University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2024

Explainable Deep Learning for Medical Imaging Classification

Courtman, Megan

https://pearl.plymouth.ac.uk/handle/10026.1/22598

http://dx.doi.org/10.24382/5231
University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Copyright statement

Copyright and Moral rights arising from original work in this thesis and (where relevant), any accompanying data, rests with the Author unless stated otherwise[1].

Re-use of the work is allowed under fair dealing exceptions outlined in the Copyright, Designs and Patents Act 1988 (amended)[2], and the terms of the copyright licence assigned to the thesis by the Author.

In practice, and unless the copyright licence assigned by the author allows for more permissive use, this means,

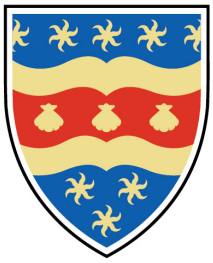
- that any content or accompanying data cannot be extensively quoted, reproduced or changed without the written permission of the author/rights holder; and
- that the work in whole or part may not be sold commercially in any format or medium without the written permission of the author/rights holder.

Any third-party copyright material in this thesis remains the property of the original owner. Such third party copyright work included in the thesis will be clearly marked and attributed, and the original licence under which it was released will be specified. This material is not covered by the licence or terms assigned to the wider thesis and must be used in accordance with the original licence; or separate permission must be sought from the copyright holder.

The author assigns certain rights to the University of Plymouth including the right to make the thesis accessible and discoverable via the British Library's Electronic Thesis Online Service (EThOS) and the University research repository, and to undertake activities to migrate, preserve and maintain the medium, format and integrity of the deposited file for future discovery and use.

[1] *E.g.* in the example of third party copyright materials reused in the thesis.

[2] In accordance with best practice principles such as, *Marking/Creators/Marking third party content* (2013). Available from: https://wiki.creativecommons.org/wiki/Marking/Creators/Marking_third_party_content [accessed 28th February 2022]



UNIVERSITY OF
PLYMOUTH

EXPLAINABLE DEEP LEARNING FOR MEDICAL IMAGING CLASSIFICATION

by

MEGAN COURTMAN

A thesis submitted to the University of Plymouth in partial fulfilment
for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

September 2024

This thesis is dedicated to my family.

To my parents Jenny and Michael, for instilling in me a lifelong love of learning.

To my grandparents Belle and Steve, no longer with us, for setting shining examples of intellectual curiosity.

To my grandparents Eileen and David, for the unwaveringly generous support that made my convoluted academic journey possible.

To my sister Bronwen, for being both my doctoral and Disney companion.

To Pebble, an angel in the shape of a cat, for the enormous comfort and joy she brought in her too-short life.

And to my husband Nicholas, for everything, always, but especially for believing that I would make a good coder.

Acknowledgements

I had a wonderful time conducting this research, for which I have many people to thank. Firstly, I am extremely grateful for the support of my supervisory team. My Director of Studies, Prof. Emmanuel Ifeachor, imparted much academic wisdom and I benefited greatly from his wealth of experience. Dr. Lingfen Sun volunteered her prodigious technical expertise to offer compelling and constructive advice. Dr. Mark Thurston guided me on matters of both software and hardware with endless generosity and enthusiasm. Dr. Stephen Mullin mentored me in every aspect of the PhD journey; his guidance and his confidence in me has been transformative.

I am also thankful for the opportunities I have had throughout this research to engage with patients and members of the public, and for the meaningful feedback this has generated. In particular, I am grateful to Dr. Andy Foggo, Chris Maycock, Sue Whipps, all at PenPRIG, and the Parkinson's UK Involvement Steering Group.

This research would not have been possible without significant computational resources. My thanks go to Dr. Chima Stanley Eke and Dr. Is-Haka Mkwawa for establishing and maintaining a remote server. I am also immensely grateful to the Holsworthy Rotary Club for providing the funding for a high-spec workstation.

This research also depended on the invaluable contributions of many co-authors. I am grateful to Dr. Lucy McGavin and Prof. Stephen Hall for their expertise in neuroradiology, Dr. Galaleldin Abdelhalim for his diligent ground-truthing of images, Dr. Adam Streeter for his statistical expertise, Dr. Hongrui Wang for the vital development of data pipelines, and Dr. Sube Banerjee for his guidance on publishing.

I am also grateful to the Applied Parkinson's Research Group at the University of Plymouth, who gave me many opportunities to present my work and facilitated many instructive discussions. My thanks in particular go to Dr. Camille Carroll, Dr. Marie-Louise Zeissler, Jemma Inches, Emma Edwards, Beccy Chapman, Chris Lovegrove and Katie Bounsall, for their useful feedback as well as their friendship.

It would be remiss of me not to acknowledge the countless open source developers who have contributed to the software I use. Every facet of my research has been made possible through their dedication.

Last but not least, I am grateful to my family for their support in many forms. For academic support in particular, I thank my dad, Dr. Michael Allen. He inspired me to become a data scientist and has been an indispensable source of wisdom.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was financed with the aid of a studentship from the Engineering and Physical Sciences Research Council.

Word count of main body of thesis: 20,761

Publications from thesis

Courtman Megan, Kim Daniel, Wit Huub, Wang Hongrui, Sun Lingfen, Ifeachor Emmanuel, Mullin Stephen, Thurston Mark, "Deep learning detection of aneurysm clips for magnetic resonance imaging safety", *Journal of Imaging Informatics in Medicine* (2024), https://doi.org/10.1007/s10278-023-00932-8.

Courtman Megan, Thurston Mark, McGavin Lucy, Carroll Camille, Sun Lingfen, Ifeachor Emmanuel, Mullin Stephen, "Explainable deep learning based detection of Parkinson's changes in MRI brain scans." *Movement Disorders* 38 (2023): suppl 1.

Courtman Megan, Thurston Mark, McGavin Lucy, Carroll Camille, Sun Lingfen, Ifeachor Emmanuel, Mullin Stephen, "Artificial Intelligence based detection of Parkinson's disease in Magnetic Resonance Imaging brain scans." *Movement Disorders* 37 (2022): suppl 2.

Courtman Megan, Thurston Mark, McGavin Lucy, Carroll Camille, Sun Lingfen, Ifeachor Emmanuel, and Mullin Stephen."095 Artificial Intelligence based detection of Parkinson's disease in magnetic resonance imaging brain scans." *Journal of Neurology, Neurosurgery and Psychiatry* 93, no. 9 (2022): e2, https://doi.org/10.1136/jnnp-2022-abn2.139.

Manuscripts in preparation

Courtman Megan, Thurston Mark, Wang Hongrui, Banerjee Sube, Streeter Adam, McGavin Lucy, Hall Stephen, Sun Lingfen, Ifeachor Emmanuel, Mullin Stephen, "Explainable machine learning approach classifies MRI brain scans of idiopathic and prodromal *LRRK2*/*GBA* Parkinson's disease cases with high accuracy."

Courtman Megan, Abdelhalim Galaleldin, Sun Lingfen, Ifeachor Emmanuel, Mullin Stephen, Thurston Mark, "Applying explainable natural language processing to radiology reports for dataset curation."

Signed: Megan Courtman Dated: 26th September 2024

Explainable deep learning for medical imaging classification

by Megan Courtman

Abstract

Machine learning is increasingly being applied to medical imaging tasks. However, the "black box" nature of techniques such as deep learning has inhibited the interpretability and trustworthiness of these methods, and therefore their clinical utility. In recent years, explainability methods have been developed to allow better interrogation of these approaches.

This thesis presents the novel application of explainable deep learning to several medical imaging tasks, to investigate its potential in patient safety and research. It presents the novel application of explainable deep learning to the detection of aneurysm clips in CT brains for MRI safety. It also presents the novel application of explainable deep learning to the detection of confounding pathology in radiology report texts for dataset curation. Furthermore, it makes novel contributions to Parkinson's research, using explainable deep learning to identify progressive brain changes in MRI brain scans, and to identify differences in the brains of non-manifesting carriers of Parkinson's genetic risk variants in MRI brain scans. In each case, convolutional neural networks were developed for classification of data, and Shapley Additive exPlanations (SHAP) were used to explain predictions. A novel pipeline was developed to apply SHAP to volumetric medical imaging data.

The application of explainable deep learning to various types of data and task demonstrates the flexibility of the combination of convolutional neural networks and SHAP. Additionally, these applications highlight the importance of combining explainability with clinical expertise, to check the viability of the models and to ensure that they meet a clinical need. These novel applications represent useful new tools for safety and research, and potentially for improvement of clinical care.

Contents

List of Figures

List of Tables

List of abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
CI	Confidence Interval
CSF	Cerebrospinal Fluid
CT	Computerised Tomography
DAT	Dopamine Active Transporter
DICOM	Digital Imaging and Communications in Medicine
FN	False Negative
FP	False Positive
GBA	Glucocerebrosidase
GBA nPD	*GBA* non PD manifesting carriers
GBA PD	*GBA* PD manifesting carriers
GC *GBA* nPD	Gaucher causing *GBA* non PD manifesting carriers
GPU	Graphics Processing Unit
IPD	Idiopathic Parkinson's disease
LRRK2	Leucine-Rich Repeat Kinase 2
LRRK2 nPD	*LRRK2* non PD manifesting carriers
LRRK2 PD	*LRRK2* PD manifesting carriers
MNIST	Modified National Institute of Standards and Technology
MRI	Magnetic Resonance Imaging
NHS	National Health Service
NLP	Natural Language Processing
PACS	Picture Archiving and Communication System
PD	Parkinson's Disease
PET	Positron Emission Tomography
PPMI	Parkinson's Progression Markers Initiative
RBD	REM sleep Behaviour Disorder
ReLU	Rectified Linear Units
REM	Rapid Eye Movement
RIS	Radiology Information System
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SPECT	Single-Photon Emission Computed Tomography
SQL	Structured Query Language
TN	True Negative
TP	True Positive
UHPNT	University Hospitals Plymouth NHS Trust

Chapter 1

Introduction

1.1 Motivation

Artificial intelligence (AI) is increasingly being used in radiology and radiological research, driven by a desire for greater efficacy and efficiency in medical care [1]. The volume of radiological data is outpacing the availibility of trained reporters [2], increasing radiologists' workloads and their susceptibility to errors. AI offers a potential solution to this problem: the prospect of seamless integration into the radiological workflow to speed up processes and reduce errors. In some cases, AI could augment the radiological process by providing pre-screened images and identified features. In other cases, where the performance of AI exceeds the performance of radiologists, processes could be entirely automated [1].

The promise of AI has led to some extreme projections: in 2016, Geoffrey Hinton, one of the founders of deep learning technology, infamously said "people should stop training radiologists now. It's just completely obvious than in five years deep learning is going to do better than radiologists" [3]. This has not come to pass. Despite the promise of AI and individual successful studies, extremely few applications have crossed the chasm between research and clinical practice. Many of those which have are of dubious legitimacy; a study in 2021 found that of 100 commercially available AI products in radiology, 64% had no peer-reviewed evidence for their efficacy, and only 18% had demonstrated clinical impact [4].

There are several factors that have limited the utility of AI in radiological practice, two of the most significant of which are addressed in this thesis. The first is that the development of successful models is reliant on a great volume of high-quality, labelled data [5], and such curated data is frequently unavailable and would require extensive manual ground-truthing [1]. The second is that the "black box" nature of many of the most successful deep learning techniques has inhibited their interpretability and trustworthiness [5].

The opacity of deep learning models has begun to be addressed in recent years with the introduction of explainable AI techniques. These approaches are designed to increase the interpretability of models despite the model complexity; in the case of radiology, they can be used to yield a map which can be overlaid on the image to indicate which anatomical regions have informed a given prediction [6]. Such visualisation techniques provide powerful insights into the workings of complex models. This thesis will focus on the utility of explainable AI in radiological model development: particularly how the explanation of predictions can be combined with insights from medical experts to verify the validity of models and to refine them to improve their performance. Whilst the thesis will focus on models in research and development, it is worth noting that explainable AI also holds promise for models that have been progressed to the production stage: it can be used for auditing (assessing models' conformance to regulations and procedures) and quality assurance (identifying potential weaknesses) [6].

In summary, there is a need to address issues such as data curation and model explainability, so that the demonstrable potential of deep learning techniques might be translated into effective radiological practice.

1.2 Aims and objectives

This project aimed to use explainable AI for medical imaging applications, to counter the limitations of successful but often opaque deep learning techniques.

The core objectives were to:

- conduct a literature review to identify existing applications of explainable AI to medical imaging
- identify and collect relevant data
- develop novel applications of explainable deep learning to medical imaging classification
- demonstrate the potential utility of these applications in clinical settings

1.3 Contributions to knowledge

This thesis makes the following contributions to knowledge. It:

1. presents a novel application of explainable AI for the detection of aneurysm clips in Computerised Tomography (CT) brain scans for Magnetic Resonance Imaging (MRI) safety
2. presents a novel application of explainable AI for the detection of confounding pathology in radiology report texts
3. presents a bespoke application of Shapley Additive exPlanations (SHAP) to volumetric medical imaging
4. identifies progressive Parkinson's brain changes in MRI brain scans using explainable AI
5. identifies differences in the brains of non-manifesting carriers of Parkinson's genetic risk variants in MRI brain scans using explainable AI
6. presents the application of explainable AI to a novel routinely-collected dataset of Parkinson's MRI brain imaging

1.4 Outline of thesis

This thesis consists of seven chapters. This chapter has introduced the project and highlighted motivations, aims and objectives and contributions of the thesis. Chapter 2 provides a background to the project, detailing the concepts of AI, machine learning, deep learning, explainable AI, and the impact of these technologies in radiology. Chapter 3 describes the methods used in this research, including the development environment, the model architectures, the performance metrics, the evaluation methods, and the explainability technique. Chapter 4 provides an initial illustration of the potential use of explainable AI for medical imaging, discussing explainable models developed for the detection of aneurysm clips in CT brain scans. Chapter 5 details a potential use of explainable AI for radiological data curation, discussing explainable models developed for the detection of confounding pathology in MRI brain scan reports. Chapter 6 presents a potential use of explainable AI for uncovering new insights in such research datasets, discussing explainable models developed for the detection of Parkinson's disease in MRI brain scans and how these have detected progressive changes as well as differences in the brains of non-manifesting carriers of Parkinson's genetic risk variants. Finally, Chapter 7 discusses the contributions to knowledge, limitations, future directions, and conclusions.

Chapter 2

Background

2.1 Artificial intelligence

Artificial intelligence describes a broad area of study in the field of computer science, which might be summarised as the study of agents that receive information from the environment and perform actions [7]. The term was coined in 1955 in a proposal for a Dartmouth College summer research project, which theorised that every aspect of intelligence could be simulated by a machine [8]. The outlined components of the "artificial intelligence problem" bear remarkable resemblance to areas of AI that continue to be studied nearly 70 years later: automation, calculation efficiency, abstraction, random seeding, machine self-improvement, computer understanding of language, and the arrangement of hypothetical "neurons" to form concepts.

Progress in AI has been slower and more turbulent than anticipated by the proposers of the Dartmouth summer program, who expected that a group of ten scientists could make significant advances over the course of two months [8]. In fact, over the decades the development of AI has undergone a cycle of optimistic, productive "springs" and pessimistic, unproductive "winters" [9]. Whilst the original goal of general machine intelligence remains unmet, bursts of progress have ultimately resulted in the substantial success of domain-level applications. Today AI can be seen discretely applied in nearly every aspect of life, due to the development of techniques in areas such as computer vision, natural language processing, speech recognition,

and robotic control.

2.1.1 Machine learning

A large component of the field of AI is the subfield of machine learning. Machine learning describes systems which improve their performance through exposure to data [10]. The system maps input data to targets through training rather than explicit programming: the system is shown many examples of inputs and targets, which allows it to learn rules for automating the task. The central task of the system is to meaningfully transform data by learning useful representations of input data. The "learning" in machine learning describes an automatic search for better representations of the data. The search takes place in a predefined set of operations, called a "hypothesis space", and uses guidance from a feedback signal [11]. Machine learning tasks include clustering, regression, classification, outlier detection, image generation, and text completion [12].

Machine learning started to flourish in the 1990s, and has quickly become the most popular subfield of AI due to the availability of fast hardware and large datasets. It is strongly tied to the field of statistics, but differs in its focus on engineering and empirical results rather than mathematical theory [11].

Supervised learning is a category of machine learning which is popular in medical imaging applications. This describes algorithms which induce models from labelled training data. These models can then be used to predict the labels for unlabelled data [13]. Supervised architectures include decision trees, support vector machines, linear regression models, random forests, boosting models, and neural networks [14].

2.1.2 Deep learning

Much of the recent development in medical AI has been driven by a subset of machine learning called deep learning, which describes the use of layered neural networks to build representations of complicated concepts out of simpler concepts (discussed further in section 3.2) [15]. A distinctive aspect of deep learning is that it does

not require manual feature selection or engineering, and can learn more abstract representations that might not have been anticipated by human practitioners.

The first successful practical application of neural networks was in 1989, when Yann LeCun of Bell Labs combined the earlier ideas of convolutional neural networks and backpropagation (discussed further in section 3.2) and applied them to the classification of handwritten digits [16]. This network was then used by the United States Postal Service in the 1990s [11].

The broader uptake and success of neural networks in the early 2010s was largely driven by the image classification competition ImageNet [17]. In 2012, a team led by Alex Krizhevsky and advised by Geoffrey Hinton entered and achieved a top-five accuracy of 83.6% (compared to the previous winner's 74.3%) [18]. The competition has allowed annual improvements in the performance of deep convolutional neural networks ever since. By 2015, the winner had reached an accuracy of 96.4% [19].

Deep learning has now achieved remarkable results in tasks such as speech recognition, handwriting recognition, machine translation, text-to-speech conversion, digital assistants, autonomous driving, ad targeting, search results, answering natural language questions, and game playing [11]. Several factors have facilitated the success of deep learning. The first is hardware: the development of Graphics Processing Units (GPUs) for the gaming market has made powerful hardware available for deep learning applications [11]. Another factor in the success of deep learning has been the availability of data: the exponential progress in storage hardware and the rise of the internet has made very large datasets available to fuel deep learning applications [11]. Often the best performing networks have been trained on large datasets like ImageNet, and then honed for specific applications, in a process known as transfer learning. A third factor in the success of deep learning has been advances in the algorithms that form neural networks (discussed further in section 3.2), allowing for very large and effective networks [11]. As a result of these successes, investment in deep learning has been monumental, and open-sourcing of deep learning development tools has made its implementation extremely accessible [11].

In terms of medical applications, traditional pre-defined feature machine learning systems have not generally met the stringent performance requirements for clinical utility [20], but deep learning methods have achieved higher performances and it is anticipated that they will reach a high enough standard to be clinically useful [1].

2.2 Artificial intelligence in radiology

Healthcare is an environment which has already been affected by the development of AI, in areas such as drug discovery, remote patient monitoring, wearables, risk management, hospital management, and medical diagnostics and imaging [1]. Radiological processes in particular have long been obvious candidates for AI integration; since the 1980s the automation of clinical tasks has shifted radiology to a quantifiable computable domain [21]. In the last decade, there have been significant advances in AI-based medical image classification due to increased compute power, the open-sourcing of large labelled datasets, and the development of deep learning [22]. There are now thousands of publications applying computer vision techniques to medical imaging [23].

2.2.1 Benefits

One of the demands of radiology which has invited the integration of AI is the sheer quantity of data that needs to be processed. The amount of imaging is increasingly outstripping the number of available trained readers [2]. In North America, for example, the number of imaging studies increases by up to 5% a year on average, whereas the number of radiology positions only increases by 2% a year [24]. A study in 2015 reported that radiologists had to interpret an image every 3-4 seconds on average to meet demand [25]. Fatigue is a known problem in radiology, and affects diagnostic accuracy [26]. The promise of AI is clear: some rote tasks could be entirely automated, while other tasks could be streamlined by the providing radiologists with pre-screened images and identified features [27]. Both approaches would save radiol-

ogists time, benefitting both their welfare and their performance. A study has found that most radiologists are optimistic about the impact of AI on their practice, due to the expectation that it will result in lowered risk of errors and increased time with patients [28].

Human errors can arise for reasons beyond fatigue and demand. Radiologists' assessments are informed by education and experience, and therefore can be subjective. They can also be affected by inattentional blindness, as demonstrated by a study in which radiologists were asked to perform a familiar lung nodule detection task, and 83% did not observe an inserted image of a gorilla [29]. AI tools have the potential to detect patterns that radiologists have not been trained to observe, patterns that are obfuscated to humans by context, or patterns that are not even accessible to the human visual system [30]. The quantified outputs of these tools have the benefit of being consistent and reproducible, and not subject to disparities in healthcare provision.

AI usually relies on large datasets, and large radiological datasets should theoretically be readily available due to the extensive routine collecting of imaging [1]. For example, nearly a million labelled chest X-rays have been open sourced [31–33]. Promisingly accurate radiological results have already been achieved using AI trained on such datasets. A study classifying chest X-rays as normal or abnormal achieved accuracy of over 95% [34]. A study applying deep learning to diagnosis of hip fractures in X-ray images achieved an accuracy of over 99% [35]. Studies have used machine learning to detect brain haemorrhages [36], liver mass classification [37], and vertebral compression [38] in CT scans, all with accuracies comparable to or greater than that of radiologists. Some studies have demonstrated that the highest accuracies are achieved when the findings of AI and radiologists are combined [39]. These results indicate the potential diagnostic value of incorporating AI as a clinical tool.

AI also has the potential to make healthcare more equitable by being data-driven and, theoretically, not being subject to human biases and long-standing inequalities.

For example, a study that used deep learning to predict the severity of osteoarthritis in knee X-rays found that this approach dramatically reduced unexplained racial disparities in pain [40]. However, this equity is only achievable if key steps are taken, such as ensuring high quality and lack of bias in data, addressing model limitations, and facilitating community participation [22]. The absence of such steps can lead to the realisation of risks of using AI in a medical setting.

2.2.2 Risks

There have been many cases of AI being misused or poorly understood in healthcare research. An example which illustrates many of the common pitfalls is the application of machine learning to detecting COVID-19 in chest radiographs and CT scans. Over two thousand papers were published on this subject between January and October 2020, but a review found that none of the models in the sixty-two included studies were of potential clinical use due to methodological flaws or underlying biases [41]. A broader review of prediction models for COVID-19 diagnosis screened over thirty-seven thousand titles and included 232 models in their analysis. They too found that all of the models were at high or unclear risk of bias [42].

A problem which plagues many machine learning applications is the quality of the input data. In the COVID-19 example, the collation of publicly available images into "Frankenstein datasets" led to duplication of images, and also to the likelihood of implicit biases, as unusual or severe cases of COVID-19 were more likely to appear in publications [41]. Many of the papers also failed to mention that they used a control dataset consisting of paediatric patients; the models were therefore likely to be distinguishing between children and adults rather than pneumonia and COVID-19 [41]. A study also demonstrated that it was possible for a model to classify the distinct sources of COVID-19 images with high accuracy once the lung region had been excluded entirely [43]. This indicates that the models could therefore have been distinguishing between sources rather than between pathologies. A similar finding was reported in an earlier study that used deep learning to detect pneumonia in

chest radiographs. The models were able to detect where a radiograph was acquired with extremely high accuracy and, as the hospitals had different disease burdens, they were able to exploit this information in their predictions [44].

Another problem demonstrated by the COVID-19 example was that of flawed methodology. Many of the AI models suffered from a high or unclear risk of bias in at least one domain. For example, many validation datasets were not representative of the target population, and many studies defaulted to using the image preprocessing classically used for ImageNet classification rather than using clinical judgement [41].

These problems are not without recourse. Suggested solutions include exercising caution when using publicly available datasets, using well-curated external validation datasets, using clinical judgement in the process of model development, open sourcing code, and assessing studies against established frameworks [41]. These measures should lead to more generalisable AI models, more accurate reporting of models and more awareness of model limitations.

As few machine learning applications have yet been fully deployed in clinical settings, it is unsurprising that early adoptions have encountered unforeseen practical difficulties. For example, a deep learning system developed by Google for the detection of diabetic retinopathy in eye scans achieved greater than 90% accuracy in the laboratory, but presented problems when deployed in a clinical setting. It rejected more than a fifth of the scans because it had been trained to reject low quality images, causing great difficulties for the clinicians and patients involved in the implementation of the system [45]. Such practical considerations will need to be addressed if AI applications are to be successfully realised. The involvement of clinicians in the development of models is likely to mitigate against such shortcomings.

The use of AI can also raise a number of ethical issues, such as that of privacy. An example which highlights this is DeepMind's collaboration with the Royal Free NHS Trust in London. In 2016 DeepMind was granted access to three hospitals' medical data in exchange for the development of an application to assist in the management of acute kidney injury. Outrage ensued when an investigation revealed

that DeepMind was granted access to the identifiable medical histories of 1.6 million patients without their consent having been given: data collection that was far beyond the scope of what had been stated publicly [46]. The controversy could have been avoided with the implementation of good data protection practices, such as the curation of appropriate and limited datasets, high quality public consultation, transparent Patient and Public Involvement (PPI), respect for opt-outs, and clear communication of results. Confidential data can also be vulnerable due to insecure connections between medical institutions and externally hosted AI systems [1]. Decentralised federated learning is being adopted as a solution to this: the practice of training a model across separate servers holding local data without there being any data exchange between devices [47]. It is evident that many of the potential pitfalls of using AI in a medical setting have multiple potential solutions.

A further problem lies in the interpretation of AI models. Due to the complexity and abstract feature representation inherent in deep learning, these models have been described as "black boxes" that are difficult to understand, and their opaque use in a clinical setting has been challenged. The lack of transparency affects the models' trustworthiness: it has been argued both that users cannot trust an opaque model and also that they might trust it too much, not noticing any mistakes that it might make [48]. This has led to recommendations that AI models be explainable in a way that clinical users can understand and use to justify their decision-making [49]. In fact General Data Protection Regulation laws in the European Union set out detailed transparency obligations for algorithmic decision-making, requiring the provision of "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [50].

2.3 Explainable artificial intelligence

Explainable artificial intelligence describes techniques which have been developed to counter the "black box" nature of machine learning methods such as deep learning. Explainability is the task of finding an interpretable model that approximates the

opaque model as closely as possible [51]. These methods can be global (providing an overall approximation of the behaviour of a model) or local (providing explanations for particular instances) [51]. Explainability has become an important issue to the general public as well as machine learning practitioners, as algorithmic decision making has become more impactful on society and individuals [52].

The number of studies looking at using explainable AI in healthcare has increased exponentially over the last few years [53]. Explainability is not needed when there are no significant consequences for unacceptable results or when the task is well understood, but in radiology neither of these conditions are met: a wrong diagnosis can have serious consequences, and a clinical diagnosis is not a trivial task [54]. Explainability is needed in radiology to audit systems, enhance trust and adhere to regulations [6]. It may be used for tasks such as image segmentation, lesion and organ detection, computer aided diagnosis and staging, prognosis, radiation therapy planning, triaging, and image reconstruction [6].

Various explainability methods have been evaluated for the interpretation of deep learning medical image models, many of which provide a visualisation of the features which contributed to the model's output [55]. One of the most popular methods is SHapley Additive exPlanations (SHAP) [53] (discussed further in Section 3.5), which was used in this research.

However, explainable AI has not been universally championed. The ability of current explainability methods to engender trust, provide transparency and mitigate bias has been doubted [56]. It has also been suggested that model transparency can still give rise to undue trust and hamper the user's ability to detect mistakes [57]. Explainability of AI systems is an ongoing debate and is at the forefront of discussions in the field.

2.4 Concluding remarks

This chapter has provided a background to the project, detailing the concepts of AI, machine learning, deep learning, explainable AI, and the impact of these technologies

in radiology. Deep learning has achieved remarkable results in medical imaging tasks, and offers the promise of a streamlined radiological workflow by the automation of some processes and the augmentation of others. However, there are concerns about the risks of this technology in radiology, including the "black box" nature of deep learning algorithms and arising issues of trust. This thesis will explore the use of explainable AI as a solution to this issue that will allow the potential of AI in radiology to be realised effectively and responsibly.

Chapter 3

Methods

The following describes the methods which were used universally across the radiological tasks formulated for this research.

3.1 Development environment

To make the code in this research reproducible, reliable and easy to develop, universal decisions were made with regard to the operating systems, programming languages and other tools used.

3.1.1 Linux

This research was conducted on Linux-based machines. Linux is a Unix-like operating system, the kernel of which is free and open-source [58]. This research used systems based on Ubuntu: a distribution of Linux which is designed to be easy to access and easy to use [59]. It is well-supported for machine learning applications [60].

3.1.2 Python

Python is a high-level, general-purpose, object-oriented, open source programming language, designed for ease and speed of development. It prioritises quality, produc-

tivity, portability and integration [61]. It is one of the most popular programming languages.

Python is made highly extensible by the use of modules, many of which have applications in machine learning. Several of these will be discussed in further depth in Section 3.1.3. Other libraries used in this research include NumPy for array support and mathematical functions [62], pandas for manipulating and analysing data [63], Matplotlib and seaborn for generating plots [64, 65], OpenCV, imutils, scikit-image and Pillow for image functions [66–69], SciPy for scientific computing [70], pydicom for handling DICOM medical imaging files [71], and dicom2niti and NiBabel for handling neuroimaging files [72, 73]. Environments were built and maintained using the Anaconda Python distribution [74].

3.1.3 Machine learning libraries

This research made use of scikit-learn, TensorFlow and Keras. The library scikit-learn is a Python module containing many machine learning tools [75]. TensorFlow is a free and open source machine learning library developed by the Google Brain team, and has a particular focus on the development of deep neural networks [76]. Keras is a free and open source deep learning Python library which acts as an interface to TensorFlow [77].

3.1.4 Jupyter

JupyterLab is an open source web-based interactive development environment for notebooks, code and data [78]. It supports many programming languages, including Python. It was used in this work to allow iterative changes to be made efficiently, in a format that was easy to document and understand.

3.1.5 Graphics Processing acceleration

This research made use of two different types of Graphics Processing Unit (GPU): an NVIDIA Quadro RTX 6000 and an NVIDIA GeForce RTX 3080. TensorFlow has

the capability to utilise NVIDIA GPUs in model training using a proprietary API from NVIDIA called CUDA as an interface [79]. This results in greatly accelerated training.

3.2 Models

3.2.1 Neural networks

Neural networks are mathematical frameworks for learning representations from data. These frameworks are structured in stacked layers. The term "neural network" is a reference to neurobiology, from which the concept drew inspiration [11]. An illustration of the structure is shown in Figure 3.1.

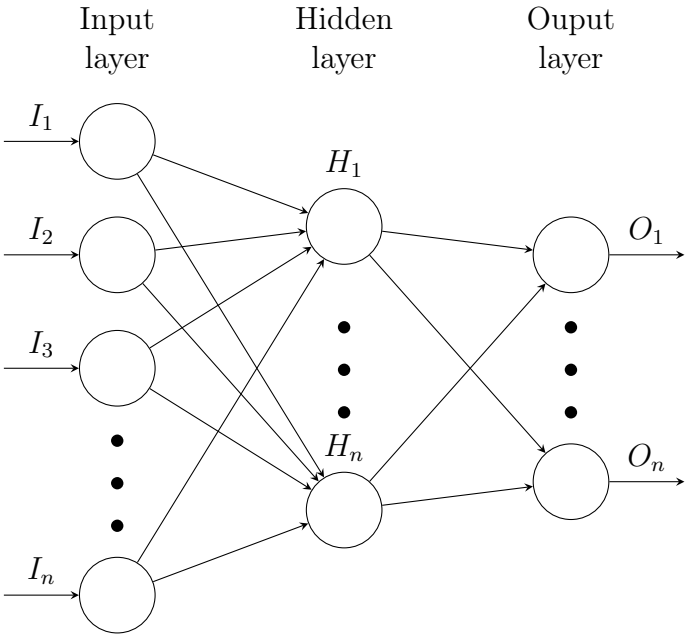


Figure 3.1: Neural network structure

The neural network, like many modern machine learning systems, uses a tensor as its basic data structure. This is a container for numerical data with any number of dimensions. All transformations learned by neural networks can be reduced to a handful of tensor operations [11].

The core building block of a neural network is the data processing module known as a layer. Most of deep learning consists of chaining together simple layers that

together will perform a process of data distillation. The data transformation implemented by a layer is parameterised by its weights. The "learning" in deep learning describes the automatic search for a set of weight values such that the network will correctly map example inputs to their associated targets [11].

The network receives information about how far its output is from the expected output via the loss function or objective function. This takes the predictions of the network and the true target to compute a distance score, capturing how well the network has performed for a specific example. The network then uses this score as a feedback signal to adjust the values of the weights marginally in a direction that will lower the loss score for the instance in question. This adjustment is performed by the optimizer, which implements an algorithm called backpropagation. All operations used in the network are differentiable but chained together: backpropagation uses the chain rule from calculus to compute the gradient values of a neural network [11]. Different optimizers have been trialed; the Adam optimisation algorithm has been used extensively throughout this research. [80].

The network is trained iteratively in loops. The weights are initially randomised and then adjusted slightly for every example seen, to minimise the loss function. A training loop consists of four stages:

1. Drawing a batch of training samples and corresponding targets.
2. Running the network on the batch to obtain predictions.
3. Computing the loss of the network.
4. Updating network weights to reduce the loss on this batch.

Each iteration over all the training data is called an epoch [11].

3.2.2 Convolutional neural networks

Convolutional neural networks, in their present form, were first successfully applied by Yann LeCun of Bell Labs, who combined the earlier idea of convolutional neural networks with the backpropagation algorithm to classify handwritten digits (the

MNIST dataset) [16]. Due to the developments enabled by the ImageNet competition (discussed in section 2.1.2), they have become the state-of-the-art for image classification, though have recently been joined by another type of neural network known as a transformer, originally designed for natural language processing [81], and now applied to imaging tasks [82].

The distinctive feature of the convolutional neural network is the convolution layer, which allows the network to learn local patterns. In the case of images, patterns are found in small 2D windows of the images. These patterns are translation invariant: a learned pattern can be recognised anywhere in the input. Spatial hierarchies of patterns can also be learned: later convolution layers will learn larger patterns made of the features of the earlier layers, allowing the network to learn increasingly abstract and complex concepts [11].

The convolution layer works by sliding windows over input data (for example, an image). At each stop, the patch is transformed into a one-dimensional vector via a convolution kernel: a tensor product with the learned weight matrix. The vectors are then reassembled into an output feature map, which represents the presence of a pattern at different locations in an input. A pooling layer is then used to aggressively downsample the feature maps by extracting windows from the feature maps and outputting the maximum value of each channel [11].

Convolutional neural networks are extremely flexible, and in this research they have been used to classify one-dimensional text data, two-dimensional image data, and three-dimensional image data. The Keras library provides implementations for these different input dimensions, as well as the capability to adjust all network hyper-parameters [77].

3.3 Performance metrics

Performance metrics were chosen to convey various aspects of the performance of classification models. In most cases, these were binary classification tasks. The computation of these metrics involves the terms true positive (TP), false positive (FP),

Table 3.1: Confusion matrix to explain equation nomenclature

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

false negative (FN) and true negative (TN). Table 3.1 displays this classification metric nomenclature.

3.3.1 Accuracy

Accuracy quantifies the proportion of all predictions that are predicted correctly. The sum of the true positives and true negatives is divided by the total number of predictions. This metric is affected by class imbalance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.3.2 Sensitivity

Sensitivity quantifies the proportion of positives that are predicted correctly. The number of true positives is divided by the sum of the true positives and false negatives.

$$Sensitivity = \frac{TP}{TP + FN}$$

3.3.3 Specificity

Specificity quantifies the proportion of negatives that are predicted correctly. The number of true negatives is divided by the sum of the false positives and true negatives.

$$Specificity = \frac{TN}{FP + TN}$$

3.3.4 Balanced accuracy

Balanced accuracy is the arithmetic mean of sensitivity and specificity. It is useful when dealing with imbalanced data, where using the accuracy metric alone might exaggerate the discriminatory ability of the model.

$$Balanced\,accuracy = \frac{Sensitivity + Specificity}{2}$$

When dealing with imbalanced data, the model predictions derived can be adjusted using Bayes' Rule to reflect the actual prevalence of each class [83].

3.3.5 Receiver Operating Characteristic curve

A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity against 1 - specificity [84]. The points on the curve correspond to the different cut-off points that could be used to determine whether the model's output, a probability, is a positive prediction.

The area under the curve (AUC) represents the overall performance of the model. A value of 1 would correspond to a perfect model, whereas a value of 0.5 would correspond to a model with no discriminatory ability. This metric is not sensitive to the prevalence of each class.

3.4 Evaluation

3.4.1 Training, test and holdout sets

Models cannot be evaluated on the same data they were trained on because they "overfit" to the training data, i.e. they effectively memorise characteristic patterns in the training data which do not generalise to a broader population. The aim in machine learning is to develop models that will generalise to unseen data, and overfitting is a central obstacle [11].

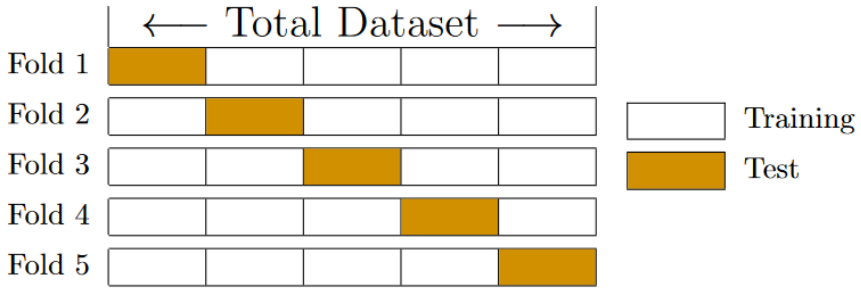


Figure 3.2: Five-fold cross-validation

To evaluate how well a machine learning model will generalise to unseen data, it is common practice to divide the available data into a training, test and holdout set. The model is iteratively trained and evaluated on the training and test sets, before finally being tested once on the holdout set. The repeated use of the training and test sets allows hyper-parameters to be tuned. However, this repeated optimisation means that information about the test data may leak into the model. Data leakage is another major obstacle in the development of machine learning models. The single use of the holdout set means that no information about that data can leak into the model, and so the model's performance on this data should be representative of the model's performance on unseen data (although if the holdout set originates from the same distribution as the training and test data, then it is not truly representative of external data) [11].

3.4.2 K-fold cross-validation

A more advanced version of this data split called k-fold cross-validation was used in this research. Where enough data was available, a holdout set was reserved from the outset. The remaining data was then used for iteratively training and testing. This data was split into `k` partitions of equal size. For each partition `i`, a model was trained on the remaining $k-1$ and evaluated on partition `i`. The final metric given is then the mean of the `k` metrics obtained. An illustrative schematic is shown in Figure 3.2.

3.4.3 Addressing overfitting

Evaluation of machine learning models frequently reveals that the model has been overfitting to the training data. There are various techniques for mitigating this effect, many of which were used in this research. Such techniques include:

- Simplifying the model (in the case of neural networks, by reducing the size of the network), as a model with more parameters has more memorization capacity [11].
- Reducing the dimensions of the input data, to reduce the impact of arbitrary patterns in the detail and enhance the impact of general patterns in the whole.
- "Early stopping" of training before convergence [85].
- Adding weight regularization: putting constraints on the complexity of the network by forcing its weights to only take small values [11].
- Adding dropout: a technique which randomly drops out a number of output features of a layer during training, thus introducing noise to break up arbitrary patterns that are not significant [86].
- Using data augmentation: artificially increasing the size of the dataset by configuring random transformations of the input data [11].

3.5 Explainability

3.5.1 Shapley Additive exPlanations

Shapley values are an idea from coalitional game theory, devised to address the issue of fair attribution in co-operative games [87]. The term was coined by Lloyd Shapley in 1953 [88] to describe a method for assigning payouts to players depending on their contribution to the total payout. For the task of interpreting machine learning models, the model's prediction is the payout, and the features are players. The

Shapley value is the average marginal contribution of a feature across all possible coalitions of features.

SHapley Additive exPlanations (SHAP) were proposed by Lundberg and Lee in 2017 [89]. An innovation of this paper was that it expressed the traditional concept of Shapley values as an additive feature attribution method: a linear function of binary variables. The additive space for probabilistic classifiers is usually taken to be the logit, which is the log-odds of the prediction [89, 90]. SHAP can be used to provide both global and individual explanations, which helped to unify the field of explainable machine learning by connecting Shapley values to local interpretable model-agnostic explanations (LIME) [91], Deep Learning Important FeaTures (DeepLIFT) [92], and Layer-Wise Relevance Propagation [93]. Another benefit of SHAP values is that they are expressed in the same units as the model prediction, which makes them intuitively accessible. SHAP is now one of the most popular methods for interpreting machine learning models, due to its flexibility, modularity and ecosystem of adaptations [94].

The SHAP explanation is specified as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where g is the explanation model, $z' \in \{0,1\}^M$ are the simplified features, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley values [89].

The open source shap library implements SHAP in Python. This includes various explainers, such as the PartitionExplainer and the DeepExplainer. The PartitionExplainer is used in this research for 2D images and text because it is particularly useful when groups of features are highly correlated or interact strongly with each other (like pixels in an image or words in a sentence). The PartitionExplainer defines a hierarchy of features for its analysis, the result of which is a set of SHAP values for each feature that take into account both the individual importance of the feature

and the importance of the feature as part of its group. The DeepExplainer is used in this research for 3D images, because it is specifically designed for deep learning models and is the fastest method for explaining these particularly computationally expensive 3D models [94].

3.5.2 SHAP for images

In the case of images, SHAP values represent the attribution of each pixel to the change of the expected model prediction when conditioning on that pixel using reference samples. In the PartitionExplainer, a masker is used to blur out pixels not in the coalition being considered. In the DeepExplainer, pixel absence is simulated by replacing pixels with pixels from a background dataset [94].

3.5.3 SHAP for text

In the case of text, words are coded as tokens, and SHAP values represent the attribution of each token to the change of the expected model prediction when conditioning on that token using reference samples. In the PartitionExplainer the absence of words is simulated by replacing these with a fixed token (e.g. "...") [94].

3.6 Concluding remarks

This chapter has described the methods used in this research, including the development environment, the model architectures, the performance metrics, the evaluation methods, and the explainability technique. The universal use of common methods demonstrates their utility and flexibility. The following chapters will outline the application of those techniques to a range of radiological tasks and data types.

Chapter 4

Magnetic Resonance Imaging safety: aneurysm clip detection

4.1 Introduction

Screening of patients for aneurysm clips and other metallic devices prior to magnetic resonance imaging (MRI) is vital to ensure that the patient and device can be scanned safely. There have been numerous makes and designs of aneurysm clip over decades [95], many of which have been categorized as MRI safe. For these particular implants, MRI is not absolutely contraindicated, but the devices need careful prior assessment to ensure that the scan takes place under manufacturer-specified conditions. However, not all historic clips are MRI safe, and even those that are safe in some conditions may not be safe in all conditions [96]. At least one fatality has been caused by the displacement of an aneurysm clip [97]. Safe examination requires review of medical records and co-ordination of multiple experts [98]. Late detection has the potential to result in last minute cancellations and wasted scanner time. Failure to perform the required checks can result in clip failure, with potentially catastrophic consequences.

MRI is the standard imaging modality for many conditions. Appropriate screening policies and procedures are essential before permitting entry to the MRI scanner

to prevent injury [99]. Best practice is to use referrer and patient questionnaires to identify patients with devices or other issues that need further investigation. Questionnaires are not fail-safe as referrer responses can be unreliable and patient responses are often not available until the day of the scan.

Deep learning has previously been used successfully to detect medical implants. Pre-trained convolutional neural networks have been used to detect pacemakers in chest radiographs with 99.67% accuracy [100] and spinal implants in lumbar spine lateral radiographs with 98.7% precision and 98.2% recall [101]. A convolutional neural network trained from scratch has been used to identify dental implants in X-ray images with 94.0% segmentation accuracy and 71.7% classification accuracy [102]. In another application, a segmentation network has been developed to identify orthopedic implants in hip and knee radiographs with 98.9% accuracy and 100% top-three accuracy, exceeding the performance of five senior orthopedic specialists [103].

This chapter describes the design of a deep learning model for the detection of the presence of aneurysm clips in computerized tomography (CT) brain scans. The vast majority of patients with aneurysm clips will have had CT brain imaging previously performed as part of their treatment or another hospital attendance, presenting the potential to screen these previous scans as part of an automatic pre-MRI safety check. This would improve MRI safety, reduce last-minute cancellations, and save time and resources.

4.2 Data

4.2.1 Subject inclusion

Data were obtained from Derriford Hospital, a large teaching hospital with a regional neurosurgery centre serving the South West of the United Kingdom. The study design was retrospective and observational using pre-existing medical image data. The date range covered was May 2011 to April 2022. A database of patients with

aneurysm clips was used to identify cases for inclusion in the study. A list of all patients undergoing aneurysm clip surgery was identified from surgical records. The radiology information system (RIS) (Cris, Wellbeing Software) was used to identify all post-surgical CT brain examinations for these patients (n=140). A custom SQL query was then used to search the RIS for matched controls (n=140). For each scan with an aneurysm clip present, a scan with no aneurysm clip present was identified. These control scans were matched according to:

- scan type
- age at time of scan, within a window of ± six months
- scan date, within a window of ± twelve months
- gender

Images for the investigations identified on the RIS were downloaded from PACS using dcmtk (OFFIS e.V.) [104]. These studies were anonymised using custom anonymisation software based on the Clinical Trials Processor (RSNA MIRC project) [105].

4.2.2 Ground truth confirmation

Manual review of images was performed by two board-certified radiologists to ensure correct labelling. In the event of any disagreement of the correct labels, a third board-certified radiologist reviewed the case to confirm the correct labelling.

4.2.3 Split

Two sets of images were extracted from the fully curated dataset: a set of localizers and a set of full CT brains. Most CT scan studies begin with one or more localizer scans. These are of poorer quality than full CT scans, but aneurysm clips can often still be clearly seen (Figure 4.1). Localizer scans acquired in the same plane were identified automatically using the DICOM tags. From the fully curated dataset, 274

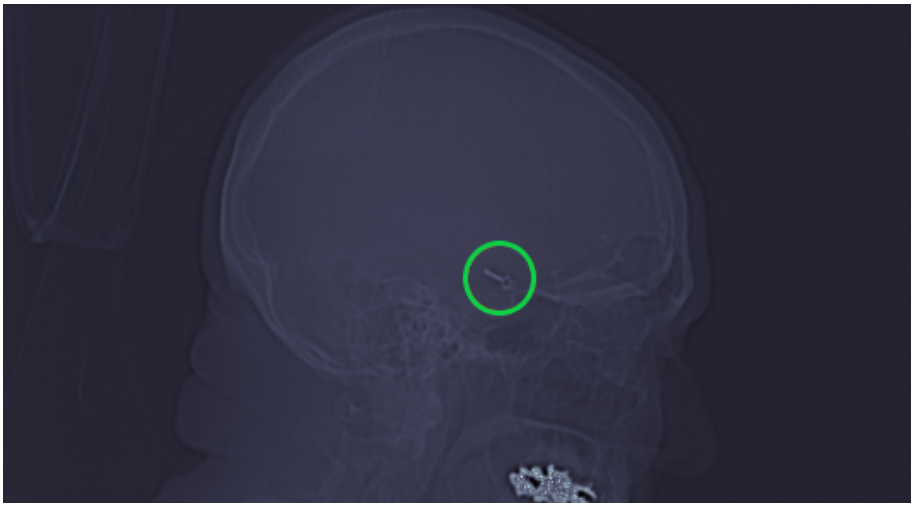


Figure 4.1: Sagittal localizer with aneurysm clip present, circled

scans were identified which contained saggital localizers: 136 with aneurysm clips and 134 without. These localizers were randomly divided at a scan level: 28 scans (10%) were reserved as a holdout set (10 with aneurysm clips and 18 without). The remaining 246 (90%) were used for model development (126 with aneurysm clips and 120 without).

To standardise the full CT brain dataset, scans reconstructed using the same kernel were identified automatically using the DICOM tags. From the fully curated dataset, 214 scans were identified which had been reconstructed using a bone kernel: 104 with aneurysm clips and 110 without. These were randomly divided at a scan level: 22 scans (10%) were reserved as a holdout set (11 with aneurysm clips and 11 without). The remaining 192 (90%) were used for model development (93 with aneurysm clips and 99 without).

For both localizers and full CT brains, five-fold cross-validation was used to develop and assess models, with the data divided into 80% training data and 20% test data in each fold.

For both types of image, the five developed models were then finally tested on the holdout set.

4.3 Image preprocessing

The images were preprocessed before model input by a deterministic automatic pipeline developed in Python using tools from OpenCV [66], SciPy [70] and scikit-image [68]. For the two-dimensional localizer scans, black borders were removed. Pixel values were rescaled between zero and one. Images were cropped to contain the head only, and the bottom of the images removed to exclude the mandible. This optimisation was included after the explainability technique revealed that models were being confounded by the presence of fillings, resulting in false positive results. Images were resized to 400×400 pixels.

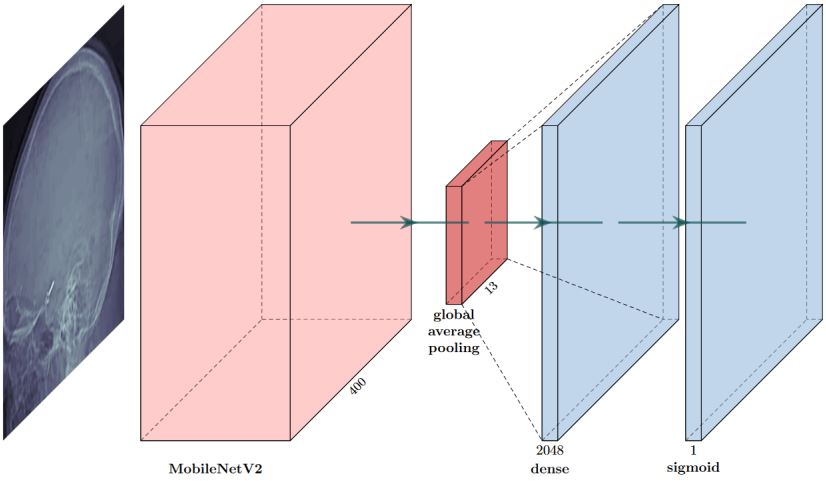
For the three-dimensional scans, the Hounsfield values were clipped with a level of 2000 and a window of 500 to optimize the visibility of metal. Pixel values were scaled between zero and one. Images were cropped to contain the head only and resized to 256×256×40 pixels.

4.4 Model development

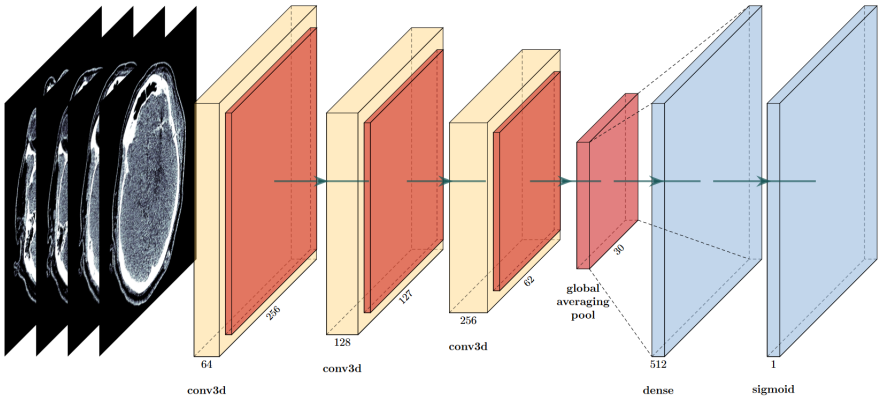
Python-based deep neural networks were built with Keras [77] using the TensorFlow backend [76]. Graphics processing unit hardware acceleration on an NVIDIA GeForce RTX 3080 was used for neural network training. Jupyter Lab [106] was used for model development to enable iterative improvements to be made efficiently.

For the classification of the two-dimensional localizer images, a convolutional neural network based on a pre-trained model was selected as a proven choice for computer vision and image classification tasks using transfer learning [23]. Several well-established pre-trained base networks were trialled, including VGG16 [107], Inception V3 [108], Xception [109], DenseNet [110] and MobileNet V2 [111]. Following analysis for each model, MobileNet V2 achieved the greatest performance and was chosen for the final models (Figure 4.2a).

For the classification of the three-dimensional CT images, a three-dimensional convolutional neural network was trained from scratch, due to a lack of available



(a) MobileNet V2 for localizers



(b) 3D-CNN for full CT brains

Figure 4.2: Network architectures

Table 4.1: Performance of different base models for localizer images

Base model	Mean ROC AUC	Parameters	Inference time (ms)	GFLOPS
VGG16	0.84	15,767,361	24.9	97.9
Inception V3	0.95	26,001,185	27.4	21.0
XCeption	0.98	25,059,881	25.5	29.4
DenseNet	0.98	22,258,241	30.7	27.4
MobileNet V2	0.99	4,883,521	26.2	2.0

pre-trained three-dimensional classification networks [112]. Several different hyper-parameter configurations were trialled. Following curve analysis for each iteration, the one which achieved the smallest loss on the validation data was chosen for the final models (Fig. 4.2b). Rectified Linear Units (ReLU) were used for the activation functions for the fully connected layers [113], and dropout of 0.3 was used before the final layer [86].

The models were trained for a maximum of 100 epochs using stochastic gradient descent with the Adam optimization algorithm (learning rate 0.001) [80]. The binary cross-entropy loss function was utilized. The batch size was 64. The images were augmented with a 50% probability of horizontal flip. Other augmentation methods were trialled, but did not result in any further increase in performance. The models achieving the lowest loss on the test sets during training were saved using checkpoints.

A classification threshold was then chosen for the models which maximized sensitivity, and therefore minimized the prevalence of false negatives.

4.5 Model evaluation

Of the pre-trained base models trialled for the localizer images, MobileNet V2 achieved the greatest mean test Receiver Operating Characteristic (ROC) area under the curve (AUC) and was chosen for the final models. Other base model results are reported in Table 4.1.

A classification threshold of 0.16 was chosen to maximize sensitivity whilst main-

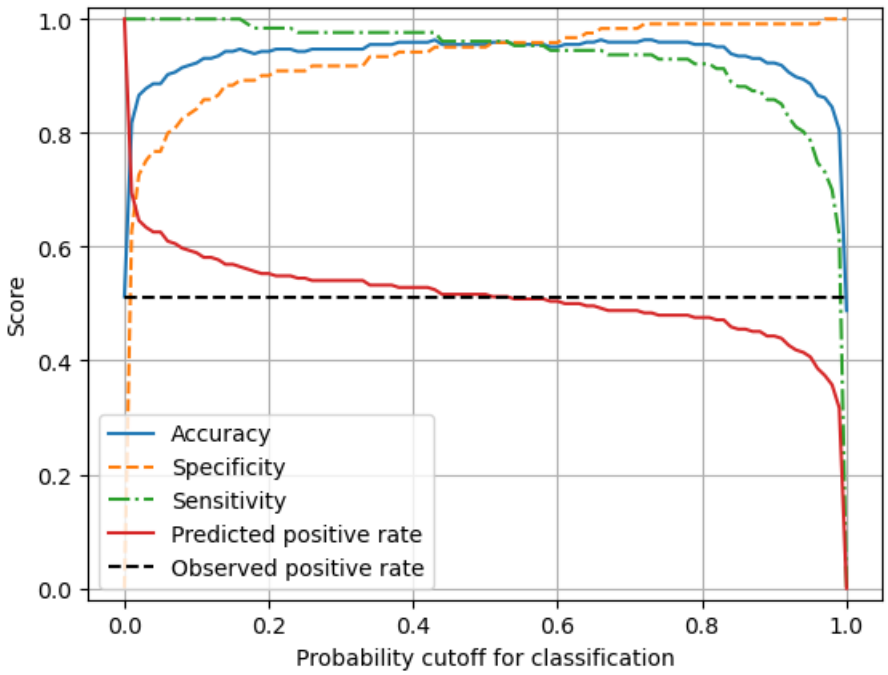


Figure 4.3: Mean test performance metrics for MobileNet V2 models in training

Table 4.2: Performance metrics for MobileNet V2 models with classification threshold of 0.16

Performance metric	Training mean	Holdout mean
ROC AUC	0.99	1.00
Accuracy	95%	82%
Sensitivity	100%	100%
Specificity	89%	82%

taining a high accuracy and specificity (Figure 4.3). The final models achieved a mean test sensitivity of 100%. Other performance metrics are reported in Table 4.2.

When tested on the holdout set of 28 localizer images, the final models achieved a sensitivity of 100%. Other performance metrics are reported in Table 4.2.

After models had been trained on three-dimensional CT images, a classification threshold of 0.30 was chosen to maximize sensitivity whilst maintaining a high accuracy and specificity (Figure 4.4). The final models achieved a mean test sensitivity of 96%. Other performance metrics are reported in Table 4.3.

When tested on the holdout set of 22 three-dimensional CT images, the final models achieved a mean sensitivity of 96%. Other performance metrics are reported

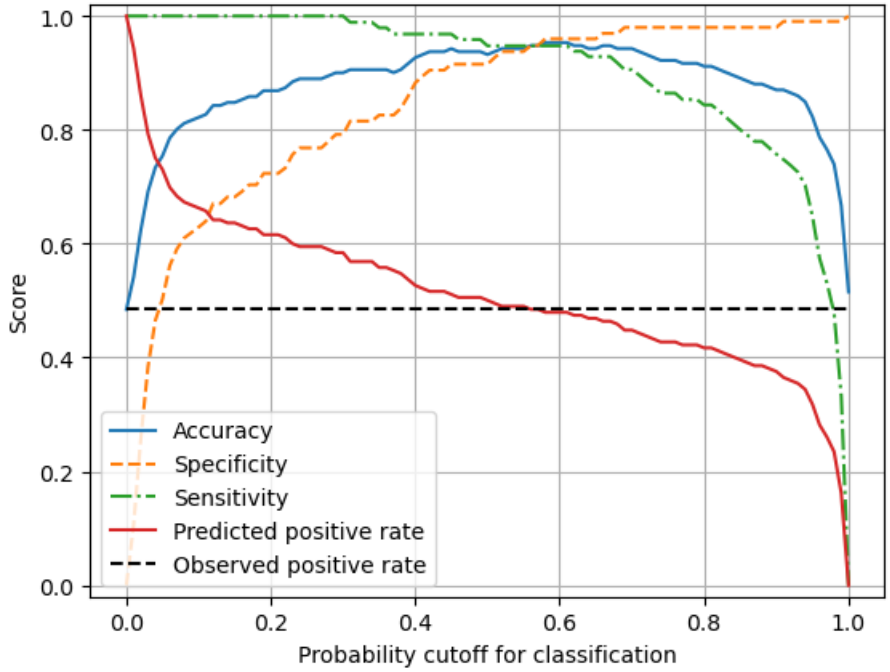


Figure 4.4: Mean test performance metrics for 3D models in training

Table 4.3: Performance metrics for 3D models with classification threshold of 0.30

Performance metric	Training mean	Holdout mean
ROC AUC	0.99	0.96
Accuracy	90%	95%
Sensitivity	100%	96%
Specificity	79%	95%

in Table 4.3. Of the 22 images, 19 were correctly classified by all five models. Of the three images that were incorrectly classified by at least one model, two were false positives and one was a false negative.

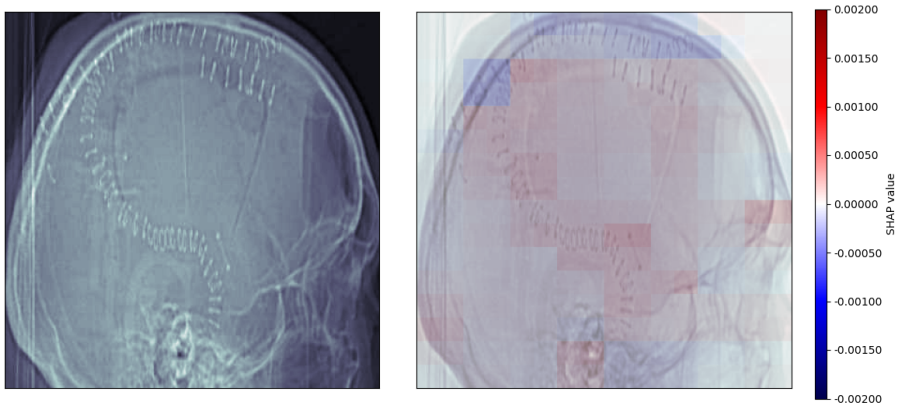
4.6 SHAP heatmaps

The incorrectly classified 2D localizer images were analysed using the SHAP explainability method. In the early stages of the research, this demonstrated the need to remove the mandible from the images, as prior to this removal the models were confounded by the presence of fillings.

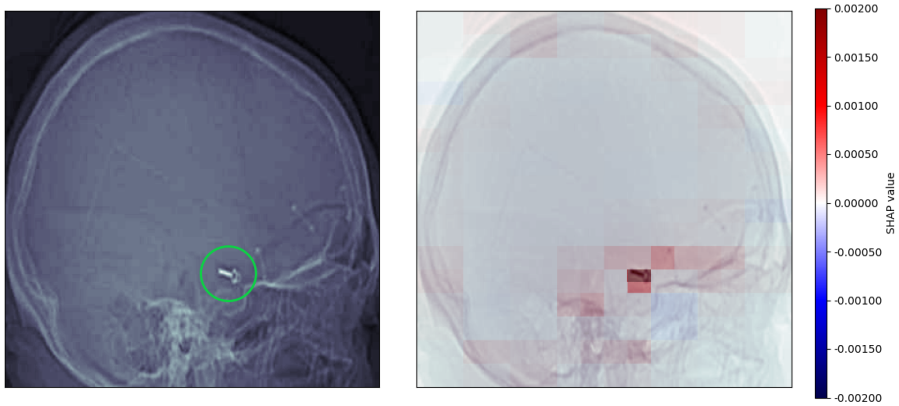
After the images had been cropped and models developed, the SHAP explainability method was used to analyse the incorrectly classified examples in the holdout test set. Three of the 28 images were incorrectly classified by all five models, and five other images were misclassified by at least one of the models. All of these errors were false positives. The average SHAP maps show that bright areas have contributed to the models' incorrect predictions, including other metal devices (Figure 4.5a). See Figure B.1 of Appendix B for all false positive average SHAP maps.

The SHAP explainability method was also used to analyse the localizer images that the models classified correctly. Of the 28 images in the holdout test set, 20 were classified correctly by all five models. The average SHAP maps for the true positives show that the pixels containing aneurysm clips contributed positively to models' correct predictions that a clip is present (Figure 4.5b). See Figure B.2 of Appendix B for all true positive average SHAP maps.

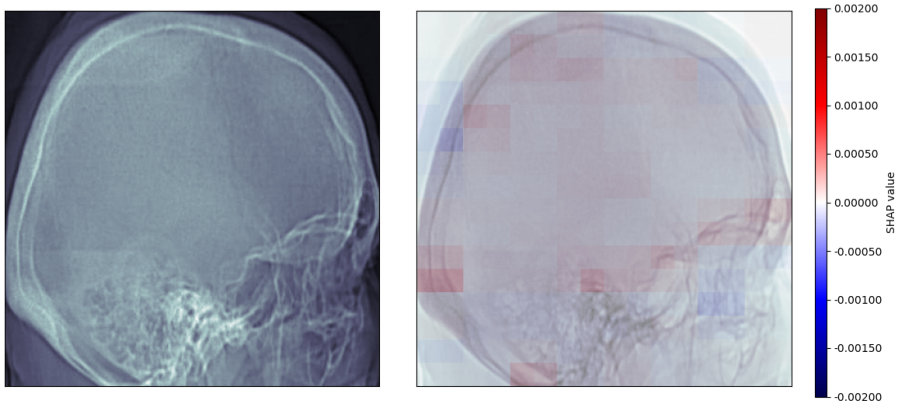
The signal was much stronger than the confounding signals in the false positive predictions, and was much stronger than any signal in the true negative predictions where no clip had been detected (Figure 4.5c). See Figure B.3 of Appendix B for all true negative average SHAP maps.



(a) False positive, as predicted by five models. The mean output probability of the image containing a clip is 0.46.



(b) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.99.



(c) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.

Figure 4.5: Maps of average SHAP values. Any pixels highlighted in red have contributed to the prediction that an aneurysm clip is present; any pixels highlighted in blue have contributed to the prediction that no aneurysm clip is present. In the case of the true positive, the aneurysm clip has been circled in green for clarity.

4.7 Discussion

The trained models exhibit excellent performance for both localizer images and full CT brain scans. Both types of model generalize well to the unseen data in the holdout sets and score particularly highly in terms of sensitivity. The sensitivity for the localizer models is 100% in both the training and the holdout data: there are no dangerous false negatives. The computational resources required to run the models are particularly low in the case of the localizer images.

The use of an explainability method is particularly valuable in this application because it demonstrates that the correct parts of the localizer image are informing the models. In general, the positive (red) signal in the images is strongly localized and more observable than the negative (blue) signal, which is weaker and more distributed. This suggests that the models are being positively informed by the presence of aneurysm clips, and are being informed on a more widespread and low level by the absence of aneurysm clips.

The use of the explainability method also enhanced the model development process. Early use of SHAP revealed that false positives were being caused by the inclusion of metal fillings. This informed the decision to develop an automatic preprocessing stage to remove the mandible from images, which resulted in better model performance.

As this application is a potential safety tool, the models have been developed and classification thresholds chosen to maximize sensitivity and minimize false negatives. As a result, they are sometimes confounded by other bright areas in the images, making some false positives likely. This could create additional work for a human operator, but it is a preferable error to dangerous false negatives. The heatmaps also demonstrate that other metal devices such as skull flap fixing plates and skin clips can be responsible for false positives (see Figure B.1 of Appendix B). These are still valuable to detect for MRI safety. Future work could assess these models on a CT brain dataset incorporating a wider range of metallic implants, to analyse

whether models trained to detect aneurysm clips specifically generalize to metal implant detection more broadly.

The detection of fillings and other implants might also indicate that the model is detecting intensity rather than the presence of aneurysm clips specifically. This may have been caused by the preprocessing stage of scaling the pixels of individual images between 0 and 1, which will have had the effect of producing detectably skewed histograms for images containing high-intensity aneurysm clips. Future work should assess whether removing or adjusting this preprocessing step results in models that are more specific to aneurysm clip detection.

It was anticipated that models developed for full CT brains might perform better than models developed for localizer scans, as the aneurysm clip would be presented in three dimensions and in greater detail. However, the performance of the three-dimensional models was slightly poorer. This may have been due to the presence of too much other confounding detail; in the case of the localizers, the use of lower-dimension input data may have mitigated against the problem of overfitting (as discussed in Section 3.4.3). The poorer performance of the three-dimensional models may also have been due to the models having been trained from scratch rather than taking advantage of pre-learned patterns. Pre-trained networks were used for the localizer scans due to their ready availability for transfer learning in two-dimensional image data. At this time, there is a notable lack of equivalent pre-trained networks available for transfer learning in three-dimensional image data. If pre-trained three-dimensional networks become available in the future, then they might be successfully leveraged in this application.

Future work could consider using an ensemble model. Ensemble methods are considered the state of the art for many machine learning applications, as they harness the power of weaker learners [114]. An ensemble model for this application could incorporate different learning algorithms, as well as bagging or boosting approaches.

The size of the data is a limitation of this research, caused by the rarity of CT scans depicting aneurysm clips. If it were possible to obtain more data this

might enable the development of even more accurate models in training, and enable more representative assessment of models in the holdout set. We have mitigated this limitation to an extent by augmenting the training data with horizontal flip, thus artificially increasing the size of the dataset.

Another limitation of this research is the lack of external validation. External validation sets are difficult to obtain as appropriate publicly available databases do not exist. We have mitigated this limitation as far as possible in this study by reserving an unseen holdout test set. However, these data originate from the same source as the training data, and the metrics reported may not be representative of the models' performance on data from a different distribution. For example, the balance of the data used in this study is not representative of the typical MRI patient population, in which only a small minority would have aneurysm clips present. An external validation set would allow for more accurate assessment of the models' capability to generalize to other populations.

4.8 Concluding remarks

A pre-trained MobileNet V2 neural network achieved high accuracy and 100% sensitivity for the detection of aneurysm clips in CT localizer scans, and the explainability method informed model development and demonstrated that the network was focusing on appropriate regions of interest in the images. A trained-from-scratch neural network also achieved high accuracy and sensitivity for the detection of aneurysm clips in full CT brain scans. This application could be a useful addition to current processes, enabling automatic safety screening for devices in advance of MRI appointments.

This chapter has provided an initial illustration of the potential application of explainable deep learning to CT imaging for patient safety. The explainability technique informed the development of the model, and allowed the validity of the model's predictions to be confirmed. The following chapters will explore the application of these techniques in a research context.

Chapter 5

Dataset curation: natural language processing for pathology detection

5.1 Introduction

The previous chapter demonstrated an application of AI for medical imaging safety, and showed how explainable AI can be used to interpret and quality check such models. The dataset was relatively easily curated: a set of CT brains containing aneurysm clips was matched with a set of CT brains not containing aneurysm clips. However, in many cases, curation of a radiological research dataset would be much more challenging. In Chapter 6, models built to differentiate Parkinson's disease (PD) scans from control scans will be discussed. Unlike the aneurysm clips task, a medical practitioner would not be able to look at standard MRI sequences and detect the presence of PD. The problem is more complicated, the potential patterns to be detected more subtle. There would appear to be a much greater risk of the model being confounded by other brain abnormalities, particularly in the case of a routinely-collected dataset, in which scans have been acquired for any number of reasons. In this chapter, the possibility of using deep learning for the automatic removal of such confounding pathology will be investigated.

Routinely acquired scans are accompanied by reports written by radiologists

which communicate findings and interpretations. We hypothesised that such reports could be used for the automatic detection of confounding pathology, and consequently the potential removal of identified scans from a radiological dataset. The raw format of radiology reports is unstructured free text, which makes them difficult to interpret and analyse computationally [115]. To convert reports into a more accessible format, studies have made use of natural language processing (NLP): a technique which allows structured information to be extracted from free text [116]. Extracted structured data has been used for a variety of purposes, such as diagnosis surveillance, case retrieval, quality assessment, patient prioritisation, and research cohort selection [115, 117, 118]. This field has expanded in recent years due to significant developments in deep learning techniques [115]. Prior to this, rule-based NLP techniques were used, which were effective for radiology report classification but required extensive manual development [119].

In previous studies which have used NLP for research dataset curation, cohorts have been positively identified for inclusion. For example, studies have identified patients with pneumonia [120], pulmonary nodules [121], pulmonary embolisms [122], abdominal aortic aneurysm [123], liver disease [124], hepatocellular cancer [125] and ureteric stones [126]. These applications have often been followed by manual case validation and data collection. In this study, by contrast, imaging data has been extracted for an existing clinical database, and we investigated whether NLP could be applied to this imaging dataset for the exclusion of potential sources of confounding data. This approach holds particular promise for the development of effective supervised machine learning models, which rely on highly curated labelled datasets [127]. Automatic removal of scans that contain confounding pathology may improve the quality of the dataset, enhancing the performance and generalisability of developed models.

5.2 Data

5.2.1 Subject inclusion

A database of PD patients was used to identify cases for inclusion in the study. The radiology information system (RIS) was used to identify all MRI brain imaging for PD patients. A custom database query was then used to search the RIS for matched controls. For each scan from a PD patient, two control scans were identified. All scans had been routinely acquired for any number of reasons. Scans were matched according to:

- scan type
- age at time of scan, within a window of ± six months
- scan date, within a window of ± twelve months
- biological sex

The final dataset contained a total of 2038 reports: 705 for scans from PD patients and 1333 control scans (Table 5.1). A non-identifiable unique identifier was assigned to each report.

Table 5.1: Demographic summary of PD and control cohorts

Cohort	Count	Median age (interquartile range)	Male/female percentage split
Parkinson's	705	69 (62-79)	62/38
Control	1333	70 (64-76)	61/39
Combined	2038	70 (63-76)	63/38

5.2.2 Ground truth confirmation

Manual review of reports was performed by two radiologists. Reference guidance on report labelling was produced to promote consistent technique between members

of the labelling team. In the event of any disagreement of the correct labels, a third member of the clinical research team reviewed the case to confirm the correct labelling.

The task was formulated as a multi-class classification problem. Each report was given a label for abnormality. The three labels given were "normal" (n=350), "abnormal" (n=1531) or "not enough information" (n=157). A demographic summary is provided in Table 5.2.

Table 5.2: Demographic summary of reports by abnormality classification

Classification	Count	Median age (interquartile range)	Male/female percentage split	PD/control percentage split
Normal	350	63 (54-70)	61/39	45/55
Abnormal	1531	71 (66-77)	63/38	32/68
Not enough information	157	68 (62-72)	63/37	34/66

Each report was also given a label for small vessel disease. The three labels given were "no small vessel disease" (n=1009), "small vessel disease" (n=869) or "not enough information" (n=160). Small vessel disease was isolated for its own category as it was by far the most common pathological finding in the dataset. It is known to be a common finding in brain imaging, particularly in elderly subjects [128] A demographic summary is provided in Table 5.3.

Table 5.3: Demographic summary of reports by small vessel disease classification

Classification	Count	Median age (interquartile range)	Male/female percentage split	PD/control percentage split
No small vessel disease	1009	67 (57-73)	61/39	32/68
Small vessel disease	869	74 (69-79)	62/38	38/62
Not enough information	160	68 (63-73)	62/38	32/68

5.2.3 Split

80% of the reports (n=1630) were used to train and develop models. Five-fold cross-validation was used, with the data divided into 80% training data and 20% test data in each fold. The five final developed models were tested on the remaining holdout set containing 20% of samples (n=408).

5.3 Text preprocessing

Reports were preprocessed to standardise text prior to model input. Punctuation and special characters were removed using pandas [63]. Text was converted to lowercase and tokenised using Keras [77].

5.4 Model development

Python-based deep neural networks were built with Keras using the TensorFlow backend [76]. One-dimensional convolutional neural networks were trained from scratch to classify the reports. Several different hyperparameter configurations were trialled. Following curve analysis for each iteration, the one which achieved the smallest loss on the validation data was chosen (Figure 5.1). ReLU was used for the activation functions for the fully connected layers [113], and dropout of 0.2 was used before the final layer [86].

The models were trained for a maximum of 100 epochs using stochastic gradient descent with the Adam optimization algorithm (learning rate 0.001) [80]. Early stopping with a patience of 50 epochs was used [85]. The sparse categorical cross-entropy loss function was utilized. The labels were encoded as integers using the LabelEncoder from scikit-learn, which does not introduce any ordinal relationship between labels [75]. The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) was calculated by generating a one-vs-the-rest ROC curve per class, and then providing the mean of the three one-vs-the-rest AUCs.

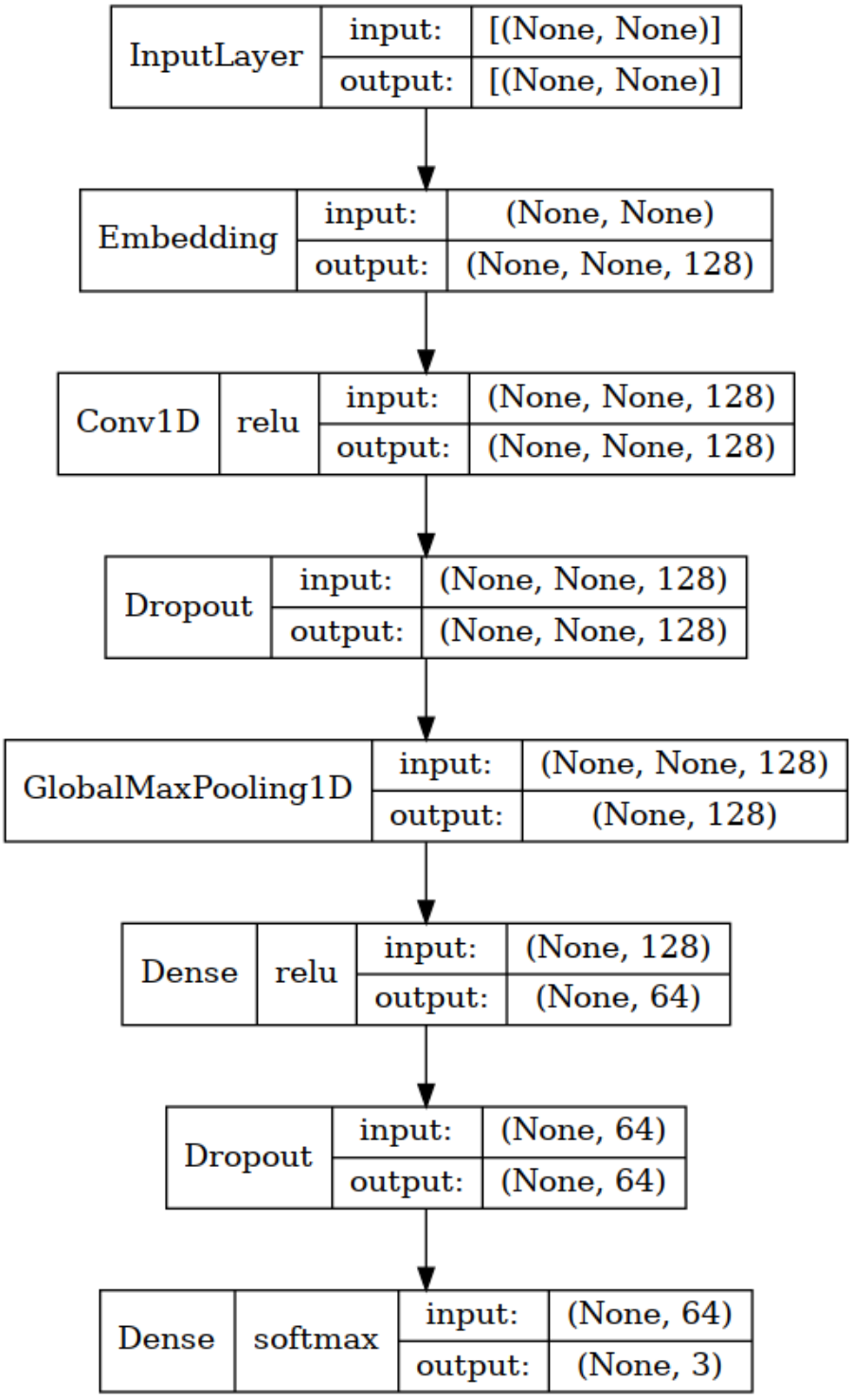


Figure 5.1: Network architecture

Table 5.4: Performance metrics for the models developed to classify radiology reports

		Mean Receiver Operating Characteristic (ROC) Area-Under-the-Curve (AUC) [95% CI]	Mean balanced accuracy [95% CI]
Abnormality	Training	0.97 [0.96, 0.98]	0.84 [0.82, 0.85]
	Holdout	0.97 [0.96, 0.97]	0.82 [0.79, 0.86]
Small vessel disease	Training	0.97 [0.96, 0.99]	0.84 [0.81, 0.88]
	Holdout	0.98 [0.97, 0.98]	0.87 [0.85, 0.88]

5.5 Model evaluation

5.5.1 Abnormal scans

In the training set, the final models achieved a mean test ROC AUC of 0.97 and a mean test balanced accuracy of 0.84 (Table 5.4). When tested on the holdout set, the five final models achieved a mean ROC AUC of 0.97, and a mean balanced accuracy of 0.82 (Table 5.4).

5.5.2 Small vessel disease

In the training set, the final models achieved a mean test ROC AUC of 0.97, and a mean test balanced accuracy of 0.84 (Table 5.4). When tested on the holdout set, the five final models achieved a mean ROC AUC of 0.98, and a mean balanced accuracy of 0.87 (Table 5.4).

5.6 SHAP plots

5.6.1 Abnormal scans

SHAP was used to explain the models' predictions in the holdout set. The most informative words are shown in Figure 5.2. Of the 408 reports, 340 (83%) were correctly classified by all five models (an example is shown in Figure 5.3) and eighteen (4%) were misclassified by all five models (an example is shown in Figure 5.4).

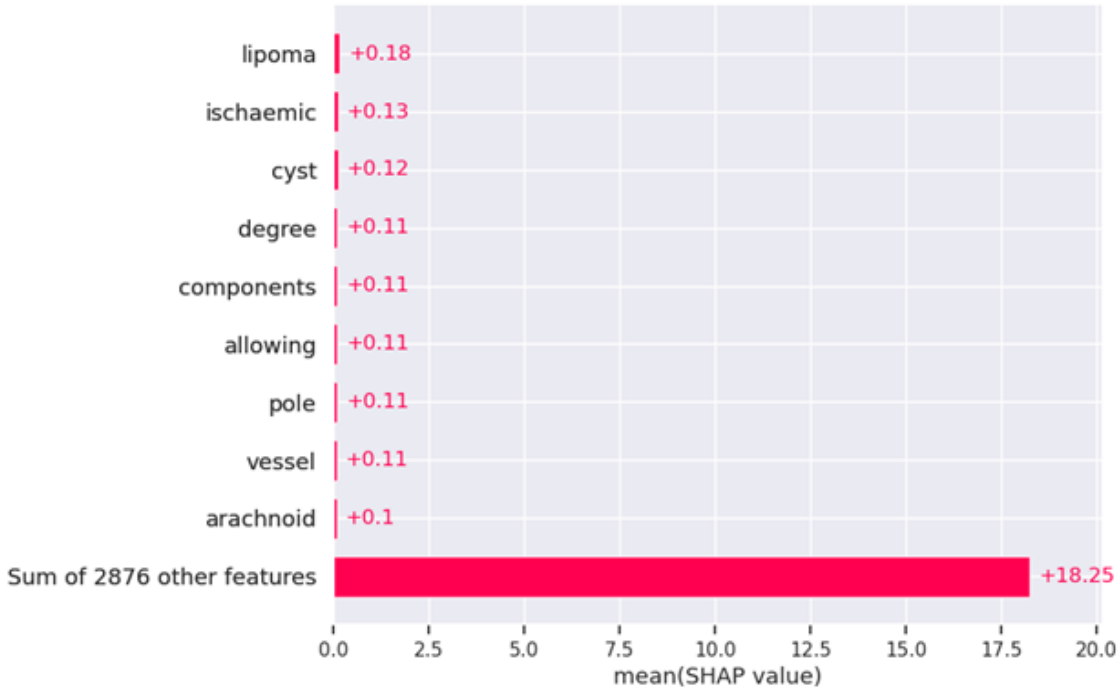


Figure 5.2: Probability SHAP values for words contributing most to the "abnormal" label

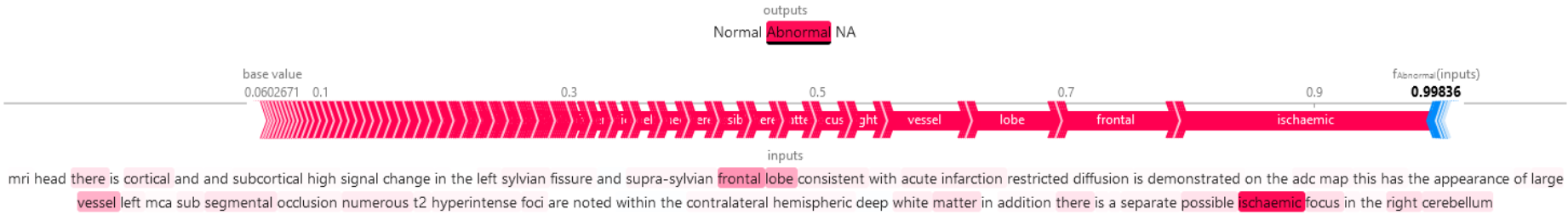


Figure 5.3: Probability SHAP values for an “abnormal” report correctly classified by all five models. Words highlighted in red have positively contributed to the model’s prediction; words highlighted in blue have negatively contributed to the model’s prediction.

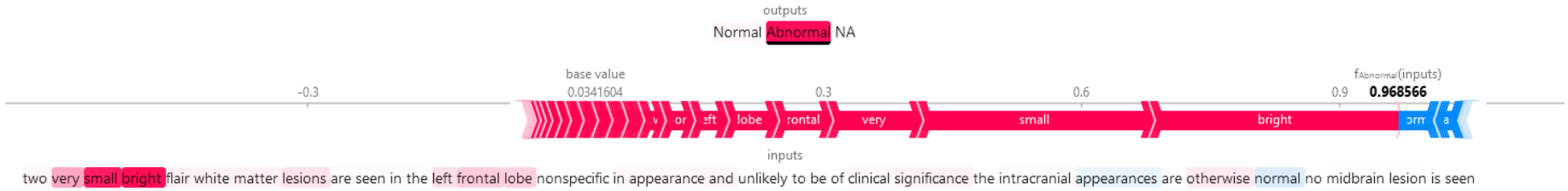


Figure 5.4: Probability SHAP values for a “normal” report misclassified as “abnormal" by all five models. Words highlighted in red have positively contributed to the model’s prediction; words highlighted in blue have negatively contributed to the model’s prediction.

5.6.2 Small vessel disease

SHAP was used to explain the models' predictions in the holdout set. The most informative words are shown in Figure 5.5. Of the 408 reports, 363 (89%) were correctly classified by all five models (an example is shown in Figure 5.6) and fifteen (4%) were misclassified by all five models (an example is shown in Figure 5.7).

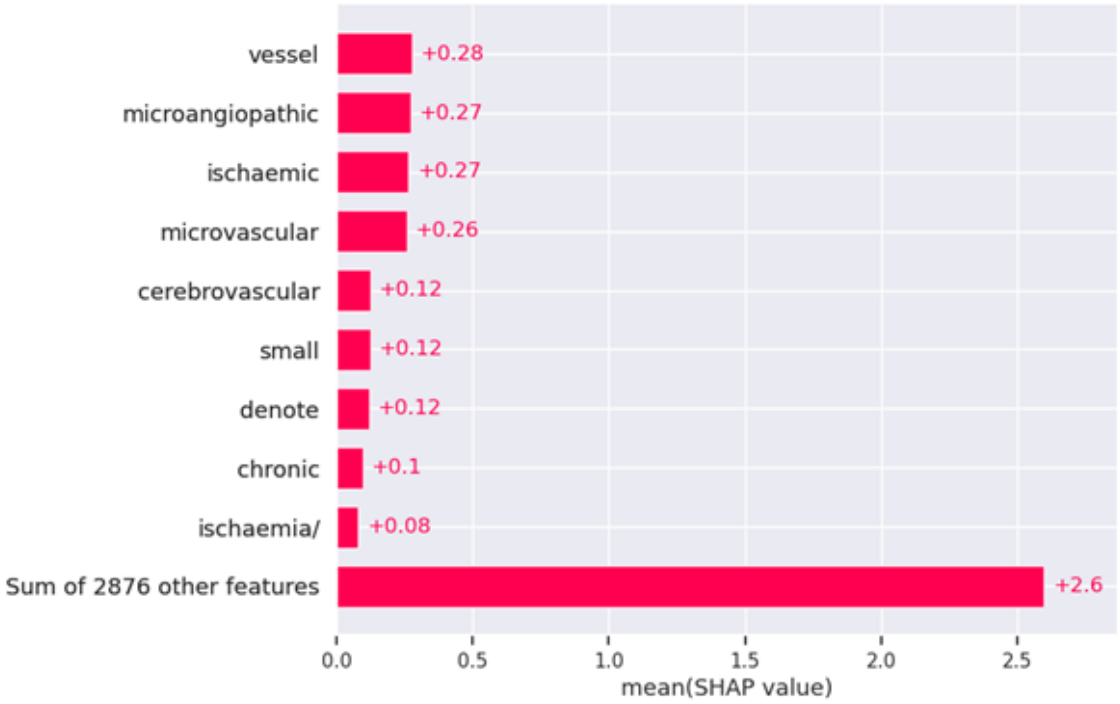


Figure 5.5: Probability SHAP values for words contributing most to the "abnormal" label

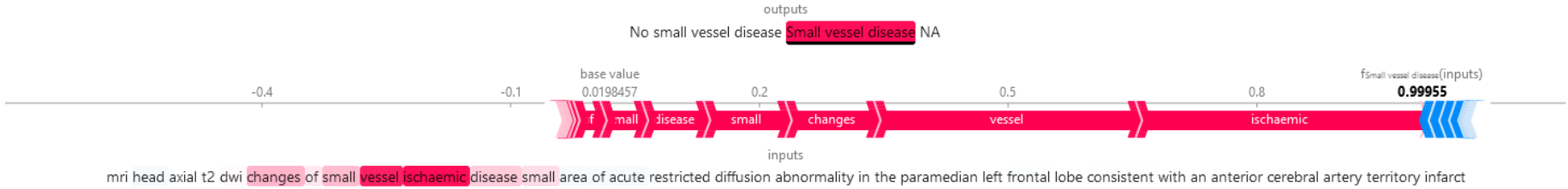


Figure 5.6: A “small vessel disease” report correctly classified by all five models. Words highlighted in red have positively contributed to the model’s prediction; words highlighted in blue have negatively contributed to the model’s prediction.

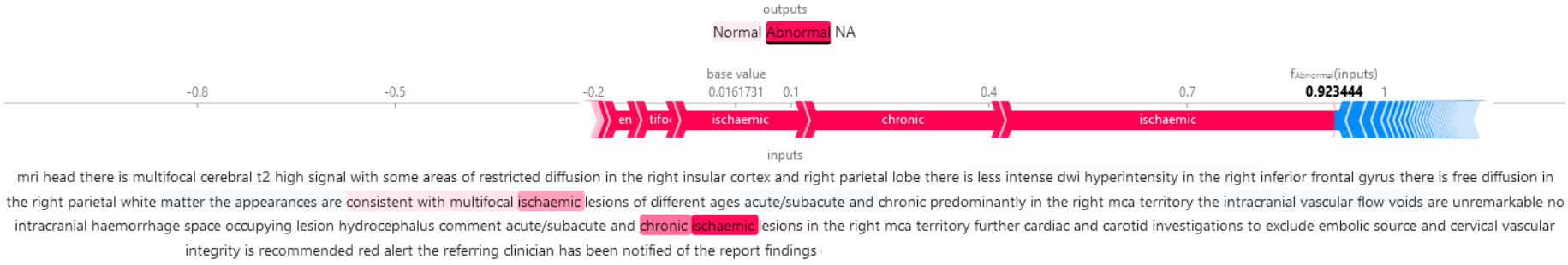


Figure 5.7: A “no small vessel disease” report misclassified as “small vessel disease" by all five models. Words highlighted in red have positively contributed to the model’s prediction; words highlighted in blue have negatively contributed to the model’s prediction.

5.7 Discussion

The deep learning models demonstrated excellent performance in classifying both abnormality and small vessel disease in radiology reports. The SHAP explainability method highlighted that relevant words were being used by the models to make these predictions, such as "lipoma" and "cyst" in the case of abnormality classification, and "vessel" and "ischaemic" in the case of small vessel disease classification. These words can also occasionally confound the models, for example when they are later negated (Figures 5.4 and 5.7). Future work could investigate whether other model architectures are more sensitive to these contexts. As this application is intended to exclude confounding data from research datasets, future work could also consider tuning the models to increase sensitivity to particular abnormalities dependent on the pathophysiology of the disease under investigation.

A limitation of this study is the lack of external validation, as such datasets are difficult to obtain. We have mitigated this limitation as far as possible in this study by reserving an unseen holdout test set. However, as these data originate from the same source as the training data, the metrics reported may not be representative of the models' performance on data from a different distribution. For example, reporting styles are likely to vary between institutions. Additionally, these data were drawn from a population of PD patients and matched controls, resulting in a dataset that is demographically unrepresentative of a general population, as there are more male than female patients and PD commonly presents later in life [129]. These scans are likely to represent the spectrum of pathology in that age group, but the older population means that there is likely to be more pathology present in this dataset than in a more random selection of scans. Within this selection, there are also demographic differences between classes which should be noted. The "abnormal" and "small vessel disease" scans are from an observably older population (see Tables 5.2 and 5.3). Future work could assess whether this age bias has had a confounding effect on the models, and could consider whether it might be helpful to include age

explicitly as a separate variable.

Nonetheless, the demonstrated accuracy of this technique makes it useful for automated processing of radiology data as part of an image analysis pipeline, allowing automated removal of investigations that may contain confounding data. In the context of the PD project, for example, these classifications could be used to automatically remove confounding "abnormal" or "small vessel disease" scans, to investigate whether this data cleaning improves the performance of PD classification models. In the future this technique may alleviate the need for costly ground-truthing of scans by researchers.

5.8 Concluding remarks

This chapter has detailed a potential use of explainable deep learning for radiological data curation, discussing explainable models developed for the detection of confounding pathology in MRI brain scans. The use of an explainability method verified the validity of the models' predictions. The next chapter will investigate whether these techniques can be used to uncover new insights in such research datasets.

Chapter 6

Parkinson's disease imaging: new insights using explainable AI

6.1 Introduction

The previous two chapters have demonstrated the utility of explainable deep learning for MRI safety (Chapter 4) and for automatic cleaning of radiological datasets (Chapter 5). These were tasks that would be achievable by medical practitioners, but training AI models to perform them might allow such processes to be automated and resources to be saved. In this chapter, we will discuss the potential of applying AI to a problem that is not currently achievable by medical practitioners: the detection of Parkinson's disease in standard MRI sequence brain imaging. This research tests the hypothesis discussed in Section 2.2.1 that AI might be able to detect patterns inaccessible to the human visual system, and investigates the utility of explainable AI in elucidating novel findings.

6.1.1 Parkinson's disease

Parkinson's disease (PD) is a neurodegenerative condition resulting from a complex array of disease mechanisms which cause the accelerated death of predominantly dopaminergic neurons [130]. The noradrenergic, serotonergic and cholinergic systems

are also affected [131]. It is a progressive movement disorder with a mean age of onset of 55 [131], and manifestation in motor symptoms such as resting tremor, bradykinesia, rigidity and postural instability [132], as well as a variety of non-motor symptoms such as mood disorders, cognitive dysfunction, pain and sensory dysfunction [133].

The condition was first characterised in 1817 by Dr. James Parkinson in *An Essay on the Shaking Palsy*, which detailed his observations of six men who exhibited tremors, bent postures, unusual gaits, and a tendency to fall [134]. Fifty years later, Dr. Jean-Martin Charcot added slowness of movement and stiffness to the list of symptoms of "la maladie de Parkinson", and observed that not all patients experienced tremors [135]. In the intervening years much has been discovered about the pathology of PD, but the underlying cause of neuron death remains unclear [136].

PD is the fastest growing neurological disorder in the world [137]. In 1855 approximately twenty-two people died of PD in England and Wales [138]; by 1990 there were 2.6 million cases globally, and by 2015 there were 6.3 million [139]. By 2020 this had increased again to an estimated 9.4 million people living with PD [140].

6.1.2 Neuropathology

The pathological hallmarks of PD are the loss of nigrostriatal dopaminergic neurons and the presence of Lewy bodies, discovered by Frederic Lewy in 1912 [141] and named by Constantin Tretieakoff in 1919 after he observed them in autopsied parkinsonian brains [142]. A major component of Lewy bodies is α-synuclein, a protein which has been associated with PD since 1997, when Dr. Mihael Polymeropoulos and colleagues identified that the p.A53T pathogenic variant of the α-synuclein gene gives rise to a form of familial PD [143]. Although this association has been established, the physiological function of α-synuclein in the pathogenesis of PD remains unknown. Lewy bodies are understood to appear sequentially in PD, first in the dorsal motor nucleus and olfactory nucleus, and spreading to the substantia nigra

pars compacta [144].

The substantia nigra, an area of the brain responsible for movement, is normally dopamine-rich. In 1958, Dr. Arvid Carlsson established dopamine's function as a neurotransmitter, and demonstrated the effect of lowering dopamine in mammalian brains and subsequently the effect of administering levodopa [145]. This paved the way for the discovery that levodopa could be effectively used to treat parkinsonian symptoms, and for the discovery that dopamine levels in the brains of people with PD are severely reduced [146]. The loss of dopaminergic neurons, which normally contain significant amounts of neuromelanin [147], explains the classic post mortem finding of substantia nigra depigmentation in the brains of people with PD. Neurodegeneration and Lewy body formation are found in the noradrenergic, serotonergic and cholinergic systems, as well as in the cerebral cortex, olfactory bulb, and autonomic nervous system [148].

6.1.3 Diagnosis

Historically, Lewy bodies could only be observed post mortem; there was no definitive diagnostic test for their presence. Recently, however, the seed aggregation assay test has been introduced for the detection of abnormal α-synuclein oligomers in serum. This is a promising step towards improving the diagnosis and management of synucleinopathies. Prior to this development, PD diagnosis had been based on clinical criteria, primarily the observation of motor symptoms: asymmetrical resting tremor, slowness of movement (bradykinesia), rigidity, and clinical improvement following administration of dopaminergic therapy [132]. Such features are included in the diagnostic criteria developed by the UK Parkinson's Disease Society Brain Bank [149] and by the National Institute of Neurological Disorders and Stroke [150]. In 2015 the Movement Disorder Society updated their diagnostic criteria to include non-motor symptoms (such as sleep dysfunction, autonomic dysfunction, hyposmia and psychiatric dysfunction) in addition to the motor symptoms [151].

Differentiating PD from other forms of parkinsonism can pose difficulties, espe-

cially in the early stage of the disease. As well as PD, there are atypical parkinsonian conditions including dementia with Lewy bodies, multiple system atrophy, progressive supranuclear palsy and corticobasal syndrome. These disorders are pathologically characterised by the abnormal deposition of the proteins α-synuclein and tau. The sites of these depositions result in differing symptoms that can overlap with symptoms of PD [152]. The diagnostic accuracy for PD is between 80% and 90%, and even lower in early-stage disease [153].

There is an ongoing search for reliable biomarkers for PD, to allow it to be distinguished from other conditions and to allow its progression to be monitored [154]. Candidates include protein biomarkers, dopamine metabolites, amino acids and other compounds found in blood, serum, and cerebrospinal fluid (CSF) [154]. Potential biomarkers may also be found in neuroimaging.

6.1.4 Imaging

The only robust diagnostic imaging test for PD is a dopamine active transporter (DAT) scan, which uses single-photon emission computed tomography (SPECT). The radioactive tracer attaches to the dopamine transporter found on dopaminergic neurons, and a visual interpretation of the scan can then be used to distinguish between normal binding and reduced or absent binding. It is particularly useful in differential diagnosis, as it can be used to distinguish between the nigrostriatal dopaminergic degeneration of PD and the non-nigrostriatal degeneration of aytpical parkinsonism [155].

Recently there have been some advances in using structural imaging in PD. Diagnosis has been augmented using high-field magnetic resonance imaging (MRI) with accelerated acquisition combined with new sequences [156]. Imaging of the substantia nigra has previously been difficult due to its low contrast in standard T1 and T2-weighted MRI [157], but new sequences sensitive to iron and nigral pigments have allowed for the assessment of pathological surrogates such as loss of dorsal nigral hyperintensity and increased nigral iron content. Iron-related changes

have also been observed using transcranial sonography [158]. In addition, diffusion-weighted imaging has been used to assess the alteration of nigral diffusivity seen in PD [159]. Outside of the substantia nigra, new methods in network analysis have allowed the tracking of subtle changes across the brain by analysing co-varying atrophy in cortical and subcortical structures [160].

6.1.5 Prodromal Parkinson's disease

PD has a prodromal stage: a period during which neurodegeneration has begun, but the motor symptoms that would allow clinical diagnosis are not defined [161]. The basis for this non-motor prodrome is that the pathologic process may not start in the substantia nigra [144]. In prodromal PD, patients experience a variety of non-motor symptoms, such as hyposmia, rapid eye movement (REM) sleep behaviour disorder (RBD), autonomic dysfunction, depression, visual changes and cognition changes [161]. These symptoms can precede diagnosis by a decade or more [162]. The speed of the progression from prodromal PD to the full clinical stages varies among patients and cannot be reliably predicted [163]. As dopaminergic deficiency is present by the time that motor symptoms appear and a diagnosis can be made, it is evident that progressive nigral and extra-nigral neurodegeneration must take place in this prodromal phase. Studies now indicate that at the time of diagnosis, up to 80% of dopaminergic neurons within the basal ganglia have degenerated [164–167].

The detection of prodromal PD holds great promise for disease treatment. If effective therapeutic interventions (such as promising prospective neuroprotective compounds) could be administered at this early stage of disease development, the death of the dopaminergic neurons could be prevented or even reversed, and thus the most debilitating features of PD avoided.

At present there is no test for prodromal PD, although some research criteria have been proposed (including markers such as RBD, olfactory loss and constipation, as well as risk factors such as having a relative with PD and pesticide or solvent exposure) [163]. Imaging biomarkers hold promise for prodromal PD detection, as

structural imaging could conceivably be used to detect the neurodegeneration which preceeds the onset of motor symptoms. Studies in cohorts considered at risk of developing PD have demonstrated visible brain changes in DAT scans [168, 169], positron emission tomography (PET) scans [169, 170], SPECT scans [171], and ultrasound scans [172, 173], but it is yet to be seen whether such brain changes would be visible in a retrospective cohort of patients who later received a diagnosis of PD.

6.1.6 Prodromal imaging biomarkers: challenges

Despite the promise of prodromal imaging biomarkers, various potential modalities entail significant clinical disadvantages. DAT and PET scans are cumbersome and expensive tests requiring the administration of a radioactive tracer, making them unsuitable for a population level screening strategy. If they were able to detect prodromal brain changes, routine brain imaging modalities such as standard sequence MRI would be more practically viable and clinically useful. MRI scans are often ordered following the onset of memory problems, and are becoming increasingly common. However, standard MRI sequences have not been used for PD detection, as they have not appeared to show the associated brain changes.

The task of assessing imaging biomarkers for prodromal PD also encounters the difficulty of gathering a suitable cohort. As there is no test for prodromal PD, the only cohorts which can be prospectively recruited are those considered at risk of developing PD (due to genetic factors or presence of non-motor symptoms, for example), and there can be no certainty if or when they might later be diagnosed. To assess imaging biomarkers in a cohort of confirmed prodromal PD patients would require the curation of a retrospective cohort, which would entail collecting past pre-diagnosis imaging data from a cohort of PD patients and a cohort of matched controls.

6.1.7 Genetic risk factors

Much research into the prodromal phase of PD has focussed on those with the PD genetic risk factors glucocerebrosidase (*GBA*) and Leucine-rich repeat kinase 2 (*LRRK2*). These risk factors are distinct from "monogenic" forms of PD, which are rare, transmit PD in a Mendelian manner with near 100% penetrance and can manifest as early as the third decade. Conversely *GBA* and *LRRK2* have a penetrance of 8–10% and 28–74% respectively [174, 175]. *LRRK2* PD manifests at an average age of 59.4 [176], and is thought to progress more slowly, often with a milder tremor predominant phenotype. *GBA* manifests at an average age of 55.8 [176], is associated with more cognitive/neuropsychiatric symptoms and tends to progress more rapidly [177]. Disease phenotype in *GBA* appears to be variant dependent. Based on a classification of symptoms documented in cases of the autosomal recessive lysosomal storage disorder Gaucher disease (caused by *GBA* variants in a biallelic state), *GBA* risk variants can be classified as "severe", "mild" and non Gaucher causing PD risk variants (from here abbreviated to "PD risk variants") [177].

6.2 Data

6.2.1 Parkinson's Progression Markers Initiative

The Parkinson's Progression Markers Initiative (PPMI) is an international observational study conducted by the Michael J. Fox Foundation, recruiting patients through outpatient neurology practices at academic centres in Austria, Canada, France, Germany, Greece, Israel, Italy, the Netherlands, Norway, Spain, the UK, and the USA, with the goal of identifying clinical and biological markers of disease heterogeneity and progression in PD [178]. The PPMI study is registered with ClinicalTrials.gov (number NCT01141023). Detailed information about inclusion criteria, informed consent, demographic data, and study design can be found on the PPMI website.

Participants in this study were included in one of four cohorts: idiopathic PD (IPD - non-carriers of genetic variants associated with PD), healthy controls, manifesting carriers (*GBA* PD or *LRRK2* PD) and non-manifesting carriers of *GBA* (*GBA* nPD) and *LRRK2* (*LRRK2* nPD) risk variants. The diagnosis for each group was made by site investigators who are movement disorder specialists and confirmed by a central consensus committee review. The PPMI study was approved by the institutional review board at each site, and participants provided written informed consent.

At baseline all PPMI subjects underwent a non-contrast enhanced T2-weighted brain MRI using a 1.5 or 3 Tesla scanner, and a non-contrast enhanced 3D volumetric T1-weighted brain MRI.

Image acquisition

All available MRI studies (n=5988) were downloaded from the PPMI website on 25 November 2020, along with demographic and clinical data, including genetic status and date of PD diagnosis. From this full MRI dataset, all T2-weighted axial scans were identified automatically using MRI parameters (Echo Time, Repetition Time) contained within the DICOM tags. These scans contained twenty-four unique sequence descriptions; most were described as "Axial PD-T2 TSE FS". In all cases voxels were anisotropic, with the most common dimensions being 0.94×0.94×3mm. Twenty-eight unique institution names were present in the DICOM tags; in a minority of cases the institution was not recorded.

Data organisation

The data were grouped into pairs of cohorts for the constructing of binary classification models. Cohorts to be compared were matched by age and sex. Ten-fold cross-validation was used to develop and assess models, with the data divided into 90% training data and 10% test data in each fold. As many subjects have contributed more than one scan to the dataset, scans were grouped by subject before

being divided so that the same subject never appeared in both the training and the test data. In most cases, a holdout dataset was not reserved as the cohort sizes were small and available training data needed to be maximised. However, to assess the possible impact of overfitting in the cross-validation strategy, a model trained for the largest cohort (all IPD scans and matched controls) was tested on a reserved holdout set which comprised 20% of the scans.

6.2.2 University Hospitals Plymouth NHS Trust

Routinely collected National Health Service (NHS) data were acquired from the University Hospitals Plymouth NHS Trust (UHPNT). This has a secondary care catchment population of 475,000 [179]. The area served, the South West of England, is the oldest population in the UK with a high disease prevalence.

Image acquisition

A database of PD patients was used to identify cases for inclusion in the study. The radiology information system (RIS) was used to identify all MRI brain imaging for PD patients. A custom database query was then used to search the RIS for matched controls. For each scan from a PD patient, three control scans were identified. All scans had been routinely acquired for any number of reasons. Scans were matched according to:

- scan type
- age at time of scan, within a window of ± six months
- scan date, within a window of ± twelve months
- biological sex

From the full MRI dataset, all T2-weighted axial scans were identified automatically using MRI parameters (Echo Time, Repetition Time) contained within the DICOM tags. Scan sequence descriptions were mostly absent from the DICOM

tags, but six unique descriptions were present, of which the most common was "*ep_b0". In all cases voxels were anisotropic, with the most common dimensions being 0.45×0.45×5mm. The scans were all acquired from one institution. The final dataset consisted of 244 case scans (from 203 patients) and 744 control scans (from 724 patients). Date of diagnosis was also acquired where available for PD patients.

Data organisation

The data were grouped into pairs of cohorts for the constructing of binary classification models. Ten-fold cross-validation was used to develop and assess models, with the data divided into 90% training data and 10% test data in each fold. As some patients have contributed more than one scan to the dataset, scans were grouped by patient before being divided so that the same patient never appeared in both the training and the test data.

6.3 Image preprocessing

Scans were skull-stripped using the FMRIB Software Library [180]. The output of the skull-stripping pipeline was manually audited, and parameters adjusted for optimal results. All further preprocessing was carried out using Python. Pixel values were clipped to the 2.5–97.5% range, to minimise the influence of extreme outliers. Volumes were cropped to the outermost dimensions of the brain and resized to 32×32×16 pixels. Pixel values were scaled between zero and one.

6.4 Model development

Python-based deep neural networks were built with Keras [77] using the TensorFlow backend [76]. Graphics processing unit hardware acceleration was used for neural network training.

For each pair of cohorts, a three-dimensional convolutional neural network was trained from scratch, due to a lack of available pre-trained three-dimensional clas-

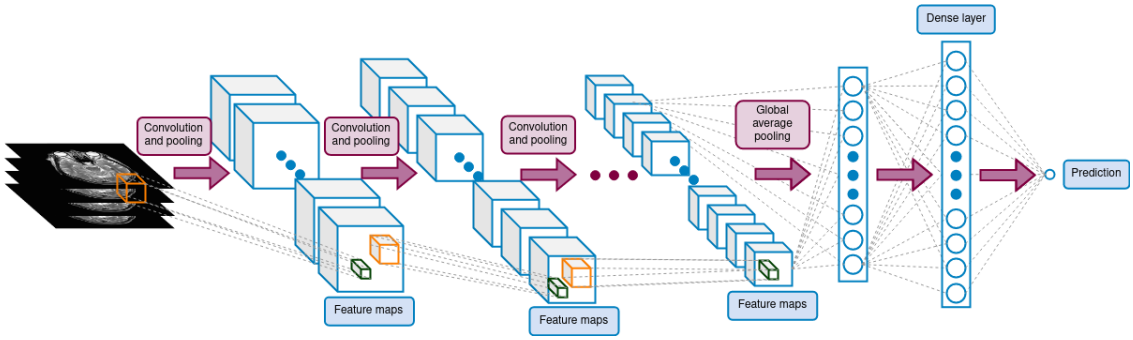


Figure 6.1: 3D convolutional neural network architecture overview

sification networks. To approximate the optimal network structure for these data, different hyperparameter configurations were trialled in the early stages. These hyperparameters were tuned following curve analysis at each iteration. Once no further reductions in the validation loss could be achieved, the hyperparameter configuration was finalised, and this architecture was used for all models (Figures 6.1 and 6.2). ReLU was used for the activation functions for the fully connected layers [113], and dropout of 0.2 was used after each convolution block and before the final layer [86].

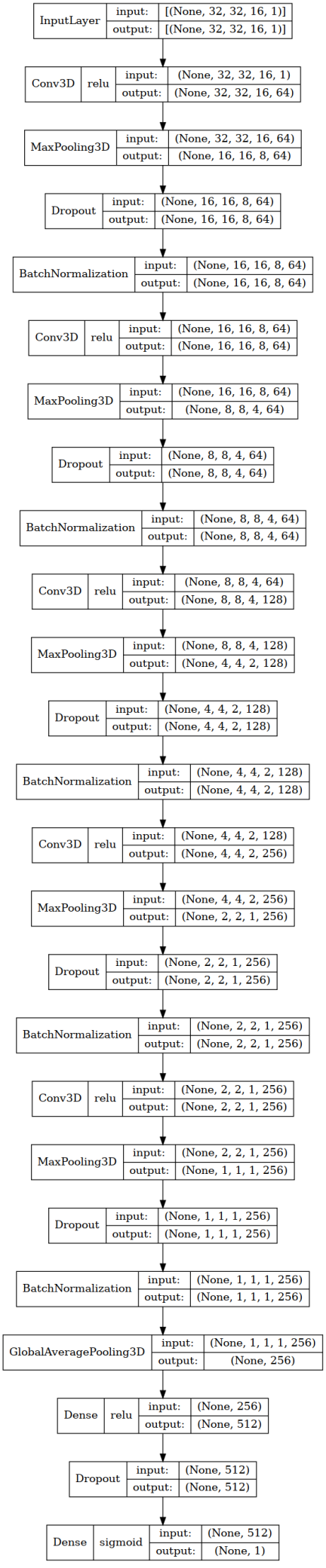


Figure 6.2: 3D convolutional neural network architecture detail

6.5 Model evaluations

6.5.1 PPMI

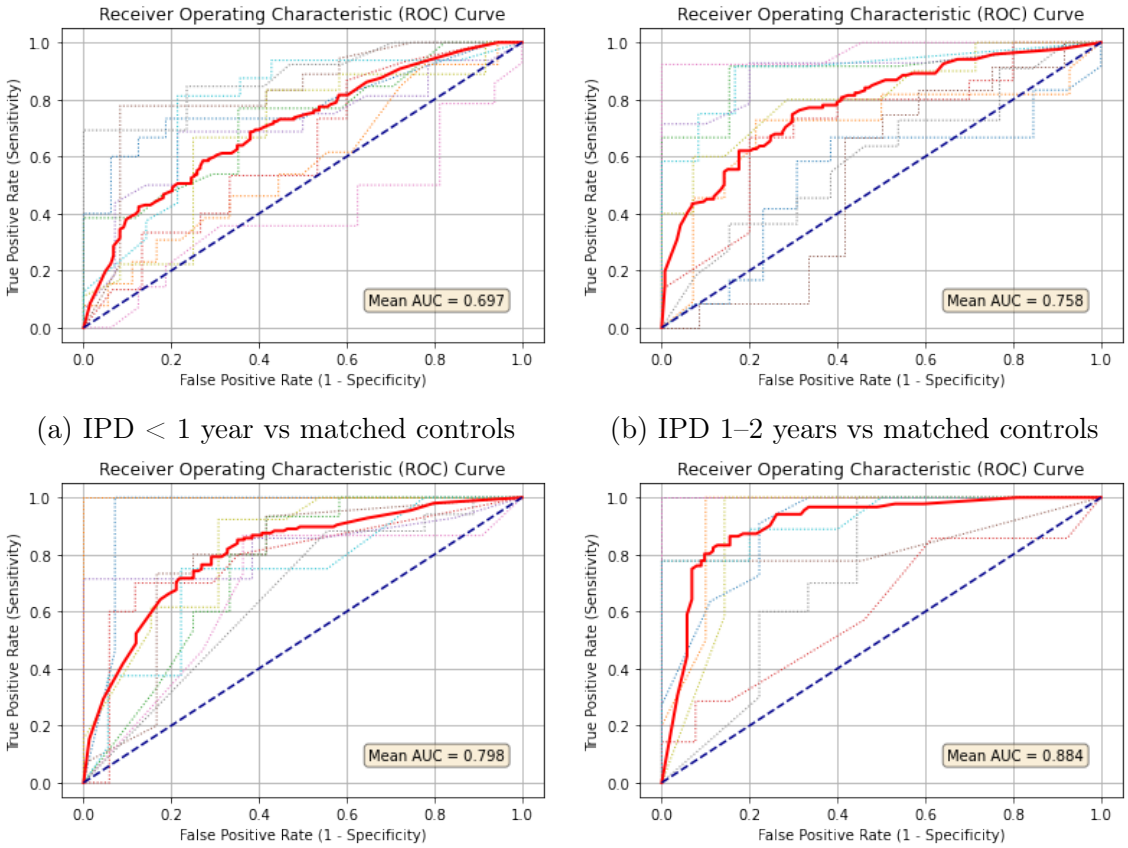
Idiopathic Parkinson's disease

In this analysis, 504 scans from 193 subjects with IPD were used. Within this cohort, 34% of subjects had one scan, 8% had two scans, 21% had three scans, 36% had four scans, and 1% had five scans. All subjects had undergone genetic testing for *LRRK2*, *GBA* or α-synuclein (*SNCA*) mutations with no pathological or PD risk factor variants found. To investigate whether model performance is affected by disease progression, these scans were stratified by time since diagnosis: those acquired more than four years after diagnosis (n=98), those acquired two to four years after diagnosis (n=133), those acquired one to two years after diagnosis (n=122), and those acquired less than a year after diagnosis (n=151). Each of these cohorts was matched on age and sex with healthy control scans in a ratio of 1:1. Demographic data for these cohorts are shown in Table 6.1. Classification thresholds were chosen to maximise accuracy and balance sensitivity and specificity (Figure A.1 in Appendix A).

In IPD subjects who had been diagnosed more than 4 years previously, relatively high accuracies (86%, 95% CI [79%, 93%]) and AUC scores (0.88, 95% CI [0.79, 0.98]) were achieved. These scores reduced successively as duration from diagnosis decreased (Table 6.2), with scans undertaken less than one year from diagnosis yielding an accuracy of 65%, 95% CI [56%, 74%], and an AUC of 0.70, 95% CI [0.60, 0.80] (Figure 6.3). All IPD models demonstrated similar regions of interest, notably in CSF voxels surrounding the brainstem (Figure 6.4). The importance of brainstem information is further explored later in this section.

Table 6.1: Demographic and scan data for all compared PPMI cohorts.
IPD = idiopathic PD
LRRK2 PD = *LRRK2* PD manifesting carriers
LRRK2 nPD = *LRRK2* non PD manifesting carriers
GBA PD = *GBA* PD manifesting carriers
GBA nPD = *GBA* non PD manifesting carriers
GC *GBA* nPD = Gaucher causing *GBA* variants non manifesting carriers

Compared cohorts	Cohort size	Median age (interquartile range)	% Male	Median symptom duration in months (interquartile range)	% 3 T scans	% Dementia	% Mild cognitive impairment
IPD < 1 year	151	63 (55-69)	69	14 (10-25)	87	0	16
Matched controls	151	62 (56-69)	69	-	97	0	1
IPD 1–2 years	122	63 (54-69)	69	28 (23-40)	99	0	19
Matched controls	122	62 (54-69)	69	-	98	0	2
IPD 2–4 years	133	65 (56-71)	67	41 (35-55)	100	1	21
Matched controls	133	65 (56-70)	67	-	96	0	2
IPD > 4 years	98	65 (56-73)	66	67 (59-83)	100	0	21
Matched controls	98	65 (56-72)	66	-	100	0	3
All IPD	513	64 (56-71)	66	37 (23-60)	98	0	18
Matched controls	513	64 (56-71)	66	-	96	0	2
LRRK2 PD	98	65 (59-70)	55	36 (20-52)	93	0	7
Matched controls	98	63 (56-69)	55	-	96	0	1
LRRK2 nPD	115	60 (56-65)	49	-	95	0	3
Matched controls	115	60 (56-65)	49	-	97	0	0
LRRK2 nPD < average age of onset	52	56 (53-57)	58	-	90	0	0
Matched controls	52	56 (53-57)	58	-	98	0	0
LRRK2 nPD > average age of onset	63	64 (62-68)	40	-	98	0	6
Matched controls	63	64 (60-69)	40	-	97	0	0
LRRK2 PD	95	65 (57-70)	56	55 (37-82)	96	0	7
Matched IPD	95	65 (57-70)	56	35 (23-59)	96	1	18
LRRK2 PD	95	65 (57-70)	56	55 (37-82)	96	0	7
Matched *LRRK2* nPD	95	65 (57-70)	56	-	93	0	7
GBA PD	128	63 (54-73)	53	33 (14-56)	92	0	14
Matched controls	128	60 (54-69)	53	-	97	0	1
GBA nPD	109	63 (57-67)	46	-	95	0	2
Matched controls	109	63 (57-67)	46	-	95	0	1
GC *GBA* nPD	101	63 (57-67)	47	-	95	0	2
Matched controls	101	60 (55-65)	47	-	98	0	0
GBA nPD > average age of onset	91	64 (61-69)	45	-	96	0	2
Matched controls	91	62 (57-66)	45	-	97	0	1
GBA PD	127	62 (55-71)	58	41 (23-65)	96	1	18
Matched IPD	127	62 (55-71)	58	35 (24-62)	97	1	15
GBA PD	109	63 (57-67)	54	58 (26-77)	96	0	12
Matched *GBA* nPD	109	63 (57-67)	54	-	95	0	2



(a) IPD $<$ 1 year vs matched controls

(b) IPD 1–2 years vs matched controls

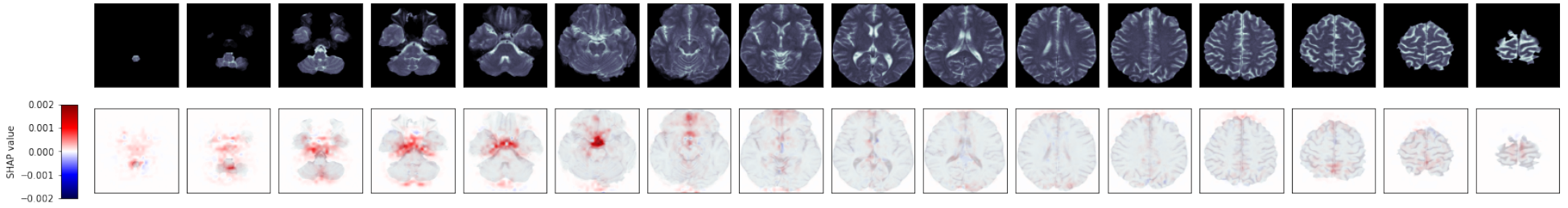
(c) IPD 2–4 years vs matched controls

(d) IPD $>$ 4 years vs matched controls

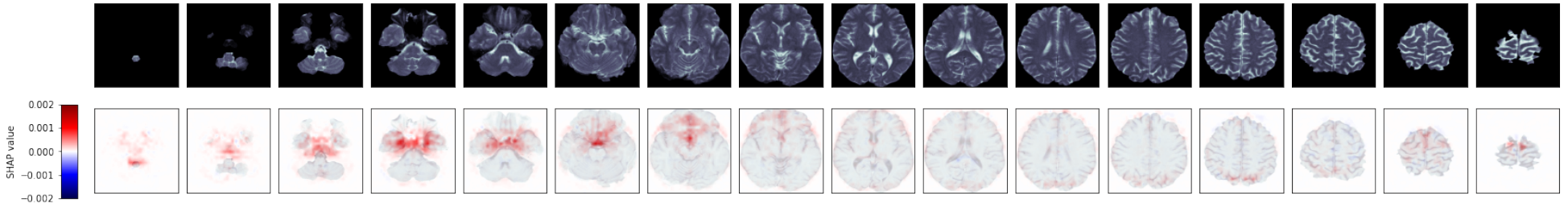
Figure 6.3: Receiver Operating Characteristic (ROC) curves for idiopathic PD (IPD) models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.

Table 6.2: Model performance results for all compared PPMI cohorts.
IPD = idiopathic PD
LRRK2 PD = *LRRK2* PD manifesting carriers
LRRK2 nPD = *LRRK2* non PD manifesting carriers
GBA PD = *GBA* PD manifesting carriers
GBA nPD = *GBA* non PD manifesting carriers
GC *GBA* nPD = Gaucher causing *GBA* variants non manifesting carriers

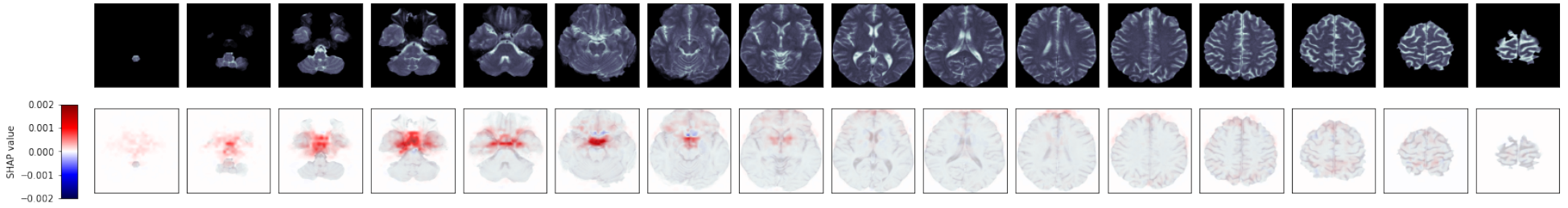
Compared cohorts	Mean test ROC AUC [95% CI]	Mean test accuracy [95% CI]	Mean test sensitivity [95% CI]	Mean test specificity [95% CI]
IPD < 1 year vs controls	0.70 [0.60, 0.80]	65% [56%, 74%]	64% [53%, 75%]	65% [48%, 82%]
IPD 1–2 years vs controls	0.76 [0.64, 0.87]	72% [63%, 81%]	70% [60%, 81%]	72% [57%, 86%]
IPD 2–4 years vs controls	0.80 [0.72, 0.88]	77% [70%, 84%]	74% [59%, 90%]	74% [55%, 92%]
IPD > 4 years vs controls	0.88 [0.79, 0.98]	86% [79%, 93%]	85% [70%, 100%]	84% [75%, 94%]
LRRK2 PD vs controls	0.94 [0.89, 0.99]	92% [88%, 97%]	92% [82%, 100%]	92% [84%, 100%]
LRRK2 nPD vs controls	0.95 [0.91, 0.99]	94% [89%, 98%]	93% [88%, 98%]	93% [85%, 100%]
LRRK2 nPD < average age of onset vs controls	0.95 [0.86, 0.99]	90% [85%, 92%]	89% [81%, 100%]	90% [84%, 100%]
LRRK2 nPD > average age of onset vs controls	0.99 [0.98, 1.00]	98% [93%, 100%]	97% [93%, 100%]	98% [92%, 100%]
LRRK2 PD vs IPD	0.88 [0.82, 0.93]	83% [74%, 91%]	82% [69%, 95%]	81% [72%, 90%]
LRRK2 PD vs *LRRK2* nPD	0.73 [0.61, 0.86]	79% [70%, 88%]	78% [62%, 94%]	78% [66%, 90%]
GBA PD vs controls	0.84 [0.69, 0.98]	81% [70%, 92%]	79% [62%, 96%]	79% [65%, 93%]
GBA nPD vs controls	0.92 [0.83, 1.00]	89% [79%, 98%]	88% [82%, 93%]	89% [73%, 100%]
GC *GBA* nPD vs controls	0.96 [0.91, 1.00]	93% [88%, 98%]	92% [86%, 97%]	92% [81%, 100%]
GBA nPD > average age of onset vs controls	0.93 [0.90, 0.97]	89% [84%, 94%]	88% [81%, 94%]	88% [79%, 97%]
GBA PD vs IPD	0.69 [0.61, 0.78]	69% [62%, 77%]	69% [59%, 78%]	68% [55%, 81%]
GBA PD vs *GBA* nPD	0.82 [0.74, 0.90]	81% [76%, 87%]	78% [61%, 96%]	78% [65%, 91%]



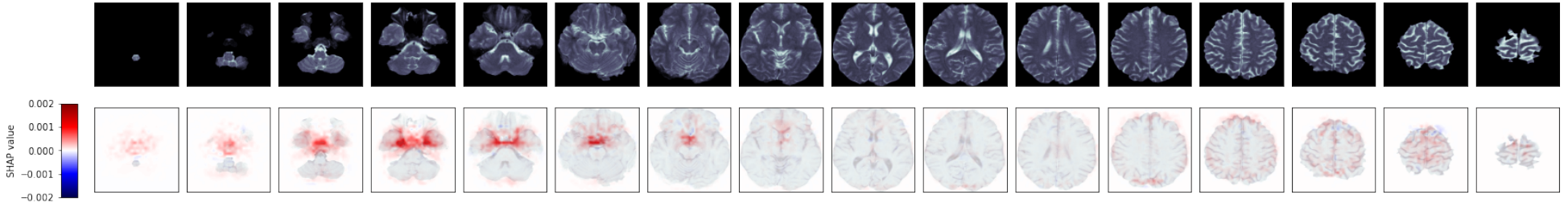
(a) IPD $<$ 1 year from diagnosis vs matched controls



(b) IPD 1–2 years from diagnosis vs matched controls



(c) IPD 2–4 years from diagnosis vs matched controls



(d) IPD $>$ 4 years from diagnosis vs matched controls

Figure 6.4: Mean SHapley Additive exPlanation (SHAP) maps for correct predictions of idiopathic PD (IPD). Pixels highlighted in red have contributed to the prediction.

Holdout testing

In this analysis, all IPD scans were matched on age and sex with healthy control scans in a ratio of 1:1. Demographic data are shown in Table 6.1. Of the total 1026 scans, 80% (n=818) were used for model development, and 20% (n=208) were reserved as a holdout test set for final evaluation of the model. The classification threshold was chosen in the development stage to maximise accuracy and balance sensitivity and specificity (Figure A.1 in Appendix A).

In development, the models yielded medium performance results (Table 6.3), with an AUC of 0.71 (95% CI [0.57, 0.85]) and an accuracy of 71% (95% CI [60%, 81%]). This performance is not surprising given the previously demonstrated decreasing performance for earlier disease stages (Table 6.2) and the skew in this cohort's population towards earlier disease stages (Table 6.1).

The holdout results were similar, albeit slightly lower (Table 6.3, with an AUC of 0.69 (95% CI [0.65, 0.73]) and an accuracy of 63% (95% CI [58%, 69%]). This demonstrates that there has not been major overfitting in the k-fold model development strategy.

Table 6.3: Holdout performance results for models trained on all idiopathic Parkinson's disease and matched controls

Validation strategy	Mean ROC AUC [95% CI]	Mean accuracy [95% CI]	Mean sensitivity [95% CI]	Mean specificity [95% CI]
K-fold test score	0.71 [0.57, 0.85]	71% [60%, 81%]	71% [54%, 88%]	70% [53%, 87%]
Holdout score	0.69 [0.65, 0.73]	63% [58%, 69%]	68% [56%, 80%]	56% [39%, 73%]

Masking experiments

This same dataset of all IPD scans and matched controls was used to conduct some experiments to further investigate the regions of interest identified in the SHAP maps (Figure 6.4). In one experiment, the whole brain was masked out of the image

to assess the impact of non-brain information. In the other, the brainstem (identified using the FMRIB Software Library tool FIRST [180]) was masked out of the image, to see whether brainstem atrophy was informing the models.

Table 6.4: Performance results for models trained on masked images

Masked out	Mean test ROC AUC [95% CI]	Mean test accuracy [95% CI]
Whole brain	0.50	53% [47%, 59%]
Brainstem	0.79 [0.71, 0.86]	74% [64%, 84%]

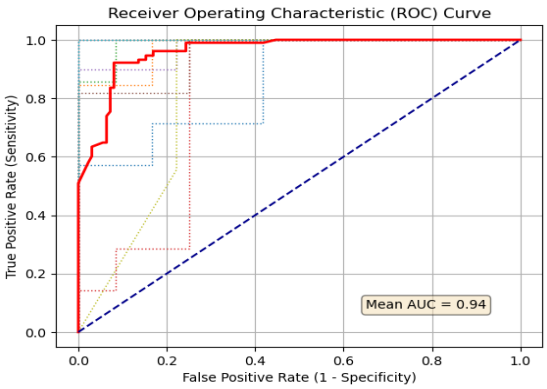
When the whole brain was masked out, the model performance was no better than random guessing (Table 6.4), indicating that no non-brain information has confounded the previously reported models. When just the brainstem was masked out, the performance of the models was somewhat higher than the previously reported models (Tables 6.3 and 6.4), indicating that the brainstem, and brainstem atrophy in particular, has not been a source of discriminative information. The higher performance results might be attributed to the removal of non-informative data, and the focusing of the models on more informative data, likely cortical changes.

LRRK2

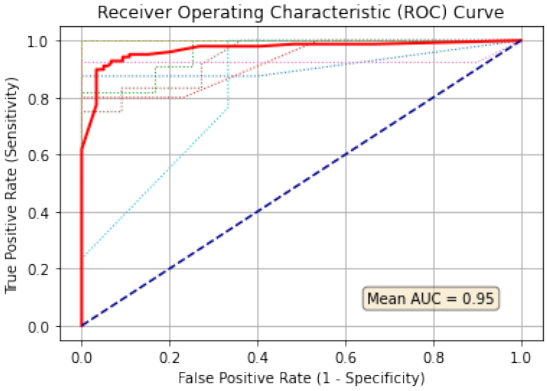
In this analysis, 210 scans from 159 carriers of *LRRK2* risk variants were used. These were stratified into PD manifesting carriers (*LRRK2* PD, n=95) and non PD manifesting carriers (*LRRK2* nPD, n=115). In the case of the non-manifesting carriers, scans were further stratified by time of scan and divided into those acquired after the age of 59.4 (the average age of onset of *LRRK2* PD) [176] (n=63), and those taken before (n=52). Each pair of compared cohorts was matched on age and sex in a ratio of 1:1. Demographic data for these cohorts are shown in Table 6.1. Classification thresholds were chosen to maximise accuracy and balance sensitivity and specificity (Figure A.2 in Appendix A).

Models performed well in all cases (Table 6.2 and Figure 6.5). Ninety two percent of *LRRK2* PD/control scans, 95% CI [88%, 97%], (AUC 0.94, 95% CI [0.89, 0.99])

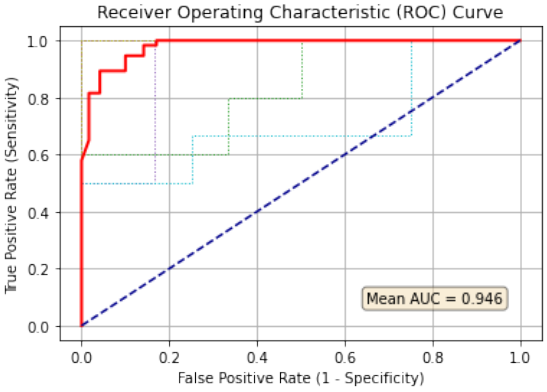
were predicted correctly. Ninety four percent of *LRRK2* nPD/control scans, 95% CI [89%, 98%], (AUC 0.95, 95% CI [0.91, 0.99]) were predicted correctly. This rose to 98%, 95% CI [93%, 100%] (AUC 0.99, 95% CI [0.98, 1.00]) in the scans from *LRRK2* nPD subjects over the age of onset. Three of the 115 *LRRK2* nPD scans came from subjects who converted to motor Parkinson's during the study period. The model correctly predicted these scans in all cases, with probability estimates of 93%, 99% and 97% respectively. Notably the model comparing *LRRK2* PD and IPD performed better than equivalent *GBA* PD models, predicting *LRRK2* scans with 83% accuracy, 95% CI [74%, 91%] (AUC 0.88, 95% CI [0.82, 0.93]). Average SHAP heatmaps demonstrated predominant interest in pixels immediately adjacent to the brainstem parenchyma (Figure 6.6). In the *LRRK2* nPD/control model, there appeared also to be additional interest in pixels adjacent to the cerebellum, particularly in scans from subjects over the average age of onset.



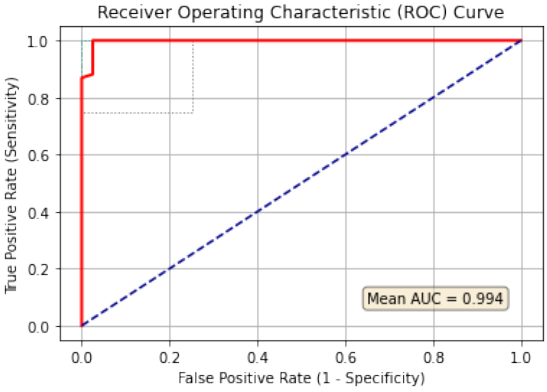
(a) *LRRK2* PD vs matched controls



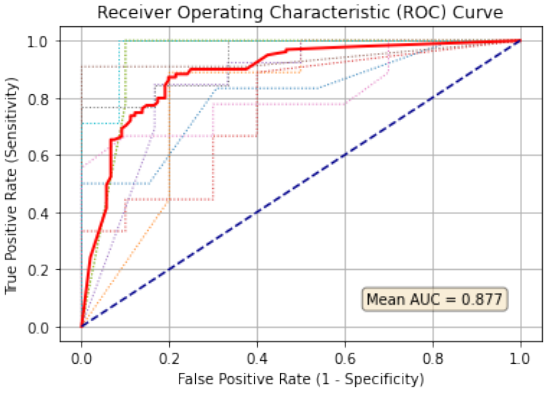
(b) *LRRK2* nPD vs matched controls



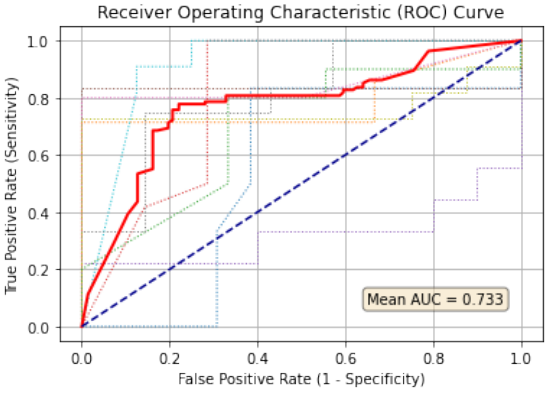
(c) *LRRK2* nPD $<$ average age of onset vs matched controls



(d) *LRRK2* nPD $>$ average age of onset vs matched controls

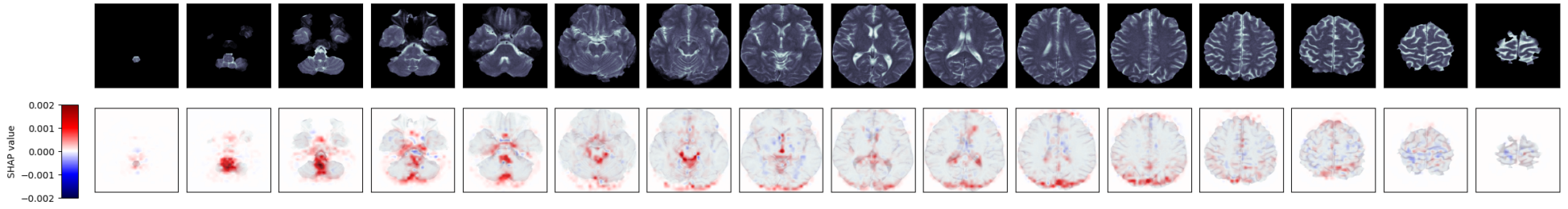


(e) *LRRK2* PD vs matched IPD

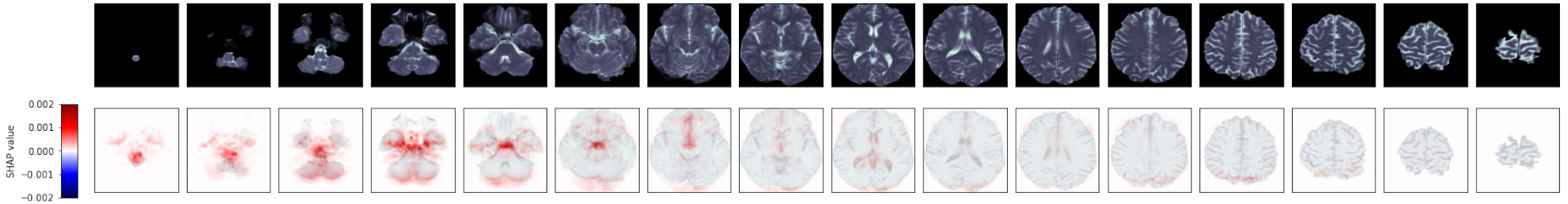


(f) *LRRK2* PD vs matched *LRRK2* nPD

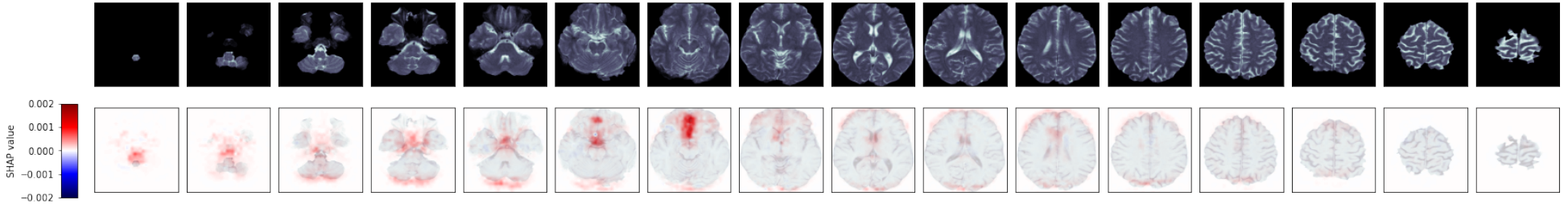
Figure 6.5: Receiver Operating Characteristic (ROC) curves for *LRRK2* models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.



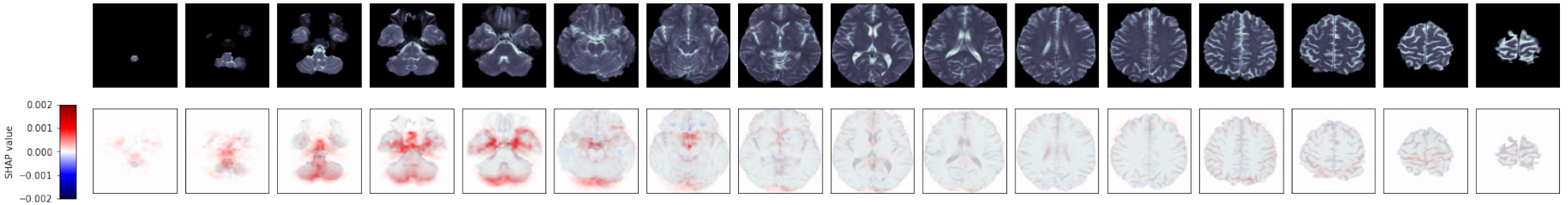
(a) *LRRK2* PD vs matched controls



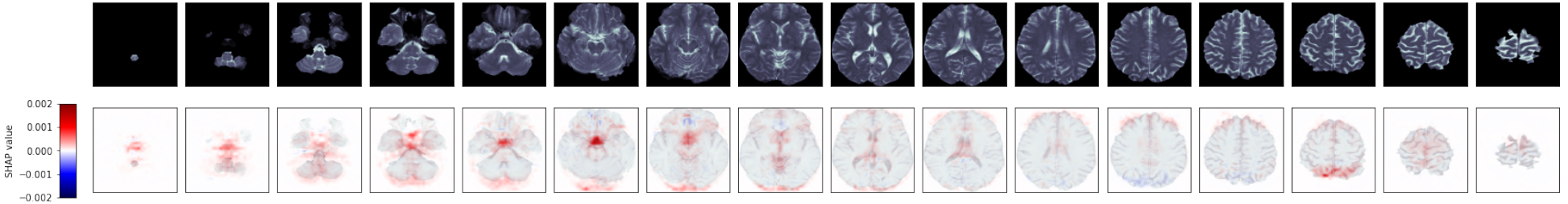
(b) *LRRK2* nPD vs matched controls



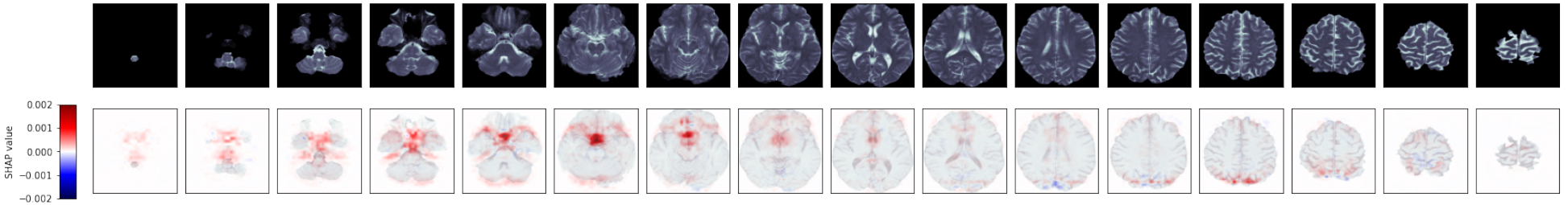
(c) *LRRK2* nPD $<$ average age of onset vs matched controls



(d) *LRRK2* nPD > average age of onset vs matched controls



(e) *LRRK2* PD vs matched IPD



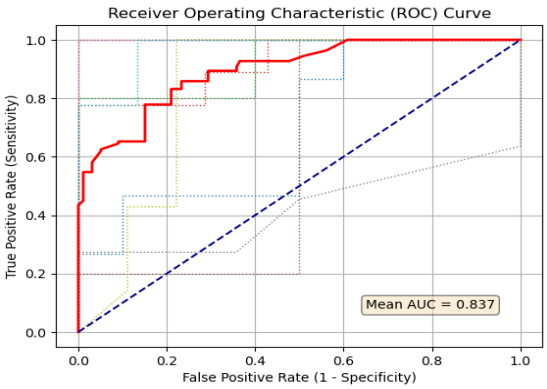
(f) *LRRK2* PD vs matched *LRRK2* nPD

Figure 6.6: Mean SHapley Additive exPlanation (SHAP) maps for correct predictions of *LRRK2* carriers. Pixels highlighted in red have contributed positively to the prediction; pixels highlighted in blue have contributed negatively to the prediction.

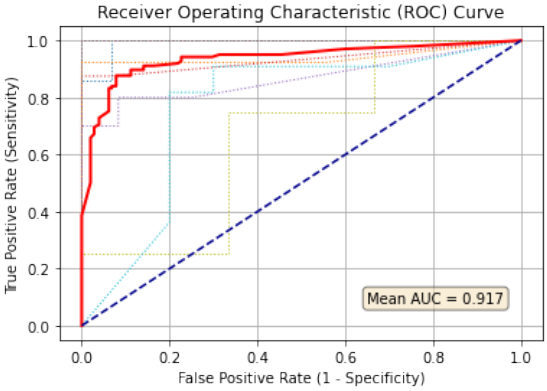
GBA

In this analysis, 236 scans from 159 carriers of a *GBA* variant were used. These were stratified into PD manifesting carriers (*GBA* PD, n=127) and non PD manifesting carriers (*GBA* nPD, n=109). The scans were also stratified into 184 scans from carriers of Gaucher causing *GBA* variants (GC *GBA*) and 51 scans from carriers of the non Gaucher causing PD variants p.E326K (alternative nomenclature p.E365K) and p.T369M (p.T409M). There were insufficient numbers to build models using cases stratified into "mild" and "severe" variants. The *GBA* nPD group was also further stratified by time of scan: divided into those acquired after the age of 55.8 (the average age of onset of *GBA* related PD) [176] (n=91), and those taken before (n=18). The latter cohort was not large enough to build a model. Each pair of compared cohorts was matched on age and sex in a ratio of 1:1. Demographic data for these cohorts are shown in Table 6.1. Classification thresholds were chosen to maximise accuracy and balance sensitivity and specificity (Figure A.3 in Appendix A).

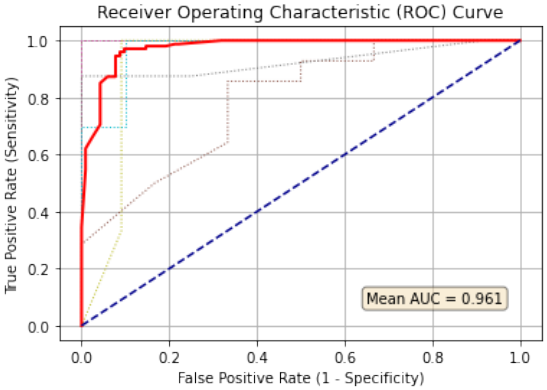
GBA models performed well (Table 6.2 and Figure 6.7). The model built for the combined *GBA* PD cohort was able to successfully predict *GBA* PD scans with an accuracy of 81%, 95% CI [70%, 92%] (AUC 0.84, 95% CI [0.69, 0.98]). The model built for the combined *GBA* nPD cohort was able to successfully predict *GBA* nPD scans with an accuracy of 89%, 95% CI [79%, 98%] (AUC 0.92, 95% CI [0.83, 1.00]). This increased marginally to 93%, 95% CI [84%, 94%] (AUC 0.96, 95% CI [0.90, 0.97]) in the combined 'mild' and 'severe' *GBA* nPD cohort. In the cohort of *GBA* nPD participants over the average age of onset, performance was almost identical to the main model (89% accuracy, 95% CI [88%, 98%], AUC 0.93, 95% CI [0.91, 1.00]). In common with idiopathic PD, SHAP heatmaps demonstrated interest in non-parenchymal pixels surrounding the brainstem (Figure 6.8). Additionally there was a focus on pixels adjacent to the posterior occipital lobe.



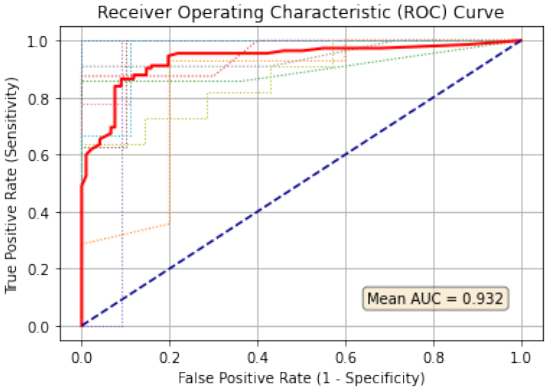
(a) *GBA* PD vs matched controls



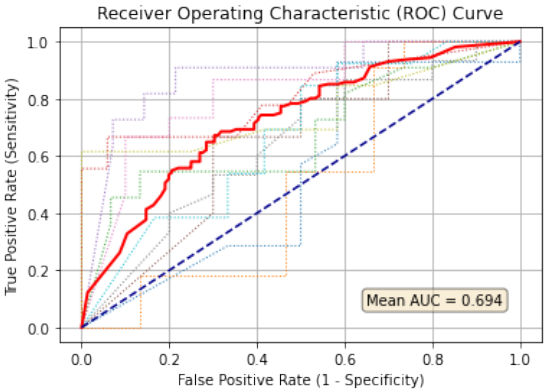
(b) *GBA* nPD vs matched controls



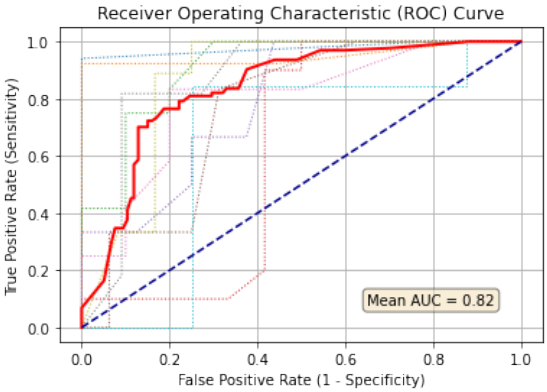
(c) GC *GBA* nPD vs matched controls



(d) *GBA* nPD > average age of onset vs matched controls

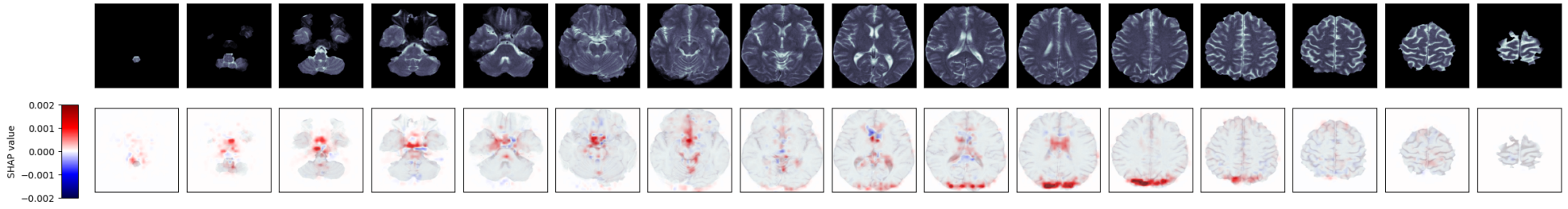


(e) *GBA* PD vs matched IPD

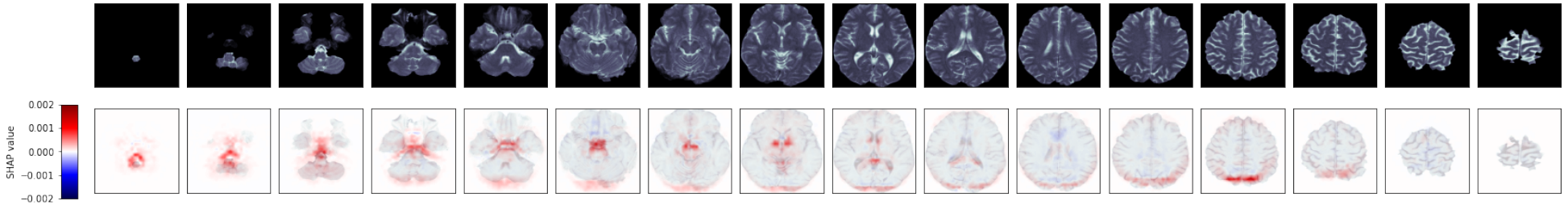


(f) *GBA* PD vs matched *GBA* nPD

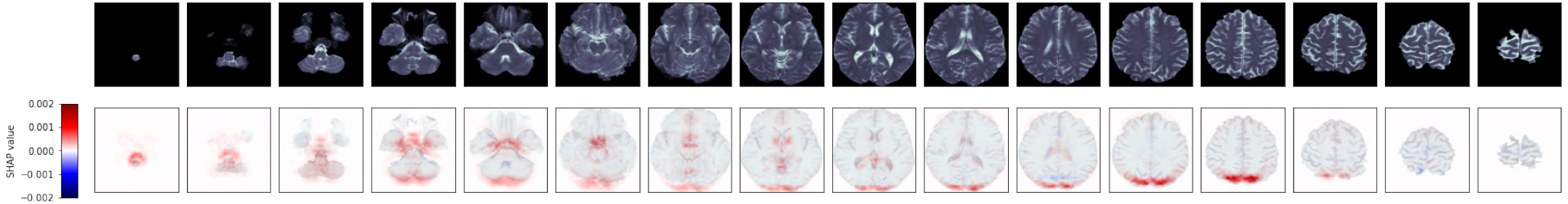
Figure 6.7: Receiver Operating Characteristic (ROC) curves for *GBA* models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.



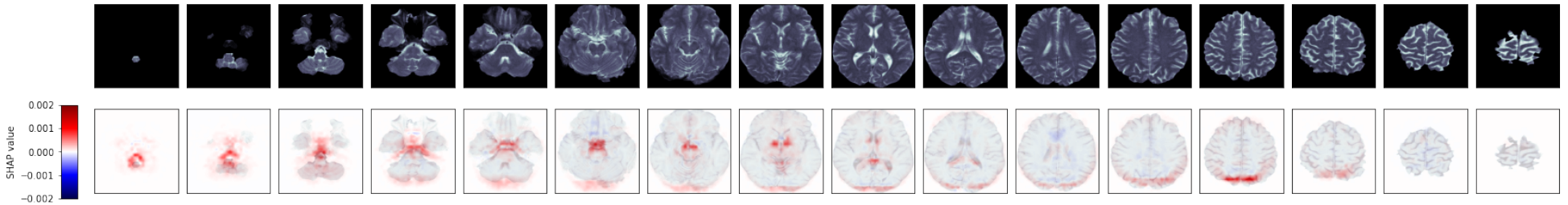
(a) *GBA* PD vs matched controls



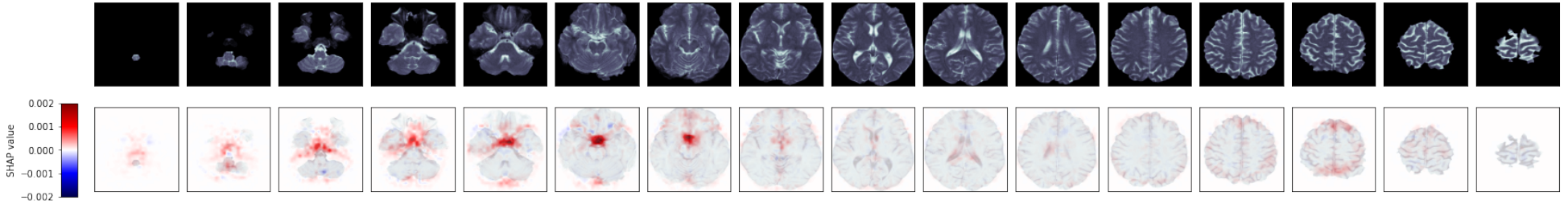
(b) *GBA* nPD vs matched controls



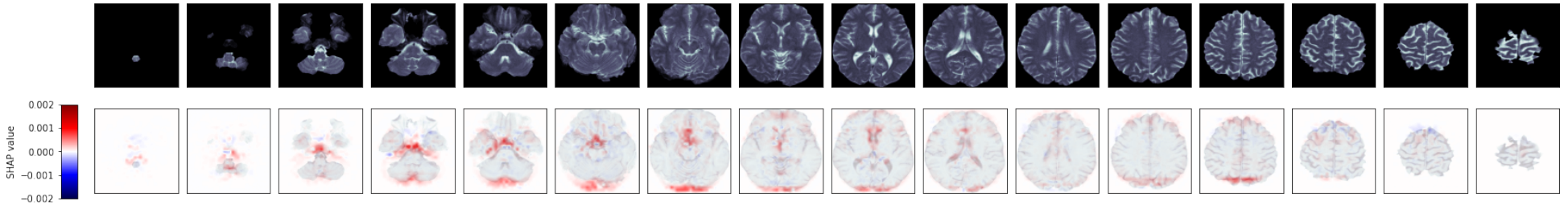
(c) GC *GBA* nPD vs matched controls



(d) *GBA* nPD > average age of onset vs matched controls



(e) *GBA* PD vs matched IPD



(f) *GBA* PD vs matched *GBA* nPD

Figure 6.8: Mean SHapley Additive exPlanation (SHAP) maps for correct predictions of *GBA* carriers. Pixels highlighted in red have contributed positively to the prediction; pixels highlighted in blue have contributed negatively to the prediction.

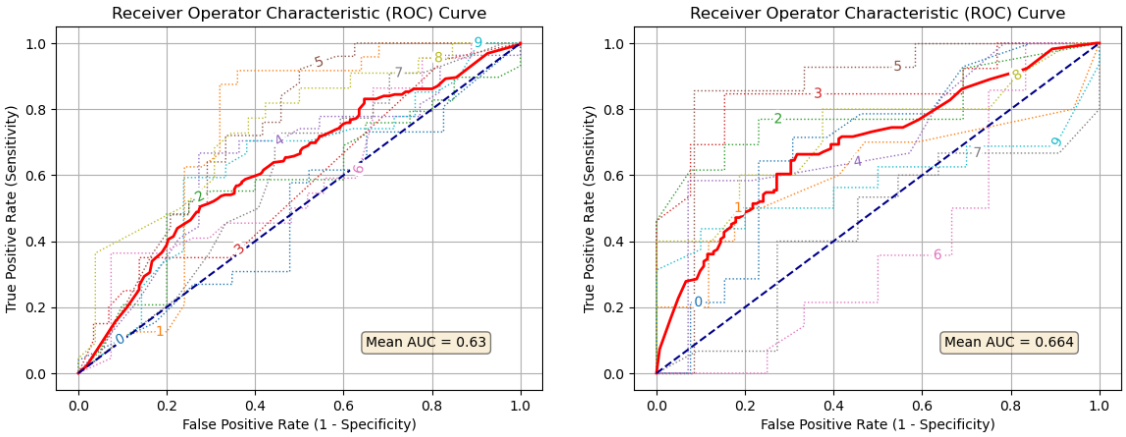
6.5.2 UHPNT

In this analysis, 244 scans from 203 patients with PD were used. The 236 scans for which a date of diagnosis was present were then stratified by time of scan: divided into those acquired after diagnosis (n=131) and those acquired before diagnosis (n=105). Each pair of compared cohorts was matched on age in a ratio of 1:1. Demographic data for these cohorts are shown in Table 6.5. Classification thresholds were chosen to maximise accuracy and balance sensitivity and specificity (Figure A.4 in Appendix A).

Table 6.5: Demographic and scan data for all compared UHPNT cohorts

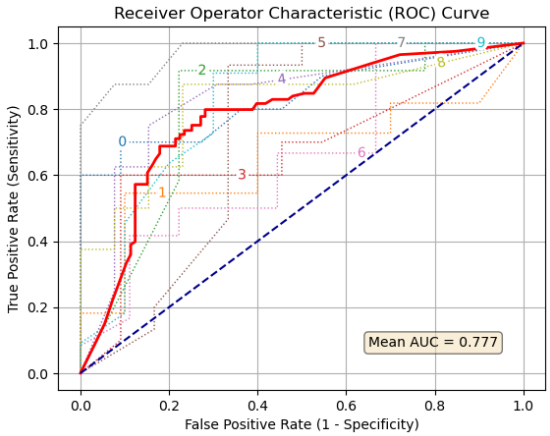
Compared cohorts	Cohort size	Median age (interquartile range)	Median disease duration in months (interquartile range)	% 1.5 T scans
All PD	244	70 (62-76)	1 (0-41)	96
Matched controls	244	70 (62-76)	-	98
PD post-diagnosis	131	73 (65-78)	36 (3-71)	95
Matched controls	131	73 (65-78)	-	97
PD pre-diagnosis	105	68 (58-71)	-	97
Matched controls	105	68 (58-71)	-	98

These models had some predictive power (Table 6.6 and Figure 6.9), but exhibited notably inferior performance when compared to the models developed for the PPMI dataset. The models built for all PD scans predicted PD with an accuracy of 61%, 95% CI [56%, 66%] (AUC 0.63, 95% CI [0.57, 0.69]). The models built for the stratified scans performed better: those built for scans taken after diagnosis were able to predict PD scans with an accuracy of 68%, 95% CI [62%, 73%] (AUC 0.66, 95% CI [0.55, 0.78]). Unexpectedly, the performance was higher for the models built for scans taken before diagnosis, with an accuracy of 77%, 95% CI [68%, 85%] (AUC 0.78, 95% CI [0.70, 0.85]). In common with the PPMI models, the SHAP heatmaps demonstrated interest in the pixels surrounding the brainstem (Figure 6.10), but the signal is weaker and less focused.



(a) PD vs matched controls

(b) PD post-diagnosis vs matched controls

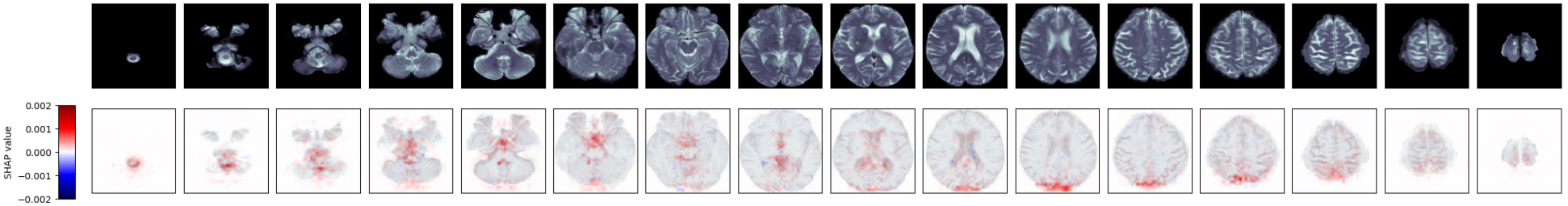


(c) PD pre-diagnosis vs matched controls

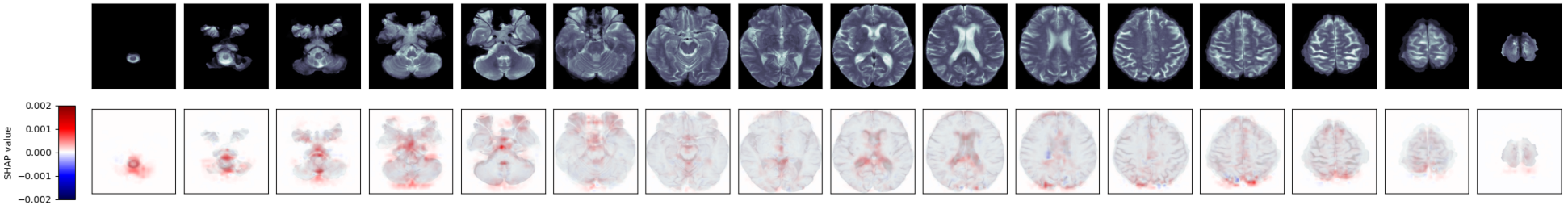
Figure 6.9: Receiver Operating Characteristic (ROC) curves for UHPNT models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.

Table 6.6: Model performance results for all compared UHPNT cohorts.

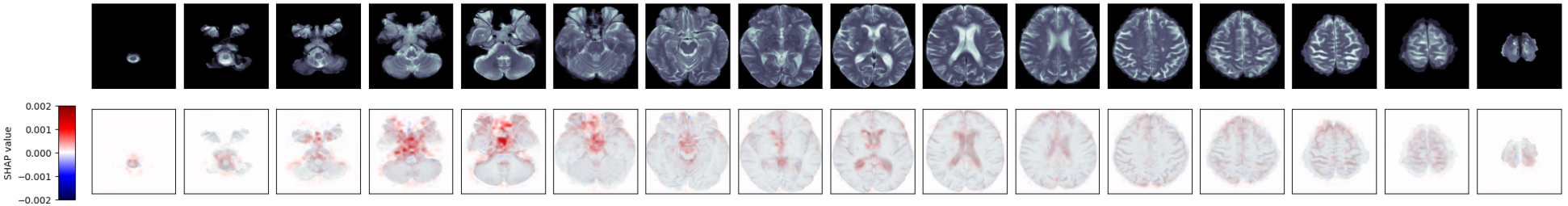
Compared cohorts	Mean test ROC AUC [95% CI]	Mean test accuracy [95% CI]	Mean test sensitivity [95% CI]	Mean test specificity [95% CI]
All PD vs controls	0.63 [0.57, 0.69]	0.61 [0.56, 0.66]	0.59 [0.44, 0.74]	0.61 [0.48, 0.75]
PD post-diagnosis vs controls	0.66 [0.55, 0.78]	0.68 [0.62, 0.73]	0.66 [0.53, 0.79]	0.67 [0.52, 0.81]
PD pre-diagnosis vs controls	0.78 [0.70, 0.85]	0.77 [0.68, 0.85]	0.75 [0.62, 0.88]	0.75 [0.64, 0.86]



(a) PD vs matched controls



(b) PD post-diagnosis vs matched controls



(c) PD pre-diagnosis vs matched controls

Figure 6.10: Mean SHapley Additive exPlanation (SHAP) maps for correct predictions of PD in UHPNT cohorts. Pixels highlighted in red have contributed positively to the prediction; pixels highlighted in blue have contributed negatively to the prediction.

Table 6.7: Demographic and scan data for external validation cohorts

Compared cohorts	Cohort size	Median age (interquartile range)	Median disease duration in months (interquartile range)	% 3 T scans
PPMI post-diagnosis	511	64 (56-71)	20 (8-34)	96
Matched controls	511	64 (56-71)	-	92
UHPNT > 4 years post-diagnosis	51	74 (67-79)	91 (62-125)	8
Matched controls	51	74 (67-79)	-	6
UHPNT 2-4 years post-diagnosis	24	74 (62-78)	38 (34-42)	4
Matched controls	24	74 (62-78)	-	0
UHPNT 1-2 years post-diagnosis	9	76 (68-78)	18 (14-20)	1
Matched controls	9	76 (68-78)	-	0
UHPNT < 1 year post-diagnosis	71	69 (61-73)	1 (0-2)	3
Matched controls	71	69 (61-73)	-	1

6.5.3 External validation

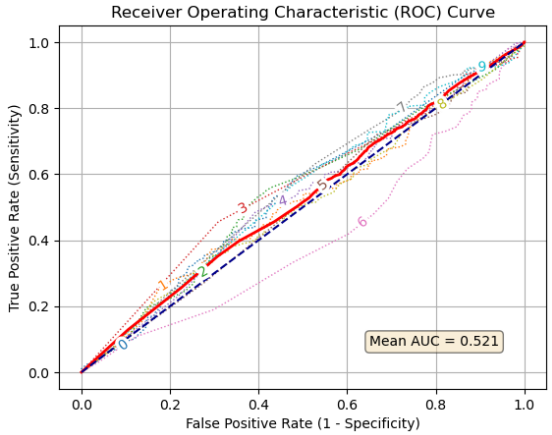
In this analysis, the models that had been trained on the PPMI data were tested on equivalent UHPNT data, and the models that had been trained on the UHPNT data were tested on equivalent PPMI data. In the case of the UHPNT models, only the models developed for post-diagnosis scans could be tested on the PPMI data, as the PPMI data does not contain scans acquired prior to diagnosis. These models were tested on 511 scans from PPMI subjects with IPD, with PPMI controls matched on age and sex in a ratio of 1:1. In the case of the PPMI models, only the IPD models could be tested on the UHPNT dataset. The appropriate models were tested on 51 UHPNT scans acquired more than four years after diagnosis, 24 UHPNT scans acquired two to four years after diagnosis, nine UHPNT scans acquired one to two years after diagnosis, 71 UNPHT scans acquired less than a year after diagnosis, and in each case UHPNT controls matched on age in a ratio of 1:1. Demographic data for these cohorts are shown in Table 6.7.

These models all performed notably poorly (Table 6.8 and Figure 6.11). The models trained on UHPNT data predicted PD with an accuracy of 51%, 95% CI [50%, 52%] (AUC 0.52, 95% CI [0.49, 0.55]). This is no better than random guessing. Likewise the models trained on PPMI data predicted PD with an accuracy of AUC

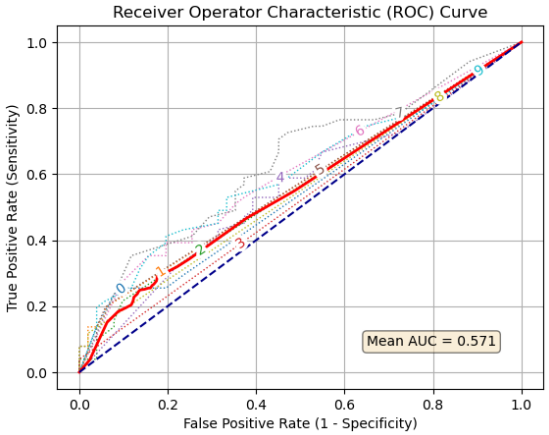
Table 6.8: Model performance results for external validation of PPMI and UHPNT models

Model	Test cohort	Mean ROC AUC [95% CI]	Mean accuracy [95% CI]
UHPNT post-diagnosis vs controls	PPMI post-diagnosis vs controls	0.52 [0.49, 0.55]	0.51 [0.50, 0.52]
PPMI $>$ 4 years post-diagnosis vs controls	UHPNT $>$ 4 years post-diagnosis vs controls	0.57 [0.55, 0.60]	0.50
PPMI 2-4 years post-diagnosis vs controls	UHPNT 2-4 years post-diagnosis vs controls	0.55 [0.51, 0.59]	0.50
PPMI 1-2 years post-diagnosis vs controls	UHPNT 1-2 years post-diagnosis vs controls	0.53 [0.46, 0.60]	0.50
PPMI $<$ 1 year post-diagnosis vs controls	UHPNT $<$ 1 year post-diagnosis vs controls	0.51 [0.50, 0.53]	0.50

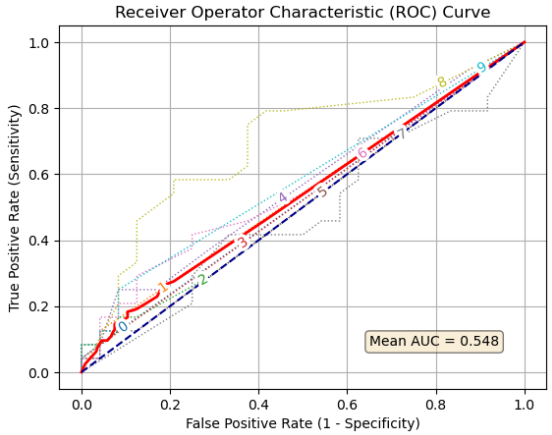
of 0.57, 95% CI [0.55, 0.60] for scans acquired more than four years after diagnosis, 0.55, 95% CI [0.51, 0.59] for scans acquired two to four years after diagnosis, 0.53, 95% CI [0.46, 0.60] for scans acquired one to two years after diagnosis, and 0.51, 95% CI [0.50, 0.53] for scans acquired less than a year after diagnosis. It is notable that the AUC was lower for earlier cases of PD, just as it was for the PPMI data (see Table 6.2). The accuracy was always 50%. This was because the output probabilities that the scans were from a PD patient were always extremely tightly grouped; this meant that all the scans would fall on one side of the classification threshold chosen and exactly half were correctly predicted. In contrast, the UHPNT model that was tested on the PPMI data output a much greater range of probabilities, but still performed as poorly.



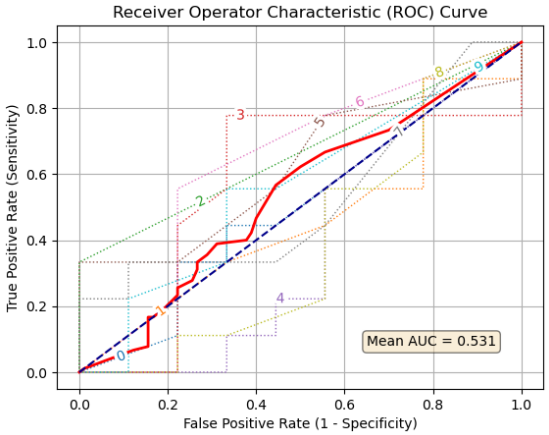
(a) UHPNT post-diagnosis vs controls tested on PPMI post-diagnosis vs controls



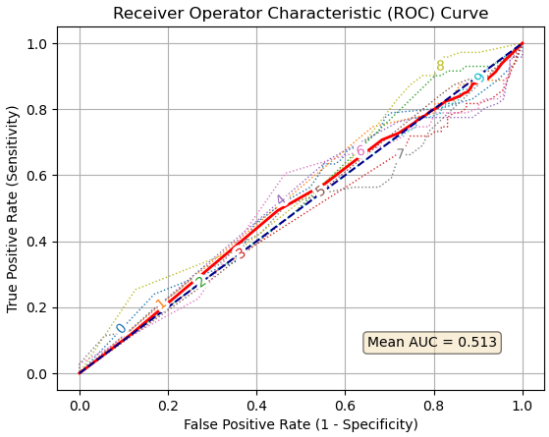
(b) PPMI $>$ 4 years vs controls tested on UHPNT $>$ 4 years vs controls



(c) PPMI 2-4 years vs controls tested on UHPNT 2-4 years vs controls



(d) PPMI 1-2 years vs controls tested on UHPNT 1-2 years vs controls



(e) PPMI $<$ 1 year vs controls tested on UHPNT $<$ 1 year vs controls

Figure 6.11: Receiver Operating Characteristic (ROC) curves for external validation of PPMI and UHPNT models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.

6.5.4 Combined PPMI and UHPNT data

In this analysis, scans from PPMI subjects with IPD were combined with scans from UHPNT patients with PD. These scans were stratified by time since diagnosis so that results could be compared with previous models as closely as possible. UHPNT scans acquired more than four years after PD diagnosis (n=51) were age-matched with UHPNT controls (n=51), with PPMI scans acquired more than four years after diagnosis (n=51), and with PPMI controls (n=51). UHPNT scans acquired one to four years after PD diagnosis (n=35) were age-matched with UHPNT controls (n=35), with PPMI scans acquired one to four years after diagnosis (n=35), and with PPMI controls (n=35). There were not enough scans to divide this cohort into scans acquired two to four years after diagnosis and one to two years after diagnosis, as had been done with the PPMI dataset. UHPNT scans acquired less than a year after PD diagnosis (n=71) were age-matched with UHPNT controls (n=71), with PPMI scans acquired less than a year after diagnosis (n=71), and with PPMI controls (n=71). Demographic data for these cohorts are shown in Table 6.9. Models were built to differentiate between PD scans and control scans. Classification thresholds were chosen to maximise accuracy and balance sensitivity and specificity (Figure A.5 in Appendix A).

Table 6.9: Demographic and scan data for combined PPMI and UHPNT cohorts

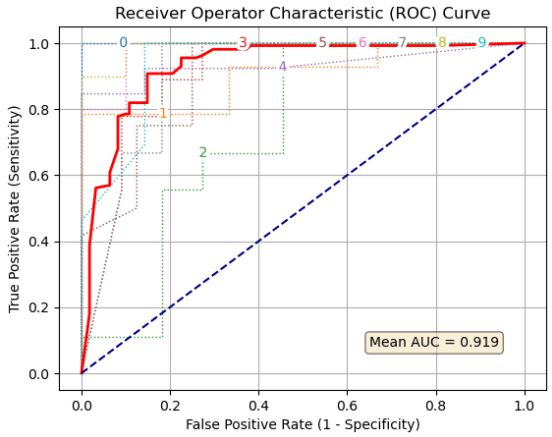
Compared cohorts	Cohort size	Median age (interquartile range)	Median disease duration in months (interquartile range)	% 3 T scans
PD > 4 years	102	74 (67-79)	61 (52-90)	54
Matched controls	102	74 (67-79)	-	52
PD 1-4 years	70	75 (64-78)	26 (19-35)	50
Matched controls	70	75 (64-78)	-	50
PD < 1 year	142	69 (61-74)	2 (1-5)	46
Matched controls	142	69 (61-74)	-	49

In common with the PPMI models, these combined models were better at detecting more advanced PD (Table 6.10 and Figure 6.12). The models built for scans acquired more than four years after diagnosis predicted PD with an accuracy of 86%, 95% CI [78%, 94%] (AUC 0.92, 95% CI [0.87, 0.97]). The models built for

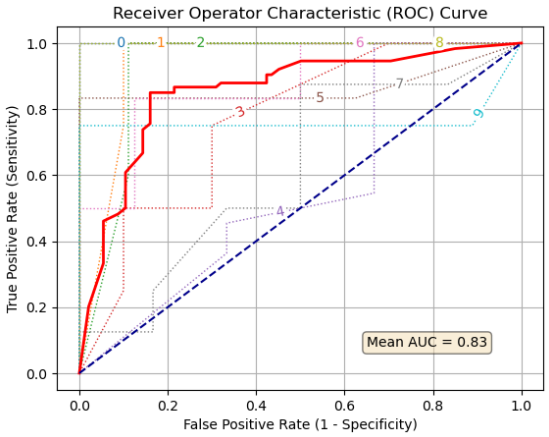
scans acquired one to four years after diagnosis predicted PD with an accuracy of 86%, 95% CI [79%, 94%] (AUC 0.83, 95% CI [0.72, 0.94]). The models built for scans acquired less than a year after diagnosis exhibited poorer performance, predicting PD with an accuracy of 72%, 95% CI [63%, 80%] (AUC 0.68, 95% CI [0.58, 0.79]). The performance of these models was usually slightly higher than the performance of the equivalent models trained on PPMI data alone. When the metrics were calculated for the PPMI and UHPNT scans separately, the performance of the models was usually notably higher for the PPMI scans. In common with other models, the SHAP heatmaps demonstrated interest in the pixels surrounding the brainstem (Figure 6.13).

Table 6.10: Model performance results for compared combined PPMI and UHPNT cohorts.

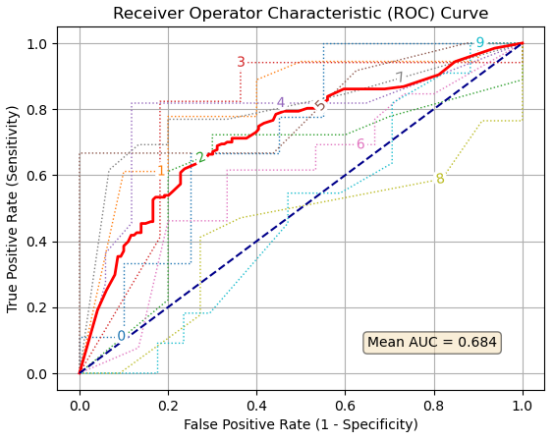
Compared cohorts	Overall mean test ROC AUC [95% CI]	Mean test ROC AUC by data source (PPMI/ UNPHT)	Overall mean test accuracy [95% CI]	Mean test accuracy by data source (PPMI/ UNPHT)	Overall mean test sensitivity [95% CI]	Mean test sensitivity by data source (PPMI/ UNPHT)	Overall mean test specificity [95% CI]	Mean test specificity by data source (PPMI/ UNPHT)
PD $>$ 4 years vs controls	0.92 [0.87, 0.97]	0.98/0.82	0.86 [0.78, 0.94]	0.94/0.77	0.85 [0.85, 1.00]	0.95/0.78	0.85 [0.68, 0.87]	0.97/0.72
PD 1-4 years vs controls	0.83 [0.72, 0.94]	0.80/0.78	0.86 [0.79, 0.94]	0.89/0.84	0.83 [0.71, 0.95]	0.85/0.85	0.84 [0.67, 1.00]	0.85/0.79
PD $<$ 1 year vs controls	0.68 [0.58, 0.79]	0.70/0.67	0.72 [0.63, 0.80]	0.76/0.68	0.69 [0.54, 0.84]	0.74/0.66	0.69 [0.49, 0.89]	0.72/0.64



(a) PD $>$ 4 years vs matched controls

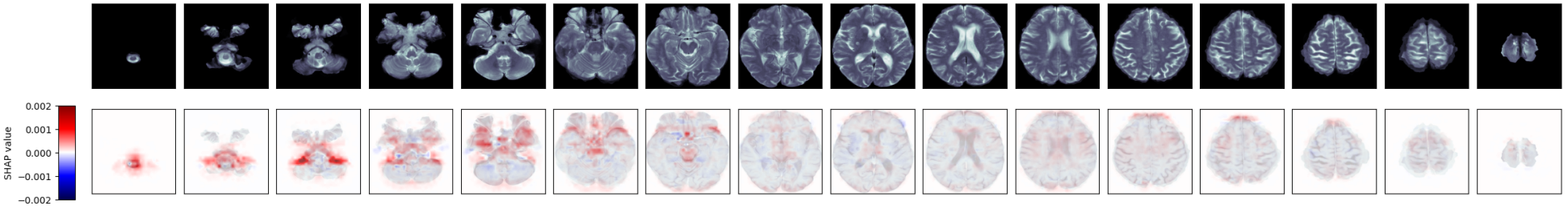


(b) PD 1-4 years vs matched controls

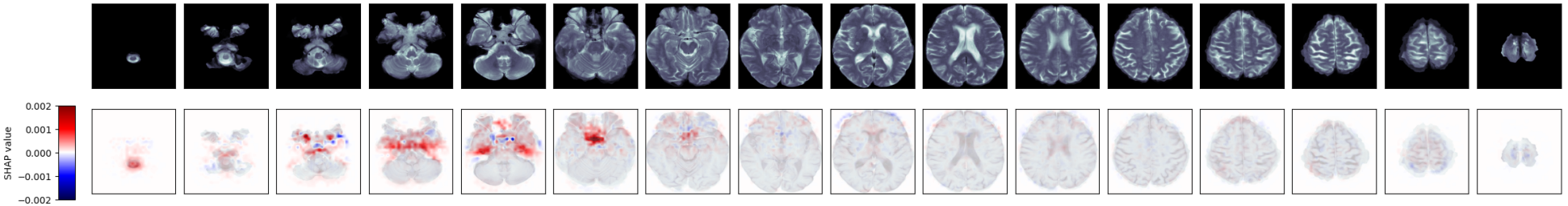


(c) PD $<$ 1 year vs matched controls

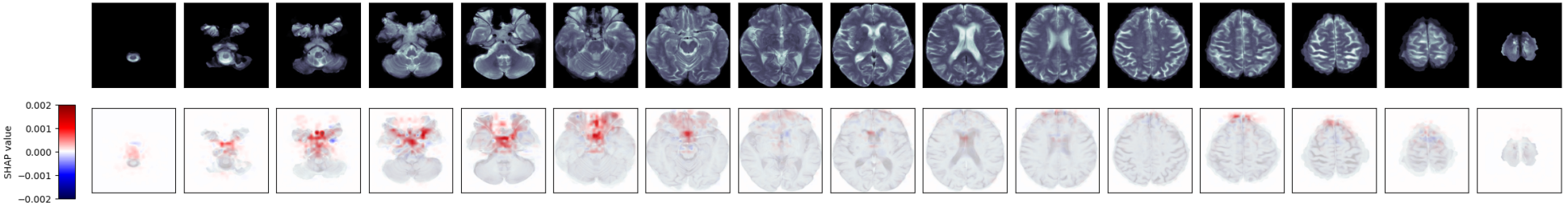
Figure 6.12: Receiver Operating Characteristic (ROC) curves for combined PPMI and UHPNT models. The bold red line represents the mean ROC curve; the dotted lines represent the ROC curve per k-fold.



(a) PD $>$ 4 years vs matched controls



(b) PD 1-4 years vs matched controls



(c) PD $<$ 1 year vs matched controls

Figure 6.13: Mean SHapley Additive exPlanation (SHAP) maps for correct predictions of PD in combined PPMI and UHPNT cohorts. Pixels highlighted in red have contributed positively to the prediction; pixels highlighted in blue have contributed negatively to the prediction.

6.6 Discussion

6.6.1 PPMI

Our models reliably differentiated PD from control scans. The findings suggest that deep learning can identify progressive changes in both idiopathic and genetic PD. The trained models exhibited good performance for established cases of idiopathic PD, which steadily decreased for models trained on earlier cases of PD.

Most of the interest of our models appears focused on CSF spaces surrounding the brainstem. Masking experiments suggested that it was not the brainstem itself which was informing the model predictions; it therefore seems likely that the regions of interest represent areas that have been affected by cortical changes. Models might have detected enlargement in the CSF spaces caused by cortical atrophy. Cerebral atrophy has previously been found in PD, with reported reductions in grey matter volume in the frontal and temporal lobes [181], and diffuse gyral atrophy throughout the temporal, parietal and frontal cortices [182]. A recent study has found enlargement of the interpeduncular and right ambient cisterns in patients with PD [183]. Another possibility is that the models have detected ventricular enlargement caused by cortical changes. Assymetric lateral ventricular enlargement has been reported in PD, associated with progression [184]. Ventricular enlargement has also been associated with cognitive decline in PD [185]. Such progressive patterns might explain the models' higher performance for later stages of PD.

To investigate whether these changes were visible in the genetic cohorts, we built models for the non-manifesting carriers of *LRRK2* and *GBA* variants, both genetic risk factors for PD. Between 28–74% of *LRRK2* and 8–10% of *GBA* carriers will develop PD [174, 175], hence a significant portion of these subjects would be expected to be within the prodromal phase of PD. Given the findings of our IPD models, we predicted that these models would not perform well, as any brain changes were anticipated to be early and subtle. However, the performance of these models was remarkably high for both *GBA* nPD and *LRRK2* nPD groups. Once again,

explainability techniques suggested a focus on pixels adjacent to the brainstem.

In the case of the non-manifesting *GBA* carriers, the models also demonstrated interest in pixels adjacent to the posterior parieto-occipital lobe. *GBA*-associated PD has been shown to be associated with a higher frequency of cognitive deficits in the visuospatial domain as well as visual hallucinations [186], hence atrophic changes within the wider visual processing regions would seem to be of particular relevance. Of less clear significance is the highlighting in *LRRK2* nPD groups of pixels adjacent to the cerebellum, particularly amongst older subjects. This has not previously been reported. It is notable that our models were able to correctly identify the brains of all the *LRRK2* nPD group who subsequently developed motor PD and in all cases the predictions were made with a high degree of certainty (Figure B.4 in Appendix B). In the case of both *GBA* nPD and *LRRK2* nPD, we are unable to make more certain judgement on the drivers of these very high model performances. It does however suggest that there is scope to use such techniques to identify carriers of these genetic risk factors using machine learning models. It may also suggest that early brain changes in *GBA* and *LRRK2* carriers are distinct from early brain changes in idiopathic PD, which would support speculation that genetic and idiopathic PD are separate disease pathways that converge on a broadly shared phenotype [187].

To further investigate these findings for potential progressive changes, we subdivided the unaffected *GBA* and *LRRK2* cohorts again by age, to assess whether there was any difference in the performance of models built to distinguish scans before and after the average age of onset of *GBA*/*LRRK2* PD. Unfortunately, among the *GBA* carriers, there were only enough scans taken after the average age of onset to build models for these. These had a slightly higher performance than models built for all *GBA* unaffected ages. For the *LRRK2* carriers, models built for scans from the older carriers yielded a marginally higher performance than the younger carriers. In both cases this may suggest that the brain differences become more pronounced with time.

A limitation of this work was that the impact of movement was not quantitatively

considered. Movement artefacts (from tremor, for example) are likely to have been present in the images and might have informed the models. Future work could consider extracting an estimation of intrascan motion using the diffusion-weighted imaging present in the PPMI dataset and the EDDY tool from the FMRIB Software Library [180]. This would allow the impact of motion upon the models to be analysed.

6.6.2 UHPNT

The models trained on UHPNT data were less reliably able to differentiate PD from control scans than those models trained on PPMI data alone. There are several likely reasons for this. The first is the size of the data: in the PPMI dataset there are double the number of total IPD scans (albeit from a slightly smaller number of individuals) than are present in the UHPNT dataset. In addition, the fewer UHPNT scans are more diverse in terms of PD progression, with many scans acquired prior to diagnosis, and therefore containing subtle (if any) early brain changes. The PPMI scans, in contrast, were all acquired after diagnosis, so on average represent more advanced cases of PD.

The external validation of the models trained on PPMI and UHPNT scans respectively yielded notably poor results. When tested in external datasets, the models performed no better than random guessing. This could suggest that the model is exhibiting extreme overfitting to the training data. However, the SHAP heatmaps for the trained models indicate that this is unlikely, given that they demonstrate focused interest in areas associated with the neuropathology of PD (rather than more diffuse interest that would suggest the learning of spurious patterns). It seems more likely in this case that the poor generalisability is due to fundamental differences between the two datasets, which made patterns learned in one ungeneralisable to the other. One such potentially insurmountable difference is the magnetic field strength of the scanners used. The vast majority of the PPMI scans were acquired using 3 T scanners, whereas the vast majority of the UHPNT scans were acquired using 1.5 T scanners. The greater magnetic field strength used to acquire the PPMI scans may

have increased the diagnostic capacity of the imaging, as the signal-to-noise ratio is increased and small structures can be seen more clearly [188]. In other studies, for example, 3 T has been shown to be superior at detecting brain atrophy [189] and at visualising the subthalamic nucleus [190].

Another key difference between the datasets is in the resolution of the images. The voxel dimensions varied between the datasets (and indeed within datasets); for example, the slice thickness in the PPMI dataset was most commonly 3mm, whereas in the UHPNT dataset it was most commonly 5mm. Images were resized to uniform dimensions prior to model input, but this step will have been conducted for vastly different quantities of voxels, and the results might not have been comparable. Future work should consider registering all the images to a common template to unify the resolutions.

A further significant difference between the two datasets is in the preprocessing. In the PPMI study, all T2 scans were acquired on the same day as a T1 scan. T1 scans allow for more precise brain extraction - so in order to extract the brain material from T2 scans, the T2 scan was registered to its matching T1 scan, and the extracted brain mask from the T1 scan was applied to the T2 scan. For the UHPNT dataset this stage was not possible, as T2 scans were not reliably accompanied by a T1 scan. Brain extraction was performed on the T2 scans alone, which is likely to have been less successful. This may have had effects such as obscuring atrophy.

Another possible difference is in demographics. The cohorts stratified by time since diagnosis are older in the UHPNT dataset (see Table 6.8) than in the PPMI dataset (see Table 6.2). The PPMI dataset is geographically more diverse. Other demographic data are anonymised in the UHPNT dataset, but it is likely that the subjects of the research PPMI dataset are wealthier and more highly educated than a sample of the general population.

Finally, the PPMI and UHPNT scans were acquired in very different environments. The PPMI scans were acquired according to a strict research protocol, with cohorts and exclusion criteria clearly defined. The UHPNT scans were acquired

routinely for any number of reasons in much more disparate circumstances, and probably contained a wide variety of confounding data. The nature of the controls is likely to be extremely different, as these are not subjects who have been chosen as healthy controls, but rather patients in need of a scan for a wide variety of reasons. Given more data, the NLP application developed in Chapter 5 might be used to automatically remove confounding data. However, in this case the size of the data was already too small to reduce further.

It is noteworthy that the models trained on the combined PPMI and UHPNT data performed slightly better than models trained on PPMI data alone, and that performance was notably higher for test PPMI scans than for test UHPNT scans. This suggests that the most informative data came from the PPMI images, but that the UHPNT images contained enough informative data to enhance the performance of the models beyond what could be learned from PPMI data alone. It is possible that training models on such diverse datasets allow more generalisable patterns to be learned. It would be valuable to validate these results in a further external dataset.

These varying results and these differences in datasets highlight the importance of not assuming that a model that performs well in a research setting will perform equally well in a routine clinical setting. External validation is rarely performed in applications of machine learning in medical imaging; these results demonstrate why this is a major limitation of these studies. The poor results for routinely collected data are suggestive that this kind of data is sub-optimal for this task, and that a more appropriate approach would be to develop optimal diagnostic tools in the clinic.

6.7 Concluding remarks

This chapter has presented a potential use of explainable AI for uncovering new insights in research datasets: discussing explainable models developed for the detection of PD in MRI brain scans and how these have detected progressive changes as well as differences in the brains of non-manifesting carriers of Parkinson's genetic risk

variants. The next, concluding chapter will discuss the contributions to knowledge, limitations, future directions, and conclusions from this thesis.

Chapter 7

Contributions to knowledge, future work and conclusion

7.1 Contributions to knowledge

7.1.1 Detection of aneurysm clips

In Chapter 4, deep learning models were developed to detect aneurysm clips in CT sagittal localizer scans with 100% sensitivity. This is the first safety tool developed for the automatic flagging of aneurysm clips prior to MRI appointments. The computational resources required to run the models are low, and the absence of dangerous false negatives makes these models particularly suitable for a safety application (although some false positives are likely). The use of SHAP in model development demonstrated the many benefits of explainable AI: early in development, it demonstrated that the models were being confounded by the presence of fillings, informing the decision to exclude the mandible from images. Later, SHAP highlighted that the pixels containing aneurysm clips were indeed contributing very strongly to the models' prediction that an aneurysm clip was present. SHAP also highlighted that other metal devices were responsible for false positives, but these are still valuable to detect for MRI safety. Explainable AI allowed for the development of a better model, and confirmed that the appropriate parts of the images

were informing model predictions, allowing the model to be more interpretable and trustworthy.

7.1.2 Detection of pathology in radiology reports

In Chapter 5, deep learning models were developed to classify abnormality and small vessel disease in radiology reports, with 91% accuracy and 93% accuracy respectively. This is the first natural language processing tool developed with the aim of automatically removing confounding pathology from radiology research cohorts. SHAP again enhanced the interpretability and trustworthiness of the models by highlighting that biologically relevant words were contributing most to model predictions.

7.1.3 Application of SHAP to 3D medical imaging

In Chapter 6, deep learning models were developed to classify the presence of Parkinson's disease in MRI brain scans. These were 3D scans, with multiple 2D frames, for which a bespoke explainability pipeline had to be built as the shap Python library does not contain the functionality to visualise explanations for 3D images. The provided explainers can calculate the Shapley values for 3D convolutional networks, as confirmed by the creator of the library [191], but this does not appear to be a common usage. In the literature review conducted for this thesis, no other examples were found of SHAP being applied to 3D medical imaging.

7.1.4 Identification of progressive Parkinson's brain changes

In Chapter 6, deep learning models developed to differentiate idiopathic Parkinson's disease from control scans exhibited good performance for established cases, but steadily lower performance for earlier cases. The SHAP explainability method highlighted that the same regions of interest, likely affected by cortical changes, were informing the different models trained on stratified cohorts, suggesting that progressive changes have been identified.

7.1.5 Identification of differences in the brains of non-manifesting carriers of Parkinson's genetic risk variants

In Chapter 6, deep learning models were developed to differentiate carriers of genetic risk variants for Parkinson's disease from control scans. Unexpectedly high performances were achieved, especially for models developed to detect scans from non-manifesting carriers of genetic variants. The SHAP explainability method highlighted that, as with the idiopathic Parkinson's cohorts, pixels surrounding the brainstem were informing these models, as well as pixels adjacent to the posterior parieto-occipital lobe in the case of non-manifesting *GBA* carriers, which is consistent with the known behaviour of *GBA*-associated Parkinson's. There is still much scope to explore the drivers of these high model performances.

7.1.6 Routinely collected dataset of Parkinson's imaging

In Chapter 6, deep learning models were developed to classify the presence of Parkinson's disease in a routinely-collected NHS dataset. The collation of and development of models for such a dataset is unprecedented. This retrospective dataset included imaging acquired from Parkinson's patients prior to their diagnosis, and so represented the prodromal as well as the more advanced stages of the disease. It is valuable to note that these routinely-collected images - acquired using a weaker magnetic field strength and covering a greater breadth of Parkinson's disease stages, and likely confounding pathology - did not yield models that performed as well as those trained on PPMI, a tightly-controlled research dataset. The additional external validation of the models trained on the PPMI and UHPNT datasets respectively represents a rare undertaking in AI in healthcare, due to the difficulty of acquiring these datasets. The results demonstrated the vital importance of external validation and a known issue of AI in healthcare: the difficulty of translating applications developed in research to clinical practice.

7.2 Limitations and future directions

7.2.1 Data

In several of the reported tasks, the sample size of data was smaller than would be hoped for the development of effective deep learning applications. This was largely due to the difficulty of acquiring radiological datasets. The aneurysm clips CT data, the radiological report data and the UHPNT MRI data were all acquired from a single institution, limiting the size of the data available. Further stratifying the data (e.g. by time of scan in relation to Parkinson's diagnosis) further reduced the size. This limitation was mitigated to an extent by the use of data augmentation (e.g. horizontal flip in the case of images), but the acquisition of more data might enable the development of even more accurate and more generalisable models.

7.2.2 External validation

A frequent limitation in AI in healthcare research is a lack of external validation. Cross-validation techniques allow for an approximation of how the developed models will perform on unseen data, but as the test data originate from the same distribution as the training data, there are likely to be commonalities that will not be seen in truly external data. This was seen in this work in the difference between the research PPMI dataset and the routinely-collected UHPNT dataset (discussed in Section 6.6.2). For the aneurysm clip models in Chapter 4, no external validation dataset could be acquired. Likewise for the radiology report text models in Chapter 5, no external validation dataset could be acquired. In the case of the models developed in Chapter 6, the PPMI and UHPNT datasets could serve as external validation datasets for models trained on the other, but such tests resulted in poor results, likely due to the discussed fundamental differences in the datasets. Additional models were trained on a combination of the datasets, and no further external validation dataset could be acquired to test these. In the future, further work should be done to validate all

of these models in external datasets.

7.2.3 Explainable AI

Explainable AI is extremely useful, as demonstrated, but it has its limitations. For the models developed for aneurysm clip and pathology detection, the use of explainable AI confirmed that the model was focusing on expected features; the models were focusing on pixels containing aneurysm clips as seen in Section 4.6, and biologically relevant words as seen in Section 5.6. This demonstrated the models' trustworthiness. However, the task of developing models for Parkinson's disease detection in conventional MRI sequences was different in nature, as it represented a task that could not be performed by clinicians; there were no known discriminatory features to be confirmed. Explainable AI was still extremely valuable in highlighting the parts of the images that were informing model predictions. However, there is only so much that the explainability method can reveal about the inner workings of the model. We can visualise which pixels were informing model predictions, but we can only speculate as to what about the nature of those pixels was informative. Causation is not considered by the explainability method, and highlighted features may be a result of the presence of noise, artifacts, or group differences from the underlying data [192]. There is much scope to further interrogate these results, especially in the cases of regions of interest that have not previously been reported.

SHAP

The flexibility and utility of the SHAP method was demonstrated in this research: it was implemented for three different tasks which used three different types of data. However, SHAP still has limitations. For example, any explainability method involving occlusion is strongly affected by sample size. The more examples seen in training, the more robust the model will be to changes in non-disease relevant areas [192]. As discussed in Section 7.2.1, some of the datasets used in this research are smaller than would be optimal, which may have affected the quality of the

explainability as well as the model performance.

The shap Python library also has weaknesses, such as the lack of functionality for visualising explanations for 3D convolutional neural networks. In medical imaging applications, which are likely to use volumetric imaging data, this is a major drawback. The development of a bespoke 3D explanation pipeline for this work has demonstrated that such functionality is possible. Its inclusion in the shap library as default would make 3D explanations much more accessible to researchers applying machine learning to medical imaging. Some more minor issues with the shap library include a lack of versatility in plot functions and sometimes unclear documentation. These again could be addressed to make SHAP more accessible to all.

Clinical expertise

Of vital importance to the future of AI in medical applications is the involvement of clinical experts. The research conducted for this thesis has been proposed and guided by clinicians who have the domain knowledge to know where AI could be useful, to interpret the outputs of explainability methods, and to propose refinements to the models based on those interpretations. For the aneurysm clip models developed in Chapter 4, it was radiological expertise which suggested that this would be a useful automatic safety check, which informed the decision to try lower-dimensional localizer images in which aneurysm clips could be clearly seen, and which noted that other metal devices were being highlighted in the heatmaps. For the pathology detection models developed in Chapter 5, it was radiological and neurological expertise which suggested that report texts could be used for automatic removal of confounding pathology from datasets, and which confirmed that biologically relevant words were informing the model predictions. For the Parkinson's disease detection models developed in Chapter 6, it was neurological and neuroradiological expertise which suggested that AI might usefully be able to detect early brain changes in MRI inaccessible to the human visual system, which proposed the stratification of cohorts by disease stage and genetic status, which interpreted the heatmaps as highlight-

ing biologically relevant regions, and which suggested that a fundamental difference between the UHPNT and PPMI datasets might be magnetic field strength.

This is a demonstration of a possible future of medical AI. As discussed in Section 2.2.1, AI has the potential to allow the radiological process to be streamlined, and in some cases entirely automated. It has the potential to lower the risk of human errors, and to allow clinicians to spend more time with patients. It has the potential to make healthcare more equitable. However, such a future is put in jeopardy by the risks of AI, discussed in Section 2.2.2. Previous applications of AI in healthcare have been beset by problems such as poor quality data, poor methodology, poor consideration of ethical issues, and poor interpretability of models, which has led to poorly performing models and mistrust of AI in healthcare. The combination of explainable AI and clinical expertise, as demonstrated in this research, is a potent pairing which can counter these issues. Using explainable AI to explain model predictions and using medical acumen to interpret those explanations in a clinical context is a powerful formula for detecting errors, verifying the validity of models, and engendering trust.

7.3 Conclusion

AI has enormous potential to save time and reduce errors in the field of radiology, as well as to open up new insights into disease mechanisms. However, little AI research has been translated into radiological practice due to difficulties such as access to high-quality, curated datasets and the limited interpretability and trustworthiness of "black box" deep learning technology. This thesis outlined several novel applications of explainable deep learning to medical imaging classification, demonstrating the potential utility of these applications in clinical settings. Deep learning models were developed to automatically flag aneurysm clips in advance of MRI appointments, and the use of explainable AI demonstrated that the models were focusing on the correct parts of the image. Deep learning models were also developed to detect abnormality in radiology report texts, allowing for the automatic removal of

confounding pathology in the curation of radiological datasets for the development of AI models, and the use of explainable AI demonstrated that the models were focusing on biologically relevant words. Finally, deep learning models were developed to detect Parkinson's disease in MRI scans, and the use of explainable AI demonstrated that the models were focusing on regions of interest that adhere to known neuropathology. This research has made contributions to MRI safety, medical imaging dataset curation and Parkinson's research. It highlights the immense potential of explainable AI techniques in radiological safety and research.

Bibliography

[1] Ahmed Hosny et al. "Artificial intelligence in radiology". en. In: *Nature Reviews Cancer* 18.8 (Aug. 2018), pp. 500–510. ISSN: 1474-1768. DOI: 10.1038/s41568-018-0016-5. URL: https://www.nature.com/articles/s41568-018-0016-5 (visited on 10/30/2023).

[2] Giles W. L. Boland, Alexander S. Guimaraes, and Peter R. Mueller. "The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization". en. In: *European Radiology* 19.1 (Jan. 2009), pp. 9–11. ISSN: 1432-1084. DOI: 10.1007/s00330-008-1159-7. URL: https://doi.org/10.1007/s00330-008-1159-7 (visited on 01/03/2024).

[3] Creative Destruction Lab. *Geoff Hinton: On Radiology*. Nov. 2016. URL: https://www.youtube.com/watch?v=2HMPRXstSvQ (visited on 01/03/2024).

[4] Kicky G. van Leeuwen et al. "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence". en. In: *European Radiology* 31.6 (June 2021), pp. 3797–3804. ISSN: 1432-1084. DOI: 10.1007/s00330-021-07892-z. URL: https://doi.org/10.1007/s00330-021-07892-z (visited on 01/03/2024).

[5] June-Goo Lee et al. "Deep Learning in Medical Imaging: General Overview". In: *Korean Journal of Radiology* 18.4 (May 2017). Publisher: The Korean Society of Radiology, pp. 570–584. DOI: 10.3348/kjr.2017.18.4.570. URL: https://synapse.koreamed.org/articles/1027354 (visited on 01/03/2024).

[6] Mauricio Reyes et al. "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities". In: *Radiology: Artificial Intelligence* 2.3 (May 2020). Publisher: Radiological Society of North America, e190043. DOI: 10.1148/ryai.2020190043. URL: https://pubs.rsna.org/doi/full/10.1148/ryai.2020190043 (visited on 01/17/2024).

[7] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010.

[8] John McCarthy et al. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955". en. In: *AI Magazine* 27.4 (Dec. 2006), pp. 12–12. ISSN: 2371-9621. DOI: 10.1609/aimag.v27i4.1904. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904 (visited on 10/26/2023).

[9] Melanie Mitchell. *Why AI is Harder Than We Think*. Apr. 2021. DOI: 10.48550/arXiv.2104.12871. URL: http://arxiv.org/abs/2104.12871 (visited on 10/30/2023).

[10] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (July 2015), pp. 255–260. DOI: 10.1126/science.aaa8415. URL: https://www.science.org/doi/full/10.1126/science.aaa8415 (visited on 10/30/2023).

[11] Francois Chollet. *Deep Learning with Python, Second Edition*. en. Google-Books-ID: mjVKEAAAQBAJ. Simon and Schuster, Dec. 2021. ISBN: 978-1-63835-009-5.

[12] Christoph Molnar. *Modeling Mindsets: The Many Cultures of Learning from Data*. Mucbook Clubhouse, 2022.

[13] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised Learning". en. In: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Ed. by Matthieu Cord and Pádraig Cunningham. Cognitive Technologies. Berlin, Heidelberg: Springer, 2008, pp. 21–49. ISBN: 978-3-540-75171-7. DOI: 10.1007/978-3-540-75171-7_2. URL: https://doi.org/10.1007/978-3-540-75171-7_2 (visited on 10/30/2023).

[14] Z. Q. John Lu. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 173.3 (July 2010), pp. 693–694. ISSN: 0964-1998. DOI: 10.1111/j.1467-985X.2010.00646_6.x. URL: https://doi.org/10.1111/j.1467-985X.2010.00646_6.x (visited on 10/30/2023).

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. en. MIT Press, Nov. 2016. ISBN: 978-0-262-33737-3.

[16] Yann LeCun, Yoshua Bengio, et al. *The handbook of brain theory and neural networks*. 1998.

[17] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". en. In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252. ISSN: 1573-1405. DOI: 10.1007/s11263-015-0816-y. URL: https://doi.org/10.1007/s11263-015-0816-y (visited on 11/01/2023).

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://dl.acm.org/doi/10.1145/3065386 (visited on 11/01/2023).

[19] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 2015. DOI: 10.48550/arXiv.1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 12/19/2023).

[20] Bram van Ginneken, Cornelia M. Schaefer-Prokop, and Mathias Prokop. "Computer-aided Diagnosis: How to Move from the Laboratory to the Clinic". In: *Radiology* 261.3 (Dec. 2011), pp. 719–732. ISSN: 0033-8419. DOI: 10.1148/radiol.11091710. URL: https://pubs.rsna.org/doi/abs/10.1148/radiol.11091710 (visited on 10/30/2023).

[21] Edward P. Ambinder. "A History of the Shift Toward Full Computerization of Medicine". In: *Journal of Oncology Practice* 1.2 (July 2005), pp. 54–56. ISSN: 1554-7477. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2793587/ (visited on 10/30/2023).

[22] Andre Esteva et al. “Deep learning-enabled medical computer vision”. en. In: *npj Digital Medicine* 4.1 (Jan. 2021). Number: 1 Publisher: Nature Publishing Group, pp. 1–9. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00376-2. URL: https://www.nature.com/articles/s41746-020-00376-2 (visited on 10/31/2023).

[23] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (Dec. 2017), pp. 60–88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135 (visited on 10/31/2023).

[24] Rebecca Smith-Bindman et al. “Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016”. In: *JAMA* 322.9 (Sept. 2019), pp. 843–856. ISSN: 0098-7484. DOI: 10.1001/jama.2019.11456. URL: https://doi.org/10.1001/jama.2019.11456 (visited on 03/18/2024).

[25] Robert J McDonald et al. “The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload”. In: *Academic radiology* 22.9 (2015). Publisher: Elsevier, pp. 1191–1198.

[26] Nadia Stec et al. “A Systematic Review of Fatigue in Radiology: Is It a Problem?” In: *American Journal of Roentgenology* 210.4 (Apr. 2018). Publisher: American Roentgen Ray Society, pp. 799–806. ISSN: 0361-803X. DOI: 10.2214/AJR.17.18613. URL: https://ajronline.org/doi/full/10.2214/AJR.17.18613 (visited on 03/18/2024).

[27] Michael Recht and R. Nick Bryan. “Artificial Intelligence: Threat or Boon to Radiologists?” In: *Journal of the American College of Radiology* 14.11 (Nov. 2017), pp. 1476–1480. ISSN: 1546-1440. DOI: 10.1016/j.jacr.2017.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1546144017308347 (visited on 11/01/2023).

[28] Q. Waymel et al. “Impact of the rise of artificial intelligence in radiology: What do radiologists think?” In: *Diagnostic and Interventional Imaging* 100.6 (June 2019), pp. 327–336. ISSN: 2211-5684. DOI: 10.1016/j.diii.2019.03.015. URL: https://www.sciencedirect.com/science/article/pii/S2211568419300907 (visited on 03/18/2024).

[29] Trafton Drew, Melissa L.-H. Võ, and Jeremy M. Wolfe. “The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Observers”. en. In: *Psychological Science* 24.9 (Sept. 2013). Publisher: SAGE Publications Inc, pp. 1848–1853. ISSN: 0956-7976. DOI: 10.1177/0956797613479386. URL: https://doi.org/10.1177/0956797613479386 (visited on 11/01/2023).

[30] Katie Chockley and Ezekiel Emanuel. “The End of Radiology? Three Threats to the Future Practice of Radiology”. In: *Journal of the American College of Radiology* 13.12, Part A (Dec. 2016), pp. 1415–1420. ISSN: 1546-1440. DOI: 10.1016/j.jacr.2016.07.010. URL: https://www.sciencedirect.com/science/article/pii/S1546144016305907 (visited on 11/01/2023).

[31] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. arXiv:1901.07042 [cs, eess]. Nov. 2019. DOI: 10.48550/arXiv.1901.07042. URL: http://arxiv.org/abs/1901.07042 (visited on 11/01/2023).

[32] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019). Number: 01, pp. 590–597. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.3301590. URL: https://ojs.aaai.org/index.php/AAAI/article/view/3834 (visited on 11/01/2023).

[33] X Wang et al. "Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *IEEE CVPR*. Vol. 7. 2017.

[34] E. J. Yates, L. C. Yates, and H. Harvey. "Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification". In: *Clinical Radiology* 73.9 (Sept. 2018), pp. 827–831. ISSN: 0009-9260. DOI: 10.1016/j.crad.2018.05.015. URL: https://www.sciencedirect.com/science/article/pii/S000992601830206X (visited on 11/01/2023).

[35] William Gale et al. *Detecting hip fractures with radiologist-level performance using deep neural networks*. arXiv:1711.06504 [cs, stat]. Nov. 2017. DOI: 10.48550/arXiv.1711.06504. URL: http://arxiv.org/abs/1711.06504 (visited on 11/01/2023).

[36] Mohammad R. Arbabshirani et al. "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration". en. In: *npj Digital Medicine* 1.1 (Apr. 2018). Number: 1 Publisher: Nature Publishing Group, pp. 1–7. ISSN: 2398-6352. DOI: 10.1038/s41746-017-0015-z. URL: https://www.nature.com/articles/s41746-017-0015-Z (visited on 11/01/2023).

[37] Koichiro Yasaka et al. "Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study". In: *Radiology* 286.3 (Mar. 2018). Publisher: Radiological Society of North America, pp. 887–896. ISSN: 0033-8419. DOI: 10.1148/radiol.2017170706. URL: https://pubs.rsna.org/doi/abs/10.1148/radiol.2017170706 (visited on 11/01/2023).

[38] Amir Bar et al. "Compression fractures detection on CT". In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. SPIE, Mar. 2017, pp. 1036–1043. DOI: 10.1117/12.2249635. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10134/1013440/Compression-fractures-detection-on-CT/10.1117/12.2249635.full (visited on 11/01/2023).

[39] Ju Gang Nam et al. "Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs". In: *Radiology* 290.1 (Jan. 2019). Publisher: Radiological Society of North America, pp. 218–228. ISSN: 0033-8419. DOI: 10.1148/radiol.2018180237. URL: https://pubs.rsna.org/doi/full/10.1148/radiol.2018180237 (visited on 11/01/2023).

[40] Emma Pierson et al. “An algorithmic approach to reducing unexplained pain disparities in underserved populations”. en. In: *Nature Medicine* 27.1 (Jan. 2021). Number: 1 Publisher: Nature Publishing Group, pp. 136–140. ISSN: 1546-170X. DOI: 10.1038/s41591-020-01192-7. URL: https://www.nature.com/articles/s41591-020-01192-7 (visited on 11/01/2023).

[41] Michael Roberts et al. “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. en. In: *Nature Machine Intelligence* 3.3 (Mar. 2021). Number: 3 Publisher: Nature Publishing Group, pp. 199–217. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00307-0. URL: https://www.nature.com/articles/s42256-021-00307-0 (visited on 11/01/2023).

[42] Laure Wynants et al. “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal”. en. In: *BMJ* 369 (Apr. 2020). Publisher: British Medical Journal Publishing Group Section: Research, p. m1328. ISSN: 1756-1833. DOI: 10.1136/bmj.m1328. URL: https://www.bmj.com/content/369/bmj.m1328 (visited on 11/01/2023).

[43] Gianluca Maguolo and Loris Nanni. “A critic evaluation of methods for COVID-19 automatic detection from X-ray images”. In: *Information Fusion* 76 (Dec. 2021), pp. 1–7. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.04.008. URL: https://www.sciencedirect.com/science/article/pii/S1566253521000816 (visited on 11/01/2023).

[44] John R. Zech et al. “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study”. en. In: *PLOS Medicine* 15.11 (Nov. 2018). Publisher: Public Library of Science, e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683. URL: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683 (visited on 11/01/2023).

[45] Emma Beede et al. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376718. URL: https://dl.acm.org/doi/10.1145/3313831.3376718 (visited on 01/10/2024).

[46] Julia Powles and Hal Hodson. “Google DeepMind and healthcare in an age of algorithms”. en. In: *Health and Technology* 7.4 (Dec. 2017), pp. 351–367. ISSN: 2190-7196. DOI: 10.1007/s12553-017-0179-1. URL: https://doi.org/10.1007/s12553-017-0179-1 (visited on 11/01/2023).

[47] HB McMahan et al. “Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)”. In: (2017). Publisher: Ft. Lauderdale, FL, USA.

[48] Fei Wang, Rainu Kaushal, and Dhruv Khullar. “Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine?” In: *Annals of Internal Medicine* 172.1 (Jan. 2020). Publisher: American College of Physicians, pp. 59–60. ISSN: 0003-4819. DOI: 10.7326/M19-2548. URL:

https://www.acpjournals.org/doi/full/10.7326/M19-2548 (visited on 11/01/2023).

[49] Christine M. Cutillo et al. "Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency". en. In: *npj Digital Medicine* 3.1 (Mar. 2020). Number: 1 Publisher: Nature Publishing Group, pp. 1–5. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0254-2. URL: https://www.nature.com/articles/s41746-020-0254-2 (visited on 11/01/2023).

[50] *Art. 13 GDPR – Information to be provided where personal data are collected from the data subject*. en-US. URL: https://gdpr-info.eu/art-13-gdpr/ (visited on 03/11/2024).

[51] Roberto Confalonieri et al. "A historical perspective of explainable Artificial Intelligence". en. In: *WIREs Data Mining and Knowledge Discovery* 11.1 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1391, e1391. ISSN: 1942-4795. DOI: 10.1002/widm.1391. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391 (visited on 01/17/2024).

[52] Plamen P. Angelov et al. "Explainable artificial intelligence: an analytical review". en. In: *WIREs Data Mining and Knowledge Discovery* 11.5 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1424, e1424. ISSN: 1942-4795. DOI: 10.1002/widm.1424. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1424 (visited on 01/17/2024).

[53] Hui Wen Loh et al. "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)". In: *Computer Methods and Programs in Biomedicine* 226 (Nov. 2022), p. 107161. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2022.107161. URL: https://www.sciencedirect.com/science/article/pii/S0169260722005429 (visited on 01/17/2024).

[54] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608 [cs, stat]. Mar. 2017. DOI: 10.48550/arXiv.1702.08608. URL: http://arxiv.org/abs/1702.08608 (visited on 01/17/2024).

[55] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. "Explainable Deep Learning Models in Medical Image Analysis". en. In: *Journal of Imaging* 6.6 (June 2020). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 52. ISSN: 2313-433X. DOI: 10.3390/jimaging6060052. URL: https://www.mdpi.com/2313-433X/6/6/52 (visited on 11/01/2023).

[56] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care". English. In: *The Lancet Digital Health* 3.11 (Nov. 2021). Publisher: Elsevier, e745–e750. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(21)00208-9. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext?tpcc=nleyeonai (visited on 11/01/2023).

[57] Forough Poursabzi-Sangdeh et al. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–52. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445315. URL: https://dl.acm.org/doi/10.1145/3411764.3445315 (visited on 11/01/2023).

[58] Robert Love. *Linux Kernel Development*. en. Google-Books-ID: 3MWRMYRwulIC. Pearson Education, June 2010. ISBN: 978-0-7686-9679-0.

[59] Keir Thomas. *Beginning Ubuntu Linux: From Novice to Professional*. en. Google-Books-ID: YkptQV6f7L8C. Apress, Dec. 2006. ISBN: 978-1-4302-0137-3.

[60] *Open source AI*. en. URL: https://ubuntu.com/ai (visited on 11/29/2023).

[61] Mark Lutz. *Programming Python*. en. "O’Reilly Media, Inc.", 2001. ISBN: 978-0-596-00085-1.

[62] *NumPy*. URL: https://numpy.org/ (visited on 11/29/2023).

[63] Wes McKinney. “Data Structures for Statistical Computing in Python”. en. In: Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a. URL: https://conference.scipy.org/proceedings/scipy2010/mckinney.html (visited on 11/01/2023).

[64] John D Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in science & engineering* 9.03 (2007). Publisher: IEEE Computer Society, pp. 90–95.

[65] Michael Waskom. “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60 (Apr. 2021), p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021. URL: https://joss.theoj.org/papers/10.21105/joss.03021 (visited on 11/29/2023).

[66] Gary Bradski, Adrian Kaehler, et al. “OpenCV”. In: *Dr. Dobb’s journal of software tools* 3.2 (2000).

[67] *imutils*. original-date: 2015-01-11T20:05:39Z. Nov. 2023. URL: https://github.com/PyImageSearch/imutils (visited on 11/29/2023).

[68] Stefan Van der Walt et al. “scikit-image: image processing in Python”. In: *PeerJ* 2 (2014). Publisher: PeerJ Inc., e453.

[69] *Pillow*. en. URL: https://pillow.readthedocs.io/en/stable/index.html (visited on 11/29/2023).

[70] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020). Publisher: Nature Publishing Group, pp. 261–272.

[71] *Pydicom /*. URL: https://pydicom.github.io/ (visited on 11/29/2023).

[72] *dicom2nifti — dicom2nifti documentation*. URL: https://dicom2nifti.readthedocs.io/en/latest/readme.html (visited on 11/29/2023).

[73] *Neuroimaging in Python — NiBabel 5.1.0 documentation*. URL: https://nipy.org/nibabel/ (visited on 11/29/2023).

[74] *Anaconda | The World's Most Popular Data Science Platform*. en-US. URL: https://www.anaconda.com/ (visited on 11/29/2023).

[75] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". en. In: *MACHINE LEARNING IN PYTHON* ().

[76] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. arXiv:1603.04467 [cs]. Mar. 2016. DOI: 10.48550/arXiv.1603.04467. URL: http://arxiv.org/abs/1603.04467 (visited on 10/31/2023).

[77] François Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[78] *Project Jupyter*. en. URL: https://jupyter.org (visited on 11/29/2023).

[79] *CUDA Zone - Library of Resources*. en-US. URL: https://developer.nvidia.com/cuda-zone (visited on 11/29/2023).

[80] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980 [cs]. Jan. 2017. DOI: 10.48550/arXiv.1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on 10/31/2023).

[81] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 11/01/2023).

[82] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929 [cs]. June 2021. DOI: 10.48550/arXiv.2010.11929. URL: http://arxiv.org/abs/2010.11929 (visited on 11/01/2023).

[83] P. J. G Lisboa, A Vellido, and H Wong. "Bias reduction in skewed binary classification with Bayesian neural networks". In: *Neural Networks* 13.4 (June 2000), pp. 407–410. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(00)00022-8. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000228 (visited on 06/05/2024).

[84] Jayawant N. Mandrekar. "Receiver Operating Characteristic Curve in Diagnostic Test Assessment". In: *Journal of Thoracic Oncology* 5.9 (Sept. 2010), pp. 1315–1316. ISSN: 1556-0864. DOI: 10.1097/JTO.0b013e3181ec173d. URL: https://www.sciencedirect.com/science/article/pii/S1556086415306043 (visited on 11/29/2023).

[85] Lutz Prechelt. "Early Stopping - But When?" en. In: *Neural Networks: Tricks of the Trade*. Ed. by Genevieve B. Orr and Klaus-Robert Müller. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1998, pp. 55–69. ISBN: 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_3. URL: https://doi.org/10.1007/3-540-49430-8_3 (visited on 03/11/2024).

[86] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". en. In: ().

[87] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[88] Lloyd S Shapley et al. "A value for n-person games". In: (1953). Publisher: Princeton University Press Princeton.

[89] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 12/18/2023).

[90] Bradley Walters et al. "How to Open a Black Box Classifier for Tabular Data". en. In: *Algorithms* 16.4 (Apr. 2023). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 181. ISSN: 1999-4893. DOI: 10.3390/a16040181. URL: https://www.mdpi.com/1999-4893/16/4/181 (visited on 06/05/2024).

[91] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *Model-Agnostic Interpretability of Machine Learning*. arXiv:1606.05386 [cs, stat]. June 2016. DOI: 10.48550/arXiv.1606.05386. URL: http://arxiv.org/abs/1606.05386 (visited on 12/18/2023).

[92] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". en. In: *Proceedings of the 34th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2017, pp. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html (visited on 12/18/2023).

[93] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". en. In: *PLOS ONE* 10.7 (July 2015). Publisher: Public Library of Science, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140 (visited on 12/18/2023).

[94] C Molnar. "Interpreting machine learning models with SHAP". In: *Lulu. com* (2023).

[95] Joseph T McFadden. "Magnetic resonance imaging and aneurysm clips: a review". In: *Journal of neurosurgery* 117.1 (2012). Publisher: American Association of Neurological Surgeons, pp. 1–11.

[96] Mary F Dempsey, Barrie Condon, and Donald M Hadley. "MRI safety review". In: *Seminars in Ultrasound, CT and MRI* 23.5 (Oct. 2002), pp. 392–401. ISSN: 0887-2171. DOI: 10.1016/S0887-2171(02)90010-7. URL: https://www.sciencedirect.com/science/article/pii/S0887217102900107 (visited on 10/31/2023).

[97] R P Klucznik et al. "Placement of a ferromagnetic intracerebral aneurysm clip in a magnetic field with a fatal outcome." In: *Radiology* 187.3 (June 1993). Publisher: Radiological Society of North America, pp. 855–856. ISSN: 0033-8419. DOI: 10.1148/radiology.187.3.8497645. URL: https://pubs.rsna.org/doi/abs/10.1148/radiology.187.3.8497645 (visited on 10/31/2023).

[98] A. Cunqueiro et al. "Performing MRI on patients with MRI-conditional and non-conditional cardiac implantable electronic devices: an update for radiologists". In: *Clinical Radiology* 74.12 (Dec. 2019), pp. 912–917. ISSN: 0009-9260. DOI: 10.1016/j.crad.2019.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0009926019303502 (visited on 10/31/2023).

[99] Frank G Shellock and Alberto Spinazzi. "MRI safety update 2008: part 2, screening patients for MRI." In: *American Journal of Roentgenology* 191.4 (2008), p. 1140.

[100] Mark Daniel Vernon Thurston, Daniel H. Kim, and Huub K. Wit. "Neural Network Detection of Pacemakers for MRI Safety". en. In: *Journal of Digital Imaging* 35.6 (Dec. 2022), pp. 1673–1680. ISSN: 1618-727X. DOI: 10.1007/s10278-022-00663-2. URL: https://doi.org/10.1007/s10278-022-00663-2 (visited on 10/31/2023).

[101] Hee-Seok Yang et al. "Deep Learning Application in Spinal Implant Identification". en-US. In: *Spine* 46.5 (Mar. 2021), E318. ISSN: 0362-2436. DOI: 10.1097/BRS.0000000000003844. URL: https://journals.lww.com/spinejournal/abstract/2021/03010/deep_learning_application_in_spinal_implant.12.aspx (visited on 10/31/2023).

[102] Aviwe Kohlakala et al. "Deep learning-based dental implant recognition using synthetic X-ray images". en. In: *Medical & Biological Engineering & Computing* 60.10 (Oct. 2022), pp. 2951–2968. ISSN: 1741-0444. DOI: 10.1007/s11517-022-02642-9. URL: https://doi.org/10.1007/s11517-022-02642-9 (visited on 10/31/2023).

[103] Ravi Patel et al. "Automated Identification of Orthopedic Implants on Radiographs Using Deep Learning". In: *Radiology: Artificial Intelligence* 3.4 (July 2021). Publisher: Radiological Society of North America, e200183. DOI: 10.1148/ryai.2021200183. URL: https://pubs.rsna.org/doi/full/10.1148/ryai.2021200183 (visited on 10/31/2023).

[104] *DCMTK - dicom.offis.de*. URL: https://dicom.offis.de/dcmtk/ (visited on 10/31/2023).

[105] K. Y. E. Aryanto et al. "A Web-Based Institutional DICOM Distribution System with the Integration of the Clinical Trial Processor (CTP)". en. In: *Journal of Medical Systems* 39.5 (Mar. 2015), p. 45. ISSN: 1573-689X. DOI: 10.1007/s10916-014-0186-y. URL: https://doi.org/10.1007/s10916-014-0186-y (visited on 10/31/2023).

[106] Thomas Kluyver et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." In: *Elpub* 2016 (2016), pp. 87–90.

[107] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 [cs]. Apr. 2015. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 10/31/2023).

[108] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: 2016, pp. 2818–2826. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html (visited on 10/31/2023).

[109] Francois Chollet. "Xception: Deep Learning With Depthwise Separable Convolutions". In: 2017, pp. 1251–1258. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html (visited on 10/31/2023).

[110] Gao Huang et al. “Densely Connected Convolutional Networks”. In: 2017, pp. 4700–4708. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html (visited on 10/31/2023).

[111] Mark Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: 2018, pp. 4510–4520. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html (visited on 10/31/2023).

[112] Sihong Chen, Kai Ma, and Yefeng Zheng. *Med3D: Transfer Learning for 3D Medical Image Analysis*. arXiv:1904.00625 [cs]. July 2019. DOI: 10.48550/arXiv.1904.00625. URL: http://arxiv.org/abs/1904.00625 (visited on 10/31/2023).

[113] Abien Fred Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. arXiv:1803.08375 [cs, stat]. Feb. 2019. DOI: 10.48550/arXiv.1803.08375. URL: http://arxiv.org/abs/1803.08375 (visited on 03/19/2024).

[114] Omer Sagi and Lior Rokach. “Ensemble learning: A survey”. en. In: *WIREs Data Mining and Knowledge Discovery* 8.4 (2018), e1249. ISSN: 1942-4795. DOI: 10.1002/widm.1249. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249 (visited on 10/31/2023).

[115] Vera Sorin et al. “Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review”. In: *Journal of the American College of Radiology* 17.5 (May 2020), pp. 639–648. ISSN: 1546-1440. DOI: 10.1016/j.jacr.2019.12.026. URL: https://www.sciencedirect.com/science/article/pii/S154614402030003X (visited on 11/01/2023).

[116] Erik Cambria and Bebo White. “Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]”. In: *IEEE Computational Intelligence Magazine* 9.2 (May 2014). Conference Name: IEEE Computational Intelligence Magazine, pp. 48–57. ISSN: 1556-6048. DOI: 10.1109/MCI.2014.2307227. URL: https://ieeexplore.ieee.org/abstract/document/6786458 (visited on 11/01/2023).

[117] Ewoud Pons et al. “Natural Language Processing in Radiology: A Systematic Review”. In: *Radiology* 279.2 (May 2016). Publisher: Radiological Society of North America, pp. 329–343. ISSN: 0033-8419. DOI: 10.1148/radiol.16142770. URL: https://pubs.rsna.org/doi/abs/10.1148/radiol.16142770 (visited on 11/01/2023).

[118] Arlene Casey et al. “A systematic review of natural language processing applied to radiology reports”. en. In: *BMC Medical Informatics and Decision Making* 21.1 (June 2021), p. 179. ISSN: 1472-6947. DOI: 10.1186/s12911-021-01533-7. URL: https://doi.org/10.1186/s12911-021-01533-7 (visited on 11/01/2023).

[119] Tianrun Cai et al. “Natural Language Processing Technologies in Radiology Research and Clinical Applications”. In: *RadioGraphics* 36.1 (Jan. 2016). Publisher: Radiological Society of North America, pp. 176–191. ISSN: 0271-5333. DOI: 10.1148/rg.2016150080. URL: https://pubs.rsna.org/doi/abs/10.1148/rg.2016150080 (visited on 11/01/2023).

[120] Sascha Dublin et al. "Natural Language Processing to identify pneumonia from radiology reports". en. In: *Pharmacoepidemiology and Drug Safety* 22.8 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.3418, pp. 834–841. ISSN: 1099-1557. DOI: 10.1002/pds.3418. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3418 (visited on 11/01/2023).

[121] Kim N. Danforth et al. "Automated Identification of Patients With Pulmonary Nodules in an Integrated Health System Using Administrative Health Plan Data, Radiology Reports, and Natural Language Processing". In: *Journal of Thoracic Oncology* 7.8 (Aug. 2012), pp. 1257–1262. ISSN: 1556-0864. DOI: 10.1097/JTO.0b013e31825bd9f5. URL: https://www.sciencedirect.com/science/article/pii/S1556086415326915 (visited on 11/01/2023).

[122] Sheng Yu et al. "Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing". In: *Journal of Biomedical Informatics*. Special Section: Methods in Clinical Research Informatics 52 (Dec. 2014), pp. 386–393. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2014.08.001. URL: https://www.sciencedirect.com/science/article/pii/S1532046414001828 (visited on 11/01/2023).

[123] Guergana K. Savova et al. "Discovering Peripheral Arterial Disease Cases from Radiology Notes Using Natural Language Processing". In: *AMIA Annual Symposium Proceedings* 2010 (2010), pp. 722–726. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041293/ (visited on 11/01/2023).

[124] Joseph S. Redman et al. "Accurate Identification of Fatty Liver Disease in Data Warehouse Utilizing Natural Language Processing". en. In: *Digestive Diseases and Sciences* 62.10 (Oct. 2017), pp. 2713–2718. ISSN: 1573-2568. DOI: 10.1007/s10620-017-4721-9. URL: https://doi.org/10.1007/s10620-017-4721-9 (visited on 11/01/2023).

[125] Yvonne Sada et al. "Validation of Case Finding Algorithms for Hepatocellular Cancer from Administrative Data and Electronic Health Records using Natural Language Processing". In: *Medical care* 54.2 (Feb. 2016), e9–e14. ISSN: 0025-7079. DOI: 10.1097/MLR.0b013e3182a30373. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875602/ (visited on 11/01/2023).

[126] Andrew Yu Li and Nikki Elliot. "Natural language processing to identify ureteric stones in radiology reports". en. In: *Journal of Medical Imaging and Radiation Oncology* 63.3 (2019), pp. 307–310. ISSN: 1754-9485. DOI: 10.1111/1754-9485.12861. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.12861 (visited on 11/01/2023).

[127] Hadley Wickham. "Tidy Data". en. In: *Journal of Statistical Software* 59 (Sept. 2014), pp. 1–23. ISSN: 1548-7660. DOI: 10.18637/jss.v059.i10. URL: https://doi.org/10.18637/jss.v059.i10 (visited on 11/01/2023).

[128] Alida A. Gouw et al. "Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations". en. In: *Journal of Neurology, Neurosurgery & Psychiatry* 82.2 (Feb. 2011). Publisher: BMJ Publishing Group Ltd Section: Review, pp. 126–135. ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp.2009.204685. URL: https://jnnp.bmj.com/content/82/2/126 (visited on 11/01/2023).

[129] Stephen K. Van Den Eeden et al. "Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity". In: *American Journal of Epidemiology* 157.11 (June 2003), pp. 1015–1022. ISSN: 0002-9262. DOI: 10.1093/aje/kwg068. URL: https://doi.org/10.1093/aje/kwg068 (visited on 11/02/2023).

[130] Bastiaan R. Bloem, Michael S. Okun, and Christine Klein. "Parkinson's disease". English. In: *The Lancet* 397.10291 (June 2021). Publisher: Elsevier, pp. 2284–2303. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(21)00218-X. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00218-X/fulltext (visited on 03/13/2024).

[131] William Dauer and Serge Przedborski. "Parkinson's Disease: Mechanisms and Models". English. In: *Neuron* 39.6 (Sept. 2003). Publisher: Elsevier, pp. 889–909. ISSN: 0896-6273. DOI: 10.1016/S0896-6273(03)00568-3. URL: https://www.cell.com/neuron/abstract/S0896-6273(03)00568-3 (visited on 11/01/2023).

[132] Joseph Jankovic. "Parkinson's disease: clinical features and diagnosis". In: *Journal of neurology, neurosurgery & psychiatry* 79.4 (2008). Publisher: BMJ Publishing Group Ltd, pp. 368–376.

[133] Abdul Qayyum Rana et al. "Parkinson's disease: a review of non-motor symptoms". In: *Expert Review of Neurotherapeutics* 15.5 (May 2015). Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14737175.2015.1038244, pp. 549–562. ISSN: 1473-7175. DOI: 10.1586/14737175.2015.1038244. URL: https://doi.org/10.1586/14737175.2015.1038244 (visited on 11/01/2023).

[134] James Parkinson. "An Essay on the Shaking Palsy". In: *The Journal of Neuropsychiatry and Clinical Neurosciences* 14.2 (May 2002). Publisher: American Psychiatric Publishing, pp. 223–236. ISSN: 0895-0172. DOI: 10.1176/jnp.14.2.223. URL: https://neuro.psychiatryonline.org/doi/full/10.1176/jnp.14.2.223 (visited on 11/01/2023).

[135] Christopher G. Goetz. "The History of Parkinson's Disease: Early Clinical Descriptions and Neurological Therapies". In: *Cold Spring Harbor Perspectives in Medicine:* 1.1 (Sept. 2011), a008862. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a008862. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234454/ (visited on 11/01/2023).

[136] Stephen Mullin and Anthony H. V. Schapira. "Pathogenic Mechanisms of Neurodegeneration in Parkinson Disease". English. In: *Neurologic Clinics* 33.1 (Feb. 2015). Publisher: Elsevier, pp. 1–17. ISSN: 0733-8619, 1557-9875. DOI: 10.1016/j.ncl.2014.09.010. URL: https://www.neurologic.theclinics.com/article/S0733-8619(14)00078-4/fulltext (visited on 11/01/2023).

[137] E. Ray Dorsey et al. "The Emerging Evidence of the Parkinson Pandemic". en. In: *Journal of Parkinson's Disease* 8.s1 (Jan. 2018). Publisher: IOS Press, S3–S8. ISSN: 1877-7171. DOI: 10.3233/JPD-181474. URL: https://content.iospress.com/articles/journal-of-parkinsons-disease/jpd181474 (visited on 11/01/2023).

[138] FC Rose. *James Parkinson His Life and Times*. Springer Science & Business Media, 2013.

[139] Valery L Feigin et al. “Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015”. In: *The Lancet Neurology* 16.11 (2017). Publisher: Elsevier, pp. 877–897.

[140] N Maserejian, L. Vinikoor-Imler, and A. Dilley. “Estimation of the 2020 Global Population of Parkinson’s Disease (PD) [abstract]”. In: *Mov Disord* 35 (suppl 1) (2020).

[141] FH Lewy and M Lewandowsky. “Handbuch der Neurologie”. In: *Berlin: Julius Springer* (1912).

[142] Constantin Trétiakoff. “Contribution a l’etude de l’Anatomie pathologique du Locus Niger de Soemmering avec quelques deduction relatives a la pathogenie des troubles du tonus musculaire et de la maladie de Parkinson”. In: *Theses de Paris* (1919).

[143] Mihael H Polymeropoulos et al. “Mutation in the α-synuclein gene identified in families with Parkinson’s disease”. In: *science* 276.5321 (1997). Publisher: American Association for the Advancement of Science, pp. 2045–2047.

[144] Heiko Braak et al. “Staging of brain pathology related to sporadic Parkinson’s disease”. In: *Neurobiology of Aging* 24.2 (Mar. 2003), pp. 197–211. ISSN: 0197-4580. DOI: 10.1016/S0197-4580(02)00065-9. URL: https://www.sciencedirect.com/science/article/pii/S0197458002000659 (visited on 11/01/2023).

[145] Arvid Carlsson. “NOBEL LECTURE: A Half-Century of Neurotransmitter Research: Impact on Neurology and Psychiatry”. In: *Bioscience Reports* 21.6 (2001). Publisher: Portland Press Ltd., pp. 691–710.

[146] H. Bernheimer et al. “Brain dopamine and the syndromes of Parkinson and Huntington Clinical, morphological and neurochemical correlations”. In: *Journal of the Neurological Sciences* 20.4 (Dec. 1973), pp. 415–455. ISSN: 0022-510X. DOI: 10.1016/0022-510X(73)90175-5. URL: https://www.sciencedirect.com/science/article/pii/0022510X73901755 (visited on 11/01/2023).

[147] L. Maroteaux, J. T. Campanelli, and R. H. Scheller. “Synuclein: a neuron-specific protein localized to the nucleus and presynaptic nerve terminal”. en. In: *Journal of Neuroscience* 8.8 (Aug. 1988). Publisher: Society for Neuroscience Section: Articles, pp. 2804–2815. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.08-08-02804.1988. URL: https://www.jneurosci.org/content/8/8/2804 (visited on 11/01/2023).

[148] Oleh Hornykiewicz. “Biochemical aspects of Parkinson’s disease”. en. In: *Neurology* 51.2 Suppl 2 (Aug. 1998). Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Articles, S2–S9. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.51.2_Suppl_2.S2. URL: https://n.neurology.org/content/51/2_Suppl_2/S2 (visited on 11/01/2023).

[149] A. J. Hughes et al. "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases." en. In: *Journal of Neurology, Neurosurgery & Psychiatry* 55.3 (Mar. 1992). Publisher: BMJ Publishing Group Ltd Section: Research Article, pp. 181–184. ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp.55.3.181. URL: https://jnnp.bmj.com/content/55/3/181 (visited on 11/01/2023).

[150] Douglas J. Gelb, Eugene Oliver, and Sid Gilman. "Diagnostic Criteria for Parkinson Disease". In: *Archives of Neurology* 56.1 (Jan. 1999), pp. 33–39. ISSN: 0003-9942. DOI: 10.1001/archneur.56.1.33. URL: https://doi.org/10.1001/archneur.56.1.33 (visited on 11/01/2023).

[151] Ronald B. Postuma et al. "MDS clinical diagnostic criteria for Parkinson's disease". en. In: *Movement Disorders* 30.12 (2015), pp. 1591–1601. ISSN: 1531-8257. DOI: 10.1002/mds.26424. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.26424 (visited on 11/01/2023).

[152] Johannes Levin et al. "The Differential Diagnosis and Treatment of Atypical Parkinsonism". In: *Deutsches Ärzteblatt International* 113.5 (Feb. 2016), pp. 61–69. ISSN: 1866-0452. DOI: 10.3238/arztebl.2016.0061. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782269/ (visited on 11/01/2023).

[153] Quin Massey et al. *Refining the Diagnostic Accuracy of Parkinsonian Disorders using Metaphenomic Annotation of the Clinicopathological Literature*. en. Pages: 2023.12.12.23299891. Dec. 2023. DOI: 10.1101/2023.12.12.23299891. URL: https://www.medrxiv.org/content/10.1101/2023.12.12.23299891v1 (visited on 04/29/2024).

[154] Fatemeh N. Emamzadeh and Andrei Surguchov. "Parkinson's Disease: Biomarkers, Treatment, and Risk Factors". English. In: *Frontiers in Neuroscience* 12 (Aug. 2018). Publisher: Frontiers. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00612. URL: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00612/full (visited on 03/27/2024).

[155] Stuart H Isaacson et al. "Clinical utility of DaTscan™ imaging in the evaluation of patients with parkinsonism: a US perspective". In: *Expert Review of Neurotherapeutics* 17.3 (2017). Publisher: Taylor & Francis, pp. 219–225.

[156] Stéphane Prange, Elise Metereau, and Stéphane Thobois. "Structural Imaging in Parkinson's Disease: New Developments". en. In: *Current Neurology and Neuroscience Reports* 19.8 (June 2019), p. 50. ISSN: 1534-6293. DOI: 10.1007/s11910-019-0964-5. URL: https://doi.org/10.1007/s11910-019-0964-5 (visited on 11/01/2023).

[157] Stéphane Lehericy et al. "The role of high-field magnetic resonance imaging in parkinsonian disorders: Pushing the boundaries forward". en. In: *Movement Disorders* 32.4 (2017), pp. 510–525. ISSN: 1531-8257. DOI: 10.1002/mds.26968. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.26968 (visited on 11/01/2023).

[158] Anyu Tao et al. "Accuracy of transcranial sonography of the substantia nigra for detection of parkinson's disease: a systematic review and meta-analysis". In: *Ultrasound in medicine & biology* 45.3 (2019). Publisher: Elsevier, pp. 628–641.

[159] D. E. Vaillancourt et al. "High-resolution diffusion tensor imaging in the substantia nigra of de novo Parkinson disease". en. In: *Neurology* 72.16 (Apr. 2009). Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Articles, pp. 1378–1384. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/01.wnl.0000340982.01727.6e. URL: https://n.neurology.org/content/72/16/1378 (visited on 11/01/2023).

[160] Yashar Zeighami et al. "Network structure of brain atrophy in de novo Parkinson's disease". In: *eLife* 4 (Sept. 2015). Ed. by David C Van Essen. Publisher: eLife Sciences Publications, Ltd, e08440. ISSN: 2050-084X. DOI: 10.7554/eLife.08440. URL: https://doi.org/10.7554/eLife.08440 (visited on 11/01/2023).

[161] Ronald B. Postuma et al. "Identifying prodromal Parkinson's disease: Pre-Motor disorders in Parkinson's disease". en. In: *Movement Disorders* 27.5 (2012). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.24996, pp. 617–626. ISSN: 1531-8257. DOI: 10.1002/mds.24996. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.24996 (visited on 03/20/2024).

[162] Anette Schrag et al. "Prediagnostic presentations of Parkinson's disease in primary care: a case-control study". English. In: *The Lancet Neurology* 14.1 (Jan. 2015). Publisher: Elsevier, pp. 57–64. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(14)70287-X. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(14)70287-X/fulltext (visited on 11/01/2023).

[163] Daniela Berg et al. "MDS research criteria for prodromal Parkinson's disease". eng. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 30.12 (Oct. 2015), pp. 1600–1611. ISSN: 1531-8257. DOI: 10.1002/mds.26431.

[164] Lazaros C. Triarhou. "Dopamine and Parkinson's Disease". en. In: *Madame Curie Bioscience Database [Internet]*. Landes Bioscience, 2013. URL: https://www.ncbi.nlm.nih.gov/books/NBK6271/ (visited on 07/22/2024).

[165] JULIAN M. FEARNLEY and ANDREW J. LEES. "AGEING AND PARKINSON'S DISEASE: SUBSTANTIA NIGRA REGIONAL SELECTIVITY". In: *Brain* 114.5 (Oct. 1991), pp. 2283–2301. ISSN: 0006-8950. DOI: 10.1093/brain/114.5.2283. URL: https://doi.org/10.1093/brain/114.5.2283 (visited on 11/01/2023).

[166] Sandrine Greffard et al. "Motor Score of the Unified Parkinson Disease Rating Scale as a Good Predictor of Lewy Body–Associated Neuronal Loss in the Substantia Nigra". In: *Archives of Neurology* 63.4 (Apr. 2006), pp. 584–588. ISSN: 0003-9942. DOI: 10.1001/archneur.63.4.584. URL: https://doi.org/10.1001/archneur.63.4.584 (visited on 11/01/2023).

[167] Shuang Yong Ma et al. "Correlation between neuromorphometry in the substantia nigra and clinical features in Parkinson's disease using disector counts". In: *Journal of the Neurological Sciences* 151.1 (Oct. 1997), pp. 83–87. ISSN: 0022-510X. DOI: 10.1016/S0022-510X(97)00100-7. URL: https://www.sciencedirect.com/science/article/pii/S0022510X97001007 (visited on 11/01/2023).

[168] Moran Artzi et al. "DaT-SPECT assessment depicts dopamine depletion among asymptomatic G2019S LRRK2 mutation carriers". en. In: *PLOS ONE* 12.4 (Apr. 2017). Publisher: Public Library of Science, e0175424. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0175424. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175424 (visited on 11/01/2023).

[169] Morten Gersel Stokholm et al. "Imaging dopamine function and microglia in asymptomatic LRRK2 mutation carriers". en. In: *Journal of Neurology* 267.8 (Aug. 2020), pp. 2296–2300. ISSN: 1432-1459. DOI: 10.1007/s00415-020-09830-3. URL: https://doi.org/10.1007/s00415-020-09830-3 (visited on 11/01/2023).

[170] Morten Gersel Stokholm et al. "Assessment of neuroinflammation in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a case-control study". English. In: *The Lancet Neurology* 16.10 (Oct. 2017). Publisher: Elsevier, pp. 789–796. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(17)30173-4. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(17)30173-4/fulltext (visited on 11/01/2023).

[171] Alex Iranzo et al. "Decreased striatal dopamine transporter uptake and substantia nigra hyperechogenicity as risk markers of synucleinopathy in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a prospective study". English. In: *The Lancet Neurology* 9.11 (Nov. 2010). Publisher: Elsevier, pp. 1070–1077. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(10)70216-7. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(10)70216-7/fulltext (visited on 03/20/2024).

[172] Xudong Li et al. "Transcranial sonography in idiopathic REM sleep behavior disorder and multiple system atrophy". In: *Psychiatry and Clinical Neurosciences* 71.4 (2017). Publisher: Wiley Online Library, pp. 238–246.

[173] Mariel Pullman et al. "Increased Substantia Nigra Echogenicity in LRRK2 Family Members Without Mutations". In: *Movement disorders : official journal of the Movement Disorder Society* 33.9 (Sept. 2018), pp. 1504–1505. ISSN: 0885-3185. DOI: 10.1002/mds.27443. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6346431/ (visited on 11/01/2023).

[174] Roy N. Alcalay et al. "Comparison of Parkinson Risk in Ashkenazi Jewish Patients With Gaucher Disease and GBA Heterozygotes". In: *JAMA Neurology* 71.6 (June 2014), pp. 752–757. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2014.313. URL: https://doi.org/10.1001/jamaneurol.2014.313 (visited on 11/06/2023).

[175] Daniel G. Healy et al. “Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson’s disease: a case-control study”. English. In: *The Lancet Neurology* 7.7 (July 2008). Publisher: Elsevier, pp. 583–590. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(08)70117-0. URL: https://www.thelancet.com/journals/lancet/article/PIIS1474-4422(08)70117-0/fulltext (visited on 11/06/2023).

[176] Volha Skrahina et al. “The Rostock International Parkinson’s Disease (ROPAD) Study: Protocol and Initial Findings”. en. In: *Movement Disorders* 36.4 (2021), pp. 1005–1010. ISSN: 1531-8257. DOI: 10.1002/mds.28416. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.28416 (visited on 11/06/2023).

[177] Laura Smith and Anthony H. V. Schapira. “GBA Variants and Parkinson Disease: Mechanisms and Treatments”. en. In: *Cells* 11.8 (Jan. 2022). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 1261. ISSN: 2073-4409. DOI: 10.3390/cells11081261. URL: https://www.mdpi.com/2073-4409/11/8/1261 (visited on 11/06/2023).

[178] Kenneth Marek et al. “The Parkinson’s progression markers initiative (PPMI) – establishing a PD biomarker cohort”. en. In: *Annals of Clinical and Translational Neurology* 5.12 (2018), pp. 1460–1477. ISSN: 2328-9503. DOI: 10.1002/acn3.644. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acn3.644 (visited on 11/06/2023).

[179] University Hospitals Plymouth NHS Trust. *About Us*. en. URL: https://www.plymouthhospitals.nhs.uk/about-us (visited on 01/03/2024).

[180] Mark Jenkinson et al. “FSL”. In: *NeuroImage*. 20 YEARS OF fMRI 62.2 (Aug. 2012), pp. 782–790. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.09.015. URL: https://www.sciencedirect.com/science/article/pii/S1053811911010603 (visited on 11/06/2023).

[181] Emma J. Burton et al. “Cerebral atrophy in Parkinson’s disease with and without dementia: a comparison with Alzheimer’s disease, dementia with Lewy bodies and controls”. In: *Brain* 127.4 (Apr. 2004), pp. 791–800. ISSN: 0006-8950. DOI: 10.1093/brain/awh088. URL: https://doi.org/10.1093/brain/awh088 (visited on 07/15/2024).

[182] M. T. M. Hu et al. “Correlating rates of cerebral atrophy in Parkinson’s disease with measures of cognitive decline”. en. In: *Journal of Neural Transmission* 108.5 (May 2001), pp. 571–580. ISSN: 1435-1463. DOI: 10.1007/s007020170057. URL: https://doi.org/10.1007/s007020170057 (visited on 07/15/2024).

[183] Bahar Say et al. “Evaluation of putamen area and cerebral peduncle with surrounding cistern in patients with Parkinson’s disease: is there a difference from controls in cranial MRI?” In: *Neurological Research* 46.3 (Mar. 2024), pp. 220–226. ISSN: 0161-6412. DOI: 10.1080/01616412.2023.2281088. URL: https://doi.org/10.1080/01616412.2023.2281088 (visited on 07/15/2024).

[184] M. M. Lewis et al. "Asymmetrical lateral ventricular enlargement in Parkinson's disease". en. In: *European Journal of Neurology* 16.4 (2009). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-1331.2008.02430.x, pp. 475–481. ISSN: 1468-1331. DOI: 10.1111/j.1468-1331.2008.02430.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-1331.2008.02430.x (visited on 07/15/2024).

[185] Liana G. Apostolova et al. "Hippocampal, caudate, and ventricular changes in Parkinson's disease with and without dementia". en. In: *Movement Disorders* 25.6 (2010), pp. 687–695. ISSN: 1531-8257. DOI: 10.1002/mds.22799. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.22799 (visited on 07/15/2024).

[186] Carlo Tessa et al. "Progression of brain atrophy in the early stages of Parkinson's disease: A longitudinal tensor-based morphometry study in de novo patients without cognitive impairment". en. In: *Human Brain Mapping* 35.8 (2014). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.22449, pp. 3932–3944. ISSN: 1097-0193. DOI: 10.1002/hbm.22449. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.22449 (visited on 11/06/2023).

[187] Leonor Correia Guedes et al. "Are genetic and idiopathic forms of Parkinson's disease the same disease?" en. In: *Journal of Neurochemistry* 152.5 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.14902, pp. 515–522. ISSN: 1471-4159. DOI: 10.1111/jnc.14902. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jnc.14902 (visited on 03/27/2024).

[188] Juan Alvarez-Linera. "3T MRI: Advances in brain imaging". In: *European Journal of Radiology*. Spanish Radiology - Spanish Researchers Open MR to Clinical Applications 67.3 (Sept. 2008), pp. 415–426. ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2008.02.045. URL: https://www.sciencedirect.com/science/article/pii/S0720048X08001708 (visited on 01/10/2024).

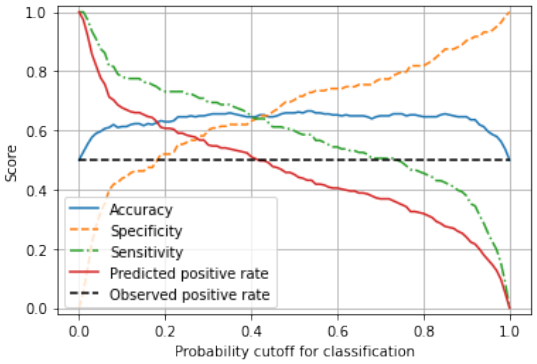
[189] N. Chow et al. "Comparing 3T and 1.5T MRI for Mapping Hippocampal Atrophy in the Alzheimer's Disease Neuroimaging Initiative". en. In: *American Journal of Neuroradiology* 36.4 (Apr. 2015). Publisher: American Journal of Neuroradiology Section: Brain, pp. 653–660. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A4228. URL: https://www.ajnr.org/content/36/4/653 (visited on 01/10/2024).

[190] Cheng-Hsin Cheng et al. "1.5T versus 3T MRI for targeting subthalamic nucleus for deep brain stimulation". In: *British Journal of Neurosurgery* 28.4 (Aug. 2014). Publisher: Taylor & Francis, pp. 467–470. ISSN: 0268-8697. DOI: 10.3109/02688697.2013.854312. URL: https://doi.org/10.3109/02688697.2013.854312 (visited on 01/10/2024).

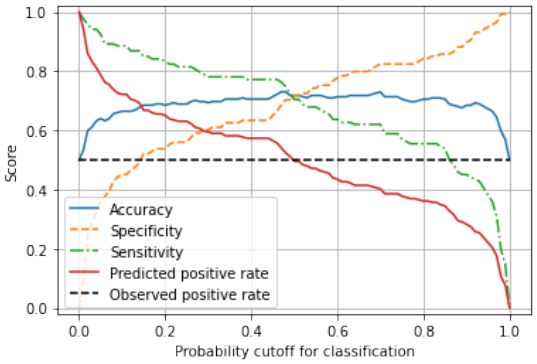
[191] *3D convolutional nueral networks · Issue #220 · shap/shap*. en. URL: https://github.com/shap/shap/issues/220 (visited on 01/12/2024).

[192] Sophie A. Martin et al. “Interpretable machine learning for dementia: A systematic review”. en. In: *Alzheimer's & Dementia* 19.5 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/alz.12948, pp. 2135–2149. ISSN: 1552-5279. DOI: 10.1002/alz.12948. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.12948 (visited on 03/28/2024).

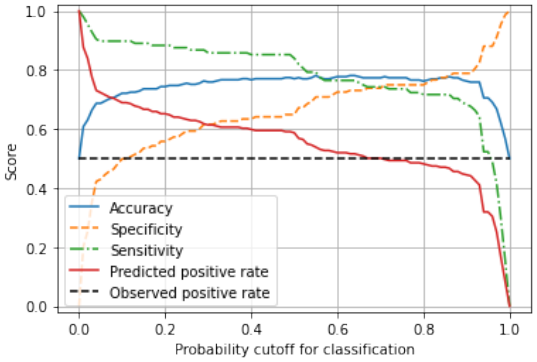
Appendix A: Threshold plots



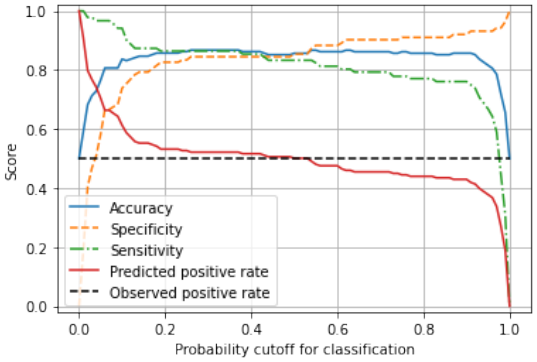
(a) IPD < 1 year vs controls. Threshold chosen: 0.43.



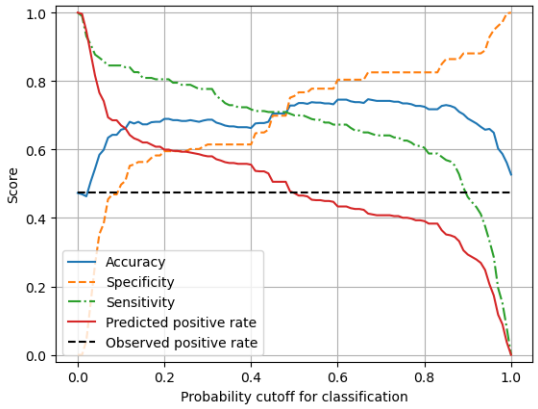
(b) IPD 1–2 years vs controls. Threshold chosen: 0.50.



(c) IPD 2–4 years vs controls. Threshold chosen: 0.70.

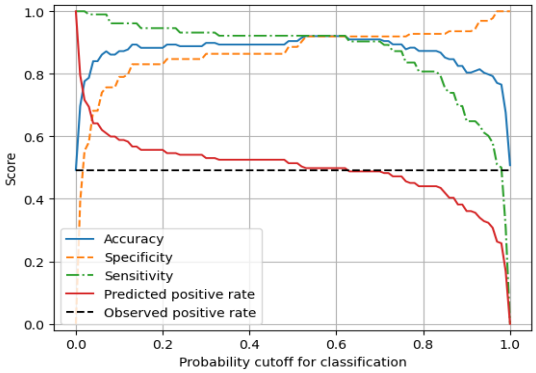


(d) IPD > 4 years vs controls. Threshold chosen: 0.42.

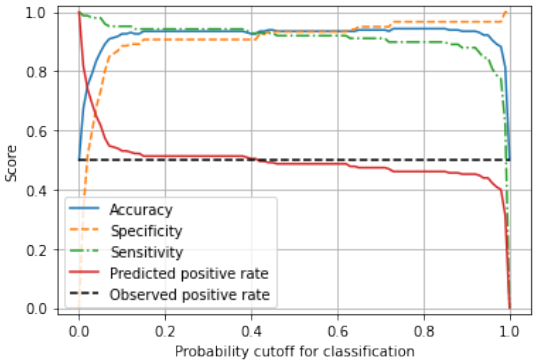


(e) All IPD vs controls. Threshold chosen: 0.48.

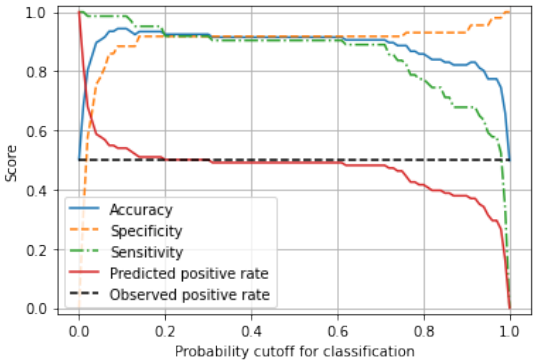
Figure A.1: Mean test performance metrics for PPMI idiopathic Parkinson's disease (IPD) models



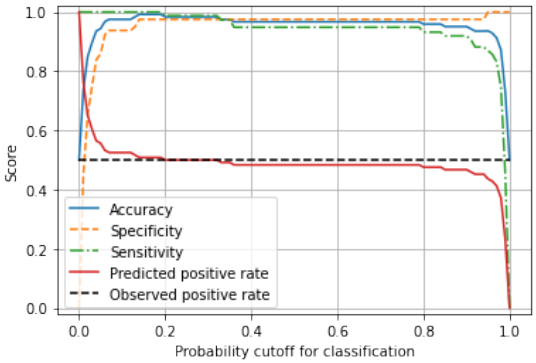
(a) *LRRK2* PD vs controls. Threshold chosen: 0.53.



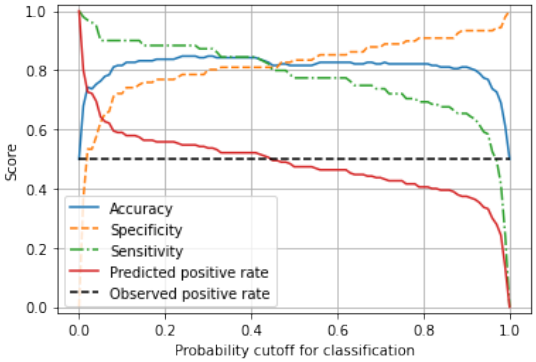
(b) *LRRK2* nPD vs controls. Threshold chosen: 0.45.



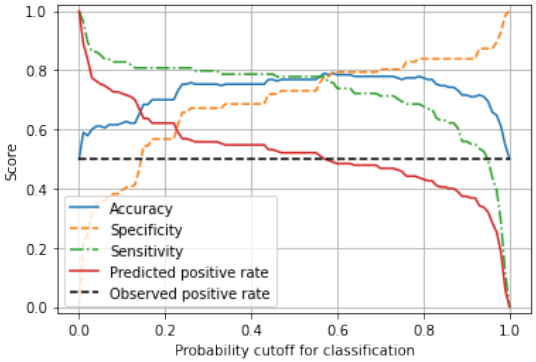
(c) *LRRK2* nPD $<$ average age vs controls. Threshold chosen: 0.30.



(d) *LRRK2* nPD $>$ average age vs controls. Threshold chosen: 0.35.

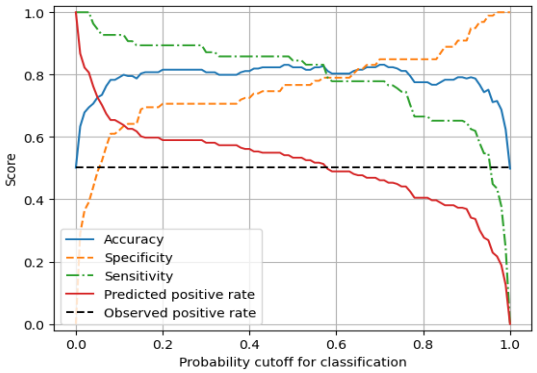


(e) *LRRK2* PD vs IPD. Threshold chosen: 0.44.

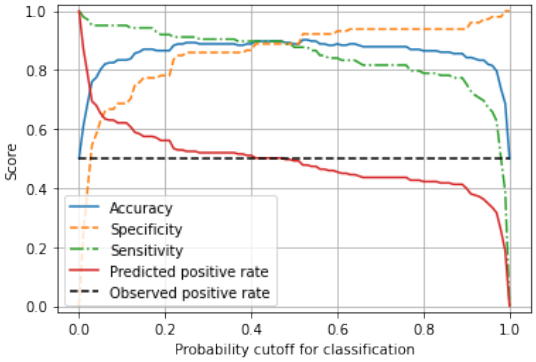


(f) *LRRK2* PD vs *LRRK2* nPD. Threshold chosen: 0.57.

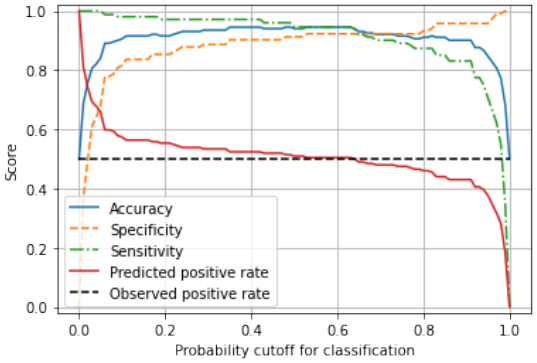
Figure A.2: Mean test performance metrics for PPMI *LRRK2* models.
IPD = Idiopathic Parkinson's disease
LRRK2 nPD = LRRK2 non Parkinson's disease manifesting carriers
LRRK2 PD = *LRRK2* Parkinson's disease manifesting carriers



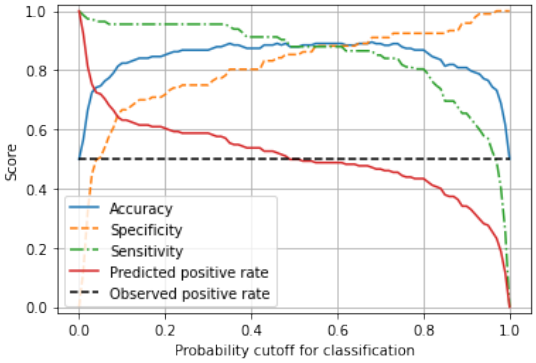
(a) *GBA* PD vs controls. Threshold chosen: 0.58.



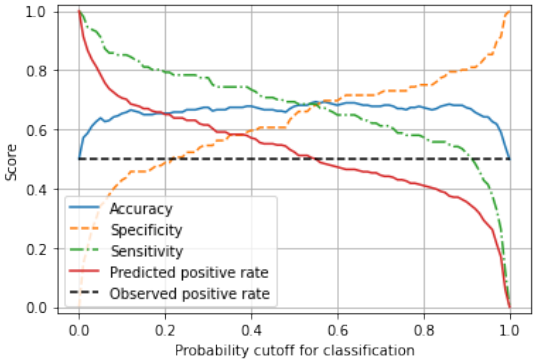
(b) *GBA* nPD vs controls. Threshold chosen: 0.50.



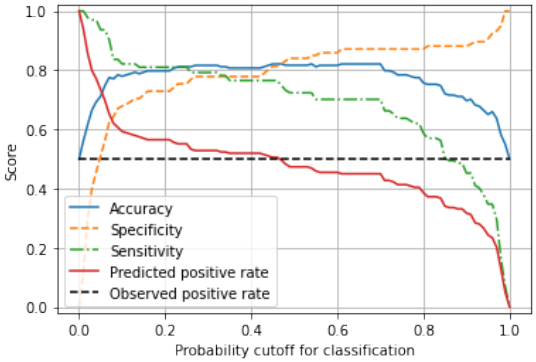
(c) GC *GBA* nPD vs controls. Threshold chosen: 0.65.



(d) *GBA* nPD > average age vs controls. Threshold chosen: 0.60.

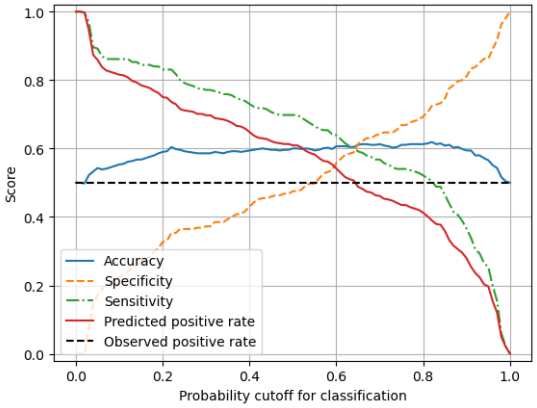


(e) *GBA* PD vs IPD. Threshold chosen: 0.55.

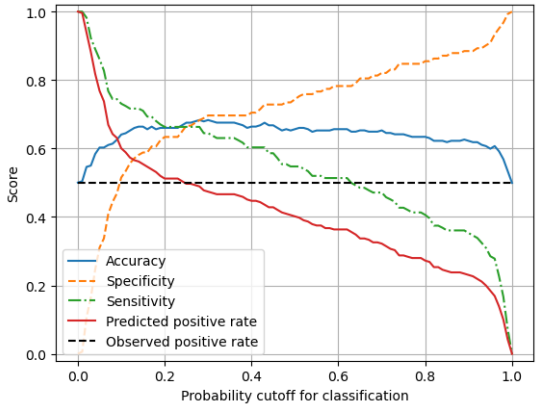


(f) *GBA* PD vs *GBA* nPD. Threshold chosen: 0.34.

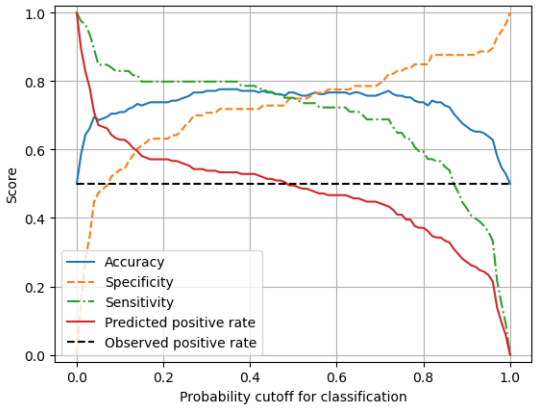
Figure A.3: Mean test performance metrics for PPMI *GBA* models.
IPD = Idiopathic Parkinson's disease
GBA nPD = *GBA* non Parkinson's disease manifesting carriers
GC *GBA* nPD = Gaucher causing *GBA* variants non-manifesting carriers
GBA PD = *GBA* Parkinson's disease manifesting carriers



(a) PD vs controls.
Threshold chosen: 0.65.

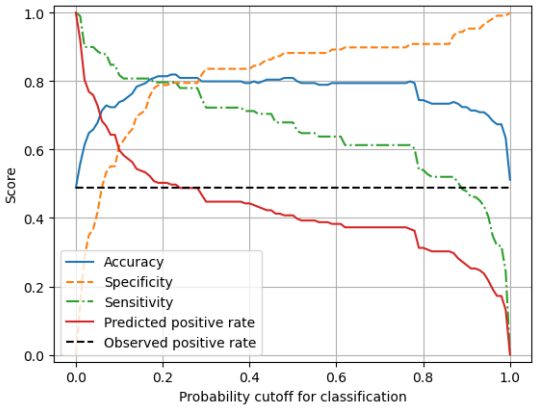


(b) PD post-diagnosis vs controls.
Threshold chosen: 0.25.

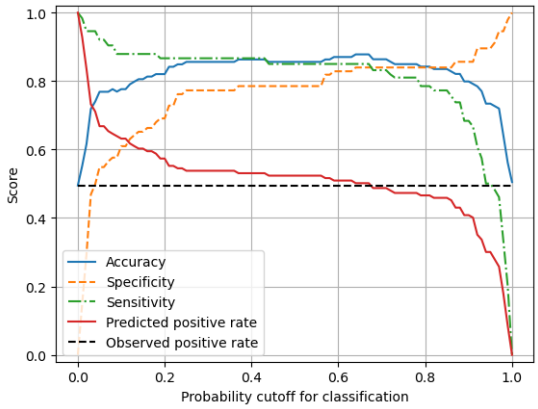


(c) PD pre-diagnosis vs controls.
Threshold chosen: 0.50.

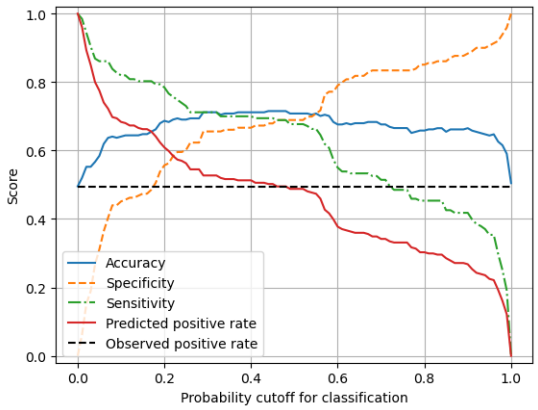
Figure A.4: Mean test performance metrics for UHPNT models.
PD = Parkinson's disease



(a) PD > 4 years vs controls. Threshold chosen: 0.23.



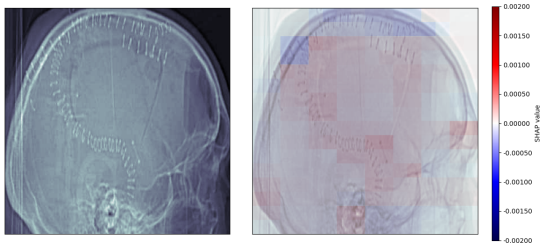
(b) PD 1-4 years vs controls. Threshold chosen: 0.68.



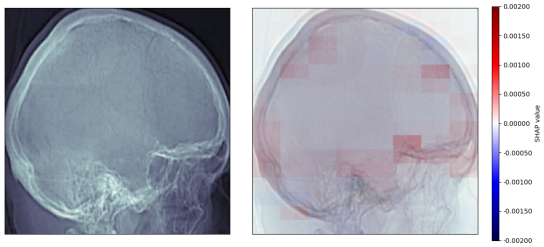
(c) PD < 1 year vs controls. Threshold chosen: 0.48.

Figure A.5: Mean test performance metrics for combined PPMI and UHPNT models.
PD = Parkinson's disease

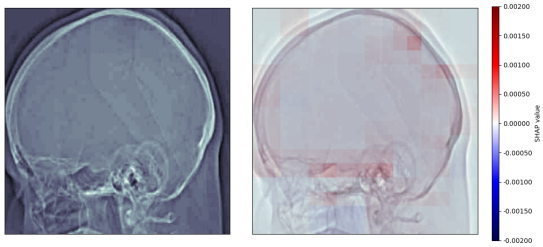
Appendix B: SHAP maps



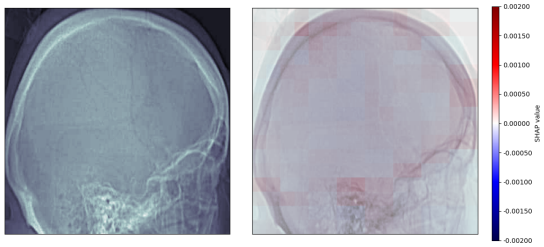
(a) False positive, as predicted by five models. The mean output probability of the image containing a clip is 0.46.



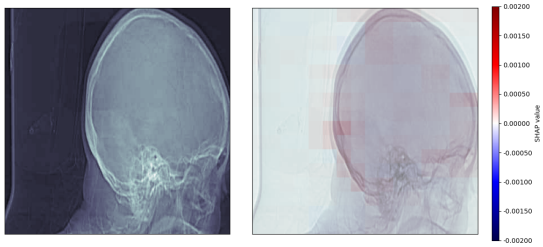
(b) False positive, as predicted by five models. The mean output probability of the image containing a clip is 0.45.



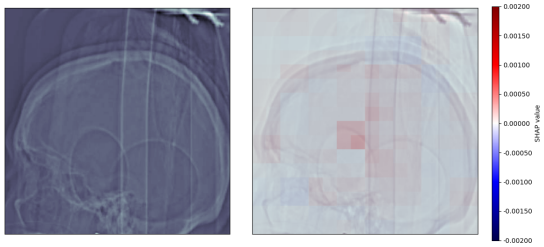
(c) False positive, as predicted by five models. The mean output probability of the image containing a clip is 0.27.



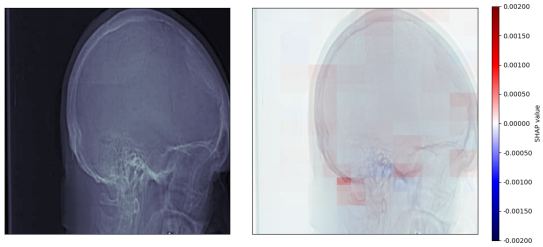
(d) False positive, as predicted by one model. The mean output probability of the image containing a clip is 0.10.



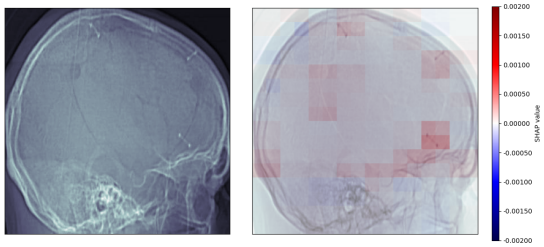
(e) False positive, as predicted by one model. The mean output probability of the image containing a clip is 0.17.



(f) False positive, as predicted by one model. The mean output probability of the image containing a clip is 0.10.

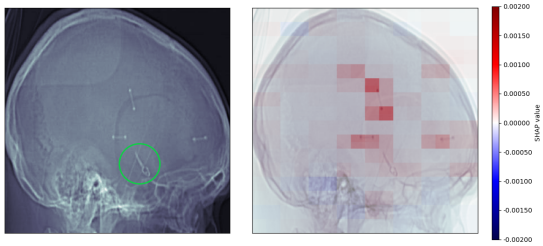


(g) False positive, as predicted by two models. The mean output probability of the image containing a clip is 0.15.

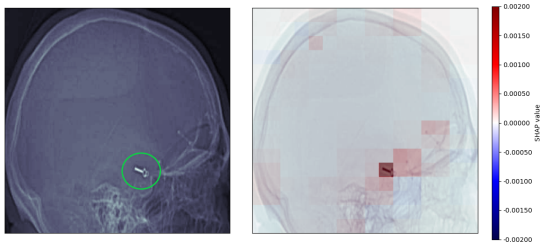


(h) False positive, as predicted by four models. The mean output probability of the image containing a clip is 0.30.

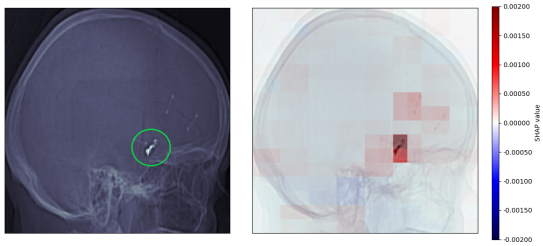
Figure B.1: Maps of average SHapley Additive exPlanation (SHAP) values for false positive aneurysm clip predictions. Any pixels highlighted in red have contributed to the prediction that an aneurysm clip is present; any pixels highlighted in blue have contributed to the prediction that no aneurysm clip is present.



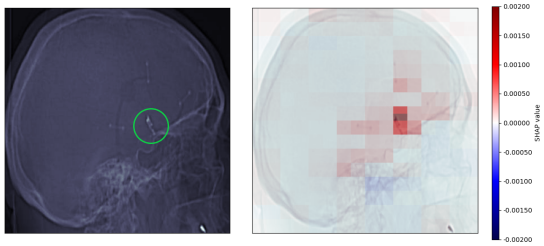
(a) True positive, as predicted by five models. The mean output probability of the image containing a clip is 1.00.



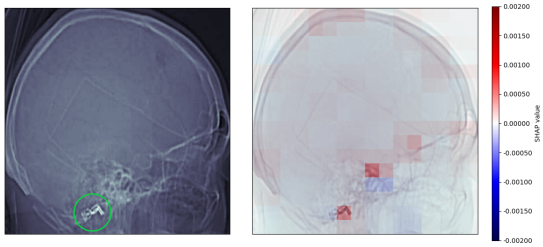
(b) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.98.



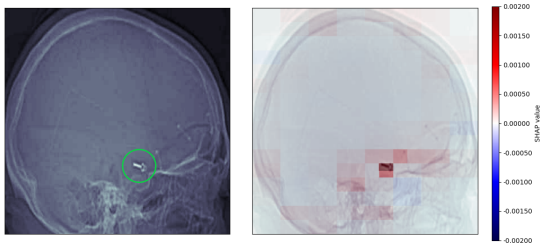
(c) True positive, as predicted by five models. The mean output probability of the image containing a clip is 1.00.



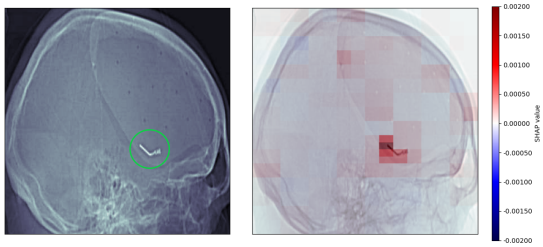
(d) True positive, as predicted by five models. The mean output probability of the image containing a clip is 1.00.



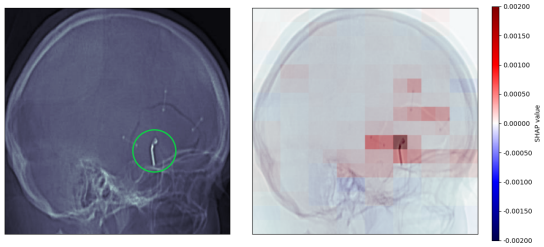
(e) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.84.



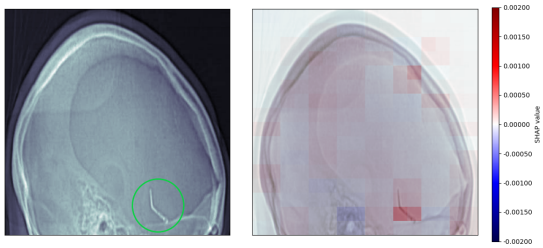
(f) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.99.



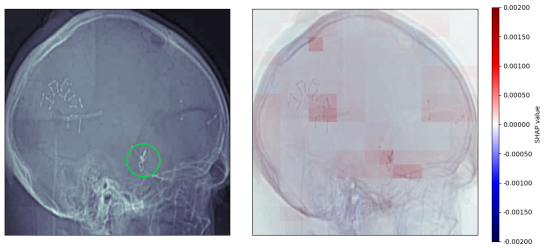
(g) True positive, as predicted by five models. The mean output probability of the image containing a clip is 1.00.



(h) True positive, as predicted by five models. The mean output probability of the image containing a clip is 1.00.

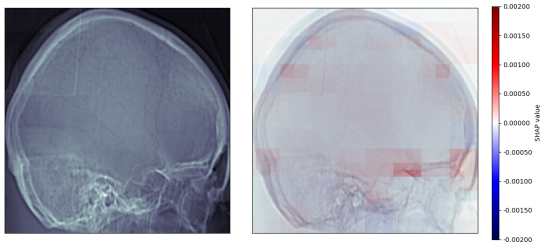


(i) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.75.

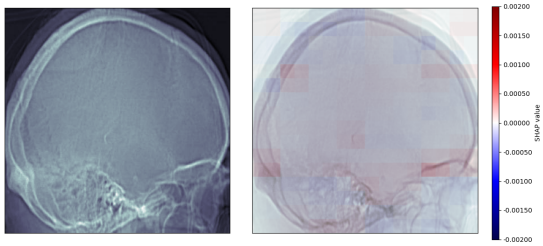


(j) True positive, as predicted by five models. The mean output probability of the image containing a clip is 0.99.

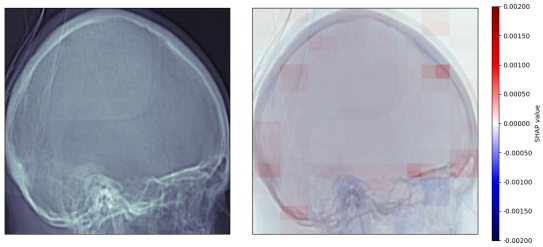
Figure B.2: Maps of average SHapley Additive exPlanation (SHAP) values for true positive aneurysm clip predictions. Aneurysm clips are circled in green. Any pixels highlighted in red have contributed to the prediction that an aneurysm clip is present; any pixels highlighted in blue have contributed to the prediction that no aneurysm clip is present.



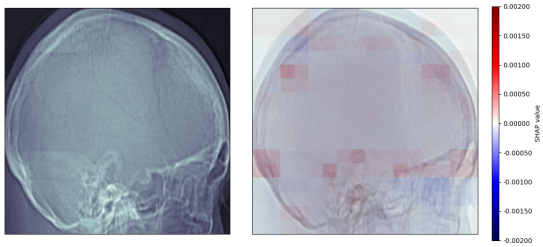
(a) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



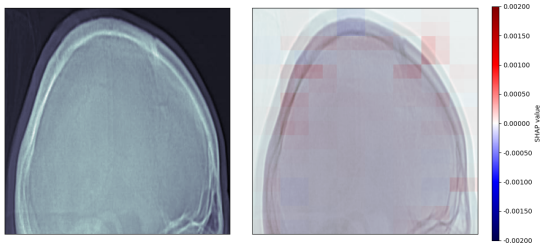
(b) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



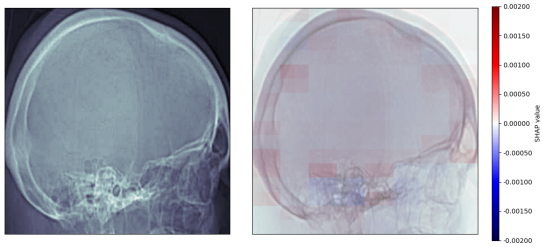
(c) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



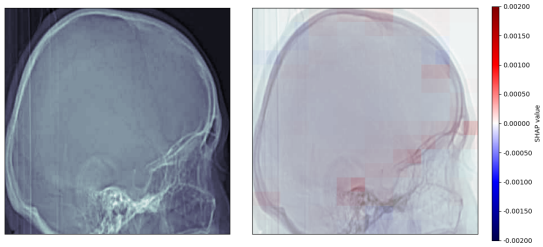
(d) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



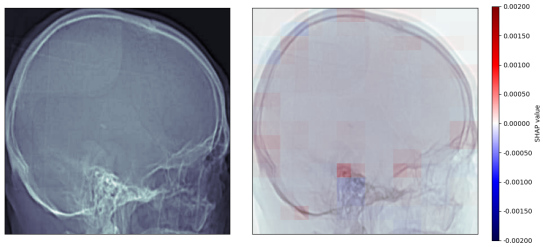
(e) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



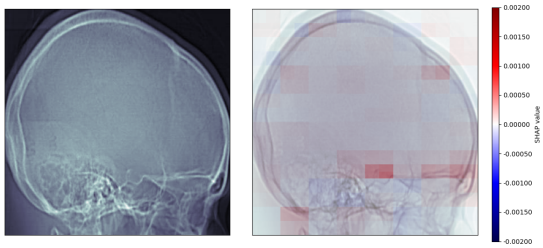
(f) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.01.



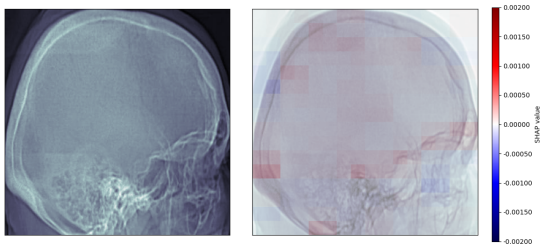
(g) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.



(h) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.

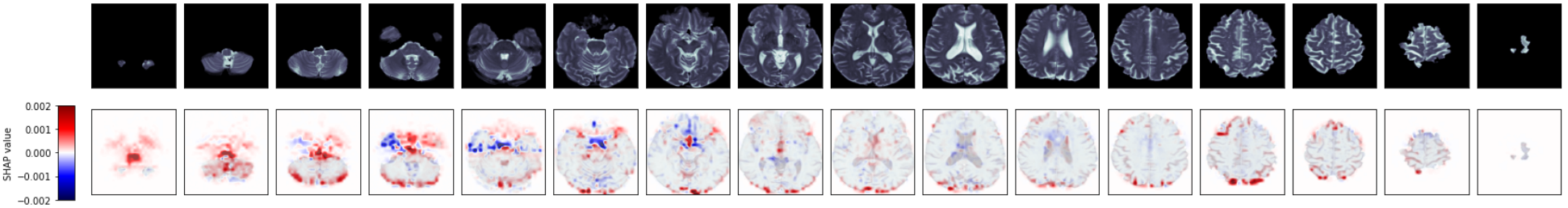


(i) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.01.

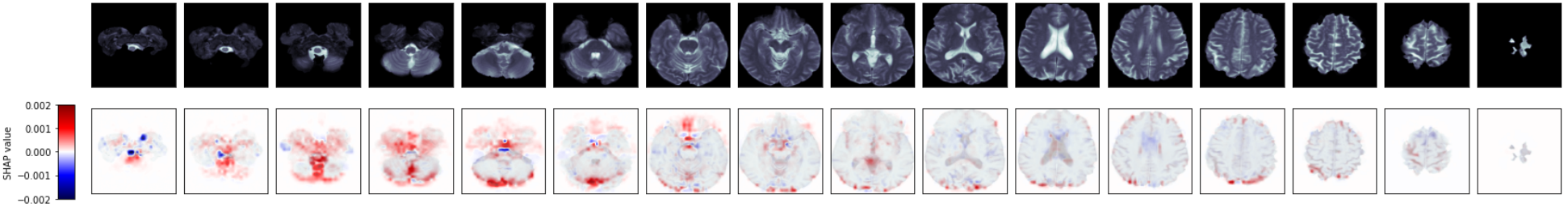


(j) True negative, as predicted by five models. The mean output probability of the image containing a clip is 0.00.

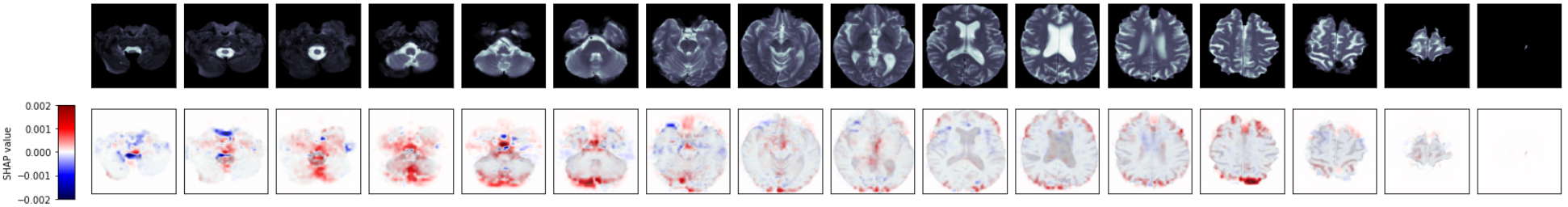
Figure B.3: Maps of average SHapley Additive exPlanation (SHAP) values for true negative aneurysm clip predictions. Any pixels highlighted in red have contributed to the prediction that an aneurysm clip is present; any pixels highlighted in blue have contributed to the prediction that no aneurysm clip is present.



(a) Predicted probability that image is from *LRRK2* carrier: 93.4%



(b) Predicted probability that image is from *LRRK2* carrier: 99.4%



(c) Predicted probability that image is from *LRRK2* carrier: 96.8%

Figure B.4: SHapley Additive exPlanation (SHAP) maps for *LRRK2* non PD manifesting carriers (*LRRK2* nPD) who subsequently developed motor Parkinson's disease, predicted by models trained to differentiate between *LRRK2* nPD and controls. Pixels highlighted in red have contributed positively to the prediction; pixels highlighted in blue have contributed negatively to the prediction.

Appendix C: PPMI data

Table C.1: *LRRK2* and *GBA* variants present in PPMI cohorts

Variant	Participant Count
LRRK2	
p.S1647T/p.S1647T / p.G2019S / p.M2397T/p.M2397T	37
p.S1647T / p.G2019S / p.M2397T	31
p.S1647T / p.G2019S / p.M2397T/p.M2397T	20
p.I723V / p.S1647T / p.G2019S / p.M2397T	12
p.N551K / p.R1398H / p.S1647T / p.G2019S / p.M2397T/p.M2397T	11
p.G2019S	11
p.S1647T / p.G2019S / p.N2081D / p.M2397T	6
p.N551K / p.S1647T/p.S1647T / p.G2019S / p.M2397T/p.M2397T	4
p.P1542S / p.S1647T / p.G2019S / p.M2397T/p.M2397T	3
p.R1441G / p.M1646T / p.M2397T	3
p.M1646T / p.S1647T / p.G2019S / p.M2397T/p.M2397T	2
p.N551K / p.R1398H / p.R1441G / p.M1646T / p.M2397T/p.M2397T	2
p.S1647T/p.S1647T / p.G2019S/p.G2019S / p.M2397T/p.M2397T	2
p.I723V / p.S1647T / p.G2019S / p.M2397T/p.M2397T	2
p.I723V / p.M1646T / p.S1647T / p.G2019S / p.M2397T/p.M2397T	2
p.R1441G	2
p.R1441G / p.M1646T / p.S1647T / p.M2397T/p.M2397T	2
p.R1514Q / p.S1647T / p.G2019S / p.M2397T/p.M2397T	2
p.N59K / p.S1647T / p.G2019S / p.M2397T	1
p.N551K / p.S1647T / p.G2019S / p.M2397T	1
p.N551K / p.R1398H / p.R1441C / p.S1647T / p.M2397T/p.M2397T	1
p.I723V / p.S1647T/p.S1647T / p.G2019S / p.M2397T/p.M2397T	1
p.V1330M / p.S1647T / p.G2019S / p.M2397T	1
GBA	
PD risk factor *GBA* variant	
p.E326K (p.E365K)	11
p.T369M (p.T408M)	8
p.T336S	1
"Mild" Gaucher causing *GBA* variant	
p.N370S (p.N409S)	113
p.N370S/p.N370S (p.N409S/p.N409S)	8
p.K13R	4
p.R44C (p.R83C)	1
p.A456P (p.A495P)	1
p.R39C (p.R78C)	1
p.E365K/p.N370S (p.E365K/p.N409S)	1
p.G115R/p.G193E (p.G154R/p.G232E)	1
p.I489L (p.I528L)	1
"Severe" Gaucher causing *GBA* variant	
p.L444P (p.483P)	5
p.IVS2+1G>A	1
p.T369M/p.R120W (p.T408M/p.R159W)	1
p.R502C	1

Table C.2: Counts of participants included in PPMI analysis and reasons for withdrawal

	Participant count			
	PD	Control	*LRRK2*	*GBA*
Total included in analysis	193	193	159	159
Withdrawal reason				
Subject withdrew consent	29	18	10	15
Death	15	2	3	10
Lost to follow up	12	13	4	2
Other	10	3	5	8
Informant/Caregiver decision	3	2	2	4
Investigator decision	3	1	0	0
Decline in health	2	1	0	0
Transportation/Travel issues	2	0	0	1
Family, care-partner, or social issues	1	0	2	0
Adverse Event	1	0	0	0
Burden of study procedures (other than travel)	0	1	0	0
Sponsor decision	0	0	1	1
Institutionalised	0	0	0	1
Total withdrawals	78	41	27	42
		188		