

Dissertation

Towards Automated Clinical Diagnostic Decision Support using Machine Learning

Gerome Vivar





Technische Universität München
TUM School of Computation, Information and Technology

Towards Automated Clinical Diagnostic Decision Support using Machine Learning

Gerome Vivar

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Rückert

Prüfer der Dissertation: Prof. Dr. Nassir Navab
Prof. Dr. Christian Wachinger
Prof. Dr. Andreas Zwergal

Die Dissertation wurde am 23.03.2022. bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 04.11.2022 angenommen.

Gerome Vivar

Towards Automated

Clinical Diagnostic Decision Support using Machine Learning

Dissertation, Version 0.0

Technische Universität München

Fakultät für Informatik

Lehrstuhl für Informatikanwendungen in der Medizin

Boltzmannstraße 3

85748 and Garching bei München

Abstract

In the clinical setting, one of the major tasks of a clinician is to perform diagnosis. This process involves an iterative collection, organization, and interpretation of information which results in a working diagnosis. Only when enough information has been gathered the clinician will decide what will be the diagnosis. Here, support could be provided to the clinician using modern tools. This aspect is within the area of computer-aided diagnosis (CADx) where data is used together with an algorithm to provide support to the clinician. This decision is data-driven and is only created to support the clinician with their decision. Recently, with the resurgence and success of artificial neural networks in different domains such as computer vision, deep learning (DL) methods have been widely used and have proliferated applications in the clinics. Applications in the clinics particularly related to diagnostic decision support have received a lot of attention recently and research works are ongoing. Despite recent successes, there are still open challenges which remain.

First is the challenge of incomplete information in clinical data. Most recent CADx methods in DL do not give a lot of emphasis on the issue of data missingness. Often the assumption is that all data is available during model training and evaluation. Such an assumption could be restrictive in tackling real-world clinical data. Although there are simple techniques to address this limitation such as using instance deletion or mean imputation, such solutions could lead to a more biased analysis. In addition, often the solution is a two step process requiring more specification, such as which data missingness to assume at the first step and then which deep learning-based model to use. In this thesis, we propose a streamlined approach that simultaneously performs data imputation and classification for diagnostic decision support.

Next, the challenge that requires attention in CADx is the notion of cost-aware and peri-diagnostic support. Most recent deep learning-based methods for CADx are tailored to provide support to the clinician at the end of the diagnostic process. This means all acquisitions for all patients have to be performed to provide diagnostic support. Such an approach could also be inefficient as all examinations for all patient are required even if they might not need it or even if it could be costly. Further, the information regarding the cost of the examination is not considered. However, in a real-world setting, clinicians could also need support during the acquisition phase. Further, deciding which examination to select next should also consider the associated cost for that examination. To this end, we propose a gradient-based feature acquisition method for peri-diagnostic decision support that is cost-aware, called Accumulated Integrated Gradients (AIG).

Lastly, translation of these approaches to the clinic is essential. To this end, we evaluated different ML and DL methods for applications in neurological disease diagnosis, to showcase the feasibility of these approaches in the clinic and highlight their strengths and limitations.

Zusammenfassung

Im klinischen Umfeld besteht eine der Hauptaufgaben des Arztes darin, eine Diagnose zu stellen. Dieser Prozess beinhaltet eine iterative Sammlung, Organisation und Interpretation von Informationen, die zu einer Arbeitsdiagnose führen. Erst wenn genügend Informationen gesammelt wurden, entscheidet der Arzt, wie die Diagnose lautet. Hier könnte der Kliniker durch moderne Hilfsmittel unterstützt werden. Dieser Aspekt fällt in den Bereich der computergestützten Diagnose, bei der Daten zusammen mit einem Algorithmus verwendet werden, um den Kliniker zu unterstützen. Diese Entscheidung ist datengesteuert und dient dazu, den Kliniker bei seiner Entscheidung zu unterstützen, nicht aber, ihn zu ersetzen. In jüngster Zeit sind künstliche neuronale Netze in verschiedenen Bereichen wie Computer Vision wieder auf dem Vormarsch und sehr erfolgreich. Diese Methoden sind weit verbreitet und haben zu einer Vielzahl von Anwendungen in Kliniken geführt. Anwendungen in der Klinik, insbesondere im Zusammenhang mit der Unterstützung von Diagnoseentscheidungen, haben in letzter Zeit viel Aufmerksamkeit erregt und zahlreiche Forschungsarbeiten in Gang gesetzt. Trotz der jüngsten Erfolge gibt es immer noch offene Herausforderungen.

Die erste Herausforderung sind unvollständige Informationen in klinischen Daten. In aktuellen DL-basierten CADx-Methoden wird diesem Problem nicht viel Aufmerksamkeit geschenkt. Oft wird davon ausgegangen, dass alle Daten während des Modelltrainings und der Auswertung verfügbar sind. Eine solche Annahme könnte bei der Bearbeitung von realen klinischen Daten einschränkend sein. Einfache Kompensationsansätze, wie z. B. die Löschung von Instanzen oder die Imputation von Mittelwerten, könnten zu einer verzerrten Analyse führen. Darüber hinaus ist die Lösung oft ein zweistufiger Prozess, der eine genauere Spezifizierung erfordert, z. B. welche fehlenden Daten im ersten Schritt verwendet werden sollen und welches Deep-Learning-basierte Modell dann zum Einsatz kommt. In dieser Arbeit schlagen wir einen vereinfachten Ansatz vor, der gleichzeitig eine Datenimputation und eine Klassifizierung zur Unterstützung diagnostischer Entscheidungen durchführt.

Die nächste Herausforderung, die bei CADx beachtet werden muss, ist der Begriff der kostenbewussten Peri-Diagnostik. Aktuelle DL-Methoden für CADx sind darauf zugeschnitten, den Kliniker am Ende des Diagnoseprozesses zu unterstützen. Das setzt eine vollständige Datenerfassung für alle Patienten voraus, um die Diagnose zu unterstützen. Ein solcher Ansatz könnte ebenfalls ineffizient sein, da nicht alle Untersuchungen für alle Patienten erforderlich sind oder kostspielig sein könnten. In der Praxis könnten die Kliniker jedoch auch während der Akquisitionsphase Unterstützung benötigen. Außerdem sollten bei der Entscheidung, welche Untersuchung als nächstes ausgewählt wird, auch die damit verbundenen Kosten berücksichtigt werden. Zu diesem Zweck stellen wir Accumulated Integrated Gradients (AIG) vor, eine kostenbewusste, gradientenbasierte Methode zur Erfassung von Merkmalen für die peri-diagnostische Entscheidungshilfe.

Schließlich ist die Übertragung dieser Ansätze auf die Klinik unerlässlich. Zu diesem Zweck haben wir verschiedene ML- und DL-Methoden für diagnostische Anwendungen bei neurologischen Erkrankungen evaluiert, um die Durchführbarkeit dieser Ansätze in der Klinik zu zeigen und ihre Stärken und Grenzen hervorzuheben.

Acknowledgments

At this point, I would like to express my gratitude to all the people who made this work possible: Dr. rer. nat. Seyed-Ahmad Ahmadi for being an exceptional mentor to me throughout this PhD journey and Prof. Dr. Nassir Navab for his guidance and wisdom. I would also like to extend my sincere thanks to PD Dr. med. Andreas Zwergal for his invaluable insight into the medical domain.

I am also extremely grateful to all my colleagues and friends at the Chair for Computer-aided Medical Procedures at TUM whom I have shared this exciting PhD journey with. Special thanks go to my co-authors Anees Kazi, Hendrik Burwinkel, Shadi Albarqouni, and Kamilia Mullakaeva. Many thanks to all my colleagues at DSGZ, Brainlab, and Helmholtz-AI Munich who have made this whole experience smoother. Special thanks also go to all my friends outside work for keeping me sane.

Last but not the least, I am grateful to my family for their tireless love and support. To my wife Jistine Sanchez, thank you for the love, support, and patience. This journey would have not been worthwhile without you.

Contents

I	Introduction and Fundamentals	3
1	Introduction	5
1.1	Background	5
1.2	The necessity of automated Machine Learning-based CADx	6
1.2.1	Improving diagnosis and reducing diagnostic errors	6
1.2.2	Access to quality healthcare for resource-limited areas	7
1.2.3	Holistic decision support based on multimodal clinical data	7
1.3	Main Objectives	8
1.3.1	Addressing data missingness during model training and testing in Diagnostic Decision Support Systems.	8
1.3.2	Transitioning from CADx to Computer-aided Peri-diagnosis	9
1.3.3	Translating Computer-aided Diagnosis (CADx) into clinical research settings using benchmarks of shallow and deep learning models.	9
1.4	Outline of the thesis	10
2	Fundamentals	11
2.1	Data Representations	11
2.1.1	Incomplete data in CADx	13
2.2	Machine Learning Models	14
2.2.1	Deep Neural Networks	14
2.2.2	Graph Neural Networks	15
2.3	Model Training	20
2.3.1	Supervised Learning	20
2.3.2	Transductive and Inductive Learning	22
2.4	Feature Attribution Methods	22
2.5	Geometric Matrix Completion	23
II	Contributions	25
3	Addressing data missingness during model training and testing in Diagnostic Decision Support Systems	27
3.1	Problem Definition and Motivation	27
3.2	Related works	28
3.3	Contributions	30
3.3.1	Addressing missingness in GNN-based CADx using Geometric Matrix Completion (GRAIL-MICCAI 2018)	30

3.3.2	Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification. (AI in Medicine 2021)	42
4	Peri-diagnostic decision support in CADx using Deep Learning (MICCAI 2020)	53
4.1	Motivation	53
4.1.1	Towards peri-diagnostic decision support	54
4.1.2	Efficient peri-diagnostic decision support	54
4.1.3	Patient-specific peri-diagnostic decision support	55
4.2	Problem Definition	55
4.3	Related work	56
4.4	Contribution	56
5	Translating Computer-aided Diagnosis (CADx) into clinical research settings using benchmarks of shallow and deep learning models.	69
5.1	Motivation and Problem Definition	69
5.2	Contributions	70
5.2.1	Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway (Journal of Neurology 2019)	70
5.2.2	Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders (Journal of Neurology 2020)	72
5.2.3	Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data (Frontiers in Neurology 2021)	84
III	Conclusion and Outlook	103
6	Conclusion & Outlook	105
6.1	Summary of Findings	105
6.1.1	Addressing incomplete clinical data for CADx	105
6.1.2	Cost-efficient peri-diagnostic support	106
6.1.3	Translation of CADx to clinical research	106
6.2	Outlook	107
6.2.1	Data missingness in CADx	107
6.2.2	Peri-diagnostic decision support	107
6.2.3	Translation of ML or DL methods for CADx in clinical research	108
	Bibliography	109
	List of Figures	115
	List of Tables	117

List of Authored and Co-authored Publications

Discussed in this thesis

2021

Vivar, Gerome, Anees Kazi, Hendrik Burwinkel, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi. "Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification." *Artificial Intelligence in Medicine* 117 (2021): 102097.

Vivar, Gerome, Ralf Strobl, Eva Grill, Nassir Navab, Andreas Zwergal, and Seyed-Ahmad Ahmadi. "Using base-ml to learn classification of common vestibular disorders on DizzyReg registry data." *Frontiers in Neurology* 12 (2021).

2020

Vivar, Gerome*, Kamilia Mullakaeva*, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi. "Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 572-581. Springer, Cham, 2020.

Ahmadi, Seyed-Ahmad*, **Gerome Vivar***, Nassir Navab, Ken Möhwald, Andreas Maier, Hristo Hadzhikolev, Thomas Brandt et al. "Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders." *Journal of Neurology* 267, no. 1 (2020): 143-152.

2019

Ahmadi, Seyed-Ahmad*, **Gerome Vivar***, Johann Frei, Sergej Nowoshilow, Stanislav Bardins, Thomas Brandt, and Siegbert Krafczyk. "Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway." *Journal of neurology* 266, no. 1 (2019): 108-117.

2018

Vivar, Gerome, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi. "Multi-modal disease classification in incomplete datasets using geometric matrix completion." In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, pp. 24-31. Springer, Cham, 2018.

Not discussed in this thesis

2019

Kazi, Anees, Shayan Shekarforoush, S. Arvind Krishna, Hendrik Burwinkel, **Gerome Vivar**, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab. "Inceptiongen: receptive field aware graph convolutional network for disease prediction." In *International Conference on Information Processing in Medical Imaging*, pp. 73-85. Springer, Cham, 2019.

Kazi, Anees, Shayan Shekarforoush, S. Arvind Krishna, Hendrik Burwinkel, **Gerome Vivar**, Benedict Wiestler, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab. "Graph convolution based attention model for personalized disease prediction." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 122-130. Springer, Cham, 2019.

Burwinkel, Hendrik, Anees Kazi, **Gerome Vivar**, Shadi Albarqouni, Guillaume Zahnd, Nassir Navab, and Seyed-Ahmad Ahmadi. "Adaptive image-feature learning for disease classification using inductive graph networks." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 640-648. Springer, Cham, 2019.

Part I

Introduction and Fundamentals

Introduction

Contents

1.1	Background	5
1.2	The necessity of automated Machine Learning-based CADx	6
1.2.1	Improving diagnosis and reducing diagnostic errors	6
1.2.2	Access to quality healthcare for resource-limited areas	7
1.2.3	Holistic decision support based on multimodal clinical data	7
1.3	Main Objectives	8
1.3.1	Addressing data missingness during model training and testing in Diagnostic Decision Support Systems.	8
1.3.2	Transitioning from CADx to Computer-aided Peri-diagnosis	9
1.3.3	Translating Computer-aided Diagnosis (CADx) into clinical research settings using benchmarks of shallow and deep learning models.	9
1.4	Outline of the thesis	10

1.1 Background

Data has become more and more important in science, and approaches have shifted to being data-driven. Recent estimate suggests that there is about 40,000 exabytes of digital data in 2020 [18]. Harnessing such data to gain insights could potentially create value across different domains. One domain where gaining new insights could create very valuable impact is in healthcare [18]. Generating new insight in this domain is valuable as people's general well-being, and potentially their lives, are at stake.

Recently, Machine Learning (ML) using Deep Neural Networks (DNN) has been widely used as a tool to gain insights from data. In the medical domain, for example, we have seen widespread use of ML particularly for diagnostic applications. Such applications include heart disease prediction [60], skin cancer diagnosis [21, 27], and thoracic disease diagnosis [53], just to name a few.

We also see a lot of healthcare related companies being built, promising to provide the next healthcare solution with the term AI or machine learning on their hat [18]. This trend also includes large established companies heavily focusing on AI and machine learning. Those who do not "embrace" AI could be left out and other competitors could take the lead. Small to large companies are using machine learning in one way or the other, either to replace existing tool-chains to faster and potentially more accurate methods, or to innovate entirely new approaches.

In literature, the use of data and machine learning for decisions related to clinical settings are within the scope of Clinical Decision Support Systems (CDSS) [78]. To be more specific, CDSS are categorized into knowledge-based and non-knowledge-based systems [78]. Under knowledge-based systems are those which make use of rule-based systems in order to provide decision support, while those that make use of "Artificial Intelligence" (AI) such as machine learning are considered non-knowledge-based systems.

The umbrella term CDSS comprises of different levels of decision support in the clinic and when the decision support is targeted towards diagnosis this falls within Diagnostic Decision Support Systems (DDSS) or often referred to as Computer-aided Diagnosis (CADx) systems [78]. The term CADx first appeared in scientific literature in the late 1950s where it was often referred to as "expert systems in medicine" [94]. The more broader term CDSS can be traced back to 1970s, and it encompasses not only diagnosis but all sorts of computer-based decision support for the clinician such as disease management, prescription control, alert systems, and many more [78]. Currently, CADx is more commonly used when referring to computer-based decision support for medical diagnosis. For the purposes of this thesis, we refer to automated CADx as a system which takes patient information as input and provides suggestions as output, which the doctor can then use to make clinically informed expert-decisions.

In terms of the current state of CADx research, recent developments have been driven by several factors such as the availability of a large amount of complex data, breakthroughs in the field of Computer Science (especially AI), the existence of large amounts of diagnostic knowledge, and the complexity of the medical diagnosis process itself [94]. However, despite recent developments, there still remain key challenges in advancing CADx systems in medicine. A major part of these challenges mainly concern the algorithmic aspect of a CADx system [95]. To be specific, these challenges include the development of better classification and data mining approaches, development of advanced feature extraction/selection approaches, and dealing with big data, just to name a few. In this regard, the CADx research environment is currently at a phase where ML/DL could potentially improve diagnostic decision support in the medical domain if such massive data is managed and analyzed properly [18].

1.2 The necessity of automated Machine Learning-based CADx

In this section, we will lay out the motivation for having an automatic diagnostic decision support system. We focus on three important factors why we would need such a system.

1.2.1 Improving diagnosis and reducing diagnostic errors

From the healthcare providers point of view, such as a primary care provider, clinical diagnosis is one of the most critical tasks [75], since the diagnosis lays the foundation for all the following steps in the clinical workflow, such as pharmaceutical or surgical treatment, therapy and rehabilitation. An accurate and timely diagnosis can have a profound influence on a patient's journey towards a positive health outcome. To this end, healthcare providers follow

certain healthcare guidelines as to how to diagnose or treat certain conditions [46, 47]. If confounding factors are present, such as co-morbidity, the clinical diagnosis becomes even more complex [5]. In short, the clinical diagnostic task is not so straightforward. Despite their education, theoretical knowledge, and year-long practical experience, there is still the possibility for clinicians to diagnose a patient incorrectly, resulting to a diagnostic error which could be harmful for the patient, and/or costly for the healthcare providers [75].

In this regard, the WHO recognizes diagnostic errors as an important problem that we should prioritize [75]. A number of potential interventions to reduce diagnostic errors have been reported [75] and ML-based CADx has the potential to be used as a technique for these interventions. With such a CADx system, diagnostic errors could be reduced and potentially result in a better diagnosis.

1.2.2 Access to quality healthcare for resource-limited areas

Another reason to use such a CADx system is to better ensure equality and inclusion in healthcare. Medicine should treat all humans equally, and ideally should be available to all humans on a similarly high quality level around the world. Unfortunately, not everyone has the same quality of healthcare as healthcare systems around the world vary, resulting in different quality of delivered care [62, 90].

Not only could such a CADx system potentially improve diagnostic decisions, but also potentially enable access to quality healthcare. This applies particularly to countries where the healthcare system is not as advanced as developed countries like Germany, the U.K, and the USA, for example, healthcare systems in resource-limited countries in Asia, Africa, or South America [56]. With such a CADx system, diagnostic support could be accessible to healthcare providers, who lack the time and resources to diagnose every patient in need of healthcare.

In particular, in areas where there is limited to no access to healthcare providers such as in developing countries or remote areas where one cannot easily find a clinical expert, CADx could enable doctors to benefit from diagnostic knowledge of the medical community. Knowledge from clinical experts could be accessible through CADx even for rare diseases, especially if there are already existing signs and symptoms of a particular disease [68, 85]. With such a CADx system, access to quality healthcare for everyone could be possible especially for resource-limited areas.

1.2.3 Holistic decision support based on multimodal clinical data

Lastly, with the increasing amount of data being collected and stored, it is important that holistic decision-making is still achieved. A recent estimate suggests that there were 40,000 exabytes of digital data in 2020 [18] and this amount continuously grow. In healthcare alone, every interaction of a patient to the healthcare providers their data are stored in one way or the other resulting to enormous and complex patient information. Estimates by the World Economic Forum for example indicate that 50 petabytes of data per year are produced by

hospitals alone. With this wealth of information there could be data overload [18], which could lead to cognitive overload and affect the diagnostic decision [75]. On the other hand, this wealth of information has a big potential to improve clinical decision support if utilized with advance methods in ML/DL [95].

In the clinic, this wealth of information is usually represented in a multimodal manner containing imaging, non-imaging, and sequential-type information [94]. This multimodal clinical data could be used to build ML-based CADx models. Such an approach could augment the knowledge of clinical experts and pave the way for holistic decision support.

1.3 Main Objectives

To truly have an automated diagnostic decision support system, we would require an end-to-end solution from the backend system up to the frontend interface of the end-user. For this thesis, we mainly focus on the backend, i.e. on key challenges relevant to building the "brain" of the system, which is the machine learning model, rather than focusing on matters of user interface (UI) or user experience (UX) with the backend algorithms.

Specifically, we focused on three important challenges related to automated machine learning-based CADx. First, an important challenge is how to handle incomplete data in CADx using modern deep learning methods [20]. Second, we address how to provide diagnostic decision support to the clinician during the diagnostic process including the acquisition phase in an efficient manner. Finally, we looked at translation of CADx into the clinical research setting using shallow and deep learning models in order to get the feedback from the clinical experts whether the CADx system outputs made sense from a clinical point-of-view.

1.3.1 Addressing data missingness during model training and testing in Diagnostic Decision Support Systems.

One challenge when dealing with healthcare data is data missingness [94]. A lot of publicly available datasets, such as those implemented in common DL frameworks [1, 59], are often curated such that data scientists have all available features for every patient instance. However, in the medical domain, this is not always realistic. Often, real-world clinical datasets are incomplete, since some information for a particular instance is not available. During model training and testing we can not directly process instances with empty entries as input to the model. This scenario adds complexity to the problem and makes the data analysis more challenging.

Recent ML-based CADx methods often do not focus on this challenge of incomplete data [94]. Often, the focus of these approaches is on improving the disease classification performance with the assumption that the input data is complete or has been fully observed. The most common approach to handle missingness in data is via a pre-processing step that imputes missing values [20]. Such an approach requires the ML practitioner to select the appropriate imputation method. Once the missing values are numerically represent, this data can then be

used for succeeding ML analysis steps. This approach involves a two-step process of imputation and then ML model fitting. Such an approach could be time consuming as one has to try different imputation methods that fit the classification task.

Further, in terms of classification methods for CADx, recent methods utilizing Graph Neural Networks (GNN) have been shown to outperform standard ML models. Here, the focus is on the classification task and again, the assumption is that the input data is feature-complete. To this end, the first objective is to address data missingness in ML-based CADx.

1.3.2 Transitioning from CADx to Computer-aided Peri-diagnosis

In the context of CADx, we use the prefix “peri-” to indicate the analogy to the term “peri-operative” in Surgical Data Science [44], which takes into account all phases of the operative process. In this context, peri-diagnosis refers to the overall workflow of the diagnostic process. In this work, the focus is on providing diagnostic decision support to the clinician even during the acquisition phase. Current methods in CADx are more tailored towards providing decision support *at the end* of the diagnostic workflow. However, it is also possible that clinicians would also need diagnostic support *during* the diagnostic workflow. Most recent CADx methods do not focus on the *timing* of the decision support.

In this work, we address this challenge in CADx, i.e. to provide decision support to the clinician during the diagnostic process. At this point we have to differentiate, what we mean when using the terms “during” and “at the end” in the diagnostic workflow. We use the phrase “during the diagnostic workflow” when providing support with decisions which data to acquire next. While the phrase “at the end of the diagnostic workflow” is used when we refer to the setting where observations have been done and all patient data is available.

Furthermore, the focus of recent works is non-cost aware CADx approaches. Here the assumption is that patient examinations/observations are available for free or “cost-free”. In this context, the cost could mean money or anything valuable that needs to be optimized such as time, hospital resources, or patient comfort. One drawback of such an assumption is that it could make the method less effective in real-world applications. Therefore, the second objective is to address peri-diagnostic decision support in a cost-aware and efficient manner.

1.3.3 Translating Computer-aided Diagnosis (CADx) into clinical research settings using benchmarks of shallow and deep learning models.

Another aspect that is also important in the domain of CADx is translational research into the clinic. We want to ensure that the CADx system we are building is accurate and usable in a real-world clinical setting. In particular, the CADx system we are building should be applicable to real-world clinical datasets. Therefore, as part of the efforts of this thesis, we strove to perform translational research into the clinic. To this end, we applied various ML

and DL algorithms, both established ones and new ones proposed in this thesis. In particular, we compared ML and DL models to datasets from our clinical partner institute, the German Center for Vertigo and Balance Disorders (DSGZ) in Munich.

1.4 Outline of the thesis

The following thesis is organized into three parts.

In **Part I**, we provide background information as well as fundamental methodology relevant to understanding recent machine learning-based CADx approaches.

Chapter 1 is this chapter, where we elaborated on key background information about CADx. We discussed in this chapter the following: the definition of CADx in the context of this thesis, specifically the scope of CADx which we consider, the necessity or motivation for automated ML/DL methods in CADx, and the main objectives which we plan to address in this work.

In **Chapter 2**, we discuss key background information needed to provide additional context. The following topics are discussed in this chapter: data representation in clinical settings, machine learning models particularly Neural Network-based, training of Machine Learning models, feature attribution methods, and the topic of Geometric Matrix Completion.

In **Part II**, we present and elaborate on our contributions. Six articles are presented in this thesis: four journal articles, one conference article, and one award-winning workshop article.

In **Chapter 3**, we elaborate on the topic of data missingness and present our contributions in addressing data missingness during model training and testing in Diagnostic Decision Support Systems.

We discuss in **Chapter 4** peri-diagnostic decision support using Deep Learning. In this chapter, we also present our contribution for peri-diagnostic decision support.

In **Chapter 5**, we elaborate on our key contribution towards translational research in clinical settings focusing on applications related to Vertigo and Balance disorders.

Finally in the last part, **Part III**, we summarize our findings and provide future research directions.

This last part contains **Chapter 6** where we discuss key findings and provide a future outlook regarding the three focus areas addressed in this thesis.

Fundamentals

Contents

2.1	Data Representations	11
2.1.1	Incomplete data in CADx	13
2.2	Machine Learning Models	14
2.2.1	Deep Neural Networks	14
2.2.2	Graph Neural Networks	15
2.3	Model Training	20
2.3.1	Supervised Learning	20
2.3.2	Transductive and Inductive Learning	22
2.4	Feature Attribution Methods	22
2.5	Geometric Matrix Completion	23

In this chapter, we lay the theoretical foundations for understanding the following chapters in this dissertation. Throughout, we will clarify the notations used in this dissertation to avoid ambiguity. First, we cover how data is typically represented in the context of CADx. Next, we discuss the notion of models wherein we cover key information about Deep Neural Networks (DNN) and then discuss Graph Neural Network (GNN), which are networks that are suited to process input data with an associated graph-information. We briefly cover how we learn the parameters of these models. Then we explore Gradient-based feature attribution methods which are mainly used to give explanations how a model does its prediction. Finally, we cover geometric matrix completion. We expound these topics from a high-level to the point where it would be sufficient to understand most recent Deep Learning-based CADx methods.

2.1 Data Representations

Clinical data can come in various forms, depending on the modality. In the context of CADx, the most basic and common form is a collection of key characteristics which are considered as random variables. These are often expressed in form of binary, ordinal or numeric variables. For a single patient, this can be summarized as a vector. For multiple patients, the observations can be summarized in form of a table or matrix. As shown in Figure 2.1, observations from a single patient represented as a single row vector in this table can have different observations. These observations are commonly also referred to as features.

For the purpose of this thesis, we will be considering three categories of features: imaging, non-imaging, and meta-features. First, imaging features are typically values derived from an imaging modality describing the morphological properties or first-order statistics of an anatomy of interest. Second, non-imaging features can be observations or summary of observations

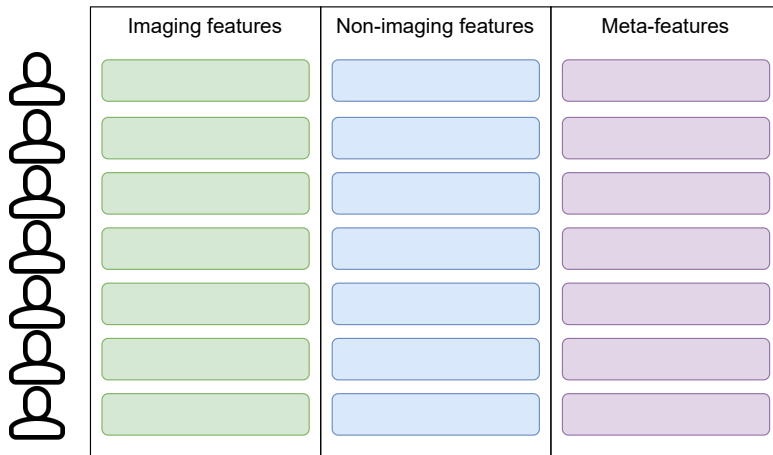


Fig. 2.1 Illustration of tabular data from clinical observations containing a complete set of imaging, non-imaging, and meta-features where each row denote patient information.

describing a patient not derived from imaging, for example clinical scores. Third, any feature that does not fall within the first two categories will be referred to as meta-features, for example, demographic information. We intend to use this separation in order to simplify the explanation on the graph construction section at the latter part of this chapter.

More concretely, given N instances or patients with D features or attributes, we have a table which can be represented as a matrix or two-dimensional tensor $\mathbf{X} \in \mathbb{R}^{N \times D}$. Here, when referring to a single instance it can be represented as a D -dimensional vector $x \in \mathbb{R}^D$. When we have image inputs in 2D or 3D these are represented as three or four-dimensional tensors in the system, respectively, wherein the additional dimension is for the channel intensities. We can also have sequential data. For example, given N instances of a D -dimensional sequence with length T time steps this is often represented as a tensor $\mathbf{X} \in \mathbb{R}^{N \times D \times T}$. Numerically, in computer languages and in memory, these are typically represented as 32-bit or 64-bit floating-point format. Note that we denote an input feature representation as a matrix or tensor \mathbf{X} for brevity. We will keep using this notation as much as possible to avoid any ambiguity.

In addition, before these data representations are used as input to the model, it is often common to standardize the input representations. This is often referred to as normalization as well. We will cover what we mean by the model in 2.2, but for now, this can be considered as a function or a “black-box” that takes in these data representations as input. To standardize these data representations, the most common approach for feature vector or sequential input is to center each scalar random variable in the data to have a zero-mean and unit-variance scaling. For image input, there are numerous ways to normalize the data, which often is dependent on the imaging modality [70]. A generic and very simple way for normalization is to scale the pixel intensities to a fixed range, e.g. $[0, 1]$.

There is also a notion of shallow features and deep features. In the context of CADx, shallow features are those features that are “hand-crafted” or clinically derived. These “hand-crafted” features are very important as these often represent domain expertise, which is derived from evidence-based research regarding a certain disease. On the other hand, are deep features,

these are feature representations taken from a layer of a pre-trained deep neural network. One common example is when taking the output feature vectors at the middle-most layer in an Autoencoder architecture. Here, the feature vector is a compressed representation of the raw input data.

2.1.1 Incomplete data in CADx

We also have to mention that the illustration in Figure 2.1 often does not depict real-world medical datasets as such data missingness contributes to the main challenges in CADx research [95]. Real-world datasets in the medical domain often contain incomplete observations resulting in incomplete data representations during analysis. For this thesis, we will refer to two forms of data missingness: (1) block-level missingness and (2) feature-level missingness.

We illustrate in Figure 2.2 how a real-world clinical dataset could look like when a “block” or set of features are missing. This is what we will refer to as block-level missingness or blockwise missingness. Such a setting could happen when there are no available imaging observations from a patient or when clinical scores are not available.

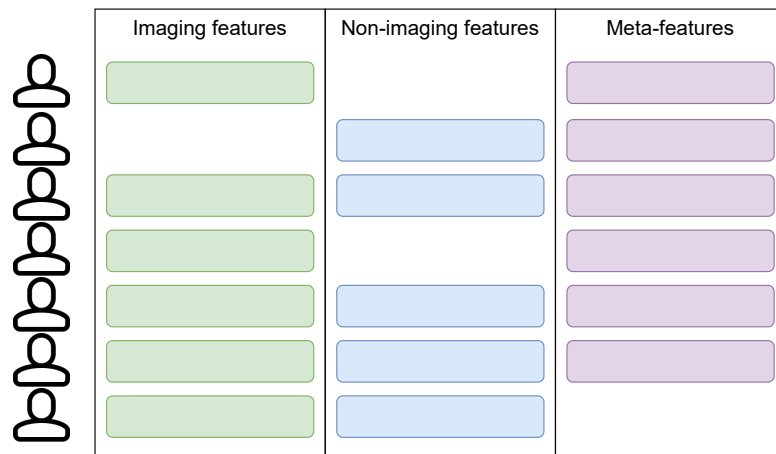


Fig. 2.2 Illustration of block-level missingness in clinical datasets containing incomplete set of imaging, non-imaging, and meta-features.

Another scenario is when there are certain elements from the set of observations are missing. Such a setting is what we will refer to as feature-level missingness as shown in Figure 2.3. Such a setting in CADx could happen when certain observations from a questionnaire for example is missing or when certain clinical observations are missing. Lastly, one could also encounter a mixture of both forms of data missingness in certain datasets in the medical domain or in CADx research specifically. There are certain simple strategies on how to address these forms of data missingness and we will tackle this in more detail when we present our work in 3. Now, we have a notion of how data is represented including the notion shallow/deep features and block-level/feature-level data missingness. We consider next the notion of what a model is.

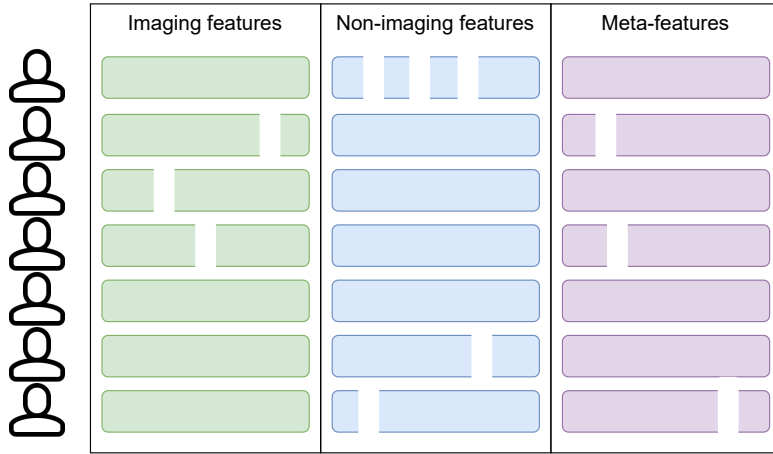


Fig. 2.3 Illustration of feature-level missingness in clinical datasets containing an incomplete set of imaging, non-imaging, and meta-features.

2.2 Machine Learning Models

From Perceptrons [64] to Graph Convolutions [10], the ML field has seen its ups and downs, including periods where artificial neural networks were left nearly unexplored in favour of other ML models, a period that is often referred to as “winter seasons” [52]. Currently, Deep Neural Networks (DNNs) [67] have mainly been the widely-used models in ML. Although there have been a lot of other models which were proposed in between Perceptrons and Graph Convolutions such as Tree-based models [8] and Support-Vector Machines [17], for the purposes of this thesis we will only cover recent models related to DNN. In the next section, we will elaborate on these DNN and also cover models which can process unstructured data called Graph Neural Network (GNN) [93].

2.2.1 Deep Neural Networks

The first type of models we will cover are Deep Neural Networks (DNNs). In its simplest case, we can view a DNN as a composition of functions that contains a linear function followed by a non-linear function [26]. We can compose multiple different functions which we would then call the neural network model architecture. For example, given L number of functions f_i containing both the linear and non-linear function we can have a composition of functions:

$$f := f_L \circ f_{L-1} \circ \dots \circ f_1 \tag{2.1}$$

where f_L is the last layer (output layer) and f_1 is the first layer (input layer).

Another way to describe a DNN is by grouping this composition of functions into three layers as depicted in Figure 2.4, i.e. input, hidden or intermediate, and output layers. The input layer is the initial layer which will process the input data representations and will then pass that information to the succeeding layers. The intermediate layers take the representations

from the input layer and process them further. We can stack multiple hidden representations together into layers of more than two or three NNs. Finally, the output layer typically is the layer that outputs the prediction of the model, which then gets fed-forward into the loss function. A gradient descent optimization algorithm is then used to find the optimal parameters for the model that will minimize this loss function.

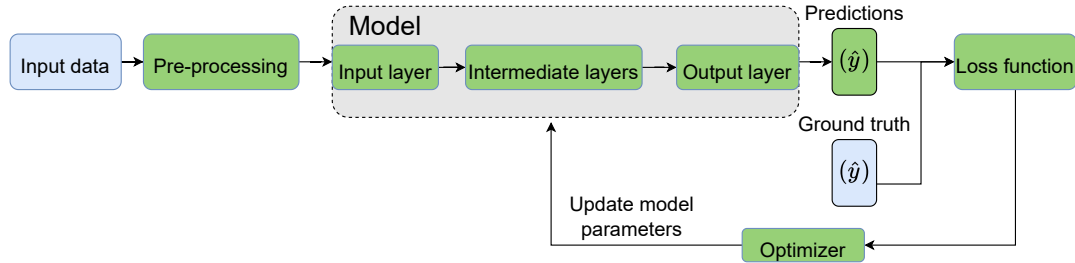


Fig. 2.4 Illustration of deep learning workflow from input data to loss function calculation including model parameter updates using an optimizer (gradient descent optimization algorithm).

For our purposes, when we say DNN we are referring to the following models: fully-connected feed-forward neural networks or Multi-layer Perceptrons (MLP) [26], Convolutional Neural Networks (CNN) [42], Recurrent Neural Networks (RNN) [30], and Transformers (TN) [84]. Depending on the input feature being processed and how you would model your problem, one is better suited over the other. A common way to match problems with model architectures is to use MLPs for vector-valued inputs, CNNs for images, and RNNs/TN for sequential.

Driven by the advancement in computational power and availability of large labelled datasets [52], DNNs have shown to be widely used models to model most learning-based problems. One aspect of their wide use is their ability to learn good representations automatically in an end-to-end manner while optimizing an objective function that satisfies a downstream task. At its core, DNNs are composed of multiple artificial neurons which pass on information from one neuron to the other neurons [26]. This is inspired by biological neural networks wherein neurons activate or send information to other neurons [52]. In non-linear DNNs, artificial neurons get “activated” using an activation function which determines how much information should be passed on to the next artificial neuron. When multiple of these artificial neurons are composed together they form an Artificial Neural Network (ANN). Multiple composition of these ANN results to “deep” architecture resulting to DNN. For more in-depth discussion on the different components of DNN we refer the reader the Deep Learning book by Goodfellow et al. [26].

2.2.2 Graph Neural Networks

The next type of models which we will cover are Graph Neural Networks (GNN) [93]. These are NN models which are suited to process graph-structured input. We first introduce some background information about graph theory before we elaborate further on GNN.

A graph is represented as $G = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ consisting of three entities, a set of nodes/vertices \mathcal{V} , a set of edges \mathcal{E} , and an adjacency matrix \mathbf{A} . The set of vertices or nodes are numbered from 1 to n ($\mathcal{V} = \{1, \dots, n\}$). The set \mathcal{E} contains edges $e = (i, j)$, which specify whether there

is a link between the i -th and j -th node. The entry $A_{i,j}$ denotes the weight $w_{i,j}$ of the edge between the two nodes. In this thesis, we only consider scalar-valued edge features but note that vector-valued edge features are completely possible [7]. When the edges are undirected (cf. Fig. 2.5), we have a symmetric graph adjacency matrix as illustrated in Fig. 2.6. The sum of the rows of the adjacency matrix can be represented in a diagonal matrix to denote the degree matrix \mathbf{D} . Additionally, a node within a graph may have attributes represented as a vector \mathbf{x}_i . In matrix form, all node features can be denoted as a feature matrix \mathbf{X} . We also have the notion of graph Laplacian matrix \mathbf{L} which is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ [10]. An example graph is shown in 2.5, including its adjacency matrix, degree matrix, and graph Laplacian matrix in 2.6 to illustrate these concepts.

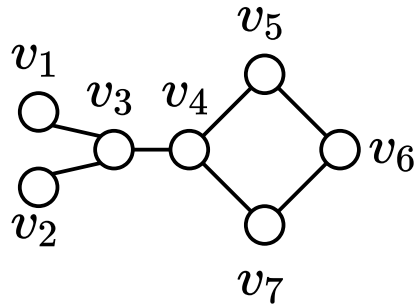


Fig. 2.5 Example undirected graph with seven nodes.

$$\begin{array}{ccc}
 \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} &
 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} &
 \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{bmatrix} \\
 \text{Adjacency Matrix (A)} & \text{Degree Matrix (D)} & \text{Graph Laplacian Matrix (L)}
 \end{array}$$

Fig. 2.6 Corresponding Adjacency matrix (left), Degree matrix (middle), and Graph Laplacian matrix (right) of an example seven-node graph shown in 2.5.

For our purposes, we categorize recent approaches into spectral-based and spatial-based approaches. Under spectral-based approaches, Bruna et al. [11] defined the convolutional operation in the Fourier domain using spectral graph theory by taking the eigen-decomposition of the graph Laplacian. There were two issues with this approach. First, the learned filters are not localized. Second, when dealing with large matrices, it becomes computationally expensive to compute the eigen-decomposition. Subsequent works proposed solutions to these issues. One such work by Defferrard et al. [19] proposed to use Chebyshev Polynomials to approximate the Graph Laplacian. Such approximation allowed the filters to be more localized and reduced the computational requirement. Further works [41, 92] proposed other improved solutions to address these issues and other approximations. Kipf and Welling [41] propose an efficient layer-wise propagation rule that is based on a first-order approximation of spectral

convolutions on graphs. Wu et al. [92] propose to successively remove nonlinearities and collapsing weight matrices between consecutive layers that is faster and more scaleable to larger datasets when compared with previous GCN formulations. As described in [9], the hidden representation under this category can be calculated as,

$$\vec{h}_v = \phi \left(\vec{x}_v, \bigoplus_{u \in \mathcal{N}_v} c_{vu} \psi(\vec{x}_u) \right) \quad (2.2)$$

where \vec{h}_v denotes node v 's hidden representation, ϕ a non-linear function, node v 's current feature representation denoted as \vec{x}_v , aggregation operator \bigoplus such as summation, $u \in \mathcal{N}_v$ denotes node v 's neighbours or nodes with connection to it, c_{vu} denotes constant weighting factor derived from the adjacency matrix specifying the contribution of node u to node v , and a local transformation function ψ .

When convolutions are applied directly on the graph such as in [28, 49, 86], we call these approaches spatial-based. One of the key ideas is to define an operator which works with arbitrarily sized neighborhoods and in a permutation-invariant manner, and at the same time has the weight sharing property of CNNs [86]. One such approach is GraphSAGE by Hamilton et al. [28] which samples a local neighbourhood and aggregates its neighborhoods information. This mainly operates on a fixed-size neighborhood of a node and learns to aggregate node information using different aggregation functions. Another approach proposed by Veličković et al. [86] is Graph Attention Networks (GAT). In GAT, all neighbors of a center node are taken into account by learning an attention mechanism. This also works on arbitrary sized neighborhoods, but learns self-attention scores on the full one-hop neighborhood. Unlike GraphSAGE, all the neighbors of the node in GAT are allowed to contribute to the target node. Their contributions are weighted via a single-layer feed forward neural network. As described by Bronstein et al. [9], the hidden representation under this category can be calculated as,

$$\vec{h}_v = \phi \left(\vec{x}_v, \bigoplus_{u \in \mathcal{N}_v} \alpha(\vec{x}_v, \vec{x}_u) \psi(\vec{x}_u) \right) \quad (2.3)$$

where \vec{h}_v denotes the hidden representation of node v , ϕ a non-linear function, \vec{x}_v current feature representation of node v , \bigoplus aggregation operator such as summation, $\alpha(\vec{x}_v, \vec{x}_u)$ computes the coefficient for the contribution of node u to node v , and a local transformation function ψ . Here, the difference with the previous spectral-based approach is that the weighting factor for the contribution of the neighbouring node is a learnable parameter.

For completeness, GNNs have also recently been categorized into three categories: (1) Convolutional, (2) Attentional, and (3) Message-Passing GNN [9]. First, Convolutional GNNs aggregate neighbouring features of a target node using a fixed weight that is directly dependent on the adjacency matrix. Such examples under this category include ChebNet [19], GCN [41], and Simple Graph Convolution (SGC) [92]. Second, when the aggregation of neighboring features for a target node are implicit weights such as an attention-like mechanism these can be categorized as Attentional GNNs [9]. Recent works such as Mixture Model Networks (MoNet)

[49] and GAT [86] fall under this category. The third category are the Message-Passing GNNs which are the most generic [9]. Here, instead of just aggregating the neighboring features of a target node, arbitrary feature vectors (“messages”) are computed based on the sending and receiving nodes’ feature vectors which are then sent across edges [9]. Such examples under this category include Interaction Networks [6], Message-Passing Neural Networks (MPNN) [24], and GraphNets [7]. Under the message-passing category, as described in [9], the hidden representation can be calculated as,

$$\vec{h}_v = \phi \left(\vec{x}_v, \bigoplus_{u \in \mathcal{N}_v} \psi(\vec{x}_u, \vec{x}_v) \right) \quad (2.4)$$

where \vec{h}_v is the hidden representation of node v , non-linear function ϕ , current feature representation of node v denoted as \vec{x}_v , learnable message function $\psi(\vec{x}_u, \vec{x}_v)$, and aggregation operator \bigoplus as a form of message passing on the graph. We illustrate this message passing concept visually in Figure 2.7 where corresponding “messages” (depicted in colour for brevity) as shown on the left panel of this figure. In the same figure in 2.7 (right), messages from the adjacent nodes are passed on.

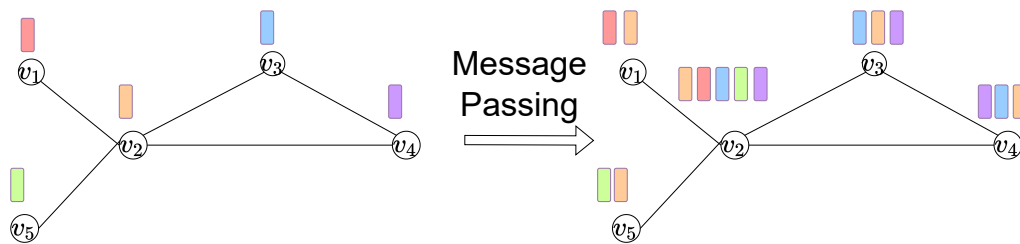


Fig. 2.7 Simplified illustration of message passing concept wherein at first step each node contains a feature vector that represents the “message” it wants to send to all its neighbours (left panel). In the next step (right panel), the messages are passed on to their neighbours.

Graph Construction for Population Modelling

The other important element in GNN-based CADx is the graph structure itself. In CADx, one way to model a disease classification problem is via node classification. Here, the nodes in the graph represent patients and the edge connections capture some form of similarity between pairs of patients [58]. In addition, by creating a graph we can capture additional inductive biases that represent certain medical knowledge about the task at hand. For example, we create a population graph that captures the connectivity of patients in relation to their risk of developing a certain disease based on clinical risk factors that are clinically known to play a role in developing such a disease. Such factors could be the age of the patient, the gender, general health markers like body-mass-index, bloodwork parameters, or the presence of genetic markers that indicate predispositions for a disease. In such construction, we can have a sparse graph which captures this higher level information regarding the disease.

An example graph is shown in Figure 2.8 to illustrate this idea. Every node in the graph represents a patient with their corresponding feature vector. The binary edge weight can be constructed using the meta-features using:

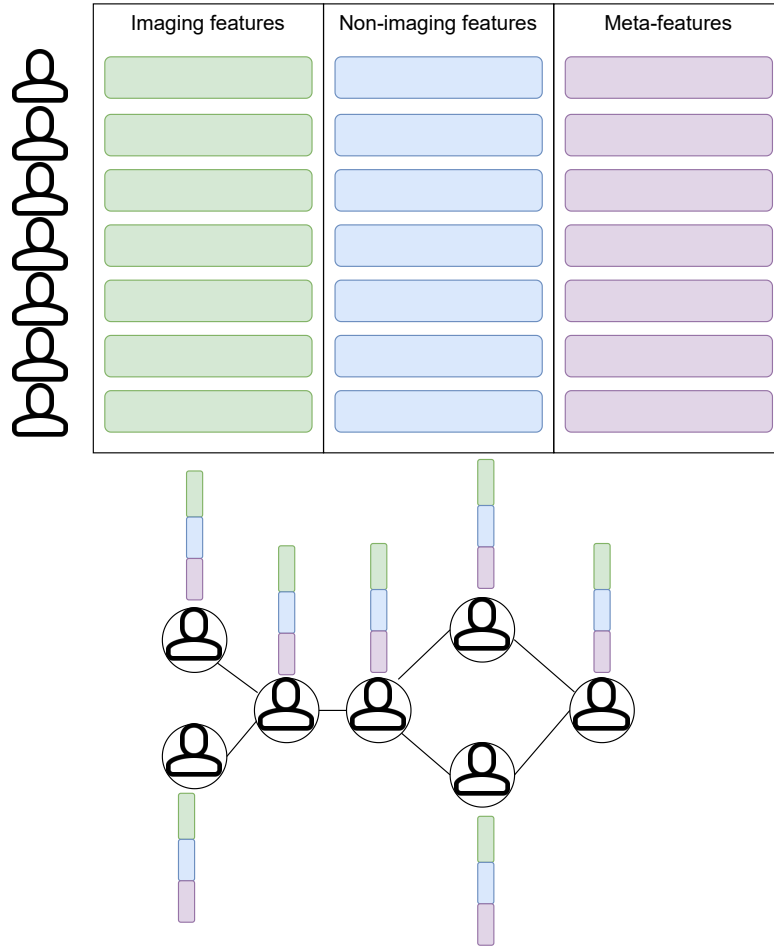


Fig. 2.8 Population graph modelling of a tabular dataset in CADx where every row in the table (top) are patients with corresponding imaging-, non-imaging-, and meta-features. Every patient is modelled as a node in the graph (bottom) with their corresponding feature vector representations (row vectors).

$$A_{i,j} = \begin{cases} 1, & \text{if } |dist(i,j)| \leq d \\ 0, & \text{otherwise} \end{cases} \quad (2.5)$$

wherein we link two nodes if the distance $dist(i,j)$ between two nodes are below or equal a given threshold d [58]. We can calculate the distance using the elements of the meta-features to compare similarities between two patients. Alternatively, one could also use all the features and link patients with their K number of neighbours in the feature space. Instead of using KNN, one can also calculate a thresholded Gaussian kernel [73] using the feature matrix to obtain non-binary weights.

$$A_{i,j} = \begin{cases} \exp(-\frac{|dist(i,j)|^2}{2\theta^2}), & \text{if } dist(i,j) \leq d \\ 0, & \text{otherwise} \end{cases} \quad (2.6)$$

Finally, for completeness, there are more recent methods wherein we can instead learn this graph in an end-to-end manner as part of the training process, as proposed in [16]. Instead of calculating the graph in the feature space, they propose to learn latent embedding features along with a differentiable soft-kNN layer which is used to link patients and derive the graph structure.

Modelling problem in GNN

Given the feature matrix and graph structure, the next question is how to model the problem. There are three common tasks related to GNNs [93]. The first one is node-level tasks. This could either be a node-level classification or regression depending on the target variable. Most GNN-based CADx methods so far have focused on node-level classification. The goal in a node-level classification task is to predict the class labels of every node in the graph. In our previous example in 2.8, we want to assign class labels \hat{y} to every node as shown in 2.9.

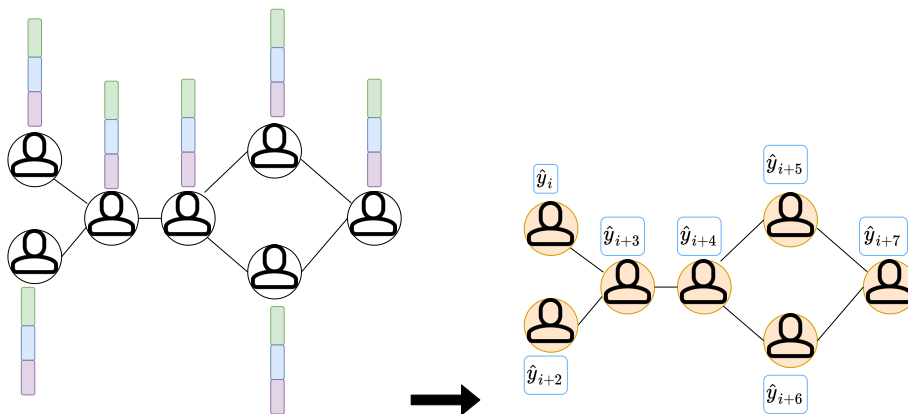


Fig. 2.9 Node-level classification task assigning class label \hat{y} to every node. Given a graph and node's feature representations (left) the goal is to assign class label \hat{y} to every node in the graph (right).

The other two tasks related to GNN are edge-level task and graph-level task [93]. For our purposes, we will not go into much more detail about these two tasks as we will not be talking about such tasks in this thesis. For a brief overview, edge-level tasks are related to link prediction. Here, the task is to decide where there is a link between two nodes. Typical applications include recommender systems in e-commerce, or friend suggestion in social networks [93]. For graph-level tasks, every input sample is a graph itself as shown in Figure 2.10 and the goal is to either perform graph-level classification or regression. A typical application domain is molecular property prediction [72] or drug discovery [51]

2.3 Model Training

2.3.1 Supervised Learning

When there is a target label $y_i \in \mathbb{R}$ associated with every input instance x_i , we can train a model in a supervised manner. The goal is to learn parameters using data that can generalize to unseen samples [26]. This is typically done by partitioning your data into training, validation, and test sets. We use the training set to estimate the parameters of a model and select optimal

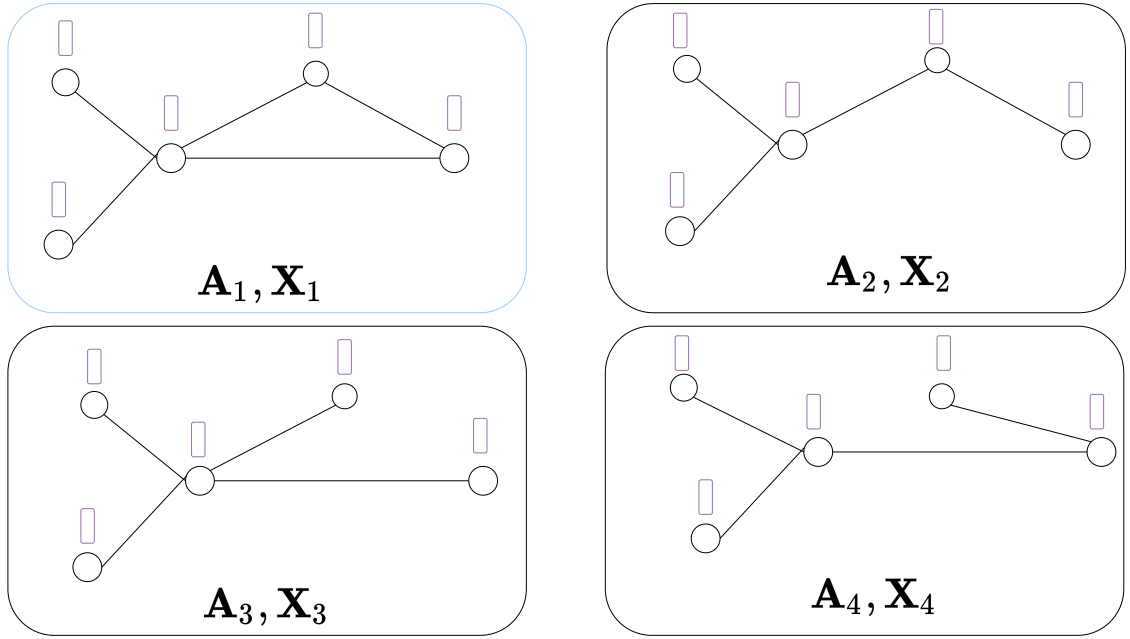


Fig. 2.10 Graph-level classification task assigning class label \hat{y} (illustrated based on the colour of the bounding box) to every graph given their corresponding adjacency matrices \mathbf{A}_i and feature matrices \mathbf{X}_i .

hyper-parameters using the validation set. Then we evaluate the model's generalization performance on the test set [26].

In the context of CADx, this implies how to pick the model which can produce good results given input from the left figure in 2.9 to the right figure in 2.9. To this end, we can follow the same training workflow as in regular deep learning. Given an independent and identically distributed dataset $\{(x_i, y_i), \dots, (x_n, y_n)\}$ we want to pick the parameters Θ^* of the model that minimizes a given loss function as shown in Equation 2.7 and Equation 2.8 respectively.

$$\Theta^* = \arg \min_{\Theta} \mathcal{L}(\Theta) \quad (2.7)$$

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_i \ell(f_{\Theta}(x_i), y_i) \quad (2.8)$$

Both the DNN and GNN use the same training workflow. The parameters of the DNN or GNN can be updated using stochastic gradient descent algorithms [66] or its variant such as Adam optimizer [40]. We update the parameters of the model by following the gradient in the opposite direction as in Equation 2.9. For simplicity, we denote all learnable parameters of a model with Θ .

$$\Theta = \Theta - \delta(\nabla_{\Theta} \mathcal{L}(\Theta)) \quad (2.9)$$

For completeness, when there are no associated target labels (or supervised ground truth labels) and no rewards from the environment given for learning this is known as Unsupervised learning [23], which will not be covered in this dissertation. Another category of learning is self-supervised learning wherein an auxiliary task is performed using the training data as a supervised information to initialize the weights of the model [32]. Another category of learning which will not be discussed in this dissertation is Reinforcement Learning wherein the goal is to have an agent that is able to maximize its rewards when placed in an unknown environment [79].

2.3.2 Transductive and Inductive Learning

In the context of GNN-based CADx, one would also come across the notion of Transductive Learning and Inductive Learning. In Transductive Learning, the setting typically involves training the model by including the features and labels of the training set as well as the features of the test set. In contrast, Inductive Learning only uses the features and labels of the training set. The test dataset is held out fully and are not used to learn the parameters of the model. In this regard, Transductive learning requires that the number of samples/instances are known a priori including the test set. Such setting often can be observed in spectral-based GNNs approaches.

2.4 Feature Attribution Methods

Deep learning models such as DNNs and GNNs have been successfully applied to different machine learning problems. However, often such models are considered “black boxes” since it is often not known what these models are doing to form the decision that an instance is labelled as healthy and the other as having a disease. In the medical domain, understanding how a network arrives at a particular prediction or decision is of great importance [3]. Healthcare providers would like to know why a patient was predicted to have a disease.

To “understand” what the model is doing, we can provide model explanations in the form of feature attributions. Feature attributions indicate how much a feature contributed to the model for a given instance [45]. Although there has been a lot of criticism regarding the “black-box” nature of neural network models, a growing body of knowledge has started to tackle this issue. This new research field of “AI interpretability”, or “Explainable AI (XAI)” is summarized by Molnar [48] in a comprehensive overview of this field. For our purposes, we will mainly focus on a few selected gradient-based feature attribution methods.

One such approach is *Saliency* [74] which calculates the feature attribution by using the gradient of the output with respect to the input. An extension of this called *Gradient * Input* [71] which calculates feature attributions by taking the (signed) partial derivatives of the output with respect to the input and multiplying them with the input feature values. Both approaches has one drawback, that is it violates the sensitivity axiom for feature attribution. This could result to having inaccurate feature attributions. To address this limitations, an axiomatic feature attribution approach was introduced by Sundararajan et al. [77] called Integrated Gradients (IG).

The idea in IG is to integrate all the gradients of the outputs with respect to different input features along a linear path from a given baseline starting point, up to the actual input features. A baseline in this context means a neutral input. For example in images, a baseline input would mean using a constant black image [77]. Compared to other gradient-based methods for feature attribution, IG is superior as it satisfies axiomatic properties [77]. For instance, two important axioms that attribution methods must satisfy, as identified by Sundararajan et al. [77] are *Sensitivity* and *Implementation Invariance*. *Sensitivity* axiom states that non-zero attribution must be given to a feature if that feature changes the output prediction. *Implementation Invariance* states that feature attributions should be identical for two functionally equivalent networks.

Using such an axiomatic feature attribution method could be more sensitive and accurate in terms of giving attribution to certain input features, based on the predicted class output. In practice, IG is approximated using:

$$\text{IG}_{i,k}^{\text{approx}}(\bar{x}_i, \text{class}_k) = (\bar{x}_i - x'_i) \times \sum_{s=1}^m \frac{\partial F(x' + \frac{s}{m} \times (\bar{x} - x'))}{\partial \bar{x}_i} \times \frac{1}{m} \quad (2.10)$$

where \bar{x} is the input vector and x'_i the baseline at the i -th dimension, $\frac{\partial F(\cdot)}{\partial \bar{x}_i}$ is the partial derivative of the network's output with respect to input \bar{x}_i , and m is the number of approximation steps of the path integral in IG

2.5 Geometric Matrix Completion

When dealing with clinical data it is common that healthcare experts provide the data in form of tabular data and is represented numerically as a matrix. Here, every column represents observations or clinically derived features and every row represents a patient. The ideal setting is to have a matrix where all entries are filled, i.e. all observations are available, and the resulting data matrix is complete, as illustrated in the left panel of Fig. Figure 2.8. However, in practice, this is often not the case [20]. It is more common that not all observations will be available for every patient resulting in a dataset that is incomplete or a matrix which contain “holes” visually.

One approach to complete matrices with incomplete information is by matrix completion [15]. In general, the matrix completion problem is not well-posed [50] and one way to deal with this is to impose certain assumptions. One such assumption is to assume that the matrix is a low-rank matrix [13]. More formally, given a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ be an incomplete feature matrix where a certain proportion of values is missing at random. The goal is to recover the missing values in this matrix. One solution to this problem is by using rank minimization as described in [50],

$$\min_{\hat{\mathbf{X}} \in \mathbb{R}^{n \times m}} \text{rank}(\hat{\mathbf{X}}) \text{ s. t. } \hat{x}_{i,j} = x_{i,j}, \forall (i,j) \in \Omega, \quad (2.11)$$

where $\hat{\mathbf{X}}$ is the predicted matrix of size $n \times m$ with values $\hat{x}_{i,j}$, Ω denotes the set of known entries in matrix \mathbf{X} with $x_{i,j}$ values. Here, we try to find a low-rank matrix such that the values in matrix $\hat{\mathbf{X}}$ are equal to the known values in matrix \mathbf{X} .

However, rank minimization is known to be computationally intractable and instead the solution can be relaxed by solving for the nuclear norm $\|\hat{\mathbf{X}}\|_*$ [15]. In addition, to make the model robust to noise, the equality constraint on the entries of the matrix can be replaced with the squared Frobenius norm $\|\cdot\|_F^2$ [14],

$$\min_{\hat{\mathbf{X}} \in \mathbb{R}^{n \times m}} \|\hat{\mathbf{X}}\|_* + \frac{\gamma}{2} \|\Omega \circ (\mathbf{X} - \hat{\mathbf{X}})\|_F^2, \quad (2.12)$$

where γ is a hyperparameter controlling the contribution of the second loss term, Ω is the masking matrix of known entries in \mathbf{X} with values 1 (known) and 0 (missing), and \circ is the Hadamard product.

An alternative approximation to the previous solution is to assume that the entries of the matrix are smooth with respect to some geometric structure such as a graph [36, 50, 61]. Here, a graph structure is built based on the rows or columns of the matrix. Unlike previous approaches that make use of the graph as a regularization term, the most recent formulation to this approximation makes use of graph signal processing to complete the matrix, which is referred as geometric matrix completion. One such approach was published by Monti et al. [50], who proposed to complete the matrix using geometric matrix completion on graphs, through a combination of GCN and LSTM networks. Here, the GCNs are used to learn filters which compute feature representations that are smooth with respect to the graph structure, while the LSTM is used to model the diffusion process. The problem boils down to minimizing the loss,

$$\ell(\Theta, \delta) = \|\hat{\mathbf{X}}_{\Theta, \delta}\|_{D,r}^2 + \|\hat{\mathbf{X}}_{\Theta, \delta}\|_{D,c}^2 + \frac{\gamma}{2} \|\Omega \circ (\hat{\mathbf{X}}_{\Theta, \delta} - \mathbf{X})\|_F^2 \quad (2.13)$$

where Θ and δ denote parameters of the GCNs and LSTMs, respectively, $\hat{\mathbf{X}}_{\Theta, \delta}$ is the predicted matrix which is dependent on the parameters of the GCN and LSTM, $\|\cdot\|_{D,r}^2$ denotes the Dirichlet norm on the row-graph, $\|\cdot\|_{D,c}^2$ denotes the Dirichlet norm on the column-graph \mathbf{X} is the known matrix, Ω is the masking matrix of known entries in \mathbf{X} , $\|\cdot\|_F^2$ denotes the squared Frobenius norm, and γ is a hyperparameter to weight the second term. In equation (2.13), the first term on the right is defined as $\text{tr}(\hat{\mathbf{X}}^T \mathbf{L} \hat{\mathbf{X}})$ which contains a rescaled graph Laplacian ($\mathbf{L} \in \mathbb{R}^{n \times n}$) term such that its eigenvalues are in the interval $[-1, 1]$. The first two terms keep the prediction smooth with respect to the row and column graph structure, respectively. Additionally, the last term will keep the imputed values as close as possible to the observed values.

Part II

Contributions

Addressing data missingness during model training and testing in Diagnostic Decision Support Systems

Contents

3.1	Problem Definition and Motivation	27
3.2	Related works	28
3.3	Contributions	30
3.3.1	Addressing missingness in GNN-based CADx using Geometric Matrix Completion (GRAIL-MICCAI 2018)	30
3.3.2	Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification. (AI in Medicine 2021)	42

3.1 Problem Definition and Motivation

One of the major challenges when handling real-world data in healthcare is that data often contain missing information [87]. This can happen when certain observations from a patient are not available due to different reasons such as the invasiveness of an observation, compliance of a patient, resource limitation or random unexpected scenarios. These random occurrences result to an incomplete dataset. In literature [65], incomplete datasets can be categorized into three types, based on the mechanism of missingness: (1) missingness completely at random (MCAR), (2) missingness at random (MAR), and (3) missingness not at random (MNAR).

The first one is MCAR and happens if the probability of missingness is the same for all cases [83]. This means that the missing data has no relationship to any of the observed features. An example is a device failure, e.g. a broken ultrasound machine that prohibited the collection of images and features for one or several patients. This is one form of missingness completely at random and it has nothing to do with the observed features. Such assumption is often considered unrealistic for the data [83] which makes the MAR assumption more favourable in practice. When the mechanism of missingness has some form of relationship to the observed features then it is considered as MAR [20]. This is often the assumption in practice wherein the missing feature could be related to the characteristic of a certain group. One example is when a certain group of individuals are more likely to do this observation than others. When the mechanism of missingness is neither MCAR nor MAR, it is considered MNAR [20]. This

implies that the missing data is related to an unobserved variable/feature. Here, the observed data has no relationship with the missing data. In practice, understanding which mechanism of missingness could be useful and in practice the MAR assumption is often used.

Beyond these three categorization on the mechanism of missingness, data missingness could also come in another form based on the type of input. We often have data missingness at a feature level or at a modality level when dealing with cross-sectional data. Missingness at the feature level happens when individual features are missing. In contrast, when a block or group of features from a multi-modal data is missing, this leads to a modality-level missingness. One example of modality-level missingness in the context of CADx is when we are dealing with a multi-modal dataset wherein the input features are coming from multiple input sources such as T1-weighted MRI, T2-weighted MRI, and clinical scores but one full modality observation is missing. Then we view this as a modality-level missingness. Depending on how we model the problem, we can also view longitudinal data as modality level missing as there could be missing modality at certain acquisition time points. As we are not only limited to these two forms, it could also happen that both forms of missingness are present in the dataset.

There are also other forms of missingness that we will not cover in this thesis. For example, when given an image but a part of the image is blocked or corrupted, this could also be considered another form of missingness. When we are given a sequential data such as measurements from the Intensive Care Unit (ICU) like heart rate, blood pressure, or body temperature measured at different time points this is again a form of missingness in sequential data.

In the context of CADx, often the issue with data missingness is not fully given much attention [94]. The input data is often assumed to be complete. For example, methods developed for automated disease diagnosis using GNN has shown to be successful. Particularly, GNN-based models have been shown to outperform linear and non-linear ML models [58]. In their seminal work, Parisot et al. [58] introduced how to leverage imaging and non-imaging modality for population based modeling in CADx by using GNN. They successfully showed that by modelling a disease classification problem as a node level classification this could outperform previous AD and ADHD classification approaches. Subsequent works followed to address limitations, such as [57]. In order to perform analysis, it is common to either exclude the sample or impute the missing value using the expected value of that variable which is the mean of that feature using the training set. A more recent approach is to use geometric matrix completion to address data missing in CADx. In this thesis, we propose such streamlined approach that instead simultaneously imputes the missing information and performs classification using geometric matrix completion, which we will elaborate in the next sections.

3.2 Related works

Recent works, dealing with missing information for CADx can be categorized into non-learning-based and learning-based approaches [20] for this thesis. Those approaches that do not make use of any learnable parametric function to deal with missing information are under a non-learning-based approach. Those approaches that make use of the observed data to learn a

parametric function to estimate the missing entries in the data are under the learning-based approaches.

We start by looking at non-learning-based approaches. One of the simplest non-learning based approaches to handle data missingness is to remove all instances that contain missing information, which is also known as sample or row deletion [20]. This approach is very simple but one drawback of this approach is that we could end up losing important information from those deleted samples, or worse, it might even prohibit data analysis. Alternatively, one could also just include all the features which are complete (i.e. column deletion), but this might result in a biased analysis. In addition, we could also end up discarding valuable information from instances which may contain useful information. So instead of removing the instance, we can instead calculate what is the expected value of that missing information by taking the mean of that particular feature using the training set. This is one example of simple imputation approach. The advantage of this is that it is fast and simple, however the resulting imputation could reduce the variance of the dataset and lead also to a biased analysis [20]. A disadvantage is also that the mean over the entire dataset weighs all samples equally, regardless of their distance to the current sample. Instead of taking the mean of all the samples in the dataset equally, we can take the mean of the K-nearest neighbors in Euclidean space and use that value to impute the missing information [82]. This requires that the input are noise-free and that the proximity of instances in the input feature space is useful. In addition, this approach requires the right balance as to how many neighbors should be considered in calculating the mean.

Under learning-based methods, we can instead learn a parametric function to estimate the values of the missing data. Unlike the previous methods which use heuristics, we can instead impute the missing information using the observed values in the data. One approach that imputes the missing data using the observed entries in the data is called Multiple Imputation of Chained Equations (MICE) [12]. MICE works by iteratively building predictive models to regress the missing values per instance. One drawback of this is that it is computationally time consuming as every instance will be imputed one at a time. Another approach is Probabilistic Principal Component Analysis (PPCA) [81], which is a learning based approach that makes use of the Expectation-Maximization algorithm to estimate values of missing data points.

Another approach is Matrix Completion (MC) [15]. The MC problem is known to be ill-posed unless certain assumptions about the matrix are imposed. One assumption is that the matrix is of low-rank. However, rank-minimization turns out to be an NP-hard combinatorial problem and as a proxy solution is to relax the rank-minimization problem into minimizing the sum of the singular values of the matrix previously. Another method to relax the solution is to impose that the solution is smooth with respect to some geometric structure like a graph. Previous works which incorporate some graph information make use of it to regularize the solution while others make use of graph signal processing. Motivated by the success of GNNs, we utilized geometric matrix completion for CADx on datasets which contain missing information. MGMC learns to simultaneously impute and classify the target class labels. We describe our contributions on addressing data missingness during model training and testing in CADx in the next section.

3.3 Contributions

3.3.1 Addressing missingness in GNN-based CADx using Geometric Matrix Completion (GRAIL-MICCAI 2018)

Modelling clinical diagnosis as a node-level classification problem using GNN has been successfully applied [57, 58]. However, these approaches assume that the input data is complete. To address the issue of missingness for GNN-based CADx we proposed a Geometric Matrix Completion (GMC) that simultaneously imputes and classifies the target class of every instance, following the approach from [50]. One difference of our proposed approach to Monti et al. [50] is that we add a classification label as part of the matrix completion problem. The inclusion of class labels for simultaneous imputation and classification has already been explored previously by Goldberg et al. [25], which makes their approach similar to our proposed GMC method. However, they did not include any graph information in their analysis, neither did they consider graph signals within the nodes.

Multi-modal Disease Classification in Incomplete Datasets Using Geometric Matrix Completion (GRAIL-MICCAI 2018)

Gerome Vivar^{1,2}, Andreas Zwergal², Nassir Navab^{1,3}, Seyed-Ahmad Ahmadi^{1,2}

¹Technische Universität München (TUM), Munich, Germany

²German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-Universität (LMU), Munich, Germany

³Johns Hopkins University, Baltimore, USA.

Copyright Statement. © Springer Nature Switzerland AG 2018. Reprinted, with permission, from Gerome Vivar , Andreas Zwergal , Nassir Navab , and Seyed-Ahmad Ahmadi, Multi-modal Disease Classification in Incomplete Datasets Using Geometric Matrix Completion, September 2018. Original publication can be found at: (https://doi.org/10.1007/978-3-030-00689-1_3). Included in this dissertation is the authors' accepted version due to copyrights.

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.

Multi-modal Disease Classification in Incomplete Datasets Using Geometric Matrix Completion

Gerome Vivar^{1,2}, Andreas Zwergal², Nassir Navab¹, and Seyed-Ahmad Ahmadi² for the Alzheimer’s Disease Neuroimaging Initiative*

¹ Technical University of Munich (TUM), Munich, GER

² German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-Universität (LMU), Munich, GER

Abstract. In large population-based studies and in clinical routine, tasks like disease diagnosis and progression prediction are inherently based on a rich set of multi-modal data, including imaging and other sensor data, clinical scores, phenotypes, labels and demographics. However, missing features, rater bias and inaccurate measurements are typical ailments of real-life medical datasets. Recently, it has been shown that deep learning with graph convolution neural networks (GCN) can outperform traditional machine learning in disease classification, but missing features remain an open problem. In this work, we follow up on the idea of modeling multi-modal disease classification as a matrix completion problem, with simultaneous classification and non-linear imputation of features. Compared to methods before, we arrange subjects in a graph-structure and solve classification through geometric matrix completion, which simulates a heat diffusion process that is learned and solved with a recurrent neural network. We demonstrate the potential of this method on the ADNI-based TADPOLE dataset and on the task of predicting the transition from MCI to Alzheimer’s disease. With an AUC of 0.950 and classification accuracy of 87%, our approach outperforms standard linear and non-linear classifiers, as well as several state-of-the-art results in related literature, including a recently proposed GCN-based approach.

1 Introduction

In clinical practice and research, the analysis and diagnosis of complex phenotypes or disorders along with differentiation of their aetiologies rarely relies on a single clinical score or data modality, but instead requires input from various modalities and data sources. This is reflected in large datasets from well-known

* Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

multi-centric population studies like the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and its derived TADPOLE grand challenge ³. TADPOLE data, for example, comprises demographics, neuropsychological scores, functional and morphological features derived from MRI, PET and DTI imaging, genetics, as well as histochemical analysis of cerebro-spinal fluid. The size and richness of such datasets makes human interpretation difficult, but it makes them highly suited for computer-aided diagnosis (CAD) approaches, which are often based on machine learning (ML) techniques [10, 11, 16]. Challenging properties for machine learning include e.g. subjective, inaccurate or noisy measurements or a high number of features. Linear [11] and non-linear [16] classifiers for CAD show reasonable success in compensating for such inaccuracies, e.g. when predicting conversion from mild-cognitive-impairment (MCI) to Alzheimer’s disease (AD). Recent work has further shown that an arrangement of patients in a graph structure based on demographic similarity [12] can leverage network effects in the cohort and increase robustness and accuracy of the classification. This is especially valid when combined with novel methods from geometric deep learning [1], in particular spectral graph convolutions [7]. Similar to recent successes of deep learning methods in medical image analysis [8], deep learning on graphs shows promise for CAD, by modeling connectivity across subjects or features.

Next to noise, a particular problem of real-life, multi-modal clinical datasets is missing features, e.g. due to restrictions in examination cost, time or patient compliance. Most ML algorithms, including the above-mentioned, require feature-completeness, which is difficult to address in a principled manner [4]. One interesting alternative to address missing features is to model CAD and disease classification as a matrix completion problem instead. Matrix completion was proposed in [5] for simultaneously solving the three tasks of multi-label learning, transductive learning, and feature imputation. Recently, this concept was applied for CAD in multi-modal medical datasets for the first time [15], for prediction of MCI-to-AD conversion on ADNI data. The method introduced a pre-computed feature weighting term and outperformed linear classifiers on their dataset, however it did not yet leverage any graph-modeled network effects of the population as in [12]. To this end, several recent works incorporated a geometric graph structure into the matrix completion problem [6, 9, 13]. All these methods were applied on non-medical datasets, e.g. for recommender systems [9]. Hence, their goal was solely imputation, without classification. Here, we unify previous ideas in a single stream-lined method that can be trained end-to-end.

Contribution. In this work, we follow up on the idea of modeling multi-modal CAD as a matrix completion problem [5] with simultaneous imputation and classification [15]. We leverage cohort network effects by integrating a population graph with a solution based on geometric deep learning and recurrent neural networks [9]. For the first time, we demonstrate geometric matrix completion (GMC) and disease classification from multi-modal medical data, towards MCI-to-AD prediction from TADPOLE features at baseline examination. In this difficult task, GMC significantly outperforms regular linear and non-linear

³ <http://adni.loni.usc.edu> || <https://tadpole.grand-challenge.org/>

machine learning methods as well as three state-of-the-art results from related works, including a recent approach based on graph-convolutional neural networks.

2 Methods

2.1 Dataset and Preprocessing

As an example application, we utilize the ADNI-based TADPOLE dataset, with the goal of predicting whether an MCI subject will convert to AD given their baseline information. We select all unique subjects with baseline measurements from ADNI1, ADNIGO, and ADNI2 in the TADPOLE dataset which were diagnosed as MCI including those diagnosed as EMCI and LMCI. Following [15], we retrospectively label those subjects whose condition progressed to AD within 48 months as cMCI and those whose condition remained stable as sMCI. The remaining MCI subjects who progressed to AD after month 48 are excluded from this study. We use multi-modal features from MRI, PET, DTI, and CSF at baseline, i.e. excluding longitudinal features. We use all numerical features from this dataset to stack with the labels and include age and gender to build the graph, following the intuition and methodology from [12].

2.2 Matrix Completion

We will start by describing the matrix completion problem. Suppose there exists a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ where the values in this matrix are not all known. The goal is to recover the missing values in this matrix. A well-defined description of this problem is to assume that the matrix is of low rank [2],

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \text{rank}(\mathbf{X}) \text{ s.t. } x_{ij} = y_{ij}, \forall ij \in \Omega, \quad (1)$$

where \mathbf{X} is the $m \times n$ matrix with values x_{ij} , Ω is the set of known entries in matrix \mathbf{Y} with y_{ij} values. However, this rank minimization problem (1) is known to be computationally intractable. So instead of solving for $\text{rank}(\mathbf{X})$, we can replace it with its convex surrogate known as the nuclear norm $\|\mathbf{X}\|_*$ which is equal to the sum of its singular values [2]. In addition, if the observations in \mathbf{Y} have noise, the equality constraint in equation (1) can be replaced with the squared Frobenius norm $\|\cdot\|_F^2$ [3],

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \|\mathbf{X}\|_* + \frac{\gamma}{2} \|\Omega \circ (\mathbf{Y} - \mathbf{X})\|_F^2, \quad (2)$$

where Ω is the masking matrix of known entries in \mathbf{Y} and \circ is the Hadamard product. Alternatively, a factorized solution to the representation of the matrix \mathbf{X} was also introduced in [13, 14], as the formulation using the full matrix makes it hard to scale up to large matrices such as the famous Netflix challenge. Here,

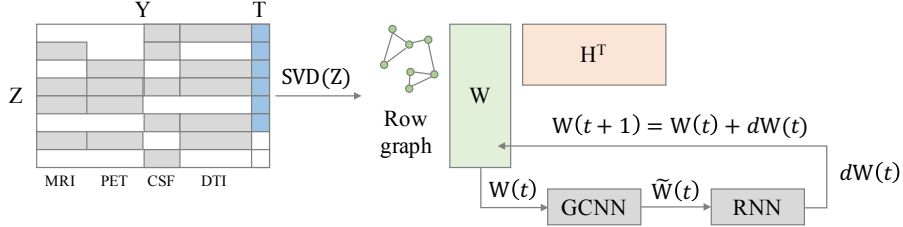


Fig. 1. Illustration of the overall approach: the matrix Z comprising incomplete features and labels is factorized into $Z = WH^T$. A connectivity graph is defined over rows W . During optimization, GCNN filters are learned along with RNN parameters and weight updates for W , towards optimal matrix completion of Z and simultaneous inference of missing features and labels in the dataset.

the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is factorized into 2 matrices \mathbf{W} and \mathbf{H} via SVD, where \mathbf{W} is $m \times r$ and \mathbf{H} is $n \times r$, with $r \ll \min(m, n)$. Srebro et al. [14] showed that the nuclear norm minimization problem can then be rewritten as:

$$\min_{W, H} \frac{1}{2} \|\mathbf{W}\|_F^2 + \frac{1}{2} \|\mathbf{H}\|_F^2 + \frac{\gamma}{2} \|\Omega \circ (\mathbf{W}\mathbf{H}^T - \mathbf{Y})\|_F^2 \quad (3)$$

2.3 Matrix Completion on Graphs

The previous matrix completion problem can be extended to graphs [6, 13]. Given a matrix \mathbf{Y} , we can assume that the rows/columns of this matrix are on the vertices of the graph [6]. This additional information can then be included into the matrix completion formulation in equation (2) as a regularization term [6]. To construct the graph, we can use meta-information out of these rows/columns or use the row/column vectors of this matrix to calculate a similarity metric between pairs of vertices. Given that every row in the matrix has this meta-information, Kalofolias et al. [6] showed that we can build an undirected weighted row graph $G_r = (V_r, E_r, A_r)$, with vertices $V_r = \{1, \dots, m\}$. Edges $E_r \subseteq V_r \times V_r$ are weighted with non-negative weights represented by an adjacency matrix $A_r \in \mathbb{R}^{m \times m}$. The column graph $G_c = (V_c, E_c, A_c)$ is built the same way as the row graph, where the columns are now the vertices in G_c . Kalofolias et al. [6] showed that the solution to this problem is equivalent to adding the Dirichlet norms, $\|\mathbf{X}\|_{D,r}^2 = \text{tr}(X^T L_r X)$ and $\|\mathbf{X}\|_{D,c}^2 = \text{tr}(X L_c X^T)$, where L_r and L_c are the unnormalized row and column graph Laplacian, to equation (2),

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \|\mathbf{X}\|_* + \frac{\gamma}{2} \|\Omega \circ (\mathbf{Y} - \mathbf{X})\|_F^2 + \frac{\alpha_r}{2} \|\mathbf{X}\|_{D,r}^2 + \frac{\alpha_c}{2} \|\mathbf{X}\|_{D,c}^2 \quad (4)$$

The factorized formulation [9, 13] of equation (4) is

$$\min_{\mathbf{W}, \mathbf{H}} \frac{1}{2} \|\mathbf{W}\|_{D,r}^2 + \frac{1}{2} \|\mathbf{H}\|_{D,c}^2 + \frac{\gamma}{2} \|\Omega \circ (\mathbf{Y} - \mathbf{W}\mathbf{H}^T)\|_F^2 \quad (5)$$

2.4 Geometric Matrix Completion with Separable Recurrent Graph Neural Networks

In [9], Monti et al. propose to solve the matrix completion problem as a learnable diffusion process using Graph Convolutional Neural Networks (GCNN) and Recurrent Neural Networks (RNN). The main idea here is to use GCNN to extract features from the matrix and then use LSTMs to learn the diffusion process. They argue that combining these two methods allows the network to predict accurate small changes \mathbf{dX} (or \mathbf{dW} , \mathbf{dH} of the matrices \mathbf{W} , \mathbf{H}) to the matrix \mathbf{X} . Further details regarding the main ideas in geometric deep learning have been summarized in a review paper [1], where they elaborate how to extend convolutional neural networks to graphs. Following [9], we use Chebyshev polynomial basis on the factorized form of the matrix $\mathbf{X} = \mathbf{WH}^T$ to represent the filters on the respective graph to each matrix \mathbf{W} and \mathbf{H} . In this work, we only apply GCNN to the matrix \mathbf{W} as we only have a row graph and leave the matrix \mathbf{H} as a changeable variable. Figure 1 illustrates the overall approach.

2.5 Geometric Matrix Completion for Heterogeneous Matrix Entries

In this work, we propose to solve multi-modal disease classification as a geometric matrix completion problem. We use a Separable Recurrent GCNN (sRGCNN) [9] to simultaneously predict the disease and impute missing features on a dataset which has partially observed features and labels. Following Goldberg et al. [5], we stack a feature matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ and a label matrix $\mathbf{T} \in \mathbb{R}^{m \times c}$ as a matrix $\mathbf{Z} \in \mathbb{R}^{m \times n+c}$, where m is the number of subjects, n is the dimension of the feature matrix, and c is the dimension of the target values. In the TADPOLE dataset, we stack the $m \times n$ feature matrix to the $m \times 1$ label matrix, where the feature matrix contains all the numerical features and the label matrix contains the encoded binary class labels for cMCI and sMCI. We build the graph by using meta-information from the patients such as their age and gender, similar to [12], as these information are known to be risk factors for AD. We compare two row graph construction approaches, first from age and gender information using a similarity metric [12] and second from age information only, using Euclidian distance-based k-nearest neighbors. Every node in a graph corresponds to a row in the matrix \mathbf{W} , and the row values to its associated feature vector. Since we only have a row graph, we leave the matrix \mathbf{H} to be updated during backpropagation. To run the geometric matrix completion method we use the loss:

$$\ell(\Theta) = \frac{\gamma_a}{2} \|\mathbf{W}\|_{D,r}^2 + \frac{\gamma_b}{2} \|\mathbf{W}\|_F^2 + \frac{\gamma_c}{2} \|\mathbf{H}\|_F^2 + \frac{\gamma_d}{2} \|\Omega_a \circ (\mathbf{Z} - \mathbf{WH}^T)\|_F^2 + \gamma_e (\ell_{\Omega_b}(\mathbf{Z}, \mathbf{X})), \quad (6)$$

where Θ are the learnable parameters, where \mathbf{Z} denotes the target matrix, \mathbf{X} is the approximated matrix, $\|\cdot\|_{D,r}^2$ denotes the Dirichlet norm on a normalized row graph Laplacian, Ω_a denotes the masking on numerical features, Ω_b is the masking on the classification labels, and ℓ is the binary cross-entropy.

3 Results

We evaluate our approach on multi-modal TADPOLE data (MRI, PET, CSF, DTI) to predict MCI-to-AD conversion and compare it to several other multi-modal methods as baseline. We use a stratified 10-fold cross-validation strategy for all methods. Hyperparameters were optimized using Hyperopt⁴, through nested cross-validation, targeting classification loss (binary cross-entropy) on a hold-out validation set (10% in each fold of training data). Following [9], we use the same sRGCNN architecture with parameters: rank=156, chebyshev polynomial order=18, learning rate=0.00089, hidden-units=36, $\gamma_a=563.39$, $\gamma_b=248.91$, $\gamma_c=688.85$, $\gamma_d=97.63$, and $\gamma_e=890.14$.

It is noteworthy that at baseline, the data matrix Y with above-mentioned features is already feature-incomplete, i.e. only 53% filled. We additionally reduce the amount of available data randomly to 40%, 30% etc. to 5%. Figure 2 shows a comprehensive summary of our classification results in terms of area-under-the-curve (AUC). Methods we compare include mean imputation with random forest (RF), linear SVM (SVC) and multi-layer-perceptron (MLP), as well as three reference methods from literature [10, 12, 15], which operated on slightly different selections of ADNI subjects and on all available multi-modal features. While implementations of [10, 15] are not publicly available, we tried to re-evaluate the method [12] using their published code. Unfortunately, despite our best efforts and hyperparameter optimization on our selection of TADPOLE data, we were not able to reproduce any AUC value close to their published value. To avoid any mistake on our side, we provide the reported AUC results rather than the worse results from our own experiments.

⁴ <http://hyperopt.github.io/hyperopt/>

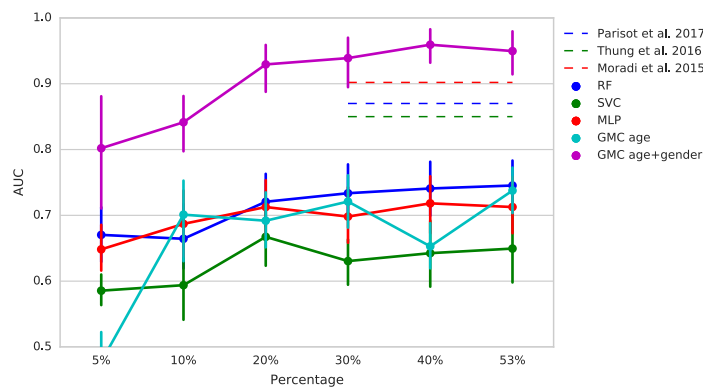


Fig. 2. Classification results: Area under the curve (AUC) of our method, for different amounts of feature-completeness and in comparison to linear/non-linear standard methods, and three state-of-the-art results in literature (Parisot et al. [12], Thung et al. [15], Moradi et al. [10]).

At baseline, our best-performing method with a graph setup based on age and gender ("GMC age-gender") [12] achieves classification with an AUC value of 0.950, compared to 0.902 [10], ~ 0.87 [12] and 0.851 [15]. In terms of classification accuracy, we achieved a value of 87%, compared to 82% [10] and 77% [12] (accuracy not reported in [15]). Furthermore, our method significantly outperforms standard classifiers RF, MLP and SVC at all levels of matrix completeness. The second graph configuration for our method ("GMC age" only) performs significantly worse and less stable than ("GMC age-gender"), confirming the usefulness of the row graph construction based on the subject-to-subject similarity measure proposed in [12]. Due to lower complexity of the GMC approach [9], training a single fold on recent hardware (Tensorflow on Nvidia GTX 1080 Ti) is on average 2x faster (11.8s) than GCN (25.9s) [12].

4 Discussion and Conclusion

In this paper, we proposed to view disease classification in multi-modal but incomplete clinical datasets as a geometric matrix completion problem. As an exemplary dataset and classification problem, we chose MCI-to-AD prediction. Our initial results using this method show that GMC outperforms three competitive results from recent literature in terms of AUC and accuracy. At all levels of additional random dropout of features, GMC also outperforms standard imputation and classifiers (linear and non-linear). There are several limitations which are worthy to be addressed. Results in Figure 2 demonstrate that GMC is still sensitive to increasing amounts of feature incompleteness, in particular at feature presence below 15%. This may be due to our primary objective of disease classification during hyper-parameter optimization. For the same reason, we did not evaluate the actual imputation performed by GMC. However, an evaluation in terms of RMSE and a comparison to principled imputation methods [4] would be highly interesting, if this loss is somehow incorporated during hyperparameter optimization. Furthermore, we only evaluated GMC on ADNI data as represented in the TADPOLE challenge, due to the availability of multiple reference AUC/accuracy values in literature. As mentioned, however, disease classification in high-dimensional but incomplete datasets with multiple modalities is an abundant problem in computer-aided medical diagnosis. In this light, we believe that the promising results obtained through GMC in this study are of high interest to the community.

Acknowledgments

The study was supported by the German Federal Ministry of Education and Health (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) (grant number 01 EO 0901).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Bibliography

- [1] Michael M. Bronstein, Joan Bruna, Yann Lecun, Arthur Szlam, and Pierre Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- [2] E.J. Candes and B. Recht. Exact low-rank matrix completion via convex optimization. *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 1–49, 2008.
- [3] Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [4] Y. Dong and C. Y. Peng. Principled missing data methods for researchers. *Springerplus*, 2(1):222, Dec 2013.
- [5] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems (NIPS)*, pages 757–765. 2010.
- [6] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. *arXiv:1408.1717*, 2014.
- [7] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, arXiv:1609.02907, 2016.
- [8] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Snchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.
- [9] Federico Monti, Michael M. Bronstein, and Xavier Bresson. Geometric matrix completion with recurrent multi-graph neural networks. *CoRR*, arXiv:1704.06803, 2017.
- [10] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, and Jussi Tohka. Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects. *NeuroImage*, 104:398–412, 2015.
- [11] Kenichi Oishi, Kazi Akhter, Michelle Mielke, Can Ceritoglu, Jiangyang Zhang, Hangyi Jiang, Xin Li, Laurent Younes, Michael Miller, Peter van Zijl, Marilyn Albert, Constantine Lyketsos, and Susumu Mori. Multi-Modal MRI Analysis with Disease-Specific Spatial Filtering: Initial Testing to Predict Mild Cognitive Impairment Patients Who Convert to Alzheimers Disease. *Frontiers in Neurology*, 2:54, 2011.
- [12] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero Moreno, Ben Glocker, and Daniel Rueckert. Spectral graph convolutions for population-based disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 177–185, 2017.

- [13] Nikhil Rao, Hsiang-Fu Yu, Pradeep Ravikumar, and Inderjit S Dhillon. Collaborative Filtering with Graph Information: Consistency and Scalable Methods. *Neural Information Processing Systems (NIPS)*, pages 1–9, 2015.
- [14] Nathan Srebro, Jason D M Rennie, and Tommi S Jaakkola. Maximum-Margin Matrix Factorization. *Advances in Neural Information Processing Systems (NIPS)*, 17:1329–1336, 2005.
- [15] Kim-Han Thung, Ehsan Adeli, Pew-Thian Yap, and Dinggang Shen. Stability-weighted matrix completion of incomplete multi-modal data for disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–96, 2016.
- [16] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen. Multimodal classification of alzheimer’s disease and mild cognitive impairment. *NeuroImage*, 55(3):856 – 867, 2011.

3.3.2 Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification. (AI in Medicine 2021)

Following our previous GMC approach, we improved GMC to be more robust in terms of classification and imputation. We also included an additional evaluation of the imputation results. We propose Multigraph Geometric Matrix Completion (MGMC) for disease classification in clinical datasets with missing information. Previous approaches made use of single graph and/or auto-regressive Recurrent Graph Neural Networks (RGCN), meaning that the output from an LSTM cell is fed directly to the next LSTM cell block. Instead, our proposed MGMC approach uses a multi-graph approach as well as non-autoregressive RGCN. We showed the effectiveness and superiority of this approach including the use of a self-attention mechanism to weight which information should be given more weight in order to yield better disease classification results.

Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification (Artificial Intelligence in Medicine 2021)

Gerome Vivar^{1,2}, Anees Kazi¹, Hendrik Burwinkel¹, Andreas Zwergal², Nassir Navab^{1,3}, Seyed-Ahmad Ahmadi^{1,2} for the Parkinson's Progression Markers and Alzheimer's Disease Neuroimaging Initiatives

¹Technische Universität München (TUM), Munich, Germany

²German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-Universität (LMU), Munich, Germany

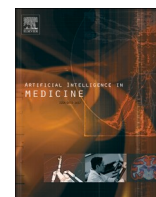
³Johns Hopkins University, Baltimore, USA.

Copyright Statement. © 2021 Elsevier B.V. Reprinted, with permission, from Gerome Vivar, Anees Kazi, Hendrik Burwinkel, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi, Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification, May 2021. Original publication can be found at: (<https://doi.org/10.1016/j.artmed.2021.102097>)

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification[☆]

Gerome Vivar^{a,b}, Anees Kazi^a, Hendrik Burwinkel^a, Andreas Zwergal^b, Nassir Navab^a, Seyed-Ahmad Ahmadi^{a,b,*}, for the Parkinson's Progression Markers and Alzheimer's Disease Neuroimaging Initiatives¹

^a Department of Computer Aided Medical Procedures (CAMP), Technical University of Munich (TUM), Boltzmannstr. 3, 85748 Garching, Germany

^b German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians University (LMU), Fraunhoferstr. 20, 82152, Planegg, Germany

ARTICLE INFO

Keywords:

Computer-aided diagnosis
CADx
Deep learning
Multimodal medical data
Population-based studies

ABSTRACT

Large-scale population-based studies in medicine are a key resource towards better diagnosis, monitoring, and treatment of diseases. They also serve as enablers of clinical decision support systems, in particular computer-aided diagnosis (CADx) using machine learning (ML). Numerous ML approaches for CADx have been proposed in literature. However, these approaches assume feature-complete data, which is often not the case in clinical data. To account for missing data, incomplete data samples are either removed or imputed, which could lead to data bias and may negatively affect classification performance. As a solution, we propose an end-to-end learning of imputation and disease prediction of incomplete medical datasets via Multi-graph Geometric Matrix Completion (MGMC). MGMC uses multiple recurrent graph convolutional networks, where each graph represents an independent population model based on a key clinical meta-feature like age, sex, or cognitive function. Graph signal aggregation from local patient neighborhoods, combined with multi-graph signal fusion via self-attention, has a regularizing effect on both matrix reconstruction and classification performance. Our proposed approach is able to impute class relevant features as well as perform accurate and robust classification on two publicly available medical datasets. We empirically show the superiority of our proposed approach in terms of classification and imputation performance when compared with state-of-the-art approaches. MGMC enables disease prediction in multimodal and incomplete medical datasets. These findings could serve as baseline for future CADx approaches which utilize incomplete datasets.

1. Introduction

Large population-based studies in medicine, acquired at multiple institutions, are instrumental resources for a better clinical understanding of the diagnosis, progression and treatment of diseases. In

medical health informatics, they serve as fundamental enablers for the design and analysis of novel clinical decision support systems (CDSS) and CADx [1]. Often, such datasets incorporate multimodal data, both imaging and non-imaging, in order to capture as many aspects of the disease as possible.

^{*} This work was supported by the German Federal Ministry of Education and Health (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) [grant 01 EO 0901], with partial support of "Freunde und Förderer der Augenklinik München", Germany.

^{*} Corresponding author at: German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-University Munich, Fraunhoferstr. 20, 82152, Planegg, Germany.

E-mail address: ahmadi@cs.tum.edu (S.-A. Ahmadi).

¹ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Further data were obtained from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/data. List of all PPMI funding partners can be found at www.ppmi-info.org/fundingpartners.

<https://doi.org/10.1016/j.artmed.2021.102097>

Received 16 November 2020; Received in revised form 4 May 2021; Accepted 5 May 2021

Available online 8 May 2021

0933-3657/© 2021 Elsevier B.V. All rights reserved.

Two prominent examples for such datasets in neurology and neuroscience were published by the Alzheimer’s Disease (AD) Neuroimaging Initiative (ADNI) [2] and the Parkinson’s disease (PD) Progressive Marker Initiative (PPMI) [3]. Together, AD and PD are the most common neurodegenerative diseases, with AD accounting for 60–80% of dementia cases, and PD affecting 1–2% of the global population over the age of 65. Neurodegenerative diseases result in a progressive decay and death of nerve cells [4]. Increasing rates of up to a million new AD cases per year [4], along with the prospect of novel models and care frameworks for dementia [5] as well as novel neuroprotective and disease-modifying therapeutics, in both AD and PD [6], motivate an early diagnosis of these diseases, ideally already at a pre-symptomatic stage.

Population-based datasets in medicine are often feature-incomplete, due to missing examinations of patients. Most ML-based CADx approaches require imputation before classification [7], and treat these steps sequentially and independently. Incomplete features are categorized into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), with MAR often lying at the basis of most modern imputation methods [8].

Related works: A recent review paper [9] on ML techniques for AD diagnosis has found that most recent methods treated multimodal feature modeling and classification separately, with a focus on the former. In addition, they suggested that more work is required in multimodal ML methods towards early AD diagnosis. In line with these findings, our proposed method addresses multimodal CADx for AD, with simultaneous feature imputation and classification.

Two commonly used methods to treat missing values in datasets are sample deletion or Mean-imputation, which either result in expensive loss of data or in biased and sub-optimal features. More advanced methods use multiple imputation or ML. Hedge et al. [7] compared Multiple Imputation Using Chained Equations (MICE) [10] with Probabilistic Principal Component Analysis (PPCA) on healthcare data, and found PPCA to be superior. A fundamentally different approach is matrix completion. Thung et al. [11,12] use Low-Rank Matrix Completion (LRMC) to predict conversion of the disease in patients with Mild Cognitive Impairment (MCI) to Alzheimer’s Disease (AD). Zhou et al. [13] proposed to solve AD diagnosis using latent representation learning, by projecting both complete and incomplete modalities onto a common subspace. Both approaches by [12] and [13] assume a linear relationship between the input features and the target variable, and latent embeddings and linear classification are trained in two separate steps [13], which does not take advantage of end-to-end learning.

Recently, graph convolutional networks (GCN) have been introduced for CADx on multimodal medical datasets. Parisot et al. [14] introduced a novel concept for modeling patient populations as a graph: patient meta-data like demographics (e.g. sex, age, etc.) are used to compute similarities between patients, leading to an adjacency matrix with an associated graph Laplacian. Intuitively, the graph is akin to a “social network” of patients in the cohort. Several works since then have demonstrated that GCNs can significantly improve the accuracy of CADx in medicine [15–19]. Importantly, the graph definition crucially affects the CADx accuracy, and we have shown previously that parallel multi-graph models with attention, i.e. one graph for each meta-feature, can make GCNs more robust [16,17].

Importantly, like most other ML methods, GCNs assume feature-completeness and depend on imputation as a pre-processing step. Regarding incomplete datasets, Monti et al. [20] showed that geometric deep learning provides a principled framework for non-linear imputation, through geometric matrix completion (GMC). In our own previous work [21], we extended upon this work through multi-target training, which combined GMC with supervised classification, into a Recurrent Graph Convolutional Network (RGCN). Similar to [15], we constructed a patient graph from clinical meta-data (e.g. age and sex of patients). We concatenated the incomplete feature matrix and incomplete labels, and trained a GCN for signal diffusion, along with a Long-Short Term

Memory (LSTM) network for iterative matrix reconstruction. Both GCN and LSTM were combined into a single-graph RGCN, which was trained end-to-end towards MCI to AD conversion prediction, with two weighted losses for simultaneous classification and imputation.

Contribution: We propose to solve disease classification in multimodal and incomplete datasets using Multi-graph Geometric Matrix Completion (MGMC). The contributions of this work are threefold: (1) we formulate the disease classification problem in multimodal and incomplete datasets using MGMC; (2) we propose a novel method which uses multiple non-autoregressive Recurrent Graph Convolutional Networks (RGCN) and a transformer-inspired self-attention mechanism for multi-graph fusion; (3) we validate the superiority of the proposed approach on two publicly available medical datasets and evaluate the effect of autoregressive LSTMs on MGMC architectures.

2. Materials and methods

We first introduce the notation used throughout the rest of the paper in Table 1, then elaborate on key background information in order to provide more context on our proposed approach.

2.1. Dataset and preprocessing

We used two publicly available datasets in this work: The Alzheimer’s Disease Prediction Of Longitudinal Evolution (TADPOLE) [2] obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Parkinson’s Progressive Marker Initiative (PPMI) dataset [3]. TADPOLE requires classification of subjects into three categories, normal control (NC), mild cognitive impairment (MCI), and Alzheimer’s disease (AD). PPMI requires detection of Parkinson’s disease (PD) vs. normal controls (NC).

In TADPOLE, we used 813 subjects coming from the ADNI protocol with 229 NC, 396 MCI and 188 AD diagnosed at baseline. This dataset contains pre-processed features [2] from cerebro-spinal fluid (CSF) markers, magnetic resonance imaging (MRI), positron emission tomography FDG (PET), diffusion tensor imaging (DTI), cognitive assessment scores, genetic information such as alipoprotein E4 (APOE4), and demographic information. Further pre-processing entailed a normalization of real-valued TADPOLE features to zero-mean and unit-variance. To match the classification task, we selected only features at baseline, and excluded features containing longitudinal information. We further

Table 1
Description of notations.

Notation	Dimension	Description
X	$n \times m$	Observed feature matrix with n samples and m features
Y	$n \times c$	Class label matrix with n samples and c number of class
Z	$n \times (m + c)$	Concatenated X and Y matrices
\hat{X}	$n \times m$	Predicted feature matrix X
\hat{Z}	$n \times (m + c)$	Predicted matrix Z
\bar{Z}	$n \times (m + c)$	Predicted matrix Z from a single RGCN
$\ \cdot\ _F^2$	–	Frobenius norm
$\ \cdot\ _0$	–	Dirichlet norm
$\mathcal{L}_{ce}(\cdot)$	–	Cross-entropy loss
$\mathcal{L}_R(\cdot)$	–	Reconstruction loss from GMC
$M^{(i)}$	–	The i th meta-information
M	–	Set containing $\{M^{(1)}, \dots, M^{(l)}\}$
G_i	–	The i th graph constructed using meta-information $M^{(i)}$
Ω_x, Ω_y	–	Denote whether input features and class labels, respectively, are known (1) or missing (0)
Θ, δ	–	Parameters from GCN and LSTM, respectively
$\gamma_{(a,b,c)}$	–	Hyper-parameters weighting loss terms
\circ	–	Hadamard product

removed features that were available for less than 10% of the available entries. In the end, the feature matrix had a dimensionality of 813×435 , excluding label information.

In the PPMI dataset, we used all 75 healthy controls (HC) and 249 subjects with PD. PPMI data consists of brain MRI as well as non-imaging information such as Unified Parkinson's Disease Rating Scale (UPDRS), Montreal Cognitive Assessment (MoCA) scores, and demographic information (age and gender). The MRI information is used as input to the network while non-imaging information is used for the graph construction. As described in our previous GCN CADx approach [17], we pre-processed MRI volumes by co-registering each images to a normative space (SRI24 atlas [22]) to reduce variability in appearance, and further performed skull stripping using ROBEX [23]. Then we scaled each volume to an intensity range of [0,1]. Finally, to obtain a lower dimensional representation as input to the graph network, we used encoded raw image intensities coming from a 3D-autoencoder, which was pretrained towards anomaly detection. We refer the reader to [24] for a detailed discussion on the implementation of the pre-processing and 3D-autoencoder. The output at the bottleneck layer of the 3D-autoencoder was then used as the feature representation of the brain MRI volume.

Notably, our pre-processed PPMI dataset was 100% feature complete. In contrast, the TADPOLE dataset is inherently incomplete in native form, and was 83% feature-complete after our pre-processing pipeline. In the experimental section, we further removed known features artificially, to test classification and imputation robustness at various levels of data missingness. For better clarity throughout the rest of the paper, when denoting e.g. 50% data availability, we refer to the amount of data available at baseline (e.g. 50% for PPMI, and 41.5% for TADPOLE).

2.2. Graph construction

We use meta-information to construct separate graphs for each dataset. In the TADPOLE dataset, we use meta-information such as age, gender, and genetic risk factor (APOE4), all of which are known risk factors related to AD. For every given meta-information feature M , we calculate a separate graph using a pairwise similarity function. An edge between nodes i and j is defined using $W(i, j) = f(M(i), M(j))$ where

$$f(M(i), M(j)) = \begin{cases} 1 & \text{if } |M(i) - M(j)| \leq \theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$M(i)$ and $M(j)$ denote meta information of node i and j of a given meta-information M , and θ denotes a threshold value which is chosen empirically by the user, given domain expertise and depending on what can be regarded as a similar trait across patients [14,15].

To construct the graphs for the PPMI dataset, we use the same formulation in Eq. (1) and build graphs for every meta-information. Here we again use age and gender, along with two PD-related clinical scores of motor function (UPDRS) and of cognitive function (MoCA) to build the graph, following [17].

2.3. Geometric matrix completion

Consider an incomplete feature matrix $X \in \mathbb{R}^{n \times m}$ where a certain proportion of values is missing at random. The goal is to recover the missing values in this matrix. One solution to this problem is by using rank minimization. However, as this is known to be computationally intractable, an alternative approximation is to constrain the predicted values to be smooth with respect to some geometric structure [25,26,20]. Here a graph structure is built based on the rows or columns of the matrix. Monti et al. [20] proposed to solve this using geometric deep learning on graphs, through a combination of GCN and LSTM networks. Compared to GMC recommender systems in [20], our CADx problem does not allow us to build a semantically meaningful column graph,

especially since features stem from different modalities. Therefore, we modify the GMC approach to consider only a row graph derived from patient similarities to model the population. Nodes within a graph are the patient instances, their corresponding row vectors are the nodes' feature vectors, and the graph edges are based on patient similarities which are computed from meta-features, according to the metric in Eq. (1). Pair-wise similarities between nodes in the population graph connect patients that share the same risk-factor characteristics. The graph is then represented as $G = (V, E, W)$, with vertices $V = \{1, 2, \dots, n\}$, and edges $E \subseteq V \times V$, which are weighted with non-negative weights. We represent the graph with a symmetric adjacency matrix $W \in \mathbb{R}^{n \times n}$. The geometric matrix completion problem reduces to minimizing the loss:

$$\ell(\Theta, \delta) = \|\widehat{X}_{\Theta, \delta}\|_D^2 + \frac{\gamma}{2} \|\Omega_x \circ (\widehat{X}_{\Theta, \delta} - X)\|_F^2 \quad (2)$$

where $\widehat{X}_{\Theta, \delta}$ is the predicted matrix conditioned on the parameters of the GCN and LSTM, and \circ denotes the Hadamard product. In Eq. (2), the first term on the right can be expressed as $\text{tr}(\widehat{X}^T L \widehat{X})$ [27] which contains a rescaled graph Laplacian ($L \in \mathbb{R}^{n \times n}$) term such that its eigenvalues are in the interval $[-1, 1]$. This term keeps the prediction smooth with respect to the graph structure.

GMC can also be extended to multi-target training on heterogeneous matrix entries. Consider a matrix $Z \in \mathbb{R}^{n \times (m+c)}$, which contains a mixture of feature and label information, which is implemented by concatenation of the feature matrix $X \in \mathbb{R}^{n \times m}$ and class label matrix $Y \in \mathbb{R}^{n \times c}$, similarly to Goldberg et al. [28]. Following Eq. (2), we can add a classification loss term on the imputed class label matrix [21]. The combined loss for completion of matrix Z is then:

$$\ell(\Theta, \delta) = \frac{\gamma_a}{2} \|\widehat{Z}_{\Theta, \delta}\|_D^2 + \frac{\gamma_b}{2} \|\Omega_x \circ (\widehat{Z}_{\Theta, \delta} - Z)\|_F^2 + \gamma_c (\mathcal{L}_{cc}(\widehat{Z}_{\Theta, \delta} \circ \Omega_y, Z \circ \Omega_y)) \quad (3)$$

where $\widehat{Z}_{\Theta, \delta}$ is the predicted matrix containing predictions for both \widehat{X} and \widehat{Y} .

2.4. Multigraph Geometric Matrix Completion

MGMC² consists of multiple non-autoregressive RGCNs and Transformer-like self-attention. We first describe the motivation why we use multiple RGCNs then elaborate on the self-attention inspired aggregation scheme including the use of non-autoregressive RGCNs. First, as we described in our previous works [16,17], the rules for constructing a population graph from a medical dataset are crucial to the accuracy of a GCN's downstream task, e.g. diagnostic classification accuracy. Instead of collapsing all meta-features into a single patient similarity measure, we therefore construct multiple graphs, one for each meta-feature. We then propose to integrate multi-graph GCNs into matrix completion by training a dedicated GCN and LSTM for each graph in an end-to-end manner. We do this to learn better imputed feature representations for each graph which could be useful in the downstream classification task.

To aggregate separate signals from parallel RGCNs, we use a self-attention aggregation mechanism inspired by Transformer networks called Scaled Dot-Product Attention [29]. We do this by training separate RGCNs (which consists of GCN and LSTM) in an end-to-end manner as shown in Fig. 1, then aggregate every RGCN outputs using the weights learned from the self-attention layer. We calculate self-attention weights for every RGCN by first stacking the outputs of RGCN ($\widehat{Z}_{\Theta, \delta}^{(i)}$) into a tensor of size $(B \times M \times F)$ where B is the full-batch-size, M denotes number of RGCNs, and F the dimensionality of the RGCN output. Weights for every

² Code: <https://github.com/pydgsz/MGMC>.

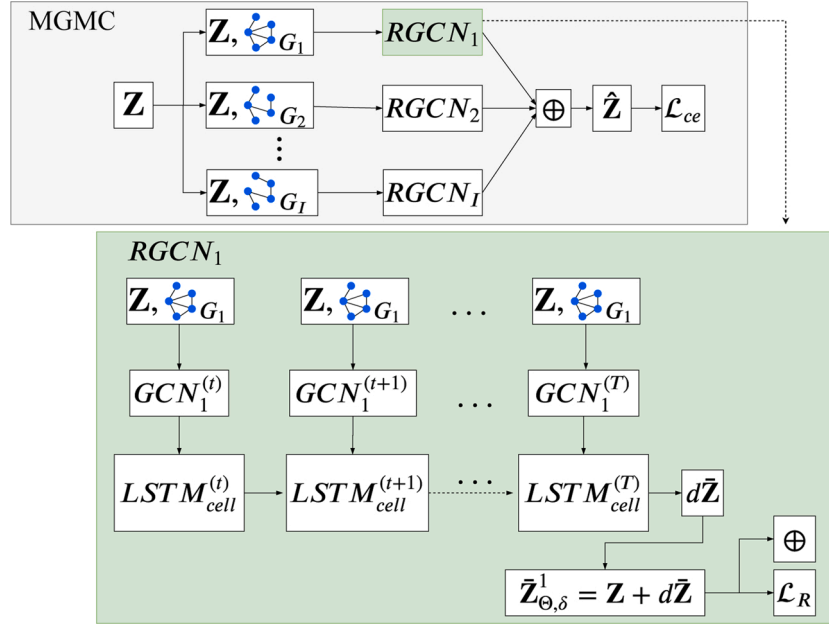


Fig. 1. Network architecture of MGMC which uses multiple Recurrent Graph Convolutional Network (RGCN) (top) including non-autoregressive RGCN layer (bottom). Information from a single RGCN branch will be aggregated (\oplus) together with the other outputs from other RGCN branches using a Scaled Dot-Product Attention mechanism. This output from a single RGCN is also used to calculate the reconstruction loss \mathcal{L}_R , which is the first term of the right-hand side of Eq. (5).

graph output are then calculated using Scaled-Dot-Product Attention [29]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q, K and V are the linearly transformed outputs after stacking using learnable weight matrices (W_Q, W_K, W_V). In the end, this self-attention aggregation mechanism (denoted as \oplus in Fig. 1) for outputs of every RGCN will yield an output \hat{Z} . Furthermore, we use multiple RGCNs, wherein each (unrolled) RGCN consists of a GCN and a non-autoregressive LSTM. Although multiple graphs and LSTMs have been used in previous methods ([20,19]), one important difference of our proposed approach is the use of non-autoregressive LSTMs. As shown in Fig. 1, we only use the original input feature as input to the next timestep including the learned parameters from the previous LSTM cell-block. Such a non-autoregressive strategy is motivated in several ways. First, it limits the number of neighborhood hops and graph signal diffusion steps, as the input feature matrix to the GCN layer is the same at every time-step in the RGCN. Second, it allows the model to have better control on which graph-relevant information is useful for the imputation and downstream classification task. Third, by using the original input features as prior information at every optimization step, we reinforce the reconstruction of the input data, and prevent the model from diverging from the input data. As a result, this strategy prevents the model from suggesting non-realistic features as outputs. For the GCN layers, we use a Cheb-Net implementation [30,20]. This uses a Chebyshev polynomial basis ($\sum_{k=0}^K T_k(\tilde{L})X\Theta_k$) to represent the spectral filters. For a more in-depth discussion regarding deep learning on graphs we refer the reader to [27]. The optimization loss for multi-graph GMC then boils down to minimizing the loss:

$$\ell(\Theta, \delta) = \sum_i^M \left(\frac{\gamma_a}{2} \|\bar{Z}_{\Theta, \delta}^{(i)}\|_{D, r}^2 + \frac{\gamma_b}{2} \|\Omega_{\Theta, \delta}(\bar{Z}_{\Theta, \delta}^{(i)} - Z)\|_F^2 \right) + \gamma_c (\mathcal{L}_{ce}(\hat{Z}_{\Theta, \delta}, Z; \Omega_{\gamma})) \quad (5)$$

where $\bar{Z}_{\Theta, \delta}^{(i)}$ is the i th predicted matrix from the i th graph (noting that this is conditioned on the parameters of the i th GCN and LSTM) and $\hat{Z}_{\Theta, \delta}^{(i)}$ is

the aggregated predicted matrix coming from all GCNs and LSTMs.

3. Results

3.1. Implementation details

We used a 10-fold stratified cross-validation strategy to split the dataset into 10% test and 90% train (of which 10% as validation set) on all methods. For all deep learning based methods we use Adam optimization [31], with implementations in PyTorch [32], on a workstation with a single GPU (Nvidia GTX 1080 Ti). We automatically determine hyperparameters in Eq. (5) using hyperparameter optimization on the validation set with 120 iterations [33], with the following search spaces for the Chebyshev Polynomial parameters ($K \in \text{range}(1, 20)$), learning rate = $\text{uniform}([0.00001, 0.1])$, intermediate layers' hidden units $\in \text{range}(8, 512)$, and $\gamma_{(a,b,c)} = \text{uniform}([0.001, 1000])$.

We compared the proposed method with shallow learning methods in machine learning, gradient-based Matrix Completion (MC), and state-of-the-art (SOTA) graph-based methods which have shown to be highly effective for disease prediction. For shallow learning, we used Logistic Regression (LR) as the linear baseline, and Random Forest (RF) [34] as a competitive non-linear baseline. We also compared against MC which is a simple non-graph-based gradient based matrix completion approach. Previous graph-based methods included GCNs [14,15], GMC [21], and MG-RGCN [19]. As several algorithms (LR, RF and GCN) assume feature-completeness, we first need to impute the missing values in the feature matrix. We used five approaches to accomplish this: the commonly used Mean-imputation method, kNN imputation [35], MICE with linear regression (MICE_LR) [10], MICE with random regression forest (MICE_RF) [10], and PPCA [36]. For GCN, we use the empirically best-performing imputer. To test imputation performance, we artificially reduce the percentage of known data in the ADNI/PPMI feature matrices and perform imputation/classification at {100,75,50,25}% data availability (MAR assumption [8]). At each percentage level, we report the worst and the best performance for each imputer + classifier combination, to give an indication of the spread of possible outcomes. To report and compare classification outcomes, we visualize the three metrics Accuracy/F-measure/ROC-AUC in Fig. 2, and compare them

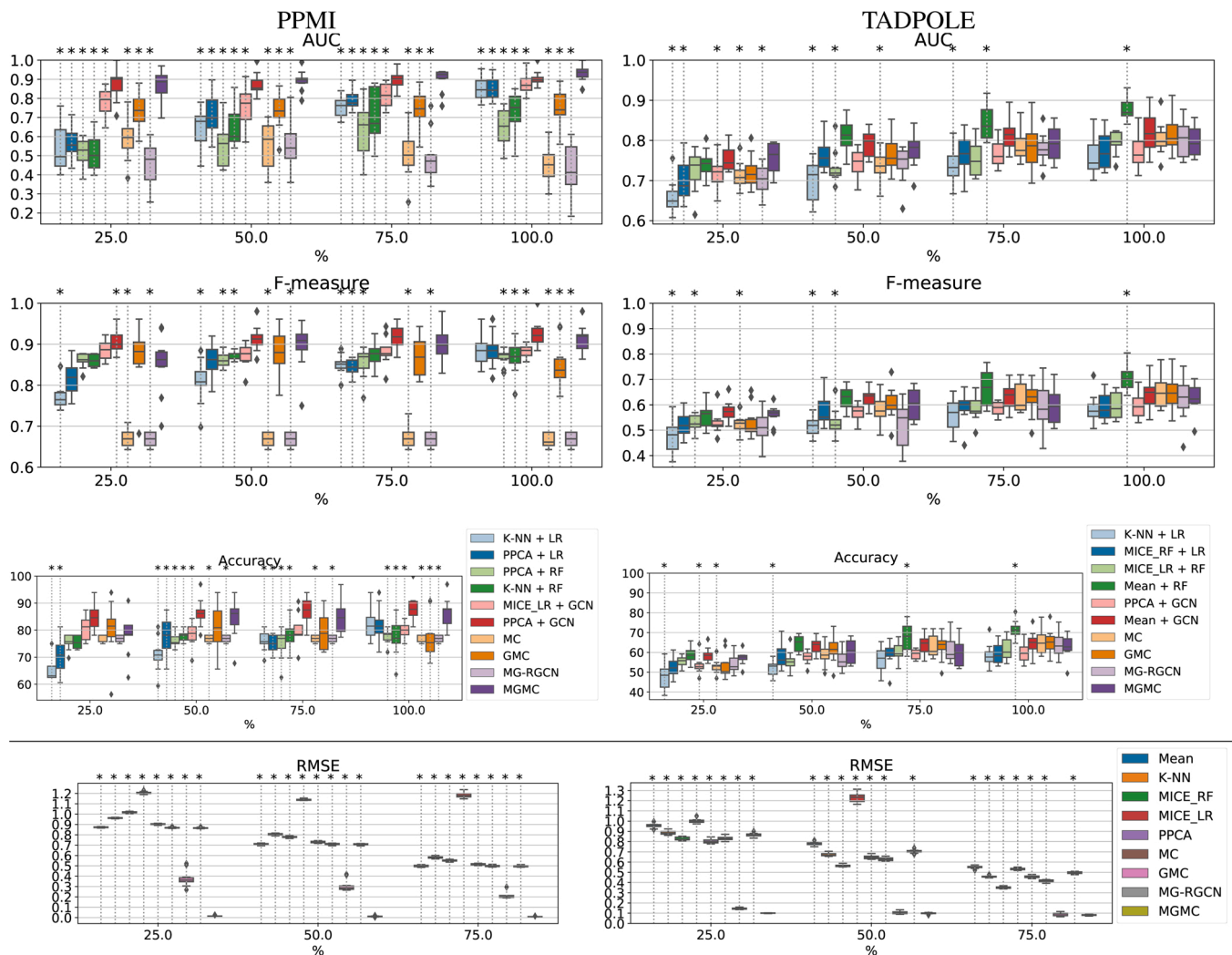


Fig. 2. PPMI (left panel) and TADPOLE (right panel) classification results (boxplots indicate the distribution of metrics over the 10 folds for each model): ROC-AUC (first-row), F -measure (second-row) and Accuracy (third-row). Imputation results (fourth-row) for PPMI and TADPOLE. Asterisk symbols (*) and dotted vertical lines denote that the tested model is statistically significantly different (two-tailed Wilcoxon rank-sum test, $p \leq 0.05$) to our proposed model (MGMC). X -axis values denote the percentage of available/known features prior to imputation and model training. (Best viewed in digital format).

quantitatively via a two-tailed Wilcoxon rank-sum hypothesis test, at an alpha-level of $p \leq 0.05$.

We use Scikit-learn [37] implementations for cross-validation, pre-processing, imputation (PPCA [38]) and shallow classifier models (LR and RF). To make baseline algorithms as competitive as possible, we also perform hyperparameter optimization (also 120 max. iterations) for the standard machine learning models (i.e. LR and RF) [39]. We concatenate the meta-features (e.g. demographics) with the feature vectors for all baseline methods, to further ensure fairness, as our proposed graph-based method utilizes this information as well (i.e. for graph construction).

3.2. PPMI and TADPOLE dataset results

We plot classification (Fig. 2 rows 1–3) and imputation (Fig. 2 row 4) results on the PPMI dataset (Fig. 2 left panels) and on the TADPOLE dataset (Fig. 2 right panels).

In PPMI, the classification metrics show that our proposed MGMC method is consistently among the top-performing methods. In terms of ROC-AUC and Accuracy (cf. Fig. 2, rows 1 and 3), MGMC is often significantly better than other classifiers, at all levels of data availability. In the following, we will describe our results by focusing mainly on the aggregate metric F -measure, as it reflects the harmonic mean between

precision and recall and is therefore better suitable to assess the classification of rare positives (as it is often desired in medicine). In Fig. 2 (middle-left panel), we can see that the average F -measure over the 10 folds for MGMC stays consistently high at 0.852/0.897/0.904/0.913 (25/50/75/100% data availability, respectively). The only other method that performs comparably high is another graph deep learning method, GCN with PPCA-imputation. The difference is significant at 25% data availability (0.905, $p < 0.05$), but not at the other levels (0.914/0.918/0.926, $p > 0.05$). It is important to note that all algorithms that require prior imputation have a noticeable difference of performance, given the same amount of available data. For example, LR combined with kNN performs on average lower than when combined with PPCA, especially at 25% of data (F -measure difference: 0.044) and 50% of data (F -measure difference: 0.053). The best vs. worst imputation combination of imputer + classifier is not consistent across models: for RF, PPCA is on average worst, kNN is best, while for GCN, MICE_LR is worst and PPCA is best. Compared to our previously proposed GMC method, MGMC performs significantly better at 100% data availability (0.913 vs. 0.850, $p < 0.05$), not significantly better on average (not significant, $p > 0.05$) at 50% (0.896 vs. 0.881) and 75% (0.904 vs. 0.872) data availability, and not significantly worse at 25% (0.852 vs. 0.870, $p > 0.05$). Another striking result in PPMI is that the matrix completion methods MC and MG-RGCN more or less failed to learn a

good classification, at all levels of data availability (F -measure < 0.7). The implications of this low performance will be discussed in Section 4.

In TADPOLE, compared to PPMI, the classification accuracy does not benefit as clearly from the population graph or imputation in our method. Similar to PPMI, the metrics ROC-AUC and Accuracy show some cases where MGMC is significantly better than other methods, notably at 25% and 50% data availability, and compared to LR or algorithms that are matched with the worst imputation method. For a further analysis, as in PPMI, we focus on the F -measure. As Fig. 2 (middle-right panel) shows, most classifiers perform in a similar range if matched with a suitable imputation method. As with PPMI, the choice of imputation method can have a noticeable effect though. Again, this choice is not consistent across classifier models. For LR/RF/GCN, the worst/best classifiers are kNN/MICE_RF, MICE_LR/Mean, and PPCA/Mean, respectively. A noteworthy performance is achieved by the combination of RF classifier with Mean-imputation. This combination achieves a significantly higher F -measure than MGMC (and all other methods) at 100% data availability (and a significantly higher ROC-AUC/Accuracy also at 75% data availability). However, RF paired with the worst imputer MICE_LR leads to a significantly worse performance for 25% and 50% data availability.

In terms of imputation quality (RMSE), Fig. 2 (bottom row) shows that in both PPMI and TADPOLE, our method imputes better (i.e. lower RMSE) than all other methods, and highly significantly ($p < 0.001$) in all comparisons, except when comparing to our previously proposed method GMC at 50% and 75% data availability. Among the other methods, the best-performing imputers for PPMI were Mean-imputation and the two matrix completion methods MC and MG-RGCN, while for TADPOLE, the best imputer was PPCA at 25%, and MICE_RF at 50% and 75% data availability. MICE_LR was the worst-performing data imputer in both datasets. Furthermore, the trend is visible that all imputation methods impute with higher RMSE errors as fewer data is available in the feature matrix, whereas our proposed MGMC method provides fairly robust imputation results.

In our ablation experiments, we investigated how non-autoregressive LSTMs affect the imputation and classification performance. In Fig. 3 top, we observe that for the PPMI dataset, the non-autoregressive model yields significantly better results in terms of ROC-AUC, F -measure, and Accuracy at all levels of data missingness. For the TADPOLE dataset (Fig. 3 bottom), the proposed method classifies comparably well at 50%, 75% and 100% data availability, but significantly outperforms the autoregressive model at 25% data availability, demonstrating better classification robustness at lower levels of data availability.

4. Discussion

4.1. Classification performance when using all available data

In PPMI, we observed that our proposed approach achieved a consistently high classification performance in terms of ROC-AUC, F -measure, and Accuracy for PD prediction when compared with standard ML models (LR and RF), MC, MG-RGCN and GMC approaches, as shown in Fig. 2 (left panel row 1–3). The only method that was able to perform equally well (and significantly better at 25% data availability) was GCN, when optimally paired with PPCA imputation. In TADPOLE, we observed that our approach is mostly at par with baseline ML methods and SOTA approaches from literature, and could only be significantly outperformed by RF and at 75–100% data availability, and only if RF was optimally paired with Mean-imputation. As mentioned in the dataset descriptions, PPMI is 100% feature-complete at baseline, whereas TADPOLE is only 83% complete at baseline. It is noteworthy that at 100% data availability, MGMC already performs imputation in TADPOLE, but we cannot validate the imputed values due to a lack of groundtruth data for those missing features. Compared to previous studies, Zhou et al. [13] reported $\sim 60\%$ classification accuracy and ~ 0.6 ROC-AUC for the same AD classification problem posed in this paper for the TADPOLE dataset. Gray et al. [40] reported $\sim 60\%$ classification accuracy and ~ 0.7 ROC-AUC. In our study, we also achieve a classification accuracy on the order of $\sim 60\%$, however with higher ROC-AUC values on the order of ~ 0.8 . To interpret these results, we recall that the Accuracy metric represents the number of true positive and true negative cases among the total population, at a fixed threshold of the model's posterior. In comparison, the ROC-AUC gives an estimate of the likelihood that a classifier simultaneously achieves a high true positive rate and low false positive rate. This indicates that MGMC, compared to related works, and compared to baseline models at 25% data availability, achieves a more robust classification outcome, not only in terms of sensitivity, but also in form of a lower likelihood for type I errors. A likely reason for the ROC-AUC difference of ~ 0.1 compared to [40] is that earlier (2013) versions of the ADNI dataset had a smaller sample size, which also makes comparisons to our work somewhat unfair. Compared to [13], the ROC-AUC difference of ~ 0.2 can be likely attributed to the use of multi-graph convolutions in our work, which are trained end-to-end in a semi-supervised manner.

4.2. Classification performance with artificially removed data

To investigate the robustness of MGMC and baseline methods with respect to missing data, we randomly reduced the amount of available data in the feature matrix relative to the number of observed entries at

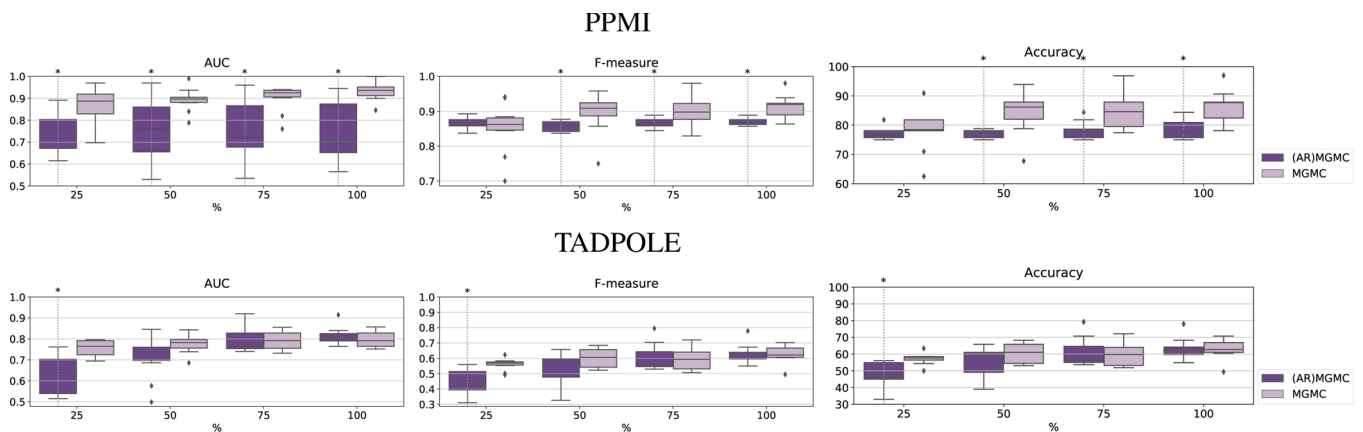


Fig. 3. PPMI (top) and TADPOLE (bottom) ablation results. ROC-AUC (left), F -measure (middle), and Accuracy (right) results on test dataset. Asterisk (*) and dotted vertical line denote model is statistically significantly different ($p \leq 0.05$) to proposed model. Values on x-axis denote relative percentage of features which are available to the network.

baseline, as shown in Fig. 2. We observed that the proposed approach has better and more stable classification and imputation results for PD prediction in PPMI when more information is missing. This effect is particularly visible in the ROC-AUC values, which may increase in standard deviation over the ten cross validation folds, but stay relatively stable in terms of median values above 0.9, even at low level of data availability around 25%. In comparison, LR, RF, MC, and MG-RGCN suffer from a noticeable drop in classification robustness. Interestingly, the single-graph GMC also yields relatively constant ROC-AUC values, but at a significantly lower level than MGMC. Furthermore, MC and MG-RGCN have an unstable and lower classification performance. This has two important implications. First, end-to-end learning of simultaneous imputation and classification, e.g. via geometric matrix completion, can improve the robustness of the CADx model towards the level of incompleteness in datasets up to a certain degree. Second, multiple RGCNs in parallel, e.g. fused by self-attention, improve both downstream tasks significantly, compared to using a single-graph or multiple graphs with a single RGCN. It is important to mention that we re-implemented MG-RGCN for comparison [19], as no reference implementation was available open-source. The on-par performance with many other algorithms on TADPOLE demonstrates a working re-implementation, however we have no clear explanation for the comparably low performance on PPMI. One factor that could partially contribute is that each graph in [19] utilizes a different feature set due to a graph-wise feature selection step as pre-processing. However, as none of the other algorithms in our comparison experiments used any sort of feature selection in the pre-processing stage, we also applied the full feature matrix to each branch of the RGCN, to make the comparison on same grounds. In MGMC, it is important to note that the two downstream tasks do not always benefit equally. In TADPOLE, for example, we observe a comparably stable classification performance at 75%, 50% and 25% data availability. However, a similar behaviour is observed for all other classifiers, and all classifiers in general classify similarly well. The only exception is the combination of Mean + RF where we observe a higher classification performance in TADPOLE. We hypothesize that one reason for this advantage could be due to the transductive imputation bias introduced in this model, since we performed imputation of the training set features together with the test set features. Another reason could also be the fact that we performed a hyperparameter tuning with nested cross-validation for all classifiers, including RF. For certain problems or datasets, apparently including TADPOLE, such hyperparameter optimization can achieve a noticeable performance boost, but not all translational studies of ML in medicine apply this step during their analyses. Only at 25–50% data availability, MGMC significantly outperforms other classifiers like GCN and LR, but only if these classifiers are matched with the worst-performing imputer (LR + kNN, and GCN + PPCA). As such, we consider this a negligible advantage for MGMC. Clearly, the main benefit of our proposed method on TADPOLE data lies not in an improved classification, but in a significantly more accurate imputation of missing values.

4.3. Joint classification and imputation performance

Most related literature in CADx naturally puts a focus on classification performance. Imputation is an often overlooked factor, even though it plays an important role in population-based and multimodal studies in medicine, as data missingness is a common problem here [41]. Considering the imputation performance in Fig. 2, our proposed approach is able to significantly outperform standard imputations (such as mean, kNN, and MICE, and PPCA) and other matrix completion approaches (MC and MG-RGCN) at all levels of missingness, on both datasets. This suggests that the proposed approach is able to take advantage of using known (semi-supervised) class label information in order to impute the features while simultaneously predicting the unknown class labels. It further suggests that the proposed method learns more class relevant feature representations compared to standard

imputation approaches (mean, kNN, MICE, and PPCA) and other matrix completion methods (MC, MG-RGCN). We can also observe that population modeling and graph incorporation cannot always compensate for sub-optimal imputation, we would always have to find the right combination of imputer and classifier in order to achieve a comparable result with MGMC. Interestingly, even though MC and MG-RGCN also make use of the class label information, just like our GMC or MGMC approaches, their model performance did not significantly improve on both datasets. We hypothesize that this could be due to the feature representational capacity of MC and MG-RGCN. Additionally, MG-RGCN only makes use of a single RGCN which is autoregressive, just like GMC, and our experiments have shown that this could have a significant influence as can be seen in Fig. 3. One limitation to note is that we were not able to compare the imputation results to further matrix completion works in literature, e.g. [11–13], as those works do not report imputation fidelity, e.g. via RMSE. However, as a surrogate, we implemented an MC approach which is gradient-based and non-graph-based learning MC approach, and its results can serve as a stand-in for this family of methods. Furthermore, we can compare classification performance on TADPOLE data with [13], who used the same subjects (examinations at baseline) and classes (NC, MCI and AD) in TADPOLE as we did in our study. Here, authors explored classification performance of their proposed method, given 10% and 20% data missingness on either the MRI or SNP modality. As authors in [13] report, the results of our proposed approach are in line with their classification accuracy results at 20% data missingness (~60% Accuracy) which corroborates our results on 75% data availability in Fig. 2 middle row. Finally, it is noteworthy that our proposed approach achieved a more accurate and stable classification performance for the PPMI prediction task than for the TADPOLE prediction task. A possible explanation is that distinguishing healthy controls from PD may be a simpler classification task than the three-class classification problem in TADPOLE (NC vs. MCI vs. AD). This notion is supported by clinical studies arguing that distinguishing NC, MCI, and AD based on clinical characteristics is a difficult problem at baseline [42].

4.4. Ablation experiments

In Section 2.4, we described our proposed improvements for usage of multiple RGCNs, specifically the usage of non-autoregressive LSTMs over autoregressive ones. Autoregressive RGCNs always use the output from the previous timestep and information from the previous LSTM cell-block as input. In contrast, non-autoregressive RGCNs always use the original input features as input at every timestep. Our motivation for using non-autoregressive LSTMs in MGMC is that the current output is always conditioned on the original input features. Intuitively, this should help the reconstructed output to avoid diverging from the input data, which is a desirable behaviour in matrix completion. Here, we perform and discuss an ablation experiment, where we compare the effect of both, as shown for PPMI and TADPOLE in Fig. 3. We observe that by using non-autoregressive LSTMs, we obtain a significantly better classification performance for all levels of data availability in PPMI. In TADPOLE, this tendency is not as clear, and a significant improvement is only achieved at 25% relative available data. At 50%, 75% and 100% available data, non-autoregressive LSTMs do not improve classification, but neither do they worsen the performance. This result suggests that it is indeed preferable to use non-autoregressive LSTMs in each parallel graph branch in MGMC. We attribute this to the intuitive notion explained above: by conditioning the reconstructed output on the original input data at every optimization timestep, we stabilize the reconstruction and achieve a better classification performance.

4.5. Overall implications

The main differences of our proposed approach to recent works that use RGCNs for matrix completion [21,19,20] are three-fold, namely (i)

the use of multiple LSTMs which are non-autoregressive, (ii) the use of self-attention weighting to aggregate information from (iii) multiple graphs representing different neighborhood relationships between patients in the population. Previous RGCN/GMC methods [21,19] use a single LSTM, while in our approach, we utilize one separate LSTM for every graph, which results to multiple recurrent graph convolutional networks. Notably, Monti et al. [20] also use a multi-graph formulation, but their approach differs from our method, since they consider both the rows and columns of the feature matrix as two separate graph structures. Instead, in this work, we consider multiple meta-information as separate graphs that contain rows of a feature matrix as the node features, similarly to [17].

A general take-away from our experiments is that the best choice of the imputation method is apparently not really dependent on the data, but mostly depends on the classification algorithm following imputation instead. Almost every imputation method that we tested in this work (Mean, kNN, MICE_RF and PPCA) appeared either as the best or worst imputation method, depending on which data it was applied and in combination with which classification algorithm. Only MICE_LR was consistently a bad match, for any classifier, and the RMSE analyses revealed that it was probably due to a consistently bad imputation performance. Overall, the data under observation, the chosen imputation and the classification models together form a complex interplay, which makes a careful examination and benchmarking necessary. In translational ML works on medical data, e.g. for CADx, such exhaustive analyses are rarely made. This is probably due to the fact that an exhaustive testing of all possible combinations of classifiers and imputation methods can quickly lead to very large numbers of experiments. When adding hyper-parameter optimization for every possible combination (as we did in our experiments), the required computational effort for nested cross-validation and the evaluation of all model setups may become a challenge. It is precisely this variability that highlights the attractiveness of our proposed MGMC approach. Imputation and classification are learned end-to-end, in a single model. Although it is not guaranteed that MGMC always achieves the best classification performance, our experiments provide evidence that the imputation is significantly better in all settings, and the classification is top-ranking compared to a wide range of classification methods, both shallow and deep, both transductive and inductive, and using matrix completion or not.

Finally, our work has certain limitations, which may suggest interesting avenues for future contributions. Following [14,15], our graph construction heuristic assumes a simple static graph. Recently, it has been shown that is possible to learn a clinical population graph end-to-end, along with the classification downstream task [43]. The resulting graph is optimally suited for e.g. classification. Consequently, an alternative approach would be to use the meta-information and the feature matrix information in parallel to build or learn the graph adjacency. The advantages could be potentially several-fold: (i) the classification accuracy might benefit from a better graph, (ii) the robustness might increase even further, compared to our applied heuristics for graph construction, (iii) no domain expertise would be necessary to manually define the optimal thresholds θ (cf. Eq. (1)) that determine patient similarity and connectedness in the graph, and (iv) the learned graph might be an end in itself, and serve as a form of knowledge discovery in medicine (e.g. discovery of previously unknown, yet connected sub-populations) [43]. Both approaches could potentially lead to better performances of the downstream tasks (classification and imputation). Another limitation is that we benchmarked our proposed MGMC method to several baseline methods (LR and RF) which are all inductive learning approaches. In contrast, our approach is inherently transductive, as we rely on spectral graph convolutions in the parallel graph-convolutional layers. We believe that it should be possible to incorporate imputation losses into the objective functions of GraphSAGE [44] or GAT [45] to obtain an inductive form of MGMC, and it is worth investigating whether the same benefits can be observed as in our

experiments. Furthermore, future works could compare against other non-deep learning based techniques that tackle missing data such as [46] and [47] and address non-MAR scenarios of missingness.

5. Conclusion

In conclusion, we propose a novel automatic disease classification method which can handle multimodal data with missing information, a common setup in medical population based studies and datasets. We accomplish this by using Multi-graph Geometric Matrix Completion (MGMC). We train our architecture through Multiple Recurrent Graph Convolutional Networks, which are optimized in an end-to-end manner. Experimental results suggest the effectiveness of our proposed approach on two well-known and challenging population based studies of neurodegenerative Parkinson's and Alzheimer's diseases. Furthermore, ablation experiments highlight the importance of using non-autoregressive LSTM including the effect of self-attention weighting. These results could serve as a baseline for future works on disease classification in incomplete datasets. In addition, this could be useful in other domains where incomplete, multimodal, and high-dimensional data is an issue.

Conflicts of interest

None declared.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

- [1] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data* 2019 Dec;6:54.
- [2] Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, Weiner MW, et al. Tadpole challenge: prediction of longitudinal evolution in alzheimer's disease. 2018 (arXiv preprint), arXiv:1805.03909.
- [3] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, et al. The parkinson progression marker initiative (ppmi). *Prog Neurobiol* 2011;95(4):629–35.
- [4] Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's Dement* 2019;15(3):321–87.
- [5] Koumakis L, Chatzaki C, Kazantzaki E, Maniadi E, Tsiknakis M. Dementia care frameworks and assistive technologies for their implementation: a review. *IEEE Rev Biomed Eng* 2019;12:4–18.

- [6] Kim K-S. Toward neuroprotective treatments of Parkinson's disease. *Proc Natl Acad Sci USA* 2017 Apr;114:3795–7.
- [7] Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. Mice vs ppca: missing data imputation in healthcare. *Inform Med Unlocked* 2019;17:100275.
- [8] Van Buuren S. Flexible imputation of missing data. CRC Press; 2018.
- [9] Tanveer M, Richhariya B, Khan R, Rashid A, Khanna P, Prasad M, et al. Machine learning techniques for the diagnosis of alzheimer's disease: a review. *ACM Trans Multimed Comput Commun Appl (TOMM)* 2020;16(1s):1–35.
- [10] Buuren Sv, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *J Stat Softw* 2010;1–68.
- [11] Thung K-H, Adeli E, Yap P-T, Shen D. Stability-weighted matrix completion of incomplete multi-modal data for disease diagnosis. *Intl. conf. on medical image computing and computer-assisted intervention* 2016:88–96.
- [12] Thung KH, Yap PT, Adeli E, Lee SW, Shen D. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Med Image Anal* 2018;45:68–82.
- [13] Zhou T, Liu M, Thung KH, Shen D. Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. *IEEE Trans Med Imaging* 2019;38(10):2411–22.
- [14] Parisot S, Ktena SI, Ferrante E, Lee M, Moreno RG, Glocker B, et al. Spectral graph convolutions for population-based disease prediction. *Intl. conf. on medical image computing and computer-assisted intervention* 2016:177–85.
- [15] Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Med Image Anal* 2018;48:117–30.
- [16] Kazi A, Krishna S, Shekarforoush S, Kortuem K, Albarqouni S, Navab N. Self-attention equipped graph convolutions for disease prediction. 2019 IEEE 16th intl. symposium on biomedical imaging (ISBI 2019) 2019:1896–9.
- [17] Kazi A, Shekarforoush S, Arvind Krishna S, Burwinkel H, Vivar G, Wiestler B, et al. Graph convolution based attention model for personalized disease prediction. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, editors. *Medical image computing and computer assisted intervention – MICCAI 2019*. Springer Intl. Publishing; 2019. p. 122–30 (Cham).
- [18] Kazi A, Shekarforoush S, Arvind Krishna S, Burwinkel H, Vivar G, Kortuem K, et al. InceptionGCN: receptive field aware graph convolutional network for disease prediction. *Information processing in medical imaging*, vol. 11492. Springer Intl. Publishing; 2019. p. 73–85.
- [19] Valenchon J, Coates M. Multiple-graph recurrent graph convolutional neural network architectures for predicting disease outcomes. *ICASSP 2019 – 2019 IEEE intl. conf. on acoustics, speech and signal processing (ICASSP)* 2019:3157–61.
- [20] Monti F, Bronstein MM, Bresson X. Geometric matrix completion with recurrent multi-graph neural networks. *Proc. intl. conf. neural information processing systems (NeurIPS)* 2017:3700–10.
- [21] Vivar G, Zwergal A, Navab N, Ahmadi S-A. Multi-modal disease classification in incomplete datasets using geometric matrix completion. *Graphs in biomedical image analysis (GRAIL)*, vol. 11044; 2018. p. 24–31.
- [22] Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The sri24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp* 2010;31(5):798–819.
- [23] Iglesias JE, Liu C-Y, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging* 2011;30(9):1617–34.
- [24] Baur C, Wiestler B, Albarqouni S, Navab N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *Intl. MICCAI brainlesion workshop* 2018:161–9.
- [25] Rao N, Yu H-F, Ravikumar P, Dhillon IS. Collaborative filtering with graph information: consistency and scalable methods. *Neural Inf Process Syst (NIPS)* 2015:1–9.
- [26] Kalofolias V, Bresson X, Bronstein M, Vandergheynst P. Matrix completion on graphs. 2014. arXiv:1408.1717.
- [27] Bronstein MM, Bruna J, Lecun Y, Szlam A, Vandergheynst P. Geometric Deep Learning: going beyond Euclidean data. *IEEE Signal Process Mag* 2017;34(4):18–42.
- [28] Goldberg A, Recht B, Xu J, Nowak R, Zhu X. Transduction with matrix completion: three birds with one stone. *Advances in neural information processing systems (NIPS)*. 2010. p. 757–65.
- [29] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in neural information processing systems*. 2017.
- [30] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems (NIPS)*. 2016. p. 3844–52.
- [31] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
- [32] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. *Advances in neural information processing systems*. 2019. p. 8024–35.
- [33] Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a python library for model selection and hyperparameter optimization. *Comput Sci Discov* 2015;8(1):014008.
- [34] Breiman L. Random forests. *Mach Learn* 2001;45(1):5–32.
- [35] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for dna microarrays. *Bioinformatics* 2001;17(6):520–5.
- [36] Tipping ME, Bishop CM. Probabilistic principal component analysis. *J R Stat Soc: Ser B (Stat Methodol)* 1999;61(3):611–22.
- [37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12(Oct):2825–30.
- [38] Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcamethods – a bioconductor package providing pca methods for incomplete data. *Bioinformatics* 2007;23(9):1164–7.
- [39] Komer B, Bergstra J, Eliasmith C. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. In: *ICML workshop on AutoML*, vol. 9; 2014.
- [40] Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Initiative ADN, et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. *NeuroImage* 2013;65:167–75.
- [41] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med* 2012;367(October):1355–60.
- [42] Langa KM, Levine DA. The diagnosis and management of mild cognitive impairment: a clinical review. *JAMA* 2014;312(23):2551.
- [43] Cosmo L, Kazi A, Ahmadi S-A, Navab N, Bronstein M. Latent-graph learning for disease prediction. *Medical image computing and computer assisted intervention – MICCAI 2020*, vol. 12262. Springer International Publishing; 2020. p. 643–53. Series Title: Lecture Notes in Computer Science.
- [44] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Advances in neural information processing systems*. 2017. p. 1024–34.
- [45] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. *Intl. conf. on learning representations* 2018.
- [46] Wang G, Deng Z, Choi K-S. Tackling missing data in community health studies using additive ls-svm classifier. *IEEE J Biomed Health Inform* 2016;22(2):579–87.
- [47] Venugopalan J, Chananani N, Maher K, Wang MD. Novel data imputation for multiple types of missing data in intensive care units. *IEEE J Biomed Health Inform* 2019;23(3):1243–50.

Peri-diagnostic decision support in CADx using Deep Learning (MICCAI 2020)

Contents

4.1	Motivation	53
4.1.1	Towards peri-diagnostic decision support	54
4.1.2	Efficient peri-diagnostic decision support	54
4.1.3	Patient-specific peri-diagnostic decision support	55
4.2	Problem Definition	55
4.3	Related work	56
4.4	Contribution	56

4.1 Motivation

Medical diagnosis is the process of determining a medical condition based on a patient's signs and symptoms [5]. A patient complaining of headaches could be a result of different conditions. Paleness or redness in the skin could also be attributed to a number of conditions. Through years of practice including evidence-based knowledge, medical doctors have mastered what to do based on the patient's signs and symptoms. They are able to do this by following medical guidelines including their own experiences in order to support their decisions.

In a clinical setting, this is what is known as the diagnostic process [5]. This process is an iterative process of gathering, integration, and interpretation of information resulting in a working diagnosis [5]. This iterative process involves the gathering of clinical history and interview, physical examination, diagnostic testing, referral and consultation. Only when sufficient information has been collected a diagnosis will be communicated. For certain diseases, this diagnostic process is laid out in the form of guidelines. In the UK for example, they have the National Institute of Health and Care Excellence (NICE) that sets out these guidelines to ensure evidence-based medicine [22].

Medical diagnosis is one of the most important processes in a clinical workflow [75]. Being able to arrive at the correct diagnosis at the right time is critical as people's general well-being could be affected. The diagnosing clinician will have to interpret all the information presented to them and make sense of it in relation to the situation of the patient. With the complexity of the clinical presentation of the disease, any tool which can reduce the uncertainty of the condition will be very useful for the clinician.

Recent methods in CADx research has been mostly deep learning-based [94] and do not consider the cost of every examination. It is assumed that the examination/feature from a patient comes for free [69]. This could be due in part that most CADx methods using deep learning are mainly focused on improving the classification performance targeted to a particular medical application. Furthermore, since it is assumed that these examination or features are freely available, most recent CADx methods assume that all examinations of a patient are all present before CADx methods start to be applied. However, the diagnosing physician can benefit from decision support beforehand, i.e. during the diagnostic process and while the diagnostic information or sensor data is still being acquired. The notion of decision support during the sequential acquisition of information about the patient is the focus of this chapter.

4.1.1 Towards peri-diagnostic decision support

Providing decision support throughout the diagnostic process is what we refer to as peri-diagnostic decision support [88]. We use the prefix “peri-” to indicate the analogy to the term “peri-operative” in Surgical Data Science [44], which takes into account all phases of the operative process, including pre-operative planning, intra-operative support, and post-operative therapy decision support. From a doctor’s perspective it answers the question "Given what is known so far about the medical status of the patient, which diagnostic examination should be taken next?". In short, peri-diagnostic decision support is a step further in CADx, wherein the system could provide support even at the start of the diagnostic workflow unlike previous CADx methods. An algorithm that provides decision support throughout the diagnostic process or even at the start of the diagnostic process could be very helpful. The process of diagnosing a condition is complex and a system that could provide diagnostic support could reduce the uncertainty of the condition and support the clinician with their decisions. Such a system could optimize the diagnostic process and lead to an efficient peri-diagnostic decision support.

4.1.2 Efficient peri-diagnostic decision support

In the context of peri-diagnostic decision support, efficiency means that we can achieve the best classification with the least cost and fewest examinations possible. The cost could be in the form of monetary expense, time spent on the diagnostic step, the burden imposed on the patient (e.g. pain, or risk), or any other variable which is very important and has value depending on the domain. One example in CADx would be assigning a monetary cost for every examination of a patient and considering this information when providing diagnostic decision support. Assigning more "cost" to an invasive procedure than a non-invasive procedure is another form of cost. If a medical observation would use more hospital resources than another medical observation then the latter would have a lower cost. Cost allocation could vary depending on what is valuable in a domain.

Next to cost, the number of diagnostic iterations is also important. It is also desirable to reach a final diagnostic decision in the quickest possible way, by using an optimal sequence of examinations among the known set of possible examinations. For example, if a model could

already have enough certainty that on average the 4th or 5th examination would no longer improve the classification performance of a model then we would no longer need to perform more than five examinations for a particular disease classification problem.

Such efficiency has the potential to optimize hospital resources. Efficiency is at the core of healthcare services. With the recent global COVID-19 pandemic, we have seen that healthcare systems around the world have been put to the test and pushed to the limits. There were not enough spaces for people to receive healthcare services during the global pandemic of 2019. Demand was too high and healthcare service supply could not cope with it. Having a diagnostic decision support system that is geared towards efficiency could potentially be very useful in such situations. Even though the focus of this thesis is on accurate and efficient CADx in medicine, an algorithm which considers a cost within the domain and provides the best decision support could be even be applied in other non-medical related domains. This raises further motivation to consider efficiency in peri-diagnostic decision support.

4.1.3 Patient-specific peri-diagnostic decision support

Next to efficiency, the notion of patient-specific CADx is highly valuable and thus requires attention as well as this has the potential to positively impact healthcare [29]. This means that decisions are specialized towards an individual rather than for a group of individuals. The decision which examination to examine next is not only based on guidelines for a group of individuals. Instead, it is also based on the currently known information about the individual. This means the whole diagnostic process should be individualized. The way information is gathered, combined, and interpreted requires individualized approaches.

An individualized or patient-specific approach is necessary as every individual is unique. Not only are we unique, but the human body is also complex. Every individual is different from one another. We also have to mention the complexity of a clinical condition if co-morbidities are present, which often results in a more challenging diagnostic process [5].

4.2 Problem Definition

Ultimately, we want a model that can suggest which examination to do next in an efficient manner, given what is known about the patient so far. Efficiency, as previously described, applies in terms of cost and the number of examinations requested. Not only should this model suggest the most efficient feature but it should also increase the classification performance best. Given this new examination including the previous examinations from a particular patient of interest, this should increase the classification performance the most. Only when there are no longer any examination methods available, we can stop the examination process or when a pre-determined budget has been exhausted.

Concretely, we formulate this problem in form of cost-sensitive feature acquisition. In this problem, the decision on which examination to do next can be posed as a feature acquisition problem wherein for every acquired feature a cost is also taken into consideration. In the context of CADx, the model is given a feature vector representing observations that are known

so far about the patient. Note that this vector could be incomplete at the start. Based on this information, the model should be able to suggest which examination should be done next, by suggesting which index in the remaining set of examination should be done next. We use this index to get the new observation for the patient of interest and evaluate if this new feature will increase the disease classification performance the most. We repeat this until there are no more remaining set of examinations to be done or alternatively one can also pre-defined budget to determine when to stop.

4.3 Related work

Recent works dealing with the problem of cost-sensitive feature acquisition at test-time can be roughly categorized based on which learning method was used: reinforcement-learning-based and non-reinforcement-learning-based.

Under reinforcement-learning-based approaches, recent approaches were mainly based on Q-learning [31, 34, 69]. The problem on which examination to select next is formulated as a Markov decision process. An agent is trained by setting the states based on the currently known or observed features and the action is the decision on which feature to select next. This newly acquired feature is then added to the currently known observations and a cost is paid or the budget is reduced. When the agent decides to perform the final prediction as the next action, the currently known observations are used as input to perform the final prediction and the acquisition process is terminated. The agent is then rewarded if the prediction on the currently known observations are correct, otherwise it is penalized.

Among non-reinforcement-learning-based works, there are methods which are attention-based [35] and gradient-based [33]. Attention-based methods directly estimate which feature to select next via an attention vector. Gradient-based methods, on the other hand, make use of the gradients of a trained network to select which feature to select next. Here, the gradient of a trained network is used to calculate feature attributions for every feature of every instance. Using the feature attributions and scaling this value with the cost of that feature, the next feature to select next is decided based on which has the feature attributions after the cost has been considered.

Lastly, to distinguish between active learning and the problem of cost-sensitive feature acquisition at test-time, we describe the difference in terms of the goal. In the context of active learning, the goal is to find which instance should be considered next to improve the learning performance [63]. In contrast, in active feature acquisition, the goal is to find which examination/feature should be selected next. In the context of CADx, this means the active learning strategy will suggest which patient should be labelled next to improve the classification performance. In contrast, the active feature acquisition strategy will suggest which feature/examination to select next.

4.4 Contribution

Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time (MICCAI 2020)

The goal in this work is to provide guidance to the clinician in the entire diagnostic process including the acquisition phase. To tackle this, we propose a gradient-based method for active feature acquisition at test time. This method is a cost-sensitive feature acquisition method that makes use of Integrated Gradients (IG). IG is an axiomatic feature attribution method that assigns feature attributions for every class of interest. Since we do not know the actual class of interest at test-time in the diagnostic process, we have to devise a strategy how to make use of these feature attributions for active feature acquisition. We further elaborate on this topic in the next section.

Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time (MICCAI 2020)

Gerome Vivar^{1,2}, Kamilia Mullakaeva¹, Andreas Zwergal², Nassir Navab^{1,3}, Seyed-Ahmad Ahmadi^{1,2}

¹Technische Universität München (TUM), Munich, Germany

²German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-Universität (LMU), Munich, Germany

³Johns Hopkins University, Baltimore, USA.

Copyright Statement. © Springer Nature Switzerland AG 2020. Reprinted, with permission, from Gerome Vivar, Kamilia Mullakaeva, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi, Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time, September 2020. Original publication can be found at: (https://doi.org/10.1007/978-3-030-59713-9_55). Included in this dissertation is the authors' accepted version due to copyrights.

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.

Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time

Gerome Vivar^{1,2,*}, Kamilia Mullakaeva^{1,*}, Andreas Zwergal², Nassir Navab^{1,3},
and Seyed-Ahmad Ahmadi^{1,2}

¹ Technical University of Munich (TUM), Munich, GER

² German Center for Vertigo and Balance Disorders (DSGZ),
Ludwig-Maximilians-Universität (LMU), Munich, GER

³ Whiting School of Engineering, Johns Hopkins University, Baltimore, USA

Abstract. Computer-aided diagnosis (CADx) algorithms in medicine provide patient-specific decision support for physicians. These algorithms are usually applied after full acquisition of high-dimensional multimodal examination data, and often assume feature-completeness. This, however, is rarely the case due to examination costs, invasiveness, or a lack of indication. A sub-problem in CADx, which to our knowledge has not been addressed by the MICCAI community so far, is to guide the physician during the entire peri-diagnostic workflow, including the acquisition stage. We model the following question, asked from a physician’s perspective: “Given the evidence collected so far, which examination should I perform next, in order to achieve the most accurate and efficient diagnostic prediction?”. In this work, we propose a novel approach which is enticingly simple: use dropout at the input layer, and integrated gradients of the trained network at test-time to attribute feature importance dynamically. We validate and explain the effectiveness of our proposed approach using two public medical and two synthetic datasets. Results show that our proposed approach is more cost- and feature-efficient than prior approaches and achieves a higher overall accuracy. This directly translates to less unnecessary examinations for patients, and a quicker, less costly and more accurate decision support for the physician.

Keywords: Computer-aided diagnosis; peri-diagnostic decision support; cost-sensitive feature attribution; integrated gradients

1 Introduction

The diagnostic workflow in medicine is “an iterative process of information gathering, information integration and interpretation” [2]. Information is first acquired through a clinical history and interview, followed by alternating examinations and working diagnoses, until sufficient information has been aggregated for a final diagnosis. The decision which examination to perform next lies in the responsibility of the physician, who has to consider its medical indication,

* G.V. and K.M. contributed equally to this work.

its invasiveness towards the patient, and often also its financial cost. Machine learning (ML) and computer-aided diagnosis (CADx) have a large potential for decision support in the clinic [16]. From a ML perspective, CADx is the task to learn the mapping of a multimodal feature vector onto a diagnostic label. Most CADx algorithms studied so far, however, ignore the acquisition stage, and provide decision support only at the end of the diagnostic workflow when all examination data is acquired and the feature vector is complete. As such, current CADx approaches miss out on the opportunity to aid the physician during the entire, *peri-diagnostic workflow*, including the acquisition stage. In this work, we address this problem by i) iteratively suggesting the next most important examination/feature to acquire, while ii) considering the overall examination cost and aiming for a maximally accurate and efficient diagnostic prediction. To the best of our knowledge, the problem of *peri-diagnostic decision support* has not been addressed in the MICCAI community so far.

Related works: In ML literature, this problem is often described as budgeted or cost-sensitive feature acquisition. Most recent approaches can be roughly categorized into reinforcement learning (RL) and non-RL approaches. Among *RL approaches*, [14] applied cost-sensitive n-step Q learning to CADx on Physionet (2012) and proprietary data. Kachuee et al. [7] classify diabetes with Deep Q-networks (DQN) and Monte-Carlo dropout, and select the feature with the maximum confidence gain of the predictor network while considering cost. [5] classify non-medical data with a DQN-variant that penalizes accumulated feature cost and incorrect predictions. RL-approaches have two important limitations: first, agent and predictor only work in tandem, neither has any utility or generalizability on its own. Second, unless agent and predictor are perfectly tuned, the network can quickly settle on a sub-optimal final classification accuracy in favor of low cost. Among the *non-RL approaches*, [3] classify fetal heartbeat patterns using Recurrent Neural Networks (RNN), which suggest the next feature through learned attention vectors as masks at every timestep. This can lead to suggesting several or repeated features at each timestep, and requires a fixed number of timesteps before its final prediction which can be inefficient. Kachuee et al. classify non-medical data [8] and detect hypothyroidism [6] using denoising autoencoders (DAE). The DAE is trained with dropout at the input layer, and learns to reconstruct complete feature vectors from incomplete inputs. Next, the encoder part is fine-tuned and trained in tandem with a predictor network towards the final prediction task. At test time, the partial derivatives of all outputs with respect to each input feature are aggregated to form the total “feature attribution”. In this context, it is important to note that feature attribution needs to fulfill four axioms, which have been derived in [15]. The gradient-based attribution only with respect to the input as performed in [8,6] violates the “Sensitivity Axiom” of feature attribution. This can lead to an acquisition of inefficient features [15] and ultimately, unnecessary patient examinations.

Contributions: 1) We propose for the first time to apply Integrated Gradients (IG), an axiomatic feature attribution method, to the problem of dynamic, budgeted feature acquisition. 2) We propose Accumulated IG (AIG), for dynam-

ically suggesting the next most important feature to acquire at test-time, and 3) we highlight the advantages of our proposed approach on two medical datasets and two explanatory datasets for illustration of its working principles.

2 Materials and Methods

2.1 Datasets and Preprocessing

To evaluate our method, we utilized four datasets. The first two are publicly available medical datasets, to demonstrate the efficacy of our method on real-life data. The latter two datasets are non-medical, for further benchmarking as well as to illustrate the inner workings and limitations of the different methods we evaluate. For pre-processing, we perform outlier removal and scaling in NHANES and Thyroid (Winsorization of real-valued features to [5, 95]-percentile, normalization to range [0,1]). **NHANES:** The “National Health and Examination Survey” dataset [1] contains demographic information, laboratory results, questionnaire, and physical examination data. The goal here is to predict diabetes (normal, pre-diabetes, and diabetes) based on measured fasting glucose levels. Costs for features were established in a crowd-sourcing approach [9], and represent the total ‘inconvenience’ of feature acquisition from a patient-perspective (including time burden, financial cost, discomfort, etc.) [7]. The cost varies from 1 to 9 on a relative, numeric scale. We use all 92062 samples and 45 features in this dataset. **Thyroid:** The UCI Thyroid disease dataset [13,4] poses a three-class classification problem (normal thyroid function vs. hyperfunction vs. or sub-normal function). There are 21 features, representing demographic information, questionnaires and laboratory results that are important for thyroid disease classification. Feature costs are provided as part of the public dataset “ann-thyroid” [13], and range from 1.00 to 22.78. We use all 7200 samples and 21 features. **MNIST:** In the MNIST dataset [11], we classify handwritten digit images in vectorized form, to simulate a tabular dataset. We use all 70,000 images with 784 features. We further assume a uniform cost of 1 for every pixel, to make our results comparable to related works. **Synthesized:** We also use a synthesized dataset as in [6], to further explain and visualize the feature attribution process. The dataset consists of 16,000 samples with 64 dimensions. The first 32 dimensions contain salient information for classification, at a linearly increasing cost from 1 to 32. The second 32 dimensions contain no valuable information for classification, again at a linearly increasing cost of 1 to 32. Hence, intuitively, an efficient feature acquisition approach should choose only features from the first 32 dimensions. For a more detailed explanation, we refer the reader to [6].

2.2 Problem Setting

In this work, we consider the problem of patient-specific, dynamic feature acquisition at test time. The goal is to sequentially acquire features that can achieve the maximum prediction performance, as efficiently as possible. We aim for a

model that is both cost- and feature-efficient, i.e. a model that achieves the maximum prediction performance with the least accumulated cost and smallest number of features possible. Formally, we consider the problem of predicting a target value $\hat{y} \in Y$ based on a feature vector $\bar{x} \in \mathbf{R}^d$, which initially contains incomplete information about the patient at test-time. For clarity, we denote a complete feature vector as x and an incomplete feature vector as \bar{x} .

2.3 Efficient Feature Acquisition at test-time using Integrated Gradients.

To efficiently acquire features at test-time, we propose to use feature attribution by Integrated Gradients (IG) [15]. Previous works make use of backpropagation for feature acquisition [6,8], by calculating the gradients of the network at the current input value. This, however, violates the ‘‘Sensitivity(a)’’ axiom of feature attribution [15], which states that if an input differs in one feature compared to a neutral baseline input, and if this leads to a different output, then that feature should be given a non-zero attribution. IG can be shown to uniquely satisfy the axiom ‘‘Sensitivity(a)’’, as well as the axiom ‘‘Implementation Invariance’’, which states that two different networks that produce the exact same outputs for the same inputs should produce the same feature attribution [15].

Where previous gradient-based approaches [8,6] only take the gradient at the current input, IG takes a path integral of the gradients while linearly blending between a baseline input $x' \in \mathbf{R}^d$ and the actual input $x \in \mathbf{R}^d$, to avoid local gradients becoming saturated [15]. The baseline input x' represents an ‘‘absence’’ of features and can be encoded as a zero-valued vector. Importantly, IG was originally designed for inference explanation, by computing feature attributions with respect to the known correct output class and model posterior. In our scenario, however, we do not know the output label of interest at test-time. We thus propose *Accumulated IG (AIG)*, i.e. to aggregate the attributions of all input features from all possible output classes (see eqns. 1 and 2). In addition, since we have an input \bar{x} which is initially empty at test-time, we have to use a different baseline in order to be able to calculate AIG. We thus represent missing features with a neutral baseline at the central tendency (i.e. mean), analogous to mean-imputation in regular machine learning. Here, accumulating the gradients implies combining attributions from K different functions. This follows the ‘‘Linearity Axiom’’ of attribution theory, keeping AIG axiomatic as in the original IG formulation [15].

To handle missing information at test-time, previous works [6,8] proposed to use denoising autoencoders (DAE). We validate a combination of DAE with AIG in our experiments, but we also propose a simplified version without the need for auto-encoding. The simplified model is a vanilla multi-layer perceptron (MLP) trained end-to-end, while applying a Beta-distributed dropout layer to the input [6] to simulate missing information during training (see Fig. 1).

Implementation Details: We approximate the continuous IG as in [15] by a few discrete steps. We calculate the attribution along the i -th dimension with respect to one specific class (k) using:

$$\text{IG}_{i,k}^{\text{approx}}(\bar{x}_i, \text{class}_k) = (\bar{x}_i - x'_i) \times \sum_{s=1}^m \frac{\partial F(x' + \frac{s}{m} \times (\bar{x} - x'))}{\partial \bar{x}_i} \times \frac{1}{m} \quad (1)$$

where \bar{x} is the input vector and x'_i the baseline at the i -th dimension, $\frac{\partial F(\cdot)}{\partial \bar{x}_i}$ is the partial derivative of the network's output with respect to input \bar{x}_i , and m is the number of approximation steps of the path integral in IG. We then sum up all the attributions for the current feature from all classes and aggregate both positive and negative gradients. To account for cost-efficiency, we scale the attribution to a feature by the inverse feature cost:

$$f^{(i)} = \frac{|\sum_{k=1}^K \text{IG}^{\text{approx}}(\bar{x}_i, \text{class}_k)|}{c_i} \quad (2)$$

where $f^{(i)}$ denotes the AIG feature attribute of input \bar{x}_i and c_i denotes its cost. Then $f_t \in \mathbf{R}^d$ is a vector which consists of AIG attributions of all features $[f_t^{(1)}, f_t^{(2)}, \dots, f_t^{(d)}]$ at timestep t . To determine which feature to acquire next, we take the index of the feature attribute with the maximum value: $a_{f_t} = \arg \max(f_t)$, where a_{f_t} denotes the feature to acquire at timestep t as illustrated in 1. Using this newly acquired feature and previously acquired features we then perform classification (a_{c_t}) on this incomplete feature vector and obtain the label y_t . We repeat this process until there are no more remaining features to acquire. Alternatively, one can set a maximum allowed cost to constrain feature acquisition to a maximum allowed budget. The network architecture and an unrolled feature acquisition process are illustrated in Fig. 1.

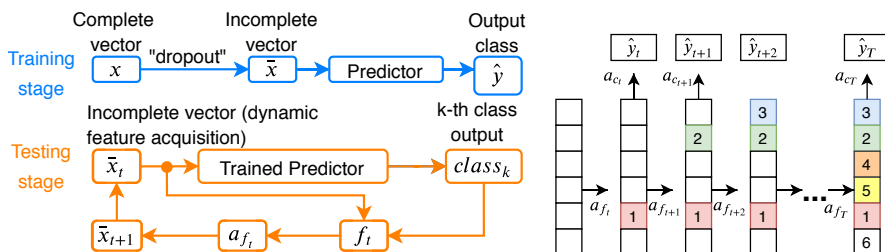


Fig. 1: Illustration of our proposed network architecture (left panel) and an unrolled feature acquisition sequence at test-time (right panel).

3 Results and Discussion

3.1 Experimental setup

Baseline model comparison: We evaluate our work against several baseline and state-of-the-art approaches in budgeted feature acquisition, including

a recent deep-RL [7], two non-RL [6,8], and a random feature selection based approaches. We split the datasets into 15% test set and 85% training set, where 15% of the latter is used for validation. We use Adam optimization [10] implemented in PyTorch [12] on a single-GPU workstation (Nvidia GTX 1080 Ti). For RL, we used the author implementation and parametrizations of Opportunistic Learning [7], to train an agent for 11,000 episodes. For comparison, all methods including our own are based on a two-layer multilayer perceptron (MLP) [64 and 32 units]. The non-RL approaches we compare against are Dynamic Feature Query (DPFQ) [8] and Feature Acquisition Considering Cost at Test-time (FACT) [6]. Again, we use the same MLP architecture for the encoder [64, 32], decoder [32, 64], and predictor [32, 16, K classes]. For the binary layer in FACT, we use the identical 8-bit representation as in [6]. Further, to randomly drop entries, we use a Beta-distribution with $\alpha = 1.5$ and $\beta = 1.5$, following [6].

Proposed: We use a vanilla MLP (encoder [64, 32], and predictor [32, 16, K classes]). We used the Adam optimizer in PyTorch with a low learning rate ($lr = 1e - 4$). We use $m = 50$ for the number of steps in the integral approximation in eqn. (1).

3.2 Feature acquisition performance

We compare our work with previous deep-RL [7] and non-RL [6,8] techniques to evaluate the effectiveness of our proposed method. We observe that our proposed AIG approach with and without DAE outperforms the SOTA methods, with a particularly large margin in the two medical datasets. Overall, our approach is the most cost- and feature-efficient (see Fig. 3) and consistently achieves the highest overall classification accuracy. The only exception is RL for Synthesized data, but otherwise RL lacks robustness. Our method’s feature-efficiency is evident e.g. in Thyroid and NHANES, on average, it is able to outperform the SOTA and reach the maximum classification accuracy after just 7 ($\sim 33\%$) and 10 ($\sim 22\%$) features, respectively (see Fig. 3). Importantly, this directly translates to the avoidance of unnecessary examinations and a much faster time-to-diagnosis, without requiring patients to undergo all examinations. Apart from feature-efficiency, our approach is also cost-efficient, e.g. spending only ~ 20 ($\sim 25\%$) units of cost in Thyroid, and ~ 50 ($\sim 29\%$) units in NHANES to achieve maximum classification performance. Further, methods like RL or DPFQ may choose cheaper features first, despite little gain in classification accuracy (see Fig. 3, right Thyroid panel), whereas our method suggests more costly features in the beginning, at the benefit of reaching the highest classification accuracy almost instantly.

3.3 Interpretation of patient-specific feature acquisition

We also use test samples of each dataset to visualize and discuss the order of feature selection by the different methods. We show heatmaps in Fig. 2, where warmer colors denote higher priority in the feature acquisition. We plot ten test samples for datasets Thyroid, NHANES and Synthesized, and one test image

from MNIST. In Fig. 2, we observe that our proposed approach initially always acquires the most informative feature, before starting to acquire features in an instance- or patient-specific manner. For example, in the Thyroid dataset, features 21, 2, and 3 are consistently selected first by the model, while at feature 1 or 19, model suggestions start to diverge which feature to acquire next. Similarly in NHANES, features 2 and 30 form an initial decision baseline, before the model diverges into patient-specific decisions at features 33 or 45. In contrast, FACT and RL may change the feature acquisition order almost instantly, already at the first or second acquisition step, which may not always be justified or effective. In MNIST, FACT heatmaps show an outlining of the digit, as FACT multiplies the output of the de-noising auto-encoder with the feature-aggregation score. This strategy prioritizes high-intensity/-amplitude features, and leads to intuitive visualizations on MNIST, but does not directly translate to an efficient feature acquisition performance, as seen e.g. in the NHANES dataset. Further, approaches like RL may choose features in random order (MNIST), or in order of least cost instead of relevance (Synthesized). In future work, we aim at investigating such phenomena from a medical perspective.

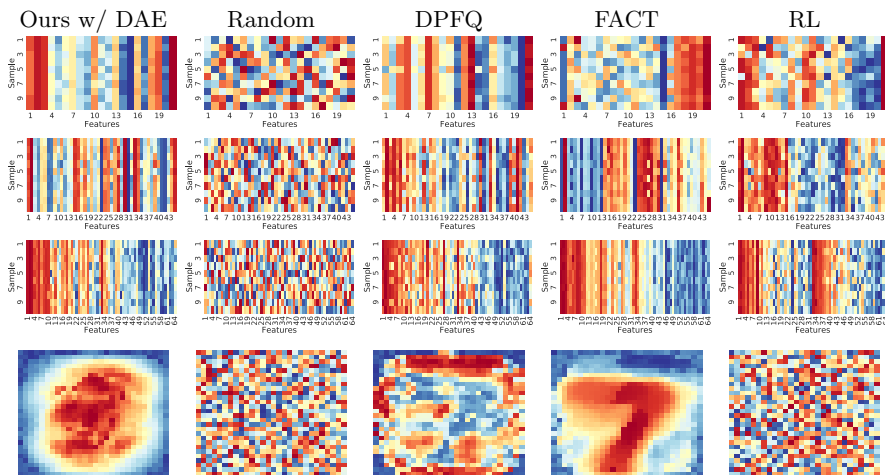


Fig. 2: Feature acquisition heatmaps for all datasets. From top to bottom: UCI-Thyroid (10 patient samples), NHANES (10 patient samples), Synthesized (10 samples), and MNIST. Warmer colors denote higher priority for feature acquisition.

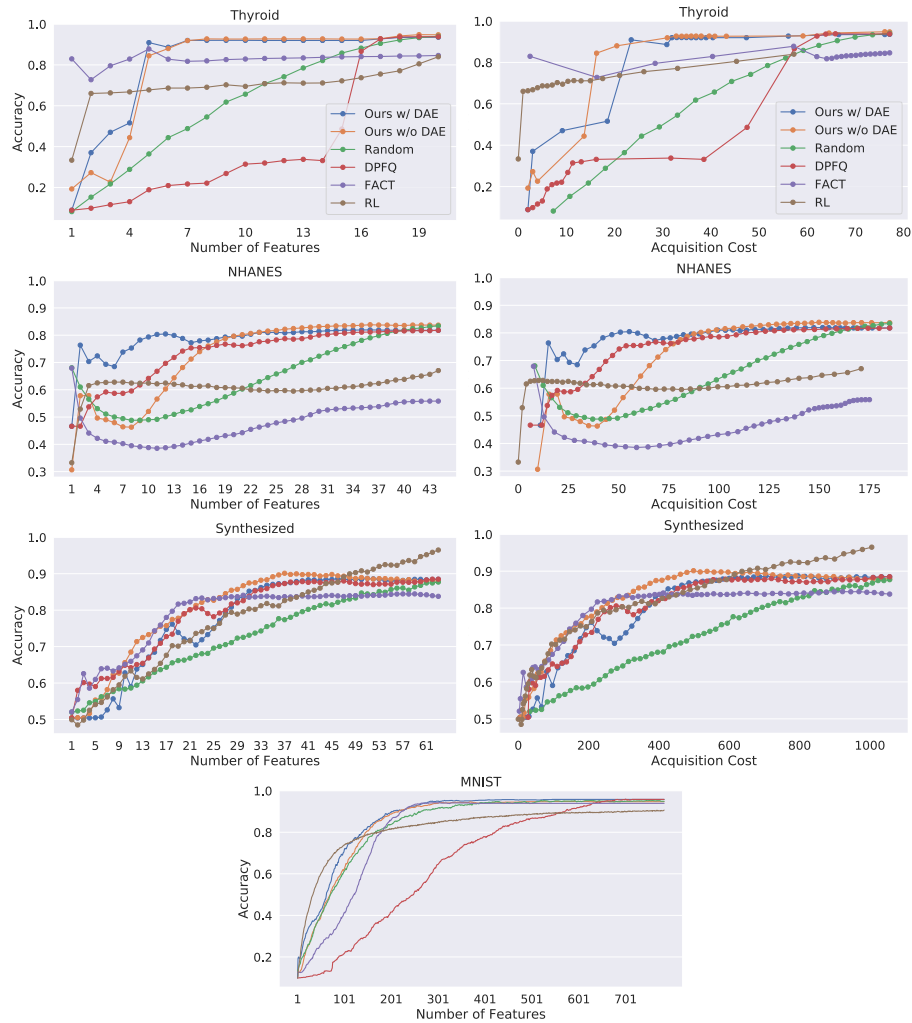


Fig. 3: Comparison of feature aggregation methods against our proposed approach, on four datasets (row 1: Thyroid; row 2: NHANES; row 3: Synthesized; row 4: MNIST with uniform feature cost). Left column: Feature count vs. accuracy curves; right column: Accumulated feature cost vs. accuracy. The compared baseline approaches denote: Random (random feature selection), DPFQ [8], FACT [6], RL [7]. Our approach is consistently most feature- and cost-efficient and achieves the highest classification final accuracy.

4 Conclusion

We propose a novel method which can efficiently acquire features at test-time, through Accumulated Integrated Gradients (AIG) and network training with dropout at the input layer. We empirically show that our approach is cost- and feature-efficient when evaluated on two medical datasets and two explanatory toy datasets. Our proposed method enables patient-specific, peri-diagnostic decision support for clinicians, which could potentially optimize spending, maximize hospital resources, and reduce examination burden for patients. Future work could address two important limitations of our work, which occur frequently in real-life clinical data, namely how to train a peri-diagnostic CADx system from data that is i) incomplete at training time and ii) made up of features from different modalities which are organized into blocks with acquisition costs that increase blockwise instead of one feature at a time.

Acknowledgments: This work was supported by the German Federal Ministry of Education and Health (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) [grant number 01 EO 0901].

References

1. Centers for Disease Control and Prevention (CDC): National health and nutrition examination survey (NHANES). <https://www.cdc.gov/nchs/nhanes/index.htm>, accessed: 2020-03-11
2. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine: Improving Diagnosis in Health Care. National Academies Press, Washington, D.C. (Dec 2015). <https://doi.org/10.17226/21794>, <http://www.nap.edu/catalog/21794>
3. Contardo, G., Denoyer, L., Artières, T.: Recurrent neural networks for adaptive feature acquisition. In: International Conference on Neural Information Processing. pp. 591–599. Springer (2016)
4. Dua, D., Graff, C.: UCI machine learning repository (2017), <http://archive.ics.uci.edu/ml>
5. Janisch, J., Pevný, T., Lisý, V.: Classification with costly features using deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3959–3966 (2019)
6. Kachuee, M., Darabi, S., Moatamed, B., Sarrafzadeh, M.: Dynamic feature acquisition using denoising autoencoders. *IEEE transactions on neural networks and learning systems* **30**(8), 2252–2262 (2018)
7. Kachuee, M., Goldstein, O., Kärkkäinen, K., Sarrafzadeh, M.: Opportunistic Learning: Budgeted Cost-Sensitive Learning from Data Streams. In: International Conference on Learning Representations (ICLR) (2019), <https://openreview.net/forum?id=S1e0Ho09KX>
8. Kachuee, M., Hosseini, A., Moatamed, B., Darabi, S., Sarrafzadeh, M.: Context-aware feature query to improve the prediction performance. In: 2017 IEEE Global

- Conference on Signal and Information Processing (GlobalSIP). pp. 838–842. IEEE (2017)
9. Kachuee, M., Karkkainen, K., Goldstein, O., Zamanzadeh, D., Sarrafzadeh, M.: Cost-Sensitive Diagnosis and Learning Leveraging Public Health Data. preprint <https://arxiv.org/abs/1902.07102> (2019)
 10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations (ICLR) (2015), <http://arxiv.org/abs/1412.6980>
 11. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (Nov 1998). <https://doi.org/10.1109/5.726791>, <http://ieeexplore.ieee.org/document/726791/>
 12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 32, pp. 8024–8035. Curran Associates, Inc. (2019)
 13. Quinlan, J.R., Compton, P.J., Horn, K.A., Lazarus, L.: Inductive knowledge acquisition: A case study. In: *Proceedings of the Second Australian Conference on Applications of Expert Systems*. p. 137–156. Addison-Wesley Longman Publishing Co., Inc., USA (1987)
 14. Shim, H., Hwang, S.J., Yang, E.: Joint active feature acquisition and classification with variable-size set encoding. In: *Advances in Neural Information Processing Systems*. pp. 1368–1378 (2018)
 15. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. pp. 3319–3328. JMLR. org (2017)
 16. Yanase, J., Triantaphyllou, E.: A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications* **138**, 112821 (Dec 2019). <https://doi.org/10.1016/j.eswa.2019.112821>

Translating Computer-aided Diagnosis (CADx) into clinical research settings using benchmarks of shallow and deep learning models.

Contents

5.1	Motivation and Problem Definition	69
5.2	Contributions	70
5.2.1	Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway (Journal of Neurology 2019)	70
5.2.2	Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders (Journal of Neurology 2020)	72
5.2.3	Using Base-ml to Learn Classification of Common Vestibular Disor- ders on DizzyReg Registry Data (Frontiers in Neurology 2021 . . .	84

5.1 Motivation and Problem Definition

Machine learning has progressed quite rapidly throughout the past decade. We have seen a lot of successful applications of ML in different domains including healthcare. However, the translation of these successes into the clinic as well as into clinical research is not often straightforward. This step is essential and there are many regulations to go through before they can be officially used in regular clinical practice. It is important to consider such compliance aspects, including the societal and ethical impact of one's approach. Nonetheless, it is very important that we are able to put these advancement in DL to good use. Therefore, as part of the efforts of this thesis, we strove to perform translational research into the clinic. To this end, we applied various ML and DL algorithms, both established ones and new ones proposed in this thesis, to datasets at our clinical partner institute, the German Center for Vertigo and Balance Disorders (DSGZ).

The goal of expert physicians at the specialized in-patient clinic of the DSGZ is to correctly diagnose the clinical presentation of patients with vertigo, dizziness, and postural/gait imbalance complaints. For certain conditions, doctors can easily perform diagnostic decisions. However, there are cases where the clinical presentation of the complaints are very difficult

to diagnose as they are confounding with symptoms of other diseases [2, 43, 89, 91]. As a result, it is not straightforward to determine whether the complaints could be attributed to a problem in the peripheral or central vestibular system making the diagnosis very difficult and potentially dangerous if not diagnosed correctly.

For example, one scenario with potential danger is the emergency department (ED), when patients present themselves with dizziness symptoms (around 4% of patients, [54]). Stroke is the underlying cause in 4–15% of those patients, but about 10% of strokes are missed at first contact [80]. If patients get discharged from the ED with a suspected benign diagnosis, they face a 50-fold increased risk of stroke in the first week compared to matched controls [4]. Naturally, this is scenario where every practitioner would like to avoid a false negative diagnosis.

In a collaboration between the DSGZ and the Chair for Computer-Aided Medical Procedures (CAMP) at TUM, we took several translational efforts to apply ML/DL-based CADx methods to address diagnostic challenges in posturography, emergency vertigo, and large dizziness registry datasets.

5.2 Contributions

5.2.1 Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway (Journal of Neurology 2019)

Neurological disorders of posturography and balance could offer clues to the underlying neurological disorder. Diagnosis can be done easily for some conditions (e.g. Persistent Postural-Perceptual Dizziness, PPPD, [76]) while others are much more difficult requiring further tests (e.g. polyneuropathy, PNP, [2]). One way to assess such difficult cases is using static posturography while undergoing different levels of controlled conditions to put the vestibular system to the test. Here, the displacement of the center-of-pressure is measured as a time-sequence. From this, clinically-relevant features can be derived, such as various frequency bands using Fourier analysis as well as accumulated sway paths and root-mean-square values.

Before the resurgence of neural networks, Support-Vector Machines and Random Forest were established methods for CADx. With the breakthroughs of deep neural networks in computer vision applications, we explore the utility of these recent advancements for medical applications such as in CADx. To this end, we explore the utility of neural- and non-neural network based models and applied ensemble methods to improve classification performance in neurological stance disorders.

Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway (Journal of Neurology 2019)

Seyed-Ahmad Ahmadi^{1,2}, Gerome Vivar^{1,2}, Johann Frei^{1,2}, Sergej Nowoshilow³, Stanislav Bardins¹, Thomas Brandt¹, Siegbert Krafczyk¹

¹German Center for Vertigo and Balance Disorders, Ludwig Maximilians Universität, Marchioninstr. 15, 81377 Munich, Germany

²Computer Aided Medical Procedures, Technical University of Munich, 85748 Garching, Germany

³IMP Research Institute of Molecular Pathology, Campus-Vienna-Biocenter 1, 1030 Vienna, Austria

Copyright Statement. © Springer-Verlag GmbH Germany, part of Springer Nature 2019. Reprinted, with permission, from Seyed-Ahmad Ahmadi, Gerome Vivar, Johann Frei, Sergej Nowoshilow, Stanislav Bardins¹, Thomas Brandt, Siegbert Krafczyk, Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway, July 2019.

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.

REPRINT DENIED. The reprint of this publication was rejected on open-access platforms. Original publication can be found at: (<https://doi.org/10.1007/s00415-019-09458-y>).

5.2.2 Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders (Journal of Neurology 2020)

In the emergency department, one of the most important aspects for the diagnosing clinician is to distinguish whether a patient suffering from acute vestibular syndrome is due to a central or peripheral acute vestibular disorder. Acute vestibular disorder or acute vestibular syndrome (AVS) is characterized by abrupt onset of acute, 'continuous' vertigo (lasting for more than 24 hours). Individuals suffering from this condition also suffer head motion intolerance, nausea, and vomiting [37]. Patients with central acute vestibular disorder are associated with stroke and peripheral acute vestibular disorder is more associated with pathology of the inner ear vestibular structures. One way to assess and detect stroke in acute vestibular syndrome is using HINTS (Head Impulse–Nystagmus–Test of Skew) [38] and ABCD² (Age, Blood, Clinical features, Duration, Diabetes) [55]. HINTS is performed to test the presence of any of the three oculomotor signs. ABCD² is used to score the risk for stroke in the days following transient ischaemic attack (TIA). Although previous works suggest that HINTS could yield sensitivity and specificity >90% [39], it still remains a challenge in the emergency department to differentiate patients central AVS from peripheral AVS due to the subjectivity of the test and reliance to clinical expertise of the examining physician. In this work, we compared state-of-the-art scoring tests with ML approaches, in order to assess their efficacy for diagnostic differentiation.

Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders (Journal of Neurology 2020)

Seyed-Ahmad Ahmadi^{1,2}, Gerome Vivar^{1,2}, Nassir Navab², Ken Möhwald^{1,3}, Andreas Maier^{1,3}, Hristo Hadzhikolev^{1,3}, Thomas Brandt^{1,4}, Eva Grill^{1,5}, Marianne Dietrich^{1,3,6}, Klaus Jahn^{1,7}, Andreas Zwergal^{1,3}

¹German Center for Vertigo and Balance Disorders, Ludwig-Maximilians-University, Munich, Germany

²Computer Aided Medical Procedures, Technical University, Munich, Germany

³Department of Neurology, Ludwig-Maximilians-University, Marchioninistrasse 15, 81377 Munich, Germany

⁴Clinical Neurosciences, Ludwig-Maximilians-University, Munich, Germany

⁵Institute for Medical Information Processing, Ludwig-Maximilians-University, Biometry, and Epidemiology, Munich, Germany

⁶Munich Cluster of Systems Neurology, SyNergy, Munich, Germany


⁷Department of Neurology, Schön Klinik Bad Aibling, Munich, Germany

Copyright Statement. © Springer-Verlag GmbH Germany, part of Springer Nature 2020. Reprinted, with permission, from Seyed-Ahmad Ahmadi, Gerome Vivar, Nassir Navab, Ken Möhwald, Andreas Maier, Hristo Hadzhikolev, Thomas Brandt, Eva Grill, Marianne Dietrich, Klaus Jahn, Andreas Zwergal, Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders, May 2020. Original publication can be found at: (<https://doi.org/10.1007/s00415-020-09931-z>). This article is licensed under a Creative Commons Attribution 4.0 International License <http://creativecommons.org/licenses/by/4.0/>.

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.



Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders

Seyed-Ahmad Ahmadi^{1,2} · Gerome Vivar^{1,2} · Nassir Navab² · Ken Möhwald^{1,3} · Andreas Maier^{1,3} · Hristo Hadzhikolev^{1,3} · Thomas Brandt^{1,4} · Eva Grill^{1,5} · Marianne Dieterich^{1,3,6} · Klaus Jahn^{1,7} · Andreas Zwergal^{1,3} 

Received: 14 April 2020 / Revised: 15 May 2020 / Accepted: 19 May 2020 / Published online: 11 June 2020
© The Author(s) 2020

Abstract

Background Diagnostic classification of central vs. peripheral etiologies in acute vestibular disorders remains a challenge in the emergency setting. Novel machine-learning methods may help to support diagnostic decisions. In the current study, we tested the performance of standard and machine-learning approaches in the classification of consecutive patients with acute central or peripheral vestibular disorders.

Methods 40 Patients with vestibular stroke (19 with and 21 without acute vestibular syndrome (AVS), defined by the presence of spontaneous nystagmus) and 68 patients with peripheral AVS due to vestibular neuritis were recruited in the emergency department, in the context of the prospective EMVERT trial (EMERGENCY VERTigo). All patients received a standardized neuro-otological examination including videooculography and posturography in the acute symptomatic stage and an MRI within 7 days after symptom onset. Diagnostic performance of state-of-the-art scores, such as HINTS (Head Impulse, gaze-evoked Nystagmus, Test of Skew) and ABCD² (Age, Blood, Clinical features, Duration, Diabetes), for the differentiation of vestibular stroke vs. peripheral AVS was compared to various machine-learning approaches: (i) linear logistic regression (LR), (ii) non-linear random forest (RF), (iii) artificial neural network, and (iv) geometric deep learning (Single/MultiGMC). A prospective classification was simulated by ten-fold cross-validation. We analyzed whether machine-estimated feature importances correlate with clinical experience.

Results Machine-learning methods (e.g., MultiGMC) outperform univariate scores, such as HINTS or ABCD², for differentiation of all vestibular strokes vs. peripheral AVS (MultiGMC area-under-the-curve (AUC): 0.96 vs. HINTS/ABCD² AUC: 0.71/0.58). HINTS performed similarly to MultiGMC for vestibular stroke with AVS (AUC: 0.86), but more poorly for vestibular stroke without AVS (AUC: 0.54). Machine-learning models learn to put different weights on particular features, each of which is relevant from a clinical viewpoint. Established non-linear machine-learning methods like RF and linear methods like LR are less powerful classification models (AUC: 0.89 vs. 0.62).

Conclusions Established clinical scores (such as HINTS) provide a valuable baseline assessment for stroke detection in acute vestibular syndromes. In addition, machine-learning methods may have the potential to increase sensitivity and selectivity in the establishment of a correct diagnosis.

Keywords Acute vestibular syndrome · HINTS · Machine-learning · MRI · Vestibular neuritis · Vestibular stroke

Abbreviations

ABCD² Age, blood pressure, clinical features, duration, diabetes
ANN Artificial neural network

AUC Area-under-the-curve
AVS Acute vestibular syndrome
CVRF Cardiovascular risk factors
DT Decision tree
DWI Diffusion weighted images
ED Emergency department
EMVERT EMERGENCY VERTigo
FLAIR Fluid attenuated inversion recovery
GMC Geometric matrix completion
HINTS Head impulse, gaze-evoked nystagmus, test of skew

Seyed-Ahmad Ahmadi and Gerome Vivar have contributed equally to this work.

✉ Andreas Zwergal
andreas.zwergal@med.uni-muenchen.de

Extended author information available on the last page of the article

LR	Logistic regression
ML	Machine-learning
MLP	Multilayer perceptron
MultiGMC	Multi-graph geometric matrix completion
RF	Random forest
ROC	Receiver operating characteristic
SingleGMC	Single-graph geometric matrix completion
SPN	Spontaneous nystagmus
SPV	Slow phase velocity
STD	Standard deviation
SVV	Subjective visual vertical
vHIT	Video-based head impulse test
VOG	Videoculography
VOR	Vestibulo-ocular reflex

Introduction

Patients with acute vertigo and dizziness account for about 4% of all visits to the emergency department (ED) [1]. Stroke is the underlying cause in 4–15% of all patients, and up to 25% of patients with the presentation of acute vestibular syndrome (AVS, defined by the presence of spontaneous nystagmus) [1, 2]. About 10% of strokes are missed at first contact [3]. Patients discharged from the ED with a suspected benign diagnosis of acute vertigo or dizziness have a 50-fold increased risk of stroke in the first week compared to matched controls [4]. Reasons for this deplorable situation are an overreliance on symptom quality and intensity as distinctive features, inadequate knowledge or application of bedside ocular motor examinations, and a blind trust in cerebral imaging results [5]. Consequently, ED physicians worldwide rank vertigo and dizziness as one of the top priorities for the development of better diagnostic algorithms [6].

Different concepts exist to differentiate peripheral and central etiologies of acute vertigo and dizziness [7, 8]. One strategy relies on a comprehensive examination of vestibular, ocular motor, and postural functions. For AVS, the HINTS test (Head Impulse, gaze-evoked Nystagmus, Test of Skew) has a high sensitivity and specificity (> 90%) for identification of stroke [9]. The diagnostic accuracy of HINTS can be further improved by video oculographic quantification of the head impulse test (vHIT) [10, 11]. Examination-based classification approaches require a profound knowledge of examination techniques and expertise in interpretation of findings. Another idea is to stratify the risk of vestibular stroke by diagnostic index tests, which aggregate information on symptom characteristics (such as symptom onset and duration, triggers, accompanying complaints) and cardiovascular risk factors (CVRF). For example, the ABCD² score (Age, Blood pressure, Clinical features, Duration, Diabetes) can help to estimate the risk of vestibular stroke, but is inferior to HINTS in diagnostic accuracy [12, 13]. The advantage of

index tests based on history taking is that they are easy to apply and not restricted to clinical subtypes such as AVS. Diagnostic approaches by magnetic resonance imaging (MRI) only, have a high rate of false-negative results (50% for lesions < 10 mm) in the first 48 h after symptom onset and are, therefore, not reliable during the acute stage [5, 14].

In the current study, we applied modern machine-learning algorithms to classify vestibular stroke vs. peripheral AVS due to vestibular neuritis based on a multimodal data set (including a standardized assessment of symptom features, CVRF, and detailed quantitative testing of ocular motor, vestibular, and postural functions). Machine-learning approaches were compared to state-of-the-art tests (such as HINTS, ABCD²) to evaluate their feasibility and value for diagnostic decision support.

Methods

Patient cohorts and study protocol

In total 108 patients, who were admitted to the ED of the University Hospital (LMU Munich), were included in this study and received a standardized assessment (of symptom features, CVRF, and vestibular, ocular motor and postural functions) following the EMVERT trial protocol [15]. Based on the findings of MRI (performed within 7 days after symptom onset) and videoculography (vHIT gain threshold: 0.7, refixation saccades, gaze-evoked nystagmus, skew deviation), 40 patients were diagnosed as having vestibular stroke (64.1 ± 12.2 years, 67.5% men, 19 with presentation of AVS), and 68 as having peripheral AVS due to vestibular neuritis (55.6 ± 14.6 years, 64.7% men). Classification algorithms (established index tests vs. modern machine-learning techniques) were applied post hoc to test their diagnostic accuracy for differentiation of both groups.

Protocol approval and patient consent

The study was approved by the Ethics Committee of the University of Munich on February 23, 2015 (57–15). The study was conducted according to the Guideline for Good Clinical Practice (GCP), the Federal Data Protecting Act and the Helsinki Declaration of the World Medical Association in its current version (revision of Fortaleza, Brazil, October 2013). All subjects gave their informed, written consent to participate in the study.

Assessment of symptom characteristics and cardiovascular risk factors

In all patients, a standardized history was taken in the ED, including the following features: symptom quality (vertigo, dizziness, double vision), symptom onset (acute, lingering),

symptom duration (10–60 min, > 60 min), symptom intensity (by visual analogue scale), preceding triggers (yes, no), accompanying features (ear symptoms, central neurological symptoms), and CVRF (diabetes, high blood pressure (> 140 mmHg), nicotine abuse, atrial fibrillation, family history, prior stroke or myocardial infarction). Health-related quality of life and functional impairment was assessed by questionnaires: European Quality of Life Score—5 dimensions—5 levels (EQ-5D-5L), including subscores for anxiety, pain, activity, self-care, and mobility (ranging from 1–5 each with 5 indicating worst impairment) [16], EQ visual analogue scale (EQ-VAS) (ranging from 0–100 with 100 being the best status), Dizziness Handicap Inventory (DHI) (ranging from 0–100 points (maximum)) [17], and modified Rankin scale (mRS) (ranging from 0–6 points).

Quantitative assessment of vestibular, ocular motor and postural functions

Videoculography (VOG): Vestibular and ocular motor signs were documented by VOG (EyeSeeCam®, EyeSeeTec GmbH, Munich, Germany) during the acute stage of symptoms, including nystagmus in straight ahead position (slow phase velocity (SPV) (°/sec), amplitude (°), horizontal and vertical component, with and without fixation), horizontal vestibulo-ocular reflex (VOR) function by vHIT (gain, presence of refixation saccades), gaze-evoked nystagmus (SPV (°/sec), horizontal and vertical component, lateral and vertical gaze positions), saccades (velocity (°/sec), horizontal and vertical direction), smooth pursuit (gain, horizontal and vertical direction), fixation suppression of the VOR (gain, horizontal direction), and skew deviation (cover test in six gaze positions). VOR gain was rated as pathological for values < 0.7. Suppression of spontaneous nystagmus (SPN) was positive, if the horizontal or vertical component of the SPV decreased by at least 40% on fixation.

Testing of subjective visual vertical (SVV): The SVV was measured by the bucket test method as described previously [18, 19]. Ten repetitions (5 clockwise/ 5 counter clockwise rotations) were performed and a mean of the deviations was calculated. The normal range was defined as $0 \pm 2.5^\circ$ [19].

Posturography: A posturographic measurement of body sway was performed using a mobile device (Wii Balance Board®, Nintendo Co. Ltd., Kyoto, Japan). Four conditions were tested: bipedal standing with eyes open/closed, upright tandem standing with eyes open/closed. For each condition, the sway pattern, normalized path length, root mean square, and peak-to-peak values in medio-lateral and anterior–posterior direction were analyzed.

MRI protocol

The standardized protocol included whole brain and brainstem fine slice (3 mm) diffusion-weighted images (DWI),

whole brain fluid attenuated inversion recovery (FLAIR)- and T2-weighted images including brainstem fine slicing (3 mm), T2*-weighted images, 3D-T1-weighted sequences (FSPGR 1 mm isovoxel) and time-of-flight angiography. All images were evaluated for the presence of ischemic stroke or bleeding by two specialized neuro-radiologists.

Classification methods

We prospectively evaluated two established diagnostic index tests, the HINTS and ABCD² clinical scores for stroke detection, to establish a baseline classification performance. We compared these baselines against the performance of various modern machine-learning techniques. The latter learn the mapping of 305 input features (from history taking, questionnaires, and instrumentation-based examinations) to the output class of stroke vs. peripheral AVS. The classification performance is quantified with three diagnostic test measures [20], namely the area-under-the-curve of a receiver-operating-characteristic (ROC-AUC), accuracy, and F1-score, defined as:

$$\text{Accuracy} = \frac{TP + TN}{N}$$

$$\text{F1 - score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}};$$

$$\text{precision} = \frac{TP}{TP + FP}; \text{recall} = \frac{TP}{TP + FN}$$

Here, TP/TN/FP/FN indicate the number of true-positive/true-negative/false-positive/false-negative detections, respectively, and N indicates the number of test samples overall. The established diagnostic index tests and each of the machine-learning techniques are described briefly in the following.

HINTS: The HINTS clinical scoring system aggregates a risk score for detection of vestibular stroke, as proposed in [9]. HINTS constitutes a 3-step examination, based on Head Impulse, gaze-evoked Nystagmus, and Test of Skew. HINTS indicates a central pattern, if horizontal head impulse test is normal, and/or a direction-changing nystagmus in eccentric gaze, and/or a skew deviation is detected. Consequently, in our data set we give 1 point per central HINTS item and define a HINTS score cutoff value of ≥ 1 as indicative for vestibular stroke. From this binary value for stroke diagnosis, we compute the detection accuracy and F1-score. Additionally, we perform a receiver-operator-characteristic (ROC) analysis, varying the HINTS cutoff over our data set, to obtain an area-under-the-curve (AUC) score.

ABCD²: ABCD² is an aggregative scoring system for clinical detection of stroke as proposed in [21] and validated in [22]. ABCD² is based on the following features:

age ≥ 60 years (1 point); blood pressure $\geq 140/90$ mm Hg (1 point); clinical features: unilateral weakness (2 points), speech impairment without weakness (1 point); duration ≥ 60 min (2 points) or 10–59 min (1 point); and diabetes (1 point). For stroke detection in our study, we consider ABCD² scores at a cutoff value of ≥ 3 . We apply this cutoff to our dataset prospectively, and obtain the accuracy and F1-score, as well as a ROC-AUC score.

Logistic Regression (LR): In descriptive statistics, LR is used to report the goodness-of-fit of a linear set of equations, mapping a set of input features (i.e., observations) to a binary descriptor variable (e.g., stroke indicator variable). In this work, we use LR in a prospective/predictive manner. We regularize LR with a combined L1 and L2 loss, which allows learning of a Lasso-like sparse model, while still maintaining the regularization properties of a ridge classifier [23, 24]. The balancing ratio between the L1 and L2 losses is optimized during learning as a hyper-parameter. After fitting the LR parameters to samples in a training set, we apply the fitted model to samples in a holdout test set, to obtain a logistical posterior probability of stroke. We binarize the soft decision output of LR at a posterior probability $p(\text{stroke}|\text{features}) > 0.5$, from which accuracy and F1-score are calculated. The AUC value is obtained by computing an ROC analysis on the probabilistic predictions for all samples.

Random Forest (RF): RF bundles an ensemble of decision tree (DT) models to compensate for tree overfitting [25] by vote aggregation [26]. In this work, we tune the number of DTs within the range of 5 to 50 trees towards optimal prediction performance. Due to the vote aggregation from the ensemble, an RF yields a probabilistic posterior. Accuracy, F1-score, and ROC-AUC are calculated on this posterior.

Artificial neural network (ANN): Computer-aided diagnosis has advanced due to the application of machine-learning techniques [27]. In particular, our own previous work [28–30], as well as numerous works in related literature [31] have demonstrated the effectiveness and modeling flexibility of ANNs for computer-aided diagnosis in medicine. Here, we apply a multilayer perceptron (MLP) with 305 input neurons, two hidden layers (128 and 64 neurons each), and two softmax-activated output neurons for classification. Due to the non-linear activation at the output layer, our ANN also yields a probabilistic posterior, allowing the calculation of accuracy, F1-score and ROC-AUC.

Geometric matrix completion (GMC): Geometric deep learning [32] is a novel field of deep learning, and has been introduced for computer-aided diagnosis in medicine only recently [33]. In previous work, we have shown that it is advantageous to construct multiple population graphs from meta-features of patients [34, 35]. We further proposed

GMC [36] (denoted in the following as SingleGMC) to alleviate the common problem of missing values in medical data sets [37]. Recently, we have combined these ideas into multi-graph matrix completion (MultiGMC) [38]. Here, we apply both the original SingleGMC approach [36] and MultiGMC to our data set. In SingleGMC, we used a single graph and constructed it using age and ABCD² scores. Graph connections are calculated based on similarity measures using age (age difference ± 5 years) and ABCD² scores (± 1 score). For SingleGMC, the graph connectivity is the sum of these similarity measures. In MultiGMC, instead of taking the sum, we use them as two separate graphs. We learn separate patient representations within these two graphs (a single spectral convolutional layer per graph) and aggregate them via self-attention, before computing the classification posterior [38]. The calculation of accuracy, F1-score, and ROC-AUC is performed as for LR/RF/ANN.

The models LR, RF, and ANN were based on implementations in the scikit-learn machine-learning library [39], while GMC [36] and MultiGMC [38] are custom implementations, based on PyTorch [40].

Statistical analysis

Compared to HINTS and ABCD², which are evaluated prospectively on the entire data set, the training of machine-learning models on the entire data set would result in overfitting and an overly optimistic performance estimate. Instead, we split the data into a training set and a test set, to obtain a prospective classification performance for our investigated models. All machine-learning based classification results were thus obtained following a rigorous ten-fold cross-validation scheme [41], with stratified label sampling to account for class imbalance, and a data split ratio of 90% training vs. 10% testing data. To perform hyper-parameter tuning for all methods, we monitored the tuned model performances on a withheld validation set (10% of the training set). We compared the best-performing model to the other four models, in terms of classification accuracy by pair-wise, two-tailed, non-parametric hypothesis tests (Wilcoxon signed-rank test) at a level of significance $p < 0.05$.

Furthermore, to make the results of the machine-learning classifier more explainable, we used the RF classifier to compute, which features contribute the most towards the detection of stroke. Such analysis constitutes a fundamental technique in the domain of machine-learning interpretability [42]. Feature importance was calculated according to the Mean Decrease in Impurity (MDI) measure [43], as implemented in scikit-learn [39]. We ranked the discriminative power of features by sorting the MDI coefficients, and reported the top 10 most important features utilized by the

RF during classification. For these features, univariate analysis of quantitative values was performed for patients with vestibular stroke and vestibular neuritis (% for categorical variables, mean \pm SD for continuous variables). The parameters were compared between groups using either the Chi-square test or Mann–Whitney *U*-test applying a significance level of $p < 0.05$.

Results

Prospective evaluation of HINTS and ABCD² diagnostic performance

In a prospective analysis, we validated the classification scores of HINTS and ABCD² for detection of all vestibular strokes (AVS and non-AVS presentation) against

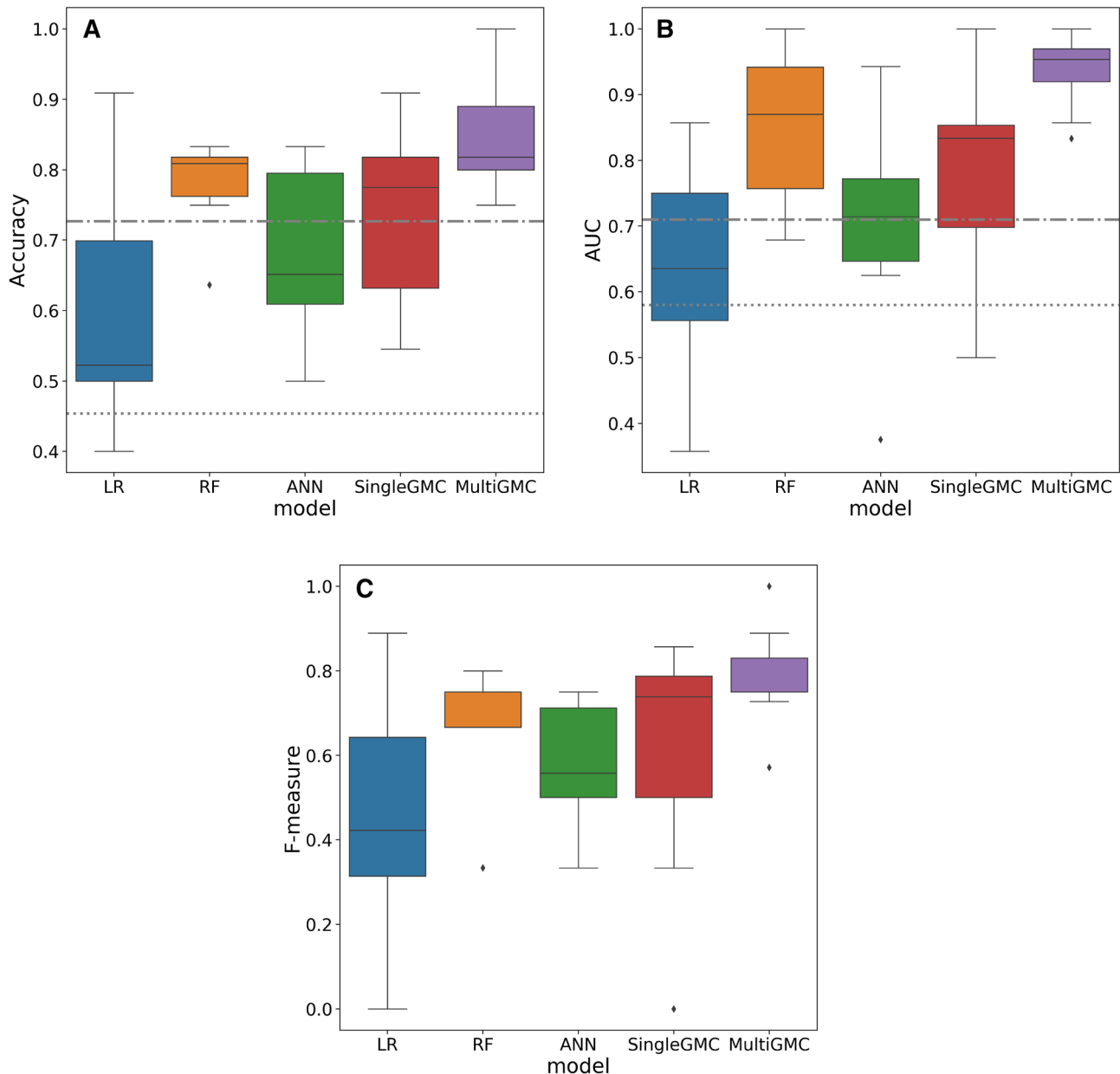


Fig. 1 **a** Accuracy, **b** ROC-AUC, and **c** F1-score (F-measure) of five machine-learning classifiers used in this work (LR: Logistic regression, RF: Random Forest, ANN: Artificial neural network, SingleGMC: Single-graph geometric matrix completion [36], MultiGMC: Multi-graph geometric matrix completion). As a baseline

comparison we additionally indicate HINTS and ABCD² performances (accuracy, ROC-AUC). The prospective validation of univariate clinical scores is illustrated as grey horizontal baselines (HINTS: dash-dotted line, ABCD²: dotted line)

peripheral AVS. In our data set, HINTS was able to detect all strokes with an accuracy of 72.7%, at a ROC-AUC of 0.71. In comparison, ABCD² detected stroke with a lower accuracy of 45.4%, at a ROC-AUC of 0.58. We indicate these univariate baseline methods as dashed horizontal lines in Fig. 1, to which we compare our machine-learning based models. HINTS had a diagnostic accuracy of 82.8%, at a ROC-AUC of 0.86 for stroke with AVS, and a diagnostic accuracy of 66.7%, at a ROC-AUC of 0.54 for stroke without AVS. ABCD² performed with an accuracy of 37.7 (ROC-AUC of 0.59) for stroke with AVS, and 38.6% (ROC-AUC of 0.62) for stroke without AVS.

Machine-learning models for vestibular stroke detection

The median accuracy of all machine-learning methods ranged between 52% (LR) and 82% (MultiGMC). Two models, the linear LR and the non-linear ANN, achieved lower classification accuracy than the univariate measures HINTS and ABCD² for all vestibular strokes vs. peripheral AVS, while RF/SingleGMC/MultiGMC were able to achieve better accuracy (Fig. 1a). Similar results were obtained for AUC and F1-score (Fig. 1b, c). Notably, two methods (RF and MultiGMC) were also able to achieve perfect classification accuracy, F1-score, and

ROC-AUC for one of the five cross-validation folds, while LR and ANN, achieved a zero (0.0) F1-score for one of five folds. In general, MultiGMC yields comparably stable results over all five folds, with consistently high accuracy, F1-score, and ROC-AUC values. Comparing machine-learning classifiers statistically, MultiGMC classifies significantly better than LR ($p < 0.01$), ANN ($p < 0.05$), and SingleGMC ($p < 0.05$), but not significantly better than RF ($p = 0.69$).

Feature importance ranking

We used RF to rank features according to their discriminative performance. The top 10 selected features can be seen in Table 1. Features from device-based measurements such as VOG, SVV testing, and posturography, were considered as single parameters (e.g., vHIT-gain right, vHIT-gain left) or in an aggregated manner (vHIT pathological or normal based on a gain cutoff of 0.7 or presence of refixation saccades). No posturographic or SVV features were selected by the RF classifier as being among the top 10 important features. Instead, two aggregated VOG features (vHIT pathological, presence of horizontal SPN) and eight VOG-based single features were identified (e.g., vHIT gain, gaze-evoked nystagmus left, right).

Quantitative univariate analysis of the 10 most important features revealed significant intergroup differences for all

Table 1 Top 10 most important features, ranked by RF classifier (i.e., ranked by discriminative power for classification) (left side)

Rank in RF	Feature	Feature type	Vestibular neuritis	Vestibular stroke	P value
1	vHIT pathological (gain < 0.7/refixation saccades)	VOG (aggregated)	100%	12.5%*	< 0.0001
2	vHIT gain (right)	VOG (single feature)	0.6 ± 0.3**	0.9 ± 0.3	< 0.0001
3	Fixation suppression of VOR gain (horizontal)	VOG (single feature)	0.03 ± 0.03	0.09 ± 0.06	< 0.0001
4	Smooth pursuit gain (downward direction)	VOG (single feature)	0.75 ± 0.17	0.67 ± 0.2	0.01
5	SPN present without fixation (horizontal)	VOG (aggregated)	95.5%***	47.5%	< 0.0001
6	SPV of SPN (0° position, vertical component)	VOG (single feature)	2.0 ± 2.5°/s	1.0 ± 1.5°/s	0.09
7	SPV of GEN (15° right, horizontal component)	VOG (single feature)	1.2 ± 1.5°/s	0.4 ± 0.6°/s	0.004
8	SPV of SPN (0° position, horizontal component)	VOG (single feature)	4.7 ± 4.0°/s	1.0 ± 1.0°/s	< 0.0001
9	SPV of GEN (15° left, horizontal component)	VOG (single feature)	1.6 ± 2.5°/s	0.3 ± 0.4°/s	0.002
10	STD of SPN amplitude (0° position, horizontal)	VOG (single feature)	2.3 ± 1.4°	1.8 ± 0.8°	0.0005

Quantification of the respective features (as % or mean ± STD) in patients with vestibular neuritis or vestibular stroke and statistical intergroup comparison (Mann–Whitney *U* test for features 2–4 and 6–10, Chi-square test for features 1 and 5) (right side). *GEN* gaze-evoked nystagmus, *SPN* spontaneous nystagmus, *SPV* slow phase velocity, *STD* standard deviation, *VOG* videooculography, *VOR* vestibulo-ocular reflex, *vHIT* video head impulse test; *vHIT was pathological in vestibular stroke lesions affecting the vestibular nucleus or medial longitudinal fascicle; **Gain is depicted for the affected side in vestibular neuritis; ***In three patients without apparent SPN, symptoms of vestibular neuritis had already started ≥ 3 days before VOG recording

but one feature (i.e., rank 6, vertical component of SPN in 0° position, $p=0.09$) (Fig. 1). The following features discriminated best between groups: (1) vHIT was pathologic in 100% of patients with vestibular neuritis (gain: 0.6 ± 0.3 at affected side), but only in 12.5% of patients with vestibular stroke (gain: 0.9 ± 0.3) ($p < 0.0001$). (2) SPN was found more frequently in vestibular neuritis (95.5%) than in vestibular stroke (47.5%), and was more intense (horizontal SPV in 0° position: $4.7 \pm 4.0^\circ/s$ vs. $1.0 \pm 1.0^\circ/s$) ($p < 0.0001$). (3) Fixation suppression of the VOR was abnormal in vestibular stroke (gain: 0.09 ± 0.06), but intact in vestibular neuritis (gain: 0.03 ± 0.03). SPN was suppressed by fixation in 94% of patients with vestibular neuritis. Ranking of feature importance by RF reflected clinically important parameters with significant intergroup differences.

Discussion

Analysis of various approaches for the detection of patients with vestibular stroke (with the clinical presentation of AVS or non-AVS) vs. patients with peripheral AVS due to vestibular neuritis revealed the following findings: HINTS achieves better classification than ABCD² and two of the tested machine-learning methods (LR, ANN), but is not as accurate as the more modern tested machine-learning methods (RF, Single-/MultiGMC) for differentiation of all vestibular strokes against peripheral AVS. In the following, we discuss the methodological and clinical implications of these findings.

Comparison of the different methodological approaches

In the current study, we compared two established clinical classification scores (HINTS, ABCD²) to a number of machine-learning techniques, both classical methods (LR, RF, ANN) and deep learning techniques based on population-modeling with graphs (SingleGMC, MultiGMC). In terms of median accuracy and area-under-the-curve (AUC), all machine-learning classifiers outperformed the detection rate of stroke as indicated by ABCD². Compared to HINTS, however, several machine-learning classifiers performed similarly (LR, ANN, SingleGMC), while only RF and MultiGMC were able to reliably outperform HINTS. For vestibular stroke with AVS, the diagnostic accuracy of HINTS was comparable to MultiGMC. From a methodological perspective, our results provide a reliable estimate of a potential prospective classification performance for future validation studies, due to the usage of a rigorous cross-validation scheme and hyper-parameter optimization of all

machine-learning models. More training data in prospective studies may improve results further, as data set size is usually a limiting factor in machine-learning studies [41]. The RF models yielded satisfactory results, while deep learning models, particularly MultiGMC, were able to improve results further. In general, the possibility to incorporate a semantic population model built from disease-relevant meta-features in form of a graph is attractive from a clinical point of view. The efficacy of this approach in everyday life clinical scenarios needs to be further validated in future studies.

Clinical implications

There is increasing discussion about the use of computer-aided diagnostic support systems in the context of complex clinical scenarios. The differentiation of central and peripheral etiologies of acute vertigo and dizziness poses such a challenge. Established diagnostic algorithms such as HINTS perform very well for AVS, which accounts for about half of acute presentations of vertigo or dizziness [9, 10]. Stroke detection remains particularly difficult, if patients have non-AVS presentations, transient or mild symptoms [3]. Therefore, in the current study we analyzed all vestibular stroke patients (AVS, non-AVS) against peripheral AVS. In our data set, ABCD² had a low diagnostic performance to indicate vestibular stroke and HINTS outperformed ABCD². Nevertheless, for all vestibular stroke patients (AVS, non-AVS), the diagnostic accuracy of HINTS was lower than previously reported for AVS only [9]. Modern machine-learning techniques (such as MultiGMC) had the highest diagnostic accuracy in separating vestibular stroke from peripheral AVS. Interestingly, ranking of feature importance by machine-learning algorithms (such as RF) closely resembled existing clinical experience. The top two features are derived from head impulse testing (vHIT pathologic, vHIT gain). In accordance, HIT has been previously considered the most important component of HINTS with a 18-fold stroke probability if normal in presence of SPN [44]. Two other features (ranks 7, 9) are concerned with gaze-evoked nystagmus, which is also part of HINTS. Skew deviation was not included in the 10 top features, which may be due to its low rate of manifestation (present in only about one quarter of vestibular stroke patients) [45]. Intensity of SPN was weighted prominently (ranks 5, 6, 8, 10). An additional feature with a high importance was a disturbed fixation suppression of the VOR (rank 3). This sign is regularly found in cerebellar lesions involving the uvula, pyramis, nodulus, and flocculus, which are common in patients with vestibular stroke [46, 47]. Notably, all the top-ranked features resulted from VOG examination, while SVV testing and posturography seemed to be less important. It is well-known that SVV deviation is found both in peripheral and

central vestibular lesions, because it reflects a peripheral or central tone imbalance of graviceptive input originating from the vertical semicircular canals and otoliths [48]. The underrepresentation of postural parameters in our data set is in partial contrast to previous clinical studies, which have shown a high diagnostic relevance of the extent of falling tendency in AVS [49]. This discrepancy may be explained by the fact that the overall sway pattern cannot be derived from one or two features, but rather from a complex interplay of parameters [30].

Conclusions

This feasibility study shows the potential of modern machine-learning techniques to support diagnostic decisions in acute vestibular disorders. The current algorithm is tailored for the differentiation of vestibular neuritis vs. vestibular stroke only, and heavily depends on a quantitative and comprehensive assessment of vestibular and ocular motor functions by VOG, which limits its application under everyday life conditions in the ED. Therefore, future studies should focus on tailored VOG-protocols, include other qualitative factors (like triggers, acuity of onset, accompanying symptom features), and test the validity of machine-learning approaches in larger multicenter data sets for a wider range of differential diagnoses, such as Menière's disease and vestibular migraine.

Acknowledgements Open Access funding provided by Projekt DEAL. We thank Katie Göttlinger for copyediting of the manuscript.

Funding The study was supported by the German Federal Ministry of Education and Research (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) (Grant Number 01 EO 0901).

Compliance with ethical standards

Conflicts of interest None of the authors has potential conflicts of interest to be disclosed.


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Newman-Toker DE, Hsieh Y-H, Camargo CA et al (2008) Spectrum of dizziness visits to US emergency departments: cross-sectional analysis from a nationally representative sample. *Mayo Clin Proc* 83:765–775. <https://doi.org/10.4065/83.7.765>
- Royl G, Ploner CJ, Leithner C (2011) Dizziness in the emergency room: diagnoses and misdiagnoses. *Eur Neurol* 66:256–263. <https://doi.org/10.1159/000331046>
- Tarnutzer AA, Lee S-H, Robinson KA et al (2017) ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: a meta-analysis. *Neurology* 88:1468–1477. <https://doi.org/10.1212/WNL.0000000000003814>
- Atzema CL, Grewal K, Lu H et al (2016) Outcomes among patients discharged from the emergency department with a diagnosis of peripheral vertigo: outcomes in patients discharged with peripheral vestibular disorders. *Ann Neurol* 79:32–41. <https://doi.org/10.1002/ana.24521>
- Saber Tehrani AS, Kattah JC, Mantokoudis G et al (2014) Small strokes causing severe vertigo: frequency of false-negative MRIs and nonlacunar mechanisms. *Neurology* 83:169–173. <https://doi.org/10.1212/WNL.0000000000000573>
- Eagles D, Stiell IG, Clement CM et al (2008) International survey of emergency physicians' priorities for clinical decision rules. *Acad Emerg Med* 15:177–182. <https://doi.org/10.1111/j.1553-2712.2008.00035.x>
- Zwergal A, Dieterich M (2020) Vertigo and dizziness in the emergency room. *Curr Opin Neurol* 33:117–125. <https://doi.org/10.1097/WCO.0000000000000769>
- Choi K-D, Kim J-S (2019) Vascular vertigo: updates. *J Neurol* 266:1835–1843. <https://doi.org/10.1007/s00415-018-9040-3>
- Kattah JC, Talkad AV, Wang DZ et al (2009) HINTS to diagnose stroke in the acute vestibular syndrome: three-step bedside oculomotor examination more sensitive than early MRI diffusion-weighted imaging. *Stroke* 40:3504–3510. <https://doi.org/10.1161/STROKEAHA.109.551234>
- Newman-Toker DE, Tehrani ASS, Mantokoudis G et al (2013) Quantitative video-oculography to help diagnose stroke in acute vertigo and dizziness: toward an ECG for the eyes. *Stroke* 44:1158–1161. <https://doi.org/10.1161/STROKEAHA.111.000033>
- Mantokoudis G, Saber Tehrani AS, Wozniak A et al (2015) VOR gain by head impulse video-oculography differentiates acute vestibular neuritis from stroke. *Otol Neurotol* 36:457–465. <https://doi.org/10.1097/MAO.0000000000000638>
- Navi BB, Kamel H, Shah MP et al (2012) Application of the ABCD² score to identify cerebrovascular causes of dizziness in the emergency department. *Stroke* 43:1484–1489. <https://doi.org/10.1161/STROKEAHA.111.646414>
- Newman-Toker DE, Kerber KA, Hsieh Y-H et al (2013) HINTS outperforms ABCD² to screen for stroke in acute continuous vertigo and dizziness. *Acad Emerg Med* 20:986–996. <https://doi.org/10.1111/acem.12223>
- Choi J-H, Oh EH, Park M-G et al (2018) Early MRI-negative posterior circulation stroke presenting as acute dizziness. *J Neurol* 265:2993–3000. <https://doi.org/10.1007/s00415-018-9097-z>
- Möhwald K, Bardins S, Müller H-H et al (2017) Protocol for a prospective interventional trial to develop a diagnostic index test for stroke as a cause of vertigo, dizziness and imbalance in the emergency room (EMVERT study). *BMJ Open* 7:e019073. <https://doi.org/10.1136/bmjopen-2017-019073>
- Herdman M, Gudex C, Lloyd A et al (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 20:1727–1736. <https://doi.org/10.1007/s11136-011-9903-x>

17. Jacobson GP, Newman CW (1990) The development of the dizziness handicap inventory. *Arch Otolaryngol Head Neck Surg* 116:424–427. <https://doi.org/10.1001/archotol.1990.01870040046011>
18. Zwergal A, Rettinger N, Frenzel C et al (2009) A bucket of static vestibular function. *Neurology* 72:1689–1692. <https://doi.org/10.1212/WNL.0b013e3181a55ecf>
19. Dieterich M, Brandt T (2019) Perception of verticality and vestibular disorders of balance and falls. *Front Neurol* 10:172. <https://doi.org/10.3389/fneur.2019.00172>
20. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recogn Lett* 27:861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
21. Johnston SC, Rothwell PM, Nguyen-Huynh MN et al (2007) Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *The Lancet* 369:283–292. [https://doi.org/10.1016/S0140-6736\(07\)60150-0](https://doi.org/10.1016/S0140-6736(07)60150-0)
22. Josephson SA, Sidney S, Pham TN et al (2008) Higher ABCD2 score predicts patients most likely to have true transient ischemic attack. *Stroke* 39:3096–3098. <https://doi.org/10.1161/STROKEAHA.108.514562>
23. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Soft.* <https://doi.org/10.18637/jss.v033.i01>
24. Kim S-J, Koh K, Lustig M et al (2007) An interior-point method for large-scale -regularized least squares. *IEEE J Sel Top Signal Process* 1:606–617. <https://doi.org/10.1109/JSTSP.2007.910971>
25. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*. Springer, New York
26. Criminisi A, Konukoglu E, Shotton J (2011) *Decision forests for classification, regression, density estimation*. *Manifold Learning and Semi-Supervised Learning*, Microsoft Technical Report
27. Yanase J, Triantaphyllou E (2019) A systematic survey of computer-aided diagnosis in medicine: past and present developments. *Expert Syst Appl* 138:112821. <https://doi.org/10.1016/j.eswa.2019.112821>
28. Krafczyk S, Tietze S, Swoboda W et al (2006) Artificial neural network: a new diagnostic posturographic tool for disorders of stance. *Clin Neurophysiol* 117:1692–1698. <https://doi.org/10.1016/j.clinph.2006.04.022>
29. Pradhan C, Wuehr M, Akrami F et al (2015) Automated classification of neurological disorders of gait using spatio-temporal gait parameters. *J Electromyograph Kinesiol* 25:413–422. <https://doi.org/10.1016/j.jelekin.2015.01.004>
30. Ahmadi S-A, Vivar G, Frei J et al (2019) Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway. *J Neurol* 266:108–117. <https://doi.org/10.1007/s00415-019-09458-y>
31. Lin D, Vasilakos AV, Tang Y, Yao Y (2016) Neural networks for computer-aided diagnosis in medicine: a review. *Neurocomputing* 216:700–708. <https://doi.org/10.1016/j.neucom.2016.08.039>
32. Bronstein MM, Bruna J, LeCun Y et al (2017) Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process Mag* 34:18–42. <https://doi.org/10.1109/MSP.2017.2693418>
33. Parisot S, Ktena SI, Ferrante E et al (2018) Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease. *Med Image Anal* 48:117–130. <https://doi.org/10.1016/j.media.2018.06.001>
34. Kazi A, Shekarforoush S, Arvind Krishna S et al (2019) Graph convolution based attention model for personalized disease prediction. In: Shen D, Liu T, Peters TM, et al. (eds) *Medical image computing and computer assisted intervention – MICCAI 2019*. Springer, Cham, pp 122–130
35. Kazi A, Krishna SA, Shekarforoush S et al (2019) Self-attention equipped graph convolutions for disease prediction. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, Venice, Italy, pp 1896–1899
36. Vivar G, Zwergal A, Navab N, Ahmadi S-A (2018) Multi-modal disease classification in incomplete datasets using geometric matrix completion. In: Stoyanov D, Taylor Z, Ferrante E, et al. (eds) *Graphs in biomedical image analysis and integrating medical imaging and non-imaging modalities*. Springer International Publishing, Cham, pp 24–31
37. Little RJ, D’Agostino R, Cohen ML et al (2012) The prevention and treatment of missing data in clinical trials. *N Engl J Med* 367:1355–1360. <https://doi.org/10.1056/NEJMsrl203730>
38. Vivar G, Kazi A, Zwergal A, et al Simultaneous imputation and disease classification in incomplete medical datasets using Multi-graph Geometric Matrix Completion (MGMC). [arXiv:2005.06935 v1 \[cs.LG\]](https://arxiv.org/abs/2005.06935)
39. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
40. Paszke A, Gross S, Massa F, et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, et al (eds) *Advances in neural information processing systems* 32. Curran Associates, Inc., pp 8026–8037
41. Bishop CM (2006) *Pattern recognition and machine learning (information science and statistics)*. Springer, New York
42. Molnar C (2019) *Interpretable machine learning: a guide for making black box models explainable*
43. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and regression trees*, ed. 1. Chapman and Hall/CRC, New York
44. Tarnutzer AA, Berkowitz AL, Robinson KA et al (2011) Does my dizzy patient have a stroke? A systematic review of bedside diagnosis in acute vestibular syndrome. *Can Med Assoc J* 183:E571–E592. <https://doi.org/10.1503/cmaj.100174>
45. Brandt T, Dieterich M (2017) The dizzy patient: don’t forget disorders of the central vestibular system. *Nat Rev Neurol* 13:352–362. <https://doi.org/10.1038/nrneurol.2017.58>
46. Baier B, Stoeter P, Dieterich M (2009) Anatomical correlates of ocular motor deficits in cerebellar lesions. *Brain* 132:2114–2124. <https://doi.org/10.1093/brain/awp165>
47. Zwergal A, Möhwald K, Salazar Lopez E et al (2020) A prospective analysis of lesion-symptom relationships in acute vestibular and ocular motor stroke. *Front Neurol*. (accepted)
48. Glasauer S, Dieterich M, Brandt T (2018) Neuronal network-based mathematical modeling of perceived verticality in acute unilateral vestibular lesions: from nerve to thalamus and cortex. *J Neurol* 265:101–112. <https://doi.org/10.1007/s00415-018-8909-5>
49. Carmona S, Martínez C, Zalazar G et al (2016) The diagnostic accuracy of truncal ataxia and HINTS as cardinal signs for acute vestibular syndrome. *Front Neurol*. <https://doi.org/10.3389/fneur.2016.00125>

Affiliations

Seyed-Ahmad Ahmadi^{1,2} · Gerome Vivar^{1,2} · Nassir Navab² · Ken Möhwald^{1,3} · Andreas Maier^{1,3} · Hristo Hadzhikolev^{1,3} · Thomas Brandt^{1,4} · Eva Grill^{1,5} · Marianne Dieterich^{1,3,6} · Klaus Jahn^{1,7} · Andreas Zwergal^{1,3} 

¹ German Center for Vertigo and Balance Disorders, Ludwig-Maximilians-University, Munich, Germany

² Computer Aided Medical Procedures, Technical University, Munich, Germany

³ Department of Neurology, Ludwig-Maximilians-University, Marchioninistrasse 15, 81377 Munich, Germany

⁴ Clinical Neurosciences, Ludwig-Maximilians-University, Munich, Germany

⁵ Institute for Medical Information Processing, Ludwig-Maximilians-University, Biometry, and Epidemiology, Munich, Germany

⁶ Munich Cluster of Systems Neurology, SyNergy, Munich, Germany

⁷ Department of Neurology, Schön Klinik Bad Aibling, Munich, Germany

5.2.3 Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data (Frontiers in Neurology 2021)

Machine learning with massive datasets has the potential to provide clinical decision support. One such application is in vestibular research. Particularly, the disease classification of vestibular disorders. In the context of ML-based CADx, there exist a large selection of ML/DL methods available one could build for disease classification. This increases the complexity of the data analysis. To this end, we build a Python software package called *base-ml* that combines the use of different open-source python packages for fast-prototyping of baseline models and baseline state-of-the-art approaches in CADx. To evaluate different ML models, we utilize a comprehensive clinical database of patients with symptoms of vertigo and dizziness from the German Center for Vertigo and Balance Disorders (DSGZ).

Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data (Frontiers of Neurology 2021)

Gerome Vivar^{1,2}, Ralf Strobl^{1,3}, Eva Grill^{1,3}, Nassir Navab², Andreas Zwergal^{1,4}, and Seyed-Ahmad Ahmadi^{1,2}

¹German Center for Vertigo and Balance Disorders, Ludwig-Maximilians-University, Munich, Germany

²Computer Aided Medical Procedures, Technical University, Munich, Germany

³Department of Biometry and Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany

⁴Department of Neurology, University Hospital Munich, Ludwig-Maximilians-University, Munich, Germany

Copyright Statement. Copyright © 2021 Vivar, Strobl, Grill, Navab, Zwergal and Ahmadi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). Reprinted, with permission, from Gerome Vivar, Ralf Strobl, Eva Grill, Nassir Navab, Andreas Zwergal, and Seyed-Ahmad Ahmadi. Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data, August 2021. Original publication can be found at: (<https://doi.org/10.3389/fneur.2021.681140>)

Contribution. The core idea, code implementation, design and evaluation of the experimental setup are all part of the author's contribution to this publication.



Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data

Gerome Vivar^{1,2}, Ralf Strobl^{1,3}, Eva Grill^{1,3}, Nassir Navab², Andreas Zwergal^{1,4†} and Seyed-Ahmad Ahmadi^{1,2*†}

¹ German Center for Vertigo and Balance Disorders, University Hospital Munich, Ludwig-Maximilians-University, Munich, Germany, ² Computer Aided Medical Procedures, Department of Informatics, Technical University Munich, Munich, Germany, ³ Department of Biometry and Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany, ⁴ Department of Neurology, University Hospital Munich, Ludwig-Maximilians-University, Munich, Germany

OPEN ACCESS

Edited by:

Carey David Balaban,
University of Pittsburgh, United States

Reviewed by:

Marcos Rossi-Izquierdo,
Lucus Augusti University
Hospital, Spain
Denise Utsch Gonçalves,
Federal University of Minas
Gerais, Brazil
Marty Slade,
Yale University, United States

*Correspondence:

Seyed-Ahmad Ahmadi
ahmadi@cs.tum.edu

† These authors share
senior authorship

Specialty section:

This article was submitted to
Neuro-Otology,
a section of the journal
Frontiers in Neurology

Received: 16 March 2021

Accepted: 30 June 2021

Published: 02 August 2021

Citation:

Vivar G, Strobl R, Grill E, Navab N,
Zwergal A and Ahmadi S-A (2021)
Using Base-ml to Learn Classification
of Common Vestibular Disorders on
DizzyReg Registry Data.
Front. Neurol. 12:681140.
doi: 10.3389/fneur.2021.681140

Background: Multivariable analyses (MVA) and machine learning (ML) applied on large datasets may have a high potential to provide clinical decision support in neuro-otology and reveal further avenues for vestibular research. To this end, we build base-ml, a comprehensive MVA/ML software tool, and applied it to three increasingly difficult clinical objectives in differentiation of common vestibular disorders, using data from a large prospective clinical patient registry (DizzyReg).

Methods: Base-ml features a full MVA/ML pipeline for classification of multimodal patient data, comprising tools for data loading and pre-processing; a stringent scheme for nested and stratified cross-validation including hyper-parameter optimization; a set of 11 classifiers, ranging from commonly used algorithms like logistic regression and random forests, to artificial neural network models, including a graph-based deep learning model which we recently proposed; a multi-faceted evaluation of classification metrics; tools from the domain of “Explainable AI” that illustrate the input distribution and a statistical analysis of the most important features identified by multiple classifiers.

Results: In the first clinical task, classification of the bilateral vestibular failure ($N = 66$) vs. functional dizziness ($N = 346$) was possible with a classification accuracy ranging up to 92.5% (Random Forest). In the second task, primary functional dizziness ($N = 151$) vs. secondary functional dizziness (following an organic vestibular syndrome) ($N = 204$), was classifiable with an accuracy ranging from 56.5 to 64.2% (k-nearest neighbors/logistic regression). The third task compared four episodic disorders, benign paroxysmal positional vertigo ($N = 134$), vestibular paroxysmia ($N = 49$), Menière disease ($N = 142$) and vestibular migraine ($N = 215$). Classification accuracy ranged between 25.9 and 50.4% (Naïve Bayes/Support Vector Machine). Recent (graph-) deep learning models classified well in all three tasks, but not significantly better than more traditional ML methods. Classifiers reliably identified clinically relevant features as most important toward classification.

Conclusion: The three clinical tasks yielded classification results that correlate with the clinical intuition regarding the difficulty of diagnosis. It is favorable to apply an

array of MVA/ML algorithms rather than a single one, to avoid under-estimation of classification accuracy. Base-ml provides a systematic benchmarking of classifiers, with a standardized output of MVA/ML performance on clinical tasks. To alleviate re-implementation efforts, we provide base-ml as an open-source tool for the community.

Keywords: chronic vestibular disorders, classification, machine learning, multivariable statistics, clinical decision support (cdss), episodic vestibular symptoms

INTRODUCTION

Multivariable statistical analysis (MVA), and modern machine learning (ML) methods have the potential to serve as clinical decision support systems (CDSS) (1–3), including the computer-aided diagnosis (CADx) of vestibular disorders (4–8). In combination with large datasets and multi-site cohorts, MVA/ML classification algorithms allow for investigating interactions between patient variables, which is why recent works advocate that these methods should be used more widely in neuro-otology and vestibular neuroscience (9). However, there is a wide variety of MVA/ML methods available, and recent advances in deep learning (DL) with artificial neural networks (ANN) (10) add to the complexity of the field.

In this work, we followed three clinical scenarios in the differential diagnosis of vestibular disorders, and defined three respective classification problems with increasing difficulty. We applied a wide variety of MVA/ML/DL methods to investigate the suitability of automated classification for these clinical questions, and to compare the algorithmic outcomes with clinical expert intuition, both from the perspective of supposed task difficulty, and from the perspective of how the algorithms weighted feature importances toward diagnostic classification. For validation, we took advantage of the DizzyReg dataset, a large prospective registry of patients with vestibular disorders (11). The dataset is multimodal and contains three main categories of variables, namely patient characteristics, symptom characteristics, and quantitative parameters from vestibular function tests.

The first classification problem addresses two groups of patients, suffering either from bilateral damage to peripheral vestibular afferents (i.e., bilateral vestibular failure), or functional dizziness without evidence for relevant structural or functional vestibular deficits. Both syndromes present with the chief complaint of persistent dizziness. However, additional symptom features (e.g., triggers, extent of concomitant anxiety and discomfort) may vary considerably. We expected that machine learning can reliably differentiate both disorders based on patient characteristics (e.g., different age spectra), symptom characteristics, and vestibular function test (e.g., head impulse test or caloric testing).

The second classification task is, whether patients with primary functional dizziness (based on psychological triggers and stressors) can be separated against patients with secondary functional dizziness following a preceding organic vestibular disorder (such as acute unilateral vestibulopathy, or benign paroxysmal positional vertigo) (8). This setting is more complex,

as patient and symptom characteristics may be similar, but the vestibular function tests may differ.

The third problem is directed to the differentiation of four episodic vestibular disorders, namely benign paroxysmal positional vertigo (BPPV), vestibular paroxysmia (VP), Menière disease (MD) and vestibular migraine (VM). This multi-class problem is supposed to be the most complex, because the demographic characteristics of patients and the spectrum of symptoms can be diverse and may overlap (e.g., between MD and VM), and vestibular function tests may be normal (e.g., in VP or VM).

To investigate classification on these three clinical objectives, we developed base-ml, a comprehensive test-bench for initial ML experimentation on clinical data. With this tool, we aim to provide clinical experts with a better intuitive feeling for the range of ML outcomes that can be expected on the given data. For better transparency, several methods can and should be investigated at the same time, subject to a comparable data pre-processing and cross-validation strategy. To this end, we compare several linear, non-linear and neural-network based ML algorithms, along with a novel graph deep learning method that we recently proposed (6, 12, 13). Following insights from multiple classification experiments for diagnostic decision support in our research over the last few years (4, 6, 13, 14), we also provide a multi-faceted analysis of algorithm outcomes, including an examination of class imbalance, multiple classification metrics, patient feature distributions, and feature importances as rated by the classifiers. To alleviate the implementation burden for multi-algorithm comparison and multivariate evaluation, we provide base-ml as an open-source tool¹ to the vestibular research community, as a starting point for further studies in this direction.

MATERIALS AND METHODS

DizzyReg Registry and Dataset

The objective of the DizzyReg patient registry is to provide a basis for epidemiological and clinical research on common and rare vertigo syndromes, to examine determinants of functioning and quality of life of patients, to identify candidate patients for future clinical research, to integrate information of the different apparatus measurements into one data source, and to help understanding the etiology of the vestibular disorders.

The DizzyReg patient registry is an ongoing prospective clinical patient registry which collects all information currently stored in electronic health records and medical discharge

¹Base-ml source code and documentation: <https://github.com/pydsgz/base-ml>.

letters to create a comprehensive clinical database of patient characteristics, symptoms, diagnostic procedures, diagnosis, therapy, and outcomes in patients with vertigo or dizziness (11). Study population includes patients with symptoms of vertigo and dizziness referred to the specialized out-patient center for vertigo and balance disorders. Recruitment into the registry commenced in December 2015 at the German Center for Vertigo and Balance Disorders (DSGZ), Munich University Hospital of the Ludwig-Maximilians-Universität. Inclusion criteria into the registry are symptoms of vertigo and dizziness, age 18 years and above, signed informed consent and sufficient knowledge of German.

Questionnaires were issued on first day of presentation to the study center to assess lifestyle and sociodemographic factors as well as self-reported perception of vertigo symptoms, attack duration and the time since first occurrence. Lifestyle and sociodemographic factors assessed using questionnaires include age, gender, education, physical activity, alcohol, smoking, sleep quality. The type of symptoms of patients included: vertigo, dizziness, postural instability, problems while walking, blurred vision, double vision, impaired vision, nausea, vomiting. Concomitant ontological or neurological symptom are documented with a focus on otological symptoms, i.e., hearing loss, tinnitus, aural fullness, pressure, hyperakusis, and neurological symptoms, i.e., headache, type of headache, photo-/phonophobia, double vision, other symptoms (ataxia, sensory loss, paresis, aphasia).

The evolution of symptoms was reconstructed by the frequency and duration of attacks. All aspects of history taking in the DizzyReg follow established concepts such as “So stoned” (15), the “Five Keys” (16) and the “Eight questions” (17). Frequency or time of onset of symptoms was included as a categorical variable with the following categories: “less than 3 months,” “3 months to 2 years,” “more than 2 years,” “more than 5 years,” and “more than 10 years.” The duration of symptoms is registered in the categories “seconds to minutes,” “minutes to hours,” “hours to days,” “days to weeks,” “weeks to months,” “continuous.”

The registry further collects information on symptoms, quality of life (EQ5D) and functioning (DHI and VAP) in a few standardized questionnaires. Information on triggers is gathered by the respective categories of the Dizziness Handicap Inventory and by elements of the Vertigo Activity and Participation Questionnaire (VAP) (e.g., head movement, position change, physical activity etc).

DHI

The Dizziness Handicap Inventory (DHI) is a well-known and widely used measure to assess self-perceived limitations posed by vertigo and dizziness (18). A total of 25 questions are used to evaluate functional, physical, and emotional aspects of disability. Total score ranging from 0 to 100 is derived from the sum of responses (0 = No, 2 = sometimes, 4 = Yes).

Quality of Life

Health-related quality of life was assessed with the generic EuroQol five-dimensional questionnaire (EQ-5D-3L). This is subdivided into five health state dimensions namely

mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, with each dimension assessed in three levels: no problem, some problem, extreme problems. These health states were converted into EQ5D scores using the German time trade-off scoring algorithm (19). The resulting total EQ5D score ranges from 0 to 1 with higher scores indicating better quality of life.

Vertigo Activity and Participation Questionnaire (VAP)

Functioning and participation were assessed based on the Vertigo Activity and Participation Questionnaire (VAP). The VAP is specifically designed for persons with Vertigo and Dizziness and can be used for people of different countries (20–22). The VAP measures functioning and participation in two scales consisting of six items each. Using weights derived from Rasch analysis the first scale has a range of 0–23 points and the second of 0–20 points with higher scores indicating more restrictions.

Data protection clearance and institutional review board approval has been obtained (Nr. 414-15).

Classification Tasks and Cohorts

As mentioned in the introduction, three classification problems with increasing complexity were tested: (1) bilateral vestibular failure vs. functional dizziness; (2) primary vs. secondary functional dizziness; (3) BPPV vs. VP vs. MD vs. VM. **Table 1** provides information about the group cohorts for each task.

Classification Pipeline

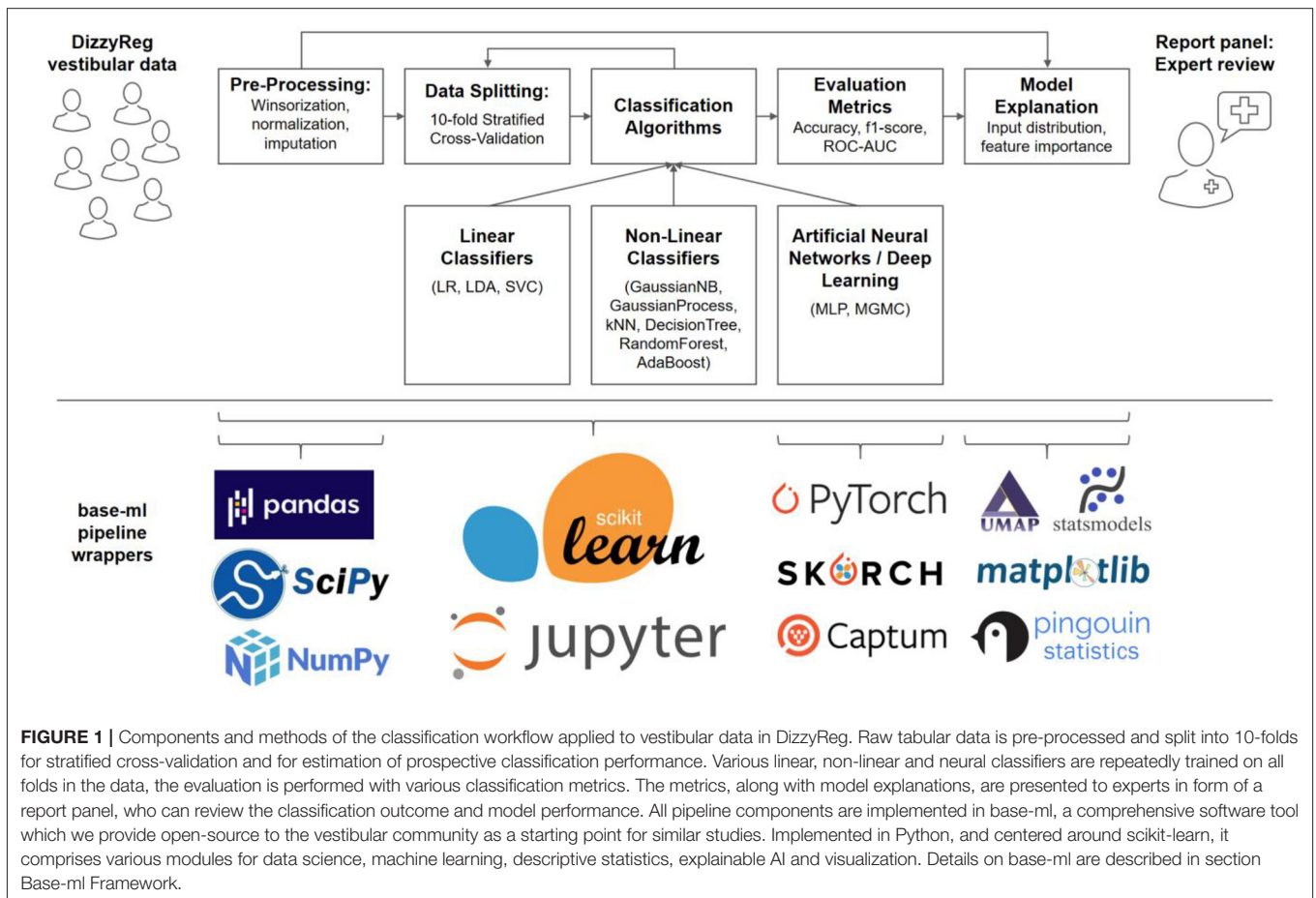
A typical machine learning pipeline comprises several steps that interplay toward a high-accuracy prediction (23). After data import, a set of pre-processing routines are applied to patient features, before data is split into several folds for training and testing, using one or several classification algorithms. The classifier performance is evaluated using several quantitative metrics, and finally presented and explained to a clinical expert on vestibular disorders, for a critical review. **Figure 1** presents an overview of our methodological pipeline in this work.

Pre-processing

Multimodal medical datasets commonly pose several challenges for CADx algorithms, including noisy or missing patient features with spurious outliers (24–26), a mixture of categorical and continuous variables (27), and different statistical distribution of variables (23). To account for outliers and different data ranges in DizzyReg variables with continuous distributions, we perform a 90% winsorization which sets extreme values to the 5th and 95th percentiles, before applying a z-transformation (27) which normalizes all variables into a comparable zero-mean and unit-variance data range. Categorical variables are binarized where possible, or represented in form of a one-hot encoding (a.k.a. one-of-K encoding), which creates a binary column for each category and sparsely represents the categories with a value of 1 in the respective column and 0 in all the other columns. To account for missing values, we perform a mean-imputation (24) if <50% of values are missing in the population, otherwise the feature is omitted from the patient representation.

TABLE 1 | Clinical tasks with respective classes of chronic/episodic vestibular disorders, and respective cohort details.

	Diagnosis abbreviation	N	Age mean (s.d.)	EQ5D	DHI	Female/Male
Task 1						
Bilateral vestibular failure	BVF	66	65.0 (17.0)	0.8 (0.2)	46.2 (22.6)	27/39
Functional dizziness	FD	346	47.2 (14.5)	0.8 (0.2)	43.3 (18.4)	178/168
Task 2						
Functional dizziness (Secondary)	FDS	204	52.1 (14.7)	0.8 (0.2)	48.0 (18.8)	130/74
Functional dizziness (Primary)	FDP	151	45.4 (14.6)	0.8 (0.2)	42.6 (17.6)	77/74
Task 3						
Benign Parox. Pos. Vertigo	BPPV	134	57.0 (12.1)	0.8 (0.2)	45.0 (19.6)	88/46
Menière disease	MM	142	53.4 (13.3)	0.9 (0.2)	43.9 (19.8)	78/64
Vestibular migraine	VM	215	44.5 (14.0)	0.8 (0.2)	41.8 (18.6)	145/70
Vestibular paroxysmia	VP	49	51.6 (14.2)	0.9 (0.2)	38.8 (22.5)	20/29



Data Splitting

In predictive statistics, in particular in the machine learning community, it is common to assess the prediction performance via hold-out test datasets, which are often randomly sampled and kept separate from the training dataset until the time of pseudo-prospective evaluation (27). Sampling a single test set could result in a biased selection and thus in an overly optimistic or pessimistic test evaluation. To avoid this, it is

recommendable to evaluate with multiple test sets, which are sampled either through random shuffling, or through a k-fold splitting. Following common recommendations, we set k to 10 in this work (28). This yields exactly one prediction for each subject in DizzyReg, and exactly ten estimates for the prospective classification performance of each classifier. As recommended by Kohavi in (29), we additionally apply a stratified cross-validation to make sure that each fold has

approximately the same percentage of subjects from each class, which is important especially in the case of class imbalance in the dataset. To ensure that individual classifiers are being trained in a suitable parametrization, we additionally perform hyper-parameter optimization using random search, in a nested cross-validation setup (for details, see section **Appendix C**).

Classification Algorithms and Metrics

Intuitively, ML classifiers try to assign class labels to samples (e.g., patients, represented as multivariable numerical vectors), by fitting separation boundaries between classes in high-dimensional space. Mathematically, these boundaries are expressed in form of a classification function $= f(x)$, which separate the statistical distributions of classes C in the input space X . The past decades of ML research have yielded a diverse set of mathematical models for separation boundaries, and algorithms to fit them to a set of training data X , including linear regression boundaries, rule-based, instance-based, tree-based, kernel-based or Bayesian methods (23), as well as the recent renaissance of artificial neural networks and deep learning (10). Importantly, no single method is guaranteed to perform best on all datasets (30), which is why it is recommendable to test multiple algorithms and let their performances be compared and critically reviewed by a domain expert, instead of deciding on a single algorithm a priori. Therefore, as described in the introduction, we compare several linear, non-linear and neural-network based ML algorithms, along with a novel graph deep learning method that we recently proposed (6, 12, 13). Details on all classifier models and their parametrization are given in section Overview of Selected Classification Algorithms. We quantitatively evaluate the classification performance with three metrics: area-under-the-curve of a receiver-operating-characteristic (ROC-AUC), as well as accuracy and f1-score, defined as (TP/TN/FP/FN denote true or false positives or negatives):

$$\text{Accuracy} = \frac{TP + TN}{N}; \quad \text{f1-score} = \frac{2 \text{Prec} \text{Rec}}{\text{Prec} + \text{Rec}};$$

$$\text{Prec} = \frac{TP}{TP + FP}; \quad \text{Rec} = \frac{TP}{TP + FN}$$

Model Explanation

A necessary tradeoff in predictive statistics and ML is to choose between model accuracy and model interpretability (31). While linear methods like logistic regression are typically more interpretable, non-linear models, depending on their complexity, are often compared to black boxes. By now, however, “Explainable AI” is a dedicated branch in ML research, and numerous model-specific and model-agnostic methods are available that can partially explain ML prediction outcomes (32). Two common ways to explain model performance is to analyze the distribution of input samples (4, 33), and to analyze feature importance (34), especially in a clinical setting (35).

First, we perform a non-linear mapping of the d -dimensional input distribution after pre-processing onto the 2D plane, and we visualize whether class distributions were already visible in the input data, or whether the input data distribution has unexpected or undesired properties, a technique which has been elucidating in our research before, e.g., in the mapping of posturography data

(4). To this end, we utilize “Uniform Manifold Approximation and Projection” (UMAP) (33), a topology-preserving manifold learning technique for visualization and general non-linear dimensionality reduction.

Second, we analyze which patient features contributed to classification outcomes the most, which is a clinically interesting aspect of classifiers. We obtain the “feature importances” for non-ANN-based models and “feature attributions” for ANN-based models. For linear classifiers (see section Linear Classifiers), these can be obtained through the model coefficients (27). For non-linear classifiers (see section Non-linear Classifiers), such as tree-based models, we obtain their feature importance using the Gini-impurity criterion (36). For neural-network based models such as MLP and MGMC (see section Neural Network and Deep Learning Classifiers), we use the Integrated Gradients algorithm (37) and calculate the feature importance by taking the feature attributions of every sample in the training dataset toward their respective ground truth class labels. Obviously, not every classification algorithm yields the same ranking for feature importances. It is argued that a combination of several feature importance rankings can provide more reliable and trustworthy (34). Therefore, for our report to the expert, we aim at presenting a single table with the top 10 most important features for the given classification problem. To merge the feature importance rankings of the different classifiers into a single list, we propose and apply a heuristic for Relative Aggregation of Feature Importance (RAFI), which comprises the following three steps. First, we take the absolute values of all feature importances, to account for algorithms with negative weights (e.g., negative coefficients in linear regression). Second, we normalize the range of importance scores across different classifiers, by computing the percentual importance. Third, we aggregate all normalized global importances by summation, and report the top 10 most important features across all classifiers to the experts for review. In detail, for each feature φ_i ($i \in [1, \dots, d]$), and across F different classifiers, each with feature importances $I_j(\varphi_i)$ ($j \in [1, \dots, F]$), we calculate the global feature importance $I_0(\varphi_i)$ as follows:

$$I_0(\varphi_i) = \sum_{j=1}^F \frac{\text{abs}(I_j(\varphi_i))}{\sum_{i=1}^d \text{abs}(I_j(\varphi_i))}$$

Overview of Selected Classification Algorithms

In this work, we apply and compare the outcomes for a total of 11 classification methods, which we chose to represent a wide range of algorithmic approaches. This collection is larger than what is typically encountered in CDSS research, as mentioned, to provide the expert with a better intuitive feeling for the range of outcomes that can be expected on the given data. The algorithms are grouped into three general categories: linear, non-linear, and ANN-based classifiers. Since explaining the inner workings of all methods in detail is out of scope for this work, each algorithm will be outlined only briefly in the following, with its most important parametrizations (if any), and a reference to explanatory material for the interested reader.

Linear Classifiers

As linear classifiers, we apply *Linear Discriminant Analysis (LDA)*, *Logistic Regression (LR)* and *Support Vector Classifiers (SVC)*. All three methods try to fit a set of linear hyperplanes between the d -dimensional distributions of the classes. *LDA* [(19), chapter 4.3] models the distribution for each class with a Gaussian and calculates the probability of belonging to a class as the maximum posterior probability in a Bayesian manner. We apply LDA in a default parametrization, without additional regularizations such as shrinkage. *LR* [(19), chapter 4.4] directly learns the posterior distribution of the target class and models it using a sigmoid-activated linear function. We apply LR with simple L2 regularization to avoid overfitting the parameters of the model on the training set. *SVC* (38) is a support-vector machine (SVM) with a linear kernel, which learns a hyperplane that maximizes the gap between the classes, giving slack to key samples (“support vectors”) to account for class overlap in the joint distribution. To avoid overfitting, we apply a standard squared L2 penalty term using a regularization parameter of 0.25.

Non-linear Classifiers

Gaussian Naïve Bayes (GNB)

GNB [(19), chapter 6.6.3] is a variant of Naïve Bayes (NB) that allows continuous input features, under the assumption of Gaussian distribution and mutual independence. Class posterior probabilities for new samples are calculated using Bayes Rule. We parametrize *GNB* to estimate class prior probabilities directly from training data, rather than imposing them a-priori.

Gaussian Process Classifier (GP)

GP (39) are a Bayesian alternative to kernel methods like non-linear SVMs. In classification, it models and approximates the class posterior probability as a Gaussian distribution. We set the initial kernel used for GP fitting to a zero-mean, unit-variance radial basis function (RBF), which is then refined during the fitting to training data.

K-Nearest Neighbors Classifier (KNN)

KNN [(19), chapter 2.3.2] classification is an instance-based method, where a sample’s class is determined by the majority class label vote of the sample’s k -nearest neighbors. We compute similarity as Euclidean distance between two patients’ feature vectors, and we use 10 nearest neighbors in the training set to predict the class label of a test input.

Decision Tree Classifier (DT)

DT (36) are a form of rule-based classifiers. A tree represents a hierarchical set of rules or decisions, each decision splitting the feature space in a single feature dimension, using an optimal splitting threshold which is calculated using information-theoretic criteria. Each new sample is passed down the tree, following splitting rules, until a leaf is hit in which a class distribution and majority class is stored. In this work, we use trees with Gini impurity as the splitting criterion, and we allow trees to expand up to a maximum depth of five.

Random Forest Classifier (RF)

RF (40) are an ensemble of multiple decision trees, where each tree is trained using a random subset of training data and a random subset of features. Due to the randomization, the individual trees are highly uncorrelated. Therefore, the ensemble output, which is calculated as an average vote from all trees, weighted by their confidences, is highly robust against various data challenges, such as high dimensional input spaces, noisy data, or highly different data distributions across variables. In this work, we use an ensemble of 10 trees, each with a maximum depth of 5 decision levels.

Adaptive Boosting Classifier (AB)

AB (41), similar to RF, is another ensemble method that combines multiple “weak” classifiers in order to form a much “stronger” classifier. A key difference is the boosting mechanism, i.e., the ensemble is allowed to iteratively add new weak classifiers, which are trained with a higher weight on those input instances that are still being misclassified. In this work, we use decision stumps (i.e., decision trees with a depth of (1) as the weak base classifiers, and we allow the maximum number of classifiers to reach up to 50.

Neural Network and Deep Learning Classifiers

Multi-Layer Perceptron (MLP)

MLP [(19), chapter 11] consider input features as activated neurons followed by one or several fully connected layers (so-called hidden layers) of artificial neurons which weight and sum incoming neuronal connections, before applying a non-linear activation function. The network weights are estimated using the backpropagation algorithm. In this work, we parametrized an ANN with two hidden layers (64 and 32 neurons), and protect every layer against overfitting, as is commonly achieved by applying dropout ($p = 0.3$) (42), followed by batch normalization (43).

Multi-Graph Geometric Matrix Completion (MGMC)

MGMC (13) is a graph-based neural network (GNN) model which we proposed recently, as an extension to our previously published geometric matrix completion approach for multimodal CADx (12). It models the classification problem as a transductive geometric matrix completion problem. Importantly, MGMC is designed to deal with the common problem of missing values in large medical datasets (25), by simultaneously learning an optimal imputation of missing values, along with the optimal classification of patients. MGMC models the patients as nodes in a graph, and computes the edges in the graph through a similarity metric between patients. The similarity is based on a few meta-features (e.g., sex, age, genetic markers etc.), which allows MGMC to span a graph between patients akin to a social network. In previous works, GNNs have shown promising results and a complementary approach in the field of CADx. In this work, we compute multiple patient graphs, each based on similarity measures of a single meta-feature, namely gender (same gender), age (age difference ± 6 years), EQ5D score (score difference of ± 0.06), and DHI score (score difference of ± 11). As advanced model parameters, we use five timesteps for the recurrent graph convolutional network, Chebyshev Polynomials of order five, and

a single hidden layer before the output (16, 32, or 64 neurons, depending on the classification task).

Statistical Methods

The most important features detected by RAFI (cf. section Classification Pipeline) are presented for expert review and interpretation. Each of these features is compared across patient classes via hypothesis tests, to provide a first glance whether there are significant differences across groups. For continuous variables, and in the case of two classes, we first test each variable for normal distribution in each of the patient group with a Shapiro-Wilk test (44). If so, we apply an unpaired two-tailed *t*-test (27), if not, we apply a Mann-Whitney U test (45). For more than two classes, we apply a one-way ANOVA test (27), or a Kruskal-Wallis (46) as an alternative for non-parametric testing, and report the group-level *p*-value. For categorical values, we apply a Chi-squared independence test (47). We report *p*-values for hypothesis tests on all variables, and assume significance at an alpha-level of $p < 0.05$.

Base-ml Framework

As described in the previous sections Classification Pipeline-Statistical Methods numerous methods are necessary to implement a full data science and machine learning pipeline, for a multimodal clinical problem like vestibular classification, and in a multi-site dataset like DizzyReg. Naturally, re-implementing this stack of methods is a time-consuming effort, which should ideally be avoided across research groups. To alleviate future classification experiments similar to this work, and to provide the community with a starting point, we developed base-ml, an open-source Python package¹ provided by the German Center of Vertigo and Balance Disorders. The package can enable a rapid evaluation of machine learning models for prototyping or research. As illustrated in **Figure 1** (lower panel), it is built around scikit-learn (48) as a backbone, which is a reference toolkit for state-of-the-art machine learning and datascience. We complement scikit-learn with various Python modules: *pandas* (49) for data IO and analysis; *scipy* and *numpy* (50) for fast linear algebra on array-shaped data; *PyTorch* (51) for implementation of ANNs and more advanced deep learning models like MGMC; *skorch*² for integration of PyTorch models into the scikit-learn ecosystem; the Captum³ library for model interpretability and understanding, which we use for calculation of feature importance in ANNs using Integrated Gradients (37); UMAP (33) for non-linear 2D mapping and visualization of the patients' input distribution; *statsmodels* (52) and *pingouin* (53), two Python libraries for descriptive statistics and hypothesis testing; and *matplotlib* for plotting and scientific visualization. Importantly, using *skorch*, we enable potential adopters of base-ml to integrate both inductive and transductive neural training workflows and even deep learning models into a comparative benchmark with more traditional ML methods. *Skorch* combines the ease of use of scikit-learn training workflows and *PyTorch*'s

GPU-enabled neural network models. In addition, with base-ml, one can easily evaluate graph-based neural network models.

RESULTS

The following sections reproduce the classification reports produced by base-ml on the three clinical tasks described in the introduction. It is important to note that base-ml is not restricted to vestibular classification scenarios. As a sanity check for base-ml, regarding classification outcomes, and comparability to baseline results in literature, we perform two additional experiments. Those two base-ml experiments are performed on non-vestibular datasets, i.e., one artificially generated dataset, and one Alzheimer's disease classification dataset, which has been widely studied in literature. To keep the main body of this manuscript dedicated to vestibular analysis, we report on non-vestibular results in the **Appendix**.

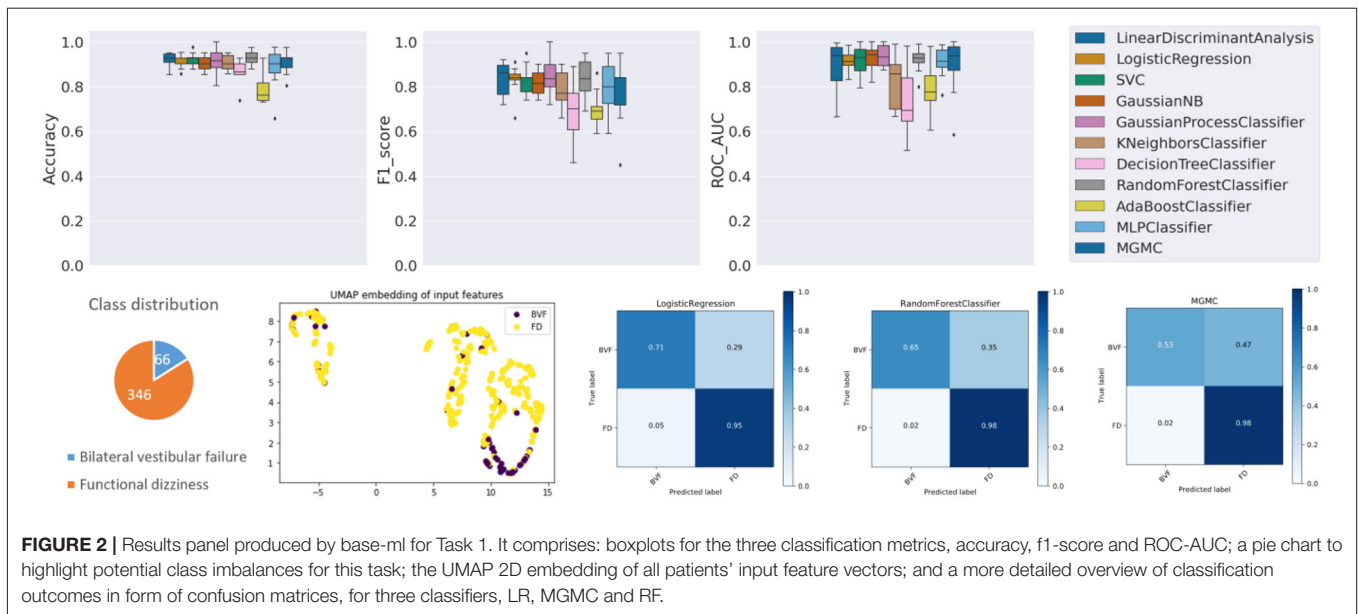
Results on Task 1 (Bilateral Vestibular Failure vs. Functional Dizziness)

The results panel for this classification task, as produced by the base-ml framework, is visible in **Figure 2**. The boxplots with metrics illustrate a wide range of classification performances for all classifiers, with an accuracy over the 10 folds between $78.7\% \pm 6.4\%$ (AdaBoost) and $93.0\% \pm 3.5\%$ (RF), an f1-score between 0.683 ± 0.144 (DecisionTree) and 0.848 ± 0.091 (GaussianProcess), and an average ROC-AUC between 0.727 ± 0.145 (DecisionTree) and 0.937 ± 0.050 (GaussianProcess), followed closely by a ROC-AUC of 0.921 ± 0.056 (RF). Quantitatively, Gaussian Process classifiers are the top-performing model on this task, and slightly outperform the best-performing neural network model MGMC (mean accuracy/f1-score/ROC-AUC: 90.8%/0.782/0.893). In fact, on this task, even one of the best linear models, LR, performs better than MGMC and almost as good as RF (mean accuracy/f1-score/ROC-AUC: 91.3%/0.831/0.917). The confusion matrices reveal that the group with functional dizziness was detected with a very high sensitivity between 95% (LR) and 98% (MGMC/RF), compared to a much lower sensitivity between 53% (MGMC) and 71% (LR) for patients with bilateral vestibular failure. Notably, hyper-parameter optimization had a positive effect on the outcomes of Task 1, and the average accuracy of all classifiers increased from 87.0 to 89.6% after parameter tuning.

Regarding class imbalance, which is important to consider in context with classification performance, the pie chart (cf. **Figure 2**, bottom left) shows that BVF is strongly under-represented in this DizzyReg subset, at 66 vs. 346 patient samples (16.0% of patients). Finally, the UMAP embedding shows that the FV subjects (colored in yellow) are already clustered and topologically separated from the BVF subjects (colored in purple) at the level of normalized input data. This underlines that the patients have clearly separate characteristics at a feature level, and classifiers have a good chance at fitting decision boundaries between the two groups. The UMAP plot reveals another interesting point, namely that the input data is clearly separated into two clusters, the implications of which are discussed below.

²Skorch source code and documentation: <https://github.com/skorch-dev/skorch>

³Captum source code and documentation: <https://github.com/pytorch/captum>



The base-ml output also produces **Table 2**, with feature importance scores aggregated with the RAFI heuristic (cf. section Classification Pipeline). Among the top ten features, six features are related to (Video-) Head Impulse Testing (HIT/vHIT; HIT left/right abnormal, vHIT normal result, vHIT gain left/right) or caloric testing, all of which are also statistically significantly different between the two groups at a level of $p < 0.001$. The most important feature is patient age, also with a significantly different expression between the two groups (63.8 ± 15.6 vs. 47.3 ± 14.1 years, $p < 0.0001$). The remaining three features are related to subjective judgement of disability by patients, namely the depression score in EQ5D ($p < 0.001$), a perceived handicap in DHI ($p < 0.01$), and the actual perceived health condition ($p = 0.133$).

Results on Task 2 (Primary vs. Secondary Functional Dizziness)

Compared to task 1, the performance of the 11 classifiers on task 2 is more homogeneous (cf. **Figure 3**), i.e., all classifiers classify with a within a similar accuracy range between 55.2% (DecisionTree) and 62.8% (GaussianProcess), a f1-score range between 0.498 (MLP) and 0.596 (SVC), and ROC-AUC range between 0.571 (DecisionTree) and 0.689 (SVC). Overall, this classification task is dominated by the linear classification algorithm SVC and the non-linear GaussianProcess classifiers, while the DecisionTree and neural network classifier MLP/ANN are the worst-performing algorithms in terms of accuracy and f1-score. The graph neural network method MGMC and RF had an accuracy of 60.6 and 62.2%, both are close to the average accuracy of all classifiers (60.4%). The confusion matrices reveal that LR and RF have an equally high sensitivity for secondary functional dizziness (77%), compared to MGMC (65%), but a comparably lower sensitivity for primary function dizziness (LR/RF: 42%, MGMC: 54%). Notably, hyper-parameter optimization had very little effect on the outcomes of Task 2, as the average accuracy

of all classifiers stayed at 60.4% both with and without the parameter tuning.

Again, the lower classification performance could partly be due to class imbalance, i.e., a slight underrepresentation of primary functional dizziness in this DizzyReg subset (42.5% primary vs. 57.5% secondary), however the class imbalance is not as severe as in task 1. The UMAP feature embedding shows that after pre-processing, two clearly separated clusters emerge in the topology of the data. Again, the source for this data separation is not clear and will be discussed further below. However, in the smaller cluster, most of patients are from the group with secondary functional dizziness (purple points), while in the larger cluster, there is a mix of both groups, and this mix is not clearly separable by data topology alone. The classification algorithms still can achieve a certain level of data separability in high-dimensional space, but it is noteworthy that the UMAP embedding reflects that task 2 is more challenging compared to task 1, even before the classifiers are applied.

The top 10 most important features for task 2 (cf. **Table 3**) are largely different from task 1. Expectedly, a normal caloric result (rank 1) and the vHIT gain left/right (ranks 4 and 2) and abnormal HIT result on the right (rank 9) differ in both groups. Patients with primary functional dizziness are younger (rank 3) and tend to drink more alcohol (≥ 1 drink in the last week, rank 6). One item from the DHI plays an important role for separation, related to problems turning over while in bed (rank 7), and another life quality factor, LIFEQ Q7, i.e., the actual perceived health condition, is relevant as well (rank 8). The duration of vertigo is important as well, in particular whether the duration is between 20 and 60 min (rank 6). Finally, the depression/fear score in the EQ5D questionnaire is relevant (rank 10). All features except EQ5D fear/depression and LIFEQ Q7 are significantly different between the two groups. It is important to note though that multivariable classifiers do not need to depend on univariate feature significance. In high-dimensional

TABLE 2 | Top 10 most important features in Task 1, aggregated over multiple classifiers.

Rank	Feature	Feature Type	Bilateral vestibular failure	Functional dizziness	P-Value
1	Age (yrs)	Questionnaire	63.83 ± 15.64	47.33 ± 14.12	<0.0001
2	HIT: right, abnormal	Neurological investigation P1	77.40%	3.40%	<0.0001
3	HIT: left, abnormal	Neurological investigation P1	77.40%	2.30%	<0.0001
4	vHIT: normal result	Apparative tests	14.30%	92.20%	<0.0001
5	vHIT: gain left	Apparative tests	0.8 ± 0.04	0.97 ± 0.12	<0.0001
6	EQ5D: fear, depression	Questionnaire	28.60%	66.40%	<0.0001
7	Caloric: normal result	Apparative tests	31.90%	91.80%	<0.0001
8	vHIT: gain right	Apparative tests	0.71 ± 0.09	0.92 ± 0.15	<0.001
9	DHI: Q21, perceived handicap	DHI	81.20%	92.60%	<0.01
10	LIFEQ: Q7, Actual perceived health condition	LIFEQ	62.51 ± 18.48	58.11 ± 18.9	0.133

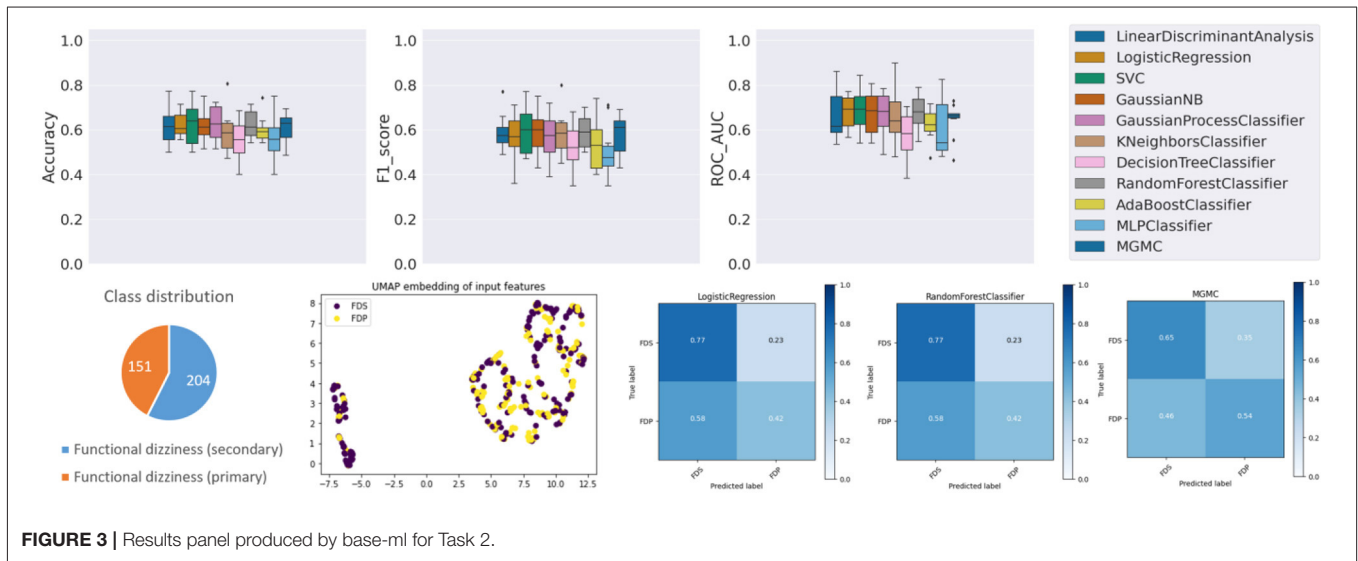


FIGURE 3 | Results panel produced by base-ml for Task 2.

TABLE 3 | Top 10 most important features in Task 2, aggregated over multiple classifiers.

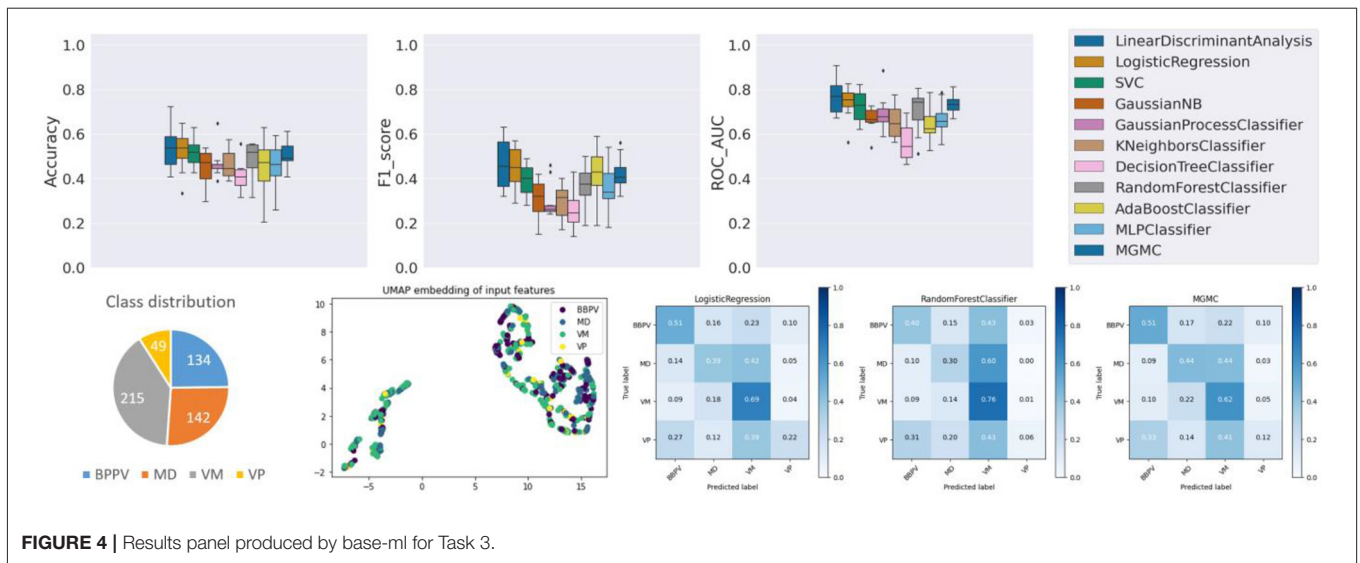
Rank	Feature	Feature type	Functional dizziness (secondary)	Functional dizziness (primary)	P-Value
1	Caloric: normal result	Apparative tests	73.10%	96.20%	<0.0001
2	vHIT: gain right	Apparative tests	0.87 ± 0.18	0.92 ± 0.19	<0.0001
3	Age (yrs)	Questionnaire	51.79 ± 13.91	45.61 ± 14.21	<0.0001
4	vHIT: gain left	Apparative tests	0.92 ± 0.13	0.97 ± 0.12	<0.0001
5	Vertigo time: 20–60 min	Questionnaire	13.20%	5.30%	<0.05
6	>= 1 alcoholic drink last week	Questionnaire	43.60%	58.30%	<0.01
7	DHI: Q13, problems turning over in bed	DHI	43.80%	25.70%	<0.001
8	LIFEQ: Q7, Actual perceived health condition	LIFEQ	57.28 ± 19.61	59.34 ± 18.53	0.111
9	HIT: right, abnormal	Neurological investigation P1	13.70%	1.40%	<0.0005
10	EQ5D: fear, depression	Questionnaire	60.0%	70.0%	0.069

space, these two univariately non-significant features may still contribute to a better separation boundary.

Results on Task 3 (BPPV vs. VP vs. MD vs. VM)

Already at first glance (cf. **Figure 4**), and as clinical intuition suggested, task 3 is the most challenging of the three classification

tasks. Compared to the average classifier accuracy of task 1 (89.6%) and task 2 (60.4%), the accuracy on task 3 is much lower (48.0%). Individually, the classifiers have an accuracy range between 40.6% (DecisionTree) and 54.3% (LDA), a f1-score range between 0.269 (DecisionTree) and 0.461 (LDA), and a ROC-AUC range between 0.564 (DecisionTree) and 0.764 (LDA). Overall on task 3, linear classifiers, and LDA in particular, classify with



the highest accuracy. The RF classifier, on the other hand, only has an average performance on task 3 (accuracy/f1-score/ROC-AUC: 48.5%/0.372/0.702), in comparison to tasks 1 and 2. The confusion matrices reveal that the disorders VM, BPPV, MD and VP can be classified with a decreasing order of classification sensitivity (e.g., for LR approximately: 70%, 50%, 40%, 20%). On task 3, hyper-parameter optimization had a much higher effect on the classifier outcomes than in tasks 1 and 2, i.e., after parameter tuning, the average classification accuracy of all models increased from 44.2 to 48.0%.

Class imbalance probably plays a role here as well, as this ordering almost coincides with the class representation in the dataset (VM: 39.8%, BPPV: 24.8%, MD: 26.3%, VP: 9.1%). Looking at the UMAP embedding, the same separation of the data cloud into two clusters is clearly visible, and the four episodic vestibular disorders are visually not clearly separable within the two clusters, which again anticipates the difficulty of the classification task.

Regarding the 10 most important features (cf. Table 4), mean patient age ranks on the top (BPPV oldest, VM youngest). Second most important is vertigo time <2 min (which is most frequent in BPPV and VP). Expectedly, several features are related to body relocation, e.g., problems getting into, out of, or turning over inside the bed (DHI Q13, rank 3; VAP Q2, rank 4), bending over (DHI Q25, rank 7), or vertical climbing (VAP Q7, rank 10). Accompanying headache is ranked in 6th position and indicative for VM. There is only one apparatusive feature relevant for task 3 (normal caloric test, rank 5), with MD being the only group with relevantly abnormal results.

DISCUSSION

In this paper, we have described several approaches for multivariable analysis and machine learning classification of three different patient cohorts from the vestibular registry dataset DizzyReg, i.e., functional dizziness vs. bilateral vestibular

failure, primary vs. secondary functional dizziness, and BPPV vs. Menière’s disease vs. vestibular migraine vs. vestibular paroxysmia. Clinically, the three tasks were rated with an increasing difficulty and the machine learning classifier performances reflected this grading, with an average accuracy of 87.0, 60.5, and 44.3%, respectively. Using results produced by base-ml, we put these accuracy scores into context with class imbalance, input feature embeddings, confusion matrices and sensitivity scores, as well as tables with the top 10 most important features, aggregated over several classifiers using the proposed RAFI heuristic. In the following, we are going to discuss these results, both from a technical and clinical perspective.

Technical Aspects

The results of the three classification experiments highlight several important points. We believe it to be apparent from the results that it is beneficial to run and benchmark several classification algorithms, ideally from different categories, such as linear, non-linear and neural models. Even a supposedly easy task from a medical perspective does not necessarily lead to a matching classifier performance, depending on which model is used (e.g., 78% classification accuracy in task 1 with Naïve Bayes), hence an a-priori selection could result in too pessimistic an assessment of classification potential using machine learning. Therefore, a wide range of methods in one comprehensive framework might benefit research groups that are new to the field of ML on clinical data. Further, linear models should always be tested along with non-linear and neural network models, as the best linear model (e.g., in task 1, SVC with mean accuracy/f1-score/ROC-AUC: 91.7%/0.819/0.926) may match or even outperform the performance of more complex models, especially if the task has a wide, rather than long data matrix, or if the classes are clearly separable.

Analyzing classifier performance purely using quantitative metrics provides only a narrow view, however. Our analysis reports additionally provide plots on class imbalance, input

TABLE 4 | Top 10 most important features in Task 3, aggregated over multiple classifiers.

Rank	Feature	Feature type	BBPV	MD	VM	VP	P-Value
1	Age (yrs)	Questionnaire	56.6±11.4	53.3±13.0	44.7±13.3	51.6±13.6	<0.0001
2	Vertigo time: < 2 min	Questionnaire	44.80%	12.70%	17.20%	71.40%	<0.0001
3	DHI: Q13, problems turning over in bed	DHI	87.90%	47.50%	44.20%	34.70%	<0.0001
4	VAP: Q2, problems to get in/out/turn over in bed.	VAP	93.30%	68.60%	58.50%	49.00%	<0.0001
5	Caloric: normal result	Apparative tests	85.90%	49.50%	84.80%	100.00%	<0.0001
6	Accompanying headache	Questionnaire	16.80%	19.00%	53.50%	15.00%	<0.0001
7	DHI: Q25, bending over increases problems	DHI	76.10%	60.30%	61.20%	61.20%	<0.05
8	DHI: Q6, restricted participation in social activities	DHI	71.40%	82.90%	75.50%	65.30%	<0.05
9	DHI: Q22, increased stress on family/friend relationships	DHI	23.10%	48.90%	45.60%	38.80%	<0.0001
10	VAP: Q7, Vertical climbing (stairs/lift)	VAP	60.00%	64.90%	62.00%	45.70%	0.139

data distribution, and confusion matrices, all of which provide different insights into the experiment. Class representation in the dataset correlated with the sensitivity for each class in all three experiments, which the confusion matrices highlighted. The input data distribution additionally revealed that DizzyReg data in our study had a fundamental separation into two clusters (cf. UMAP embeddings in **Figures 2–4**). At least in task 1 this did not affect classification outcomes to match the clinical intuition, however, for future ML-based studies, this separation would need to be investigated further. Counteracting such a data separation, e.g., with input data transforms (54), or more advanced techniques like domain adaptation (55), could improve classification results further. As such, the results obtained through the base-ml tool provide not only information about which machine learning models to pursue further, but they also indicate starting points regarding the optimization of the input data with classical data science and statistical methods. For clinicians, an important part of the results are the most important features selected by the classifiers, which we present in an aggregated form using the proposed RAFI heuristic. These features will be discussed in more detail and put into a clinical context in section Clinical Implications.

The method presented in this work, and comprised in the base-ml tool have several noteworthy limitations. In general, base-ml is intended as a first screening tool for ML experiments, rather than as a complete ML solution that leads to a trained model for prospective studies and/or deployment. It has been shown previously that hyper-parameter optimization using nested cross-validation can lead to significant improvements of classification performance (6, 12, 13). In our study, while hyper-parameter tuning had no noticeable effects on Task 2, there were noticeable improvements in the average classification outcomes across all models in Tasks 1 and 3. Further, not only the models themselves have hyper-parameters, but every part of the ML pipeline in base-ml could be individually optimized further. This could include alternative input normalization strategies [e.g., power transforms (54, 56)] and imputation methods [e.g., kNN imputation or multiple imputation by chained equations, MICE (57, 58)] or the inclusion of feature selection methods (e.g., based on univariate hypothesis testing), all of which are important toward optimal classifier performance (9). A default

treatment made in our experiments, for example, is to discard variables that were recorded for <50% of the population. In clinical practice, however, some variables may be missing because the according examinations or apparative tests were not ordered by the physician, maybe due to time, cost, lack of indication, or expected inefficacy toward diagnosis. In that case, individual rules for variable rejection, imputation and/or normalization may be necessary. For base-ml, we chose to avoid such in-depth treatment, in favor of an ease-of-use at the exploratory stage. However, base-ml is built on top of scikit-learn and already provides an interface to modern deep learning methods with skorch, and explainable AI solutions through Captum. This makes it easy to include many further methods for feature selection, imputation and normalization, as well as further classification explainable AI algorithms (32). However, at a certain level of complexity that aims at deployment rather than exploration, it is recommendable to consider more in-depth analyses and tool, ideally in close collaboration with data science and ML experts, and potentially starting off from insights obtained with base-ml. A particularly interesting avenue is the current research direction of Automated Machine Learning (AutoML), which aims at an optimization of the entire classification pipeline end-to-end (59). Importantly though, small to medium-size datasets might not provide enough data samples to train such complex pipelines. Until more cross-institutional vestibular registry datasets like DizzyReg come to existence, and with sufficient data to apply AutoML, the methods which we wrapped in base-ml and presented in this work still provide a solid starting point for ML-based analysis. As such, and for the time being, we believe these tools to be a valuable contribution for the vestibular research community.

Clinical Implications

Clinical reasoning in the diagnostic differentiation of common vestibular disorders is based on a “mental aggregation” of information from patient characteristics (such as age and gender), symptom characteristics (namely quality, duration, triggers, accompanying symptoms), clinical examination (e.g., positioning maneuvers), and quantitative tests of vestibular function (such as vHIT, calorics) (16). It is an open and relevant question, whether ML-based methods are able to identify features

from a multimodal vestibular patient registry, which resemble this clinical thinking and feature weighting. In the current study, we tested three clinical scenarios of different complexity on the DizzyReg database to further address this issue.

The first classification task represented two groups of patients suffering from chronic dizziness of almost diametrical etiology. In bilateral vestibular failure, imbalance can be directly assigned to an organic damage of vestibular afferents, which is accompanied by a low degree of balance-related anxiety (60, 61), while in functional dizziness the vestibular system is physiologically intact, but the subjective perception of balance is severely disturbed due to fearful introspection (62). It can be expected that ML-based algorithms will predominantly select features as most important for the segregation of both disorders, which represent either measurements of vestibular function or scales for anxiety and perceived disability. Indeed, the top 10 important features exactly meet this assumption with six of them reflecting low and high frequency function of the vestibular-ocular reflex (HIT left/right normal, vHIT gain left/right, bilateral vHIT normal, caloric response normal), and further three features healthy-related quality of life, depression and fear. Furthermore, age was an important differential feature, which is in good accordance to the fact that bilateral vestibular failure appears more frequently in older patients and functional dizziness in younger and mid-aged patients.

In the second classification task, two groups of patients with functional dizziness were compared, who were presumably very similar in their symptomatic presentation, but differed in the evolution of their symptoms: patients with primary functional dizziness, where chronic psychological stress or anxiety is the driving force, and patients with secondary functional dizziness, which develops after a preceding somatic vestibular disorders (e.g., BPPV) due to altered balance perception and strategies of postural control (8). Accordingly, top 10 features for classification included vestibular function tests (such as vHIT gain left/right and caloric response normal). The subtle differences between groups may speak for a partially recovered acute unilateral vestibulopathy or MD as some causes underlying secondary functional dizziness. Furthermore, symptom provocation by position changes in bed may point to BPPV as another vestibular disorder triggering secondary functional dizziness. This findings agree with previous literature (8). Interestingly, patients with primary functional dizziness had higher fear and depression scales, which may indicate a more intense psychological symptom burden. Indeed, previous studies have shown a psychiatric comorbidity in primary functional dizziness in 75 vs. 42% in secondary functional dizziness (63). The more frequent consumption of alcohol in primary functional dizziness may also show that those patients subjectively profit from its relaxing effects to a higher extent than patients with secondary functional dizziness, who have some degree of vestibular deficits, which may exacerbate on alcohol (e.g., partially compensated unilateral vestibulopathy or vestibular migraine).

The third classification task was designed to differentiate common episodic vestibular disorders like BPPV, MD, vestibular migraine and vestibular paroxysmia. Expectedly, a set of features

was most indicative for BPPV, namely short attack duration and provocation by position changes. MD as compared to the other vestibular disorders was associated with the highest rate of pathological vestibular function tests (caloric test abnormal). It is well-known that long-standing MD can cause vestibular function deficits (64), while this is less frequent in vestibular migraine (65). The latter was associated with the highest frequency of headache and the youngest mean patient age, in accordance to literature (66). Vestibular paroxysmia was mostly defined by a short-symptom duration. The overall moderate accuracy for classification of the four episodic vestibular disorders can be explained by several factors: (i) one methodological explanation could be that this was a multi-class task, which is more challenging; (ii) despite the exhaustive history taking and examination details for patients recorded in DizzyReg, it is possible that not all relevant information is included. For example, systematic audiological test results are only available for patients with Menière's disease and vestibular migraine, but not for BPPV or vestibular paroxysmia. Therefore, audiological test results could not be generally included in the third classification task as a variable; (iii) there are potential overlaps of symptom characteristics and features. A prominent example is an overlap syndrome of MD and vestibular migraine, which could point toward a common pathophysiology (67); (iv) although the guidelines "International Classification of Vestibular Disorders (ICVD)" of the Barany Society give clear criteria for diagnosis mostly based on history taking, complex clinical constellations such as overlapping syndromes or atypical presentations appear regularly in the practice of a tertiary referral center, which may cause some difficulties in clear-cut classification. Limited classification accuracy may be partly explained by this selection bias, and further testing in primary care settings will be needed; (v) given the difficulty of task 3, the low ML classification performance is neither surprising nor a sign of a failure of ML classification approaches. Instead, our results suggest that ML algorithms, even given considerable data to learn from, may not automatically be able to solve difficult clinical tasks. The wide range of tuned ML algorithm performances presented by base-ml can reveal such difficulty better than a narrow selection of ML results without tuning; (vi) previous studies suggest that expert consensus may not always be unanimous, and may indicate the difficulty of patient diagnosis, despite clear guidelines and diagnostic criteria. For example, authors in (68) tried to validate diagnostic classifications through multi-rater agreement between several experienced otoneurological raters, and an acceptable consensus was achieved only in 62% of the patients. This study indicates that some diagnostic inaccuracy persists in the clinical setting, despite established international classification criteria. This could be taken as a further argument to augment clinical decision making by ML-based support systems.

CONCLUSION

Analysis of large multimodal datasets by novel ML/MVA-methods may contribute to clinical decision making

in neuro-otology. Important features for classification can be identified and aligned with expert experience and diagnostic guidelines. The optimal ML/MVA-method depends on the classification task and data structure. Base-ml provides an innovative open source toolbox to test different methods and clinical tasks in parallel. The multi-faceted presentation of results and explainable AI features, including an identification of clinically relevant features and their statistical analysis, enables clinicians to better understand ML/MVA outcomes, and identify avenues for further investigation. Future research needs to be extended to larger multicenter datasets and new data sources to improve the performance of automated diagnostic support tools.

DATA AVAILABILITY STATEMENT

The data analyzed in this study was obtained from the DSGZ DizzyReg, the following licenses/restrictions apply: The DSGZ provides application forms that must be completed before the data in the DizzyReg may be accessed. Please contact the DSGZ for more details on the application process. Requests to access these datasets should be directed to Ralf Strobl, ralf.strobl@med.uni-muenchen.de.

REFERENCES

- Dagliati A, Tibollo V, Sacchi L, Malovini A, Limongelli I, Gabetta M, et al. Big data as a driver for clinical decision support systems: a learning health systems perspective. *Front Digit Humanit.* (2018) 5:8. doi: 10.3389/fdigh.2018.00008
- Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data.* (2019) 6:54. doi: 10.1186/s40537-019-0217-0
- Gamache R, Kharrazi H, Weiner J. Public and population health informatics: the bridging of big data to benefit communities. *Yearb Med Inform.* (2018) 27:199–206. doi: 10.1055/s-0038-1667081
- Ahmadi S-A, Vivar G, Frei J, Nowoshilow S, Bardins S, Brandt T, et al. Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway. *J Neurol.* (2019) 266:108–17. doi: 10.1007/s00415-019-09458-y
- Pradhan C, Wuehr M, Akrami F, Neuhaeuser M, Huth S, Brandt T, et al. Automated classification of neurological disorders of gait using spatio-temporal gait parameters. *J Electromyogr Kinesiol.* (2015) 25:413–22. doi: 10.1016/j.jelekin.2015.01.004
- Ahmadi S-A, Vivar G, Navab N, Möhwald K, Maier A, Hadzhikolev H, et al. Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders. *J Neurol.* (2020) 267:143–52. doi: 10.1007/s00415-020-09931-z
- Groezinger M, Huppert D, Strobl R, Grill E. Development and validation of a classification algorithm to diagnose and differentiate spontaneous episodic vertigo syndromes: results from the DizzyReg patient registry. *J Neurol.* (2020) 267:160–7. doi: 10.1007/s00415-020-10061-9
- Habs M, Strobl R, Grill E, Dieterich M, Becker-Bense S. Primary or secondary chronic functional dizziness: does it make a difference? A DizzyReg study in 356 patients. *J Neurol.* (2020) 267:212–22. doi: 10.1007/s00415-020-10150-9
- Smith PE, Zheng Y. Applications of multivariate statistical and data mining analyses to the search for biomarkers of sensorineural hearing loss, tinnitus, and vestibular dysfunction. *Front Neurol.* (2021) 12:627294. doi: 10.3389/fneur.2021.627294
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board University Hospital Munich Ludwig Maximilian University Munich, Germany. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GV, AZ, RS, and S-AA contributed to conception and design of the study and wrote the first draft of the manuscript. NN and EG contributed to study refinement. RS and GV organized the database. GV and S-AA developed base-ml and performed the data and statistical analyses. S-AA, AZ, NN, and EG provided funding. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the German Federal Ministry of Education and Research (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) (grant number 01 EO 0901).

- Grill E, Müller T, Becker-Bense S, Gürkov R, Heinen F, Huppert D, et al. DizzyReg: the prospective patient registry of the German center for vertigo and balance disorders. *J Neurol.* (2017) 264:34–6. doi: 10.1007/s00415-017-8438-7
- Vivar G, Zwergal A, Navab N, Ahmadi S-A. Multi-modal disease classification in incomplete datasets using geometric matrix completion. In: Stoyanov D, Taylor Z, Ferrante E, Dalca AV, Martel A, Maier-Hein L, et al. editors. *Graphs in Biomedical Image Analysis Integrating Medical Imaging Non-Imaging Modalities*. Cham: Springer International Publishing (2018). p. 24–31.
- Vivar G, Kazi A, Burwinkel H, Zwergal A, Navab N, Ahmadi S-A. Simultaneous imputation and classification using multigraph geometric matrix completion (MGMC): application to neurodegenerative disease classification. *Artif Intell Med.* (2021) 117:102097. doi: 10.1016/j.artmed.2021.102097
- Vivar G, Mullakaeva K, Zwergal A, Navab N, Ahmadi S-A. Peri-diagnostic decision support through cost-efficient feature acquisition at test-time. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al. editors. *Medical Image Computing Computer Assisted Intervention – MICCAI 2020 Lecture Notes in Computer Science*. Cham: Springer International Publishing (2020). p. 572–81.
- Wuyts FL, Van Rompaey V, Maes LK. “SO STONED”: common sense approach of the dizzy patient. *Front Surg.* (2016) 3:32. doi: 10.3389/fsurg.2016.00032
- Brandt T, Strupp M, Dieterich M. Five keys for diagnosing most vertigo, dizziness, and imbalance syndromes: an expert opinion. *J Neurol.* (2014) 261:229–31. doi: 10.1007/s00415-013-7190-x
- Strobl R, Grözinger M, Zwergal A, Huppert D, Filippopoulos F, Grill E. A set of eight key questions helps to classify common vestibular disorders—results from the DizzyReg patient registry. *Front Neurol.* (2021) 12:670944. doi: 10.3389/fneur.2021.670944
- Jacobson GP, Newman CW. The development of the dizziness handicap inventory. *Archiv Otolaryngol Head Neck Surg.* (1990) 116:424–7. doi: 10.1001/archotol.1990.01870040046011
- Greiner W, Weijnen T, Nieuwenhuizen M, Oppe S, Badia X, Busschbach J, et al. A single European currency for EQ-5D health states. *Eur J Health Eco.* (2003) 4:222–31. doi: 10.1007/s10198-003-0182-5

20. Alghwiri AA, Whitney SL, Baker CE, Sparto PJ, Marchetti GF, Rogers JC, et al. The development and validation of the vestibular activities and participation measure. *Archiv Phys Med Rehabil.* (2012) 93:1822–31. doi: 10.1016/j.apmr.2012.03.017
21. Grill E, Furman JM, Alghwiri AA, Müller M, Whitney SL. Using core sets of the international classification of functioning, disability and health (ICF) to measure disability in vestibular disorders: study protocol. *J Vestib Res.* (2013) 23:297–303. doi: 10.3233/VES-130487
22. Mueller M, Whitney SL, Alghwiri A, Alshehber K, Strobl R, Alghadir A, et al. Subscales of the vestibular activities and participation questionnaire could be applied across cultures. *J Clin Epidemiol.* (2015) 68:211–9. doi: 10.1016/j.jclinepi.2014.10.004
23. Bishop CM. *Pattern Recognition and Machine Learning.* New York, NY: Springer (2006).
24. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med.* (2010) 50:105–15. doi: 10.1016/j.artmed.2010.05.002
25. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med.* (2012) 367:1355–60. doi: 10.1056/NEJMsr1203730
26. Pesonen E, Eskelinen M, Juhola M. Treatment of missing data values in a neural network based decision support system for acute abdominal pain. *Artif Intell Med.* (1998) 13:139–46. doi: 10.1016/S0933-3657(98)00027-X
27. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY: Springer New York (2009).
28. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* (2005) 21:3301–7. doi: 10.1093/bioinformatics/bti499
29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2.* Montreal, QC (1995). p. 1137–43.
30. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Computat.* (1997) 1:67–82. doi: 10.1109/4235.585893
31. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci.* (2001) 16:199–231. doi: 10.1214/ss/1009213726
32. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* (2019). Available online at: <https://christophm.github.io/interpretable-ml-book/> (accessed July 11, 2021).
33. McInnes L, Healy J, Melville J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* (2020). Available online at: <http://arxiv.org/abs/1802.03426> (accessed March 2, 2021).
34. Saarela M, Jauhainen S. Comparison of feature importance measures as explanations for classification models. *SN Appl Sci.* (2021) 3:272. doi: 10.1007/s42452-021-04148-9
35. Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. *Stud Health Technol Inform.* (2010) 160:861–5.
36. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* 1th ed. Boca Raton, FL: Chapman and Hall/CRC (1984).
37. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70.* Sydney, NSW: ICML'17. p. 3319–28 (2017).
38. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018
39. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning.* Cambridge: MIT Press (2008).
40. Criminisi A, Konukoglu E, Shotton J. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Micro Tech Rep.* (2011). doi: 10.1561/9781601985415
41. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comp Syst Sci.* (1997) 55:119–39. doi: 10.1006/jcss.1997.1504
42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* (2014) 15:1929–58. Available online at: <https://dl.acm.org/doi/10.5555/2627435.2670313>
43. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift (2015). Available online at: <http://arxiv.org/abs/1502.03167> (accessed March 3, 2021).
44. Shapiro SS, Wilk MB. An analysis of variance test for normality (Complete Samples). *Biometrika.* (1965) 52:591. doi: 10.1093/biomet/52.3-4.591
45. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist.* (1947) 18:50–60. doi: 10.1214/aoms/1177730491
46. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Statist Assoc.* (1952) 47:583–621. doi: 10.1080/01621459.1952.10483441
47. Cressie N, Read TRC. Multinomial goodness-Of-Fit tests. *J Royal Statist Soc Series.* (1984) 46:440–64. doi: 10.1111/j.2517-6161.1984.tb01318.x
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* (2011) 12:2825–30. Available online at: <https://dl.acm.org/doi/10.5555/1953048.2078195>
49. Reback J, McKinney W, Jbrockmendel, Bossche JVD, Augspurger T, Cloud P, et al. *Pandas-dev/pandas: Pandas 1.2.3.* Zenodo. (2021). Available online at: <https://zenodo.org/record/3509134>
50. SciPy 1.0 Contributors, Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* (2020) 17:261–72. doi: 10.1038/s41592-019-0686-2
51. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems 32.* Vancouver, BC: Curran Associates, Inc. p 8026–37.
52. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: *9th Python in Science Conference.* Austin, TX (2010).
53. Vallat R. Pingouin: statistics in Python. *JOSS.* (2018) 3:1026. doi: 10.21105/joss.01026
54. Box GEP, Cox DR. An analysis of transformations. *J Royal Statist Soc Series B.* (1964) 26:211–43. doi: 10.1111/j.2517-6161.1964.tb00553.x
55. Wilson G, Cook DJ. A survey of unsupervised deep domain adaptation. *ACM Trans Intell Syst Technol.* (2020) 11:1–46. doi: 10.1145/3400066
56. Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika.* (2000) 87:954–9. doi: 10.1093/biomet/87.4.954
57. Mandel JSP. A comparison of six methods for missing data imputation. *J Biom Biostat.* (2015) 06:1–6. doi: 10.4172/2155-6180.1000224
58. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Multiple imputation by chained equations. *Int J Methods Psychiatr Res.* (2011) 20:40–9. doi: 10.1002/mpr.329
59. He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. *Knowledge Based Syst.* (2021) 212:106622. doi: 10.1016/j.knsys.2020.106622
60. Strupp M, Kim J-S, Murofushi T, Straumann D, Jen JC, Rosengren SM, et al. Bilateral vestibulopathy: diagnostic criteria consensus document of the classification committee of the bárány society. *VES.* (2017) 27:177–89. doi: 10.3233/VES-170619
61. Decker J, Limburg K, Henningsen P, Lahmann C, Brandt T, Dieterich M. Intact vestibular function is relevant for anxiety related to vertigo. *J Neurol.* (2019) 266:89–92. doi: 10.1007/s00415-019-09351-8
62. Dieterich M, Staab JP. Functional dizziness: from phobic postural vertigo and chronic subjective dizziness to persistent postural-perceptual dizziness. *Curr Opin Neurol.* (2017) 30:107–13. doi: 10.1097/WCO.0000000000000417
63. Lahmann C, Henningsen P, Brandt T, Strupp M, Jahn K, Dieterich M, et al. Psychiatric comorbidity and psychosocial impairment among patients with vertigo and dizziness. *J Neurol Neurosurg Psychiatry.* (2015) 86:302–8. doi: 10.1136/jnnp-2014-307601
64. Huppert D, Strupp M, Brandt T. Long-term course of Ménière's disease revisited. *Acta Oto Laryngol.* (2010) 130:644–51. doi: 10.3109/00016480903382808
65. Radtke A, von Brevern M, Neuhauser H, Hottenrott T, Lempert T. Vestibular migraine: long-term follow-up of clinical symptoms and vestibulo-cochlear findings. *Neurology.* (2012) 79:1607–14. doi: 10.1212/WNL.0b013e31826e264f
66. Lempert T, Olesen J, Furman J, Waterston J, Seemungal B, Carey J, et al. Vestibular migraine: diagnostic criteria. *J Vest Res.* (2012) 22:167–72. doi: 10.3233/VES-2012-0453

67. Lopez-Escamez JA, Długaiczek J, Jacobs J, Lempert T, Teggi R, von Brevern M, et al. Accompanying symptoms overlap during attacks in menieres disease and vestibular migraine. *Front Neurol.* (2014) 5:265. doi: 10.3389/fneur.2014.00265
68. Soto-Varela A, Arán-González I, López-Escámez JA, Morera-Pérez C, Oliva-Domínguez M, Pérez-Fernández N, et al. Peripheral vertigo classification of the otoneurology committee of the spanish otorhinolaryngology society: diagnostic agreement and update (Version 2-2011). *Acta Otorrinolaringol.* (2012) 63:125–31. doi: 10.1016/j.otoeng.2012.03.011
69. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, Weiner MW, et al. TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease (2018). Available online at: <http://arxiv.org/abs/1805.03909> (accessed June 18, 2021).
70. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage.* (2013) 65:167–75. doi: 10.1016/j.neuroimage.2012.09.065

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Vivar, Strobl, Grill, Navab, Zwergal and Ahmadi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

Appendix A. Supplementary Experiments on TADPOLE Dataset

Data Description

TADOLE (69) is an ADNI-based dataset consisting of imaging-derived features and non-imaging features. The task is to classify whether observations at a baseline timepoint are from healthy normal controls (NC), patients with mild cognitive impairment (MCI), and Alzheimer’s disease (AD). It consists of 813 instances (229 NC, 396 MCI, and 188 AD). Imaging features are computed using standard ADNI feature extraction pipelines.

Results and Discussion

We evaluated all models on this dataset as supplementary experiment to understand the strengths and limitations of our proposed model. For our purposes we only look at the F1-score, as this metric is more robust to class imbalance, which is present in TADPOLE. We observe that the best performing models are the hyper-parameter-optimized tree-based models such as Random Forest and AdaBoostClassifier. Furthermore, neural network based models such as MLPClassifier and MGMC yield comparable results but do not outperform other models. We also observe from the confusion matrices that the biggest source of error in most models is to distinguish patients with diagnosed AD from patients with MCI. Likewise, the confusion matrices reveal that models almost never mistake healthy controls with AD patients and vice-versa. Overall almost all models perform comparably, except notable mis-classification rates in KNeighborClassifier and GaussianProcessClassifier. Our obtained classification results of ~0.6–0.7 F1-score are in line with recent literature, e.g., our previous comparison of MGMC

with regular machine learning classifiers [cf. results in (13), not yet computed with base-ml], or RF-based AD classification by Gray et al. (70).

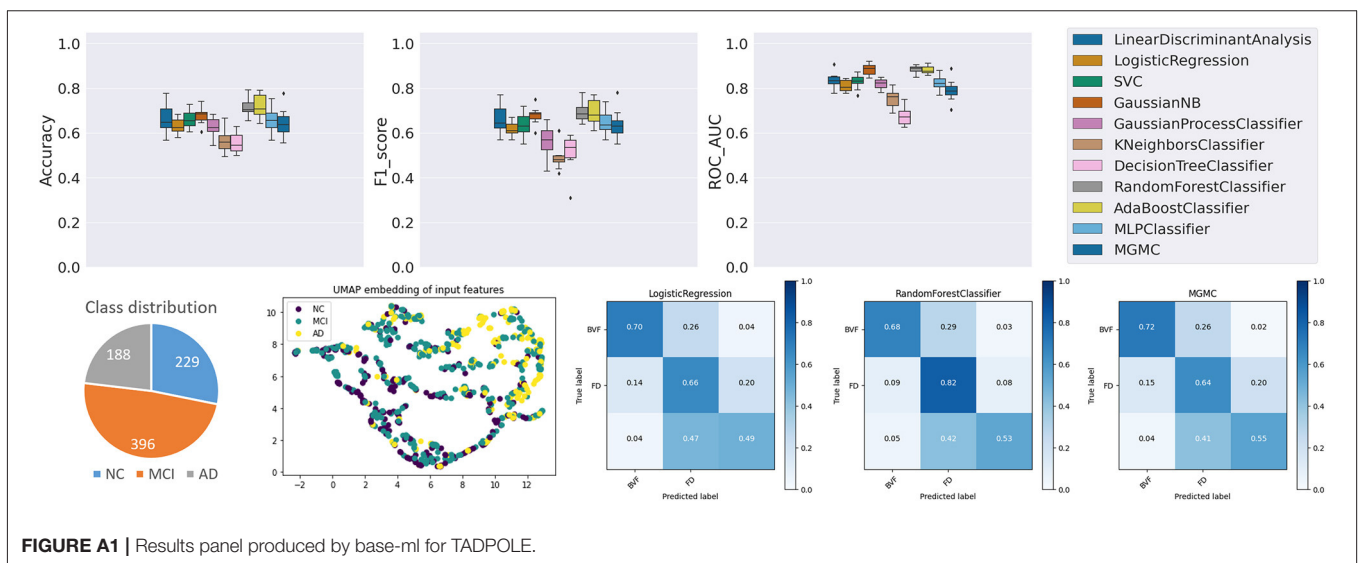
Appendix B. Supplementary Experiments on Generated Dataset

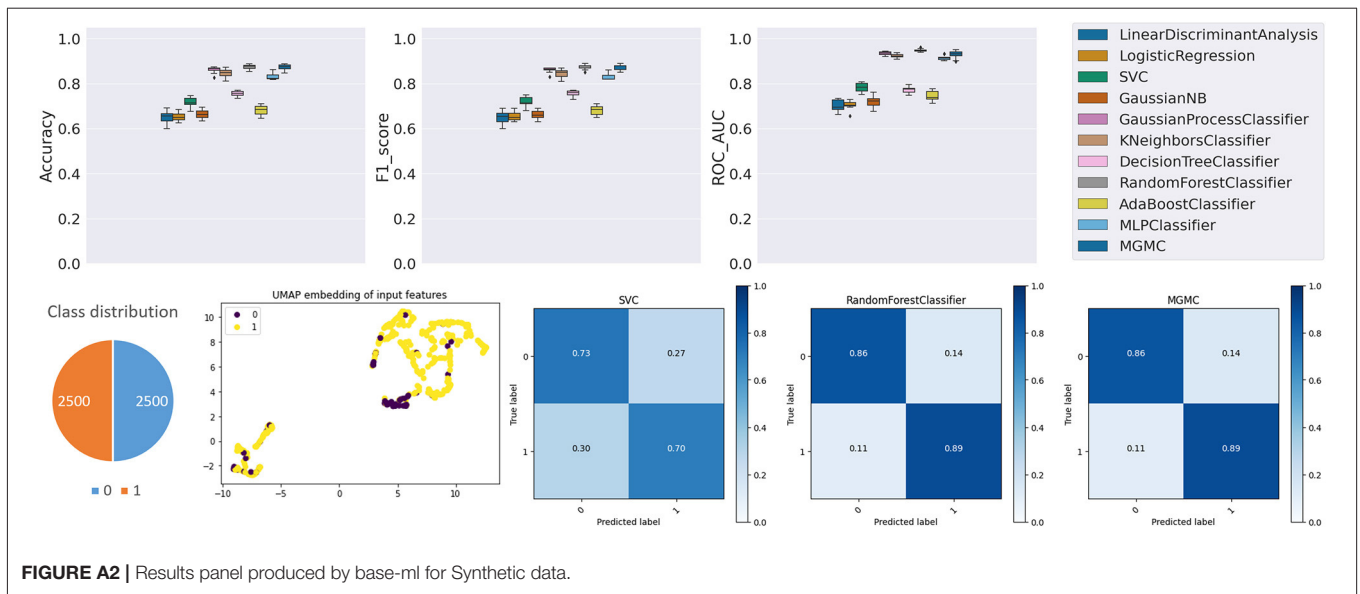
Data Description

To further illustrate the utility of base-ml, we created a synthetic dataset for a binary classification task. We generated 5,000 samples with 20 features of which 10 features are informative and the remaining 10 are uninformative using Scikit-learn (48) (using the built-in function <make_classification>). It is important to note that by design, this classification task has a non-linear separation boundary between the two classes and can therefore not be solved with high accuracy by linear classifier models.

Results and Discussion

As can be seen in **Figure A2**, most non-linear models based on neural networks and properly tuned tree-based models such as Random Forest could yield comparable performance. When looking at the classification accuracy of both MGMC and Random Forest, both perform nearly identically, and with the highest accuracies among all models. As expected, the linear models such as Logistic Regression and Linear Discriminant Analysis obtained the lowest classification accuracy. Overall, we observe that base-ml properly reflects the statistical properties and the difficulty of this artificial classification problem. The source data distributions are not simply separable by topology mapping (see UMAP embedding), and the separation is only resolvable by selected and properly tuned non-linear models – this characteristic would not have been detected by an analysis that was limited to linear models, or less suited non-linear models (e.g., for this dataset: Decision Tree Classifier or AdaBoost Classifier).





Appendix C. Implementation Details: Hyperparameter Search Ranges

To have a more comparable analysis, we selected the best hyperparameters using the validation set, before reporting performance metrics on a with-held test-set (nested cross-validation). We do this by randomly searching the hyperparameter space for 100 iterations for every model and select the best hyperparameters which yields the best validation set classification performance.

For Logistic Regression we used the following hyperparameters (C: randint(1, 11); penalty: {"elasticnet"}, solver: {"saga"}, l1-ratio: uniform(0, 1));

Random Forest (max_depth: {3, None}; max_features: randint(1, 11); min_sample_split: randint(2, 11); bootstrap: {True, False}; criterion: {"gini", "entropy"}, n_estimators: randint(5, 50));

K-Neighbors Classifier (n_neighbors: randint(3, 100); weights: {"uniform", "distance"});

SVC (C: log_uniform(1e-6, 1e+6); gamma log_uniform(1e-6, 1e+6); degree: randint(1, 8), kernel: {"linear", "poly", "rbf"});

Decision trees (max_depth: {3, None}; max_features: randint(1, 11); min_samples_split: randin(2, 11); criterion: {"gini", "entropy"});

Gaussian Process Classifier (kernel: {1*RBF(), 1*DotProduct(), 1*Matern(), 1*RationalQuadratic(), 1*WhiteKernel()});

AdaBoostClassifier: (n_estimators: {50, 60, 70, 80, 90, 100}, learning-rate: {0.5, 0.8, 1.0, 1.3});

GaussianNB (var_smoothing: logspace(0, 9, num=100)); Linear Discriminant Analysis (solver: {"svd", "lsqr", "eigen"}; shrinkage: numpy.arange(0, 1, 0.01));

MLP Classifier (learning-rate: {1e-1, 1e-2, 1e-3, 1e-4}; hidden-units: {32, 64, 128}, dropout probability: {0.0, 0.1, 0.2, 0.3, 0.4, 0.5});

MGMC ([cross-entropy, Frobenius-norm, Dirichlet-norm weighting]: uniform(0.001, 1000)).

Part III

Conclusion and Outlook

Conclusion & Outlook

Contents

6.1	Summary of Findings	105
6.1.1	Addressing incomplete clinical data for CADx	105
6.1.2	Cost-efficient peri-diagnostic support	106
6.1.3	Translation of CADx to clinical research	106
6.2	Outlook	107
6.2.1	Data missingness in CADx	107
6.2.2	Peri-diagnostic decision support	107
6.2.3	Translation of ML or DL methods for CADx in clinical research	108

6.1 Summary of Findings

The main goal of this thesis was to develop and apply machine-learning-based diagnostic decision support algorithms for healthcare providers, to deliver quality healthcare at the right time. We outlined why this thesis is necessary in the first place, i.e. because of its potential to improve and reduce diagnostic errors, to enable access to quality care for resource-limited areas, and to enable holistic decision support based on multi-modal clinical data. We have addressed important challenges one would typically face when handling real-world healthcare data for machine learning-based CADx such as data missingness. We also proposed a novel CADx paradigm, namely to provide diagnostic decision support throughout the whole diagnostic workflow which we call peri-diagnostic decision support. Finally, we took several efforts towards translation of shallow and deep learning based models into clinical research.

6.1.1 Addressing incomplete clinical data for CADx

To enable CADx for clinical datasets with incomplete information, we proposed to simultaneously impute the missing information and classify the target disease label. We approached this by using RGCN and showed that non-autoregressive RGCN with multi-graph approach are more robust than their autoregressive counterparts in terms of classification and imputation performance. The proposed model improved in accuracy and significantly outperformed all other approaches in terms of imputation performance. Eventually, this means that we have a streamlined approach which enables CADx even in settings where we have datasets with considerable amounts of missing information. Providing more informative information to the learning algorithm typically would be much more advantageous as the model could generalize more to unseen datasets. One limitation of this proposed approach is that all data points must be available during training, meaning it is transductive. Additionally, models

are often highly dependent on the graph which could make or break the learning algorithm. Defining the graph connectivity is challenging, especially if there is no domain knowledge about the problem. As a result, as a future research direction, one could look at automating the graph construction by learning it automatically in an end-to-end manner. Other works could also target at introducing an inductive learning-based GMC which can allow training in a mini-batch manner, without requiring fully feature-complete samples in the dataset during training.

6.1.2 Cost-efficient peri-diagnostic support

ML models often assume that all the dataset features are available, and that their acquisition was for free. This means the learning algorithm does not take into account that certain observations could be expensive for the patient or the healthcare provider. Such a system could be very inefficient. We argue that most ML models assume feature-completeness at training time, and research towards cost-efficient models has not been a priority so far. However, in the context of CADx it is desirable to have a system which is cost-efficient. This means we only acquire an examination if it is needed, i.e. if its acquisition is expected to result in a better classification performance. Most CADx models only provide decision support to the clinician once all examinations are already available. Again, this makes the clinical diagnostic workflow inefficient to both the patient and healthcare provider. Our goal in this part of the thesis was to have a system which is trained with feature-incompleteness and cost-efficiency in mind, and which can provide diagnostic decision support to the clinician at test time. In order to do this, we propose Accumulated Integrated Gradients (AIG), which is a derivation of the Integrated Gradients (IG) feature attribution method. AIG improved in performance when compared to previous methods. We also showed that AIG is both cost- and feature-efficient. This implies a potential to reduce financial burden to the clinic and healthcare system, as well as to reduce unwanted examinations for the patient. As a result, hospital resources could also be optimized. One limitation of the current approach is that it is mainly targeted to feature-level acquisition of examination. Future works could address datasets which contain features that are in a block or grouped together as a set (e.g. grouped by modality).

6.1.3 Translation of CADx to clinical research

The penultimate goal of modern CADx methods is their translation into clinical routine. Translation of recent advancements in CADx using deep learning is a step closer to this goal. With the utilization of shallow and deep-learning-based models, we can answer diagnostic decision support for CADx. We showed that this could be successfully applied to CADx for neurological disorders in clinical research of neurological movement and balance disorders. We explored shallow- and deep-learning-based models for disease diagnosis using a simple python package called *base-ml*, which we open-sourced to enable easy experimentation of disease classification problems using non-DNN based models and DNN-based models, including more advanced models such as recently proposed GNN methods (MGMC). We enable this through the utilization of multiple open-sourced python packages used in ML/DL. We showed that it is possible to answer CADx questions which are important to the diagnosing clinician, such as providing diagnostic decision support or providing model explanations or

interpretations. One limitation of the current proposed approaches is that it mainly focuses on neurological disorders, and future work should continue translating current state-of-the-art CADx approaches to datasets from various clinical domains. As demonstrated in our works, the evaluation should be made in close collaboration with domain-specific clinical experts.

6.2 Outlook

Although the solutions and findings presented here are promising, there are certainly a number of open questions one could address, including potential research directions which continue the research questions or objectives in this thesis.

6.2.1 Data missingness in CADx

The problem of data missingness in clinical data should certainly be given attention when coming up with a CADx system. Data-driven diagnostic decision support could provide more accurate diagnostic support to clinicians. One approach to utilize such data for CADx is using ML or DL. However, many recent advancements in ML or DL for CADx do not explore settings where data contains missing information. Pre-prepared (e.g. challenge) datasets are often prepared in a way that suggests to data scientists that clinical data is usually feature-complete.

In this work, we address data missingness in the area of feature level missingness. A more realistic setting in multi-modal CADx using DL is having feature-level missingness where missingness could come from a single feature of observation of a patient or from a set of features that belongs to a modality of observation. This means missingness at the modality-level where blocks of features could be not available for analysis. We recommend to explore scenarios where a combination of feature- and modality-level missingness is present in the data, which would be much more representative of real-world clinical datasets. A solution towards this is certainly of great importance.

6.2.2 Peri-diagnostic decision support

CADx approaches using ML/DL often provide support when all data is available at the end of the acquisition phase. In other words, it is assumed that the phase where diagnostic data is acquired on the patient has already been concluded. However, often clinicians would also need support even during the acquisition phase. The AIG approach proposed in this work addresses active feature acquisition in a feature-level setting. An important extension of this is to enable feature acquisition at the modality-level. Here, a set of features coming from a modality could be acquired at a single step instead of breaking these into multiple levels of the acquisition step. Enabling peri-diagnostic decision support at the feature- and modality-level could address even more real-world clinical data.

Another aspect that could also be given attention is the notion of online learning. This goes in the direction of the field of reinforcement learning (RL). Here, the system is allowed

to explore and exploit the environment during model training. This means potential features/observations other than the current finite set of features could also be exploited by the system. This area could also allow one to propose other forms of dynamic cost functions into the system other than just scaling the feature attributions with the given cost. Such an approach could be possible by assigning appropriate rewards to the features of interest.

Additionally, the proposed AIG method could be improved and explored even further. One aspect of this is to extensively explore other forms of attribution methods, instead of calculating feature attributions only. In a sense, the fact that our currently proposed AIG method requires model differentiability to calculate gradients could be considered a limitation. Instead, future research could include works on model-agnostic explainability methods [48]. This could further enable peri-diagnostic decision support, even when the models are no longer differentiable or not even neural-network based.

6.2.3 Translation of ML or DL methods for CADx in clinical research

The clinical translation of CADx approaches is certainly very important, to make sure that the innovation and new techniques reach the everyday diagnostic routine of doctors. In this work, we analyzed the efficacy of ML/DL approaches for CADx in clinical research, in particular related to neurological diseases such as vestibular disorders. Although the applications evaluated in this thesis are specific to neurological disorders, these approaches could also be applied to other disease diagnosis problems.

Another aspect that also needs more attention from the CADx community is the creation of benchmark datasets for disease diagnosis. Several different solutions have been proposed in the literature and reproducing such results is not always straightforward. Although we use publicly available datasets in this work to enable reproducibility, it could still be challenging for other researchers to have access to these different datasets as there is no single point of access. Having a single point of access where different benchmark datasets for CADx are accessible could unify efforts in CADx algorithm development. This could also avoid CADx researchers from building the same solutions just applied to other disease classification or domain. Such a solution could potentially improve reproducibility in CADx research as well as result to more accurate CADx solutions.

Bibliography

- [1] M. Abadi, A. Agarwal, P. Barham, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015 (cit. on p. 8).
- [2] S.-A. Ahmadi, G. Vivar, J. Frei, et al. “Towards computerized diagnosis of neurological stance disorders: data mining and machine learning of posturography and sway”. In: *Journal of Neurology* (2019), pp. 1–10 (cit. on p. 70).
- [3] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai. “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective”. In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–9 (cit. on p. 22).
- [4] C. L. Atzema, K. Grewal, H. Lu, M. K. Kapral, G. S. Kulkarni, and P. C. Austin. “Outcomes among patients discharged from the emergency department with a diagnosis of peripheral vertigo”. In: *Annals of Neurology* 79 (2016) (cit. on p. 70).
- [5] E. P. Balogh, B. T. Miller, and J. R. Ball. “Improving diagnosis in health care”. In: (2015) (cit. on pp. 7, 53, 55).
- [6] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, and k. kavukcuoglu. “Interaction Networks for Learning about Objects, Relations and Physics”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016 (cit. on p. 18).
- [7] P. W. Battaglia, J. B. Hamrick, V. Bapst, et al. “Relational inductive biases, deep learning, and graph networks”. In: *ArXiv abs/1806.01261* (2018) (cit. on pp. 16, 18).
- [8] L. Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 14).
- [9] M. M. Bronstein, J. Bruna, T. Cohen, and P. Velivckovi’c. “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges”. In: *ArXiv abs/2104.13478* (2021) (cit. on pp. 17, 18).
- [10] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric deep learning: going beyond euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42 (cit. on pp. 14, 16).
- [11] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. “Spectral Networks and Locally Connected Networks on Graphs”. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2014 (cit. on p. 16).
- [12] S. v. Buuren and K. Groothuis-Oudshoorn. “MICE: Multivariate Imputation by Chained Equations in R”. In: *Journal of statistical software* (2010), pp. 1–68 (cit. on p. 29).
- [13] E. Candes and B. Recht. “Exact low-rank matrix completion via convex optimization”. In: *46th Annual Allerton Conference on Communication, Control, and Computing* (2008), pp. 1–49. arXiv: arXiv:0805.4471v1 (cit. on p. 23).

- [14] E. J. Candes and Y. Plan. “Matrix completion with noise”. In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936. arXiv: 0903.3131 (cit. on p. 24).
- [15] E. J. Candès and B. Recht. “Exact matrix completion via convex optimization”. In: *Commun. ACM* 55 (2012), pp. 111–119 (cit. on pp. 23, 24, 29).
- [16] L. Cosmo, A. Kazi, S.-A. Ahmadi, N. Navab, and M. Bronstein. “Latent-graph learning for disease prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 643–653 (cit. on p. 20).
- [17] K. Crammer and Y. Singer. “On the algorithmic implementation of multiclass kernel-based vector machines”. In: *Journal of machine learning research* 2.Dec (2001), pp. 265–292 (cit. on p. 14).
- [18] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik. “Big data in healthcare: management, analysis and future prospects”. In: *Journal of Big Data* 6.1 (2019), pp. 1–25 (cit. on pp. 5–8).
- [19] M. Defferrard, X. Bresson, and P. Vandergheynst. “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett. 2016, pp. 3837–3845 (cit. on pp. 16, 17).
- [20] T. Emmanuel, T. M. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. “A survey on missing data in machine learning”. In: *Journal of Big Data* 8 (2021) (cit. on pp. 8, 23, 27–29).
- [21] A. Esteva, B. Kuprel, R. A. Novoa, et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118 (cit. on p. 5).
- [22] M. Garbi. “National Institute for Health and Care Excellence clinical guidelines development principles and processes”. In: *Heart* 107.12 (2021), pp. 949–953 (cit. on p. 53).
- [23] Z. Ghahramani. “Unsupervised learning”. In: *Summer school on machine learning*. Springer. 2003, pp. 72–112 (cit. on p. 22).
- [24] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. “Neural message passing for quantum chemistry”. In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272 (cit. on p. 18).
- [25] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. “Transduction with Matrix Completion: Three Birds with One Stone”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2010, pp. 757–765 (cit. on p. 30).
- [26] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016 (cit. on pp. 14, 15, 20, 21).
- [27] H. A. Haenssle, C. Fink, R. Schneiderbauer, et al. “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists”. In: *Annals of oncology* 29.8 (2018), pp. 1836–1842 (cit. on p. 5).
- [28] W. L. Hamilton, R. Ying, and J. Leskovec. “Inductive representation learning on large graphs”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 1025–1035 (cit. on p. 17).
- [29] T. Heart, O. Ben-Assuli, and I. A. Shabtai. “A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy”. In: *Health policy and technology* 6 (2017), pp. 20–25 (cit. on p. 55).
- [30] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 15).

- [31] J. Janisch, T. Pevný, and V. Lisý. “Classification with Costly Features Using Deep Reinforcement Learning”. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 3959–3966 (cit. on p. 56).
- [32] L. Jing and Y. Tian. “Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), pp. 4037–4058 (cit. on p. 22).
- [33] M. Kachuee, S. Darabi, B. Moatamed, and M. Sarrafzadeh. “Dynamic feature acquisition using denoising autoencoders”. In: *IEEE transactions on neural networks and learning systems* 30.8 (2018), pp. 2252–2262 (cit. on p. 56).
- [34] M. Kachuee, O. Goldstein, K. Kärkkäinen, S. Darabi, and M. Sarrafzadeh. “Opportunistic Learning: Budgeted Cost-Sensitive Learning from Data Streams”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019 (cit. on p. 56).
- [35] M. Kachuee, A. Hosseini, B. Moatamed, S. Darabi, and M. Sarrafzadeh. “Context-aware feature query to improve the prediction performance”. In: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2017, pp. 838–842 (cit. on p. 56).
- [36] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst. “Matrix completion on graphs”. In: *arXiv:1408.1717* (2014) (cit. on p. 24).
- [37] J. C. Kattah. “Use of HINTS in the acute vestibular syndrome. An Overview”. In: *Stroke and Vascular Neurology* 3.4 (2018), pp. 190–196. eprint: <https://svn.bmj.com/content/3/4/190.full.pdf> (cit. on p. 72).
- [38] J. C. Kattah, A. V. Talkad, D. Wang, Y.-H. Hsieh, and D. E. Newman-Toker. “HINTS to Diagnose Stroke in the Acute Vestibular Syndrome: Three-Step Bedside Oculomotor Examination More Sensitive Than Early MRI Diffusion-Weighted Imaging”. In: *Stroke* 40 (2009), pp. 3504–3510 (cit. on p. 72).
- [39] J. C. Kattah, A. V. Talkad, D. Z. Wang, Y.-H. Hsieh, and D. E. Newman-Toker. “HINTS to diagnose stroke in the acute vestibular syndrome: three-step bedside oculomotor examination more sensitive than early MRI diffusion-weighted imaging”. In: *Stroke* 40.11 (2009), pp. 3504–3510 (cit. on p. 72).
- [40] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015 (cit. on p. 21).
- [41] T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017 (cit. on pp. 16, 17).
- [42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 15).
- [43] K. Limburg, A. Dinkel, G. Schmid-Mühlbauer, et al. “Neurologists’ Assessment of Mental Comorbidity in Patients With Vertigo and Dizziness in Routine Clinical Care—Comparison With a Structured Clinical Interview”. In: *Frontiers in Neurology* 9 (2018) (cit. on p. 70).
- [44] L. Maier-Hein, M. Eisenmann, D. Sarikaya, et al. “Surgical data science - from concepts toward clinical translation”. In: *Medical image analysis* 76 (2022), p. 102306 (cit. on pp. 9, 54).
- [45] A. F. Markus, J. A. Kors, and P. R. Rijnbeek. “The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies”. In: *Journal of Biomedical Informatics* 113 (2021), p. 103655 (cit. on p. 22).

- [46] T. A. McDonagh, M. Metra, M. Adamo, et al. “2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC”. In: *European heart journal* 42.36 (2021), pp. 3599–3726 (cit. on p. 7).
- [47] G. M. McKhann, D. S. Knopman, H. Chertkow, et al. “The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association work-groups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s Dementia* 7.3 (2011), pp. 263–269 (cit. on p. 7).
- [48] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020 (cit. on pp. 22, 108).
- [49] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein. “Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5425–5434 (cit. on pp. 17, 18).
- [50] F. Monti, M. M. Bronstein, and X. Bresson. “Geometric Matrix Completion with Recurrent Multi-Graph Neural Networks”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, et al. 2017, pp. 3697–3707 (cit. on pp. 23, 24, 30).
- [51] D. Morselli Gysi, Í. do Valle, M. Zitnik, et al. “Network medicine framework for identifying drug-repurposing opportunities for COVID-19”. In: *Proceedings of the National Academy of Sciences* 118.19 (2021). eprint: <https://www.pnas.org/content/118/19/e2025581118.full.pdf> (cit. on p. 20).
- [52] N. Muthukrishnan, F. Maleki, K. Ovens, C. Reinhold, B. Forghani, and R. Forghani. “Brief history of artificial intelligence”. In: *Neuroimaging Clinics* 30.4 (2020), pp. 393–399 (cit. on pp. 14, 15).
- [53] J. G. Nam, S. Park, E. J. Hwang, et al. “Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs”. In: *Radiology* 290.1 (2019), pp. 218–228 (cit. on p. 5).
- [54] D. E. Newman-Toker, Y.-H. Hsieh, C. A. Camargo, A. J. Pelletier, G. T. Butchy, and J. Edlow. “Spectrum of dizziness visits to US emergency departments: cross-sectional analysis from a nationally representative sample.” In: *Mayo Clinic proceedings* 83 7 (2008), pp. 765–75 (cit. on p. 70).
- [55] D. E. Newman-Toker, K. Kerber, Y.-H. Hsieh, et al. “HINTS outperforms ABCD2 to screen for stroke in acute continuous vertigo and dizziness.” In: *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine* 20 10 (2013), pp. 986–96 (cit. on p. 72).
- [56] W. H. Organization et al. *Delivering quality health services: A global imperative*. OECD Publishing, 2018 (cit. on p. 7).
- [57] S. Parisot, S. I. Ktena, E. Ferrante, et al. “Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer’s disease”. In: *Medical Image Analysis* 48 (2018), 117–130 (cit. on pp. 28, 30).
- [58] S. Parisot, S. I. Ktena, E. Ferrante, et al. “Spectral graph convolutions for population-based disease prediction”. In: *Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*. 2017, pp. 177–185 (cit. on pp. 18, 19, 28, 30).
- [59] A. Paszke, S. Gross, F. Massa, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *CoRR* abs/1912.01703 (2019). arXiv: 1912.01703 (cit. on p. 8).
- [60] R. Poplin, A. V. Varadarajan, K. Blumer, et al. “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning”. In: *Nature Biomedical Engineering* 2.3 (2018), pp. 158–164 (cit. on p. 5).

- [61] N. Rao, H.-F. Yu, P. Ravikumar, and I. S. Dhillon. “Collaborative Filtering with Graph Information: Consistency and Scalable Methods”. In: *Neural Information Processing Systems (NIPS)* (2015), pp. 1–9 (cit. on p. 24).
- [62] N. Reibling, M. Ariaans, and C. Wendt. “Worlds of healthcare: a healthcare system typology of OECD countries”. In: *Health Policy* 123.7 (2019), pp. 611–620 (cit. on p. 7).
- [63] P. Ren, Y. Xiao, X. Chang, et al. “A survey of deep active learning”. In: *ACM Computing Surveys (CSUR)* 54.9 (2021), pp. 1–40 (cit. on p. 56).
- [64] F. Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 14).
- [65] D. B. Rubin. “Inference and missing data”. In: *Biometrika* 63.3 (1976), pp. 581–592 (cit. on p. 27).
- [66] S. Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016) (cit. on p. 21).
- [67] I. H. Sarker. “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions”. In: *SN Computer Science* 2.6 (2021), pp. 1–20 (cit. on p. 14).
- [68] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, and S. Thun. “The use of machine learning in rare diseases: a scoping review”. In: *Orphanet Journal of Rare Diseases* 15.1 (June 2020) (cit. on p. 7).
- [69] H. Shim, S. J. Hwang, and E. Yang. “Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018 (cit. on pp. 54, 56).
- [70] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, et al. “Statistical normalization techniques for magnetic resonance imaging”. In: *NeuroImage: Clinical* 6 (2014), pp. 9–19 (cit. on p. 12).
- [71] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. “Not just a black box: Learning important features through propagating activation differences”. In: *arXiv preprint arXiv:1605.01713* (2016) (cit. on p. 22).
- [72] Z. Shui and G. Karypis. “Heterogeneous Molecular Graph Neural Networks for Predicting Molecule Properties”. In: *2020 IEEE International Conference on Data Mining (ICDM)* (2020), pp. 492–500 (cit. on p. 20).
- [73] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains”. In: *IEEE signal processing magazine* 30.3 (2013), pp. 83–98 (cit. on p. 19).
- [74] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *In Workshop at International Conference on Learning Representations*. Citeseer. 2014 (cit. on p. 22).
- [75] H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson. “The global burden of diagnostic errors in primary care”. In: *BMJ quality & safety* 26.6 (2017), pp. 484–494 (cit. on pp. 6–8, 53).
- [76] J. P. Staab, A. Eckhardt-Henn, A. Horii, et al. “Diagnostic criteria for persistent postural-perceptual dizziness (PPPD): Consensus document of the committee for the Classification of Vestibular Disorders of the Bárány Society.” In: *Journal of vestibular research : equilibrium & orientation* 27 4 (2017), pp. 191–208 (cit. on p. 70).
- [77] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic attribution for deep networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 3319–3328 (cit. on pp. 22, 23).
- [78] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. “An overview of clinical decision support systems: benefits, risks, and strategies for success”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–10 (cit. on p. 6).

- [79] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 22).
- [80] A. S. S. Tehrani, J. C. Kattah, K. Kerber, et al. “Diagnosing Stroke in Acute Dizziness and Vertigo: Pitfalls and Pearls.” In: *Stroke* 49 3 (2018), pp. 788–795 (cit. on p. 70).
- [81] M. E. Tipping and C. M. Bishop. “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622 (cit. on p. 29).
- [82] O. G. Troyanskaya, M. N. Cantor, G. Sherlock, et al. “Missing value estimation methods for DNA microarrays”. In: *Bioinformatics* 17 6 (2001), pp. 520–5 (cit. on p. 29).
- [83] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018 (cit. on p. 27).
- [84] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 15).
- [85] F. M. D. L. Vega, S. Chowdhury, B. Moore, et al. “Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases”. In: *Genome Medicine* 13.1 (Oct. 2021) (cit. on p. 7).
- [86] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. “Graph Attention Networks”. In: *International Conference on Learning Representations* (2018). accepted as poster (cit. on pp. 17, 18).
- [87] G. Vivar, A. Kazi, H. Burwinkel, A. Zwergal, N. Navab, and S.-A. Ahmadi. “Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification”. In: *Artificial intelligence in medicine* 117 (2021), p. 102097 (cit. on p. 27).
- [88] G. Vivar, K. Mullaeva, A. Zwergal, N. Navab, and S.-A. Ahmadi. “Peri-Diagnostic Decision Support Through Cost-Efficient Feature Acquisition at Test-Time”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II*. Lima, Peru: Springer-Verlag, 2020, 572–581 (cit. on p. 54).
- [89] G. Vivar, R. Strobl, E. Grill, N. Navab, A. Zwergal, and S.-A. Ahmadi. “Using Base-ml to Learn Classification of Common Vestibular Disorders on DizzyReg Registry Data”. In: *Frontiers in Neurology* 12 (2021) (cit. on p. 70).
- [90] A. Wagstaff, G. Flores, J. Hsu, et al. “Progress on catastrophic health spending in 133 countries: a retrospective observational study”. In: *The Lancet Global Health* 6.2 (2018), e169–e179 (cit. on p. 7).
- [91] J.-C. Warninghoff, O. Bayer, U. Ferrari, and A. Straube. “Co-morbidities of vertiginous diseases”. In: *BMC Neurology* 9 (2009), pp. 29–29 (cit. on p. 70).
- [92] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger. “Simplifying graph convolutional networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6861–6871 (cit. on pp. 16, 17).
- [93] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. “A comprehensive survey on graph neural networks”. In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24 (cit. on pp. 14, 15, 20).
- [94] J. Yanase and E. Triantaphyllou. “A systematic survey of computer-aided diagnosis in medicine: Past and present developments”. In: *Expert Systems with Applications* 138 (2019), p. 112821 (cit. on pp. 6, 8, 28, 54).
- [95] J. Yanase and E. Triantaphyllou. “The seven key challenges for the future of computer-aided diagnosis in medicine”. In: *International journal of medical informatics* 129 (2019), pp. 413–422 (cit. on pp. 6, 8, 13).

List of Figures

2.1	Illustration of tabular data from clinical observations containing a complete set of imaging, non-imaging, and meta-features where each row denote patient information.	12
2.2	Illustration of block-level missingness in clinical datasets containing incomplete set of imaging, non-imaging, and meta-features.	13
2.3	Illustration of feature-level missingness in clinical datasets containing an incomplete set of imaging, non-imaging, and meta-features.	14
2.4	Illustration of deep learning workflow from input data to loss function calculation including model parameter updates using an optimizer (gradient descent optimization algorithm).	15
2.5	Example undirected graph with seven nodes.	16
2.6	Corresponding Adjacency matrix (left), Degree matrix (middle), and Graph Laplacian matrix (right) of an example seven-node graph shown in 2.5.	16
2.7	Simplified illustration of message passing concept wherein at first step each node contains a feature vector that represents the “message” it wants to send to all its neighbours (left panel). In the next step (right panel), the messages are passed on to their neighbours.	18
2.8	Population graph modelling of a tabular dataset in CADx where every row in the table (top) are patients with corresponding imaging-, non-imaging-, and meta-features. Every patient is modelled as a node in the graph (bottom) with their corresponding feature vector representations (row vectors).	19
2.9	Node-level classification task assigning class label \hat{y} to every node. Given a graph and node’s feature representations (left) the goal is to assign class label \hat{y} to every node in the graph (right).	20
2.10	Graph-level classification task assigning class label \hat{y} (illustrated based on the colour of the bounding box) to every graph given their corresponding adjacency matrices \mathbf{A}_i and feature matrices \mathbf{X}_i	21

List of Tables

