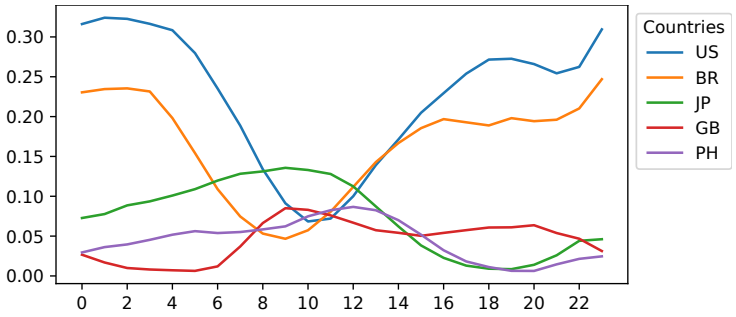


Figure 3.1: Class distribution in `geo-tweets` dataset across 24 hours for five most frequent countries (`US`: USA, `BR`: Brazil, `JP`: Japan, `GB`: Great Britain, `PH`: Philippines).

`distiluse-base-multilingual-cased-v1` multi-lingual sentence embedding network (Reimers and Gurevych, 2019). One is of the tweet text itself and one is of the text in the *location* meta-data, which is a user-provided free text that might or might not be related to the actual location of the user. As target labels, we use the tweet's country of origin, which for geo-tagged tweets can be extracted from the tweet's meta-data. Specifically, we infer the country either from the lat-long coordinates, or, if those are not provided, we use the user-specified *country_code* entry of the meta-data.

The input features can be extracted also for tweets without geo-tags. Therefore, the classifier resulting from the described training method can be used to predict the country of origin from any tweet, which is an often-required step in social media analysis (Han et al., 2014). Note, however, that in this work, we are not trying to compete with existing methods for this specific task, as the Twitter data serves just as a real-world testbed for our proposed general-purpose model adaptation technique.

As an instantiation of *LIMES*, we implement a logistic regression model in the *keras* toolkit with *tensorflow* backend. The training is performed incrementally by Equation (3.11) using the Adam optimizer with mini-batches of size 100. No regularization is used and all hyperparameters are left at their defaults (learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$).

As baseline, we use alternative methods discussed in Section 3.3.2, relying on the same model architecture and optimization routines as for *LIMES*. *Incremental* is a classifier that is trained incrementally to minimize (3.1), i.e. identically to *LIMES* but without adaptation and forecasting. *Ensemble* is an ensemble of 24 classifiers, one per hour. Each classifier is trained incrementally on the data of the respective hour. For prediction at time step $t+1$, the model with index $(t+1) \mod 24$ is used. *Restart* trains a new model at every time step t from the data S^t and uses it for predictions at step $t+1$.

All four methods are tested in 60 experimental settings: 2 subsets (*early*, *late*), 3 feature types (*tweet*, *location* or their concatenation), and 10 different data realization. In each case, we measure the accuracy of the resulting classifier. We then report average and standard deviation across the ten realization, once for the average accuracy across

all hours (*avg-of-avg*) and once for the average of the minimum accuracy within each day (*avg-of-min*). The latter value is often a quantity of interest for real-world systems, because customer satisfaction tends to depend not on the average of their experience, but rather on the worst case.

3.4.1 Results

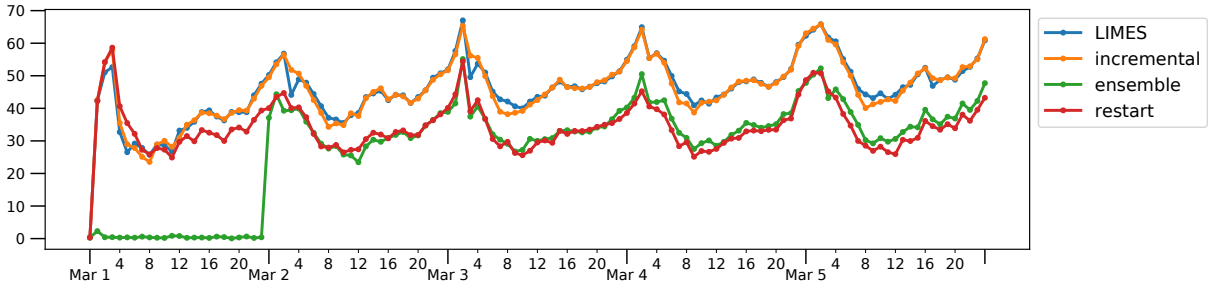
Table 3.2 summarizes the results. Overall, one can see that the task of predicting the country from the tweet text is more difficult than from the user-provided meta-data. Combining both feature types yields the highest accuracy. In all cases, the minimal accuracy across the day is substantially lower than the average. This indicates that the difficulty of the classification task varies over time, presumably due to changes in the label distribution.

Comparing the different methods, we observe a ranking that is stable across all experimental settings: *LIMES* achieves the best results, followed by *incremental*. *Ensemble* performs substantially worse, and *restart* even worse than that. Given that *incremental* differs from *LIMES* only in the lack of an adaptation and forecasting step, we can conclude that the way we propose of adapting to the class distribution indeed has a positive effect. For the *avg-of-avg accuracy* measure, the gain due to adaptation and forecasting is the biggest for the most difficult task: 0.53–0.65% when predicting the country based on the tweet text itself. It gets smaller, the easier the task becomes: 0.13–0.19% for *location* features, and 0.07–0.12% for the concatenation of both. More apparent, however, is the positive effect on the *avg-of-min accuracy*: for *tweet* features, it is 2.64–3.00%, for *location* 0.71–0.83% and for their concatenation 0.53–0.60%.

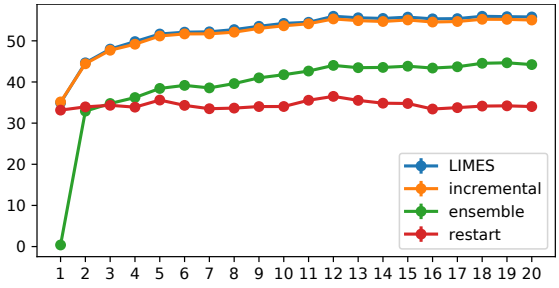
Figure 3.2 provides an illustrative explanation of these results. Subfigure (a) shows the four methods' accuracy curves over the first 120 hours of one exemplary learning run (*tweet* features, subset *early*, realization *0*). First, one can clearly see the 24-hour cycle by which the accuracies of all classifiers fluctuate. The accuracies are highest at around 2:00 GMT. As we know from Figure 3.1, this corresponds to the time the USA and Brazil are by far the most common countries from which tweets originate. The accuracy is the lowest around 9:00 GMT, which is the time when tweets come from several different countries in comparable amounts.

Second, one can also see that *ensemble*, which does not share information between different hours, starts working only after 24 hours and afterwards increases its prediction quality quite slowly. In contrast, the other models make reasonable predictions already after the first hour, because they transfer information between different hours.

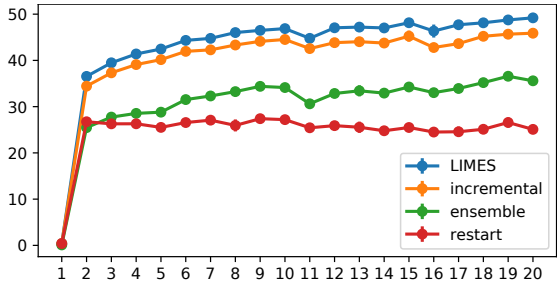
Finally, one observes that *LIMES* is superior over *incremental* mostly in the *valleys* around 9:00 GMT, i.e. during the hours where the classification problem is the hardest. Presumably, this is because *incremental* learns parameters that are adapted to the average



(a) Curve of per-hour accuracy for first 120 hours.



(b) Curve of average per-day accuracy over 20 days.



(c) Curve of minimum-across-day accuracy over 20 days.

Figure 3.2: Illustration of classifier accuracy over time for the proposed *LIMES* and baseline methods in one of the tested settings (*tweet* features, subset *early*, realization *0*).

data distribution, and that is closer to the simple situation with two dominant countries than to the difficult situations where many countries are similarly likely.

Subfigures (b) and (c) make the difference between the methods more visible by suppressing the daily fluctuations. They show the accuracy values for all 20 days aggregated as the per-day-average and per-day-minimum, respectively. Here, one can see the difference between *LIMES* and *incremental* well: it is small but consistent for the per-day-average accuracy, and bigger for the per-day-minimum accuracy.

Incremental training is clearly effective, as can be seen from the fact that the accuracy curves for the three incrementally trained methods, *LIMES*, *incremental* and *ensemble*, increase over time. In contrast, *restart*'s accuracy is approximately constant (up to the hourly fluctuations) over the runtime of the system.

3.5 Summary and Conclusions

In this work, we introduced *LIMES*, a lightweight method for continuous classifier training from streaming data with class-prior shift. Inspired by the success of recent meta-learning methods, we introduced a system that trains multiple models, each adapted to a specific class distribution, but requires only a single set of parameters, as each model is derived from a base model by analytic classifier adaptation. The adaptation parameters

are extrapolated for the next data distribution based on the data observed so far. Our experiments on the large-scale `geo-tweets` dataset show that this process results in improved prediction quality compared to common baselines, especially when judged by the worst-case measure of the lowest accuracy across a day.

Despite the promising results, the problem of continual learning from data streams with time-varying distribution is far from solved. In particular, our results suggest that new methods are required to tackle the problem that classifiers for certain data distributions are harder to learn than others. This suggests that one might have to adapt not only the model parameters, but also the model architecture. We plan to address this aspect in future work.

Chapter 4

P2: Position: Do not explain vision models without context

4.1 Preface

In the following chapters, I will introduce new methods for investigating Deep Learning models from a spatial context perspective. Before doing so, however, let us first ask the key question: ***Do existing explainable AI (XAI) methods allow for accurate model explanations when ground truth labels depend on spatial relationships?***
The short answer is 'no'. I identified the research gap which is worth further investigation.

In the paper presented below, I (1) review the most popular methods of explaining Computer Vision models by pointing out that they do not take into account contextual information, (2) show examples of failures of popular XAI methods, (3) provide examples of real-world use cases where spatial context plays a significant role such as healthcare, street surveillance, autonomous driving, (4) propose new research directions that may lead to better use of contextual information in explaining Computer Vision models, (5) argue that a change in the approach to explanations is needed (from *where* to *how*). The work provided below has been positioned along the axes defined in Chapter 1 as shown in Table 4.1. In the paper, the context is considered without restriction to any particular scale.

Table 4.1: Mapping of the paper by Tomaszewska and Biecek (2024) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska & Przemysław Biecek. (2024). "Position: Do Not Explain Vision Models Without Context". Proceedings of the International Conference on Machine Learning (ICML).

Please note that the published version of the text has been enriched below with a description of additional use cases in Section 4.4 and Machine Learning techniques in Section 4.6 and the new Section 4.5.

4.2 Introduction

The number of solutions that incorporate Deep Learning (DL) models is rapidly increasing. Recently, a significant change of the mindset in the community is observed. Many researchers agree that we should not only aim for the models with the highest performance but also the ones that behave in a responsible manner. This aspect is approached using eXplainable AI (XAI) techniques. In the work, we focus only on those for an investigation of Computer Vision models. The most attention is attributed to visual explanations. There are already many well-established post-hoc methods that generate colorful heatmaps highlighting the key regions in the input image for the model decision-making process. The major types of explanations are: permutation-based (i.e. LIME (Ribeiro et al., 2016)), gradient-based (i.e. Grad-CAM (Selvaraju et al., 2017), Integrated gradients (Sundararajan et al., 2017)) and propagation-based (i.e. Layer-wise Relevance Propagation (Binder et al., 2016), Concept Relevance Propagation (Achtibat et al., 2022)). For a systematic survey on visual XAI methods, please consult a survey (Alicioglu and Sun, 2022; Holzinger et al., 2020).

The biggest attention of the XAI community is focused on the aforementioned visual explanations. In 2019, it was assessed that the number of papers on visual explanations was about 4 times bigger than the number of papers on textual and example-based explanations of models used in medical applications. The same trend was anticipated in 2020 (van der Velden et al., 2022). We suspect that this has not changed much. The existing visual methods provide valuable insights, however, may not be enough. We identify a set of use cases in which they would fail to accurately explain the model's predictions. These are the cases where the ground truth labels depend on the spatial relationships between objects constituting images. These may be observed in critical domains, such as street surveillance systems, autonomous cars, and healthcare. **This position paper argues that we should take into account contextual information when explaining vision models – we need spatial XAI.**

From *where* to *how* The question is how to investigate whether the model takes into account spatial context. This area of model investigation seems largely unexplored.

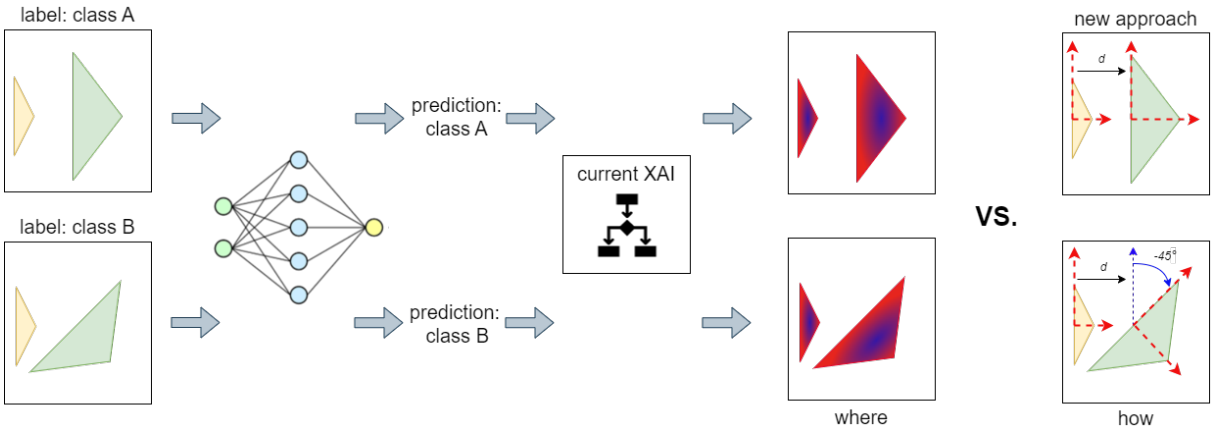


Figure 4.1: Example of two images consisting of the same objects but located differently within the scene (*angular orientation*). A DL model would correctly classify the images into separate classes. However, the common heatmap-based explanations in the two cases will highlight solely the two same triangles and would not capture the spatial relationships that are crucial factors for correct classification. Even after translating the highlighted regions within heatmaps to semantic concepts as suggested by Achtibat et al. (2022) (from *where* to *what)*, we would simply learn that in both cases there are two triangles. Therefore, there will be no difference in explanations that should be able to point out that spatial relationships are the main model's decision factor. That is why we postulate a shift of the paradigm from *where* to *how* so that the spatial relationships of how the objects are oriented towards each other will be captured within XAI methods.

Recently, Achtibat et al. (2022) proposed to change the paradigm in XAI from *where* to *what*. It means that instead of simply highlighting the regions in the input images that are key for the model's prediction, we should focus on extracting *what* semantic features within the images are important. We propose to shift the approach from *where* to *how* so that instead of only operating on the image pixel space *where* the pixels are highlighted, we should also analyze *how* the objects within images are oriented towards each other in the space (Figure 4.1). The spatial relationships may be thought of as a kind of analogy to interaction terms that are investigated in models for tabular data analysis.

As the field of XAI matures, we should focus on more complex concepts to capture the ambiguity of data and the intricate reasoning of DL models. Humans when analyzing images often take into account the composition of the scene. This information can be captured using eye trackers. The resulting saliency maps are widely studied and modeled using neural networks (Liu and Han, 2016). Therefore, investigating *how* may not only help in getting a deeper understanding of DL models, increasing safety by preventing inaccurate model predictions as we will have XAI tools to evaluate models before the launch but also in checking if there is an alignment between the way human and DL models reason about the scenes.

The structure of the work is as follows. First, we define the types of contextual information within images and provide illustrative examples (Section 4.3). Next, we elaborate

on the domains when spatial context matters (Section 4.4). Then, we provide a summary of works on DL models where the spatial context plays a central role (Section 4.6). We provide examples of failures of popular XAI techniques in Section 4.7. Later, we contrast the number of solutions addressing the topic of spatial context developed in the modelling community with the number of papers in the topic of spatial XAI (Section 4.8). Lastly, we outline the necessary research directions (Section 4.9) and draw conclusions (Section 4.11).

4.3 Contextual Information within Images

The importance of context within input data is vastly studied in Time Series and Natural Language Processing (NLP), yet much less explored in Computer Vision.

The key elements of scene compositions when analyzing images using DL models were defined by Galleguillos and Belongie (2010): *semantic context*, *spatial context* and *scale context*. This nomenclature was built upon the taxonomy proposed in a psychological study (Biederman et al., 1982).

4.3.1 Semantic context

It refers to the situation when the co-presence of the objects in an image is helpful when performing a task i.e. classification.

Likelihood

Here, the fact that objects are within the same space (regardless of spatial relations between them) is an important clue. Suppose there is an image of an empty room with one chair and the task is to classify the room. It seems that it can be a doctor's room, a classroom or a stage in a theatre etc. Without any contextual information, that would be just a pure guess. However, if there is a blackboard in the image, it becomes clear that most probably it is a classroom. In this case, the information on the co-presence of other objects in the scene makes a classification task easier.

Compound

We would like to add another category of semantic context that was not pointed out by Galleguillos and Belongie (2010): *compound*. There are objects i.e. rollerblades that are a compound of other self-existing objects (here, shoes and wheels). The information on whether there are wheels attached to shoes is crucial to classify the object correctly – shoes without wheels are not rollerblades anymore. Let us consider the possible outcome of the existing XAI methods when explaining the rollerblades prediction. When a model

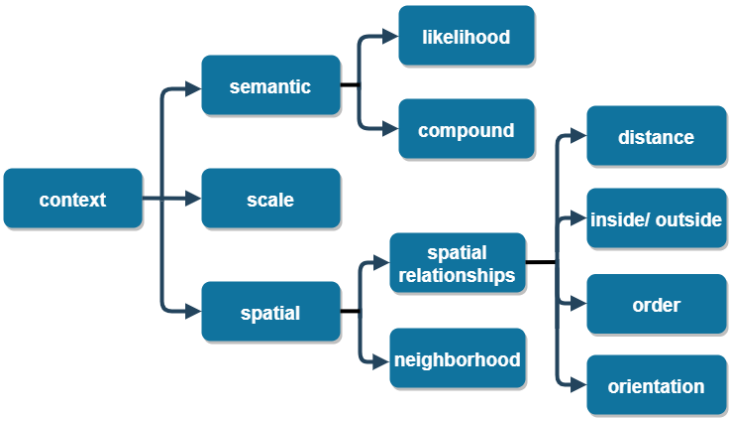


Figure 4.2: Taxonomy of contextual information within images. It is an extended version of the one proposed by Galleguillos and Belongie (2010).

is only supposed to classify shoes vs. rollerblades then it may be sufficient for the model to focus only on the detection of wheels and therefore this is the part that most probably will be highlighted within the explanation heatmap. However, if there is a classification of rollerblades vs. cats then we would rather expect to see in the heatmap both the shoes and wheels highlighted. Therefore, we may face some relativity of explanations.



(a) physician



(b) flowers in a vase



(c) train entering tunnel



(d) man giving roses



(e) patient



(f) flowers near vase



(g) train leaving tunnel



(h) man leaving woman

Figure 4.3: Examples of images where the ground truth labels depend on spatial relationships between objects: *distance* (a, e), *inside/outside* (b, f), *order* (c, g), *orientation* (d,h). The images were created with the assistance of DALL-E 3.

4.3.2 Spatial context

In this type of context, we would like to distinguish two categories: *spatial relationships* and *neighborhood*.

Spatial relationships

In this case, ground truth labels depend on the spatial relationships between objects i.e. *distance*, *inside/outside*, *order*, *orientation* (some illustrations are in Figure 4.3). We provide below illustrative real-world examples.

Distance The fact that someone is standing within some distance from a shovel does not make this person a construction worker. The fact the man is holding a shovel increases the chances of him being a construction worker. Therefore, the distance is a key. It applies to different occupations and their attributes. Consider that someone is wearing a stethoscope (most probably a physician) and the other case when someone is touched only by a small part of the stethoscope during an examination (patient) (see Figure 4.3a, 4.3e). Another example is already said rollerblades – not only there should be shoes and wheels in the picture but they should be properly attached. Similarly, an image can be classified as a bump if the cars are in contact otherwise if there is a significant distance between them, it is not a bump. This pairwise relationship can be extended to many objects i.e. when there is an image of some people spread far away from each other, we cannot classify the scene as a crowd unless people stand close to each other.

The distance can have also a more physics-informed aspect. Consider there is an image of a squared table where all the legs are attached to the same side of the table. Such a table would collapse even though it has the same elements as the typical table (table top and four legs). Note that even if the model will correctly distinguish that such a sample is out-of-distribution of the real-world tables, we would not be able to correctly explain the model's decision using current XAI methods.

Inside/ outside Suppose there are flowers in a vase. This is true only if the flowers are inside the vase, the fact that the two objects are close to each other is not enough (Figure 4.3b, 4.3f).

Order Suppose there is a train near a tunnel. If the order of the objects is (1) railway engine, (2) carriage, (3) tunnel, it means that a train leaves the tunnel but if it is (1) carriage, (2) railway engine, (3) tunnel, then a train enters the tunnel (Figure 4.3c, 4.3g). Also, when we think of a hamburger, we expect a piece of meat between two pieces of bread. If there are two pieces of bread and a piece of meat on top, this is not a proper hamburger.

Orientation There are two cases regarding orientation: an angular position between objects (when we can assign some coordinate system to each object) and the situation when objects are heading or standing back towards each other. The example for the latter is a man giving roses to a woman (they face each other) vs. a man holding roses and standing back to a woman (Figure 4.3d, 4.3h). It may seem a subtle case of *order* relationship because it can be translated to the question of whether the roses are between the two people or not, but we decide to analyse orientation as a separate subcategory. The other more severe case is whether a person with a knife/gun is heading towards the other person or is going in the other direction. In the first case, there is a high chance that the image depicts 'violence', whereas in the latter – it is more probable that there is 'no violence' in this particular frame.

Neighborhood

The spatial context can be also understood in a manner more similar to the way it is treated in NLP where it is analysed how many words surrounding the key part of the text we should input to the system. We would like to distinguish neighborhood as a separate subcategory of spatial context which can be thought of as *co-presence* but within a strictly constrained location (spatial aspect).

4.3.3 Scale context

It is difficult to assess the size of a given object if there is no reference object of a known size in the image. This may impact the correctness of a classification. Suppose a photo of a silver hoop – without contextual information, it can be either a ring or a bracelet.

In the following sections, we focus mainly on spatial context.

4.4 Application domains

Spatial context plays a significant role in many domains.

Street surveillance systems The videos from the streets and parks should be analysed in an automated way to detect danger and alarm the police. Automation is crucial as it would be difficult for people in municipal surveillance centres to analyse all of the recordings carefully in real time due to the abundance of data. The behavior of AI systems used for such an application should be trustworthy and safe (in compliance with regulations such as the EU AI Act and the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence signed by President J. Biden). When analysing situations on streets, the case with a human holding a knife/gun in a

relative position to another human is the scene that may occur. Therefore, the assessment of how severe the situations are, may depend on the orientation of the people towards each other.

Autonomous cars Safety is one of the biggest concerns in autonomous cars and so the same law regulations as in the case of surveillance systems apply. Suppose an autonomous car is approaching a crossing where a person is waiting to cross the street but there are no traffic signals. The vision system in an autonomous car identifies a human and a stick somewhere in the scene. Just based on this information, it is difficult to say whether the person is elderly and is holding a stick to facilitate walking or if it is elderly and a visible stick is just a part of a street sign or street lights. In the former, the car should predict that the needed time for the pedestrian to cross the street will be longer than in the latter and therefore, adjust its speed accordingly. Such adaptive cruise control can already benefit from the information about the surrounding cars to set the appropriate speed (Wei et al., 2020) and so should be done also in the case of approaching pedestrians.

Another example can be the case of a stop sign. Normally, it is placed in front of a junction or railways to inform the driver that he should stop and check whether it is safe for him to cross the junction. However, there is also another use case when the stop sign can be used but will have a different meaning. It is in the case of roadworks when instead of two lanes enabling the ride in two directions, only one lane is available. In such cases, it sometimes happens that there is a construction worker holding a stop sign which in this case means that the driver should stop as the cars from the other direction are driving and once the stop sign is put down, then a driver can start a ride.

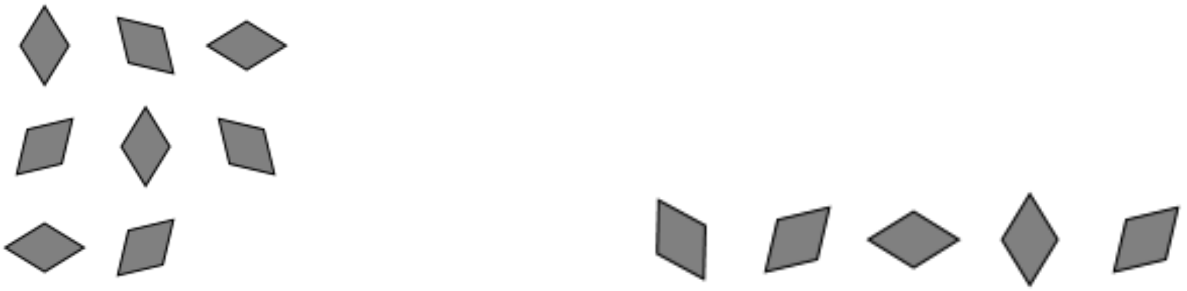
Robotics For robots working in factories and, in the future, those accompanying humans in daily life (such as cobots or robot assistants), understanding spatial relationships (particularly distance) is crucial. For instance, if a robot is asked to take an object from a conveyor belt and put it to the box on the right, the understanding of the inside/outside relationship is important. Another use case is collaborative dressing.

Healthcare From the clinical point of view, it is important to check if lesions are clustered (primary) or widespread (with metastases) as it directly translates to the assessment of the stage of the disease and impacts the choice of treatment methods. For instance, in histopathology, the data is saved in the form of Whole Slide Images (WSIs) which are of huge resolution showing the whole cuff-out lesion at the level of individual cells. Therefore, the processing of them using Deep Learning can be troublesome. There are two possible approaches. First, analysis of only separate patches taken out of the whole WSI (if local labels are available), which leads to a loss of broader contextual information. Second, to analyse the whole lesion in one go to preserve spatial context (tissue-level per-

spective) (Lu et al., 2021a; Shao et al., 2021; Lerousseau et al., 2021; Chen et al., 2022; Rymarczyk et al., 2023).

Land analysis (agriculture, environmental studies, archaeology) The satellite images are similar to the histopathological ones as they are often also of a very high resolution and so context may play an important role. The satellite images can be used to assess the parameters of the soil, detect methane leakage and identify some historic ruins hidden in the soil. In these cases, the neighborhood context may be important during analysis with the help of DL models (Liang et al., 2022; Cheng et al., 2022).

IQ tests Spatial context is important also in the case of synthetic data such as visual IQ tests in the form of Raven Progressive Matrices (RPM) (Raven and Court, 1998) where a test-taker is given a 3x3 grid of panels with one field missing (Figure 4.4a) and the task is to choose the correct panel (matching the underlying relationship in the matrix) from the set of given panels (Figure 4.4b). The relationships within the images in the panels can occur within rows, columns and/or diagonals of the matrix. An example of the underlying rule within the RPM can be that there is a diamond shape in each panel that is rotated by the same angle from one panel to another one in each row but the starting orientations of the diamonds in the first column are different. Therefore, the RPM has to be analysed as a whole, paying attention to spatial relationships between the objects in panels. For throughout survey on the application of Deep learning to abstract reasoning tasks (including visual IQ tests), consult the work by Małkiński and Mańdziuk (2023).



(a) input RPM (b) set of possible answers

Figure 4.4: Visual IQ test as an example of the task where spatial relationship understanding is required to solve it properly by choosing one answer from the set of given (b) to fill in the input RPM (a).

4.5 Deep Learning model vs. spatial reasoning – datasets

The visual reasoning is evaluated on various datasets – natural scenes and synthetic ones. The synthetic datasets (e.g. so-called visual IQ tests in a form of Raven Progressive

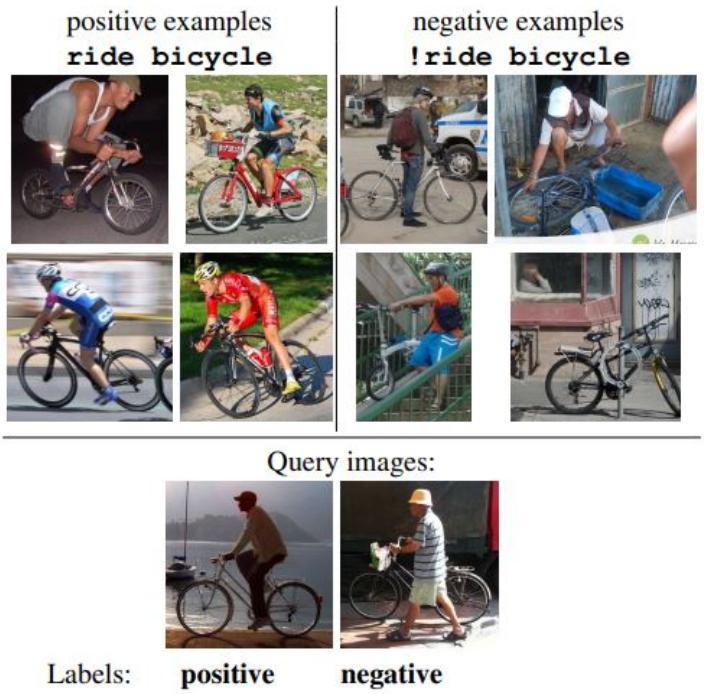


Figure 4.5: Samples from the Bongard-HOI dataset. Source: Jiang et al. (2022).

Matrices (Raven and Court, 1998) and Abstract Reasoning Corpus (Chollet, 2019)) can be used to assess abstract reasoning skills. In addition, there is the Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2020) which is a collection of various datasets used in a low data regime. The datasets in VTAB are grouped into three sets: natural, specialized, and structured. Via the structured set, it is assessed if a model comprehends a structure of a scene. Most of the tasks are generated from simulated environments. In this category, there are such datasets like CLEVR (Johnson et al., 2017) (with two tasks: object counting and depth prediction with 3D shapes located in different configurations in a space), dSprites (Matthey et al., 2017) (simple black/white shapes, with two tasks: location and orientation prediction), SmallNORB (LeCun et al., 2004) (objects viewed under varying conditions, with two tasks: object azimuth and camera-elevation prediction), DMLab (Beattie et al., 2016) (frames from a rendered 3D maze, task involves predicting the time for a pre-trained RL agent to navigate to an object), KITTI (Geiger et al., 2012) (frames captured from a car driver's perspective, task: predict distance to the nearest vehicle).

The other interesting dataset is Bongard-HOI (Jiang et al., 2022). For instance, there are images with caption 'riding a bike', showing human sitting on a bike (positive examples) and negative examples with human standing next to a bike. Therefore, the spatial relationship between a human and a bike is important (Figure 4.5).

Lastly, there is a dataset called Visual Genome (Krishna et al., 2017) where images are accompanied with graphs showing the relationships between objects, for example 'sunglasses laying **on** counter'.

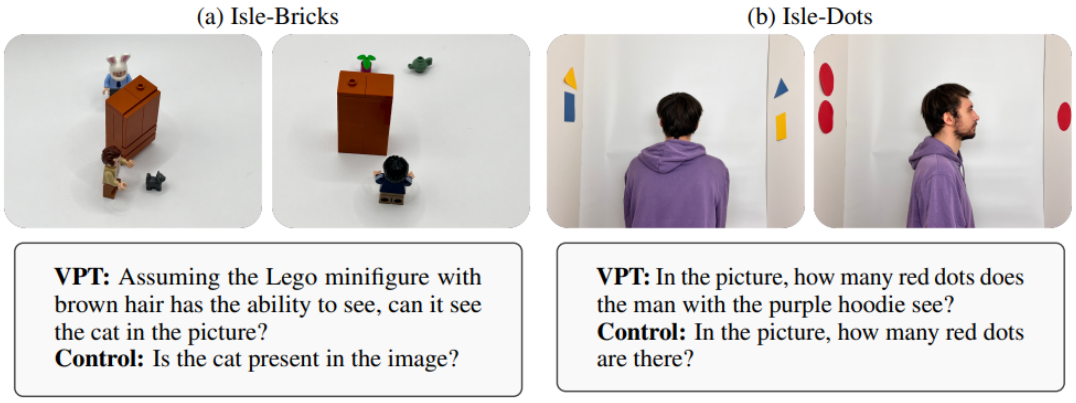
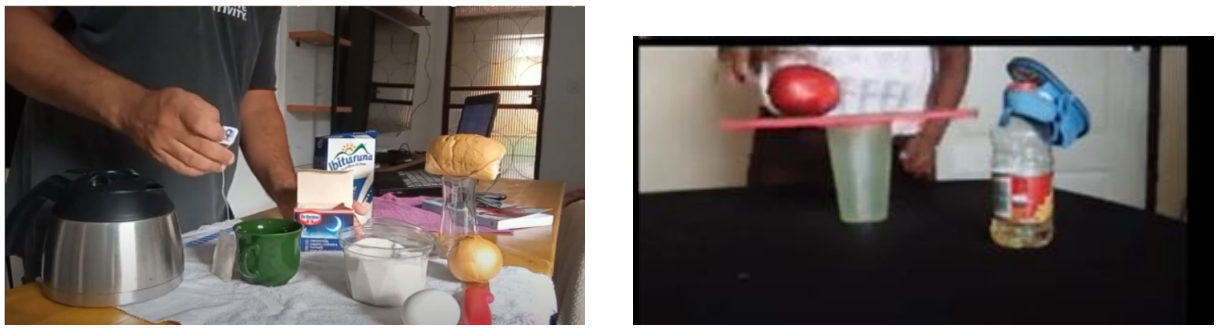


Figure 4.6: Samples from Isle-Bricks and Isle-Dots datasets for evaluation of the capability of visual perspective-taking. Source: Góral et al. (2024).

The understanding of spatial relationships by Vision Language Models (VLMs) is evaluated in the task of visual perspective taking. This task is motivated by the importance of understanding another person's perspective, which allows individuals to predict their actions. For example, a driver can prevent accidents by considering what pedestrians are able to see. To inspect the VLMs, the two datasets (Isle-Bricks and Isle-Dots) together with the evaluation protocol were released (Góral et al., 2024). The samples from the datasets are shown in Figure 4.6. The results of experiments comparing popular VLMs such as GPT-4-Turbo, GPT-4o, Llama-3.2-11B-Vision-Instruct in their ability of scene understanding, spatial reasoning and visual perspective-taking are provided in the paper by Góral et al. (2025).

A further challenge for Vision Language Models in understanding spatial context is posed by the Perception Test (Pătrăucean et al., 2023) dataset, which consists of short videos accompanied by multiple-choice questions that the model is required to answer (videoQA). The aim is to evaluate models in four areas: memory, abstraction, physics/geometry, semantics. Within each category multiple skills are investigated. For instance, in the physics category there is the spatial relationships subcategory. However, the questions from other categories also sometimes require the understanding of spatial context. The first example is a video (Figure 4.7a) where the person is pouring water from a kettle into a cup and then is putting a tea bag not inside the cup but next to it. The question is 'What action or actions did the person fail to complete and why?'. In this case, the model has to understand that in order to make tea, it is crucial to put a tea bag in the right place (the spatial relationship: inside/outside). Another example is a video where a person is writing a word 'BIKE' on the sheet of paper and the question is 'What would be the order of the letters if a person would have written them in a reverse order?'. This question is assigned by the authors of the dataset as the one from the category: Physics/Geometry, subcategory: Spatial relationships. There is also a follow-up

video which starts at the end of the aforementioned video. Here, a person is placing the shapes of different letters on the table next to the paper with the word 'BIKE' and the question is whether the letters put on the table match the ones in the paper and if so, if the order is the same. The last example is a video (Figure 4.7b) where there is a cup on a table where a person is later placing a book on it and lastly a mango on top of everything to build a tower. The model is asked to predict if the construction will be stable. For humans, it is clear that it depends of whether the fruit will be put on the central axis of all of the object. If all of the objects are coaxial, the tower should be stable.



(a) inside/outside (b) relative positions (coaxial)

Figure 4.7: Illustrative frames from the Perception Test dataset. The subcaptions specify the spatial relationship that must be understood in order to answer correctly the question to video from which the frame is taken. Source: Pătrăucean et al. (2023).

4.6 Spatial context in DL models

Spatial context can be valuable when performing Machine Learning and Deep Learning tasks.

There are the fields of Spatial Statistics and Spatial Data Science where the neighbourhood of instances plays significant role and is vastly used in the analysis often following the first law of geography saying that 'everything is related to everything else, but near things are more related than distant things'. For example, the Reed–Xiaoli Detector (RXD), a widely used anomaly detection algorithm in hyperspectral image analysis, identifies pixels that deviate significantly from the background by computing the Mahalanobis distance between each pixel and the global background distribution. Similar method was proposed by Kondmann et al. (2021) for the analysis of anomalies in time in medium- to high-resolution optical satellite images. In the method, a value of a given image pixel at time t is modelled as a linear combination of its neighbors. This way the weighting coefficients corresponding to neighbours are trained. At time $t+1$, the trained coefficients and the new values of the neighbours are used to calculate the new value at the same pixel. If the calculated value is different from the true value in a large extent, then most probably this is an anomaly. However, since this work focuses on Deep Learning models,

classical methods will not be further discussed in detail.

Belowe, I provide the analysis of whether the information on spatial aspects is concerned within the design (architectures) and training schemes of Deep Learning models.

Model architecture families In the early days of neural networks, the images were flattened before being passed to multilayer perceptrons, which led to an irrevocable loss of spatial aspect within the data. In order to overcome this downside, convolutional neural networks were proposed that analyse the images as matrices. Such a design of convolutional layers was supposed to capture local patterns. Later, a new architecture (Vision Transformer) was introduced with a self-attention mechanism as a main part. The self-attention is supposed to retrieve (even the long-range) relationships between the input patches. Additionally, the model is fed with patch positional encodings.

Within this timeline of consistent progress in the field, two solutions outside of the main track deserve attention: CapsuleNet (Sabour et al., 2017), MLP-Mixer (Tolstikhin et al., 2021). The motivation behind the idea of CapsuleNets was to propose an architecture that will take into account the spatial relationships between objects in the image as opposed to convolutional neural networks (CNNs) that were claimed to pay attention only to different elements without spatial context due to continuous downsampling operations (Pawan and Rajan, 2022). Such a by-design behaviour of CNNs was perceived as a drawback. The illustrative example is provided in Figure 4.8.

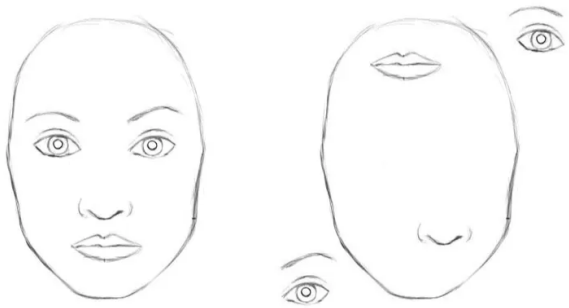


Figure 4.8: In the two images, there are the same elements so a CNN most probably will classify them as a face in both cases unlike a CapsuleNet that will not classify a *deformed face* as a *face* as spatial relationships between objects are not properly preserved. Source: Pechyonkin (2017).

Another worth-mentioning solution that went off the beaten track was MLP-Mixer. It was introduced before the Vision Transformer. The authors showed in one of the experiments, that MLP-Mixer is invariant to the order of input patches and pixels within the patches. The performance of the MLP-Mixer was the same on the said modified images as on the original ones. In the case of ResNet, the performance on the images where the spatial relationships were corrupted in the said manner, dropped significantly.

It is claimed that it is due to CNNs' strong inductive bias.

Context-oriented solutions Some solutions were designed to benefit directly from contextual information within images. The context is expected to help in emphasizing the required features while suppressing the undesired variation which is especially important when there are various orientations and scales of the same objects (Cheng et al., 2022).

The topic was analysed for natural scene images (mostly in a detection task). The early approaches are summarized in a survey by Galleguillos and Belongie (2010). The vast number of works incorporated the concept of neighborhood into the pipeline. One of those is spatial-context-aware deep neural network (Zhang et al., 2022a) used for multi-label classification when bounding boxes are provided. In the solution, there are two bounding box generation branches – the object and context ones. The first one localizes precisely the objects (tight bounding boxes) whereas the second expands the bounding boxes by incorporating the neighborhood around the object. Later, the features are extracted from the two branches and combined to perform a final classification benefitting from contextual information.

Another approach to address the topic of spatial context in models is to use Feature Pyramid Networks (Lin et al., 2017) which process the images at different scales to gain a better understanding of global and local context.
Domain-specific solutions:

- Histopathological data: The concept of incorporating information from various resolutions was a central part of several works (Zormpas-Petridis et al., 2019; Marini et al., 2021; Chen et al., 2022). Another concept was to use information from neighboring patches of the one that was under analysis at a particular moment. Zhang et al. (2020) propose two 2D-LSTM spatial context branches on top of a CNN feature extractor to obtain continuous and discrete spatial context-dependent information.

- Remote sensing data: Two main challenges are identified when trying to benefit from global and local context in land cover analysis: the ambiguity of global context and lack of efficient context combination strategy (Cheng et al., 2022). To mitigate these limitations, Cheng et al. (2022) propose a mechanism of a fusion of global and local contexts under the guidance of uncertainty. The access to the information on neighborhood pixels is provided to the system in work by Liang et al. (2022).

Pretraining The spatial aspect was also incorporated into model pretraining schemes. One class of such approaches are the ones dedicated to inpainting where the model is asked to fill up the missing parts of the input image. In such cases, the generation of missing image parts is conditioned on its surroundings (Pathak et al., 2016; He et al.,

2022). The resulting context encoder can be later used for various downstream tasks as they capture the semantics of visual structures.

The pretraining scheme where the idea of spatial context was incorporated in a more explicit form is described by Doersch et al. (2015). The model is given two neighboring patches of the image and is asked to output the spatial configuration of them out of 8 possibilities. There is a grid 3x3 where the first image is set and the position of the second within the grid is the target.

Context relationships as output In all of the aforementioned works except the one by Doersch et al. (2015), the understanding of spatial context by DL models is not the goal itself, it is a means towards better performance in the main task. In the work of Swingler and Bath (2020), the prediction of spatial context is the centre of the experiments. In the first experiment, the model is given an image of two cloth pieces and it is supposed to classify each object and classify what type of spatial relationship is present (*left, above, below left, and below right*). In the second experiment, one sample input is an image containing three objects and one-hot embeddings of two objects. The task is to output the spatial relation between the two objects mentioned within the input. The focus on spatial relationships within the output is also in generative models when relative positions of the objects are specified in the prompts (Wang et al., 2024d).

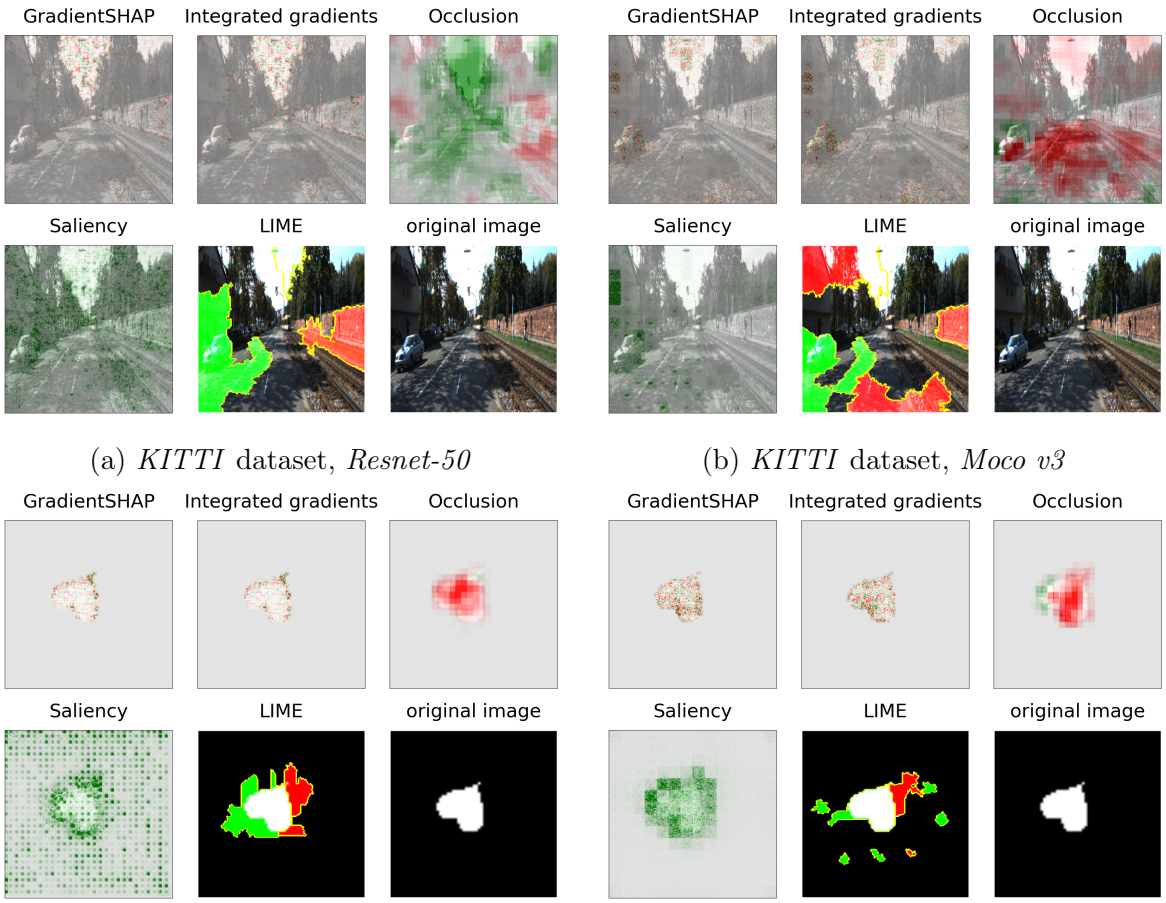
For another view on the topic of context understanding in computer vision consult a survey (Wang and Zhu, 2023).

4.7 XAI Failures

We provide examples when popular XAI techniques fail to explain in a comprehensive way the predictions of the model when spatial context is crucial. We fine-tuned *Resnet-50* and *ViT base* models in low data regime as it is done in work by Jia et al. (2022) using datasets from Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2020). The models were pretrained on Imagenet: *Resnet-50* – in a supervised manner, *ViT base* – in a contrastive manner (*Moco v3*). We focused on the subset of VTAB called 'structured' where the labels depend on spatial context. In the experiments, we use two datasets: *KITTI* (Geiger et al., 2012) where images were collected using sensors in the car – the task is to predict the binned distance to the closest vehicle in the scene, *dsprites* (Matthey et al., 2017) where images of simple shapes undergo rotations and other shifts in the space – the task is to predict binned orientation. Hence, we analyze the only real-life dataset and one of the few synthetic datasets in VTAB.

Having fine-tuned models of a satisfactory performance (similar to the one claimed by Jia et al. (2022)), we applied 5 popular XAI techniques: GradientSHAP (Lundberg and Lee, 2017), Integrated gradients (Sundararajan et al., 2017), Occlusion (Zeiler and

Fergus, 2014), Saliency (Simonyan et al., 2014), LIME (Ribeiro et al., 2016) to check if they manage to explain the model's decisions correctly. In Figure 4.9, we provide images with the models' explanations. The colors reflect the extent to which particular parts contribute to the final model's decision (green – means positive attribution, whereas, red – negative). We can see that the existing popular XAI methods fail to explain the correct model decisions i.e. that the nearest vehicle is within the distance of 8 to 20 meters.



(a) *KITTI* dataset, *Resnet-50*

(b) *KITTI* dataset, *Moco v3*

(c) *dsprites* dataset, *Resnet-50*

(d) *dspites* dataset, *Moco v3*

Figure 4.9: The explanations of correct predictions of *Resnet–50* and *Moco v3* on the samples from the *KITTI* and *dsprites* datasets. In the first case, the task is to predict a binned distance to the nearest vehicle, whereas, in the second, to predict a binned orientation of a shape. Five popular XAI methods were used for the study. The color spectrum from red to green depicts the extent to which a particular image part contributed to the model's prediction (from negative to positive impact).

The explanations fail – they are inaccurate, vague and difficult to interpret (Figure 4.9). It can be seen that the nearest vehicle is highlighted in green (using the LIME method), however, it is not clear that the distance is crucial. One may conclude that the same explanations could also be seen in the object recognition tasks i.e. cars vs. planes. Therefore, it may be important to design spatial XAI techniques that would be tailored to different ML tasks or context types.

In the XAI field, the visual explanations are the most popular. However, it seems that they may not be sufficient to properly highlight the role of spatial relationships in predictions. Hence, some other approaches are needed.

4.8 Spatial context in XAI

The topic of spatial context was considered at the stage of a training of neural networks but it was of much smaller interest in XAI. We showed that there is still plenty to do as popular XAI methods fail. Note that without proper XAI tools, we would not be able to explain the predictions of CNNs and CapsuleNets when the same features are within input images but placed differently in the scene. The existing methods in the case of 'deformed face' (Figure 4.8) will most probably highlight the same key elements of the image both in the case of CNNs and CapsuleNets even though the two DL models will return different predictions.

The investigation of whether the spatial context plays a role and to what extent in the decision-making of DL models is a niche that should be fulfilled. We call this area of research *spatial XAI*. There are only a few works in this domain. As a result of the analyses, we distinguish four main approaches: *intrinsically explainable models*, *methods introducing measures of spatial context*, *leveraging current XAI methods* and the ones *based on input-output relationships*.

Intrinsically explainable models The work by Bardool et al. (2019) encompasses two aspects as it proposes a new explainable-by-design architecture for detection and the metrics to evaluate the importance of context. The authors modify the well-known detector architecture adding a contextual learner to extract contextual cues from the scenes. The contextual feature maps for each class give insights into the way the model benefits from context. The other contribution is the two scoring functions to measure the contextual relevance of detections in relation to other objects present in the image and the general scene. The first score is calculated as a summation of the elements of the feature map corresponding to a bounding box within the mask region, divided by the area of the bounding box's mask. The second scoring method is analogous but it is calculated based on contextual feature maps generated for each class instead of bounding boxes. The metrics are later used for the comparison of different DL models. The authors challenged the proposed DL model on a set of hypotheses referring to spatial context. It was possible thanks to the fact that they had control over the generation of a synthetic dataset so that you could prepare samples violating the underlying spatial rules in the dataset and see how the model behaves in such scenarios.

Measures of spatial context Sacha et al. (2024) analyse the spatial aspect within explainable-by-design prototypical models. They define the problem of *'spatial explanation misalignment'* and introduce the metric to quantify this phenomenon. Moreover, they propose a benchmark and a method to mitigate this spatial misalignment in this particular kind of models. It was observed that sometimes the location of the explanation within the input image changes when the input image is modified in regions that should not influence the model prediction (such image modifications are not in meaningful areas and can be thought of as distractors).

Tomaszewska et al. (2025b) focus on metrics to evaluate the preservation of spatial context within DL models. The authors proposed three spatial context measures ($SCM_{features}$, $SCM_{targets}$, $SCM_{residuals}$). The quantitative measurement is possible thanks to the incorporation of spatial regression models which are run on patch features to predict the respective attention scores assigned by the DL models. The method is general and can be applied to any attention-based vision model that most often finds application in the analysis of high-resolution images. In the work, the method is illustrated in the histopathology use case where the spatial context in the input is understood as spatial autocorrelation between patches of Whole Slide Images.

Leveraging current XAI methods One of the early XAI methods is LIME (Ribeiro et al., 2016) which operates on superpixels from input images assigning them the importance score explaining the model's decision. In work by Rabold et al. (2020), the authors propose to build upon the LIME method by incorporating logic rules obtained by the Inductive Logic Programming system, Aleph. The authors, therefore, propose to combine visual and symbolic explanations to explain spatial relationships.

Input-output relationships In Deep Learning, often ablation studies are performed where the particular parts of the networks are switched off to check whether they are necessary and what their role is. This evaluation is often performed by comparing the performance of the original network and the modified one. In XAI, a similar technique is applied to images that undergo some modifications. This is done in so-called permutation-based XAI methods where parts of the input image are erased to see how it impacts predictions. In work by Tomaszewska et al. (2023), an analogous approach was used to evaluate the importance of neighborhood. The model for prediction of whether the patch from a histopathological image contains tumor or not was trained on images containing the region of interest and neighborhood. The ground truth label is provided only based on the region of interest. Later, the inference was performed with images with a gradually decreased size of the neighborhood. It was analysed how much the initial performance metrics were impacted by the limitation of contextual information. In the case of a so-defined neighborbood, a simple application of the existing XAI techniques could show

whether models pay attention to the neighborhood or not.

Another input-output relationship approach accompanied with a new XAI method is proposed by Haldekar et al., 2017. The authors cover some parts of the input image (using a sliding window) and later compute the influence it has on a value of cross entropy compared to the situation when no modifications to the input image are made. As a result, the heatmaps showing key parts of the images are generated. The authors present some heatmaps visualizing the relationships (*below*, *behind* and *above*). However, the limitation of the provided heatmaps is the fact that without the authors' captions, it would be difficult to decipher what type of relationships are highlighted. Moreover, the authors analyse by zeroing out whether there are specific groups of neurons in MLP layers responsible for detecting particular spatial relationships.

Analysis of spatial relation understanding is performed in text-image generation models where it is assessed if the spatial relations specified in the prompt are preserved within generated images (Cho et al., 2022).

4.9 Needed Research Directions

In some of the mentioned papers on spatial context (Haldekar et al., 2017; Bardool et al., 2019; Swingler and Bath, 2020; Rabold et al., 2020), the claims were investigated on synthetically generated datasets in order to have better control over the experimental setup which facilitates the verification of the research hypotheses. These datasets were not made publicly available and, moreover, were not of sophisticated compositionality as there were only some simple objects in relative relationships put on the plain background. With the advent of diffusion models i.e. stable diffusion (Rombach et al., 2021), it is possible to generate more realistic images with the desired spatial relationships between objects (samples in Figure 4.3). Therefore, we postulate for new datasets reflecting scenarios where ground truth depends on spatial relationships.

Assessment of spatial reasoning skills in text-to-image models PAINTSKILLS (Cho et al., 2022) is a compositional diagnostic evaluation dataset where images are generated using prompts containing information on spatial relationships (*above, below, left, right*) between objects. Note that these relationships are relatively simple compared to the ones we outlined in Section 4.3.2. The dataset is used as one of the scenarios verifying reasoning skills within the recently proposed benchmark, Holistic Evaluation of Text-to-Image Models (HEIM) (Lee et al., 2023). Within the reasoning scenario, it is checked whether models understand objects, counts, and spatial relations (compositionality) within images. Some task-specific benchmarks dedicated to the evaluation of the vision-language model's ability to understand scene composition were proposed (Ma et al., 2022; Ray et al., 2023). It seems that the understanding of compositionality within text-

to-image solutions is assessed only at a high level, however, in XAI the focus is on a deeper analysis of a model decision-making process. Here, we could conclude that the recent approaches are more about checking the accuracy of the model in outputting what is expected instead of tracking the model's reasoning process.

Need for spatial context benchmark for non-generative models We believe that analogous or even more in-depth benchmarks should be proposed to verify the reasoning of purely vision models with a focus on assessing whether the spatial context is properly taken into account. To the best of our knowledge, there is no such benchmark even though there are many vision models that benefit from contextual information (Section 4.6). The analysis of some reasoning capabilities is performed in the case of models solving IQ tests but not in the real-world cases that we point out.

Need for measures A benchmark incorporating spatial relations understanding could serve as an important point in a checklist before the launch of DL models in critical domains. This could fulfill some of the requirements of ensuring the safety of models. For this benchmark to be successful, some new metrics to evaluate to what extent the models take into account spatial context should be proposed. So far models' investigation mostly relies on visual XAI methods (heatmaps) which are not a sufficient way to explain models in cases when spatial context matters.

Need for diverse XAI methods One of the interesting directions seems to be incorporating new types of explanations on top of visual ones (as it was done by Rabold et al. (2020) with symbolic rules) or adding them as a separate element. Moreover, it may be valuable to focus on studying the inner workings of DL models and analyzing them using some probe tasks (Adi et al., 2017). The motivation behind probe tasks is that if we can train a probe model on top of some internal representation (extracted from the DL model) to correctly predict a property of the input data, it means that the property is encoded somewhere in the representation. Overall, we should not only focus on qualitative analyses but also try to enrich them with quantitative ones for a more holistic understanding of vision models' decisions.

4.10 Spatial XAI within Pipeline

The importance of spatial context within input data to DL models may not be black or white and some procedures may be needed when using spatial XAI in the future. There are many use cases where the spatial context is crucial. But there may be some exceptions. For instance, there can be an image of a room without any thermometer in it and we ask

a model what the temperature in the room is. In this case, the context may not help in solving this task as simply visual information is not enough to solve this particular task.

We distinguish two possible scenarios of how to use the spatial XAI in the real-life Data Science pipeline but it may not be a finite list: (1) check in advance if the spatial XAI is applicable, (2) critically analyze the output of spatial XAI techniques.

In the first case, we assume that the process will be that the evaluators of the DL systems will know whether, in the particular task/ dataset, the spatial context is crucial. In some cases (i.e. ML for healthcare), the experts' domain knowledge may be needed. This way it may be assessed if the use of spatial XAI techniques for the investigation of the model is justified. Such *'check before use'* strategy is analogous to the case with the spatial regression models that can be applied instead of the Ordinary Least Square regression only if the observations and the residuals are not independent and form some clusters. This is formally checked from the perspective of spatial autocorrelation using the Global Moran's I test. Maybe, a similar form of verification could be introduced to check if the use of spatial XAI method is justified in a particular case or not.

However, in some use cases, it may not be obvious if the spatial context is meaningful. Hence, the scenario could be that we apply the spatial XAI techniques to every use case and then evaluate the output – whether the spatial XAI methods show that indeed the spatial context plays a key role in the particular task. Such *'post-hoc'* analysis can potentially bring some valuable insights into different phenomena where domain experts do not know a priori if the spatial context is important.

4.11 Conclusions

Recently, in the XAI field the shift of the paradigm from *where* to *what* was proposed, meaning that instead of only highlighting some parts of input images, we should also be able to name them semantically. We propose to go further and focus on the notion of *how* the objects within the images are oriented towards each other (contextual information). We provide many examples where the ground truth labels depend on the spatial relationship between objects i.e. in autonomous cars, surveillance systems, healthcare, agriculture, and environment studies. In these cases, the existing XAI methods will fail to provide valid explanations of the predictions of the DL models. It may happen that despite having two different models' predictions for two images, we would get semantically the same explanations. We provide examples of failures of XAI methods (unclear explanations).

We conducted a survey on papers that focused on the topic of spatial context within DL models and XAI techniques. It turned out that there are only a few works in so-called *spatial XAI* compared to the number of DL models benefitting from spatial context. In this work, we would like to direct the attention of the XAI community to this largely

unexplored area. We propose some potential research directions in the field of spatial XAI like a holistic benchmark that will evaluate models from the spatial context point of view and allow for quantitative assessment. We believe that by inspiring a community to address the topic of spatial XAI, it is possible to move forward research in XAI in general, as it would help in going out of the realm of current development directions. Moreover, a better understanding of how the models *'see'* spatial context may lead to better DL models' design in the future.

Chapter 5

P3: Duel-based Deep Learning system for solving IQ tests

5.1 Preface

The idea of duels is known in Deep Learning mainly in the form of competition between generator and discriminator in a min-max training of Generative Adversarial Networks. The question arises: ***Can the idea of duel between candidate answers (instead of models) be helpful when designing a Deep Learning model to solve visual IQ tests in the form of Raven's Progressive Matrices that require spatial understanding?***
The short answer is 'yes'.

The duels are the core of the new system called Duel-IQ, which I introduce to solve Raven's Progressive Matrices (visual IQ test). The tasks require the understanding of spatial relationships across eight panels (semiglobal context) to select, from the provided options, the panel that best satisfies the underlying relations. The spatial relationships (such as rotations between consecutive objects and differences in object sizes) can occur in rows, columns and/or on diagonals. In the case where there are multiple candidate answers that seem to fit based on some inferred rules, the correct answer is the one following the simplest rule. This relevance was the key motivation behind the idea of a duel-based approach. Relative comparisons ('Which of these two options is better?') are easier than absolute evaluation ('How good is this option in absolute terms?'). That is why the prior knowledge about the importance of relevance when selecting the answer in IQ tests was incorporated in the design of Duel-IQ, . From this point of view, it can be called a kind of Informed Machine Learning. To summarize, the work provided below has been positioned along the axes defined in Chapter 1 as shown in Table 5.1.

Table 5.1: Mapping of the paper by Tomaszewska et al. (2022) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska, Adam Żychowski & Jacek Mańdziuk. (2022). "Duel-based Deep Learning system for solving IQ tests". Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS).

5.2 Introduction

The motivation behind the launch of Artificial Intelligence (AI) was the ability to transfer human intellectual capabilities to machines, thus accomplishing the Artificial General Intelligence (AGI) level. While this goal is still ahead of us (nowadays we experience Narrow AI) (Adams et al., 2012), some steps on this path have already been made. The well-known Turing test (Turing, 1950) serves for verification whether it is possible to differentiate between a human being and an AI system based on a conversation.

More than half of a century later, the so-called visual Turing test was introduced (Geman et al., 2015). The term refers to Visual Question Answering (VQA) tasks (Antol et al., 2015) where the system is given a question and an image that it should base its answer on. Another approach to testing the intelligence of AI systems is to use the same method as is applied to people, i.e. the IQ tests, which verify the ability of abstract visual reasoning.

5.2.1 Contribution

In this work, we propose a novel approach to solving visual IQ tests based on Raven's Progressive Matrices (RPMs), called Duel-IQ. The solution is inspired by a round-robin tournament where one candidate answer's fitness is assessed in comparison to another candidate answer.

The motivation behind the idea of a tournament is the relativism of candidate-answer fitness in IQ tests (Raven and Court, 1998). To clarify this further, let us provide an intuitive example. It may happen that the two candidate answers fit the relations within the RPMs. However, the correct answer is considered the one with a simpler explanation.

For this reason, using a binary model that given a candidate answer is asked to decide whether or not it is correct, may not be an optimal setting. Another possibility would be a classification approach with as many inputs as the number of candidate answers (K)

where the model is supposed to point the correct one.

Following our intuition that it could be easier for the model to choose a correct answer out of two choices instead of K possibilities, we propose an approach located between the two above-mentioned extremes, namely with two inputs (candidate answers) compared directly in a *duel*. A series of such duels within a round-robin tournament may, what we believe and verify in the paper, lead to finding the proper answer.

Furthermore, in a tournament scenario, the risk of an incorrect system decision is alleviated within the scoring module. Similarly to ensembles (Dietterich, 2000), some predictions can be wrong, but as the final decision is a result of the aggregation of many predictions (averaging or max voting), it is often closer to the correct one. This feature offers another advantage over using just one model with the number of inputs (filled-in RPMs) equal to the number of candidate answers. An additional asset of the proposed system is its relatively low complexity.

The focus of the paper is on the input design and on the postprocessing of the network output (scoring module). Regarding the latter, we propose two schemes of pairing up candidate answers for the duels, which is also reflected in the number of classifier outputs. Moreover, we evaluate the system performance in the out-of-distribution scenario with a different set of objects (panels) in the training and testing sets, respectively. Further analysis refers to the effects of using auxiliary training.

A performance comparison of the proposed system with three other methods of similar complexity (number of trainable parameters) from the literature is conducted. The results indicate that using a duel-based approach is a meaningful idea that leads to superior accuracy.

The code of the proposed solution and the code for dataset generation are released at `https://github.com/ptomaszewska/Duel-IQ`.

5.3 Problem description

The most popular visual IQ tests are based on Raven's Progressive Matrices (RPMs), initially introduced by Raven (1936). Typically, an RPM consists of 9 panels arranged as a 3x3 matrix. The bottom right panel is empty. The test-taker is supposed to choose the correct answer from the set of candidate answers (please refer to Figure 5.1 for an example) to fill in the empty panel. The correct answer is the one that fits the relations underlying the RPM's design. The relations can occur row-wise, column-wise or/and diagonally.

Visual IQ tests are widely used as their results can be compared worldwide. They are independent of language or culture, and require little prior knowledge (Raven, 2000).

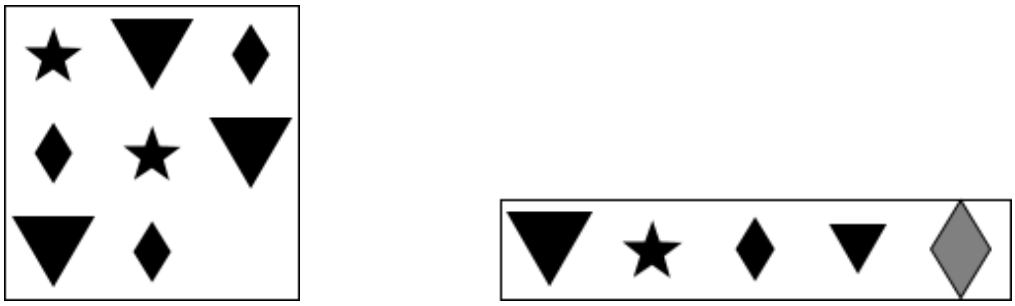


Figure 5.1: Example of an RPM task (left) with the set of candidate answers (right). The correct answer is the second from the left.

5.4 Related work

The idea of solving intelligent tasks (i.e. tasks whose solving requires intelligence) by machines is quite old. Examples of such tasks include: detection of the regularities in figures in the geometric-analogy problems (Evans, 1964) (which was the first notable attempt), Bennett's Mechanical Comprehension Tests (Klenk et al., 2011), completion of sequences of numbers (Strannegård et al., 2013) or words-related tests (Ohlsson et al., 2013).

In recent years various Machine Learning (ML) approaches, specifically Deep Learning (DL) methods were proposed to solve RPMs and the related tasks – see Małkiński and Mańdziuk (2025) for a comprehensive overview of this area.

For example, Hoshen and Werman (2017) employs convolutional neural networks to recognize the relation between two images and apply this rule to the next image (in one variant by choosing it from a given set of candidates, in another one by generating an answer from a scratch). For the sake of brevity, we will refer to the first variant as 'IQ of NN'.

DeepIQ (Mańdziuk and Żychowski, 2019) introduces a system with 3 components: a deep autoencoder which is trained to learn a feature-based representation of various figure images, an ensemble of shallow multilayer perceptrons applied to the detection of feature differences, and a scoring module inspired by the human approach used for the final assessment of candidate answers.

The most common approach to solving an RPM relies on identification of the relationships among objects (panels) and using them to identify the correct answer. This idea can be realized by various DL techniques, e.g. Relation Network (Barrett et al., 2018; Zheng et al., 2019), Multiplex Graph Networks (Wang et al., 2020), Scattering Compositional Learner (compositions of a sequence of neural modules) (Wu et al., 2020), or Hierarchical Rule Induction Network (Hu et al., 2021) imitating human induction strategies.

The Wild Relation Network (WReN) introduced by Barrett et al. (2018), solves an RPM in a few steps. First, 8 RPM panels are concatenated depthwise with all candidate answers (one at a time), resulting in K input packages (K is the number of candidate

answer panels). In each package, the panels are embedded using a convolutional neural network. Then, pairs of such embeddings are input to the relation network whose scores are later aggregated to yield the final output. The above WReN training is extended with an auxiliary training where the network learns to explicitly point the relations, objects and their attributes provided in the input.

Yet another approach (Małkiński and Mańdziuk, 2024) casts the problem of solving RPMs into a multi-label classification framework with labels determined by the set of abstract rules underlying the RPM.

There is no publicly available dataset with the real IQ test questions as they are strictly confidential. Consequently, the authors of the above-mentioned works used artificially generated datasets which vary in complexity, e.g. the number of candidate answers or the selection of geometric transformations applied to the RPM panels. This makes the performance comparison of the solutions proposed in the literature not a straightforward task. Chollet (2019) proposed the Abstraction and Reasoning Corpus (ARC) dataset for the assessment of abstract reasoning skills which, however, does not reflect the structure of IQ tests taken by humans.

A broader look at the field of Abstract Visual Reasoning (of which RPMs are part) is presented in the recent survey paper by Małkiński and Mańdziuk (2023).

5.5 Proposed solution

We propose a duel-based system (see Figure 5.2) for solving visual IQ tests involving RPMs, called Duel-IQ. The system does not require any expert knowledge and its operation is divided into four steps: preprocessing, feature extraction, classification and scoring.

During the preprocessing step, RPMs are filled with candidate answers. Therefore, one IQ test task is afterwards represented by as many filled-in RPMs as the number of candidate answers. These RPMs are inputs to subsequent modules. First, the convolutional autoencoder is used for feature extraction (see Figure 5.2). Then, in the classification network, there are fully connected layers on top of the encoder to judge the winner of the duels (in some scenarios a draw is also possible). The process of pairing up the RPMs filled in with candidate answers and the classifier network are described in more detail in Section 5.5.2. Predictions of the classifier are later aggregated in the scoring module in order to generate the final decision which of the candidate answers best fits a given RPM (see Section 5.5.3).

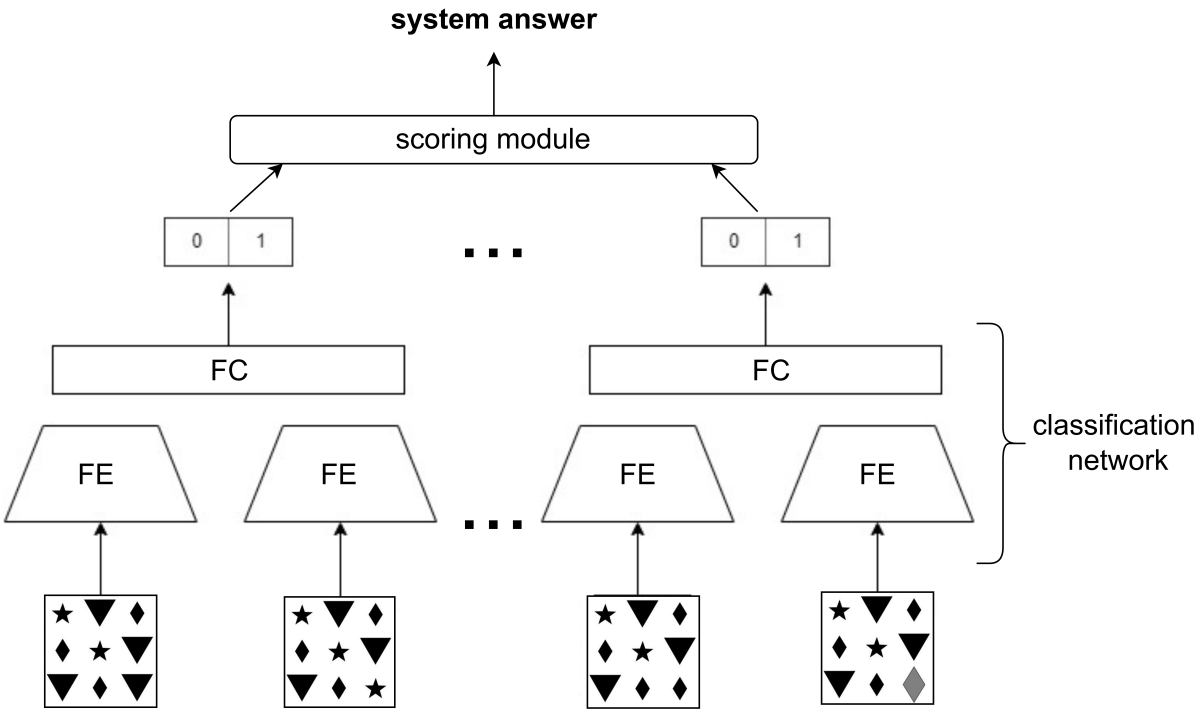


Figure 5.2: Proposed duel-based system (*winner_in* variant) in which the results of duels are aggregated in the scoring module. FE stands for feature extractor (encoder part of autoencoder), whereas FC for fully connected layers.

5.5.1 Autoencoder

The filled-in RPMs are input (one at a time) to the convolutional autoencoder whose task is to recreate the given image. The output of the encoder part provides a compressed representation of the input.

5.5.2 Classification network

The next step is the dataset preparation for the round-robin duels. Two scenarios are considered at this stage: *winner_in* and *any_in*. In the first one, in the training pairs, there is always a correct candidate answer present. Therefore, the classifier's goal is to choose the correct answer from the two (two-class classification).

In the *any_in* scenario, there is no obligation that there must be a correct answer within the input pair, therefore there may be pairs where none of the candidate answers is correct. In such a case, the model is allowed to choose a third class, meaning a draw – both answers are wrong.

Please observe that at test time (the time of solving an unknown IQ test), the system will experience scenarios where there is no correct answer within the input pair as the pairing up is done in such a way that all combinations are analyzed (irrespective of the order of candidate answers in the pair). Hence, the system trained in the *winner_in*

scenario will, at test time, experience the situations it was not trained on. The goal of such a setup was to verify whether the model, in cases where there are two incorrect candidate answers in the input pair, outputs probabilities of the two classes (after the softmax function) of around 0.5 meaning that it is largely uncertain as such a scenario was not previously observed.

The formulated duel pairs are input to the classification network. Each RPM is compressed using the encoder and then the two representations are concatenated and passed to dense layers to point a winner. Please recall that the encoder is a part of the trained autoencoder (cf. Section 5.5.1).

5.5.3 Scoring module

The results of the duels are aggregated in the scoring module. Two approaches are tested: *probability sum* and *winning frequency*.

Probability sum scoring

Let us denote the output of the classifier with softmax in the last layer as $f(x_i, x_j)$ where (x_i, x_j) is the input pair of RPMs filled with candidate-answers ($i \neq j$). Thus, $f(x_i, x_j)$ is a vector of probabilities where the first element signifies that x_i is the correct answer, and analogically the second element refers to x_j being the correct outcome. In case of the *any_in* scenario, there is the third element in $f(x_i, x_j)$ meaning that both candidate answers are incorrect.

For brevity, we define $f_{x_i}(x_i, x_j)$ as the probability returned by the model specifying that x_i is the correct answer. The probabilities that the answer x_i wins its duels are aggregated in the scoring module:

$$s_{x_i} = \sum_{j \neq i} f_{x_i}(x_i, x_j) \tag{5.1}$$

Let us denote $s = (s_{x_0}, ..., s_{x_K})$ where K is the number of candidate answers. The final decision of the proposed system is $\hat{y} = \text{argmax}_x\, s$.

Winning frequency scoring

Following the notation introduced in the previous section, we formulate the *winning frequency* approach used in the scoring module. Conceptually, it resembles max-voting in ensembles. Here, instead of probabilities, the number of duels wins is considered. Let us denote the equivalent of s_{x_i} from the previous approach:

$$freq_{x_i} = \sum_{j \neq i} I(\text{argmax}_x\, f(x_i, x_j) = x_i) \tag{5.2}$$

Later, similarly to the *probability sum* approach, the f vector is defined as $f = (freq_{x_0}, ..., freq_{x_K})$. The final answer of the proposed system for a given IQ task, in this scenario, is $\hat{y} = \text{argmax}_x f$.

There can happen a particularity when more than one answer has the same number of duels won. In such a case, the system returns the one with a smaller index as the final answer.

5.6 Experimental setup

5.6.1 Datasets

We use a data generation scheme analogous to that in the work of Mańdziuk and Żychowski (2019), but we increase the number of data samples.

The RPM as a whole (i.e. a 3x3 grid of panels) is of the size of 150x150 pixels. Each of the generated test instances reflects a change in one of the following features: shape, size, or rotation. In the case of shape, each row/column contains three different shapes and the task is to point the missing shape in the last row/column. For size and rotation tests, the respective feature (size or rotation) increases or decreases iteratively (with a constant ratio) in consecutive rows/columns. Additionally, in half of the test instances, random perturbations were applied, i.e. apart from the main feature change (described above), another feature was changed randomly (e.g. a figure which iteratively changes its size was randomly rotated).

All figures' features were generated within the following ranges: size (width and height) from 25 to 50 pixels, rotation from 0 to 2π, shading from 0 to 255. Each figure was placed centrally within a panel (a square cell). Please consult the work of Mańdziuk and Żychowski (2019) for the details.

To verify the generalization capability of the proposed system, three experiment scenarios were created:

1. the same shapes in training, validation and test datasets (14 figure shapes presented in Figure 5.3),

2. the figures in training and validation sets were the same (left set of 7 figure shapes in Figure 5.3) but different than in test set (right set of 7 figure shapes in Figure 5.3),

3. the same shapes in training, validation and test sets (left set of 7 figure shapes in Figure 5.3)

In experiment 2, we wanted to perform a cross-dataset check. To have a baseline to refer to, experiment 3 is run to assess how difficult the task was. In practice, experiment 2 is conducted in such a way that the model is trained on data from experiment 3 but

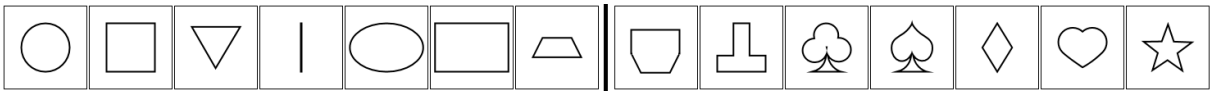


Figure 5.3: Figure shapes used in experiments.

evaluated on testing set from experiment 2. As the number of figure shapes in the training set in experiments 2 and 3 is two times smaller than in the case of experiment 1, we used a two times smaller training dataset (see Table 5.2).

Table 5.2: The number of single task questions as shown in Figure 5.1. In the preprocessing stage, each task is extended to 5 filled-in RPMs that are subsequently combined into pairs for the duels. The training set in experiments 2 and 3 is shared, likewise the validation set. In experiments 2 and 3 the same trained model is tested on two different sets.

	TRAIN	**VAL**	**TEST**
experiment 1	34 400	9 800	4 800
experiment 2	17 200	4 900	4 800
experiment 3			4 800

5.6.2 Autoencoder

The final architecture of the autoencoder (depicted in Figure 5.4) was chosen by means of a grid search where the number of layers and the numbers of filters in convolutions and transposed convolutions were selected. We analyzed architectures that resulted in different hidden representation sizes and investigated the impact of changing transposed convolution layers into a combination of upsampling and convolution.

It turned out that the proposed architecture (Figure 5.4) resulted in a lower mean squared error (MSE) and a slightly higher structural similarity index measure (SSIM) than the method with the same latent representation size but with no use of transposed convolution. At the same time, a peak signal-to-noise ratio (PSNR) of our model was slightly lower than that of its fully convolutional counterpart. In conclusion, the proposed architecture presents the best tradeoff between performance metrics (see Section 5.7) and hidden representation size.

5.6.3 Classification

The architecture of the Duel-IQ classifier is shown in Figure 5.5. The classifier takes a pair of filled-in RPMs that participate in a duel as an input. Both RPMs are compressed using the encoder and then passed through an additional trainable convolutional layer ($kernel\_size = 2, n\_filters = 4$). Next, the two representations are concatenated and processed by two fully connected layers to make the final prediction.

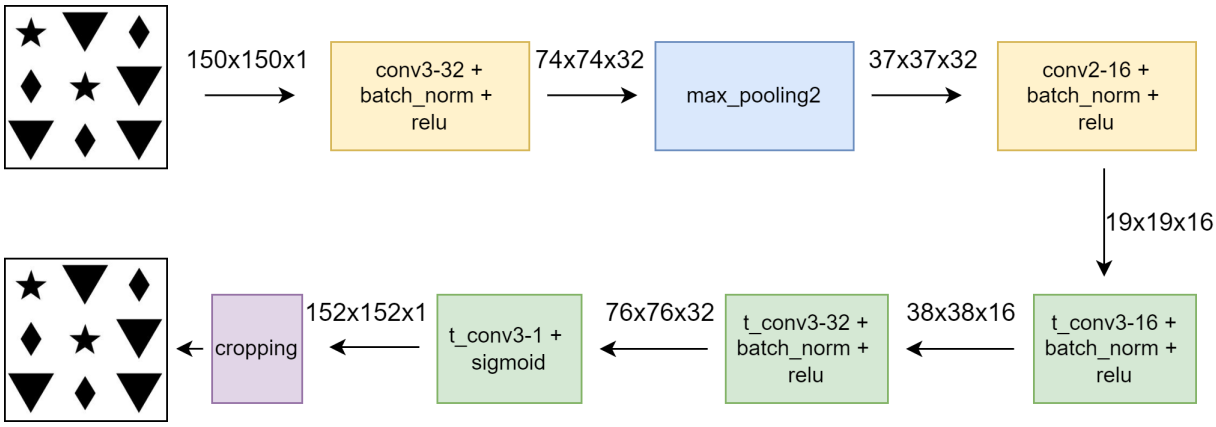


Figure 5.4: Autoencoder architecture. The convolutional layer parameters are denoted as "conv-kernel size-number of filters". Likewise transposed convolutional layers are described. The max pooling layer parameter refers to the receptive field size.

In order to choose an optimal architecture of the classification network, we checked the impact of making the weights of the encoder trainable as opposed to frozen and observed that enabling fine-tuning is beneficial. We decided to stack an additional convolutional layer on top of the encoder, followed by dense layers, in order to further decrease the size of hidden representations before their concatenation. This convolutional layer is a proxy between the encoder (meant to compress the input in possibly lossless manner) and the dense layers (that focus on comparing the two inputs).

Please note that the networks were not pretrained on Imagenet data (which is a common practice) since RPMs have a different structure than natural scene images. In filled-in RPMs, there are high frequencies and a lot of white space. Moreover, the image is not coherent since it is composed of 9 panels. Furthermore, the images in the dataset are in grayscale and therefore transforming them to RGB space would be redundant and would simply multiply the same information.

In some of the experimental modes, we employed auxiliary training where the model was asked to return the measure of dissimilarity between the candidate answers within the input RPMs. The dissimilarity measure was computed according to Algorithm 1. In the formulas for each term, we applied normalization to the [0,1] interval. The most complex is the formula for rotation similarity as it takes into consideration the cyclic nature of the issue. The symbol % means a remainder. The overall dissimilarity measure is also normalized to 0-1, i.e. 0 means exactly the same figure answers, 1 – completely different ones.

5.7 Experimental results

All experiments were run with the following parameters. The training was scheduled for 20 epochs each, however, it could have been stopped earlier if the loss on validation

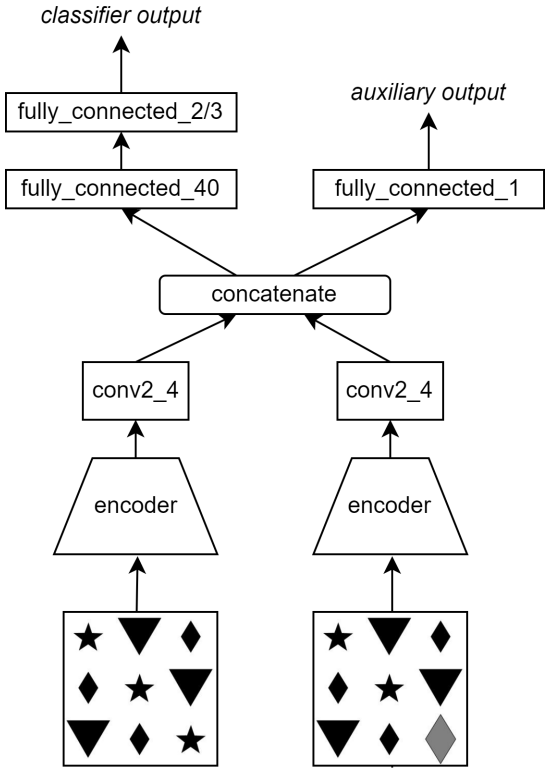


Figure 5.5: Classification network with auxiliary training applied. Convolutional layers are denoted in the same manner as in Figure 5.4. A parameter in each fully connected layer refers to the number of neurons.

set increased in more than 5 consecutive epochs. The model was optimized using Adam algorithm (Kingma and Ba, 2014) with the learning rate 0.001 and decaying coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

5.7.1 Autoencoder

The highest spatial compression ratio of the autoencoder with still satisfactory quality of the reconstructed images was 4. The quality was assessed using three metrics: MSE, SSIM and PSNR – see Table 5.3.

Table 5.3: Performance metrics of the autoencoder on test data in different experiments (MSE – the lower, the better; SSIM, PSNR – the higher, the better).

	MSE ↓	**SSIM ↑**	**PSNR ↑**
experiment 1	0.002	0.959	29.156
experiment 2	0.003	0.947	27.259
experiment 3	0.001	0.969	30.495

Reconstructed images are of high quality – all structural information is preserved. Note

Algorithm 1 Dissimilarity measure between two candidate answers

Require: Parameters of candidate answer 1 and 2

$score \leftarrow 0$

if $shape_1 == shape_2$ **then**

  $score \leftarrow score + 1$

end if

$score \leftarrow score + \frac{|\text{intensity}_1 - \text{intensity}_2|}{256}$

$score \leftarrow score + \frac{|\text{width}_1 - \text{width}_2|}{50}$

$score \leftarrow score + \frac{|\text{height}_1 - \text{height}_2|}{50}$

$first\_angle \leftarrow |rotation_1 \% 360, -rotation_2 \% 360|$

$score \leftarrow score + \frac{\min(first\_angle,\ 360 - first\_angle)}{180}$

return $dissimilarity = score/5$

that SSIM is close to the ideal value of 1, whereas PSNR of about 30 is also considered very high.

The best values are achieved in experiment 3 where the task is the easiest because the same figures are used in the training and the testing sets, and only 7 types of object types (shapes) are considered. The second-best scores are in experiment 1 which is similar to experiment 3, but the number of shapes is doubled (there are 14 of them). Experiment 2, in which the model is trained on a dataset containing shapes different from those composing the test set, exhibits the worst performance.

Apart from the verification of quality metrics, a visual inspection was also performed which indicated that despite the drop of SSIM and PSNR measures in experiment 2, the reconstructed images are almost identical to the original images except that the edges of the figures are slightly blurry (cf. Figure 5.6).

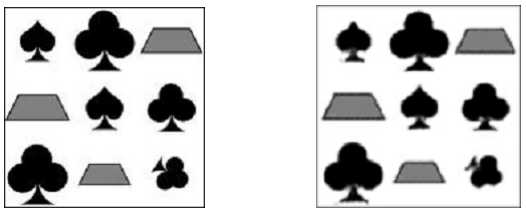


Figure 5.6: Comparison of original image (left) and the reconstructed one (right).

5.7.2 Classification network

In total, 8 different variants of classification networks were analyzed. We trained the networks for experiments 1 and 3. In experiment 2, the same model was used as in experiment 3 (training and validation sets are shared in both scenarios). Moreover, we verified whether the choice of the *winner_in* or *any_in* approach has any impact on the performance. Another combination was with and without the auxiliary training applied. Each variant was trained 10 times with different random seeds. Results of these experiments

on the test set are presented in Table 5.4.

Table 5.4: Classifier performance on test set in various scenarios – mean accuracy (%) and its standard deviation. For brevity, 'exp' stands for experiment.

		INPUT	ACC
with aux. training	exp 1	winner_in	92.7 ± 2.0
		any_in	77.1 ± 7.7
	exp 2	winner_in	70.9 ± 1.2
		any_in	53.1 ± 1.6
	exp 3	winner_in	90.8 ± 0.6
		any_in	83.6 ± 0.8
without aux. training	exp 1	winner_in	92.2 ± 2.5
		any_in	70.0 ± 15.8
	exp 2	winner_in	70.6 ± 1.8
		any_in	53.3 ± 1.3
	exp 3	winner_in	90.3 ± 1.6
		any_in	83.3 ± 1.6

The following conclusions can be drawn from the experiments. **(1)** Given the same maximum number of training epochs, there was no significant difference in the results in experiments with and without auxiliary training. **(2)** The metrics were higher in the *winner_in* scenario than in *any_in* which is not that surprising since in the former case a two-class classification was performed, whereas in the latter case, a three-class one. **(3)** The classification network was able to generalize to the test data composed of figures other than those used in the training set (experiment 2). **(4)** The results are visibly better than a random choice that would be 50% in the *winner_in* scenario and 33% in the *any_in*.

5.7.3 Scoring module

The scoring module was applied on top of the classification variants specified in Table 5.4. We analyzed the performance achieved with the two scoring approaches: *probability sum* and *winning frequency* described in Section 5.5.3. The outcomes are presented in Table 5.5.

In most of the cases, the high standard deviation in Table 5.5 is due to the fact that one of the runs resulted in significantly lower performance, whereas the rest of them featured similarly high quality metrics.

The results show that in the *probability sum* scoring, the impact of the auxiliary training is negligible only in the case of experiment 2. In experiment 1, the auxiliary training improved the accuracy, whereas in the case of experiment 3, it slightly deteriorated the performance (however, the difference is within the standard deviation limits). This can be

Table 5.5: Mean test accuracy (%) and its standard deviation of Duel-IQ in various scenarios. The best performance (in terms of mean) in corresponding experiments is presented in bold. “PROB." stands for *probability sum* and “WINNER" for *winning frequency* algorithm in the scoring module.

		INPUT	**PROB.**	**WINNER**
with aux. training	exp 1	winner_in	$\mathbf{82.8 \pm 6.2}$	59.1 ± 7.8
		any_in	79.4 ± 7.3	78.4 ± 7.4
	exp 2	winner_in	47.4 ± 1.9	36.4 ± 2.3
		any_in	$\mathbf{49.1 \pm 1.8}$	47.5 ± 1.3
	exp 3	winner_in	77.6 ± 6.6	59.8 ± 4.8
		any_in	73.0 ± 9.1	70.3 ± 10.0
without aux. training	exp 1	winner_in	60.1 ± 11.5	42.7 ± 6.5
		any_in	47.2 ± 10.3	47.3 ± 7.6
	exp 2	winner_in	47.3 ± 2.0	37.5 ± 2.0
		any_in	48.9 ± 1.9	46.8 ± 1.5
	exp 3	winner_in	$\mathbf{80.0 \pm 3.2}$	58.7 ± 5.9
		any_in	77.7 ± 8.0	75.5 ± 10.2

attributed to the fact that in experiment 1, the dataset was twice as big as in experiments 2 and 3.

The overall performance of the system with *probability sum* scoring in *any_in* and *winner_in* scenarios was almost the same in every analyzed case. Note that the classifier accuracy (Table 5.4) was higher in *winner_in* than in *any_in* scenario. A possible explanation why the advantage in terms of classifier accuracy was not directly translated to the accuracy of the overall system, can be attributed to the fact that the *winner_in* approach has a flavor of out-of-distribution, which makes the problem to be solved more difficult. Therefore, the classifier accuracy in *winner_in* and *any_in* did not transfer in the same way to the overall system's results in Table 5.5.

A comparison of the *probability sum* and *winning frequency* shows that the results are very similar in the *any_in* scenario whereas significantly different in the *winner_in* scenario where *probability sum* outperforms *winning frequency* approach. This difference can be attributed to the fact that in the *any_in* scenario, the model is not explicitly taught which probabilities it should generate. Therefore, if in the duel with two incorrect answers it favors one of the answers even a little bit, the final model decision in the *winning frequency* approach can be distorted.

5.7.4 Baseline comparison

We tested our hypothesis and the main motivation behind the proposed system saying that (in this task) it is more effective to perform a series of 2- or 3- class classifications

(duels) than a single K-class classification, where K is the number of candidate answers. The latter approach was mostly investigated in previous works. Therefore, we analyzed the analogous neural network architecture to the one from Figure 5.5 but in a 5-class classification scenario, this time with 5 filled-in RPMs as input. In this case, neither an auxiliary training nor the scoring module were incorporated. To adjust to the increased overall input dimensionality, we changed the number of neurons in the penultimate fully connected layer to 400. The other training parameters were the same as in the baseline Duel-IQ system.

We compared the performance of Duel-IQ with other methods of similar complexity: DeepIQ (Mańdziuk and Żychowski, 2019), Wild Relation Network (WReN) (Barrett et al., 2018), and "IQ of NN" (Hoshen and Werman, 2017).

In the case of WReN, we adjusted the publicly available code[1] to our dataset features. We also resigned from the auxiliary training due to its fundamentally different nature in WReN (Barrett et al., 2018) compared to our approach. While in Duel-IQ only the dissimilarity measure between the candidate answers is presented during training, in the work of Barrett et al. (2018) an information about the underlying relations in RPM is provided. The use of this key information makes the overall task of choosing the correct answer much easier for WReN, rendering a direct comparison unfair.

Regarding the "IQ of NN" approach, recall that in the original implementation (Hoshen and Werman, 2017) only the relationship in one row (not the entire RPM) was analyzed. Therefore, we first of all changed the number of input panels. Furthermore, we implemented and tested two versions of this method which are direct extensions of the solution proposed by Hoshen and Werman (2017) and differ only in the number of output classes (output neurons). The first one (*5-class*) has 5 outputs (a traditional classification approach), whereas the second one (*2-class*) has only two outputs (a duel-based approach). Results of a series of such duels are further aggregated in the scoring module using the *probability sum* algorithm from Duel-IQ. This option allows verifying the potential benefits of the duel-based approach vs traditional K-class classification in "IQ of NN" model. *winner_in* scenario was investigated as it brought better results in the case of Duel-IQ.

The above-mentioned baseline methods were run 10 times. The accuracy (mean and standard deviation) is presented in Table 5.6. We summarize the results of Duel-IQ by indicating the highest performance across 8 different variants of the experiment (two pairing schemes: *winner_in*/*any_in*, two algorithms in the scoring module: *probability sum*/*winning frequency*, two training schemes: with/without auxiliary training).

The results prove that Duel-IQ outperforms other analyzed methods when the training and testing datasets consist of the same figure shapes (experiments 1 and 3). Otherwise (experiment 2), the leading method is DeepIQ.

[1]https://github.com/Fen9/WReN, https://github.com/mikomel/wild-relation-network

Table 5.6: Performance comparison of different methods for solving IQ tests: mean, standard deviation, and maximum accuracy (in brackets). Best results in each experiment are in bold.

ALGORITHM	**EXP 1**	**EXP 2**	**EXP 3**
Duel-IQ	$\mathbf{82.8 \pm 6.2}$ (**88.8**)	49.1 ± 1.8 (54.9)	$\mathbf{80.0 \pm 3.2}$ (**85.3**)
Duel-IQ (5-class version)	72.2 ± 5.4 (75.6)	49.2 ± 1.1 (50.8)	71.5 ± 3.2 (74.1)
DeepIQ	73.0 ± 3.7 (78.3)	$\mathbf{61.7 \pm 3.9}$ (**67.4**)	68.8 ± 2.4 (72.5)
WReN	54.1 ± 1.1 (56.3)	34.1 ± 1.2 (36.0)	49.0 ± 1.5 (51.1)
IQ of NN (2-class version)	67.3 ± 20.1 (90.3)	49.1 ± 10.2 (72.8)	68.3 ± 11.2 (89.8)
IQ of NN (5-class version)	55.8 ± 11.5 (76.3)	44.3 ± 5.7 (54.8)	55.1 ± 6.7 (69.3)

A possible reason for the better performance of DeepIQ in this setting is that it learns the relations between two panels and not the whole filled-in RPM. Such an approach (focusing on two shapes) may lead to developing a more universal representations that help solving an out-of-distribution scenario. Furthermore, DeepIQ is explicitly enforced to focus on predefined aspects such as rotations, changes in size, etc.

However, the drawback of the DeepIQ solution is the need to incorporate expert knowledge for the definition of a possible set of relations within the RPMs. On the contrary, the proposed duel-based system is fully autonomous (does not require *a priori* expert information about the aspects that need to be taken into consideration while solving IQ tests).

Moreover, the experiments confirmed that Duel-IQ results in better accuracy than 5-class classification model of similar architecture (referred to as "Duel-IQ (5-class version)" in Table 5.6). This observation aligns with the initial hypothesis that classification with two input RPMs can be easier than the one with as many input filled-in RPMs as the number of candidate answers.

Similar conclusions can be drawn from an analysis of the results of the two variants of "IQ of NN". Here, the 2-class version outperforms the 5-class one, as well. However, due to the high variance of "IQ of NN" experimental results, the claim that round-robin approach is beneficial in this case should be taken with care.

The results achieved by "IQ of NN" are mediocre. The possible reason is that the authors used their algorithm for a simplified task with 2 panels and the system was supposed to choose the candidate answer that fits the relation – the third missing panel. This task is much less complex than the analyzed RPMs where there are three rows

(matrix) and the relations can occur vertically, horizontally, and/or diagonally. Another aspect is the tendency of the "IQ of NN" to overfit. The authors of the model reported its training for 30 to 100 epochs depending on a problem. When applying the algorithm to our dataset, the model was stopped after 8 to 15 epochs as a result of an early stopping mechanism – the loss on validation set with patience parameter set to 5. The 2-class version achieved the highest accuracy (in a single run), though, at the same time, the highest variance was reported for this model.

The worst results were achieved with WReN. If figure shapes in the training and testing sets were from the same distribution, the performance of WReN was similar to "IQ of NN (5-class)", otherwise (experiment 2) it was clearly lower.

Quite surprisingly, WReN results were worse in experiment 3 than in experiment 1. This is unexpected because experiment 1 is more difficult than experiment 3 due to higher number of figure shapes. The performance degradation in experiment 3 can be attributed to the fact that in order to train the relation network within WReN, a large number of samples is required (note that the number of training samples in experiment 1 was twice as big as in experiment 3).

In order to assert that the results of Duel-IQ are not biased by the figure shapes chosen for training and testing (see Figure 5.3), additional tests were performed with randomly shuffled figure shapes between these two sets. The averaged accuracy (over 10 independent runs) in experiments 2 and 3 equaled $51\% \pm 3\%$ and $84\% \pm 3\%$, respectively. The results were generally 2-3% higher compared to the baseline cases (cf. Table 5.6) which confirms that neither the choice of figures nor their particular split into training and test sets were biased. In fact, the results suggest that the particular setting in the baseline experiments can be considered "more difficult than average".

5.8 Human performance

The performance of Duel-IQ was confronted with human performance in two surveys conducted among 94 university students and PhD candidates, mostly with Computer Science (CS) background. 61 people took part in the first survey and 33 others in the second one. Each survey was composed of 21 representative RPM problems from our dataset of experiment 1 (7 problem types x 3 instances). The questions (42 in total) varied in their level of difficulty.

The average human accuracy weighted by the ratio of participants in each survey equaled $77.3\% \pm 5.6\%$ which is on par (within standard deviation ranges) with Duel-IQ in experiment 1 ($82.8\% \pm 6.2\%$). Human performance on Raven's Matrices of similar complexity was tested by Matzen et al. (2010) within the general population with 60% to 80% of correct answers depending on the test type, which is consistent with our survey, taking into account a particular education profile (CS) and age of our participants.

5.9 Conclusions

In this paper, we propose a duel-based neural network system (Duel-IQ) suitable for solving RPMs – a well-known example of abstract visual reasoning tasks. The following variants of the system are investigated: (1) the same vs different sets of figures in RPMs in training and test sets, respectively; (2) with vs without auxiliary training; (3) *winner_ in* vs *any_ in* pairing schemes; (4) *probability sum* vs *winning frequency* scoring formulas.

The results prove that an auxiliary training that enforces the system to assess the similarity between the candidate answers may improve the performance, mainly in the case of the most complex dataset (experiment 1). For *probability sum* scoring function, when the training and test data have the same distribution (experiments 1 and 3), the *winner_ in* slightly outperforms the *any_ in* approach, whereas in out-of-distribution scenario (experiment 2), it is the other way around. Further investigation of this observation is a focus of our further studies.

Overall, the results confirm the abstract reasoning abilities of the ensemble-like duel-based model and its superiority over 5 reference methods when training and test sets are composed of the same figure shapes. In out-of-distribution setting Duel-IQ is inferior to DeepIQ (Mańdziuk and Żychowski, 2019). Moreover, the performance of Duel-IQ is on par with the results of a group of 94 CS students and PhD candidates.

Chapter 6

Digital pathology

As pointed out in Chapter 4, the spatial context plays an important role when analysing tissues. This applies to spatial transcriptomics and digital pathology.

Spatial transcriptomics (Williams et al., 2022; Jain and Eadon, 2024; Isnard and Humphreys, 2025) is a recent field of research that tries to match gene expression patterns within their original tissue context, essentially providing a "map" of where genes are active in a tissue. This enables researchers to explore how gene expression corresponds to specific cell types, tissue structure, and disease-related changes. It helps reveal how various cell types are distributed and how they interact within the tissue environment. Unlike conventional transcriptomics, which analyzes gene expression without spatial reference, this approach integrates both molecular and spatial data.

In this dissertation, the focus is however on digital pathology which is the topic of the following chapters. Digital pathology (Vipul Baxi, 2022; Song et al., 2023; Waqas et al., 2023) is a rapidly growing field that automates the process of diagnosis of tissue extracted during biopsy to analyse if the lesion is benign or with tumor. Traditionally, pathologists examine tissue samples on glass slides using a microscope. Digital pathology automates the process by the use of computer-aided methods mainly Deep Learning-based.

The core elements in digital pathology are Whole Slide Images (WSIs) which are the high-resolution digital images of tissues obtained during a biopsy procedure. The tissue after the acquisition is stained using the H&E method. It employs two distinct dyes (hematoxylin and eosin) to color various cellular structures in different colors, enhancing the visibility and interpretation of tissue sections. Hematoxylin colors the nuclei blue, whereas eosin gives a pink tone to the cytoplasm and surrounding extracellular elements (Fischer et al., 2008). The WSIs are of huge size e.g. $60,000 \times 100,000$ pixels. Moreover, their size is not standardized. The analysis of WSIs is often much more difficult than classification of the central object in the image, as i.e. in popular datasets like Imagenet (Deng et al., 2009), CIFAR-100 (Krizhevsky, 2009) etc. This increased difficulty originates from the fact that in histopathology the differences between the images from different classes are much more subtle than in the aforementioned datasets. Various tasks are tackled in

digital pathology: tissue segmentation, cell detection and counting, classification (benign vs. tumor, subtyping), prognosis of metastasis or survival rate.

In the thesis, I focus on tissue classification and prognosis of time to death due to metastasis. From a Machine Learning perspective (in terms of model architecture), both tasks are essentially classification problems; however, prognosis aims to foresee disease progression rather than assess the current state.

The substantial memory requirements imposed by the large size of whole slide images make it difficult to process and classify them using end-to-end Deep Learning approaches. Consequently, Multiple Instance Learning (MIL) has emerged as the predominant approach for the analysis of WSIs.

6.1 Multiple Instance Learning

The setup in Multiple Instance Learning (MIL) is as follows. There is a collection of instances $(x_1, ..., x_k)$ that form a bag (X) which has one global label (Y). Therefore, one data sample in this setup is a bag. Since bags are not required to contain the same number of instances (k), batching data samples for efficient and accelerated training becomes challenging. As a result, training is often performed on a single bag at a time rather than on batches. Here are some examples of bags for clarification:

- a set of keys of various colors; the label is the answer to the question of whether there is a green key in a set,
- a set of patches constituting a WSI; the label is the information on whether the lesion is benign or with tumor.

The label Y is positive for a bag if at least one instance (x_i) in the bag meets criterion (has label $y_i = 1$):

$$Y = \begin{cases} 0, & \text{if } \sum_i y_i = 0, \\ 1, & \text{otherwise.} \end{cases}$$

It means that if at least one patch contains tumor cells, then the whole tissue is classified as tumor.

In Multiple Instance Learning (MIL), there are two classes of approaches: instance-level and embedding-level. In the instance-level approach, the classifier generates a score for each instance that is subsequently aggregated using pooling. A major challenge of this method is that instance-level labels are typically not readily available to train an effective classifier. Instead of performing labor-intensive fine-grained annotations at the instance level, some attempts have been made to assign the same label to every patch within a WSI. However, this strategy introduces noisy labels that hinder the training process. To mitigate this issue, techniques such as co-teaching (Han et al., 2018) can be employed,

though the final performance often remains suboptimal. Such a uniform labeling strategy is particularly problematic in cases where a very small tumor is present within a large tissue sample.

In the embedding-level approach, the pipeline is as follows: (1) features (h_i) are extracted from each instance, (2) the features are aggregated (e.g., max, mean) to form a representation of the bag, (3) the classification is performed on the bag-level. This is a popular approach in digital pathology. In this case, a bag is a Whole Slide Image and an instance is a patch.

One of the first advancements to the embedding-level MIL solutions was the replacement of the simple aggregation function (max, mean) with a weighted average:

$$\mathbf{z} = \sum_{i=1}^{k} a_i \mathbf{h}_i.$$

The coefficients are the attention scores (a_i) assigned by the model to patch features (h_i). The higher the attention score, the higher contribution of the features of a particular patch to the overall WSI representation.

The first attention mechanism in Deep Learning was proposed by Gkartzonika et al. (2023). The one-hidden layer feed-forward network was applied to calculate the attention alignment in recurrent neural networks in machine translation. The computed attention scores are scaled using softmax so that they sum up to 1. The scaling is important to make the solution invariant to the length of sentences. In the case of MIL, the scaling makes the solution invariant to the number of instances in a bag.

Various variants of attention mechanism were proposed in MIL models. The first variant was defined as follows (Ilse et al., 2018):

$$a_i = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_i^\top\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\right\}}, \tag{6.1}$$

where $\mathbf{w} \in \mathbb{R}^{d\times 1}$ and $\mathbf{V} \in \mathbb{R}^{d\times m}$ are trainable (d is the length of the feature vector, m – hyperparameter). The aim of the application of hyperbolic tangent (*tanh*) was to foster (dis)similarities among instances.

The second method, known as the gated attention mechanism (Ilse et al., 2018), was a modification of the previous approach, designed to capture more complex relationships that may not have been possible to capture earlier, since the hyperbolic tangent is nearly linear within the range $[-1, 1]$:

$$a_i = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_i^\top\right) \odot \operatorname{sigm}\left(\mathbf{U}\mathbf{h}_i^\top\right)\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot \operatorname{sigm}\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\right\}}, \tag{6.2}$$

where *sigm* is a sigmoid function. The same notation was applied as in the previous

formula (**U** matrix is analogous to **V**).

The attention scores provide interpretability, allowing visualization of tissue regions that most influenced the model's decision (Figure 6.1). Such visualizations are valuable for physicians who can inspect if the important patches from the model's point of view are indeed important from the medical perspective. For instance, in the tumor-versus-healthy classification task, when the model identifies a tissue as tumorous, it is expected to attend predominantly to the tumor region. If there is an alignment, the physicians' trust in the model is enhanced.

Note that the methods for aggregation of features from the patches within bags are permutation-invariant. There is an assumption that neither the ordering nor the dependency of instances within a bag matters (Ilse et al., 2018). Moreover, the positional encoding is not associated with patch features. **These observations were a motivation for my research as it seemed counterintuitive that when analysing tissues the inner structure and spatial relationships are not taken into account. From the perspective of histopathologists, it does not seem reasonable to analyse a whole slide image after the random permutation of its patches. Therefore, the question arises whether such counterintuitive invariance to spatial context impacts the performance and reasoning of Multiple Instance Learning models. It is investigated in the following chapters.** I focus on the digital pathology use case, however, the MIL models also find application in the analysis of other medical images (Barbosa et al., 2024) like Doppler Ultrasound frames (Struski et al., 2023, 2024), lung X-ray (Seibold et al., 2021; Avolio et al., 2024), hand X-ray (Bo et al., 2026), etc.

Below, I provide the description of two popular MIL models (CLAM and TransMIL) that differ in the way they approach the incorporation of information on patch neighbourhood and their position within WSI. These models are investigated from the spatial context point of view in the following chapters.

6.1.1 CLAM

One of the most fundamental Multiple Instance Learning models for weakly supervised classification of Whole Slide Images is CLAM (Clustering-constrained Attention Multiple Instance Learning) (Lu et al., 2021a). The model pipeline is shown in Figure 6.2.

The first step in the approach is data preprocessing. The most important element is a tissue segmentation to get rid of the unnecessary background and holes in tissue. This way the non-informative regions of WSI images are discarded which is important for two reasons. The model does not waste training time on regions that lack meaningful information (which conserves computational resources) and it reduces the risk of the model mistakenly

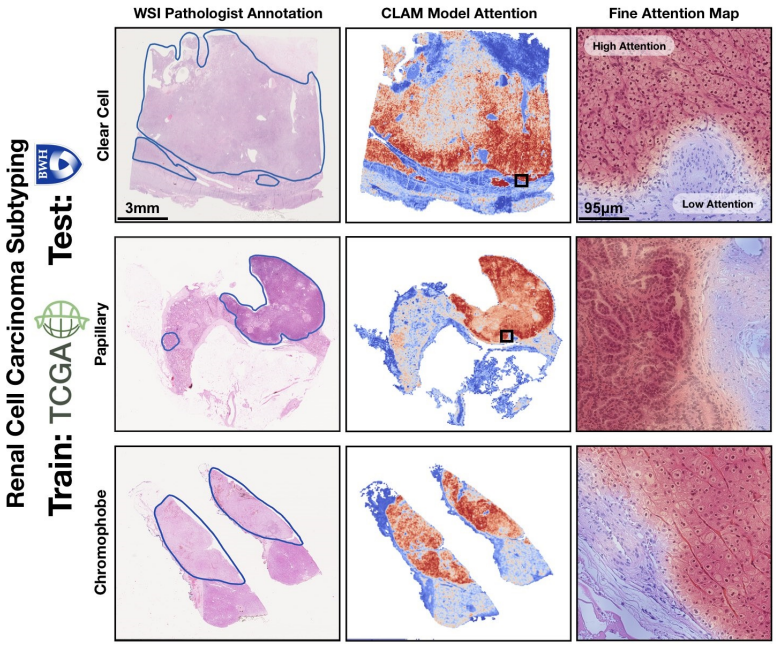


Figure 6.1: Visualization of the attention scores assigned by MIL model. Red color means high score whereas blue – low score meaning that the patch is not important from the model's point of view when solving the classification task. Source: Lu et al. (2021a).

assigning undue importance to such regions during decision-making. The second step is patching, which is performed on segmented images. Most commonly, a non-overlapping variant of patching is applied. Subsequently, features are independently extracted from the patches. Originally, it was done using ResNet-50 pretrained on Imagenet. Note that no information about positional encoding is provided. Afterwards, CLAM uses a gated attention-based MIL architecture to assign a weight (attention score) to feature vector of each patch. These are used to calculate the weighted sum of patch features to form a slide-level representation. The weights reflect the importance of the patches from the DL model's point of view.

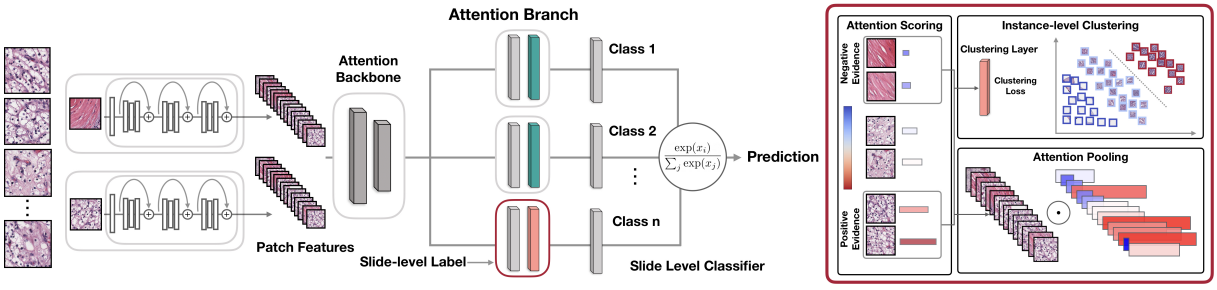


Figure 6.2: The CLAM model architecture (multi-branch). Source: Lu et al. (2021a).

The authors proposed two variants of the attention mechanism in the multiclass setup: single- and multi- branch. In the single-branch variant, one attention score is calculated per patch according to Equation 6.2 regardless of the number of possible output classes. This results in one global representation of the WSI that is used for final tissue classifi-

cation (multi-class). In the multi-branch setup, the set of attention scores (Equation 6.2) is calculated separately with respect to each class. Therefore, there are as many WSI representations computed as weighted sum of patch features and attention scores as the number of possible classes. In the multi-branch case, there is an independent binary classifier on top of each of the representations which decides if a particular class is proper or not. Lastly, on top of all of the binary predictions, the softmax is applied to output one final prediction (class).

To improve the learning of informative regions, CLAM introduces a clustering constraint. In the case of binary WSI classification, the process is as follows. Patch features corresponding to B highest attention scores and to B smallest attention scores are selected. If the model assigns high attention scores, it means that the model perceives a patch as important, whereas low attention scores mean that the model ignores the particular patch. In the former case, a positive pseudo-label is assigned to patches, whereas in the latter a negative pseudo-label. These $2B$ patch features are input to a one-layer MLP-based classifier whose task is to predict the pseudo-labels. Note that we use the term 'pseudo-labels' instead of labels as there are no fine-grained patch-level annotations of their importance. Instead, these 'important'/'non-important' classes are based on the output of the attention mechanism that is updated during model training.

In the multi-class setup, the clustering may be more sophisticated. In this case, we name the branch corresponding to the WSI ground-truth label as 'in-the-class' and all the other branches as 'out-of-the-class'. The first scenario is when there is a 'mutual exclusivity assumption' that representative patches from different classes cannot be present within one slide. It is often valid in the case of cancer subtyping problems. Then, the patch features corresponding to the B highest attention scores from the 'in-the-class' branch are assigned a positive pseudo-label whereas the negative pseudo-label is given to the B patches with lowest attention scores from the 'in-the-class' branch and B top-attended patches in all 'out-of-the-class' branches. However, in some cases the assumption of mutual exclusivity is violated. This occurs, for example, in the cancer vs. no-cancer task, where a slide may contain both tumor and normal tissue patches. In such cases, clustering supervision to highly attended patches from 'out-of-class' branches is not applied, as it is unclear whether they represent false positives.

The classification and clustering branches are trained jointly, where clustering can be treated as an auxiliary task. This differs from the multi-task setup, where the output from all of the tasks is important for further use. Here, the output of clustering is not overall important, it serves only as a way of improving the performance of the WSI classification. The total loss in CLAM training is defined as a weighted sum of two components: one for classification and the other for clustering. More emphasis is placed on the classification part. The clustering loss helps the model to better distinguish between discriminative and non-discriminative instances. It serves as additional supervision when

only one global (slide-level) label is available. In summary, CLAM combines instance-level attention, feature clustering, and slide-level classification to achieve high accuracy and interpretability in a weakly labeled data regime.

6.1.2 TransMIL

Unlike classical MIL models, TransMIL (Shao et al., 2021) (short for Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification) leverages transformer-based architecture to model spatial relationships among instances (patches). The authors introduce the term correlated Multiple Instance Learning. They replace the MIL attention pooling used i.e. in CLAM with self-attention pooling.

The TransMIL model consists of multiple steps (Figure 6.3). The first step is data preparation which is performed in the same manner as in CLAM that is segmentation, patching and feature extraction using pretrained model.

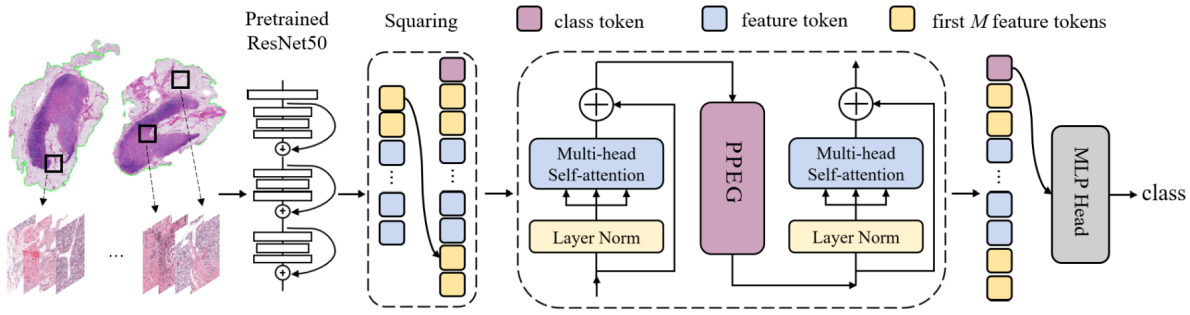


Figure 6.3: The TransMIL model architecture. Source: Shao et al. (2021).

Since the number of patches in WSIs may vary, the model standardizes them to fit a square layout. To do this, it first computes the nearest square number N that is greater than or equal to the number of WSI patches. Next, it calculates M, which is a difference between N and the actual number of patches in a WSI. Finally, it constructs a sequence (concatenation) of a special 'class token', all original patch features, and the features from the first M patches.

Afterwards, the Multi-head Self-Attention is applied to the constructed sequence to model correlations. To address the challenge of long instance sequences in WSIs, the softmax approximation from the Nyström method (Xiong et al., 2021) is employed. This reduces the computational complexity from the standard $O(n^2)$ to $O(n)$.

In the next step, the Pyramid Position Encoding Generator (PPEG) is run. It is designed to implicitly encode positional information. This is different than in the case of traditional Transformer architectures, such as Vision Transformer, where positional embeddings are explicitly added to the input. The PPEG module employs convolutional kernels of different sizes within the same layer, allowing it to implicitly encode positional information at multiple levels of granularity. This enriches the representation of each patch

with contextual information. Pyramid Position Encoding Generator works as follows. The sequence is first split into patch features and a class token. Patch features are then reshaped into the 2D grid, where group convolutions with kernel k and $\frac{k-1}{2}$ zero padding ($k = 3, 5, 7$) are applied. Later, all output feature maps are fused and added to features used as input to group convolution. Such patch features are flattened back into a sequence which is lastly concatenated with class token.

Subsequently, on top of the PPEG output, another Multi-head self-attention is applied to perform deep feature aggregation. Finally, classification is performed on the class token after applying Layer Normalization.

Chapter 7

P4: Does context matter in digital pathology?

7.1 Preface

When people analyse information without the proper amount of context, it often happens that they draw incorrect conclusions. The question arises: ***Does lack of small area (context) surrounding the region, which was used to define ground truth, result in the situations where Deep Learning models change their prediction compared to the situation when full context was provided?***
The short answer is 'yes'.

To investigate the issue, we propose the occlusion-based method, which is inspired by the Local Interpretable Model-agnostic Explanations (LIME) method (Ribeiro et al., 2016). The first step of the LIME method in the case of investigation of models processing images is to divide the input into superpixels. Later, many new perturbed image sampled are created by occluding different superpixels. The new image samples are the input to the original black-box model to obtain predictions. Such predictions are later used to train a simple interpretable model, typically linear regression, with additional weighting based on the similarity of perturbed samples to the original image. The trained model serves as a surrogate that approximates the complex model locally around the chosen instance. The coefficients of this simple model indicate which features (superpixels) influence the prediction the most.

Unlike in LIME, we do not focus in the paper on assigning importance scores to image regions but focus on stability of model predictions depending on the applied image perturbations. In our experiments, the scale of the perturbations is regulated by the size of the context. It is done by occluding some parts of the local context around the central tissue area. We provide the results of experiments showing that, indeed, models perform

worse when contextual information is limited. This work is positioned along the axes defined in Chapter 1 as shown in Table 7.1.

Table 7.1: Mapping of the paper by Tomaszewska et al. (2023) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska, Mateusz Sperkowski & Przemysław Biecek. (2023). "Does context matter in digital pathology?". Workshop on Modelling and Representing Context (MRC) at European Conference on Artificial Intelligence (ECAI).

7.2 Introduction

Deep Learning (DL) models are perceived as black-boxes. They sometimes make decisions based on different reasons than humans do. DL models tend to take shortcuts and follow spurious correlations (Geirhos et al., 2020). The popular example is the case where the model misclassified a husky as a wolf because there was snow in the background, which was a rule in the case of images with wolves within the dataset (Ribeiro et al., 2016). Although following such a rule may lead to high classification performance (if the dataset is biased), we expect the model to distinguish between wolves and husky dogs based on the animal features. Motivated by the fact that there is sometimes a mismatch between the way DL models and humans reason, we decided to investigate whether the DL models for vision follow the same good practices when diagnosing lesions based on histopathological data as expert histopathologists. The histopathologists when diagnosing a particular region of a lesion, take into account also the surrounding tissue (Zormpas-Petridis et al., 2019). We investigate whether the classification performance of DL models for vision will be higher when they have access to information about neighbouring tissues than in the case where no contextual information is given. We conduct a quantitative analysis on how the amount of contextual information within the input to the models impacts the final performance. Our contribution is as follows:

- we verify whether DL models for vision behave in a similar way to histopathologists who benefit from contextual information when diagnosing lesions,
- we measure quantitatively the impact of the amount of contextual information provided to different DL models for vision on their classification performance,

- we investigate whether it happens that the models behave in a non-stable way by changing predictions for the images many times given different amounts of contextual information.

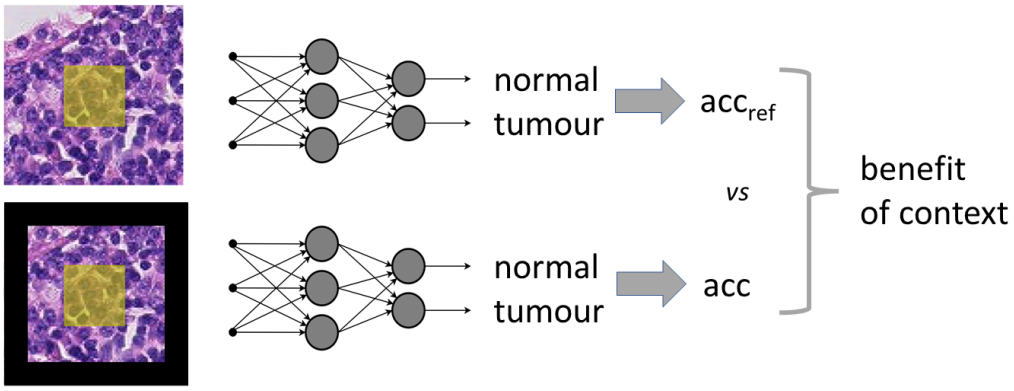


Figure 7.1: The scheme of the proposed study. The yellow squares in the center of histopathological images depict the regions that the annotations are based on (the squares are shown only for visualization purposes and are not present in dataset images). The black border is applied to remove some parts of contextual information.

The code to replicate our results is available at
`https://github.com/ptomaszewska/PCam_context`.

7.3 Motivation

The histopathological data is saved in the form of Whole Slide Images (WSIs) where the whole lesion is under huge resolution. In the popular Camelyon16 Breast Cancer dataset (Ehteshami Bejnordi et al., 2017) the images have an average resolution of 94,747x188,764 (avg. size of 1.97GB). Such huge images are difficult to load into memory and process, therefore they are most commonly split into smaller-sized patches. The advantage of WSIs is that based on them both local (cell-level) and global (tissue-level) analysis can be performed. In the work, we use the variant of the Camelyon16 dataset, PatchCamelyon (PCam) (Veeling et al., 2018), where the initial WSIs are cut into patches of size 96x96. Each such patch has a label (normal tissue or tumor lesion) and the model's task is a binary classification. However, in such a dataset, the global context is not preserved as the relationship between neighbouring patches is lost. Nevertheless, the dataset contains local context within images since the label is assigned to the whole patch only based on the central 32x32 region of the patch. Therefore, the surrounding box can be thought of as contextual information. The question is whether contextual information is useful when making predictions.

7.4 Method

The goal of the study is to check whether the DL models for vision are sensitive to different amounts of contextual information within input histopathological images. First, we use the test set of the PCam dataset as input for inference to the already trained models on the original PCam dataset. The resulting performance metrics serve as a reference point when the full available context is provided to the model (acc_{ref}, $precision_{ref}$, $recall_{ref}$, AUC_{ref}).

In the following experiments, we restrict the amount of contextual information in the image data. Let us define, the size of context (s) as a width in pixels of the area around the central 32x32 square. The maximum size of the context is $(96-32)/2 = 32$ where 96 is the length of the original image's side. The bigger the context size, the more information about the neighbouring tissue is provided to the model. To evaluate how the size of the context impacts the model prediction, we remove external layers of pixels of the context area and the image is padded to the original size of 96x96 with black pixels (we call it a border for brevity). We decided to use the black colour as it is often used as a baseline colour in explainable AI (XAI) methods i.e. Integrated Gradients (Sundararajan et al., 2017). We applied the border to the images to obscure some part of the context instead of cutting off the pixels and changing the image resolution to avoid a situation where it would be difficult to disentangle the source of the performance change – increased resolution vs. limited context size.

The images padded with a black border are the input to the DL models. We analyse the difference between the metrics obtained on padded images and the reference, original ones. This will give us information on how much contextual information is beneficial when making predictions.

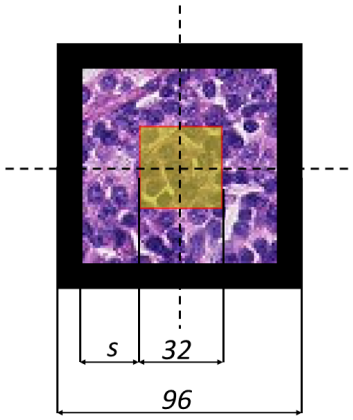


Figure 7.2: The histopathological image padded with black border with dimensions specified. The dimension s denotes context size.

7.5 Experiments

7.5.1 Deep Learning models

We apply the method described in Section 10.5 on two classes of DL models – convolutional (*ResNet18* and *DenseNet121*) and transformer-based (*Swin* (Liu et al., 2021) and *ViT* (Dosovitskiy et al., 2021)). The convolutional models trained on histopathological data (without any image standarization) were taken from Pocock et al. (2022). In the case of the transformer-based models, we took the models pretrained on Imagenet. The *Swin* model was pretrained in a supervised manner, whereas in the case of *ViT*, we used the models pretrained in different schemes: supervised (called *supViT* for brevity) and unsupervised (contrastive – *MoCo* (He et al., 2020b) and autoencoder-based – *MAE* (He et al., 2022)). We applied end-to-end finetuning using the whole PCam training set. As the pretrained transformer-based models operate on an input size of 224x224 and the images within PCam dataset are of size 96x96, we applied resizing (as it is done in (Jia et al., 2022)). We did not apply any standarization to keep the same preprocessing procedure as in the convolutional models. The hyperparameters such as base learning rate (lr_{base}) and weight decay (wd) were the same as the ones used by the authors of Jia et al. (2022) (details in Table 7.2). Compared to the paper, we decreased the batch size from 128 to 64 and shortened the training procedure. The base learning rate was linearly scaled using the formula $lr = lr_{base} * batch\_size/256$ (Goyal et al., 2018) to account for the changes. The models were finetuned for 5 epochs with Adam optimizer and with a linear learning rate scheduler increasing to the final lr value. We took the model from the epoch that resulted in the biggest accuracy on the validation set. The fact that a small number of epochs was enough to achieve satisfactory accuracy was due to the big size of the training set (262,144 samples).

Table 7.2: Hyperparameters used for finetuning of transformer-based models on PCam dataset.

model	$\boldsymbol{lr_{base}}$	$\boldsymbol{wd}$
Swin	0.0001	0.001
supViT	0.001	0.01
MAE	0.0001	0.0001
MoCo	0.0001	0.0001

7.6 Results

In the PCam dataset, there are duplicates of images due to the procedure used to create it. In all experiments, we use the test set after the removal of the duplicates. First, we perform an inference using a test set with original images (with full context size) using the

analysed models. As a result, the reference values of performance metrics are obtained (Table 7.3).

Table 7.3: Deep Learning models' performance when full available contextual information is given (so-called reference performance).

model	acc_{ref}	$precision_{ref}$	$recall_{ref}$	AUC_{ref}
ResNet18	0.8786	0.9396	0.7754	0.9477
DenseNet121	0.8941	0.9498	0.8029	0.9904
Swin	0.9172	0.9680	0.8405	0.9716
supViT	0.9160	0.9514	0.8537	0.9734
MAE	0.9173	0.9513	0.8568	0.9746
MoCo	0.9143	0.9617	0.8397	0.9708

The analysed models have similar reference performance. However, it is observed that transformer-based models perform slightly better than convolutional models. Note that the convolutional models have smaller capacity as they have much less tunable parameters in millions (*ResNet18* – 11.7, *DenseNet121* – 8) than transformer-based models (*Swin* – 86.7, *ViT* – 85.8), which can be a source of the difference.

7.6.1 Image dimensions mapping

As already mentioned transformer-based models operate on the resized PCam images (224x224) whereas the convolutional models on the original size of the images (96x96). Therefore, the maximum size of the context and the central square is increased. This mismatch is important when analysing the results on a common axis. Note that the mapping from the size of 96x96 to 224x224 is of factor 2.33 which makes the alignment difficult when trying to map the values of context size in the two scenarios. As a solution, we decrease the context size in the resized images by 7 pixels as opposed to the scenario with images of original size where it is done pixel by pixel. Cutting out the context by every 7^{th} pixel in the resized image translates to cutting every 3^{rd} pixel in the original image. As a consequence in the plots in the following sections, there will be more points on the curves referring to the convolutional models than the transformer-based ones. Note that the context size of 32 within the original images maps to the value of 74.67 within the resized images, therefore, between the 74^{th} and 75^{th} pixel (which we additionally analyse despite the policy to cut off the context by every 7^{th} pixel). To aggregate the results from the two context sizes, we apply a conservative approach. We take the score that is the furthest from the label meaning that (1) if the correct class is 1, we take the lowest probability from the two pixels (74^{th} and 75^{th}), (2) if the label is 0, we take the biggest probability. This is done under the assumption that the model outputs just one score, which is the probability of class 1.

7.6.2 Drop of performance with the decrease of context size

Having reference values of performance metrics, we limit the context size within the input images by applying a black border to hide parts of the contextual information. The performance metrics for different models are shown in Figure 7.3.

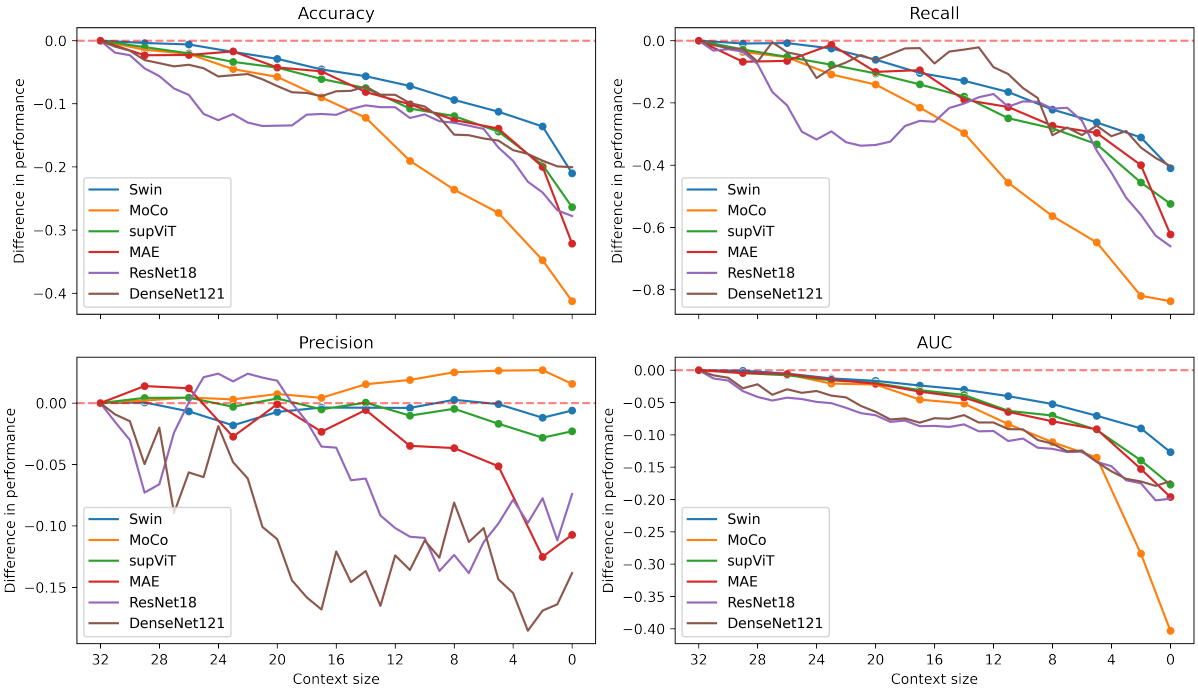


Figure 7.3: The performance gap when the context size is limited. The performance metrics of Deep Learning models, expressed as decreases from the corresponding reference values (with full context available), are shown for different context sizes. Note that the values on the x-axis are in decreasing order which makes an interpretation of the experiments easier. The y-axis is not shared within the subplots so that the variations of results for different models are more visible. The markers on the curves corresponding to transformer-based models highlight a smaller number of data points than in the case of convolutional models.

It is observed that the metric that is the most affected by limiting the amount of contextual information available at the prediction time is *recall*. This metric is especially important in healthcare where false negatives are of the greatest concern. The biggest drop in *recall* (0.84) occurs in the case of *MoCo* model which starts to predict mostly one class – normal. At the same time, in the case of this model, the *precision* curve slightly increases with the decrease of context size (up to 0.016 when $context\_size = 0$) even though the reference value was already very high (0.9617). The two models that experience the smallest drop of *recall* are *Swin* (0.41) and *DenseNet121* (0.40). The values in brackets are the ones corresponding to scenario with no contextual information within input.

Note that *precision* seems to be the least impacted metric by the limitation of contextual information. The most considerable decrease is observed in the case of *DenseNet121*

(0.14) and *MAE* (0.11), however, in the first case, the significant drop is observed even when the amount of context is only slightly limited.

In the case of *accuracy* and *AUC* plots, all the models behave similarly except *MoCo* which has the biggest drop of about 0.41. Interestingly, the *accuracy*, *recall* and *precision* curves of *ResNet18* experience significant fluctuation over different context sizes. In the case of *accuracy* and *recall*, local minimum is observed for a context size of about 24, whereas the local maximum at 8.

Note that by the fact that fewer data points are depicted in the case of transformer-based models than in the convolutional models, the smoothness of the curves cannot be compared between the two families of models, unlike the general trends. Overall, it seems that the analysed convolutional models are more sensitive to the lack of contextual information than the transformer-based models (except *MoCo*). However, the differences do not seem significant taking into account the gap in the models' capacity.

7.6.3 Consistency of predictions over different context sizes

Despite the changes in models' performance, the predictions for some images are the same over different context sizes. It turns out that the model that generates the consistent predictions most often is *Swin* and at the second place – *supViT* (see Table 7.4). The rest of the models are far behind when it comes to the percentages of images with agreeing predictions. The most sensitive to the decrease of the context size is *DenseNet121*.

Table 7.4: Percentage of images that get the same prediction regardless of the context size.

ResNet18	*DenseNet121*	*Swin*	*supViT*	*MAE*	*MoCo*
54.66	52.39	72.41	69.06	60.76	58.20

7.6.4 Misleading pieces of context

The decrease in performance results means that the models change their predictions for some images depending on the amount of contextual information. We distinguish two types of images (1) the ones that undergo the change of prediction given a particular model only once (2) the ones that experience the change of prediction more than once (referred to as *'swinging images'*). We present how the probability of class tumor changes depending on the context size for the sample images from the aforementioned two types in the case of *DenseNet121* (Figure 7.4). It is seen that the images that undergo frequent changes of the predictions are the ones that get the probabilities close to the threshold of 0.5. However, note that there is one sample (depicted in blue color) that the model was initially confident about (probability above 0.8) but with the decrease of the context size,

the probability went down and started to oscillate around the threshold. It can be seen that in the case of images that experience the change of prediction only once, it happens for different context sizes.

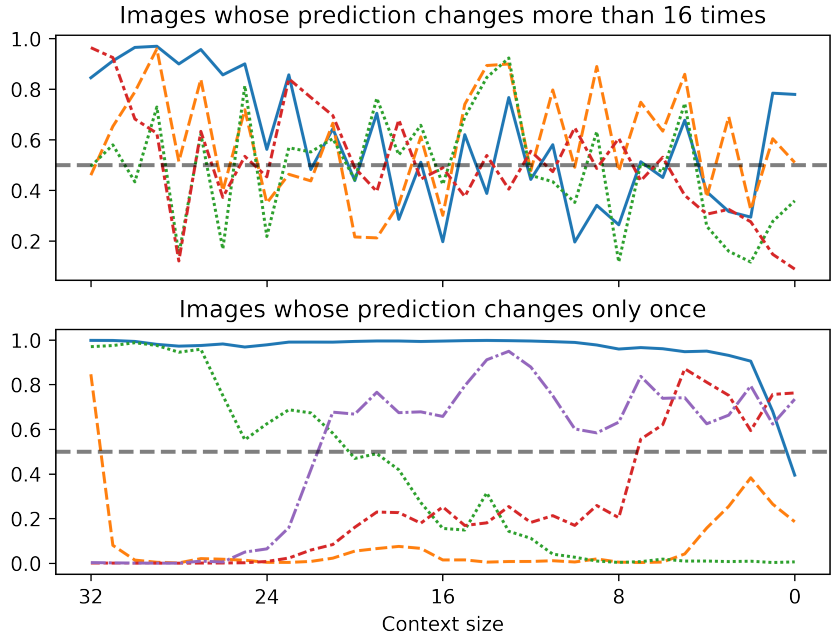


Figure 7.4: The misleading nature of context depending on its size (on the example of *DenseNet121*). The probability of the class tumour is shown in two cases: all images that experienced the change in the prediction in more than half of all context sizes (top), and a sample of images that changed the class only once (bottom).

The analysis of when (at what context size) the model changes predictions was performed. We analyse the number of images undergoing the particular direction of a change for the first time at the given context size. In the analysis, the *'swinging images'* are not included. We show results only for *DenseNet121* as for illustration (Figure 7.5). Interestingly, it can be seen that the changes of predictions occur mostly for the extreme values of context size – either the small or the big ones. This observation holds for other models except *ResNet18* where the shift TN (true negative) →FP (false positive) happens mostly for moderate values of context size, however, the scale of the shift is small as for the given moderate context size there are a maximum of 11 images that undergo a particular change of prediction.

Possible interpretation

The shift FP→TN could suggest that the model has seen some tumor cells in the context area as it predicted the class tumor but by limiting the amount of contextual information, we cut off the pixels containing these tumor cells and the model outputs the correct class 'normal'. Such behavior of the model could be understandable as the model was not explicitly said to base its opinion only on the central square.

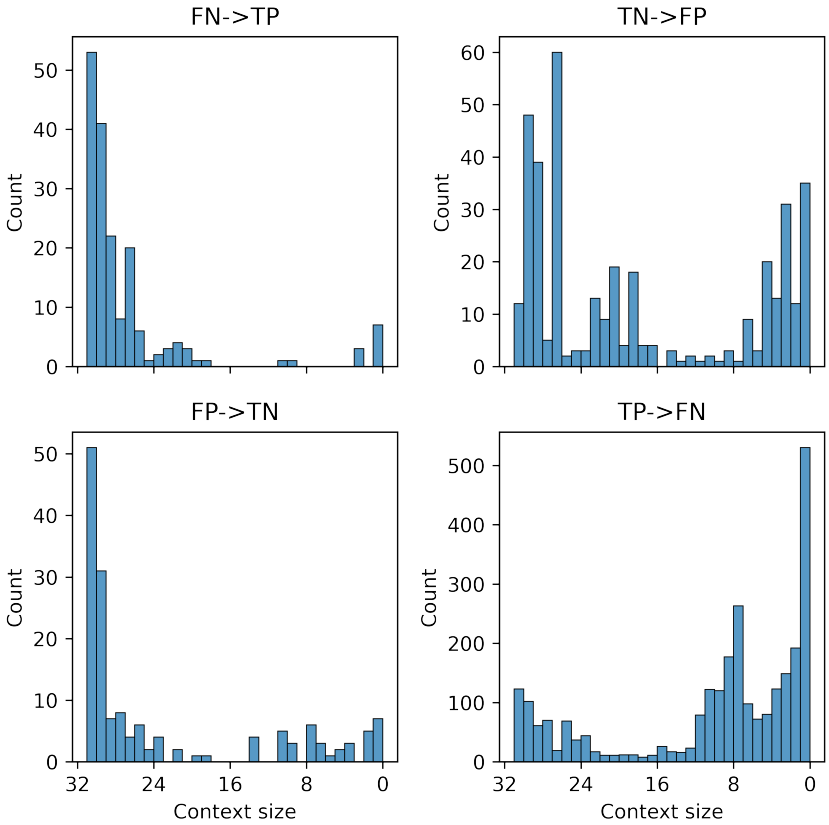


Figure 7.5: Number of images undergoing the change of prediction for the first time given the particular context size with the distinction on the initial and consecutive model prediction (*'swinging images'* not included). Note that for better visibility, the y-axis is not shared between subplots. The bin width is equal to 1. The results are provided for the *DenseNet121*.

For other shifts it is more difficult to find a reasonable explanation. For example, the shift FN→TP could mean that initially, the model focused too much on contextual information where there was only normal tissue. When this distraction in the form of contextual information was taken away, the model paid only attention to the key central part and spotted the tumor cells and as a consequence output the correct class. However, such a behavior of the model is not desirable.
In the aforementioned possible interpretations, we focused on the ones that do not require domain knowledge. However, the tissue structure and some spatial relationships that were partially covered by the black borders may potentially also play a role in the changes of model predictions.

Summarized model behavior

The summarized results from the histograms (without the distinction on the type of prediction shift) corresponding to different models are shown in Figure 7.6. It turns out that indeed (as shown in Figure 7.5) in the case of *DenseNet121*, the most changes occur

for extreme context sizes but it is even more visible in the case of *ResNet18*. In *supViT* and *MAE*, the biggest boost of the number of changes occurs when the context size is significantly reduced. The most changes in total are observed when *MoCo* is used.

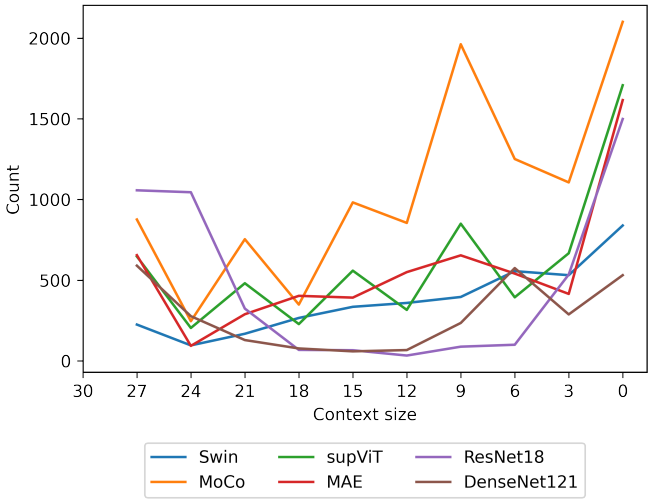


Figure 7.6: Number of images undergoing the change of prediction for the first time given the particular context size (*'swinging images'* not included). To account for the fact that in the case of transformer-based models, we cut the context size by every 3 pixels (after mapping the image size back to the original one) not by 1 pixel as in the case of convolutional models (see Section *Image dimensions mapping*), we summed the number of images undergoing the change from three consecutive context sizes. By the application of the aforementioned 'normalization', the curves of convolutional and transformer-based models are on the same scale and have the same number of data points).

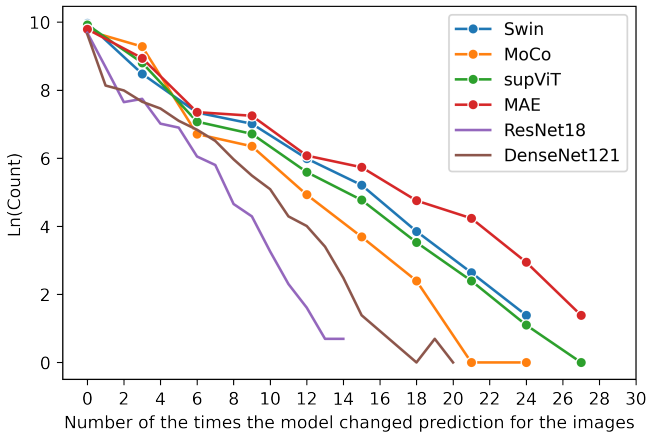


Figure 7.7: The logarithm of the number of images that experience the particular number of prediction changes given a particular model. To account for the fact that the maximum possible number of prediction changes in the case of transformer-based models was 10 and in the case of convolutional models – 31, the values on x-axis are multiplied by the factor of 3 (for transformer-based models) to put them on the same scale as convolutional models allowing a fair comparison.

Lastly, it was analysed how many images in total experience a particular number of

prediction changes given a particular model (Figure 7.7). It turns out that the models that change predictions more than 24 times per image (out of 30 analysed possibilities, after rescaling explained in the Figure's caption) are *MAE* and *supViT*. However, these are very rare in the case of *MAE* where there are only four such images, and in the case of *supViT* (only one). Note that the area under the curve metric cannot be compared between the models from the convolutional and transformer families as in the latter there are fewer data points. However, when comparing *DenseNet121* and *ResNet18*, it is visible that for the *'most swinging'* images, the changes of predictions are more frequent in the case of *DenseNet121* than *ResNet18*. The four *'most swinging'* images given *DenseNet121* are shown in Figure 7.4. The most similar models in behavior are *supViT* and *Swin* where the relationship between the logarithm of the number of images undergoing the particular number of changes and the number of changes is almost linear.

Models' agreement

It is analysed whether the same images are confusing to the models when the context size is limited. We investigate how many *'swinging images'* are in common for any pair of models regardless of context size when the changes of predictions occur. We analyse transformer-based models and convolutional ones separately. It turns out that the biggest agreement between the transformer-based models is in the case of *Swin* and *MAE* (1176 cases) whereas the smallest agreement is between *supViT* and *MoCo* (502 cases). In the case of *ResNet18* and *DenseNet121*, there are 4822 *'swinging images'* in common, therefore, they seem much more alike than transformer-based models even though three out of four analysed transformer-based models have the same architecture (ViT) but differ in pretraining scheme.

7.7 Conclusions

In the work, we investigate whether the Deep Learning models for vision are sensitive to contextual information when making predictions on histopathological data. It turns out that when the context size is limited, the models achieve worse performance than in the case when full context is available which means that context is important at prediction time. It is observed that depending on the amount of contextual information, the model can output different predictions for a given image. We evaluate the behavior of models that have similar reference performance metrics (when full access to a context is provided) in the case when the size of contextual information is decreased. It turns out that the model that is the most sensitive to the limitation of context size is *MoCo*. It may possibly be attributed to the fact that the model was pretrained in a contrastive way but it requires further investigation. We observe that there are images of two types – the ones that

undergo one change of prediction and the *'swinging images'*. For the latter, it would be interesting to consult the images with a histopathologist to verify whether indeed in these cases, the context may be misleading. Moreover, in the future, the possible interpretations of the obtained results could be complemented by the analysis with the use of heatmap-based XAI techniques.

Chapter 8

P5: From Input to Insight: Probing the Reasoning of Attention-based MIL Models

8.1 Preface

When analysing a phenomenon based on results of experiments, one of the first steps is often the removal of noise that can distract from the core of the topic. The question arises: ***Can a similar approach to investigation of the capabilities of MIL models in understanding spatial context, provide valuable insights when combined with a fine-grained evaluation protocol?***
The short answer is 'yes'.

The idea of limiting distractors in the input signals, so that the model focuses solely on the key information, was implemented in the proposed set of synthetic datasets (SyntheticSMIL). Combined with the evaluation protocol (ReasonSMIL), this setup enabled the investigation of Multiple Instance Learning (MIL) models in terms of their ability to understand global context, defined here as the compositionality of large images. This work is positioned along the axes defined in Chapter 1 as shown in Table 8.1.

The need for the analysis of MIL models from a spatial context point of view is rooted in the fact that spatial context is important in digital pathology where these models often find application. Histopathologists analyse a whole tissue structure with respect to different types of cells etc. (Kumar et al., 2017). A notable example when spatial context plays crucial role is tumor budding. A bud is a single tumor cell or a cell cluster consisting of four tumor cells or less which is detached from the main tumor region. Therefore, the identification of tumor budding in H&E slides involves multiple steps: (1) localizing the main tumor region, (2) detecting detached tumor cells, (3) assessing the size of identified

regions from previous step, (4) discarding the identified region, if the criteria are not met, otherwise counting the number of buds. The number of buds translates to the risk stratification. The bigger, the number of buds, the worse prognosis and poorer chances for recovery.

The tumor budding is an important factor analysed in the case of various types of tumor such as colorectal (Lugli et al., 2017), bladder (Fukumoto et al., 2016) and breast (Okcu et al., 2021). For instance, in the former case, tumor budding is an independent predictor of lymph node metastases and well as of the survival in stage II (Lugli et al., 2017). As a consequence, it is suggested that the tumor budding should be taken into account along with other clinicopathological features in a multidisciplinary setting. This underscores the clinical significance of tissue characteristics such as the size and number of particular cells.

Following the fact that the number of high density points (like in tumor budding) and differences in cell shapes matter when analysing tissue, we propose a set of synthetic datasets to investigate MIL models. Our approach to simplifying real-world images is analogous to the strategies used by Lucieri et al. (2020), where dermatoscopic images are simplified, and by Bodnár et al. (2024), where similar simplification is applied to facilitate product quality assessment. Bodnár et al. (2024) highlight that when visually inspecting microscopic cross-section images of solder joints, it is important to assess the spatial arrangement of anomalies (voids and cracks). It is analysed whether there is just a single void or whether there are two types of anomalies (voids and cracks) in a particular relationship (close, far, overlap). These scenarios are mimicked with the use of a dataset with simple shapes like triangles, circles and rectangles in various spatial relationships.

The set of datasets that we prepared are later used in paper described in Chapter 10.

Following the medical motivation, the goal of the paper presented below was to analyse the attention-based MIL models, commonly used in digital pathology, from the spatial point of view. The aim was to check if the features such as size and number of elements are important in the model's decision-making process. Note that despite critical applications of MIL models in healthcare, they have been poorly investigated with regard to their properties, vulnerabilities, and reasoning. To address this research gap, we propose rule-based synthetic datasets SyntheticSMIL, and ReasonSMIL protocol, for the investigation of the attention-based MIL models.

The mathematical formulation of the attention mechanism (described in Chapter 6) allow models to pay attention to multiple areas (spatial context), however, from the model point of view it would be also sufficient to pay all the attention to just one patch as long as it gives correct prediction (shortcut learning). This can be attributed to the formulation of the cost function applied during the training of such models that focuses on the correct classification of the tissue and not necessarily about the correct intermediate arguments.

In this aspect, CLAM (Lu et al., 2021a) and TransMIL (Shao et al., 2021) models, that are investigated in the study, differ. CLAM includes an auxiliary branch (clustering) that examines the attention scores assigned by the model, whereas TransMIL relies solely on a single component in the loss function which is cross-entropy loss with respect to tissue classification. In CLAM, the auxiliary loss contributes to the total loss function with a weight of 0.3.

Attention scores are frequently used as proxies for interpretability in neural networks, but their reliability, particularly in NLP, remains contested. In works by Serrano and Smith (2019); Bai et al. (2021), it is shown that attention weights can often be altered or nullified without affecting model predictions, calling into question their explanatory value. However, later study (Bibal et al., 2022) highlights that attention's interpretability may depend on the task. For example, Clark et al. (2019) and Vig and Belinkov (2019) find that attention in models like BERT and GPT-2 captures syntactic structure effectively. In work by Brunner et al. (2020), there is a theoretical analysis, showing attention can separate useful signal from noise. There were later the follow-up works (Kobayashi et al., 2020; Sun and Marasović, 2021) showing that alternative attention distributions obtained through adversarial training perform poorly, suggesting that the attention mechanism of recurrent neural networks indeed learns something useful. For detailed description of the ongoing debate on the issue of interpretability of attention scores, consult a paper by Lopardo et al. (2024). To conclude, there is still no definitive theoretical consensus on whether attention scores provide faithful explanations.

Motivated by these findings and the debate, I decided to investigate this issue in regards to models used in digital pathology, where attention score visualizations, such as heatmaps highlighting patches in Whole Slide Images (WSIs), are commonly presented in publications as a way to evaluate model behavior and foster trust.

The model internal values such as output of particular model layers and attention scores are often analysed in Mechanistic Interpretability field. There is, however, often the question from which layer the values should be extracted. In the case of the investigation of Large Language Models, the common heuristic is to pick values from one of the middle layers. There are also more methodological approaches on how to choose the optimal point in the neural network to extract the activations (e.g. Arditi et al. (2024)). However, in the case of Multiple Instance Learning models, there is no need for sophisticated selection method, as the attention mechanism is present only in one or maximum two places in the network. This is a large difference compared to Large Language Models (LLMs) where it is present in each transformer block. For instance, in relatively small LLM called Llama 3 8B (Grattafiori et al., 2024), there are 32 layers (transformers blocks).

Table 8.1: Mapping of the paper by Tomaszewska et al. (2025a) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska*, Michał Gozdera*, Elżbieta Sienkiewicz, & Przemysław Biecek. (2025). "From Input to Insight: Probing the Reasoning of Attention-based MIL Models". Proceedings of the World Conference on eXplainable Artificial Intelligence.

* – equal contribution

8.2 Introduction

By 2030, most data used in AI is expected to be artificially generated through rules, simulations, or models (Leinar Ramos and Jitendra Subramanyam, 2024). While the focus is typically on generating realistic data (Bauer et al., 2024), an alternative approach is to create simplified datasets to shed some light on model behavior. Such datasets allow for precise evaluation of whether models are learning the correct relationships. Though simplified, these datasets are crucial for testing model reasoning, particularly in complex tasks like medical imaging. By using clearly defined rules, they enable targeted investigations, help identify model weaknesses, and ensure models are reasoning correctly rather than relying on shortcuts (Geirhos et al., 2020; Dagaev et al., 2023; Wu et al., 2023). In this work, the term 'reasoning' is used in a distinct sense compared to its use in Large Language Models.

Our approach is a reverse of *'garbage in, garbage out'* in the form of *'understandable input, understandable output'* – making synthetic datasets an effective tool in explainable AI (XAI) research and for early-stage model development, ensuring transparency, scalability, and reproducibility before deployment.

In our work, we focus on attention-based Multiple Instance Learning (MIL) models (Ilse et al., 2018) which are often applied to large, high-resolution images with single global label, such as Whole Slide Images (WSIs) in pathology. Such medical application makes them an important area for research investigation. These models aggregate information extracted from image patches to make an overall prediction. Note that in real high-resolution images, more than one patch can be crucial in the decision-making process.

Our goal is to quantitatively investigate if the models truly capture all important

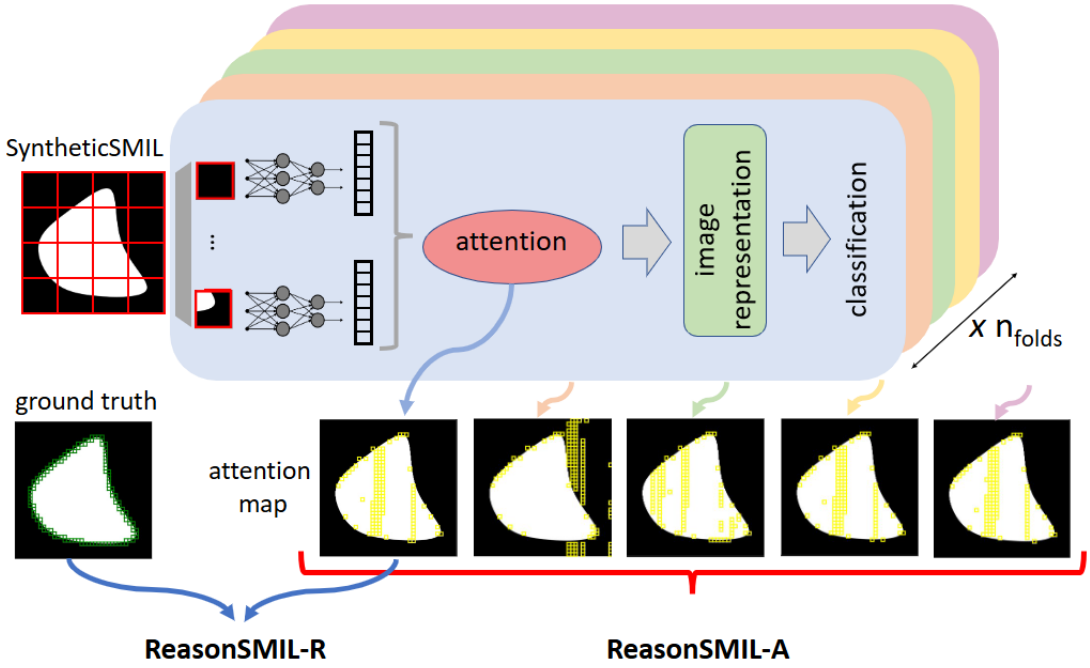


Figure 8.1: The proposed ReasonSMIL protocol evaluates the reasoning ability of attention-based MIL models trained e.g. on the SyntheticSMIL dataset. ReasonSMIL-R compares the model's patchwise attention scores against ground truth, while ReasonSMIL-A checks the agreement between attention maps from models trained on different cross-validation folds. Ground truth and model-identified important patches are highlighted in green and yellow, respectively.

patches instead of relying on just a few, i.e. shortcuts. The assumption that each image can contain multiple important patches leads to the concept of spatial context – relationships between the patches in the space, which is in line with a notion of spatial XAI (Tomaszewska and Biecek, 2024). Our proposed pipeline is shown in Figure 11.1. We address the need to investigate the behavior of the attention-based MIL models in regards to their spatial reasoning capabilities (SMIL) which are important e.g. in the analysis of histopathological images (Abousamra et al., 2023; Dee et al., 2023). We propose a term Spatial Multiple Instance Learning (SMIL). It is an extension of traditional Multiple Instance Learning (MIL) models by incorporating spatial context and relationships between instances (patches) in a bag. This makes it a powerful approach for tasks that require understanding of spatial patterns, such as in medical imaging. The notion is inspired by 'correlated MIL' (Shao et al., 2021).

Our main contributions are: (1) a set of configurable datasets (SyntheticSMIL) where multiple patches are essential for correct classification (spatial context), (2) a fine-grained, multi-criteria evaluation protocol to analyze if MIL models reason correctly (ReasonSMIL-R) and are stable within cross-validation folds (ReasonSMIL-A), (3) an illustrative investigation of the popular CLAM and TransMIL models using the proposed datasets and protocol.

The code is provided at `https://github.com/gozderam/synth_reason_smil.git`.

8.3 Related work

Multiple Instance Learning In Multiple Instance Learning (MIL), images of extremely high resolution with only global labels (weakly annotated) are processed. This approach is widely used in Whole Slide Images in digital pathology. Such images are analysed as bags of instances (here understood as patches). Each patch can be either with tumour cells or without. One patch with a tumour is enough to label the whole lesion as a tumour. In MIL, the samples (patches) from images are often first analysed individually and later the information from them is aggregated within attention-based modules to make a final decision about a bag. Common attention-based MIL models used for WSI analysis include CLAM (Lu et al., 2021a), TransMIL (Shao et al., 2021), ProtoMIL (Rymarczyk et al., 2023).

Our work focuses on analysing the reasoning of the attention-based MIL vision models, used for instance, in digital pathology for Whole Slide Images. Typically, such models are evaluated by global performance metrics (e.g., accuracy, AUC) and visualizations of attention scores via heatmaps. However, spatial context understanding through attention scores has rarely been studied, with only a few exceptions (Hense et al., 2024; Tomaszewska et al., 2025b). In the first, the coexistence of some instances (MNIST images) within bags is evaluated (only even/odd numbers in a bag or only adjacent pairs of numbers in a bag). In the latter, the spatial regression is used to quantify the role of context in a form of neighborhood in WSIs.

Model evaluation on synthetic data Synthetic datasets have been used in machine learning mainly to address data sensitivity and privacy issues (Rubin, 1993). The rise of Generative Adversarial Networks (Goodfellow et al., 2014) and diffusion models (Ho et al., 2020) has further popularized the use of synthetic datasets, including text-to-image generation (Kandwal and Nehra, 2024) and 3D environments for dataset sampling (Eftekhar et al., 2021; Martinez-Gonzalez et al., 2021). These methods aim to create realistic images for training models, ensuring good performance in real-world scenarios.

However, synthetic datasets can also be used to evaluate models by testing hypotheses, such as understanding spatial relationships (Swingler and Bath, 2020), or vulnerability to adversarial attacks (Costa et al., 2024). Such datasets can also serve for explainable AI (XAI) study (Tjoa and Guan, 2020; Arras et al., 2022). The MIL models were challenged under the adversarial attack (Yao, 2022; Zhang et al., 2023) and other synthetic tile modifications (Albuquerque et al., 2024). Yet, no rule-based datasets were used to investigate the MIL model's reasoning, vulnerabilities or properties.

The model testing using rule-based datasets in our work is inspired by the notion of

unit tests for symbolic conceptual reasoning in Deep Learning using synthetic datasets (Lovering and Pavlick, 2022) that are based on the theory of concepts by (Fodor, 1998) where one of the arguments is a significant role of compositionality in reasoning. The ability of abstract reasoning is also evaluated in visual IQ tests i.e. based on Ravens Progressive Matrices (Raven and Court, 1998).

8.4 SyntheticSMIL datasets

We propose a set of three synthetic datasets, SyntheticSMIL, to test MIL models' ability to reason using spatial context, where distinguishing classes depends on multiple patches. MIL is often used to analyse images of non-standardized size larger than typical 224×224, with weak labels, and valuable role of spatial relationships e.g. WSIs in digital pathology.

The SyntheticSMIL datasets were inspired by histopathological images that are input to the analysed MIL models. Key features in histopathological images are shapes, textures, and densities (Kumar et al., 2017). Their analysis is critical for detecting abnormalities and classifying tissue samples: (1) shape — irregular or enlarged nuclei often indicate malignancy, (2) texture — coarser chromatin texture is a common marker in cancerous nuclei, (3) density — increased cellular density typically signals high-grade tumors. We wanted to create datasets checking one data characteristic to make models' explanations as firm and straightforward as possible. Hence, to isolate the understanding of spatial context, we designed datasets with pure signal, free from distractors, ensuring straightforward interpretation without the need for disentanglement.

Inspired by configurable 3D environments (Eftekhar et al., 2021; Martinez-Gonzalez et al., 2021), the datasets allow flexible adjustments thanks to parametrization. The SyntheticSMIL implementation enables the generation of images of different sizes (Figure 9.2). We describe the generation process of the synthetic binary datasets, their configurable parameters, and methods to define important patches (ground truth). The goal is to create a flexible, diverse dataset generation framework with adjustable difficulty levels. The key values of the parameters used for the generation of datasets for the experiments are provided in Section 8.6.

Concavity The *concavity* dataset focuses on differences among concave shapes. The task is to distinguish structures based on the ratio of a given structure's area to its convex hull area. Specifically, structures with a ratio below a given threshold (*'more' concave*) must be separated from those with a ratio above it (*'less' concave*).

In Algorithm 2:

- `gen_points`$(n_{pmin}, n_{pmax}, d_m, d_s, d_{var})$ generates base points for spline creation. The number of points n_p is an integer sampled from $\{n_{pmin}, n_{pmin}+1, \ldots, n_{pmax}-1\}$

Algorithm 2 *Concavity* dataset generation (single image).

$a_f \leftarrow -1$
while $a_f \notin [a_{f_{min}}, a_{f_{max}}]$ **do**
  $b_{pts} \leftarrow$ `gen_points`$(n_{pmin}, n_{pmax}, d_m, d_s, d_{var})$
  $s_{pts} \leftarrow$ `get_spline_points`(b_{pts})
  $c_{hull} \leftarrow$ `convex_hull`(s_{pts})
  $a_f \leftarrow$ `pol_area`(s_{pts}) / `pol_area`(c_{hull})
end while
$pts_{scaled} \leftarrow$ `scale`$(s_{pts}, sf_{min}, sf_{max})$
`draw_polygon`(pts_{scaled})

with equal probabilities for each value. Later, for each point i, $i \in \{0, 1, \ldots, n_p - 1\}$ a distance d_p (from the origin of the coordinate system) is sampled from a normal distribution $d_i \sim \mathcal{N}(d_m, d_s)$. Therefore, each point i can be placed on a circle with a radius d_i. The points are placed on the corresponding circles so that the angle between the y axis of the coordinate system and the line from the origin of the coordinate system to the point i is $rot_i = 2\pi i/n_p + shift_i \pi/n_p$ where $shift_i \sim \mathcal{U}(0, 1)$ is a noise factor.

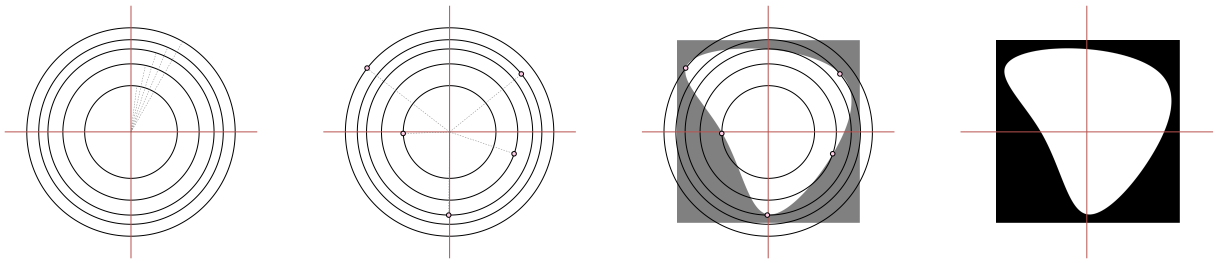
- `get_spline_points`(b_{pts}) generates the B-spline based on points b_{pts}.
- `convex_hull`(s_{pts}) returns a list of points that are the convex hull of points s_{pts}.
- `pol_area(*)` returns the areas of polygons formed by s_{pts} and c_{hull} respectively.
- Generation algorithm keeps repeating as long as the resulting area ratio a_f (computed for a generated shape) does not fall into the range $[a_{f_{min}}, a_{f_{max}}]$.
- `scale`$(s_{pts}, sf_{min}, sf_{max})$ scales generated points with a scaling factor $sf \sim \mathcal{U}(sf_{min}, sf_{max})$.

The graphical explanation of Algorithm 2 is shown in Figure 8.2.

The classes differ in the allowed values of a_f - the ratio of the polygon area to its convex hull area. For class 1, the allowed values ($[a_{f_{min_1}}, a_{f_{max_1}}]$) are lower than for class 0 ($[a_{f_{min_0}}, a_{f_{max_0}}]$). The ranges for different classes are disjoint - *'more'* concave belong to class 1, whereas the *'less'* concave to class 0. The smaller and closer a_f ranges for classes, the more difficult to classify a dataset is.

In *concavity*, key patches (ground truth, GT) for class distinction can be defined in three ways: (1) background only, (2) foreground – polygon area (without the edge), and (3) polygon edge (Figure 8.3d). Each method provides enough information for classification.

Let p_{ij} be the patch of a given image, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, M\}$, and pts_{scaled} - as in Algorithm 1 (main paper). The procedure of selecting ground-truth important patches in the *concavity* dataset is described in Algorithm 4.



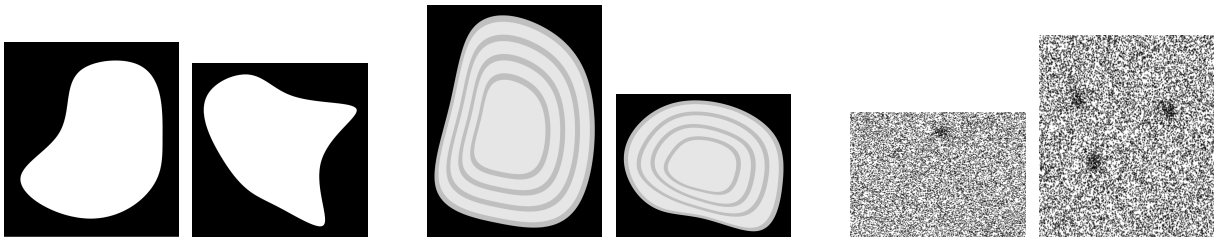
(a) Coaxial circles with random radius and centres in the origin of the coordinate system.

(b) Spline base points placed with random angles on each circle.

(c) Generated shape based on spline base points.

(d) Final image in the coordinate system.

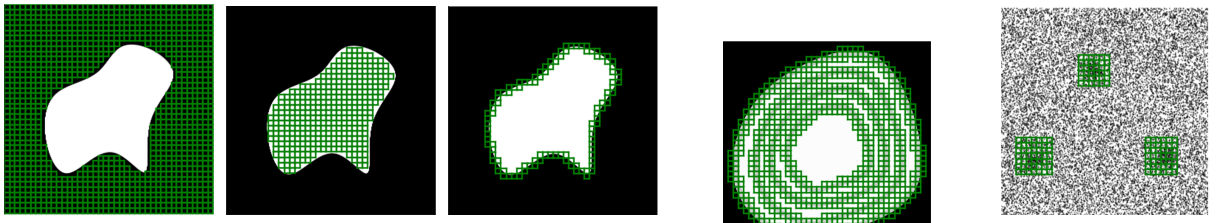
Figure 8.2: *Concavity* images generation process.



(a) *concavity* – 'less' vs. 'more' concave

(b) *onion* – same vs. different width

(c) *points density* – N_{hd_0} vs. N_{hd_1}



(d) *concavity* GT – variants: 'background', 'foreground', 'edge'

(e) *onion* GT

(f) *points density* GT

Figure 8.3: Samples from two different classes of the *concavity*, *onion*, and *points density* datasets (a–c), along with exemplary ground truth patches outlined (d–f). The size of images can be different – width and height can be sampled from normal distribution given mean and standard deviation.

Algorithm 4 Setting ground-truth for a given image from *concavity* dataset.

$pol \leftarrow \texttt{polygon}(pts_{scaled})$

$bg \leftarrow \texttt{background}(pts_{scaled})$

$gt_{patches_{bg}} \leftarrow \{p_{ij} | p_{ij} \cap pol = \emptyset\}$

$gt_{patches_{pol}} \leftarrow \{p_{ij} | p_{ij} \cap bg = \emptyset\}$

$gt_{patches_{edge}} \leftarrow \{p_{ij} | p_{ij} \cap pol \neq \emptyset \wedge p_{ij} \cap bg \neq \emptyset\}$

In Algorithm 4:

- `polygon(`pts_{scaled}`)` returns a region determined by pts_{scaled} (together with the edges).
- `background(`pts_{scaled}`)` returns the background of the region determined by pts_{scaled} (a complement of `polygon(`pts_{scaled}`)` in the context of the entire image).

Onion An image from the *onion* dataset consists of a series of nested, light-grey regions, referred to as *layers*, surrounding a central area. Each layer is enclosed by a dark-grey boundary (called *edge*). In class 0, the layers are coaxial and share the exact same shape, differing only in scale. In class 1, however, the shapes vary across layers.

Algorithm 5 *Onion* dataset generation (single image).

$edges \leftarrow \emptyset$
$e_{pts_0} \leftarrow s_{pts}$
for $i \in \{0, 1, \ldots, l_c - 1\}$ **do**
  $edges \leftarrow edges \cup$ `get_edge(`e_{pts_i}, e_w`)`
  $e_{pts_{i+1}} \leftarrow$ `transform(`$e_{pts_i}, l_w, noise_{std}$`)`
end for
`draw_layers(`$edges$`)`

In Algorithm 5:

- `get_edge(`e_{pts_i}, e_w`)` returns a set of points constituting an edge shape that is constructed based on e_{pts_i}. The edge has a width of e_w.
- `transform(`$e_{pts_i}, l_w, noise_{std}$`)` returns transformed points being the input points for the next (inner) edge. The transformation is based on layer width (l_w) and the normally distributed noise factor ns applied to points location, $ns \sim \mathcal{N}(1, noise_{std})$ ($ns = 1$ means no noise is applied).
- s_{pts} are generated as for *concavity* dataset (Algorithm 2).

The key parameter in the generation scheme is $noise_{std}$. When $noise_{std} = 0$, the noise is not applied to the subsequent layers and all layers on a single image are of the same shape disregarding a scale (class 0). With $noise_{std} \neq 0$, subsequent layers differ both in a scale and shape (class 1). The lower $noise_{std}$, the more difficult it is to properly classify images.

The important patches (ground truth) are all the patches that include pixels comprising layer edges (Figure 8.3e).

Onion Let p_{ij} be the patch of a given image, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, M\}$. The procedure of selecting important patches (ground-truth) in *onion* dataset is described in Algorithm 6.

Algorithm 6 Setting ground-truth for a given image from *onion* dataset.

$gt_{patches} \leftarrow \emptyset$
for $i \in \{1, 2, \ldots, N\}$ **do**
  for $j \in \{1, 2, \ldots, M\}$ **do**
    $n_{colors} \leftarrow$ `get_n_colors(`p_{ij}`)`
    if $n_{colors} > 1$ **then**
      $gt_{patches} \leftarrow gt_{patches} \cup \{p_{ij}\}$
    end if
  end for
end for

In Algorithm 6:

- `get_n_colors(`p_{ij}`)` returns the number of different colors (pixel RGB values) present in a patch p_{ij}.

Points density The *points density* dataset refers to images where the distinguishing factor is the number of areas with a high density of points (Figure 8.3c).

Algorithm 7 *Points density* dataset generation (single image).

$r \leftarrow$ `discrete_uniform(`$\{r_{min}, 2r_{min}, \ldots, wr_{min}\}$`)`
$pts_{bg} \leftarrow$ `get_background_points(`n_{bg}`)`
$centers_{hd} \leftarrow$ `get_high_density_centers(`N_{hd}`)`
$pts_{hd} \leftarrow \emptyset$
for $c_{hd} \in centers_{hd}$ **do**
  $pts_{c_{hd}} \leftarrow$ `get_hd_neigh(`c_{hd}`,` n_{hd}`,` r`)`
  $pts_{hd} \leftarrow pts_{hd} \cup pts_{c_{hd}}$
end for
`draw(`pts_{bg}`)`, `draw(`pts_{hd}`)`

In Algorithm 7:

- `get_background_points(`n_{bg}`)` returns n_{bg} points (x_{bg}, y_{bg}) uniformly distributed across the entire image, $x_{bg}, y_{bg} \sim \mathcal{U}(0, 1)$.
- `get_high_density_centers(`N_{hd}`)` returns N_{hd} points (x_{hdc}, y_{hdc}) that constitute the centers of areas with high densities and are uniformly distributed across the image without border parts, $x_{hdc}, y_{hdc} \sim \mathcal{U}(0.15, 0.85)$. A given minimal distance between centers $(c_{d_{min}})$ is preserved in order not to generate high-density areas in the same regions.

- `get_hd_neigh(`c_{hd}`,` n_{hd}`,` r`)` returns n_{hd} points (x_{hd}, y_{hd}) that define the dense area around point $c_{hd}(x_{hdc}, y_{hdc})$ and are normally distributed with scale r, $x_{hd} \sim \mathcal{N}(x_{hdc}, r)$, $y_{hd} \sim \mathcal{N}(y_{hdc}, r)$. For each image, r is randomly sampled from the discrete uniform distribution, $r \sim \mathcal{U}_d(\{r_{min}, 2r_{min}, \ldots, wr_{min}\})$ where w - maximal multiplicity parameter.

In the *points density* dataset, classes are distinguished by the number of dense areas (N_{hd}). The dataset difficulty can be controlled by adjusting n_{hd} – lower values increase difficulty.

The key patches (ground truth) for decision-making in *points density* dataset are the ones that cover regions of a high density (Figure 8.3f).

Let p_{ij} be the patch of a given image, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, M\}$, and $centers_{hd}$, $pts_{c_{hd}}$ - as in Algorithm 3 (main paper). The procedure of selecting important patches (ground-truth) in *points density* dataset is described in Algorithm 8.

Algorithm 8 Setting ground-truth for a given image from *points density* dataset.

$gt_{patches} \leftarrow \emptyset$
for $c_{hd} \in centers_{hd}$ **do**
  $x_{min}, x_{max} \leftarrow$ `min_x(`$pts_{c_{hd}}$`)`, `max_x(`$pts_{c_{hd}}$`)`
  $y_{min}, y_{max} \leftarrow$ `min_y(`$pts_{c_{hd}}$`)`, `max_y(`$pts_{c_{hd}}$`)`
  $center_{neigh} \leftarrow$ `rect(`$x_{min}, x_{max}, y_{min}, y_{max}$`)`
  $gt_{patches} \leftarrow gt_{patches} \cup \{p_{ij} | p_{ij} \cap center_{neigh} \neq \emptyset\}$
end for

In Algorithm 8:

- `min_x(`$pts_{c_{hd}}$`)`, `max_x(`$pts_{c_{hd}}$`)`, `min_y(`$pts_{c_{hd}}$`)`, `max_y(`$pts_{c_{hd}}$`)` return minimal and maximal x and y coordinates of a set of points $pts_{c_{hd}}$.

- `rect(`$x_{min}, x_{max}, y_{min}, y_{max}$`)` returns a rectangular area determined by vertical (x_{min}, x_{max}) and horizontal (y_{min}, y_{max}) lines.

8.5 ReasonSMIL protocol

We propose the ReasonSMIL evaluation protocol to investigate the attention-based MIL models in a fine-grained manner to understand the model's internal reasoning. The ReasonSMIL evaluation protocol has two components for the investigation of: (1) alignment between model reasoning and the important patches (ground truth) and (2) consistency across models.

8.5.1 Investigation of model reasoning

The goal of ReasonSMIL-R (ReasonSMIL-Reasoning) is to verify if the model's decision is based on the correct patches. Since the datasets are synthetic, we know the key rules for class distinction. The process involves (1) defining a ground truth (GT), (2) extracting attention scores per patch from the trained model for each image, (3) using evaluation metrics to compare attention scores with the GT, (4) averaging the metric values across all images.

Metrics We propose using diverse metrics from two domains for a comprehensive analysis – segmentation (Dice score) and classification (with binarization: balanced accuracy (BAC), Matthew's correlation; without binarization: AUC, PR AUC, Relevance Mass Accuracy (RMA) (Arras et al., 2022)). The Dice, BAC and Matthew's coefficient scores are applied after binarizing the attention scores, while metrics like AUC and PR AUC handle continuous values. BAC and Matthew's correlation are useful for imbalanced datasets. Here, a dataset is a set of patches within an image where in the case of weakly annotated data, a region of interest is often small.

Binarization To binarize the attention scores, we select the same number of top-scoring patches as there are important patches in the ground truth. Recall and precision are not used, as they yield the same values as the Dice score with this binarization method.

8.5.2 Investigation of models' agreement

The ReasonSMIL-Agreement (ReasonSMIL-A) compares models trained on different data subsets during cross-validation, focusing on internal reasoning rather than global metrics like accuracy or AUC. It analyzes the correspondence of attention scores, similar to inter-annotator agreement, using Fleiss' kappa (Fleiss, 1971) (0 = no agreement, 1 = full agreement) and RMSE for attention score maps.

8.6 Experiments

The ReasonSMIL protocol is versatile and applicable to any attention-based vision MIL model. We use CLAM and TransMIL to illustrate how ReasonSMIL and SyntheticSMIL can be employed to analyse a specific MIL architecture. Additionally, to demonstrate how these methods can facilitate the comparison of different architectures, we conduct a comparative analysis between CLAM and TransMIL.
For CLAM, we used default hyperparameters (except switching from Resnet-50 to Resnet-32 due to datasets simplicity). In TransMIL, we used the same patch features as in CLAM, with default parameters except for points density, where the learning rate was set to 1e-5.

We generated 3 versions of each of the datasets – *concavity*, *onion*, *points density*, for the experiments. The versions differ in difficulty level which was achieved by manipulation of dataset parameters: *concavity* (*easy*: $area_{fmin_0} = 0.90, area_{fmax_1} = 0.89, area_{fmin_1} = 0.80$, *medium*: $0.93, 0.92, 0.86$, *hard*: 0.97, 0.96, 0.94); *onion* (*easy*: $noise_{std} = 0.07$, *medium*: 0.06, *hard*: 0.05); *points density* (*easy*: $n_{hd} = 200$, *medium*: 110, *hard*: 90). All the values of the dataset parameters are specified in Table 8.3. Each image is of size 8960×8960 pixels (40×40 patches of size 224×224). We fixed the image size as one degree of freedom to focus on dataset difficulty. The statistics on the number of positive patches in ground truth is provided in Table 8.2. In total, the dataset used during 5-fold cross-validation contains 300 images and the test set (same for all folds) comprises 100 images. The number of samples in the datasets were inspired by the setup of the widely-used CAMELYON16 (Ehteshami Bejnordi et al., 2017) dataset in digital pathology research on MIL models which consists of 270 training and 130 test samples.

Table 8.2: The parameters of the SyntheticSMIL datasets used in the experiments. The values of the parameters in bold were modified to create datasets of different difficulty levels.

param	**value**
n_{pmin}	4
n_{pmax}	10
d_m	1.5
d_s	0.4
d_{var}	0.5
$a_{f_{max_0}}$	0.99
$\boldsymbol{a_{f_{min_0}}}$	0.97/0.93/0.90
$\boldsymbol{a_{f_{max_1}}}$	0.96/0.92/0.89
$\boldsymbol{a_{f_{min_1}}}$	0.94/0.86/0.80
sf_{min}	0.6
sf_{max}	1.0

(a) *concavity*

param	**value**
e_w	0.0555
l_w	0.1111
c	4
$\boldsymbol{noise_{std}}$	0.05/0.06/0.07

(b) *onion*

param	**value**
r_{min}	0.025
w	1
$c_{d_{min}}$	0.15
N_{hd_0}	1
N_{hd_1}	3
$\boldsymbol{n_{hd}}$	90/110/200
n_{bg}	20000

(c) *points density*

8.7 Results

8.7.1 CLAM

We provide the results (on test sets) of an illustrative investigation of the CLAM model trained on the proposed SyntheticSMIL datasets.

Global performance

The global classification performance of CLAM model trained on SyntheticSMIL datasets of various difficulty levels (defined by a set of parameters) are in Table 8.4. As expected,

Table 8.3: The number of ground truth patches (min, max, mean with standard deviation) across datasets of different difficulty levels. The whole size of the image within the datasets is 1600 patches (40×40).

dataset	level	variant	min	max	mean
concavity	hard	back	464	1282	$948_{\pm 202}$
	medium	back	445	1288	$1003_{\pm 177}$
	easy	back	672	1289	$1002_{\pm 175}$
	hard	fore	227	990	$536_{\pm 185}$
	medium	fore	220	1013	$483_{\pm 162}$
	easy	fore	221	794	$480_{\pm 159}$
	hard	edge	84	149	$116_{\pm 18}$
	medium	edge	84	148	$115_{\pm 17}$
	easy	edge	90	153	$117_{\pm 18}$
onion	hard	-	657	783	$732_{\pm 24}$
	medium	-	644	783	$730_{\pm 26}$
	easy	-	645	783	$728_{\pm 28}$
points density	hard	-	25	121	$70_{\pm 35}$
	medium	-	25	139	$74_{\pm 38}$
	easy	-	30	160	$86_{\pm 46}$

easier datasets yield higher global performance (accuracy and AUC), with models showing stability across cross-validation folds (small standard deviations). The exception is the onion dataset (*medium*), where the standard deviation of accuracy is 0.16. Next, we apply the fine-grained ReasonSMIL protocol for further investigation.

ReasonSMIL-R

The results of ReasonSMIL-R for the CLAM model are shown in Table 8.5. The highest values occur in the *concavity* dataset when the ground truth is 'foreground'. Lower values are found in the 'edge' variant of *concavity* and in *points density*, both of which have sparse ground truth. Negative Matthew's correlation coefficients appear only in *concavity* when the ground truth is 'background'.

Points density There is a clear link between higher global performance and higher fine-grained metrics, indicating that models are better at assigning correctly attention scores to important patches in easier datasets. However, in these easiest cases, despite near-maximal global performance, fine-grained metrics remain unsatisfactory, suggesting that models may focus on unimportant patches.

Onion For the most difficult onion dataset, the fine-grained metrics are significantly lower than for other variants (i.e. Dice score equals 0.04). However, in easier dataset variants, fine-grained metrics remain similar to each other despite varying global performance: in contrary to *points density* dataset. Therefore, here the reasoning metrics are

Table 8.4: The performance of the CLAM model on SyntheticSMIL datasets. Performance metrics are reported as mean and standard deviation. Hereinafter red color means the most difficult dataset/the worst metric value within a given dataset; green – the easiest dataset/the best metric value; yellow – in between.

dataset	level	accuracy	AUC
concavity	hard	$0.66_{\pm 0.03}$	$0.73_{\pm 0.03}$
	medium	$0.77_{\pm 0.02}$	$0.88_{\pm 0.01}$
	easy	$0.86_{\pm 0.02}$	$0.93_{\pm 0.01}$
onion	hard	$0.59_{\pm 0.04}$	$0.62_{\pm 0.03}$
	medium	$0.82_{\pm 0.16}$	$0.91_{\pm 0.06}$
	easy	$0.93_{\pm 0.03}$	$0.98_{\pm <0.01}$
points density	hard	$0.55_{\pm 0.05}$	$0.60_{\pm 0.02}$
	medium	$0.66_{\pm 0.05}$	$0.73_{\pm 0.01}$
	easy	$0.99_{\pm 0.02}$	$1.0_{\pm <0.01}$

not in line with global accuracy. This can be attributed to two aspects. Firstly, sometimes there are more and less important patches within ground truth. There are patches with a central density peak, as well as those scattered toward the periphery – potentially less relevant. However, some metrics in ReasonSMIL-R expect a binary decision if a patch is important or not. This could be mitigated by more selective choice of GT patches. Secondly, the feed-forward layers may make wrong predictions even though the assigned attention scores to patches are reasonable.

Concavity All models achieve very weak values in 'edge' variant in comparison with other GT variants. Adding to this very low standard deviations (stdevs), we can conclude that models tend to not choose 'edge' variant in its reasoning. For the two hardest versions of the dataset, we observe varying metric values and significant standard deviations for both the 'back' and 'fore' GT variants. Such high standard deviations suggest that different cross-validation models select different GT variants, which can be further analyzed using ReasonSMIL-A. Notably, for the 'back' variant, the best metric values are observed for the medium dataset variant, while for the 'fore' variant, the worst metrics appear in the medium variant. For the easiest dataset variant, we see low metric values for the 'back' variant and very high ones for 'fore', both with low standard deviations– leading to the conclusion that the reasoning of all cross-validation models aligns with the 'fore' GT variant. Additionally, there is no clear relationship between global accuracy and ReasonSMIL-R metrics for this dataset.

For datasets with one possible definition of ground-truth important patches (such as *points density*, *onion*), the low ReasonSMIL-R metric value means the model does not reason in a correct way. For datasets with multiple possible ground truth variants (*concavity*), ReasonSMIL-R analysis becomes more complex and insightful. Firstly, if a model for a given GT variant achieves a poor metric value, it does not necessarily mean

Table 8.5: ReasonSMIL-R on CLAM. In the first column, different datasets ('d') are pointed out ('c' – *concavity*, 'o' – *onion*, 'p' – *points density*). In the 'variant' column, different methods for defining the ground truth are indicated ('back' – background, 'fore' – foreground, 'edge').

d.	level	variant	PR AUC	AUC	BAC	Dice	M. coeff.	RMA
c.	hard	back	$0.69_{\pm 0.16}$	$0.31_{\pm 0.37}$	$0.42_{\pm 0.31}$	$0.51_{\pm 0.26}$	$-0.16_{\pm 0.62}$	$0.58_{\pm 0.03}$
	med.	back	$0.78_{\pm 0.18}$	$0.47_{\pm 0.45}$	$0.57_{\pm 0.36}$	$0.66_{\pm 0.29}$	$0.14_{\pm 0.73}$	$0.63_{\pm 0.04}$
	easy	back	$0.66_{\pm <0.01}$	$0.14_{\pm 0.01}$	$0.3_{\pm <0.01}$	$0.45_{\pm <0.01}$	$-0.40_{\pm 0.01}$	$0.60_{\pm 0.02}$
	hard	fore	$0.84_{\pm 0.27}$	$0.8_{\pm 0.40}$	$0.83_{\pm 0.32}$	$0.77_{\pm 0.41}$	$0.65_{\pm 0.64}$	$0.35_{\pm 0.03}$
	med.	fore	$0.69_{\pm 0.34}$	$0.62_{\pm 0.49}$	$0.69_{\pm 0.37}$	$0.57_{\pm 0.51}$	$0.37_{\pm 0.74}$	$0.30_{\pm 0.04}$
	easy	fore	$0.95_{\pm 0.01}$	$0.98_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.95_{\pm 0.01}$	$0.92_{\pm 0.02}$	$0.33_{\pm 0.02}$
	hard	edge	$0.21_{\pm 0.02}$	$0.21_{\pm 0.02}$	$0.56_{\pm 0.01}$	$0.18_{\pm 0.02}$	$0.11_{\pm 0.02}$	$0.07_{\pm <0.01}$
	med.	edge	$0.26_{\pm 0.06}$	$0.27_{\pm 0.04}$	$0.58_{\pm 0.03}$	$0.22_{\pm 0.06}$	$0.16_{\pm 0.07}$	$0.07_{\pm <0.01}$
	easy	edge	$0.24_{\pm 0.05}$	$0.27_{\pm 0.03}$	$0.57_{\pm 0.03}$	$0.2_{\pm 0.05}$	$0.14_{\pm 0.06}$	$0.07_{\pm <0.01}$
o.	hard	-	$0.46_{\pm 0.06}$	$0.46_{\pm 0.18}$	$0.52_{\pm 0.13}$	$0.48_{\pm 0.14}$	$0.04_{\pm 0.25}$	$0.50_{\pm 0.05}$
	med.	-	$0.53_{\pm 0.02}$	$0.63_{\pm 0.01}$	$0.57_{\pm <0.01}$	$0.53_{\pm <0.01}$	$0.14_{\pm <0.01}$	$0.58_{\pm 0.03}$
	easy	-	$0.53_{\pm 0.01}$	$0.63_{\pm <0.01}$	$0.57_{\pm <0.01}$	$0.53_{\pm <0.01}$	$0.14_{\pm <0.01}$	$0.60_{\pm <0.01}$
p.	hard	-	$0.06_{\pm <0.01}$	$0.57_{\pm 0.01}$	$0.52_{\pm <0.01}$	$0.08_{\pm 0.01}$	$0.03_{\pm 0.01}$	$0.04_{\pm <0.01}$
	med.	-	$0.11_{\pm 0.02}$	$0.6_{\pm 0.01}$	$0.55_{\pm 0.01}$	$0.13_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.05_{\pm <0.01}$
	easy	-	$0.29_{\pm <0.01}$	$0.66_{\pm 0.01}$	$0.62_{\pm <0.01}$	$0.28_{\pm <0.01}$	$0.24_{\pm <0.01}$	$0.07_{\pm <0.01}$

that it reasons in an incorrect way - it could have learnt a pattern matching a different GT variant. Secondly, models trained on different cross-validation folds can reason in line with different GT variant which directly impacts ReasonSMIL-R metrics calculated as averages of values from models trained on different cross-validation folds. To detect such phenomena, we calculate the standard deviations for all ReasonSMIL-R metrics and introduce another component of ReasonSMIL - ReasonSMIL-A.

ReasonSMIL-A

The results on reasoning stability are shown in Table 8.6. There is no clear relationship between ReasonSMIL-A metrics (Fleiss' kappa, RMSE) and dataset difficulty: for *points density* and *concavity* the Fleiss' kappa values are not inline with global accuracy, for *onion* such a relation can be observed. For *concavity*, models achieve the lowest Fleiss' kappa values for *medium* dataset in all GT variants. The small Fleiss' kappa values in the *medium* dataset and mediocre in *hard* suggest that the models tend to reason in line with different ground truth variants (Figure 8.8). The analysis of RMSE metric gives perspective on the topic of stability which does not depend on the definition of the ground truth variant (does not require binarization). Note that in the case of *points density* and *onion*, there is an alignment that the Fleiss' kappa and RMSE for the easiest dataset are the most favorable.

Table 8.6: ReasonSMIL-A on CLAM.

dataset	level	v.	Fleiss' ↑	RMSE ↓
concavity	hard	back	$0.41_{\pm 0.11}$	$0.11_{\pm 0.02}$
	hard	fore	$0.40_{\pm 0.08}$	
	hard	edge	$0.50_{\pm 0.09}$	
	medium	back	$0.16_{\pm 0.15}$	$0.07_{\pm 0.02}$
	medium	fore	$0.14_{\pm 0.12}$	
	medium	edge	$0.35_{\pm 0.05}$	
	easy	back	$0.93_{\pm 0.04}$	$0.12_{\pm 0.02}$
	easy	fore	$0.94_{\pm 0.01}$	
	easy	edge	$0.69_{\pm 0.14}$	
onion	hard	-	$0.69_{\pm 0.06}$	$0.10_{\pm 0.01}$
	medium	-	$0.84_{\pm 0.03}$	$0.14_{\pm 0.02}$
	easy	-	$0.88_{\pm 0.02}$	$0.06_{\pm 0.01}$
points density	hard	-	$0.78_{\pm 0.05}$	$0.14_{\pm 0.01}$
	medium	-	$0.70_{\pm 0.05}$	$0.13_{\pm 0.01}$
	easy	-	$0.81_{\pm 0.04}$	$0.03_{\pm <0.01}$

8.7.2 TransMIL

Global performance

The results for TransMIL are provided in Table 8.7. In the case of *concavity* CLAM achieves better or similar performance than TransMIL unlike in the case of *onion*. In *points density*, in a hard variant, CLAM outperform TransMIL which is contrary to situation in the medium variant. For the detailed comparative study between CLAM and TransMIL, we select datasets on which the models achieved similar performance (*concavity*: medium, *onion*: easy, *points density*: easy). In the following analyses, we used the attention scores from the first multi-head self-attention module of TransMIL.

Table 8.7: The performance of the TransMIL model on SyntheticSMIL datasets.

dataset	level	accuracy	AUC
concavity	hard	$0.55_{\pm 0.05}$	$0.61_{\pm 0.06}$
	medium	$0.78_{\pm 0.06}$	$0.84_{\pm 0.09}$
	easy	$0.80_{\pm 0.11}$	$0.86_{\pm 0.10}$
onion	hard	$0.94_{\pm 0.03}$	$0.99_{\pm 0.01}$
	medium	$0.96_{\pm 0.02}$	$0.99_{\pm <0.01}$
	easy	$0.96_{\pm 0.01}$	$0.99_{\pm 0.01}$
points density	hard	$0.50_{\pm <0.01}$	$0.76_{\pm 0.02}$
	medium	$0.84_{\pm 0.02}$	$0.96_{\pm 0.01}$
	easy	$0.95_{\pm 0.01}$	$0.99_{\pm <0.01}$

ReasonSMIL-R

TransMIL outperforms CLAM in the *concavity* 'back' variant in all metrics. In the 'edge' variant it is observed for all metrics except PR AUC and RMA. However, CLAM performs better than TransMIL in the *concavity* 'fore' variant and across all metrics in *onion* and *points density*. The exact results are in Table 8.8.

Table 8.8: ReasonSMIL-R on TransMIL where the model achieved similar performance metrics as CLAM.

d.	level	variant	PR AUC	AUC	BAC	Dice	M. coeff.	RMA
c.	medium	back	$0.79_{\pm 0.07}$	$0.67_{\pm 0.21}$	$0.79_{\pm 0.21}$	$0.85_{\pm 0.14}$	$0.58_{\pm 0.41}$	$0.67_{\pm 0.05}$
		fore	$0.39_{\pm 0.13}$	$0.33_{\pm 0.22}$	$0.44_{\pm 0.18}$	$0.23_{\pm 0.26}$	$-0.13_{\pm 0.35}$	$0.26_{\pm 0.05}$
		edge	$0.10_{\pm 0.03}$	$0.46_{\pm 0.09}$	$0.64_{\pm 0.13}$	$0.32_{\pm 0.25}$	$0.27_{\pm 0.27}$	$0.07_{\pm <0.01}$
o.	easy	-	$0.41_{\pm 0.02}$	$0.37_{\pm 0.03}$	$0.23_{\pm 0.07}$	$0.16_{\pm 0.08}$	$-0.54_{\pm 0.15}$	$0.42_{\pm 0.02}$
p.	easy	-	$0.07_{\pm <0.01}$	$0.51_{\pm 0.01}$	$0.57_{\pm 0.01}$	$0.19_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.06_{\pm <0.01}$

ReasonSMIL-A

The results are in Table 8.9. In the case of *concavity*, TransMIL exhibits higher RMSE values compared to CLAM, which favors CLAM since, for RMSE, lower values indicate better performance. Regarding Fleiss' kappa for *concavity*, the results are inconsistent and heavily influenced by the GT variant. In the 'back' and 'fore' variants, TransMIL achieves higher Fleiss' kappa scores than CLAM, which is favorable as higher Fleiss' kappa values are better. However, in the 'edge' variant, TransMIL's Fleiss' kappa scores are lower than CLAM's, giving an advantage to the CLAM model.
For *onion*, TransMIL yields lower Fleiss' kappa scores and higher RMSE values than CLAM. In terms of *points density*, Fleiss' kappa scores are comparable between TransMIL and CLAM, while TransMIL shows a higher RMSE.

Table 8.9: ReasonSMIL-A on TransMIL.

dataset	level	att	Fleiss' ↑	RMSE ↓
concavity	medium	back	$0.32_{\pm 0.25}$	$0.50_{\pm 0.07}$
		fore	$0.21_{\pm 0.22}$	
		edge	$0.16_{\pm 0.09}$	
onion	easy	-	$0.62_{\pm 0.06}$	$0.11_{\pm 0.04}$
points density	easy	-	$0.82_{\pm 0.06}$	$0.08_{\pm 0.11}$

8.7.3 Under the lens

Along with the analysis of ReasonSMIL metrics, we inspected visualizations of the important patches identified by CLAM and TransMIL models in the case when multiple GT



(a) fold 1 (b) fold 2

Figure 8.4: The important patches in the same test image, identified by the CLAM model but trained on different cross-validation folds. We selected the same number of patches as the ground truth in the 'fore' variant.

variants are possible (*concavity*).

For the analysis of TransMIL in Section 7, we picked the attention scores from the first attention layer (*att0*) as their visualizations seem less noisy than the ones from the consecutive layer (*att1*) (Figure 8.6), It may be due to some additional high-level patch interactions (Figure 8.5). The 'edge' variant is very rarely observed in reasoning (both in *att0* and *att1*). Note that it may happen that in *att0* the model follows another GT variant than in *att1* (Figure 8.7). From now on, we focus only on *att0*. **TransMIL tends to alternate its reasoning for different samples within the same cross-validation fold**. It is observed mainly between the 'fore' and 'back' variants (Figure 8.8b). However, there are a few exceptions. In one fold in each of the dataset difficulty levels (*easy*, *medium*, *hard*), TransMIL reasons mostly in line with just one GT – the 'back' variant.

The empirical analyses showed that the **CLAM model reasons in general in line with one GT variant within one cross-validation fold** (Figure 8.8a). **However, it turns out that the model tends to choose a different reasoning scheme across different folds** (Figure 8.4). From 5 different cross-validation folds CLAM learnt to reason in line with the following GT variants depending on the dataset difficulty: *hard* (4/5 models – 'fore', 1/5 – 'back'), *medium* (3/5 models – 'fore', 2/5 – 'back'), *easy* (5/5 models – 'fore'). In general, the most popular is 'fore' GT variant.

Overall, we can conclude that the TransMIL model is less consistent than CLAM, as it shows less ability to consistently focus on a specific GT variant within a given fold. These observations are reflected in standard deviations of ReasonSMIL metrics averaged within each cross-validation fold.

Patches visualisations

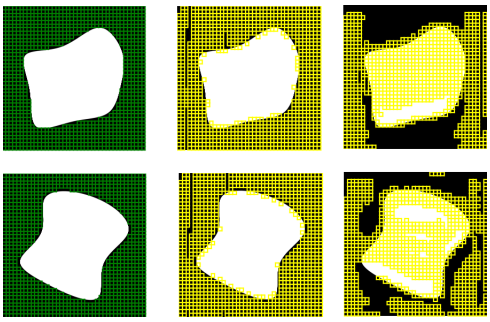


Figure 8.5: TransMIL vs. 'back' GT variant – green patches (first column). In the second column, there are visualisations of important patches based on *att0* from TransMIL, whereas, the third column – *att1*. Dataset difficulty: *hard*, upper image: class 0; lower image: class 1, cross-validation fold: 0. In this case for the same image, the reasoning in *att0* follows 'edge' whereas in *att0* it is 'foreground'.

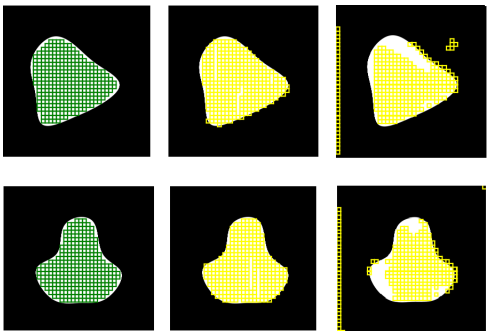


Figure 8.6: The comparison of important patches selected by TransMIL against the 'fore' GT variant in green (first column). In the second column, there are visualisations of important patches based on *att0* from TransMIL, whereas, the third column – *att1*. Dataset difficulty: *medium*, upper image: class 0; lower image: class 1, cross-validation fold: 0. It is visible that for *att0*, the choice of important patches is more reasonable.

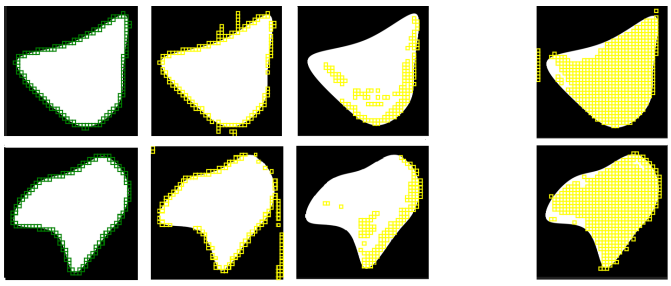
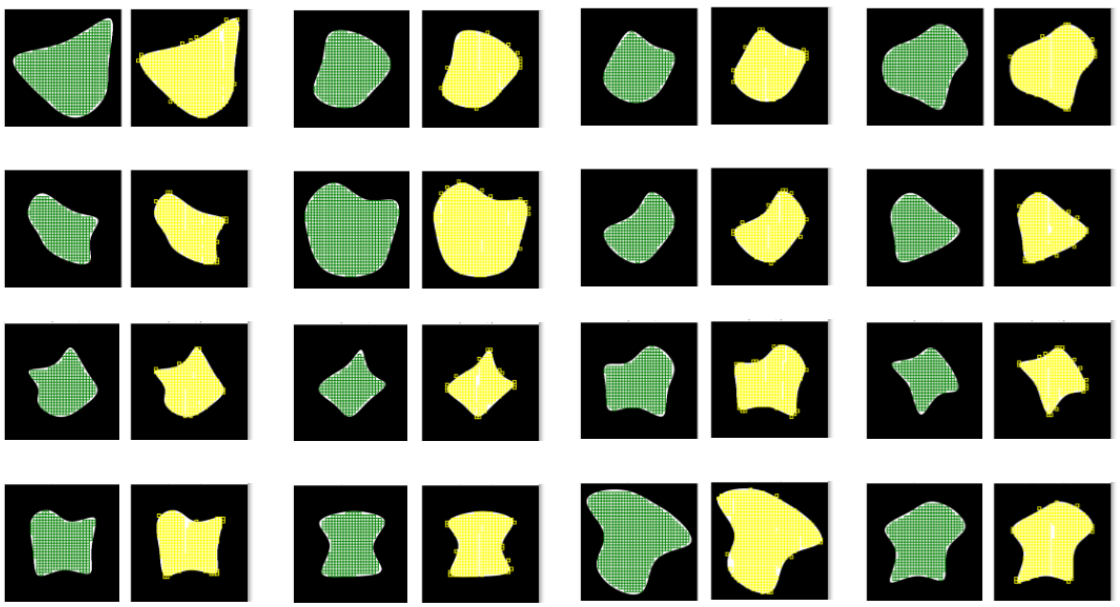
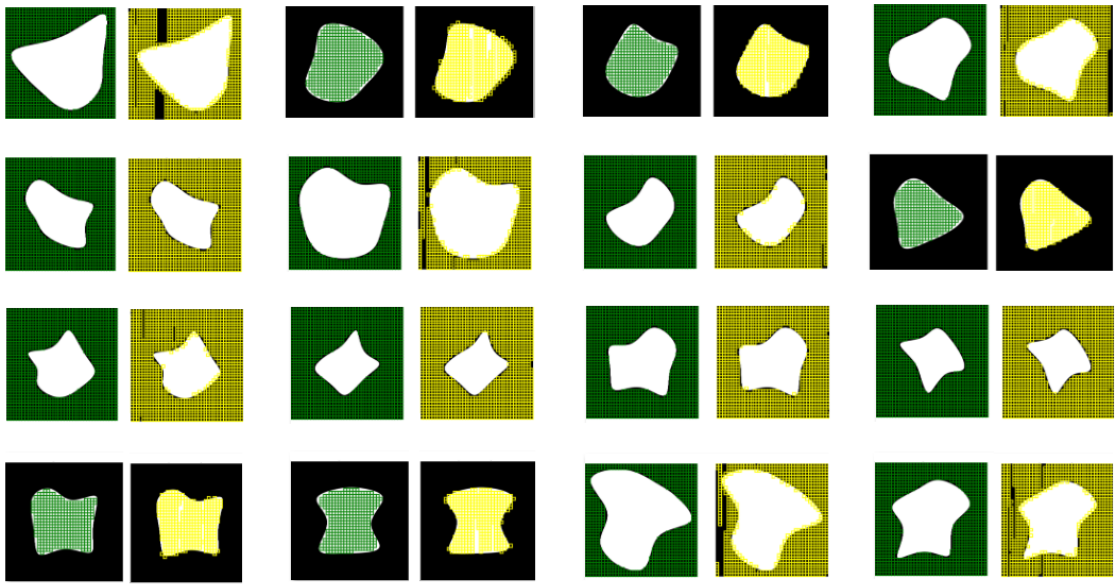


Figure 8.7: TransMIL vs. 'edge' GT variant (green patches, yellow patches - selected by TransMIL: *att0* – second column, *att1* – third column). Dataset difficulty: *medium*, upper images: class 0; lower images: class 1, cross-validation fold: 0. GT 'edge' is rarely chosen in the performed experiments, here only visible in *att0*. When checking the 'fore' variant in the case of *att1* (forth column), it seems that indeed this is the reasoning in this model layer.



(a) CLAM ground-truth variants choices.



(b) TransMIL ground-truth variants choices.

Figure 8.8: The choices of GT variants in the case of a model trained on a given cross-validation fold shown on few samples (green patches – GT, yellow patches – selected by CLAM (a)/TransMIL - *att0* (b)). Dataset difficulty: *medium*, two upper rows: class 0; two lower rows: class 1, cross-validation fold: 0. GT patches visualizations were selected to match models' choices ('fore' or 'back'). In all experiments, a given CLAM model trained on a given fold chooses a single GT variant in all test images. When trained on a particular fold, TransMIL tends to alternate with different reasoning schemes for different images.

8.7.4 Role of position encoding

In order to check the benefit of positional encoding (PE) in TransMIL, we run additional experiments. In TransMIL, the positional encoding is created in the PPEG module where convolutional layers with different kernel sizes are applied to input patch features to catch the information on neighbourhood at different granularities. We ran our evaluation pipeline on TransMIL with PE turned off meaning no convolutional kernels are applied and only identity operation is preserved.

The global performance values in case with and without positional encoding are given in Table 8.10. In general, positional encoding in all analysed cases improved global accuracy slightly (the biggest difference is in the case of concavity). The fact that the differences are small aligns with the ablation study conducted by the authors of TransMIL on Whole Slide Images.

Table 8.10: Performance comparison of TransMIL with and without positional encoding (PE).

dataset	**w/o PE**		**w/ PE**		**difference**	
	accuracy	**AUC**	**accuracy**	**AUC**	**accuracy**	**AUC**
concavity (medium)	$0.73_{\pm 0.06}$	$0.83_{\pm 0.09}$	$0.78_{\pm 0.07}$	$0.84_{\pm 0.09}$	+0.042	+0.008
onion (easy)	$0.94_{\pm 0.02}$	$0.99_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.99_{\pm 0.01}$	+0.012	+0.001
points density (easy)	$0.94_{\pm 0.01}$	$0.98_{\pm 0.01}$	$0.95_{\pm 0.01}$	$0.99_{\pm 0.01}$	+0.006	+0.006

ReasonSMIL-R The results are provided in Table 8.11. For *concavity* in the 'back' and 'edge' variants, PE improves most metrics, particularly PR AUC, AUC, and balanced accuracy (BAC). In contrast, the 'fore' variant experiences a decline in most metrics. The *onion* variant shows minimal improvement in PR AUC and AUC, while balanced accuracy, Dice, and Matthew's coefficient slightly degrade. For *points density*, all metrics exhibit slight but consistent improvements with PE, indicating that PE generally enhances model alignment with ground truth patches.

ReasonSMIL-A The results are provided in Table 8.12. The shift from the 'without PE' to the 'with PE' scenario reveals that Fleiss' kappa improves for *concavity* (all GT variants), while RMSE is worse. In the *onion* variant, Fleiss' kappa slightly drops, but RMSE improves significantly. For *points density*, Fleiss' kappa shows a notable improvement, suggesting better agreement, while RMSE remains stable.

Table 8.11: ReasonSMIL-R comparison for *concavity* (medium) with and without positional encoding (PE).

metric	**w/o PE**	**w/ PE**	**difference**
back variant			
PR AUC	$0.79_{\pm 0.11}$	$0.80_{\pm 0.07}$	+0.005
AUC	$0.66_{\pm 0.28}$	$0.67_{\pm 0.21}$	+0.008
BAC	$0.78_{\pm 0.28}$	$0.79_{\pm 0.21}$	+0.010
Dice	$0.84_{\pm 0.22}$	$0.85_{\pm 0.14}$	+0.010
Matthews Coeff.	$0.56_{\pm 0.57}$	$0.58_{\pm 0.41}$	+0.019
Relevant Mass Acc.	$0.68_{\pm 0.09}$	$0.67_{\pm 0.05}$	−0.007
edge variant			
PR AUC	$0.09_{\pm 0.03}$	$0.10_{\pm 0.03}$	+0.005
AUC	$0.44_{\pm 0.08}$	$0.46_{\pm 0.09}$	+0.015
BAC	$0.60_{\pm 0.14}$	$0.64_{\pm 0.13}$	+0.036
Dice	$0.25_{\pm 0.26}$	$0.32_{\pm 0.25}$	+0.068
Matthews Coeff.	$0.20_{\pm 0.28}$	$0.27_{\pm 0.27}$	+0.072
Relevant Mass Acc.	$0.07_{\pm 0.01}$	$0.07_{\pm 0.01}$	+0.001
fore variant			
PR AUC	$0.40_{\pm 0.19}$	$0.39_{\pm 0.13}$	−0.014
AUC	$0.34_{\pm 0.30}$	$0.33_{\pm 0.22}$	−0.013
BAC	$0.47_{\pm 0.30}$	$0.44_{\pm 0.18}$	−0.031
Dice	$0.28_{\pm 0.41}$	$0.23_{\pm 0.26}$	−0.043
Matthews Coeff.	$-0.07_{\pm 0.60}$	$-0.13_{\pm 0.35}$	−0.062
Relevant Mass Acc.	$0.25_{\pm 0.08}$	$0.26_{\pm 0.05}$	+0.006

Table 8.12: ReasonSMIL-A comparison of *concavity* (medium) with and without position encoding (PE).

variant	**Fleiss' Kappa**		
	w/o PE	**w/ PE**	**difference**
back	$0.27_{\pm 0.15}$	$0.32_{\pm 0.25}$	+0.05
edge	$0.12_{\pm 0.08}$	$0.16_{\pm 0.09}$	+0.04
fore	$0.16_{\pm 0.12}$	$0.21_{\pm 0.22}$	+0.05
	RMSE		
variant	**w/o PE**	**w/ PE**	**difference**
back/edge/fore	$0.45_{\pm 0.04}$	$0.50_{\pm 0.07}$	+0.05

8.8 Conclusions

We propose a set of configurable synthetic datasets (SyntheticSMIL) complemented with a comprehensive evaluation protocol (ReasonSMIL) for the investigation of the attention-based vision MIL models. To justify their usefulness, we provide an illustrative investigation of the CLAM and TransMIL models.

The results show that there is no straightforward mapping between the models' global performance metrics and the models' reasoning correctness evaluated using ReasonSMIL-R. It is observed that the maximal global performance metrics do not lead to maximal

fine-grained metrics. It turns out that the reasoning of models in cases when they can base their prediction on different valid reasoning schemes ('foreground', 'background', 'edge') is surprising. CLAM model seems to reason consistently on all images when trained on a particular cross-validation fold. However, this reasoning scheme may not be the same across folds. TransMIL in turn follows different reasoning patterns (GT) for different images even within a single model, trained on a given cross-validation fold. We can therefore conclude that the reasoning of the CLAM model is more consistent than the TransMIL's.

The proposed tools give valuable insights into model's internals and reasoning which may be useful in the early stage of SMIL architecture development (analogy to unit tests to explore program properties).

Chapter 9

P6: Information Injection to Deep Learning Solutions in Knowledge Transfer

9.1 Preface

Data compression is an important area of research that allows faster data transfer and efficient storage. The operation of compression can be lossy or lossless. The question arises: ***Is it possible to recover structure of the image from the patch features?*** *The short answer is 'yes'.*

In this work, I analyse whether it is possible to recover information about tissue structure within Whole Slide Image (WSIs) from the information on patch level only. The investigation of methods to process WSIs will be the main focus of the following chapters. In these methods (Multiple Instance Learning models), one of the first steps is to split the high resolution images (WSIs) into patches in a grid manner. Later, features are extracted from each patch independently using pretrained backbone. In the work, I investigate whether, despite the independent transformation applied to each patch, contextual information in the form of tissue structure can be restored.

This work is positioned along the axes defined in Chapter 1 as shown in Table 9.1. The work discusses the semiglobal spatial context defined as the neighborhood of eight patches (so-called Queen's case (Anselin, 1988)). Here, one common point between the patches is enough to call them adjacent. This is in contrast to the Rook's case, where the two regions must share a whole border to be called adjacent. The terminology originates from chessboard where Rook's neighbors refer to the four adjacent cells located directly to the North, South, East, and West, while Queen's neighbors additionally include the diagonal (corner) cells (Anselin et al., 2007). The paper is categorized into the explainable

AI category as input to compute similarity measures are the patch embeddings extracted from pretrained neural networks. Therefore, any observations that are made about the similarity scores can be directly translated to the quality of the embeddings.

In the work, the focus is on analysis of data where the Deep Learning model (here feature extractor) is fixed. It is inspired by the Data-Centric paradigm in AI which was proposed by Andrew Ng in one of the challenges[1]. The approach is that instead of having a fixed dataset and modifying the model, now the model is fixed and the data is preprocessed so that the model has best performance.

In this work, I focus on the analysis of the informativeness of patch embeddings from the perspective of tissue-level structure. The statistics that I compute based on patch-level embeddings can later potentially be injected into a Deep Learning model (*knowledge transfer*) to boost its performance.[2]

Table 9.1: Mapping of the paper by Tomaszewska (2022) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska. (2022). "Information Injection to Deep Learning Solutions in Knowledge Transfer". Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Doctoral Consortium.

9.2 Introduction

It seems an interesting research direction to combine two aspects: *knowledge transfer* models and the injection of additional knowledge into the system in order to boost its performance.

9.2.1 Knowledge Transfer

Knowledge transfer (Figure 9.1) is understood in a broader sense than *transfer learning* (TL). It includes algorithms as diverse as *domain adaptation*, *knowledge distillation*, *multi-task learning*, *continual learning* etc. They vary in the type, source, and target of

[1] https://https-deeplearning-ai.github.io/data-centric-comp/

[2] The approach with information injection (which is in line with Informed Machine Learning) postulated in the work was not later proceeded as finally the main focus in the dissertation was more on investigation of the existing models (explainable AI).

the transferred knowledge, For more details, see the survey (Lu et al., 2020). Another algorithm that could be interpreted as *knowledge transfer* is *multiple-instance learning* (Ilse et al., 2018), where the information from multiple instances is aggregated to represent the entire bag of instances in a suitable form to perform the final task. This is the bottom-up approach to *knowledge transfer*. Moreover, in the case of Computer Vision applications, the information is often extracted from the instances using pretrained neural networks.

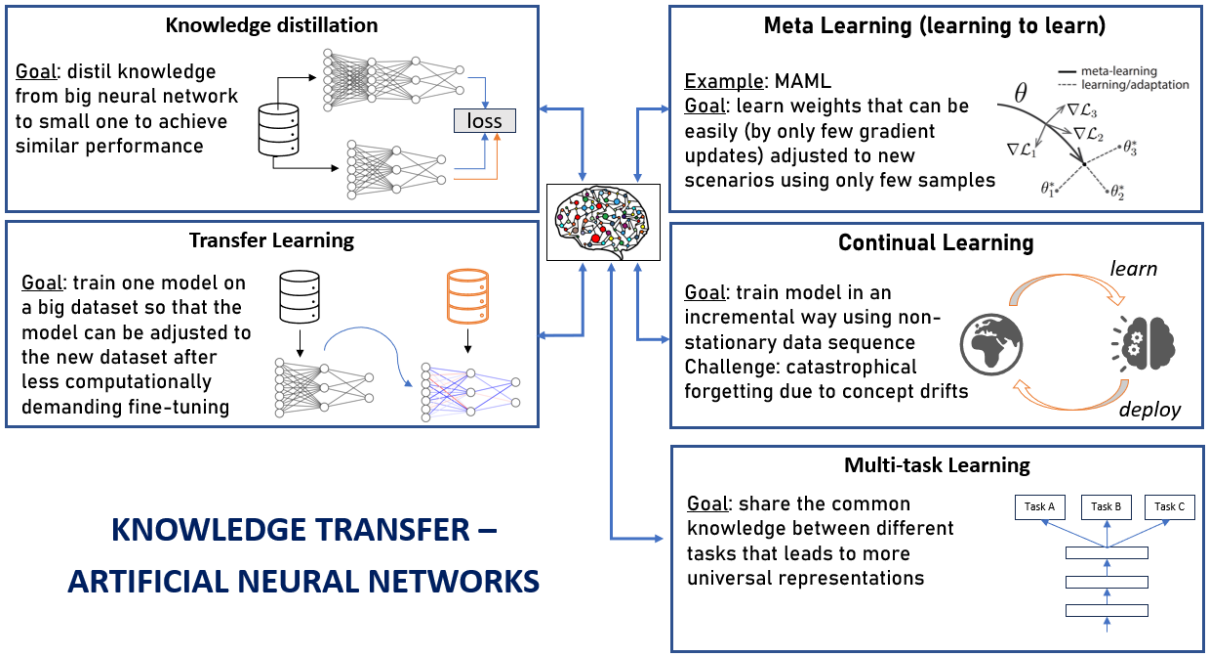


Figure 9.1: Knowledge transfer.

9.2.2 Information Injection

The injection of additional information into the Deep Learning system could be defined as providing additional information to an intelligent system. I argue that it could be expert knowledge or other information that could be derived by a human based on the data itself. The latter is a less voiced. Note that the goal is not to replicate the information that can be discovered by the neural network on its own (such as edges in convolutional neural networks). Expert knowledge can be incorporated into Deep Learning models in the form of neuro-symbolic AI (i.e., integration of rule-based AI and neural networks). This research direction is of utmost importance, especially in healthcare (Yousefirizi et al., 2022).

To make the definition of information injection less fuzzy, positive and negative examples are provided. Positive examples include adding positional encoding in Transformer architecture, injecting information about the class in Conditional Generative Adversarial Networks (cGAN) (Mirza and Osindero, 2014) and incorporating information about data distributions in the Instance Weighting Strategy in *continual learning* (Huang et al.,

2006). The information injection can be applied at different stages of the network, from an additional information added to the input (positional encoding), additional input to intermediate layers (cGAN) to the formulation of cost function (weighting). What all these solutions have in common is that the injected information is not trainable, unlike the models themselves. However, the injection is not a *multimodal learning* in which additional knowledge is inserted into the system as a new modality. In this case, in order to solve the task, all the modalities are necessary, unlike in proposed information injection concept, where it is optional but may improve the performance of the system. Moreover, *active learning* (Settles, 2009), where a human labels key data samples, is not an example of information injection. This is caused by the fact that this additional information is provided outside the model.

9.3 Proposed Contribution

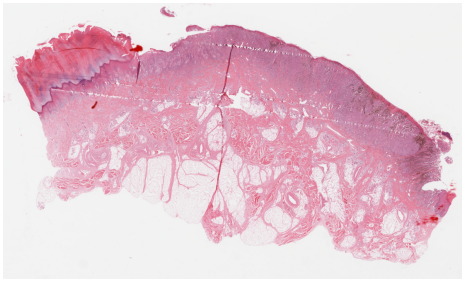
One of the planned projects on the topic of information injection into *knowledge transfer* models refers to the area of *multiple-instance learning* algorithm for histopathological data analysis. Histopathological images are of large, varying resolution e.g. 60000x100000. This makes them difficult to be processed using classical convolutional neural networks. One of the interesting works in the field of histopathological data analysis that deals with the problem of large input size is described in Lu et al. (2021a) (later referred to as CLAM for brevity). The general flow of this solution is as follows. First, features are extracted from the patches of images, and then they are passed to the attention module that returns scores. The scores are used as weights in the weighted mean that aggregates information from all patches. The result of the aggregation is supposed to be a proper representation of the overall histopathological image to perform a classification task (distinguish healthy vs. cancerous or different types of cancer). In addition, an auxiliary task (clustering) is applied where pseudolabels are created based on attention scores.

The authors stated in their conclusions that their solution does not benefit from spatial information during training. This additional information is believed to improve the performance of the system by providing valuable insights that are difficult for the network to recognize.

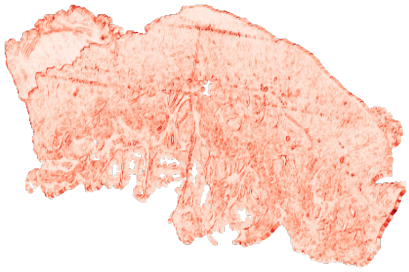
In CLAM, each patch is processed separately and then aggregated using only the attention scores with respect to the analyzed classes. Therefore, there is no spatial information about the surrounding patches, unlike in the case of convolutional neural networks, where the spatial relations are preserved by default. Consequently, in my work, the main focus would be on the potential injection of context into the CLAM solution.

The idea is to provide additional information in the form of a similarity measure between the extracted features from the patches. Given a patch, the similarity metric (such as cosine similarity or correlation) is calculated with each of the patches in Queen's

case. Later, the mean of the values is returned. The advantage of the mean is that, in the case of border patches, the score is not biased. The context could be later incorporated into the aggregation function. The idea behind such a context formulation is the result of a preliminary study. Below, due to limited space, I present only results where a similarity metric in a form of Pearson correlation was used to process Figure 9.2a into context information at Figure 9.2b.



(a) Original histopathological image of a lesion.



(b) Patchwise neighbourhood correlation.

Figure 9.2: Comparison between the original image and its neighbourhood correlation visualization.

It has been observed that by computing correlation between the features of neighboring patches, the original tissue structure (shown in Figure 9.2a) can be highly restored, which is promising from the perspective of information injection into the CLAM model. Another idea for providing context could be to incorporate a variant of a positional encoding.

Note that there is an extension of the CLAM solution by Lu et al. (2021b), where information injection is limited to the concatenation of an intermediate data representation within a neural network with information about the patient's sex. Such an approach is common where an image and a small amount of tabular data are required to make a prediction, however, it does not provide context about the spatial relationships of the patches.

Chapter 10

P7: Deep Spatial Context: When Attention-based Models Meet Spatial Regression

10.1 Preface

People often do not understand in all the details how vehicles such as cars and planes work, but this does not stop them from taking these means of transport. People believe that proper analyses and tests (e.g. crash tests) are performed by designers and manufacturers before vehicles are allowed to be used. Such tests provide valuable insight into inner workings of a vehicle. The question arises: ***Is it possible to evaluate MIL models by measuring in what extent they preserve and use information on spatial context (neighborhood) in their inner workings?***
The short answer is 'yes'.

In Chapter 9, it was shown that by computing correlation/ cosine similarity between the features of eight neighboring patches, the original tissue structure can be highly restored. This investigation showed that the information about spatial similarity (continuity) between patches is an interesting direction that should have continuation.

It was a cornerstone for my publication introducing the Deep Spatial Context ($DSCon$) method. The work was also motivated by the observation that despite the popularity of the attention-based Multiple Instance Learning (MIL) models, surprisingly little is known about the kinds of relationships they learn. To address this research gap, we propose the method to analyze the spatial relationships in these models using three quantitative measures (Spatial Context Measures) that assess the strength of the spatial structure detectable at the level of targets, features and residuals.

DSCon is inspired by contributions from various fields as little in this regard has been

made in the MIL field. Primarly *DSCon* borrows from spatial regression to provide a solid theoretical basis for analysing deep computer vision models. This allows for inspection of the importance of spatial context. Moreover, *DSCon* method draws inspiration from probing classifiers (Alain and Bengio, 2016).

Probing classifiers were introduced to better understand the role and dynamics of the intermediate model layers. The idea is as follows. There is the investigated model, which is already trained to solve a classification task. Later, for input data, an inference is performed and the internal activation values from the given model layer are extracted. The collected artifacts are used to train a linear classifier called a probe to predict the corresponding ground-truth labels of the input data. The analysis of the probe performance trained on representations extracted from various layers gives information on whether the network's ability to solve the task is increasing from layer to layer in a gradual or abrupt manner. Moreover, such an analysis can be performed after various training epochs to inspect its dynamics. Note that in this case, there is no gradient propagation between the main network and the probes so that the training schemes of the two do not interfere.

The probes find application in the XAI technique called TCAV – Testing with Concept Activation Vectors (Kim et al., 2017). Here, a linear classifier is trained on the intermediate model representation of the samples with particular concept (e.g. dotted) and without the concept. The vector orthogonal to the classifier decision boundary is called a concept activation vector (CAV). Subsequently, using CAV and directional derivatives, it is possible to express the influence of a particular concept on the model prediction as a single number. For instance, it is possible to assess the influence of the 'dotted' concept on the model predicting a class 'labybird'.

Probing classifiers are a widely used technique for analyzing Large Language Models (LLMs) through the lens of Mechanistic Interpretability. For example, the Logit Lens method (nostalgebraist, 2020) has been adopted and extended in subsequent approaches such as ReDeEP (Sun et al., 2025). Here, the probing classifier has the same output space as the investigated model. The output is a probability distribution over token vocabulary.

However, it does not always have to be the case, and the probe output can be of a different kind than the one of the main model. It means that probe and main network can solve different tasks. This was already described by Belinkov (2022). One of the provided examples was a sentiment analysis model that maps text to a sentiment label, while a probing classifier maps its intermediate representations to part-of-speech tags. This notion was also applied to analyse whether the LLM considers the input sentence to be true or false (Azaria and Mitchell, 2023), assess the LLM's truthfulness (Bao et al., 2025), determine the source of its knowledge (parametric vs. context knowledge) (Tighidet et al., 2024), or detect whether the text was included in the LLM's pre-training data (data contamination) (Liu et al., 2024b).

In the aforementioned works on the analysis of model activations, the classifier was applied on top of model internal activations, but I propose in my publication to apply regression model instead. I follow the convention established in prior research, where the probes are of small size and low complexity (typically linear). One of the few exceptions in the literature is Azaria and Mitchell (2023), where the probe consists of several fully connected layers with ReLU nonlinearities.

In this work, the spatial context is analysed as a concept. Note that in this case, the process in other than pointed out in Section 2.1.3 where the concepts of interest are named at the very end of the XAI exploration process. In the work, the choice of the concept to be analysed is a result of the collaboration with Dr. Mai P. Hoang who is currently the director of Dermatopathology at Massachusetts General Hospital and Professor of Pathology at Harvard Medical School. She described the daily process of assessment of the tissues that I later formalized into spatial context.

This work is positioned along the axes defined in Chapter 1 as shown in Table 10.1. The paper deals with both semiglobal and global context depending on the k value considered in the k-nearest neighbors algorithm used to define the spatial weight matrix. The case with $k = 8$ is called semiglobal as in Chapter 9, whereas the cases with $k > 8$ are referred to as global.

Table 10.1: Mapping of the paper by Tomaszewska et al. (2025b) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

<u>Paulina Tomaszewska</u>, Elżbieta Sienkiewicz, Mai P. Hoang & Przemysław Biecek. (2025). "Deep Spatial Context: When Attention-based Models Meet Spatial Regression (digital pathology)". Proceedings of the International Conference on Artificial Intelligence in Medicine (AIME).

Below, I provide the version of the paper extended with additional analyses and experiments that could not be included in the published version due to the page limit.

10.2 Introduction

The behavior and reasoning of Deep Learning (DL) models are often difficult to understand, especially for non-computer scientists, leading to low trust in critical areas like healthcare (Combi et al., 2022; van der Velden et al., 2022; Jin et al., 2023). For 80% of physicians, clear information on key AI model characteristics and features is crucial (survey by American Medical Association (2023)).

This need for understanding inspired the rise of eXplainable AI (XAI) (Alicioglu and Sun, 2022). As postulated in (Koh et al., 2020; Crabbé and van der Schaar, 2022; Longo et al., 2024; Chrabąszcz et al., 2025), the XAI methods should operate on human-understandable, expert-specified concepts. One such concept can be the spatial context in images, which is rarely addressed in XAI unlike in the design of the model architectures (Tomaszewska and Biecek, 2024).

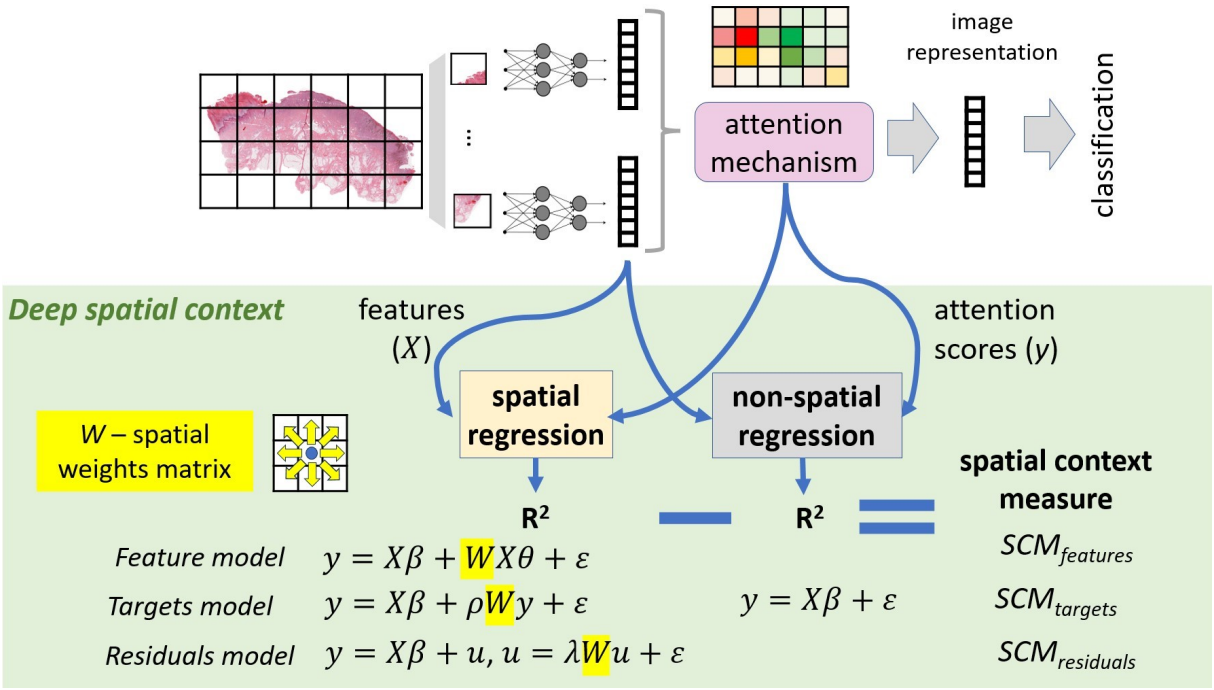


Figure 10.1: In the proposed *Deep spatial context* method patch features and attention scores extracted from the attention-based Multiple Instance Learning model are used in various regression models to evaluate spatial context. The Spatial Context Measures are computed as performance differences (e.g., R^2) between the respective spatial and non-spatial regression models. Three spatial models assess where the spatial context lies: in features, targets, or residuals.

Spatial context is important in such areas like remote sensing (Abrahams et al., 2024), spatial biology (Bissell and Radisky, 2001) and digital pathology (Abousamra et al., 2023). The latter relates to the daily practice of histopathologists who analyze tissues from both global (tissue-level) and local (cell-level) perspectives (Gurcan et al., 2009) and the fact that tumors form different spatial cellular structures (Abousamra et al., 2023).

The distribution of tumor areas within the tissue, i.e tumor budding (the process where single cancer cells or small clusters detach from the main tumor and invade surrounding tissue (Lugli et al., 2020)), is crucial for the assessment of the disease stage, which directly affects risk stratification and treatment decisions (Dee et al., 2023). In this work, we focus on the spatial context which refers to the importance of the location and relative position of elements within the image.

The goal of this work is to analyze deep attention-based MIL models from a spatial context point of view. Such models are often used in digital pathology to analyse Whole Slide Images (WSIs). The attention-based deep Multiple Instance Learning (MIL) models (Ilse et al., 2018) assign attention scores (weights) to patches extracted from the huge resolution images to form the overall representation of the image and perform the final classification. In many MIL models i.e. (Lu et al., 2021a; Rymarczyk et al., 2023) the information about the patches' location within the whole image (position encoding) is not explicitly given. This can be troublesome in the case of histopathological images that have non-standardized sizes.

To verify if the information on spatial context is preserved within the attention-based MIL models (even if the information on patch locations is not provided), we propose to investigate the relationship between the patch features and the assigned attention scores. The MIL models are not explicitly asked to learn spatial relations. However, the fact that it achieves decent classification performance suggests that some contextual information may be captured but it requires further investigation. However, the reasoning may vary even for models of similar performance (Bach et al., 2015; Watson et al., 2022). Following the claims in Biecek and Samek (2024), we propose the method to empower model developers with additional insights.

To address this research gap, we introduce a *Deep spatial context* method (*DSCon*) (Figure 10.1) which explores the MIL model through the lens of a semantic concept such as the spatial context. The method is general and can be applied to various domains.

Our main contribution is as follows.

- We propose a novel *Deep spatial context* method that allows for the investigation of DL models from the perspective of spatial context concept. The method incorporates spatial regression into the field of explainability, providing a solid mathematical foundation.

- We verify whether the spatial context is preserved in attention-based MIL models and we provide a method to assess it quantitatively using three spatial context measures ($SCM_{features}$, $SCM_{targets}$, $SCM_{residuals}$) that enable to distinguish where the spatial relationships are maintained (feature, label, residual space).

- We show use cases where the *DSCon* method finds application – the comparison on models within one model family, across model families; on different subsets of data.

The code is provided at `https://github.com/ptomaszewska/DSCon.git`.

10.3 Related work

The most popular post-hoc methods in XAI are attribution-based, which focus on highlighting the parts of the input image that mostly influence the final model decisions, i.e. Simonyan et al. (2014); Ribeiro et al. (2016); Selvaraju et al. (2017); Ancona et al. (2018). The recent works focus more on high-level, complex semantic concepts (Kim et al., 2017; Achtibat et al., 2022; Sun et al., 2023). Most of existing XAI methods do not have solid mathematical foundations. One of few exceptions are methods based on SHAP values (Ancona et al., 2018; Lundberg et al., 2018) that are rooted in game theory (Lundberg and Lee, 2017).

Few attempts were made to investigate the attention mechanism, which is crucial in DL systems, from a more theoretical point of view, i.e. via a dynamical system perspective (Huang et al., 2023), through the lens of exchangeability (Zhang et al., 2022b) or as a marginal likelihood over possible network structures (Singh and Buckley, 2023). To the best of our knowledge, we are the first to apply spatial statistics for the analysis of attention-based MIL models and Deep Learning models, in general.

In the papers on digital pathology, the analysis of the model behavior is often limited to the global performance metrics accompanied by a few heatmaps with patches' importance in the overall tissue classification. Only a few attempts are made to explain the models in more detail, i.e. Rymarczyk et al. (2023); Xie et al. (2020); Hense et al. (2024). As far as we know, no prior studies have analyzed models from a spatial context perspective before ours.

10.4 Preliminaries

Regression models Ordinary least squares (OLS) regression is formulated for observation i as $y_i = X_i\beta + \epsilon_i$ where y_i is the dependent variable, X_i – the vector of explanatory variables, β – the vector of coefficients for the explanatory variables and ϵ_i – the error term. The assumptions behind the formula are: $\mathbb{E}[\epsilon_i\epsilon_j] = 0 \quad \text{for} \quad i \neq j$ and $\mathbb{E}[X_i\epsilon_i] = 0$ which translate to the fact that there cannot be spatial dependence between the variables. However, in the context of patches derived from a Whole Slide Image, there exists an inherent spatial continuity in features. This continuity suggests that the information from neighboring patches may contain valuable contextual data, which aligns with the expertise of histopathologists.

To account for spatial dependencies, two common classes of spatial regression models are used with $W = [w_{i,j}]$ being a spatial weights matrix:

Targets model – the spatial aspect is associated with the features within the regression formula: $y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + X_i \beta + \epsilon_i$, where ρ is the spatial autoregressive coefficient. This formulation can be estimated using spatial two-stage least squares (S2SLS) methods as described by Anselin (1988). In statistics, this model is formally called the Spatial Lag Model.

Residuals model – the spatial aspect is hidden within residuals, $y_i = X_i \beta + \lambda \sum_{j=1}^{n} w_{ij} \epsilon_j + \epsilon_i$, where λ is the spatial error coefficient, capturing the strength of spatial correlation in the residuals. This model is typically estimated using the generalized method of moments (GMM) as discussed by Kelejian and Prucha (1999, 1998). In statistics, this model is called the Spatial Error Model.

Furthermore, the OLS model can be extended to incorporate spatial information by manually lagging features based on neighboring patches, resulting in the **Feature model** – the spatial aspect is considered within features, $y_i = X_i \beta + \sum_{j=1}^{n} w_{ij} X_j \theta + \epsilon_i$, where θ is the vector of coefficients for the spatially lagged explanatory variables, X_j – the vector of explanatory variables for observation j (neighbors of i). In statistics, this model is called Spatial Lag of X (SLX) (Hughes and Holt, 1970). Although sometimes referred to as 'pseudo-spatial', we will treat it equivalently with other spatial models in subsequent discussions.

In all models, the spatial structure of the data is encapsulated within W. It can be defined according to Rook's or Queen's contiguity criteria (Cliff and Ord, 1973), where Rook's criterion considers neighbors as regions that share a common border, while the Queen's considers neighbors as regions sharing at least one common point. Alternatively, a distance-based criterion, such as the k–nearest neighbors approach, can be employed.

Attention-based MIL models The high-level pipeline of the attention-based MIL models is as follows. Let an image of high resolution, x, be divided into N non-overlapping patches $\{p_1, p_2, \ldots, p_N\}$. Each patch, p_i, is independently processed by a feature extractor, yielding a corresponding feature vector, h_i, for each patch, p_i, where $i = 1, 2, \ldots, N$. These feature vectors, h_i, are subsequently input into a module, which computes attention scores, a_i, for each feature vector, h_i. The overall representation, z, of the input image, x, is computed based on feature vectors, h_i, and attention scores, a_i. This representation, z, is subsequently passed through a series of feed-forward layers, which perform the final classification of the input image, x.

10.5 Deep Spatial Context

The key mechanism of attention-based models is the assignment of attention scores to patch features. To explore this relationship from a spatial context perspective, we propose

the *Deep Spatial Context* (*DSCon*) method, which integrates spatial regression into the model analysis. The goal of the *DSCon* method is to evaluate the impact of incorporating spatial terms into regression models compared to models without such terms, thereby assessing the significance of spatial context in the predictive performance.

The detailed procedure of the *DSCon* method is as follows (Algorithm 9). Given a pre-trained attention-based model, the first step is to perform inference and extract the patch features $H = \{h_i\}_{i=1}^{N}$ and attention scores $a = \{a_i\}_{i=1}^{N}$ for N patches constituting WSI. Before applying regression analysis to predict the attention scores, a, based on the extracted features, H, it is crucial to eliminate features that cause multicollinearity, as this can adversely affect the regression model. Feature selection is performed using Pearson correlation, with a threshold γ to filter out highly correlated features.

Next, an ordinary least squares (OLS) regression model, excluding any spatial component, is applied. The input are patch feature, H, after filtering. The goal is to predict attention scores, a. The procedure is applied to each image separately. The resulting performance metric of the OLS model, specifically the coefficient of determination R^2, is recorded to accommodate variations in sizes of Whole Slide Images, serving as a baseline for subsequent analysis.

Later, the applicability of spatial regression models is then assessed by examining the residuals $\{\epsilon_i\}_{i=1}^{N}$ of the OLS model for spatial clustering. This is formally tested using Global Moran's I statistic (Moran, 1950), which evaluates the presence of global spatial autocorrelation in the data. The Global Moran's I is defined as:

$$GM_I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})}{\left(\sum_{i=1}^{N}(\epsilon_i - \bar{\epsilon})^2\right) \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}},$$

where N is the number of patches, w_{ij} are elements of the spatial weights matrix W, and $\bar{\epsilon}$ is the mean of the residuals $\{\epsilon_i\}_{i=1}^{N}$. At this point, the spatial weights matrix W is defined, for instance, using the k–nearest neighbors approach to determine the neighborhood structure around each patch.

If the p-value obtained from the Global Moran's I test is below a specified significance level α, the use of spatial regression is justified. With the same set of features used in the OLS regression, spatial *features*, *targets*, *residuals models*, are executed, and their performance metrics are saved.

The final step involves calculating the differences between the performance metrics obtained from each spatial regression model and the OLS model. The Spatial Context Measures are defined as:

$$SCM_{features} = R^2_{features} - R^2_{OLS}, \tag{10.1}$$

$$SCM_{targets} = R^2_{targets} - R^2_{OLS}, \tag{10.2}$$

Algorithm 9 *Deep spatial context (DSCon)*

Input: MIL model (f), dataset (X, Y)
Parameter: correlation threshold (γ), neighborhood size (k)
Output: spatial context measures ($SCM_{features}$, $SCM_{targets}$, $SCM_{residuals}$)

for x, y in (X, Y) **do**
  $H, a, y_{pred} \leftarrow f(x)$
  $c \leftarrow \{correlation(H_{:,i}, H_{:,j})\}$
  if $abs(c_{ij}) > \gamma$ **then**
    drop $H_{:,i}$
  end if
  $R^2_{OLS}, residuals_{OLS} \leftarrow OLS(H, a)$
  $W \leftarrow spatial\_weights(type = kNN, k)$
  $p_{Moran} \leftarrow Moran\_test(W, residuals_{OLS})$
  if $p_{Moran} > significance\_level$ **then**
    skip
  end if
  $R^2_{features} \leftarrow Feature\_model(H, a, W)$
  $R^2_{targets} \leftarrow Targets\_model(H, a, W)$
  $R^2_{residuals}, \leftarrow Residuals\_model(H, a, W)$

  $SCM_{features} \leftarrow R^2_{features} - R^2_{OLS}$
  $SCM_{targets} \leftarrow R^2_{targets} - R^2_{OLS}$
  $SCM_{residuals} \leftarrow R^2_{residuals} - R^2_{OLS}$
end for

$$SCM_{residuals} = R^2_{residuals} - R^2_{OLS}. \tag{10.3}$$

A positive spatial context measure indicates that incorporating spatial context in the regression model is beneficial, implying that the spatial relationships are effectively captured in the trained DL model. Conversely, a negative spatial context measure indicates that the model lacks a detectable spatial signal. The magnitude of the spatial context measure provides quantitative insights into the importance of spatial context.

10.6 Deep Learning models

We used the Camelyon16 Breast Cancer dataset (Ehteshami Bejnordi et al., 2017) containing 399 WSIs in total. We conducted experiments with CLAM (Lu et al., 2021a) and TransMIL (Shao et al., 2021) models. The second uses positional encoding for patches, whereas the first does not.

10.6.1 CLAM

We decided to test the proposed *Deep spatial context* method on a range of different variants of CLAM achieving similar classification performance to verify whether, the differences in capturing the spatial context are significant. The CLAM variants differ in the feature extractor, which is reflected in the names of the whole trained CLAM models mentioned afterwards. We chose diverse feature extractors – convolutional and transformer-based such as *Swin* (Liu et al., 2021), *ViT* (Dosovitskiy et al., 2021). In total, we tested 9 different variants listed in Figure 10.2.

We used the convolutional and transformer-based backbones already pretrained on Imagenet. Following the original CLAM setup, we did not fine-tune them on histopathological data. The following models served as feature extractors: *Swin (base)* (*patch_size=4, window_size=7*), *Swin v2* by Liu et al. (2022b) (*tiny, small, base* with *patch_size=4* and *window_size=8* or *16*), *ViT (base)*. The extracted patch features served as an input to the attention pooling and consecutive classification layers which were trained from scratch.

The initial parameters of the grid search were inspired by the ones used in Jia et al. (2022). Later, they were modified based on the intermediate results. After an extensive search over many hyperparameters, we chose those that for every model variant (with different backbones) resulted in a decent performance similar to the baseline with *Resnet50* backbone (denoted as *resnetCLAM*). In *resnetCLAM*, the parameters claimed in the original paper were applied. The following parameters were chosen for the final experiments on CLAM model with transformer-based backbones: *learning_rate=0.005*, *weight_decay_rate=0*, *dropout_rate=0.25*, *bag_weight=0.7*. In the first layer of the classifier, there were 512 neurons, whereas 128 in the second. Following the setup in Jia et al. (2022), the SGD optimizer was used.

The classification performance of the models is summarized in Figure 10.2. The differences in mean performance among the models are not statistically significant which was tested using the Wilcoxon-Holm method. It is highlighted in Figure 10.2 by horizontal line below the ranking axis connects all the model results.

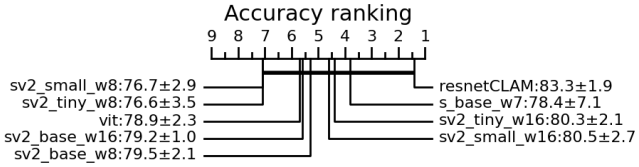


Figure 10.2: Critical distance plot[1]of models' accuracy of classification of normal tissue vs. tumor lesion. For brevity: *s* denotes *Swin*, *sv2* – *Swin v2* and *w* – *window*. In the labels, the mean accuracy (percentage) and the standard deviation over 5 trained models during the cross-validation are provided.

[1]The code used for the generation of the plot is from Ismail Fawaz et al. (2019).

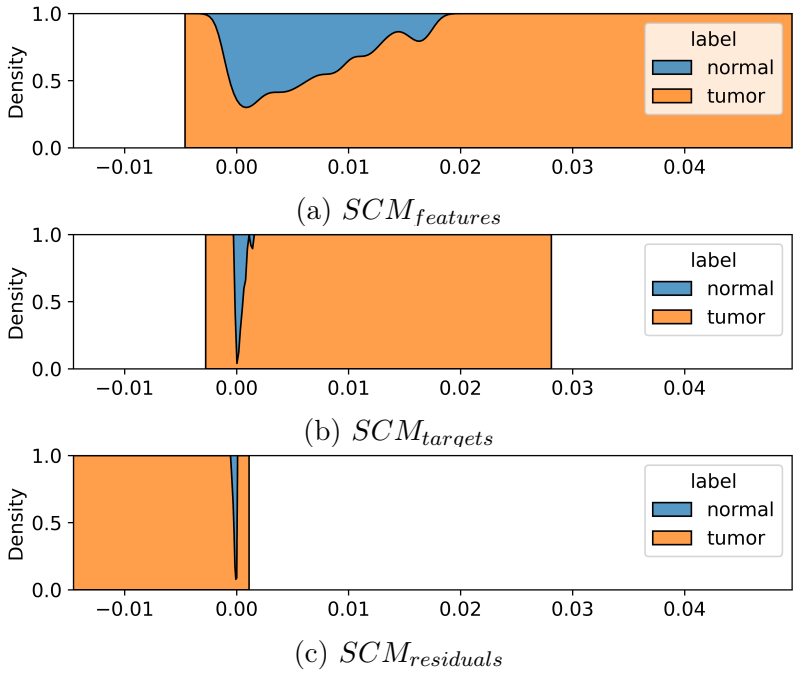
10.6.2 TransMIL

We trained the TransMIL model to compare whether its spatial context measures would be greater than those of the CLAM model. This comparison is motivated by the fact that TransMIL incorporates positional encoding, unlike CLAM. For model training, we used the default parameters. Note that in the case of CLAM and TransMIL the same extracted patch features are used for later processing within the model, therefore, we trained in total 9 model variants as in the case of CLAM. The resulting accuracy metrics of 8 model variants with transformer-based feature extractors varied from $75.3_{\pm 1.4}\%$ to $81.1_{\pm 1.2}\%$. The performance of $resnetCLAM$ was $87.3_{\pm 2.6}\%$. Therefore, in the first case the performance of TransMIL is on par with CLAM, whereas in the latter, TransMIL achieves superior performance.

10.7 Application of *Deep Spatial Context*

Depending on the backbone, the length of the extracted features was 768 or 1024. In the case of CLAM, the raw attention scores (before softmax) were used, in TransMIL, they were transformed, $a = ln(a - min(a) + \epsilon)$. In this way, the distributions were closer to Gaussian. We took the attention scores from the first multi-head self-attention module. Subsequently, the size of the features was reduced due to filtering based on the Pearson correlation coefficients with the threshold ($\gamma = 0.6$). Later, Moran's I test was performed with $\alpha = 0.05$. It turned out that in 99.4% (CLAM) and 86.3% (TransMIL) of all cases, there was a statistically significant spatial autocorrelation which justified the use of spatial regression. Here, one case is a test image analyzed using the data (features and attention scores) extracted from a particular DL model using a regression model with a spatial weights matrix defined using k–nearest neighbors. In short, it means that one test image was processed using the particular 'model–k combination'. In the experiments, a different range of neighborhood is considered within the spatial weights matrix defined using k–nearest neighbors (with *k=8, 24, 48, 99*). In WSIs, $k = 8$ leads to the same results as the Queen's contiguity matrix, since the grid of patches is regular.

The regression models were run on test data. Five different sets of attention scores corresponded to the features extracted from each patch using a particular backbone as a result of a 5-fold cross-validation run on whole MIL models. Therefore, in total, a particular regression model was fitted 5 times on a set of features and attention scores from a particular image. The performance results (R^2) were aggregated using concatenation to have more data points for the analysis.



(a) $SCM_{features}$

(b) $SCM_{targets}$

(c) $SCM_{residuals}$

Figure 10.3: Stacked kernel distribution estimation plots where the proportion of values from normal tissues and tumor lesions for a given Spatial Context Measure is shown. The results are for the *sv2_base_w8* model with *k=24* nearest neighbors within the spatial weights matrix.

10.8 Results

10.8.1 Analysis of model reasoning on different subsets of data

The analysis of results from all combinations of the CLAM model with different backbones and k values (in short model–k combinations) indicates that the increase in the R^2 metric due to the application of spatial regression reaches higher values in the case of tumor lesions than in normal tissues (for $SCM_{features}$ and $SCM_{targets}$). The results of one representative model–k variant is in Figure 10.3. Note that the range of values attributed to normal lesions is much wider in the case of $SCM_{features}$ than in $SCM_{targets}$, where the values are almost zero for normal tissues, meaning that the spatial context within the targets is negligible in this case. Moreover, $SCM_{residuals}$ almost always has negative values, which is rare in the case of $SCM_{features}$ and $SCM_{targets}$. The visible differences in the boost ranges of R^2 in different regression models are statistically significant regardless of k value within the spatial weights matrix which was verified using a t-test with $\alpha = 0.05$. These conclusions are valid for all analysed model variants except *sv2_small_w16* where the extreme values of $SCM_{features}$ and $SCM_{targets}$ are observed in the case of normal tissue.

To conclude, the spatial aspect is more observable within features than targets (bigger SCM values). However, a greater separability between the values attributed to tumor and normal tissue is visible within $SCM_{targets}$ (Figure 10.3). Applying spatial context to

residuals seems to introduce distortion causing $SCM_{residuals}$ to be almost always negative. Therefore, in the following analyses, we discard $SCM_{residuals}$.

10.8.2 Analysis of alignment between model reasoning and common sense

So far tumor data was analyzed without the distinction of whether it is widespread or clustered. Let us define the measure of tumor spread as a mean Euclidean distance between patches (of size 224x224) containing tumor cells ($dist_{mean}$). The information is extracted from the binary masks attached to the dataset.

In Figure 10.5 it is shown that the Spatial Context Measures often peak for the smallest k. To explore this further, we investigated whether this this holds true for widespread tumors, where the global context is more likely to be relevant.

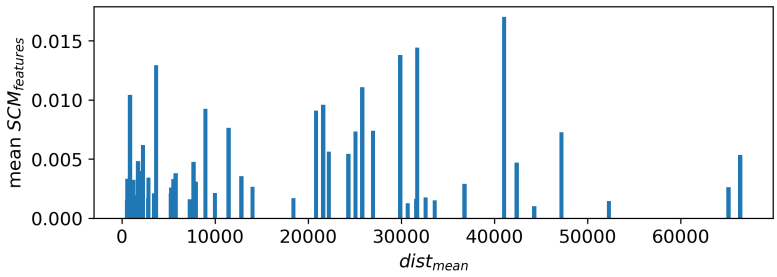


Figure 10.4: Mean of $SCM_{features}$ computed on all tumor images with all model–k combinations where $k = 99$. The images are characterized by the measure of tumor spread.

It turns out that the expected increase in Spatial Context Measure for more widespread tumors is not observed (Figure 10.4). The Pearson correlation coefficient between $dist_{mean}$ and $SCM_{features}$ ranges from 0.070 (for $k = 8$) to 0.125 (for $k = 99$) which is a monotonical increase with respect to neighborhood size. The same analysis performed on $SCM_{targets}$ shows that the correlation coefficients are about two times bigger than in the case of $SCM_{features}$ but are less dependent on k – they range from 0.214 ($k = 8$) to 0.237 ($k = 48$). Therefore, it seems that the global context even for widespread tumors is not preserved.

10.8.3 Comparison of different model variants with similar performance

We prepared a ranking based on spatial context measures given tumor data to analyze differences in CLAM models with different backbones aggregating results from (Table 10.2). It turns out that places 5–9 are the same for the mean values of $SCM_{features}$ and $SCM_{targets}$. In the first four places are models with backbones of different architectures – *resnetCLAM*, *ViT* and *Swin v2*. The order of the models varies for different spatial

context measures. It seems that the architecture type – convolutional vs. transformer-based neither the attention coverage (global – in *ViT* vs. local – in *Swin* family) do not directly impact the spatial context measures. Within 1.–4. places are mostly models that generate longer representations (1024) – *resnetCLAM*, *ViT* and *Swin v2 base*. The exception is *Swin v2 small*. For *base* and *tiny* models, the smaller the window size, the bigger the spatial context measure given the same model architecture (which is in line with Figure 10.5). The statistical significance of the differences was analyzed using a pairwise Wilcoxon test ($\alpha = 0.05$). The differences between the spatial context measures within DL models with different backbones can be directly attributed to the design of the feature extractors and the values of the resulting features because there was no significant difference in the overall classification performance of the DL models (Figure 10.2).

Impact of size of neighborhood (k)

We analyse impact of the size of the neighborhood k incorporated in the spatial weights matrix within spatial regression models on the Spatial Context Measures. In Figure 10.5, we show the average measures over all tumor images. In the case of $SCM_{targets}$, there is a clear decreasing trend (the bigger k, the smaller R^2 boost), whereas, in the case of $SCM_{features}$, it depends on the feature extractor. For some DL models, there is an initial very slight increase from $k = 8$ to $k = 24$ and then a drop, and for others, there is a decreasing trend on the whole spectrum. In general, however, all the models behave similarly as the curves are largely parallel. The decreasing trend on the plots aligns with the first law of geography saying that *'everything is related to everything else, but near things are more related than distant things'* (Tobler, 1970).

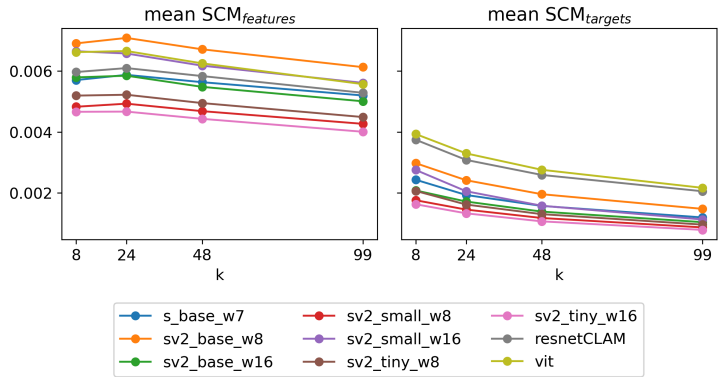


Figure 10.5: The mean value of Spatial Context Measures for models with different feature extractors with respect to the size of neighborhood (k) in the spatial weights matrix within spatial regression models over all images with tumor.

Neighborhood size vs. window size

The general claim regarding the analyzed models based on Figure 10.5 is that the continuity and spatial correlation of targets (measured using $SCM_{targets}$) occurs more in the adjacent patches than the distant ones. Therefore, the smallest possible spatial context is observed. The question is whether it is true that the bigger the window size within the $Swinv2$, the bigger extent to which the spatial relations are captured. It turns out that both in the case of $SCM_{features}$ and $SCM_{targets}$ the aforementioned relationship is true only for *small* model as the curves corresponding to small backbone with $window\_size = 16$ are above the corresponding ones with $window\_size = 8$ for all k. In the case of *base* and *tiny* models, the opposite is true meaning the smaller *window size* the bigger Spatial Context Measures. Note that the conclusions drawn regarding the two measures are consistent which is an asset.

Spatial context measures vary depending on the feature extractor used

We prepared a ranking based on spatial context measures given tumor data to analyze differences in CLAM models with different backbones aggregating results from (Table 10.2). It turns out that places 5–9 are the same for the mean values of $SCM_{features}$ and $SCM_{targets}$. In the first four places are models with backbones of different architectures – *resnetCLAM*, *ViT* and *Swin v2*. The order of the models varies for different spatial context measures. It seems that the architecture type – convolutional vs. transformer-based neither the attention coverage (global – in *ViT* vs. local – in *Swin* family) do not directly impact the spatial context measures. Within 1.–4. places are mostly models that generate longer representations (1024) – *resnetCLAM*, *ViT* and *Swin v2 base*. The exception is *Swin v2 small*. For *base* and *tiny* models, the smaller the window size, the bigger the spatial context measure given the same model architecture (which is in line with Figure 10.5). The statistical significance of the differences was analyzed using a pairwise Wilcoxon test ($\alpha = 0.05$). The differences between the spatial context measures within DL models with different backbones can be directly attributed to the design of the feature extractors and the values of the resulting features because there was no significant difference in the overall classification performance of the DL models (Figure 10.2).

Difference in spatial context measures within training and test sets

Sometimes DL models generalize poorly on the test set which is observed within the performance metrics. The question is whether it is also the case in the ability to capture the spatial context. We performed the independent samples t-test to verify whether there is a significant difference in mean spatial context measures within CLAM variants on tumour lesions in training and test sets. It turned out that in more than 86% of all

Table 10.2: Ranking of CLAM variants based on the mean Spatial Context Measures. In the brackets, the mean number of predictors (over all test images) used in regression models is given (after dropping the correlated ones). The values are provided only in one column for brevity.

ranking	**mean** $SCM_{features}$	**mean** $SCM_{targets}$
1.	*sv2_base_w8* (736)	*vit*
2.	*vit* (703)	*resnetCLAM*
3.	*sv2_small_w16* (572)	*sv2_base_w8*
4.	*resnetCLAM* (410)	*sv2_small_w16*
5.	*s_base_w7* (755)	*s_base_w7*
6.	*sv2_base_w16* (747)	*sv2_base_w16*
7.	*sv2_tiny_w8* (585)	*sv2_tiny_w8*
8.	*sv2_small_w8* (577)	*sv2_small_w8*
9.	*sv2_tiny_w16* (615)	*sv2_tiny_w16*

model-k combinations, the mean of $SCM_{features}$ and $SCM_{targets}$ were significantly greater on the training set than the test set. Therefore, it seems that DL models do not generalize well in capturing spatial context despite the use of the early stopping mechanism during model training.

'Most spatial' images

As already shown in Figure 10.3, $SCM_{features}$ and $SCM_{targets}$ tend to be much bigger for tumor lesions than normal tissues. We selected the images with tumor where the spatial context is the most significant. We calculated the threshold as a 95%-percentile of the SCM_{lag} on the normal lesions. Later, we picked images with tumor with a larger $SCM_{targets}$ than the threshold. This procedure was performed for each model-k combination separately (results from the folds were aggregated using the mean). For the selected images, we extracted information from their binary masks about (1) the number of tumor patches ($n_{tumor\_patches}$) and (2) tumor spread ($dist_{mean}$ – the same metric as in Figure 10.4). We applied an additional filter on the images as we picked only the ones that were chosen as 'the most spatial' by more than half of all model-k combinations. Moreover, we computed the mean value of k within the model-k combination that led to the decision that a particular image is the 'most spatial' (k_{mean}). It turns out that most of the images pointed as 'the most spatial' have many patches with tumor cells (Figure 10.6). The point $log(n_{tumor\_patches}) = 6$ could be perceived as a cutoff which gives about 400 tumor patches. Note that whether a tumor is widespread or clustered ($dist_{mean}$) does not play a significant role.

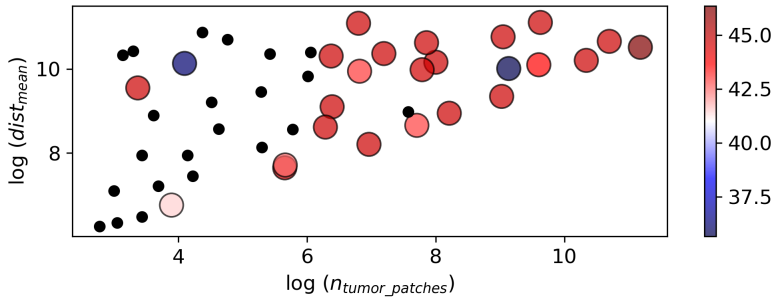


Figure 10.6: The relationship between the logarithms of $n_{tumor\_patches}$ and $dist_{mean}$. The color denotes k_{mean}. The small black dots depict the tumor images that were chosen as 'the most spatial' by less than or half of the model-k combinations.

10.8.4 Comparison of models from different families

The CLAM and TransMIL models generally had similar classification performance. However, it turns out that they reason in a different manner from the spatial context point of view (Table 10.3). Spatial Context Measures are significantly higher in the case of TransMIL than in CLAM (except the mean of $SCM_{targets}$). This can be attributed to the fact that positional encoding is within the input to TransMIL unlike in CLAM. Note that in CLAM, the maximum values of $SCM_{features}$ and $SCM_{targets}$ are bigger in the case of tumor lesions than normal tissue which is not the case given TransMIL where in $SCM_{features}$ there is a reverse relationship and in the case of $SCM_{targets}$ the maximum values in the two tissue variants are mostly similar. Note that for the CLAM model tumor context was more important than the normal context which was not always the case for TransMIL. From the clinical point of view, tumor context is more important, that is why in the next section we provide an ablation study on the CLAM model.

10.8.5 Analysis of the different variants of positional encoding within TransMIL

We run additional experiments to show that the investigation of proposed Spatial Context Measures can bring actionable insights that can have an impact on model architecture design.

We did an ablation study on the Pyramid Position Encoding Generator (PPEG) inside TransMIL. In PPEG, convolutions with different kernel sizes (3x3, 5x5, 7x7) are applied to patch features organized spatially in a way resembling their initial positions in Whole Slide Image. The output of these convolutions is aggregated and combined with the input (identity operation). Such an operation is supposed to bring information about spatial context at different granularities.

In our study, we evaluate how Spatial Context Measures change in response to modifications to the PPEG. We tested the following setups: no convolutions (meaning no

Table 10.3: The Spatial Context Measures (multiplied by 10 000) aggregated for all k for CLAM/TransMIL with different backbones. The measures are provided separately for normal and tumor images. Only the *Swin*-based backbones with *window_size* resulting in a higher SCM given the same backbone size (in CLAM) are provided.

	normal				**tumor**			
	$SCM_{features}$		$SCM_{targets}$		$SCM_{features}$		$SCM_{targets}$	
CLAM								
	mean	max	mean	max	mean	max	mean	max
sv2_base_w8	35	198	1	17	<u>67</u>	<u>479</u>	22	343
sv2_small_w16	<u>54</u>	<u>544</u>	<u>12</u>	<u>180</u>	62	380	19	179
sv2_tiny_w8	25	157	2	26	50	339	15	211
resnetCLAM	23	174	6	106	58	454	29	<u>401</u>
vit	25	97	3	44	63	412	<u>30</u>	303
TransMIL								
sv2_base_w8	<u>184</u>	1975	-40	64	<u>131</u>	1039	-21	601
sv2_small_w16	161	<u>2119</u>	-29	90	85	953	6	1078
sv2_tiny_w8	145	1879	-23	103	94	<u>1152</u>	3	1144
resnetCLAM	120	877	<u>43</u>	<u>1025</u>	124	752	<u>76</u>	1237
vit	100	1154	-12	684	92	977	20	<u>1303</u>

positional encoding is applied), conv 3x3, conv 5x5, conv 7x7 to check the role of each individual operation in the PPEG module and see if the granularity of context considered within positional encoding matters. For the study, we analysed $k = 8, 24, 48$ with the Deep Spatial Context. We focused on TransMIL applied on patch features extracted using *Resnet 50* as initially proposed by the authors of this MIL model.

First, we analyze the proportion of test samples for which Deep Spatial Context can be applied, as determined by a p-value in the Global Moran's I test being smaller than the significance level (α). It turned out that this proportion of test samples spatial regression can be applied is as follows in different setups: no convolution (no positional encoding) 97.5%, conv 3x3: 97.9%, conv 5x5: 99.4%, conv 7x7: 99.53%, full PPEG (original): 97.83%. Therefore, the spatial autocorrelation is present in the biggest number of samples when conv 7x7 is applied in PPEG. Interestingly, when full PPEG is applied (with all variants of convolutions) this ratio decreases suggesting that some negative interference between the outputs of convolutions appeared.

We also analysed mean $SCM_{features}$ and $SCM_{targets}$ for different variants of PPEG within TransMIL. It turned out that the ranking with decreasing values of mean Spatial Context Measures is as follows: conv 5x5, conv 3x3, conv 7x7, full PPEG, no conv.

Our analyses bring interesting insights into the model design and can be applied to gauge the model development process. We show that our study is a good complement to the ablation study in the TransMIL paper (section 4.2.1) where the effect of different variants of PPEG is analysed only based on AUC scores of the final tissue classification. We analyse the topic in more details investigating model internals.

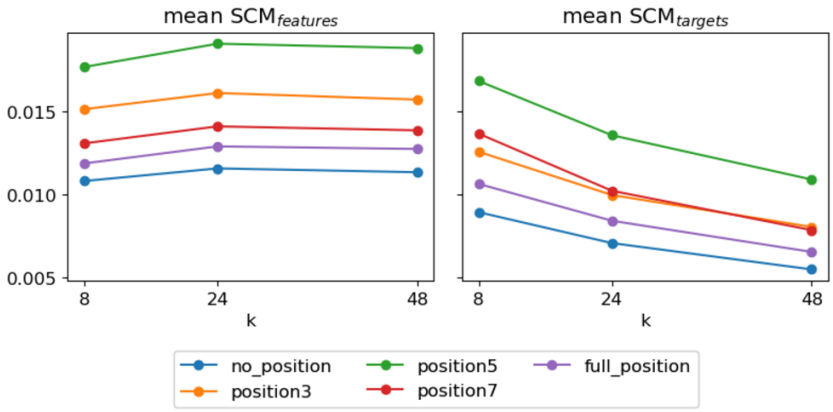


Figure 10.7: Mean Spatial Context Measures computed for TransMIL with different versions of positional encoding in PPEG module (by the application of conv 3x3, conv 5x5, conv 7x7, combination of all conv denoted as full_position) and no positional encoding (no_position).

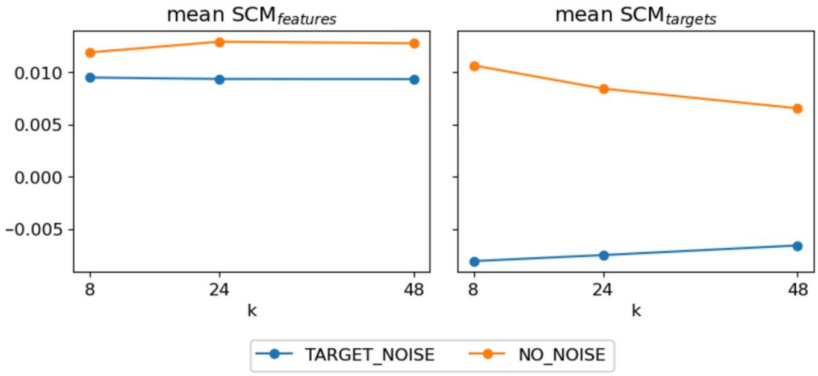


Figure 10.8: Mean Spatial Context Measures computed for original TransMIL (NO_NOISE) and TARGET_NOISE – the case where the trained attention scores are replaced with random values from Gaussian distribution $N(0, 1)$.

10.8.6 Additional baseline – TransMIL

We decided to also run additional experiments as a baseline to compare against. We took the original TransMIL model trained with full PPEG, and features extracted using *Resnet 50*. Then, we replaced the trained attention scores with random values from Gaussian distribution $N(0, 1)$ - we call it noise. This model is denoted TARGETS_NOISE. Note that attention scores are used as targets in the regression models within DSCon. The same analysis could not be performed in regards to features as replacing them with pure noise made the regression models unable to converge.

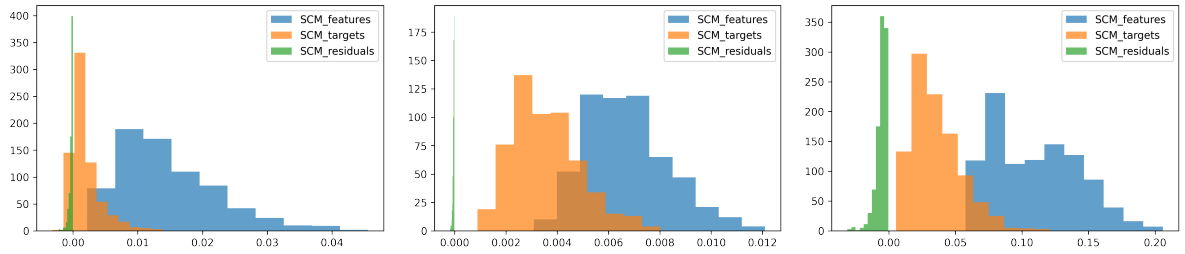
In the case of the original TransMIL, the spatial regression could have been applied to 97.8% of test samples (Global Morans' I test). In the case of random attention scores (targets in regression models), it was the case for only 5.4% of test samples. Figure 10.8 shows mean Spatial Context Measures for different values of neighbourhood k. It is observed that $SCM_{targets}$ is largely impacted by the application of random values – it has

negative values in the case of TARGETS_NOISE. In the case of $SCM_{features}$ the differences between TARGETS_NOISE and the original TransMIL are much smaller which is intuitive.

10.8.7 Evaluation on synthetic datasets

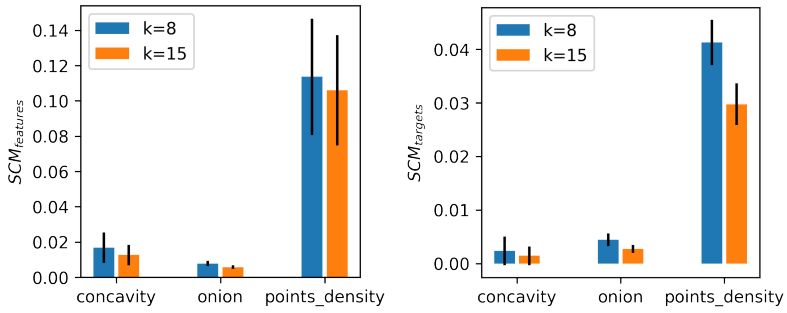
To show that the proposed *Deep spatial context* method is applicable to various scenarios, we run experiments on a set of synthetic datasets, SyntheticSMIL (Tomaszewska et al., 2025a). The datasets (*concavity*, *onion*, *points density*) are designed in such a way that the MIL model has to pay attention to multiple patches within the image to complete the task properly.
In the experiments, we followed the same procedure for dataset generation and CLAM training as the authors to get the highest model performance.

We provide the results computed on the test set using models trained on different cross-validation folds. In the study, we do not differentiate the images based on the classes they belong to as for all the classes, the understanding of the 'scene' was required. The comparison of the Spatial Context Measures for each dataset is presented in Figure 10.9. The same conclusions can be drawn as in the case of histopathological images: (1) $SCM_{residuals}$ are negligible, (2) the range of $SCM_{features}$ values is bigger than the one of $SCM_{targets}$. We analysed also the impact of the choice of k on the Spatial Context Measures (Figure 10.10). We verified only $k = 8$ and $k = 15$ as the images from the dataset consisted of 40×40 patches which is much less than in the case of analysed WSIs (Section 10.8). There is a clear relationship that the smaller k, the higher Spatial Context Measures which is in line with conclusions in previous experiments on digital pathology.



(a) *concavity* (72%) (b) *onion* (57%) (c) *points density* (100%)

Figure 10.9: Histograms of the Spatial Context Measures for different SyntheticSMIL datasets. The y axis is not shared among the plots for better visibility. In the brackets, the percentage of test images where the use of spatial regression models is justified (checked using Global Moran's I test) is provided.



(a) (b)

Figure 10.10: Mean $SCM_{features}$ (a) and $SCM_{targets}$ (b) with standard deviation. The colors refer to the neighbourhood size (k) in the spatial weight matrix. The y axis is not shared.

10.8.8 Local inspection on image level

Spatial statistics allow not only to analyze data from the global point of view but also from the local one. One of the methods is to use a local indicator of spatial autocorrelation (LISA) (Anselin, 1995) which is formally similar to the Global Moran's I statistic but the summation over all regions is removed. We provide a visualization of one of the images from the local perspective using LISA (Figure 10.11).

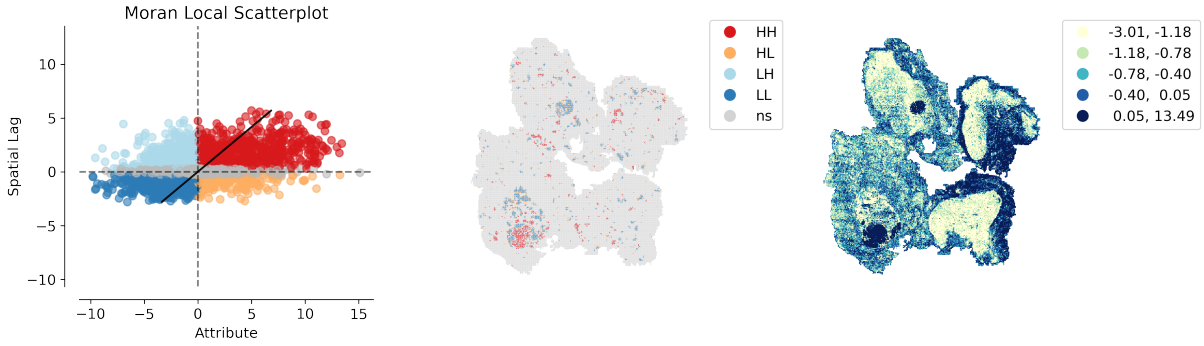


Figure 10.11: On the left, the Moran Local Scatterplot with the OLS residuals and the spatially lagged counterparts (*sv2_base_w8*, *k=24*). The slope of the linear fit corresponds to the Global Moran's I statistic. In the center, the Local Spatial Autocorrelation cluster map is provided on OLS residuals (*ns* – not statistically significant clusters, *HH* – means that the high value of the residual is surrounded by high values, *HL* – the high value is surrounded by low values which show negative spatial correlation. Note that in Moran's I statistics, we operate on deviations from the mean therefore the *high* and *low* refer to the values above and below the average. On the right, the Choropleth map is provided with the raw attention scores grouped into 5 quantiles.

10.9 Conclusion

We propose a *Deep spatial context* (*DSCon*) method with solid theoretic foundations for the investigation of the Multiple Instance Learning models from the perspective of spatial context. In the method, the inspection of the spatial aspect of the relationship between the features extracted from the patches and the attention scores within a DL classifier is performed.

The *DSCon* can be used for many use cases, i.e., inspection of different variants of models, models from different families or on subsets of data. We present the application of the method to histopathological images and synthetic data. In the former case, we observe that the spatial relationships are maintained by the models mostly in the case of tumor lesions in the CLAM model, which is in line with histopathologists' expertise. This observation does not hold for TransMIL. In the synthetic and histopathological cases, we observed that spatial relationships can be recovered mainly from the local context, $k = 8$. The spatial signal is larger within the patch features than within the neighboring attention scores (targets). Overall, the *DSCon* method enriches the toolkit for model diagnosis, helping to bridge the gap between model performance metrics and their underlying reasoning processes. We believe that the solid mathematical foundations of our methods are an important advantage.

Chapter 11

P8: Role of patch diversity while predicting time of death due to uveal melanoma metastasis

11.1 Preface

It can be difficult to find a needle in a haystack. It is often easier when the search space is smaller. The question arises: ***Will Multiple Instance Learning models trained on regions of interest selected by a physician instead of Whole Slide Images achieve higher performance?***
The short, but surprising answer is 'no'.

The 'needle-in-the-haystack' (NIAH) problem is nowadays often investigated in Large Language Models. It is analysed if models are able to identify the key information in the long input context which is not always strictly relevant. In general, the longer the context, the more difficult it is for a model to retrieve the correct information. Kamradt (2023) proposed one of the datasets to stress test models in this regard. In this case, Paul Graham essays serve as a 'haystack' with a sentence "The magic number is X" hidden somewhere within it as a 'needle'. The model is asked at the end of the prompt to finish the sentence "The magic number is". Later, many studies were conducted on the topic (Liu et al., 2024a; Wang et al., 2024b; Hengle et al., 2025). In a recent work, Wu et al. (2025) claim that the reason why some models can handle long context correctly and pass the NIAH test is the presence of the so-called retrieval heads (special set of attention heads) within LLMs. They are responsible for implementing the conditional copy algorithm from the input to the LLM output.

In Multiple Instance Learning paradigm, the models that deal with Whole Slide Images potentially can also experience the 'needle-in-the-haystack' problem as often there

are many input patches but only few contain the key information. It would suggest that limiting the context (aka number of patches) should result in increased model performance. However, still in Multiple Instance Learning, the role of individual patches remains poorly understood, even though it is critical for both model performance and clinical interpretability. In this paper, we investigate this topic using CLAM as a representative MIL model, comparing four input strategies: physician-selected regions of interest versus raw Whole Slide Images (WSIs) versus random patches versus top-k patches. We show that patch diversity is the key factor driving model performance. Importantly, the regions considered informative by the model do not align with physician-selected regions of interest, suggesting that CLAM captures complementary, non-obvious tissue patterns.

This work is positioned along the axes defined in Chapter 1 as shown in Table 11.1. I contrast the global (WSI-level) and semiglobal (based on subset of patches) approaches. The work was categorized as explainable AI because the analysis of model performance was followed by a detailed analysis of diversity measures at different granularity levels to explain the obtained performance results.

Table 11.1: Mapping of the paper by Tomaszewska et al. (2026) to the axes regarding context: type, scale, use case, ML stage.

	type	scale	use case	ML stage
context	temporal	global	histopathology	informed ML
	spatial	semiglobal	IQ tests	explainable AI
		local	tweets	
			other	

Paulina Tomaszewska*, Krzysztof Kotowski*, Piotr Donizy, Gustav Stålhammar & Przemysław Biecek. (—). "Role of patch diversity while predicting time of death due to uveal melanoma metastasis". version submitted to the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2026.

This paper has two co-first authors; specifically, Krzysztof Kotowski was responsible for preparing and analyzing the data from a medical perspective, while I was responsible for constructing AI models, analysing statistical results, and conducting a detailed analysis of the role of context in AI models.

The version of the paper provided below is an extended version of the one submitted to MICCAI where there are strict page limits.

11.2 Introduction

Uveal melanoma (UM) is the most common intraocular malignancy in adults (Krantz et al., 2017). It arises from pigment-producing cells (melanocytes) in the choroid. Approximately 50% of patients develop metastatic disease (Rietschel et al., 2005), which significantly impacts patient prognosis and results in a median survival of less than two years (Nathan et al., 2021). Primary site of uveal melanoma metastasis is liver (Carvajal et al., 2023). Importantly, the distant metastases are rarely found during primary tumor diagnosis. Instead, they are diagnosed in initial years after primary tumor treatment (median of 3 years), as well as decades later (Krantz et al., 2017; Rola et al., 2022). This underscores the critical need for Deep Learning (DL) systems capable to predict whether and when fatal metastasis will develop based on characteristics of the primary tumor. Recently, some studies have employed DL models to predict genetic factors associated with poor survival in UM patients (Sun et al., 2019; Liu et al., 2022a; Akram et al., 2024). Alternatively, few works predict vital status solely based on cyto- or histopathological images (Wan et al., 2023; Liu et al., 2025). These studies did not employ Multiple Instance Learning (MIL) models (Ilse et al., 2018). This image-only approach is more challenging, as genetic factors generally outperform clinical and histopathological predictors (Worley et al., 2007; Amaro et al., 2017).

Ensuring DL models use relevant information is essential for robustness, interpretability, and trust (Combi et al., 2022; Jin et al., 2023; Starke et al., 2025). Careful dataset design help reduce noise and spurious correlations (Geirhos et al., 2020). For example, DeGrave et al. (2021); Hryniewska et al. (2021) showed that models diagnosing COVID-19 sometimes relied on irrelevant regions outside the lungs, which could be mitigated by applying lung segmentation before classification.

Selecting regions of interest (ROIs) before inputting Whole Slide Images (WSIs) to the MIL models can also be relevant. WSIs are high-resolution scans that depict both tumor tissue (including malignant cells and non-malignant elements such as immune cells, erythrocytes, necrotic regions, subretinal fluid), as well as surrounding normal tissue (like cornea, choroid, retina, vitreous body) and slides background. Removing background or empty regions reduces computational costs and discards insignificant information. From a histopathologist's perspective, not all tissue regions are equally informative for diagnosis or prognosis. Thus, using patches selected by physician may provide a clearer signal and reduce spurious correlations (Nauta et al., 2022; Lin et al., 2024).

However, it remains unclear whether manual ROI selection improves model performance compared with learning from entire WSIs. In this study, we compare MIL models trained on the WSIs versus subsets of WSI patches (e.g selected by physician) to predict death associated with uveal melanoma metastasis in terms of its occurrence and time (Fig. 11.1). We show that (1) time of death due to uveal melanoma metastasis can be

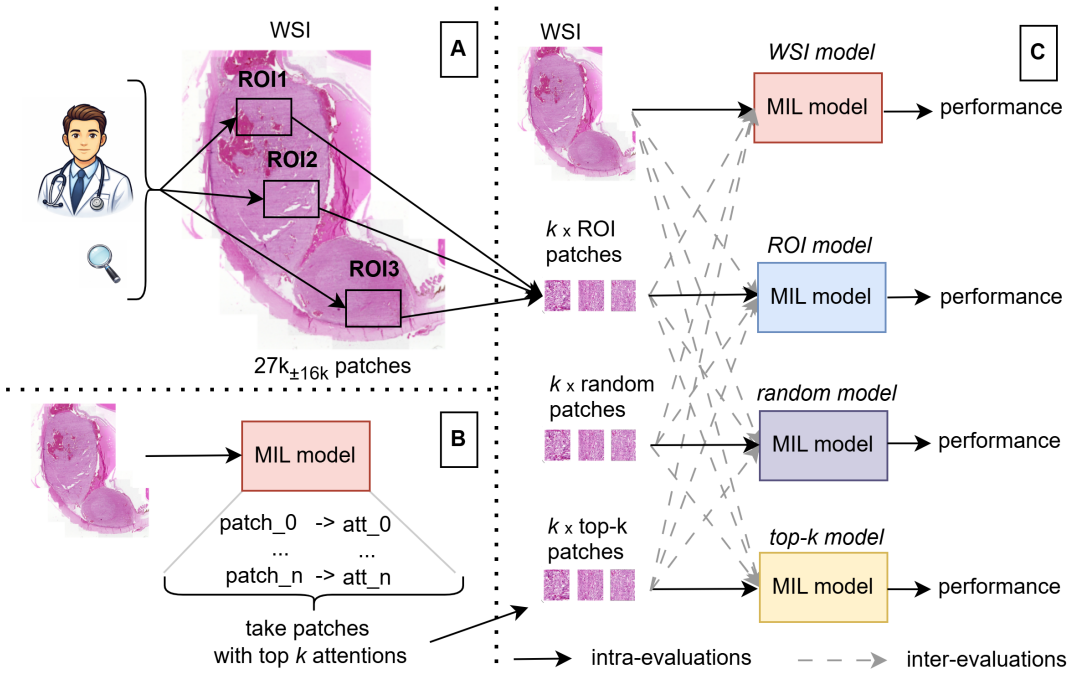


Figure 11.1: Multi-step experimental workflow. (A) A physician selects three regions of interest (ROIs) from each Whole-Slide Image (WSI). (B) A MIL model trained on WSIs (shown in the first row of panel C) is used to select the top-k patches. (C) Using the same MIL model architecture as in WSI-level training, we train the model from scratch on subsets of input patches selected using different strategies.

predicted with high performance based on primary tumor histology, (2) CLAM model is sensitive to input patch diversity, and (3) top-ranked patches identified by the model do not always correspond with the content selected by physician as region of interest. The code is available at `https://github.com/ptomaszewska/WSI_patch_diversity`.

11.3 Dataset and Task

We used a dataset of 565 Whole Slide Images from enucleated eyes stains from an archival patient cohort from St. Erik Eye Hospital and Karolinska Institutet (Stockholm, Sweden). On average, each WSI consists of $27,204_{\pm 16,380}$ patches. The dataset was collected over a period of 33 years. To the best of our knowledge, this is the largest dataset available on the topic.

Patient data was included in the study based on criteria: (1) previously confirmed diagnosis of uveal melanoma, (2) available enucleated eye stains in archive, (3) survival and metastasis information available, (4) patients were deceased. WSIs depict primary uveal melanoma tumor tissue with labels based on time of diagnosis of primary tumor and time and cause of death. Three categories were created: no metastasis ('N') – patient died on non-melanoma associated cause after more than 5 years since the initial diagnosis;

early metastasis ('EM') – patient died due to metastatic disease within 0-3 years after the initial diagnosis; and late metastasis ('LM') – patient died with melanoma-associated death occurring more than 3 years after diagnosis. Patients who deceased within 5 years on non-melanoma associated cause were excluded. The tissues slides were scanned with the Grundium Ocus40 digital scanner using the 40x objective. Scanning region was limited to uveal melanoma tumor area.

Additionally, a physician selected from each WSI three non-consecutive Regions of Interest (ROIs). Following criteria were applied: (1) region prevailingly consists of malignant uveal melanoma cells, (2) marginal tumor regions were excluded, (3) necrotic regions, subretinal fluid and lymphoid infiltrate regions were excluded, (4) areas were selected to demonstrate the most representative morphology of the malignant cells in the tumor. Each ROI contains at least 12 patches (on average $13.2_{\pm 3.7}$).

11.4 Model trained on WSI

We run numerous experiments with training different MIL models on WSIs. We investigated CLAM (Lu et al., 2021a) and TransMIL (Shao et al., 2021), two prominent attention-based models in digital pathology (Gadermayr and Tschuchnig, 2024). The two model families share the same input preprocessing procedure. We replaced feature extractor from original implementations with the ones pretrained specifically on histopathological data – UNI v1 (Chen et al., 2024) and CONCH v1 (Lu et al., 2024). In the case of CLAM, we tried the single-branch (SB) and multi-branch (MB) variants and with and without use of clustering as an auxiliary training. We investigated variants with dropout turned on and off. All other parameters in CLAM and TransMIL were set to the default values.

The mean AUC across different model variants ranges from 0.79 to 0.87, whereas mean accuracy differs from 0.63 to 0.74, which is satisfactory in the case of three-class classification on the roughly balanced dataset. The highest performance in the case of CLAM and TransMIL are achieved when UNI backbone was used. It can be potentially attributed to the fact that it extracts embedding of the bigger length (1024 vs. 768 in the case of CONCH). In our experiments, the performance of single-branch CLAM is higher than in the case of multi-branch version which is in line with the information from the CLAM's authors that SB often performs better. Overall, the best performance was in the case of CLAM with single-branch and auxiliary training in the form of clustering turned on and no dropout. The model was trained on features extracted using UNI backbone. The performance of this model was $AUC = 0.87_{\pm 0.04}$ and $acc = 0.74_{\pm 0.07}$. Additional advantage of the CLAM is the possibility to extract from the model a direct one-to-one mapping between input WSI patches and their corresponding attention scores (gated attention proposed by Ilse et al. (2018)) which gives room for further analyses.

11.5 Model trained on subsets of patches from WSIs

In subsequent experiments, the selected best model configuration was used to verify how the model performance changes depending on the subset of input WSI patches used for model training. Three scenarios are considered: ROI, top-k, and random input patches, yielding the *ROI*, *top-k*, and *random models*, respectively.

From each three ROIs selected from a particular WSI, the first n patches are selected in a grid manner. Later, the patches corresponding to the same tissue are concatenated resulting in k patches ($k \in \{\, 3n \mid n = 1, \ldots, 12 \,\}$) for each case. The set of top-k patches consists of k ($k \in \{\, 3n \mid n = 1, \ldots, 12 \,\} \cup \{\, 40 + 8m \mid m = 0, \ldots, 10 \,\}$) patches which were assigned the highest attention scores by the best model configuration (described in Sec. 11.4, called *WSI model*). There are 5 different sets of top-k patches which reflect the fact that the 5-fold cross-validation was used. Lastly, the k ($k \in \{\, 3n \mid n = 1, \ldots, 12 \,\} \cup \{\, 40{+}8m \mid m = 0, \ldots, 10 \,\}$) patches were randomly selected from WSI. This was performed in an additive manner, such that the random set of k patches included all patches present in the subset of size m ($m < k$). There are three sets of random patches as three different random seeds were used. The ranges of k were chosen to align ROI patches with other variants. In all cases, the number of labeled training samples was the same; only the number and content of patches per sample (bag) differed.

11.5.1 Results

Intra-evaluations

The best model variant from Sec. 11.4 was trained and evaluated on the same patch subset variant (top-k, random, ROI) using the same cross-validation splits. AUC and accuracy were aggregated across folds, and for random patches additionally averaged over three random seeds. For the top-k, the model trained on patches extracted using the *WSI model* for fold f was subsequently only trained and evaluated on the same fold f split.

The results (Fig. 11.2) indicate that all models perform trained and evaluated on subsets of patches perform substantially worse than the *WSI model*. For $k \leq 56$, the *top-k model* achieves the highest AUC among subset-based approaches, whereas for larger k, the *random model* becomes dominant. A similar pattern is observed for accuracy. The *top-k* and *ROI* models exhibit comparable accuracy for $k \leq 15$, but for larger k, the *ROI model* underperforms the *top-k model*. From $k = 56$ onward, the *random model* attains the highest accuracy and AUC among all subset-based methods. *Random model* is the only model that improves with increasing numbers of input patches.

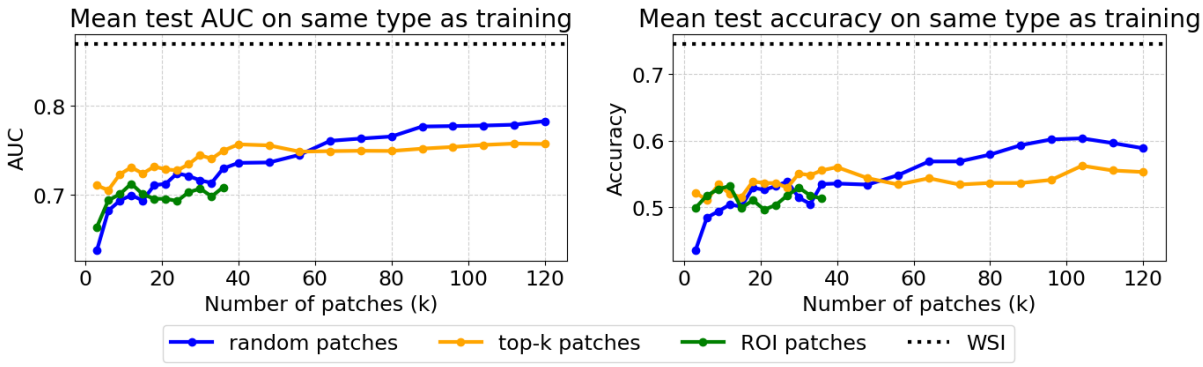


Figure 11.2: Mean performance on test set in intra-evaluations.

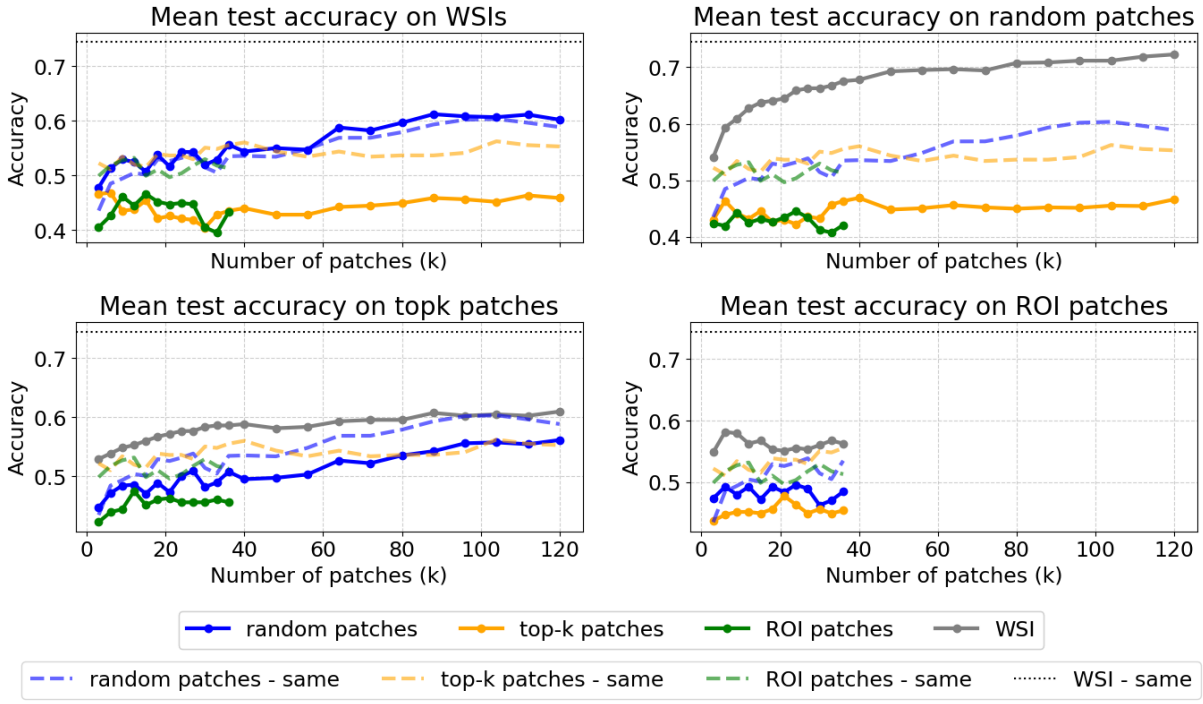


Figure 11.3: Mean accuracy on test set in inter-evaluations.

Inter-evaluations

The models were also subjected to cross-type evaluations where testing was performed on other input data variant than during training. The results are shown in Fig. 11.3. We report accuracy only, as it is more relevant from an application perspective. The dashed lines in lighter color depict the performance of intra-evaluations (from Fig. 11.2). The inter-evaluation on WSIs was performed so that the models trained on different subsets of k patches were always tested on the same set of WSIs. Meanwhile, the inter-evaluations on random, top-k and ROI patches was performed in a k-wise manner. It means that a model trained on a subset of k patches selected according to a particular strategy or WSIs was tested on a subset of k patches selected using other strategy.

Evaluation on WSIs The highest performance is achieved by *random model*. Its performance on WSIs is on par with its intra-performance. Such alignment between inter- and intra-performance is not observed in the case of other models.

Evaluation on random patches The accuracy of *WSI model* on random patches converges to its performance on WSIs. Unexpectedly, the performance of the *WSI model* is much higher than the intra-performance of *random model*. Accuracy of *top-k* and *ROI models* is similar but much lower than their corresponding intra-performance and accuracy of *WSI model*.

Evaluation on top-k patches The highest performance was achieved by the *WSI model*, exceeding the intra-performance of the *top-k model* but reaching the intra-performance of *random model* for larger k. The *random model* outperformed the *ROI model*; however, both performed worse than their respective intra-performance. Overall, the *ROI model* exhibited the lowest performance among all model variants in this inter-evaluation.

Evaluation on ROI patches Here, the smaller set of k values was analyzed due to small size of ROIs. The performance of models is stable across k. The highest performance has the *WSI model*, later in the ranking is *random model* which has performance slightly worse than intra-performance of *ROI model*. Lastly, there is *top-k model* with lower performance.

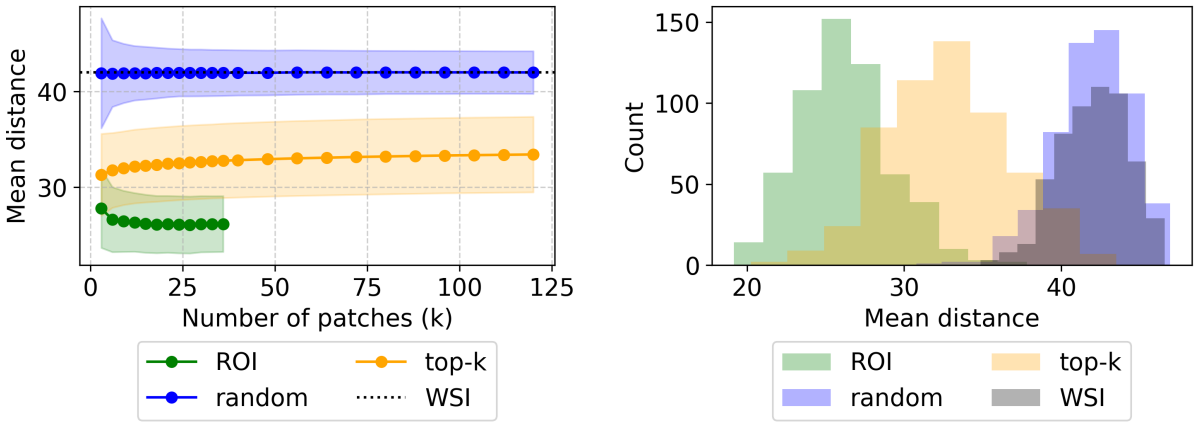
Summary Intra-evaluation performance exceeds inter-evaluation performance, except when a random-patch model is tested on WSIs. Models trained on top-k and ROI patches generalize poorly to all other input types, indicating limited ability to capture global discriminative patterns. It is likely that the distribution of patches used for training is too narrow in these scenarios, which may explain the low cross-type performance due to the out-of-distribution nature of the evaluation. These initially unexpected observations will be confronted with the diversity measures.

11.6 Patch inspection

11.6.1 Technical perspective on patch diversity

We conduct analyses of the intra- and inter-diversity measures of different input data types.

Intra-diversity The results as a mean across all samples in the dataset are in Fig. 11.4. It is clear that the highest intra-diversity is observed in random variant which is on par



(a) Intra-diversity vs. k

(b) Histogram at $k = 36$

Figure 11.4: Mean Euclidean distance between patches constituting a sample.

with WSI variant. The top-k intra-diversity is much lower, whereas the smallest is in the ROI variant.

Additionally, the sample-level diversity is defined as $D_{\text{sample-level}}(i) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|\bar{\mu}_i - \bar{\mu}_j\|_2$, where $\bar{\mu}_i$ is the mean patch representation of sample i and N is the total number of samples. The results show that $D_{sample-level}$ in the case of random patches drops from 31.2 to 19.4 at $k = 120$ and converges to the WSI baseline (19.0). The curve corresponding to ROI variant starts at 34.7 and quickly stabilizes reaching 32.3 at $k = 36$. The top-k curve begins highest (38.3) and steadily decreases to 32.1 at $k = 36$ and 30.9 at $k = 120$ but remains well above random and WSI.

Inter-diversity is computed as the mean over the entire dataset of the pairwise Euclidean distance between the mean patch representations of a given sample in two different data variants (top-k, random, ROI, WSI), i.e., $D_{\text{inter-level}}(i) = \frac{1}{N} \sum_{i=1}^{N} \|\mu_i^{(A)} - \mu_i^{(B)}\|_2$, where N is the number of samples and $\mu_i^{(A)}$ and $\mu_i^{(B)}$ are the mean patch representations of sample i in data variant A and B, respectively. The results show that mean ROI samples has the highest mean pairwise overall distance to mean samples of top-k (26.3) and moderate values with random (21.2) and WSI (19.6). In contrast, WSI and random have similar mean distance to top-k (∼21–22), with small mean distance among themselves (5.9).

The analyses of diversity scores computed at various levels of granularity show that the random patches are the most similar to WSIs which results in the highest inter-evaluation performance between the two.

11.6.2 Medical perspective on patch selection

In addition, we analyzed the content of the selected patches from a histopathological perspective. For this, we annotated all the top-k patches ($k = 3$) from fold 0, which

Table 11.2: Mean ± standard deviation of normalized label distributions across categories assigned by a physician to top-k patches ($k = 3$), computed from the training, validation, and test sets. Column names follow the $label_{prediction}$ convention; N_{EM} was not observed in fold 0. Categories include B/N/P (bleeding & necrosis & pigmentation), tumor cell types (epithelioid, spindle A/B), and fibrous strands. The last row reports the total number of patches.

Cat.	$\mathbf{EM}_N$	$\mathbf{EM}_{EM}$	$\mathbf{EM}_{LM}$	$\mathbf{LM}_N$	$\mathbf{LM}_{EM}$	$\mathbf{LM}_{LM}$	$\mathbf{N}_N$	$\mathbf{N}_{LM}$
B/N/P	$0.00_{\pm 0.00}$	$0.32_{\pm 0.05}$	$0.04_{\pm 0.06}$	$0.01_{\pm 0.02}$	$0.02_{\pm 0.04}$	$0.22_{\pm 0.10}$	$0.37_{\pm 0.05}$	$0.02_{\pm 0.04}$
epith.	$0.00_{\pm 0.00}$	$0.28_{\pm 0.10}$	$0.05_{\pm 0.02}$	$0.05_{\pm 0.06}$	$0.04_{\pm 0.03}$	$0.40_{\pm 0.02}$	$0.16_{\pm 0.05}$	$0.02_{\pm 0.02}$
fib.	$0.07_{\pm 0.08}$	$0.04_{\pm 0.07}$	$0.04_{\pm 0.08}$	$0.11_{\pm 0.09}$	$0.00_{\pm 0.00}$	$0.29_{\pm 0.16}$	$0.46_{\pm 0.07}$	$0.00_{\pm 0.00}$
sp. A	$0.00_{\pm 0.00}$	$0.32_{\pm 0.10}$	$0.04_{\pm 0.06}$	$0.06_{\pm 0.09}$	$0.04_{\pm 0.06}$	$0.16_{\pm 0.10}$	$0.31_{\pm 0.07}$	$0.07_{\pm 0.13}$
sp. B	$0.00_{\pm 0.00}$	$0.47_{\pm 0.17}$	$0.00_{\pm 0.00}$	$0.02_{\pm 0.03}$	$0.11_{\pm 0.19}$	$0.19_{\pm 0.17}$	$0.21_{\pm 0.18}$	$0.00_{\pm 0.00}$
n	9	483	45	63	30	555	486	24

features the highest patch-level intra-diversity. The patches were categorized according to Callender's classification (Callender, 1931; McLean et al., 2018).The results are in Table 11.2. UM tumors containing epithelioid cells are associated with a poorer prognosis than tumors with only spindle cells (Kroll and Kuwabara, 1965) which mostly aligns with our results. Only 18% of top-k patches with epithelioid cells were observed within 'N' class, compared to 38% of spindle A and 21% of spindle B patches. Small standard deviation means there is no significant difference between training, validation and test sets. The number of test patches assigned to particular category is as follows: bleeding & necrosis & pigmentation ($n = 84$), tumor cells (epithelioid ($n = 99$), spindle A ($n = 36$), spindle B ($n = 9$)) and fibrous strands ($n = 27$).

The results were confronted with the annotations of the ROI patches ($k = 3$) from the test set (fold 0). These are the results on normalized label distribution across categories and the corresponding number of patches: epithelioid tumor cells ('N': 0.31, 'EM': 0.28, 'LM': 0.41, $n = 123$), pigmentation ('EM': 0.33, 'LM': 0.67, $n = 18$), spindle A tumor cells ('N': 0.23, 'EM': 0.33, 'LM': 0.44, $n = 39$), spindle B tumor cells ('N': 0.41, 'EM': 0.37, 'LM': 0.21, $n = 75$).

The number of patches within histopathological categories of top-k and ROI patches differs. While epithelioid tumor cells were the most frequent category in both, top-k patches more often include non-tumor or microenvironmental elements such as necrosis, bleeding, and occasionally fibrous strands. It aligns with Fig. 11.4 showing that diversity of top-k patches is higher than of ROI patches. The aforementioned non-tumor structures were largely excluded from physician-defined ROIs, which is consistent with routine histopathology assessment, where tumor cell type and nuclear morphology are established prognostic features in UM (Kroll and Kuwabara, 1965) and guide ROI selection. It remains unclear why the model assigns high attention to non-tumor elements. One possible explanation is that necrosis and hemorrhage correlate with more aggresively growing

or larger tumors and therefore carry prognostic information captured by the WSI-level model.

11.7 Conclusions

We show that time to UM–related death due to metastases can be predicted from primary tumor histology using a MIL model with high performance when Whole Slide Images are the input. They provide full contextual information. We compare a model trained on WSIs with variants trained on subsets of WSI patches selected randomly, by physician (ROI), or based on attention scores (top-k). Our results show that CLAM model, as a representative MIL approach, is highly sensitive to the diversity of input patches. Reducing patch diversity through manual selection leads to diminished performance, highlighting the importance of preserving histopathological heterogeneity. In addition, the patches assigned the highest attention scores often contained tissue elements that were not prioritized in the physician-defined ROIs. These findings highlight the importance of carefully selecting data in DL pipelines for predicting UM prognosis.

Chapter 12

Summary

In the dissertation, I focused on the investigation of the role of temporal and spatial context in classification using deep neural networks. The context (mainly temporal and textual) is commonly addressed in the architectures of Deep Learning solutions (Chapter 2.2). However, there is much less focus on investigation of the role of context (especially within images) in models. Most of the time, it is assumed that context in models processing images is an additional set of values that can be fused together with image data. For instance, X-ray images of lungs can be accompanied with information on patient characteristics (age, sex, 'pack-years' and some blood results). However, it is overlooked that the images themselves contain many complex relationships between depicted objects that Deep Learning model can benefit from. Even if models use this information in the decision-making process, we may be unaware of it, because we lack dedicated methods to inspect models from this perspective.

I identified the limitations of the popular explainable AI techniques. I argue and give examples that the techniques do not provide satisfactory model explanations in the case where the classification task requires understanding of spatial relationships between objects in the input image. As a result, in the position paper, I postulated the change of the paradigm in XAI from 'where' the key regions in the input image are located to 'how' they are located towards each other. This is a cornerstone for the subfield called *spatial XAI*. The methods from this field should be applied for investigation of systems in such critical domains as healthcare, street surveillance, autonomous driving etc.

In the research I built upon my postulates and focused on investigation of one type of context within images that is spatial (mainly the neighborhood subcategory) in the digital pathology use case. In this field, I proposed a set of three datasets (SyntheticSMIL), evaluation protocol (ReasonSMIL), two spatial XAI methods (Deep Spatial Context and occlusion-based applied to patch-level investigation). In addition, I performed a comparative analysis of semiglobal vs. global context-based approaches when predicting time of death due to metastases and provide the observation that it is possible to restore the tissue structure based on patch features.

The SyntheticSMIL dataset collection was designed to operationalize the term 'spatial Multiple Instance Learning' (SMIL), which I introduced inspired by the earlier idea of 'correlated Multiple Instance Learning' by Shao et al. (2021). The datasets were created in such a way that the information needed to perform classification is visible on more than one patch, so that the correct classification requires the understanding of the image compositionality. Such synthetic datasets allowed for in-depth model evaluation thanks to the available fine-grained ground-truth information. The datasets were also used when evaluating the models using the proposed Deep Spatial Context method.

The proposed evaluation protocol (ReasonSMIL) is applied to the investigation of Multiple Instance Learning models to check if they understand the compositionality of the input image (ReasonSMIL-Reasoning) and if models trained on various cross-validation folds are similar in this regard (ReasonSMIL-Agreement). Although ReasonSMIL was originally applied in scenario where models were trained on synthetic datasets (SyntheticSMIL) to provide a more controlled setup, its applicability extends beyond this. It can be employed in any setting where fine-grained masks are available, including Whole Slide Images.

The proposed spatial XAI techniques differ in use case and scale of context that they are designed to investigate. In the following works, the analyses were performed on histopathological images. The Deep Spatial Context allows for an investigation of the MIL model internals (attention scores) from the semiglobal and global context point of view depending on the defined value of k in the k-nearest neighbors method used to define spatial weight matrix. In the semiglobal case, the setup with the minimal value of $k = 8$ (the Queen's case) is considered, whereas larger values of k correspond to the global context. The Deep Spatial Context method enables the examination of model reasoning across different data subsets, facilitates comparisons between model variants and architectures, and allows for assessing the extent to which model reasoning aligns with common-sense expectations.
The occlusion-based method allows the analysis of model sensitivity to the limitation of the local context defined by the size of the neighborhood tissue around the histopathological patch. The method can be used to inspect the consistency of predictions over different context sizes, resulting in identification of three groups of patches (the ones undergoing just one change of prediction by a classifier, 'swinging' and with constant prediction).
The methods expand the currently limited set of techniques available in the spatial XAI arsenal making it a valuable contribution. This is particularly important given the Deep Spatial Context method's reliance on strong mathematical foundations, which are still uncommon in the field of XAI.

Moreover, I conducted a scrutinized comparison of two approaches regarding the input data to system: whole tissue vs. preselected regions. It led to important practical observation that the time-consuming selection of regions of interest does not result in

performance gains. It suggests that contextual information in WSIs and its diversity are important.

In addition to the histopathological use case, I completed one more project on the spatial context. I introduced a Deep Learning architecture (Duel-IQ) designed to solve the visual IQ tests that inherently rely on the understanding of spatial relationships observed within rows, columns and/or diagonals e.g. orientation (relative angular position).

Lastly, I did one project on temporal context. I introduced the LIMES method for a lightweight adaptation of the continuously trained model using one informative historical record which serves as temporal context. The method includes the algorithm for selection of the temporal context and its injection into model in the adaptation scheme to tackle the class-prior shift.

To conclude, the dissertation is a multi-faceted investigation of the role of the context in deep neural networks performing classification. The summary of the outcome of my dissertation is presented in the scheme in Figure 1.2 in Chapter 1. Contextual information was observed to play an important role at various scales (from local to global) in classifying histopathological images. It was clearly seen in the case of models analysing tissue patches whose performance largely dropped when the size of local context (area surrounding tissue patch) was limited. It turned out that the Multiple Instance Learning models analysing images of high resolution can understand image compositionality and neighborhood context to some extent, which is backed with the results of Deep Spatial Context method and ReasonSMIL evaluation protocol. It is to be highlighted that Deep Learning models (CLAM and TransMIL) differ in this regard. CLAM model is more consistent than the TransMIL when paying attention to patches (compositionality) when trained on different cross-validation folds. Moreover, the analysis of Spatial Context Measures showed that on average CLAM perceives and maintains the context within targets (attention scores) in a bigger extent than TransMIL unlike in the case of the feature space. These are the interesting observations, as in CLAM there is no information about positional encoding unlike in TransMIL where thanks to the use of grouped convolution (with various kernel sizes), it is implicitly encoded.
In another study, it was shown that performance of MIL model trained on regions of interest (ROIs) extracted by a physician had much lower performance than analogous model trained on Whole Slide Images. It can be attributed to high homogeneity of the ROIs. Moreover, it turned out that when MIL model is exposed to WSIs, it pays significant attention not only to types of tissues perceived as important according to current medical guidelines but also other types which may suggest that more holistic information is preferred by a model. Additionally, it was observed that high diversity of the input patches to MIL models is crucial for the good generalization.

Apart from the investigation of the spatial context, it was shown in the thesis that

the use of temporal context for the model adaptation can result in a boost of the model performance when trained in a continuous manner under the class-prior shift.

Chapter 13

Discussion and Future Work

13.1 Other types of model architectures

In the dissertation, I focused on the investigation of purely Computer Vision models. However, images can also be analysed using Vision Graph Neural Networks (ViG) (Han et al., 2022). The fact that different parts of images are connected and there is an information flow between them that can be traced seems like an interesting setup for the exploration of interactions between the regions (perceived as spatial context). Due to a different nature of ViG networks from the vision models (convolutional- and transformer-based), they would probably require other techniques to investigate them from the explainability point of view.

Another direction for the extension of the exploration of the topic of spatial understanding in neural networks processing images would be the analysis of the Multimodal Models (Vision Language Models). The fact that the models deal jointly with text and images, would probably make their exploration easier as one can simply ask about the reasoning of the model behind a particular answer. However, as pointed out by Turpin et al. (2023), the output of the Chain-of-Thought may not provide reliable explanations.

13.2 Conflicting information

As described in Section 2.2 with the example of misperception of the said sound (McGurk effect), it may happen that the context and the main message will be in conflict. One of examples is situation where knowledge encoded in LLM weights and the one provided in the prompt are contradictory. It can lead to hallucinations (Sun et al., 2025).

Striking the right balance between contextual information and the main input signal can be challenging at times. Below I present two problematic scenarios: the lack of proper world model (context encoded within the model) and too strong learnt context overshadowing what it actually provided in the input to the model.

The illustrative example of the first scenario is a case where a model for detection is given an image where there is a cow occluded by a wall so that only a head and back of an animal is visible. It may happen that a DL model detects in the image a 5.2 meter long cow[1]. However, it is a well-known fact that such long cows do not exist. Regardless of whether the input image was modified on purpose by a user (adversarial modification) or it was an original image, the key issue here is that an intelligent DL model should have an understanding that there are no such long animals of this kind. Therefore, the model should be cautious when making the prediction. Maybe the model should state that the image is fake or come to the conclusion that due to such a large occlusion, it is not obvious if there is just one cow in the image. Such reflections would require a wider knowledge about the world and deeper reasoning. To mitigate risk of such model mistakes, probably the use of neurosymbolic AI or other methods to incorporate world model (proper context) into AI would be needed.

The contrary situation is described by Vo et al. (2025). Here, the knowledge learnt by Vision-Language Models during training dominates over the information present in the input images. As a result, if a model is provided with an image of an elephant with 5 legs, it answers that the animal has 4 legs due to the strong prior knowledge. Similarly, if a model is given a flag with stripes similar to the flag of the USA and is asked about the number of stripes, it will answer 13 as it is the number of stripes in a flag of the USA instead of 14 which is the number of stripes that are indeed visible in this modified flag in the image. It may be concluded that in these cases, a model instead of analysing solely the input image is trying to map it to the most similar scenario observed during training and replies based on the training sample instead of the actual sample. It suggests over-reliance on prior knowledge, which in these cases was not needed at all. The tasks required just the analysis of the provided images without digging into any contextual information.

These two examples show that it is important to correctly balance between contextual information (sometimes encoded already in the model weights) and actual input.

In this dissertation, the role of context was carefully studied. The information that contextual information is beneficial when analysing tissue to make diagnosis/ prognosis originated from histopathologists. That is why the SyntheticSMIL (described in Chapter 8) datasets were designed in such a way that understanding of spatial context was helpful when classifying images opposed to scenario where it would be a distractor (conflicting source of information). The slightly different approach was taken in Chapter 10. The spatial regression models (core element of Deep Spatial Context method) are given information about the neighboring patches within input, but they are not forced to use it. If the model uses the contextual information that is not beneficial in particular case, its

[1] `https://www.linkedin.com/posts/paolo-foglio-694aab174_an-ai-has-identified-a-cow-that-is-over-5-activity-7137455140386594816-WAxA/`, accessed: 2025-08-15

performance is lower than in the case of OLS regression which overall results in negative Spatial Context Measures. Such cases were also reported in a paper along the positive numbers to provide the holistic information on the phenomenon. Moreover, in Chapter 7, despite the clear trend that with the increase of the size of the occluded area, the performance of the model drops, some fluctuations were observed and reported.

In future work, I believe it would be interesting to conduct analyses of scenarios with contextual information clearly conflicting with the main signal. Potentially, research on the issue from the explainable AI/ Mechanistic Interpretability point of view could lead to follow-up, where one could propose some adaptable steering mechanisms to compensate for the conflicting situations within neural networks.

Bibliography

Abdelmalak, I., Madhusudhanan, K., Choi, J., Stubbemann, M., and Schmidt-Thieme, L. (2025). Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? *arXiv:2502.09683*.

Abousamra, S., Gupta, R., Kurc, T., Samaras, D., Saltz, J., and Chen, C. (2023). Topology-Guided Multi-Class Cell Context Generation for Digital Pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abrahams, E., Snow, T., Siegfried, M. R., and Pérez, F. (2024). A Concise Tiling Strategy for Preserving Spatial Context in Earth Observation Imagery. *Machine Learning for Remote Sensing (ML4RS) Workshop at the International Conference on Learning Representations (ICLR)*.

Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. (2022). From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5:1006 – 1019.

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S., and Sowa, J. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*, 33:25–42.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2017). Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Akram, F., de Bruyn, D. P., van den Bosch, Q. C. C., Trandafir, T. E., van den Bosch, T. P. P., Verdijk, R. M., de Klein, A., Kiliç, E., Stubbs, A. P., Brosens, E., and von der Thüsen, J. H. (2024). Prediction of molecular subclasses of uveal melanoma by deep learning using routine haematoxylin-eosin-stained tissue slides. *Histopathology*, 85(6):909–919.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Albuquerque, T., Yüce, A., Herrmann, M. D., and Gomariz, A. (2024). Characterizing the Interpretability of Attention Maps in Digital Pathology. *arXiv:2407.02484*.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805.

Alicioglu, G. and Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics*, 102:502–520.

Almeida, P. R., Oliveira, L. S., Britto, A. S., and Sabourin, R. (2018). Adapting dynamic classifier selection for concept drift. *Expert Systems with Applications*, 104:67–85.

Alves, D., Mendonça, F., Mostafa, S. S., Rodrigues, P., and Rocha, A. (2025). A deep learning approach for improving spatiotemporal resolution of numerical weather prediction forecasts. *Scientific Reports*, 15:32796.

Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8:53.

Amaro, A., Gangemi, R., Piaggio, F., Angelini, G., Barisione, G., Ferrini, S., and Pfeffer, U. (2017). The biology of uveal melanoma. *Cancer and Metastasis Reviews*, 36(1):109–140.

American Medical Association (2023). AMA Augmented Intelligence Research, Physician sentiments around the use of AI in health care: motivation, opportunities, risks, and use cases.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Springer Dordrecht.

Anselin, L. (1995). Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L., Sridharan, S., and Gholston, S. (2007). Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns. *Social Indicators Research*, 82(2):287–309.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. (2024). Refusal in language models is mediated by a single direction. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Arras, L., Osman, A., and Samek, W. (2022). CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40.

Avolio, M., Fuduli, A., Vocaturo, E., and Zumpano, E. (2024). A comparative study of linear type multiple instance learning techniques for detecting COVID-19 by chest X-ray images. *Progress in Artificial Intelligence*, 13(1).

Azaria, A. and Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP*.

Bach, S., Binder, A., Montavon, G., Müller, K.-R., and Samek, W. (2015). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bai, B., Liang, J., Zhang, G., Li, H., Bai, K., and Wang, F. (2021). Why attentions may not be interpretable? In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21.

Bai, N., Iyer, R. A., Oikarinen, T., Kulkarni, A. R., and Weng, T.-W. (2025). Interpreting neurons in deep vision networks with language models. *Transactions on Machine Learning Research*.

Balazevic, I., Shi, Y., Papalampidi, P., Chaabouni, R., Koppula, S., and Henaff, O. J. (2024). Memory consolidation enables long-context video understanding. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Balazevic, I., Steiner, D., Parthasarathy, N., Arandjelovic, R., and Henaff, O. J. (2023). Towards in-context scene understanding. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Bao, Y., Zhang, X., Du, T., Zhao, X., Feng, Z., Peng, H., and Yin, J. (2025). Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. *arXiv:2506.00823*.

Barbosa, D., Ferreira, M., Junior, G. B., Salgado, M., and Cunha, A. (2024). Multiple instance learning in medical images: A systematic review. *IEEE Access*, 12:78409–78422.

Bardool, K., Tuytelaars, T., and M., J. O. (2019). A Systematic Analysis of a Context Aware Deep Learning Architecture for Object Detection. In *Proceedings of the Benelux Conference on Artificial Intelligence (BNAIC 2019) and the Belgian Dutch Conference on Machine Learning (Benelearn)*.

Barnea, E. and Ben-Shahar, O. (2019). Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barrett, D., Hill, F., Santoro, A., Morcos, A., and Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bartlett, P. L. (1992). Learning with a slowly changing distribution. In *Workshop on Computational Learning Theory (COLT)*.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations . In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., Chard, K., and Foster, I. (2024). Comprehensive Exploration of Synthetic Data Generation: A Survey. *arXiv:2401.02524*.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., et al. (2016). Deepmind lab. *arXiv:1612.03801*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. (2022). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

Benčević, M., Habijan, M., Galić, I., Babin, D., and Pižurica, A. (2024). Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044.

Bereska, L. and Gavves, S. (2024). Mechanistic Interpretability for AI Safety – A Review. *Transactions on Machine Learning Research*.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., and Watrin, P. (2022). Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Biecek, P. and Samek, W. (2024). Position: explain to question not to justify. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Biecek, P. and Samek, W. (2025). Model science: getting serious about verification, explanation and control of ai systems. In *Frontiers in AI track at European Conference on Artificial Intelligence (ECAI) 2025*.

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177.

Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016). Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In *Proceedings of the Information Science and Applications (ICISA)*.

Bissell, M. J. and Radisky, D. C. (2001). Putting tumours in context. *Nature Reviews Cancer*, 1:46–54.

Bo, Z., Coates, L. C., and Papież, B. W. (2026). Interpretable rheumatoid arthritis scoring via anatomy-aware multiple instance learning. In *Applications of Medical Artificial Intelligence*, pages 52–62, Cham. Springer Nature Switzerland.

Bodnár, A., Gulyás, L., and Kárász, Z. (2024). Understanding geometric relationship concepts innbsp;few-shot learning. In *Proceedings of the Computational Collective Intelligence: 16th International Conference (ICCCI)*.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2020). On Identifiability in Transformers. In *Proceedings of International Conference on Learning Representation (ICLR)*.

Bu, Z., Wang, Y.-x., Zha, S., and Karypis, G. (2024). Differentially private bias-term fine-tuning of foundation models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Callender, G. R. (1931). Malignant melanotic tumors of the eye. a study of histologic types in 111 cases. *Transactions of the American Academy of Ophthalmology and Otolaryngology*, 36:131–142.

Carvajal, R. D., Sacco, J. J., Jager, M. J., Eschelman, D. J., Olofsson Bagge, R., Harbour, J. W., Chieng, N. D., Patel, S. P., Joshua, A. M., and Piperno-Neumann, S. (2023). Advances in the clinical management of uveal melanoma. *Nature Reviews Clinical Oncology*, 20(2):99–115.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., and Bloom, J. I. (2024). A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Interpretable AI: Past, Present and Future*.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. (2019). This looks like that: deep learning for interpretable image recognition. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F. (2022). Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*.

Chen, S.-A., Li, C.-L., Arik, S. O., Yoder, N. C., and Pfister, T. (2023). TSMixer: An all-MLP architecture for time series forecast-ing. *Transactions on Machine Learning Research*.

Cheng, X., He, X., Qiao, M., Li, P., Hu, S., Chang, P., and Tian, Z. (2022). Enhanced contextual representation with deep neural networks for land cover classification based on remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 107:102706.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Cho, I. and Hockenmaier, J. (2025). Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models. In Christodoulopoulos, C.,

Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28961–28973, Suzhou, China. Association for Computational Linguistics.

Cho, J., Zala, A., and Bansal, M. (2022). DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*.

Chrabąszcz, M., Baniecki, H., Komorowski, P., Płotka, S., and Biecek, P. (2025). Aggregated Attributions for Explanatory Analysis of 3D Segmentation Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Chu, X., Tian, Z., Zhang, B., Wang, X., and Shen, C. (2023). Conditional positional encodings for vision transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In *Proceedings of ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Cliff, A. D. and Ord, J. K. (1973). *Spatial Autocorrelation*. Pion.

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., and Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 133.

Costa, J. C., Roxo, T., Proença, H., and Inácio, P. R. M. (2024). How Deep Learning Sees the World: A Survey on Adversarial Attacks Defenses. *IEEE Access*, 12:61113–61136.

Crabbé, J. and van der Schaar, M. (2022). Concept Activation Regions: A Generalized Framework For Concept-Based Explanations. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., and Love, B. C. (2023). A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, 166:164–171.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., and Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, 600:70–74.

Dee, E., Goglia, A., Swami, N., Nguyen, B., Hougen, H., Khan, A., Kishan, A., Punnen, S., Nguyen, P., Mahal, B., and Alshalalfa, M. (2023). Determinants of Widespread Metastases and of Metastatic Tropism in Patients with Prostate Cancer: A Genomic Analysis of Primary and Metastatic Tumors. *International Journal of Radiation Oncology, Biology, Physics*, 117(2, Supplement):e375–e376. ASTRO 2023: 65th Annual Meeting.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619.

Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Workshop on Multiple Classifier Systems*, MCS '00, page 1–15, Berlin, Heidelberg. Springer-Verlag.

Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised Visual Representation Learning by Context Prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Dogra, V., Verma, S., Kavita, Woźniak, M., Shafi, J., and Ijaz, M. F. (2024). Shortcut learning explanations for deep natural language processing: A survey on dataset biases. *IEEE Access*, 12:26183–26195.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. (2024). A survey on in-context learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dubanowska, Z., Żelaszczyk, M., Brzozowski, M., Mandica, P., and Karpowicz, M. (2025). Representation-based broad hallucination detectors fail to generalize out of distribution. In *Proceedings of EMNLP Findings*.

Dvornik, N., Mairal, J., and Schmid, C. (2018). Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Eftekhar, A., Sax, A., Malik, J., and Zamir, A. (2021). Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., and the CAMELYON16 Consortium (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210.

European Commission (2024). Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. In *Proceedings of the spring joint computer conference*, pages 327–338.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501. Association for Computing Machinery.

Faw, M., Sen, R., Zhou, Y., and Das, A. (2025). In-Context Fine-Tuning for Time-Series Foundation Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, (5).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Fukumoto, K., Kikuchi, E., Mikami, S., Ogihara, K., Matsumoto, K., Miyajima, A., and Oya, M. (2016). Tumor budding, a novel prognostic indicator for predicting stage progression in T1 bladder cancers. *Cancer Science*, 107(10):1430–1436.

Gadermayr, M. and Tschuchnig, M. (2024). Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112.

Galleguillos, C. and Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722. Special Issue on Multi-Camera and Multi-Modal Sensor Fusion.

Garcia, C., Țurcan, A., Howman, H., and Filik, R. (2022). Emoji as a tool to aid the comprehension of written sarcasm: Evidence from younger and older adults. *Computers in Human Behavior*, 126.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Geman, D., Geman, S., Hallonquist, N., and Younes, L. (2015). Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, 112.

Ghorbani, A., Wexler, J., Zou, J., and Kim, B. (2019). Towards automatic concept-based explanations. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Giraud-Carrier, C. (2000). A note on the utility of incremental learning. *AI Communications*, 13(4):215–223.

Gkartzonika, I., Gkalelis, N., and Mezaris, V. (2023). Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2018). Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv:1706.02677*.

Graepel, T., Candela, J. Q., Borchert, T., and Herbrich, R. (2010). Web-scale Bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Grattafiori, A. et al. (2024). The Llama 3 Herd of Models. *arXiv:2407.21783*.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Lin, Z., Zhang, B., Ni, L., Gao, W., Wang, Y., and Guo, J. (2026). A survey on LLM-as-a-Judge. *The Innovation*.

Gu, X., Fan, H., Huang, Y., Luo, T., and Zhang, L. (2024). Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., and Yener, B. (2009). Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, 2:147–171.

Góral, G., Ziarko, A., Miłoś, P., Nauman, M., Wołczyk, M., and Kosiński, M. (2025). Beyond recognition: Evaluating visual perspective taking in vision language models. *arXiv:2505.03821*.

Góral, G., Ziarko, A., Nauman, M., and Wołczyk, M. (2024). Seeing through their eyes: Evaluating visual perspective taking in vision language models. *arXiv:2409.12969*.

Haldekar, M., Ganesan, A., and Oates, T. (2017). Identifying spatial relations in images using convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. (2018). Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Han, K., Wang, Y., Guo, J., Tang, Y., and Wu, E. (2022). Vision GNN: an image is worth graph of nodes. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195.

HaoChen, J. Z. and Ma, T. (2023). A Theoretical Study of Inductive Biases in Contrastive Learning. In *Proceedings of the International Conference on Representation Learning (ICLR)*.

Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74.

He, J., Mao, R., Shao, Z., and Zhu, F. (2020a). Incremental learning in online scenario. In *Proceedings ot the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020b). Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE Computer Society.

Hengle, A., Bajpai, P., Dan, S., and Chakraborty, T. (2025). Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of*

the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5165–5180. Association for Computational Linguistics.

Hense, J., Idaji, M. J., Eberle, O., Schnake, T., Dippel, J., Ciernik, L., Buchstab, O., Mock, A., Klauschen, F., and Muller, K. R. (2024). xMIL: Insightful explanations for multiple instance learning in histopathology. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., and Andreas, J. (2022). Natural language descriptions of deep visual features. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2020). Explainable AI Methods – A Brief Overview. In *Proceedings of the XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with International Conference on Machine Learning (ICML)*.

Hoshen, D. and Werman, M. (2017). IQ of neural networks. *arXiv:1710.01692*.

Hryniewska, W., Bombiński, P., Szatkowski, P., Tomaszewska, P., Przelaskowski, A., and Biecek, P. (2021). Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies. *Pattern Recognition*, 118:108035.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hu, S., Ma, Y., Liu, X., Wei, Y., and Bai, S. (2021). Stratified Rule-Aware Network for Abstract Visual Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Huang, B., Balashankar, A., and Aminifar, A. (2025). Beft: Bias-efficient fine-tuning of language models. *arXiv:2509.15974*.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *Proceedins of Findings of the Association for Computational Linguistics (ACL)*.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting Sample Selection Bias by Unlabeled Data. In *Proceeding of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Huang, Z., Liang, M., Qin, J., Zhong, S., and Lin, L. (2023). Understanding Self-attention Mechanism via Dynamical System Perspective. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hughes, G. and Holt, J. S. (1970). A Distribution-Free Test for Serial Independence of Regression Disturbances. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):297–310.

Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based Deep Multiple Instance Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Isnard, P. and Humphreys, B. D. (2025). Spatial transcriptomics. *The American Journal of Pathology*, 195(1):23–39.

Jain, S. and Eadon, M. T. (2024). Spatial transcriptomics in health and disease. *Nature Reviews Nephrology*, 20(10):659–671.

Jelassi, S., Sander, M. E., and Li, Y. (2022). Vision transformers provably learn spatial structure. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. (2022). Visual Prompt Tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jia, M., Zhao, L., Li, G., and Zheng, Y. (2025). ContextHOI: Spatial context learning for human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jiang, H., Ma, X., Nie, W., Yu, Z., Zhu, Y., and Anandkumar, A. (2022). Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin, W., Li, X., Fatehi, M., and Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kamradt, G. (2023). Needle In A Haystack - Pressure Testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`. GitHub repository.

Kandwal, S. and Nehra, V. (2024). A Survey of Text-to-Image Diffusion Models in Generative AI. In *Proceeding of the International Conference on Cloud Computing, Data Science Engineering*.

Kang, J., Wu, W., Christianos, F., Chan, A. J., Greenlee, F. D., Thomas, G., Purtorab, M., and Toulis, A. (2025). LM2: Large memory models for long context reasoning. In *Workshop on Reasoning and Planning for Large Language Models at the International Conference on Learning Representations (ICLR)*.

Karaman, B. K., Dodelzon, K., Akar, G. B., and Sabuncu, M. R. (2024). Longitudinal Mammogram Risk Prediction . In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Kelejian, H. and Prucha, I. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating A Spatial Autoregressive Model with Autoregressive Disturbances. *The Journal of Real Estate Finance and Economics*, 17:99–121.

Kelejian, H. H. and Prucha, I. (1999). A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review*, 40(2):509–33.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. (2017). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kim, S., Jung, D., and Cho, M. (2023). Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Representation Learning (ICLR)*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. B. (2023). Segment anything. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Klenk, M., Forbus, K., Tomai, E., and Kim, H. (2011). Using analogical model formulation with sketches to solve Bennett Mechanical Comprehension Test problems. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(3):299–327.

Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept Bottleneck Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kondmann, L., Toker, A., Saha, S., Schölkopf, B., Leal-Taixé, L., and Zhu, X. X. (2021). Spatial context awareness for unsupervised change detection in optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*.

Konwer, A., Xu, X., Bae, J., Chen, C., and Prasanna, P. (2022). Temporal context matters: Enhancing single image prediction with disease progression representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kopf, L., Bommer, P. L., Hedström, A., Lapuschkin, S., Höhne, M. M.-C., and Bykov, K. (2024). Cosy: Evaluating textual explanations of neurons. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Krantz, B. A., Dave, N., Komatsubara, K. M., Marr, B. P., and Carvajal, R. D. (2017). Uveal melanoma: epidemiology, etiology, and treatment of primary disease. *Clinical Ophthalmology*, 11:279–289.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., and Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

Kroll, A. J. and Kuwabara, T. (1965). Electron microscopy of uveal melanoma: A comparison of spindle and epithelioid cells. *Archives of Ophthalmology*, 73(3):378–386.

Kumar, V., Abbas, A. K., and Aster, J. C. (2017). *Robbins Basic Pathology*. Elsevier - Health Sciences Division, 10th edition.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096.

Lawsen, A. (2025). Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv:2506.09250*.

LeCun, Y., Huang, F. J., and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.-Y., Fei-Fei, L., Wu, J., Ermon, S., and Liang, P. (2023). Holistic Evaluation of Text-to-Image Models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Lee, W., Ham, B., and Kim, S. (2025). Maximizing the position embedding for vision transformers with global average pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Leinar Ramos and Jitendra Subramanyam (2024). Maverick Research: Forget About Your Real Data – Synthetic Data Is the Future of AI. Gartner. Online; 12 February 2024.

Lerousseau, M., Vakalopoulou, M., Deutsch, E., and Paragios, N. (2021). SparseConvMIL: Sparse Convolutional Context-Aware Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the MICCAI Workshop on Computational Pathology*.

Li, C., Lu, Y., Mei, Q., Wang, D., and Pandey, S. (2015). Click-through prediction for advertising in Twitter timeline. In *Conference on Knowledge Discovery and Data Mining (KDD)*.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Li, S., Zhu, S., Paul, S., Roy-Chowdhury, A., Song, C., Krishnamurthy, S., Swami, A., and Chan, K. S. (2020). Connecting the Dots: Detecting Adversarial Perturbations Using Context Inconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Li, Y., Lu, X., Xiong, H., Tang, J., Su, J., Jin, B., and Dou, D. (2023). Towards long-term time-series forecasting: Feature, pattern, and distribution. In *Proceedings of the IEE International Conference on Data Engineering (ICDE)*.

Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2935–2947.

Liang, W., Wu, Y., Li, M., and Cao, Y. (2022). A Feature Fusion-Net Using Deep Spatial Context Encoder and Nonstationary Joint Statistical Model for High-Resolution SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18.

Lin, M., Weng, N., Mikolaj, K., Bashir, Z., Svendsen, M. B. S., Tolsgaard, M. G., Christensen, A. N., and Feragen, A. (2024). Shortcut Learning in Medical Image Segmentation . In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, N. and Han, J. (2016). A Deep Spatial Contextual Long-Term Recurrent Convolutional Network for Saliency Detection. *IEEE Transactions on Image Processing*, 27:3264–3274.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024a). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Liu, T. Y. A., Chen, H., Gomez, C., Correa, Z. M., and Unberath, M. (2022a). Direct gene expression profile prediction for uveal melanoma from digital cytopathology images via deep learning and salient image region identification. *Ophthalmology Science*, 3(1).

Liu, T. Y. A., Chen, H., Koseoglu, N. D., Kolchinski, A., Unberath, M., and Correa, Z. M. (2025). Direct Prediction of 48 Month Survival Status in Patients with Uveal Melanoma Using Deep Learning and Digital Cytopathology Images. *Cancers*, 17(2).

Liu, Y., Wang, R., Shan, S., and Chen, X. (2018). Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. (2022b). Swin Transformer V2: Scaling Up Capacity and Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, Z., Zhu, T., Tan, C., Liu, B., Lu, H., and Chen, W. (2024b). Probing language models for pre-training data detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., and Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Lopardo, G., Precioso, F., and Garreau, D. (2024). Attention Meets Post-hoc Interpretability: A Mathematical Perspective. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Lovering, C. and Pavlick, E. (2022). Unit Testing for Concepts in Neural Networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208.

Lu, J., Sun, Y., and Yang, S. (2025). In-context time series predictor. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021a). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570.

Lu, M. Y., Williamson, D. F., Chen, T. Y., and Mahmood, F. (2021b). AI-based pathology predicts origins for cancers of unknown primary. *Nature*, page 1–25.

Lu, Y., Luo, L., Huang, D., Wang, Y., and Chen, L. (2020). Knowledge Transfer in Vision Recognition: A Survey. *ACM Comput. Surv.*, 53(2).

Lucieri, A., Bajwa, M. N., Dengel, A., and Ahmed, S. (2020). Explaining ai-based decision support systems using concept localization maps. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Lugli, A., Kirsch, R., Ajioka, Y., and et al. (2017). Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Modern Pathology*, 30:1299–1311.

Lugli, A., Zlobec, I., Berger, M. D., Kirsch, R., and Nagtegaal, I. D. (2020). Tumour budding in solid cancers. *Nature Reviews Clinical Oncology*.

Lundberg, S. M., Erion, G., and Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. (2022). CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Małkiński, M. and Mańdziuk, J. (2025). Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *ACM Comput. Surv.*, 57(7).

Mańdziuk, J. and Żychowski, A. (2019). DeepIQ: A human-inspired AI System for Solving IQ Test Problems. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Marini, N., Otálora, S., Ciompi, F., Silvello, G., Marchesin, S., Vatrano, S., Buttafuoco, G., Atzori, M., and Müller, H. (2021). Multi-Scale Task Multiple Instance Learning for the Classification of Digital Pathology Images with Global Annotations. In *Proceedings of the MICCAI workshop on Computational Pathology*.

Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J. A., Garcia-Garcia, A., Orts-Escolano, S., Garcia-Rodriguez, J., and Vincze, M. (2021). UnrealROX+: An Improved Tool for Acquiring Synthetic Data from Virtual 3D Environments. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and Oak, S. (2025). Artificial intelligence index report 2025. Technical report, Stanford Institute for Human-Centered Artificial Intelligence. Accessed: 2025-06-25.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., and Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior research methods*, 42(2):525–541.

Maté, A. and Dimiccoli, M. (2024). Temporal Context Consistency Above All: Enhancing Long-Term Anticipation by Learning and Enforcing Temporal Constraints. *arXiv:2412.19424*.

Małkiński, M. and Mańdziuk, J. (2023). A review of emerging research directions in Abstract Visual Reasoning. *Information Fusion*, 91:713–736.

Małkiński, M. and Mańdziuk, J. (2024). Multi-label contrastive learning for abstract visual reasoning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1941–1953.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

McLean, I. W., Foster, W. D., Zimmerman, L. E., and Gamel, J. W. (2018). Modifications of Callender's Classification of Uveal Melanoma at the Armed Forces Institute of Pathology. *American Journal of Ophthalmology*, 195.

McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., et al. (2013). Ad click prediction: a view from the trenches. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*.

Mehrabian, A. and Ferris, S. R. (1967). Inference of Attitudes from Nonverbal Communication in Two Channels. *Journal of Consulting Psychology*, 31(3):248–252.

Mei, L., Yao, J., Ge, Y., Wang, Y., Bi, B., Cai, Y., Liu, J., Li, M., Li, Z.-Z., Zhang, D., Zhou, C., Mao, J., Xia, T., Guo, J., and Liu, S. (2025). A survey of context engineering for large language models. *arXiv:2507.13334*.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Minku, L. L. and Yao, X. (2011). DDD: a new ensemble approach for dealing with concept drift. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):619–633.

Mirza, M. and Osindero, S. (2014). Conditional Generative Adversarial Nets. *arXiv:1411.1784*.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23.

Muhlbaier, M. D. and Polikar, R. (2007). Multiple classifiers based incremental learning algorithm for learning in nonstationary environments. In *Proceedings of the International Conference on Machine Learning and Cybernetics*.

Murshed, M. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., and Hussain, F. (2021). Machine learning at the network edge: A survey. *ACM Computing Surveys (CSUR)*, 54(8):1–37.

Narla, A., Kuprel, B., Sarin, K., Novoa, R., and Ko, J. (2018). Automated classification of skin lesions: From pixels to practice. *Journal of Investigative Dermatology*, 138(10):2108–2110.

Nathan, P., Hassel, J. C., Rutkowski, P., Baurain, J.-F., Butler, M. O., Schlaak, M., Sullivan, R. J., Ochsenreither, S., Dummer, R., Kirkwood, J. M., Joshua, A. M., Sacco, J. J., Shoushtari, A. N., Orloff, M., Piulats, J. M., Milhem, M., Salama, A. K. S., Curti, B., Demidov, L., Gastaud, L., Mauch, C., Yushak, M., Carvajal, R. D., Hamid, O., Abdullah, S. E., Holland, C., Goodall, H., Piperno-Neumann, S., and IMCgp100-202 Investigators (2021). Overall Survival Benefit with Tebentafusp in Metastatic Uveal Melanoma. *New England Journal of Medicine*, 385(13):1196–1206.

Nauta, M., Walsh, R., Dubowski, A., and Seifert, C. (2022). Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

nostalgebraist (2020). Interpreting gpt: the logit lens. Accessed: 2025-06-08.

Ohlsson, S., Sloan, R. H., Turán, G., and Urasky, A. (2013). Verbal IQ of a four-year old achieved by an AI system. *Age*, 4(5):6.

Oikarinen, T. and Weng, T.-W. (2023). CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Oikarinen, T. and Weng, T.-W. (2024). Linear explanations for individual neurons. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Okcu, O., Çiğdem Öztürk, Şen, B., Arpa, M., and Bedir, R. (2021). Tumor Budding is a reliable predictor for death and metastasis in invasive ductal breast cancer and correlates with other prognostic clinicopathological parameters. *Annals of Diagnostic Pathology*, 54:151792.

Olorunnimbe, K. and Viktor, H. (2023). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artificial Intelligence Review*, 56:2057–2109.

Pariza, V., Salehi, M., Burghouts, G. J., Locatello, F., and Asano, Y. M. (2025). Near, far: Patch-ordering enhances vision foundation models' scene understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Paulo, G. S., Mallen, A. T., Juang, C., and Belrose, N. (2025). Automatically interpreting millions of features in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Pawan, S. and Rajan, J. (2022). Capsule networks for image classification: A review. *Neurocomputing*, 509:102–120.

Pechyonkin, M. (2017). Understanding Hinton's Capsule Networks. Part I: Intuition. `https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-i-intuition-b4b559d1159b`.

Pocock, J., Graham, S., Vu, Q. D., Jahanifar, M., Deshpande, S., Hadjigeorghiou, G., Shephard, A., Bashir, R. M. S., Bilal, M., Lu, W., Epstein, D., Minhas, F., Rajpoot, N. M., and Raza, S. E. A. (2022). TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, vol. 2(1):120.

Pătrăucean, V., Smaira, L., Gupta, A., Continente, A. R., Markeeva, L., Banarse, D., Koppula, S., Heyward, J., Malinowski, M., Yang, Y., Doersch, C., Matejovicova, T., Sulsky, Y., Miech, A., Frechette, A., Klimczak, H., Koster, R., Zhang, J., Winkler, S., Aytar, Y., Osindero, S., Damen, D., Zisserman, A., and Carreira, J. (2023). Perception test: A diagnostic benchmark for multimodal video models. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008). *Dataset shift in machine learning*. MIT Press.

Rabold, J., Deininger, H., Siebers, M., and Schmid, U. (2020). Enriching Visual with Verbal Explanations for Relational Concepts – Combining LIME with Aleph. In *Machine Learning and Knowledge Discovery in Databases*, pages 180–192, Cham. Springer International Publishing.

Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Rauker, T., Ho, A., Casper, S., and Hadfield-Menell, D. (2023). Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks . In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive psychology*, 41(1):1–48.

Raven, J. C. (1936). Mental Tests Used in Genetic Studies: The Performances of Related Individuals in Tests Mainly Educative and Mainly Reproductive. Master's thesis, University of London.

Raven, J. C. and Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press Oxford, UK.

Ray, A., Radenovic, F., Dubey, A., Plummer, B. A., Krishna, R., and Saenko, K. (2023). Cola: A Benchmark for Compositional Text-to-image Retrieval. In *Proceeding of the Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). iCarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*.

Rietschel, P., Panageas, K. S., Hanlon, C., Patel, A., Abramson, D. H., and Chapman, P. B. (2005). Variates of Survival in Metastatic Uveal Melanoma . *Journal of Clinical Oncology*, 23(31):8076–8080.

Rigamonti, A., Feuerhake, F., Donadon, M., Locati, M., and Marchesi, F. (2021). Histopathological and immune prognostic factors in colo-rectal liver metastases. *Cancers*, 13(5):1075.

Riggi, N., Aguet, M., and Stamenkovic, I. (2018). Cancer metastasis: A reappraisal of its underlying mechanisms and their relevance to treatment. *Annual Review of Pathology: Mechanisms of Disease*, 13:117–140.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. (2024). Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Rola, A. C., Kalirai, H., Taktak, A. F. G., Eleuteri, A., Krishna, Y., Hussain, R., Heimann, H., and Coupland, S. E. (2022). A retrospective analysis of 10 years of liver surveillance undertaken in uveal melanoma patients treated at the supraregional "liverpool ocular oncology centre", uk. *Cancers*, 14(9).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Royer, A. and Lampert, C. H. (2015). Classifier adaptation at prediction time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. *arXiv:1606.04671*.

Rymarczyk, D., Pardyl, A., Kraus, J., Kaczyńska, A., Skomorowski, M., and Zieliński, B. (2023). ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification. In *Machine Learning and Knowledge Discovery in Databases*, pages 421–436, Cham. Springer International Publishing.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Sacha, M., Jura, B., Rymarczyk, D., Struski, L., Tabor, J., and Zieliński, B. (2024). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, Cham.

Schneider, K.-M. (2005). Weighted average pointwise mutual information for feature selection in text categorization. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (PKDD)*.

Schulz, D. and Bauckhage, C., editors (2025). *Informed Machine Learning*. Cognitive Technologies. Springer, Cham.

Seibold, C., Kleesiek, J., Schlemmer, H.-P., and Stiefelhagen, R. (2021). Self-guided multiple instance learning for weakly supervised disease classification and localization in chest radiographs. In Ishikawa, H., Liu, C.-L., Pajdla, T., and Shi, J., editors, *Computer Vision – ACCV 2020*, pages 617–634, Cham. Springer International Publishing.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Settles, B. (2009). Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and zhang, y. (2021). TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv:2506.06941*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Singh, R. and Buckley, C. L. (2023). Attention: Marginal Probability is All You Need? *arXiv: 2304.04556*.

Skovholt, K., Grønning, A., and Kankaanranta, A. (2014). The communicative functions of emoticons in workplace e-mails: :-)*. *Journal of Computer-Mediated Communication*, 19(4):780–797.

Song, A. H., Jaume, G., Williamson, D. F. K., Lu, M. Y., Vaidya, A., Miller, T. R., and Mahmood, F. (2023). Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1:930–949.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Starke, G., Gille, F., Termine, A., Aquino, Y. S. J., Chavarriaga, R., Ferrario, A., Hastings, J., Jongsma, K., Kellmeyer, P., Kulynych, B., Postan, E., Racine, E., Sahin, D., Tomaszewska, P., Vold, K., Webb, J., Facchini, A., and Ienca, M. (2025). Finding Consensus on Trust in AI in Health Care: Recommendations From a Panel of International Experts. *Journal of Medical Internet Research*, 27.

Strannegård, C., Amirghasemi, M., and Ulfsbäcker, S. (2013). An anthropomorphic method for number sequence problems. *Cognitive Systems Research*, 22:27–34.

Street, W. N. and Kim, Y. (2001). A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*.

Struski, Ł., Janusz, S., Tabor, J., Markiewicz, M., and Lewicki, A. (2024). Multiple instance learning for medical image classification based on instance importance. *Biomedical Signal Processing and Control*, 91.

Struski, Ł., Rymarczyk, D., Lewicki, A., Sabiniewicz, R., Tabor, J., and Zieliński, B. (2023). Promil: Probabilistic multiple instance learning for medical imaging. In *Frontiers in Artificial Intelligence and Applications, Volume 372: ECAI 2023*, pages 2210–2217. IOS Press.

Sun, A., Ma, P., Yuan, Y., and Wang, S. (2023). Explain Any Concept: Segment Anything Meets Concept-Based Explanation. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Sun, K. and Marasović, A. (2021). Effective attention sheds light on interpretability. In *Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP*.

Sun, M., Zhou, W., Qi, X., Zhang, G., Girnita, L., Seregard, S., Grossniklaus, H. E., Yao, Z., Zhou, X., and Stålhammar, G. (2019). Prediction of BAP1 Expression in Uveal Melanoma Using Densely-Connected Deep Classification Networks. *Cancers*, 11(10).

Sun, Z., Zang, X., Zheng, K., Xu, J., Zhang, X., Yu, W., Song, Y., and Li, H. (2025). ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Swingler, K. and Bath, M. (2020). Learning Spatial Relations with a Standard Convolutional Neural Network. In *Proceedings of the International Joint Conference on Computational Intelligence*.

The White House (2023). Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Tighidet, Z., Mei, J., Piwowarski, B., and Gallinari, P. (2024). Probing language models on their knowledge source. In *BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*.

Tjoa, E. and Guan, C. (2020). Quantifying Explainability of Saliency Methods in Deep Neural Networks With a Synthetic Dataset. *IEEE Transactions on Artificial Intelligence*, 4:858–870.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46(sup1):234–240.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP Architecture for Vision. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Tomaszewska, P. (2022). Information Injection to Deep Learning Solutions in Knowledge Transfer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Doctoral Consortium.

Tomaszewska, P. and Biecek, P. (2024). Position: Do Not Explain Vision Models Without Context. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Tomaszewska, P., Gozdera, M., Sienkiewicz, E., and Biecek, P. (2025a). From Input to Insight: Probing the Reasoning of Attention-based MIL Models. In *Proceedings of the World Conference on eXplainable Artificial Intelligence*.

Tomaszewska, P., Kotowski, K., Donizy, P., Stålhammar, G., and Biecek, P. (2026). Role of patch diversity while predicting time of death due to uveal melanoma metastasis. version submitted to the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

Tomaszewska, P. and Lampert, C. H. (2022). Lightweight conditional model extrapolation for streaming data under class-prior shift. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*.

Tomaszewska, P., Sienkiewicz, E., Hoang, M. P., and Biecek, P. (2025b). Deep Spatial Context: When Attention-Based Models Meet Spatial Regression (Digital Pathology). In *Proceedings of the International Conference on Artificial Intelligence in Medicine (AIME)*.

Tomaszewska, P., Sperkowski, M., and Biecek, P. (2023). Does spatial context matter in digital pathology? *International Workshop on Modelling and Representing Context at European Conference on Artificial Intelligence (ECAI)*.

Tomaszewska, P., Żychowski, A., and Mańdziuk, J. (2022). Duel-based deep learning system for solving iq tests. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Turing, A. M. (1950). I.—Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. (2023). Steering language models with activation engineering. *arXiv:2308.10248*.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Tworkowski, S., Staniszewski, K., Pacek, M., Wu, Y., Michalewski, H., and Miłoś, P. (2023). Focused Transformer: Contrastive Training for Context Scaling. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Tyagi, A. (2022). Underlying facets of cancer metastasis. *Cancers*, 14(12):2989.

van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018). Rotation Equivariant CNNs for Digital Pathology. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Conference*.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Vipul Baxi, Robin Edwards, M. M. S. S. (2022). Digital pathology and artificial intelligence in translational medicine. *Modern Pathology*, 35(1):23 – 32.

Vo, A., Nguyen, K.-N., Taesiri, M. R., Dang, V. T., Nguyen, A. T., and Kim, D. (2025). Vision Language Models are Biased: Counting legs of an animal is surprisingly hard. In *AI for Math Workshop at the International Conference on Machine Learning (ICML)*.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., and Schuecker, J. (2023). Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633.

Wan, Q., Ren, X., Wei, R., Yue, S., Wang, L., Yin, H., Tang, J., Zhang, M., Ma, K., and Deng, Y. P. (2023). Deep learning classification of uveal melanoma based on histopathological images and identification of a novel indicator for prognosis of patients. *Biological Procedures Online*, 25(1).

Wang, D., Jamnik, M., and Lio, P. (2020). Abstract Diagrammatic Reasoning with Multiplex Graph Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., and Xiao, Y. (2023). MICN: Multi-scale local and global context modeling for long-term series forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., and ZHOU, J. (2024a). Timemixer: Decomposable multiscale mixing for time series forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang, W., Zhang, S., Ren, Y., Duan, Y., Li, T., Liu, S., Hu, M., Chen, Z., Zhang, K., Lu, L., Zhu, X., Luo, P., Qiao, Y., Dai, J., Shao, W., and Wang, W. (2024b). Needle in a multimodal haystack. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Wang, X. and Zhu, Z. (2023). Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229.

Wang, Y., Xie, C., Liu, Y., and Zheng, Z. (2024c). VideoLLaMB: Long-context Video Understanding with Recurrent Memory Bridges. *arXiv:2409.01071*.

Wang, Z. and Wu, L. (2023). Theoretical analysis of inductive biases in deep convolutional networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, Z., Xie, E., Li, A., Wang, Z., Liu, X., and Li, Z. (2024d). Divide and Conquer: Language Models can Plan and Self-Correct for Compositional Text-to-Image Generation. *arXiv:2401.15688*.

Waqas, A., Bui, M. M., Glassy, E. F., El Naqa, I., Borkowski, P., Borkowski, A. A., and Rasool, G. (2023). Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Laboratory Investigation*, 103(11).

Watson, M., Shiekh Hasan, B. A., and Moubayed, N. A. (2022). Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Wei, Z., Jiang, Y., Liao, X., Qi, X., Wang, Z., Wu, G., Hao, P., and Barth, M. (2020). End-to-End Vision-Based Adaptive Cruise Control (ACC) Using Deep Reinforcement Learning. *arXiv:2001.09181*.

Wijeratne, P. A., Eshaghi, A., Scotton, W. J., Kohli, M., Aksman, L., Oxtoby, N. P., Pustina, D., Warner, J. H., Paulsen, J. S., Scahill, R. I., Sampaio, C., Tabrizi, S. J., and Alexander, D. C. (2023). The temporal event-based model: Learning event timelines in progressive diseases. *Imaging Neuroscience*, 1.

Williams, H. J. L., Asatsuma, T., Vento-Tormo, R., and Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14:68.

Worley, L. A., Onken, M. D., Person, E., Robirds, D., Branson, J., Char, D. H., Perry, A., and Harbour, J. W. (2007). Transcriptomic versus chromosomal prognostic markers and clinical outcome in uveal melanoma. *Clinical Cancer Research*, 13(5):1466–1471.

Wu, S., Chen, S., Xie, C., and Huang, X. (2023). One-Pixel Shortcut: On the Learning Preference of Deep Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. (2025). Retrieval head mechanistically explains long-context factuality. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wu, Y., Dong, H., Grosse, R., and Ba, J. (2020). The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv:2007.04212*.

Xie, C., Muhammad, H., Vanderbilt, C. M., Caso, R., Yarlagadda, D. V. K., Campanella, G., and Fuchs, T. J. (2020). Beyond Classification: Whole Slide Tissue Histopathology Analysis By End-To-End Part Learning. In *Proceedings of the International Conference on Medical Imaging with Deep Learning*.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G. M., Li, Y., and Singh, V. (2021). Nyströmformer: A nyström-based algorithm for approximating self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yang, J., Su, B., Zhao, W. X., and Wen, J.-R. (2024). Unlocking the power of spatial and temporal information in medical multimodal pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yao, G. (2022). Combating Extreme Adversarial Vulnerabilities in WSI Classification. *Workshop on Computational Biology at the International Conference on Machine Learning (ICML)*.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. (2020). On completeness-aware concept-based explanations in deep neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Yin, M., Li, S., Song, C., Asif, M. S., Roy-Chowdhury, A. K., and Krishnamurthy, S. V. (2022). Adc: Adversarial attacks against object detection that evade context consistency checks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. (2018). Lifelong learning with dynamically expandable networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yousefirizi, F., Pierre Decazes, Amyar, A., Ruan, S., Saboury, B., and Rahmim, A. (2022). AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging: Towards Radiophenomics. *PET Clinics*, 17(1):183–212. Artificial Intelligence and PET Imaging, Part II.

Yu, Q., Yang, G., Wang, X., Shi, Y., Feng, Y., and Liu, A. (2025). A review of time series forecasting and spatio-temporal series forecasting in deep learning. *Journal of Supercomputing*, 81(10):1160.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. (2020). A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *arXiv:1910.04867*.

Zhang, J., Zhang, Q., Ren, J., Zhao, Y., and Liu, J. (2022a). Spatial-Context-Aware Deep Neural Network for Multi-Class Image Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhang, W., Guo, J., He, T., Zhao, L., Xu, L., and Bian, J. (2025). Video in-context learning: Autoregressive transformers are zero-shot video imitators. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhang, W., Zhu, C., Liu, J., Wang, Y., and Jin, M. (2020). Cancer Metastasis Detection Through Multiple Spatial Context Network. In *Proceedings of the International Conference on Computing and Pattern Recognition*.

Zhang, Y., Liu, B., Cai, Q., Wang, L., and Wang, Z. (2022b). An Analysis of Attention via the Lens of Exchangeability and Latent Variable Models. *arXiv:2212.14852*.

Zhang, Y. and Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhang, Y.-X., Meng, H., Cao, X.-M., Zhou, Z., Yang, M., and Adhikary, A. R. (2023). Interpreting vulnerabilities of multi-instance learning to adversarial perturbations. *Pattern Recognition*, 142.

Zhang, Z. and Gong, X. (2023). Positional label for self-supervised vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence and Symposium on Educational Advances in Artificial Intelligence*.

Zhao, H., Sun, H., Kong, J., Li, X., Wang, Q., Jiang, L., Zhu, Q., Abdelzaher, T. F., Choi, Y., Li, M., and Shao, H. (2026). Activation steering for LLM alignment via a unified ODE-based framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2025). A Survey of Large Language Models. *arXiv:2303.18223*.

Zheng, K., Zha, Z.-j., and Wei, W. (2019). Abstract reasoning with distracting features. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021a). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhou, W., Yang, Y., Yu, C., et al. (2021b). Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. *Nature Communications*, 12.

Zhu, H., Fiondella, L., Yuan, J., Zeng, K., and Jiao, L. (2025). Neurogenpoisoning: Neuron-guided attacks on retrieval-augmented generation of LLM via genetic optimization of external knowledge. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*.

Zormpas-Petridis, K., Failmezger, H., Raza, S. E. A., Roxanis, I., Jamin, Y., and Yuan, Y. (2019). Superpixel-Based Conditional Random Fields (SuperCRF): Incorporating Global and Local Context for Enhanced Deep Learning in Melanoma Histopathology. *Frontiers in Oncology*, 9.

Żukowska, N., Goswami, M., Wiliński, M., Potosnak, W., and Dubrawski, A. (2024). Towards long-context time series foundation models. *NeurIPS Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability and NeurIPS Workshop on Time Series in the Age of Large Models*.

List of Figures

1.1 Mapping of papers and the corresponding research questions to axes regarding context (type, scale, use case, stage). 18
1.2 The key outcomes developed in this dissertation. 18

2.1 AI performance vs. human performance in benchmarks. 24
2.2 The CLIP-Dissect method. 29
2.3 The Describe-and-Dissect method. 30

3.1 Class distribution in `geo-tweets` dataset across 24 hours (March 1, 2020) for five most frequent countries. 49
3.2 Illustration of classifier accuracy over time for the proposed *LIMES* and baseline methods in one of the tested settings (*tweet* features, subset *early*, realization *0*). 51

4.1 The need for paradigm shift – instead of *from 'where' to 'what'*, we should focus on *from 'where' to 'how'*. 55
4.2 Taxonomy of contextual information within images. 57
4.3 Examples of images where the ground truth labels depend on spatial relationships between objects: *distance*, *inside/outside*, *order*, *orientation*. . . . 57
4.4 Visual IQ test as an example of the task where spatial relationship understanding is required to solve it properly by choosing one answer from the set of given to fill in the input RPM. 61
4.5 Samples from the Bongard-HOI dataset. 62
4.6 Samples from Isle-Bricks and Isle-Dots datasets for evaluation of the capability of visual perspective-taking. 63
4.7 Illustrative frames from the Perception Test dataset. 64
4.8 The same elements in elements so a CNN most probably will classify them as a face in both cases unlike a CapsuleNet that will not classify a *deformed face* as a *face* as spatial relationships between objects are not properly preserved. 65
4.9 The examples of unclear explanations of correct predictions of *Resnet–50* and *Moco v3* on the samples from the *KITTI* and *dsprites* datasets. 68

5.1 Example of an RPM task with the set of candidate answers. 78
5.2 Proposed duel-based system (*winner_in* variant) in which the results of duels are aggregated in the scoring module 80
5.3 Figure shapes used in experiments. 83
5.4 Autoencoder architecture. 84
5.5 Classification network with auxiliary training applied 85
5.6 Comparison of original image and the reconstructed one. 86

6.1 Visualization of the attention scores assigned by MIL model 97
6.2 The CLAM model architecture (multi-branch) 97
6.3 The TransMIL model architecture 99

7.1 The scheme of the proposed occlusion-based study. 103
7.2 The histopathological image padded with black border with dimensions specified 104
7.3 The performance gap when the context size is limited. 107
7.4 The misleading nature of context depending on its size 109
7.5 Number of images undergoing the change of prediction for the first time given the particular context size with the distinction on the initial and consecutive model prediction 110
7.6 Number of images undergoing the change of prediction for the first time given the particular context size 111
7.7 The logarithm of the number of images that experience the particular number of prediction changes given a particular model 111

8.1 The proposed ReasonSMIL protocol evaluates the reasoning ability of attention-based MIL models trained e.g. on the SyntheticSMIL dataset. 119
8.2 *Concavity* images generation process. 123
8.3 Samples from two different classes of the *concavity*, *onion*, and *points density* datasets, along with exemplary ground truth patches. 123
8.4 The important patches in the same test image, identified by the CLAM model but trained on different cross-validation folds. 134
8.5 TransMIL vs. 'back' GT variant. 135
8.6 The comparison of important patches selected by TransMIL against the 'fore' GT variant. 135
8.7 TransMIL vs. 'edge' GT variant. 135
8.8 The choices of GT variants in the case of a model trained on a given cross-validation fold. 136

9.1 Knowledge transfer 143

9.2 Comparison between the original image and its neighbourhood correlation visualization. 145

10.1 The proposed *Deep spatial context* method. 150
10.2 Critical distance plot of models' accuracy of classification of normal tissue vs. tumor lesion. 156
10.3 Stacked kernel distribution estimation plots where the proportion of values from normal tissues and tumor lesions for a given Spatial Context Measure. 158
10.4 Mean of $SCM_{features}$ computed on all tumor images with all model–k combinations where $k = 99$. 159
10.5 The mean value of Spatial Context Measures for models with different feature extractors with respect to the size of neighborhood (k) in the spatial weights matrix within spatial regression models over all images with tumor. 160
10.6 The relationship between the logarithms of $n_{tumor\_patches}$ and $dist_{mean}$. . . 163
10.7 Mean Spatial Context Measures computed for TransMIL with different versions of positional encoding in PPEG module. 165
10.8 Mean Spatial Context Measures computed for original TransMIL (NO_NOISE) and TARGET_NOISE. 165
10.9 Histograms of the Spatial Context Measures for different SyntheticSMIL datasets. 166
10.10 Mean $SCM_{features}$ and $SCM_{targets}$ with standard deviation. 167
10.11 The local inspection of tissue using spatial statistics. 167

11.1 Multi-step experimental workflow. (A) A physician selects three regions of interest (ROIs) from each Whole-Slide Image (WSI). (B) A MIL model trained on WSIs (shown in the first row of panel C) is used to select the top-k patches. (C) Using the same MIL model architecture as in WSI-level training, we train the model from scratch on subsets of input patches selected using different strategies. 172
11.2 Mean performance on test set in intra-evaluations. 175
11.3 Mean accuracy on test set in inter-evaluations. 175
11.4 Mean Euclidean distance between patches constituting a sample. 177

List of Tables

1.1 Axes regarding context investigated in the thesis. 15

3.1 Mapping of the paper by Tomaszewska and Lampert (2022) to the axes regarding context: type, scale, use case, ML stage. 40
3.2 Experimental Results: classifier accuracy of the proposed *LIMES* method and baselines for six experimental settings. 48

4.1 Mapping of the paper by Tomaszewska and Biecek (2024) to the axes regarding context: type, scale, use case, ML stage. 53

5.1 Mapping of the paper by Tomaszewska et al. (2022) to the axes regarding context: type, scale, use case, ML stage. 76
5.2 The number of single task questions in experiments. 83
5.3 Performance metrics of the autoencoder on test data in different experiments. 85
5.4 Classifier performance on test set in various scenarios. 87
5.5 Mean test accuracy (%) and its standard deviation of Duel-IQ in various scenarios. 88
5.6 Performance comparison of different methods for solving IQ tests: mean, standard deviation, and maximum accuracy. 90

7.1 Mapping of the paper by Tomaszewska et al. (2023) to the axes regarding context: type, scale, use case, ML stage. 102
7.2 Hyperparameters used for finetuning of transformer-based models on PCam dataset. 105
7.3 Deep Learning models' performance when full available contextual information is given (so-called reference performance). 106
7.4 Percentage of images that get the same prediction regardless of the context size. 108

8.1 Mapping of the paper by Tomaszewska et al. (2025a) to the axes regarding context: type, scale, use case, ML stage. 118
8.2 The parameters of the SyntheticSMIL datasets used in the experiments. . . 128

8.3 The number of ground truth patches (min, max, mean with standard deviation) across datasets of different difficulty levels. 129
8.4 The performance of the CLAM model on SyntheticSMIL datasets. 130
8.5 ReasonSMIL-R on CLAM. 131
8.6 ReasonSMIL-A on CLAM. 132
8.7 The performance of the TransMIL model on SyntheticSMIL datasets. . . . 132
8.8 ReasonSMIL-R on TransMIL where the model achieved similar performance metrics as CLAM. 133
8.9 ReasonSMIL-A on TransMIL. 133
8.10 Performance comparison of TransMIL with and without positional encoding (PE). 137
8.11 ReasonSMIL-R comparison for *concavity* (medium) with and without positional encoding (PE). 138
8.12 ReasonSMIL-A comparison of *concavity* (medium) with and without position encoding (PE). 138

9.1 Mapping of the paper by Tomaszewska (2022) to the axes regarding context: type, scale, use case, ML stage. 142

10.1 Mapping of the paper by Tomaszewska et al. (2025b) to the axes regarding context: type, scale, use case, ML stage. 149
10.2 Ranking of CLAM variants based on the mean Spatial Context Measures. . 162
10.3 The Spatial Context Measures (multiplied by 10 000) aggregated for all k for CLAM/TransMIL with different backbones. 164

11.1 Mapping of the paper by Tomaszewska et al. (2026) to the axes regarding context: type, scale, use case, ML stage. 170
11.2 Mean ± standard deviation of normalized label distributions across categories assigned by a physician to top-k patches ($k = 3$), computed from the training, validation, and test sets. Column names follow the $label_{prediction}$ convention; N_{EM} was not observed in fold 0. Categories include B/N/P (bleeding & necrosis & pigmentation), tumor cell types (epithelioid, spindle A/B), and fibrous strands. The last row reports the total number of patches. 178