UNIVERSIDAD DE CÓRDOBA

DOCTORADO INTERNACIONAL
Programa de Doctorado: «Computación Avanzada, Energía y Plasmas»

TÉCNICAS DE APRENDIZAJE PROFUNDO PARA DATOS ESPACIALES ORDINALES Y APLICACIONES EN IMAGEN MÉDICA

Doctorando:

JAVIER BARBERO GÓMEZ

Directores:

CÉSAR HERVÁS MARTÍNEZ

PEDRO ANTONIO GUTIÉRREZ PEÑA

Córdoba, enero de 2024

TITULO: *Deep learning techniques for ordinal spatial data and applications in medical imaging*

AUTOR: *Javier Barbero Gómez*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

UNIVERSITY OF CÓRDOBA

INTERNATIONAL DOCTORATE

PhD Programme: 'Advanced Computing, Energy and Plasmas'

DEEP LEARNING TECHNIQUES FOR ORDINAL SPATIAL DATA AND APPLICATIONS IN MEDICAL IMAGING

Author:

JAVIER BARBERO GÓMEZ

Supervisors:

CÉSAR HERVÁS MARTÍNEZ

PEDRO ANTONIO GUTIÉRREZ PEÑA

Córdoba, January 2024

Contact: jbarbero at uco.es
Learning and Artificial Neural Networks (AYRNA) research group
Department of Computer Science and Numerical Analysis
University of Córdoba
Córdoba (Spain)
http://www.uco.es/ayrna/

SUPERVISORS:
César Hervás Martínez
Pedro Antonio Gutiérrez Peña

DECLARACIÓN

La memoria titulada «Técnicas de aprendizaje profundo para datos espaciales ordinales y aplicaciones en imagen médica», que presenta D. Javier Barbero Gómez para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado «Computación Avanzada, Energía y Plasmas» de la Universidad de Córdoba bajo la dirección de los Catedráticos de Universidad D. César Hervás Martínez y D. Pedro Antonio Gutiérrez Peña.

El doctorando D. Javier Barbero Gómez y los directores de la tesis D. César Hervás Martínez y D. Pedro Antonio Gutiérrez Peña garantizamos al firmar esta Tesis Doctoral que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la Tesis y, hasta donde nuestro conocimiento alcanza, en la realización del trabajo se han respetado los derechos de otros autores a ser citados cuando se han utilizado sus resultados o publicaciones.

Córdoba, 16 de enero de 2024

El doctorando:

Firmado por BARBERO GOMEZ JAVIER - ***2686** el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios

Fdo.: Javier Barbero Gómez

Los directores de tesis:

Firmado por HERVAS MARTINEZ CESAR - ***2848** el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios

Fdo.: César Hervás Martínez

Firmado por GUTIERREZ PEÑA PEDRO ANTONIO - 45735457A el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios

Fdo.: Pedro Antonio Gutiérrez Peña

INFORME RAZONADO DE LAS/LOS DIRECTORAS/ES DE LA TESIS

Este documento se presentará junto con el depósito de la tesis en https://moodle.uco.es/ctp3/

DOCTORANDA/O

JAVIER BARBERO GÓMEZ

TÍTULO DE LA TESIS:

Técnicas de aprendizaje profundo para datos espaciales ordinales y aplicaciones en imagen médica

INFORME RAZONADO DE LAS/LOS DIRECTORAS/ES DE LA TESIS
(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma)

En su Tesis, D. Javier Barbero Gómez realiza una exploración exhaustiva de la intersección del aprendizaje profundo y la regresión ordinal, centrándose en el desafío específico del diagnóstico de imágenes médicas. El candidato ha demostrado una sólida comprensión de las complejidades inherentes a las tareas ordinales y ha realizado contribuciones significativas.

Inicialmente presenta una nueva arquitectura de redes neuronales convolucionales diseñada específicamente para la regresión ordinal. El uso de un esquema de códigos de salida de corrección de errores para la asignación de etiquetas muestra la innovación y logra un rendimiento superior en diversas tareas. La flexibilidad y eficiencia de este enfoque, especialmente en el mantenimiento de las métricas de clasificación tradicionales al tiempo que mejora el rendimiento ordinal, lo convierten en una valiosa contribución a este campo.

A continuación aplica con éxito la metodología propuesta a una aplicación médica práctica, el diagnóstico de la enfermedad de Parkinson, superando los retos relacionados con los escáneres cerebrales volumétricos. La introducción de una arquitectura capaz de procesar imágenes 3D junto con un algoritmo de aumento de datos demuestra la adaptabilidad a diversos tipos de entrada y la resistencia contra los desafíos de desequilibrio de clases. Los avances tangibles en la evaluación de la actividad cerebral dopaminérgica subrayan la utilidad práctica de las técnicas desarrolladas.

Finalmente aborda el aspecto crítico de la explicabilidad de los modelos ordinales. La rigurosa validación de los métodos de explicación existentes y la introducción de dos técnicas novedosas aportan matices a los procesos de toma de decisiones de estos modelos.

En general, la tesis no sólo avanza en la comprensión teórica de los clasificadores de regresión ordinal dentro del aprendizaje profundo, sino que también proporciona metodologías prácticas que mejoran el rendimiento, la adaptabilidad y la explicabilidad de los modelos de redes neuronales convolucionales en el dominio ordinal. La exitosa aplicación a retos médicos reales refuerza el impacto y la relevancia de la tesis.

A medida que avanza la investigación, el candidato ha identificado eficazmente futuras vías de exploración, incluyendo el perfeccionamiento de las metodologías, la ampliación de las aplicaciones, y una mayor investigación de la interpretabilidad de los modelos de redes convolucionales ordinales. La tesis sirve como una base sólida para futuras investigaciones en esta área crítica y en evolución de las aplicaciones de aprendizaje profundo.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, a 16 de enero de 2024

Las/los directoras/es

Firmado por HERVAS MARTINEZ CESAR - ***2848** el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios

Firmado por GUTIERREZ PEÑA PEDRO ANTONIO - 45735457A el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios

Fdo.: César Hervás Martínez

Fdo.: Pedro Antonio Gutiérrez Peña

MENCIÓN DE DOCTORADO INTERNACIONAL

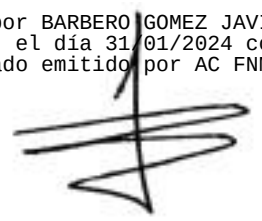
Esta tesis cumple los criterios para la obtención de la mención «Doctorado Internacional» concedida por la Universidad de Córdoba. Para ello se presentan los siguientes requisitos:

1. Estancia predoctoral realizada en otros países europeos:
   - Centro de investigación **INESC-TEC, Universidade do Porto, Oporto, Portugal** durante 3 meses desde el 5 de septiembre hasta el 5 de diciembre de 2022. Tutelado por el **Dr. Jaime dos Santos Cardoso**, *Full Profesor* en el Departamento de Ingeniería Eléctrica e Informática, Facultad de Ingeniería, Universidade do Porto (Portugal).
2. Esta tesis está avalada por los siguientes informes de idoneidad realizados por doctores de otros centros de investigación europeos:
   - El **Dr. Riccardo Rosati**, *Post-doctoral research fellow* en el Departamento de Ingeniería de la Información, Università Politecnica delle Marche (Italia).
   - El **Dr. Wilson Silva**, *Assistant Professor* en el Departamento de Ciencias de la Información y la Computación, Universiteit Utrecht (Países Bajos).
3. La defensa de tesis y el texto se han realizado parcialmente en dos idiomas europeos: español e inglés. La tesis está escrita al completo en inglés y cuenta con un resumen en español.
4. Entre los miembros del tribunal se encuentra un doctor procedente de un centro de educación superior europeo, tratándose del **Dr. Luca Romeo**, *Tenure Track Assistant Professor* en el Departamento de Economía y Derecho, Università di Macerata (Italia).

Córdoba, 16 de enero de 2024

El Doctorando:

Firmado por BARBERO GOMEZ JAVIER - ***2686** el día 31/01/2024 con un certificado emitido por AC FNMT Usuarios



Fdo.: Javier Barbero Gómez

A mi familia, por su apoyo incondicional y las tardes de Rummikub.

A mis amigos, por sus ánimos, algunos llantos y muchas risas.

A mis compañeros de laboratorio, por su trabajo incansable y su humor interminable.

A mis directores, por su paciencia, su fe en mi y sus batallitas.

A todos los que me habéis acompañado y he acompañado en este viaje.

Gracias

ABSTRACT

This thesis navigates the intersection of deep learning and ordinal regression, aiming to address the challenges inherent in ordinal tasks, particularly within medical image diagnosis. The primary contributions unfold across three interconnected chapters, each designed to tackle a distinct facet of the overarching problem.

First, a new breed of Convolutional Neural Network (CNN) architectures tailored for ordinal regression is introduced. The superior performance of the proposed Ordinal Binary Decomposition (OBD) model, coupled with an Error Correcting Output Codes (ECOC) scheme for label assignment, is demonstrated across diverse tasks. Notably, it achieves superior ordinal performance metrics without compromising traditional classification ones, offering a flexible and efficient solution to the challenges of ordinal regression problems.

The next chapter extends this methodology to a real medical application: the diagnosing of Parkinson's disease. Faced with the challenges of volumetric brain scans, a native 3D CNN architecture is introduced, along with an innovative data augmentation algorithm (OGO-SP-β) that exploits ordinal information. This methodology showcases significant advancements in assessing dopaminergic brain activity, demonstrating adaptability to diverse input types and resilience against class imbalance challenges.

Finally, the next chapter delves into the often-neglected realm of interpretability in ordinal CNN models. Existing explanation methods are rigorously validated, and two novel techniques, GradOBD-CAM and OIBA, are introduced to shed light on the decision-making processes of these models. GradOBD-CAM outperforms existing methods, providing nuanced insights into feature importance in the context of ordinal regression.

As a whole, this thesis contributes to advancing the understanding and application of ordinal regression classifiers within deep learning. The developed methodologies, spanning novel architectures, data augmentation techniques, and explanation methods, generally enhance the performance, adaptability, and explainability of CNN models in the ordinal domain. The successful application of these methodologies to real medical challenges underscores their practical utility, with implications extending beyond the medical realm.

As the research unfolds, this thesis lays the groundwork for future explorations, suggesting avenues for refining methodologies, expanding applications, and delving deeper into the interpretability of ordinal CNN models. In conclusion, this thesis provides a comprehensive and nuanced exploration of ordinal regression challenges, offering tangible solutions and insights that contribute to the evolving landscape of deep learning applications.

RESUMEN

Esta tesis navega por la intersección del aprendizaje profundo y la regresión ordinal, con el objetivo de abordar los retos inherentes a las tareas ordinales, en particular en el diagnóstico de imágenes médicas. Las principales contribuciones se desarrollan a lo largo de tres capítulos interconectados, cada uno diseñado para abordar una faceta distinta del problema general.

En primer lugar, se presenta una nueva variedad de arquitecturas de Redes Neuronales Convolucionales (CNN) adaptadas a la regresión ordinal. El rendimiento superior del modelo de Descomposición Ordinal Binaria (OBD) propuesto, junto con un esquema de asignación de etiqueta basado en Códigos de Salida de Corrección de Errores (ECOC), se demuestra a través de diversas tareas. En particular, logra métricas de rendimiento ordinal superiores sin comprometer las métricas de clasificación tradicionales, ofreciendo una solución flexible y eficiente a los retos de los problemas de regresión ordinal.

El siguiente capítulo amplía esta metodología a una aplicación médica real: el diagnóstico de la enfermedad de Parkinson. Enfrentados a los retos de los escáneres cerebrales volumétricos, se introduce una arquitectura de CNN nativa 3D así como un innovador algoritmo de aumento de datos (OGO-SP-β) que explota la información ordinal. Esta metodología muestra avances significativos en la evaluación de la actividad cerebral dopaminérgica, demostrando su adaptabilidad a diversos tipos de datos de entrada y su resistencia frente a los desafíos del desequilibrio de clases.

Por último, el siguiente capítulo se adentra en el ámbito, a menudo descuidado, de la explicabilidad en los modelos CNN ordinales. Se validan rigurosamente los métodos de explicación existentes y se introducen dos técnicas novedosas, GradOBD-CAM y OIBA, para arrojar luz sobre los procesos de toma de decisiones de estos modelos. GradOBD-CAM supera a los métodos existentes, proporcionando información matizada sobre la importancia de las características en el contexto de la regresión ordinal.

En conjunto, esta tesis contribuye a avanzar en la comprensión y aplicación de modelos de regresión ordinal dentro del aprendizaje profundo. Las metodologías desarrolladas, que abarcan arquitecturas novedosas, técnicas de aumento de datos y métodos de explicación, mejoran generalmente el rendimiento, la adaptabilidad y la interpretabilidad de los modelos CNN en el dominio ordinal. La aplicación con éxito de estas metodologías a retos médicos reales subraya su utilidad práctica, con implicaciones que se extienden más allá del ámbito médico.

A medida que se desarrolla la investigación, esta tesis sienta las bases para futuras exploraciones, sugiriendo vías para refinar las metodologías, ampliar las aplicaciones y profundizar en la interpretabilidad de los modelos CNN ordinales. En conclusión, esta tesis proporciona una exploración completa y matizada de los retos de regresión ordinal, ofreciendo soluciones tangibles y conocimientos que contribuyen al panorama en evolución de las aplicaciones de aprendizaje profundo.

PUBLICATIONS

Some of the contributions, ideas and figures of this thesis have appeared previously in the following publications:

- INTERNATIONAL JOURNAL PAPERS:
  - **Javier Barbero-Gómez**, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Error-Correcting Output Codes in the Framework of Deep Ordinal Classification'. In: *Neural Processing Letters* (May 2022). ISSN: 1573-773X. DOI: 10.1007/s11063-022-10824-7

    JCR (2022): 3.1. Ranking position in Computer Science, AI: 85/145 (Q3)
  - **Javier Barbero-Gómez**, Pedro-Antonio Gutiérrez, Víctor-Manuel Vargas, Juan-Antonio Vallejo-Casas and César Hervás-Martínez. 'An Ordinal CNN Approach for the Assessment of Neurological Damage in Parkinson's Disease Patients'. In: *Expert Systems with Applications* 182 (Nov. 2021), p. 115271. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.115271

    JCR (2021): 8.655. Ranking position in Computer Science, AI: 21/145 (Q1)
- INTERNATIONAL CONFERENCE PUBLICATIONS:
  - **Javier Barbero-Gómez**, Ricardo Cruz, Jaime S. Cardoso, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Evaluating the Performance of Explanation Methods on Ordinal Regression CNN Models'. In: *International Work-Conference on Artificial Neural Networks (IWANN 2023)*. June 2023

During the development of this thesis, the following related collaborations with other authors were also published:

- Víctor Manuel Vargas, Pedro Antonio Gutiérrez, **Javier Barbero-Gómez** and César Hervás-Martínez. 'Activation Functions for Convolutional Neural Networks: Proposals and Experimental Study'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021), pp. 1–11. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2021.3105444

  JCR (2021): 14.255. Ranking position in Computer Science, AI: 6/145 (Q1D1)
- Víctor Manuel Vargas, Pedro Antonio Gutiérrez, **Javier Barbero-Gómez** and César Hervás-Martínez. 'Soft Labelling Based on Triangular Distributions for Ordinal Classification'. In: *Information Fusion* 93 (May 2023), pp. 258–267. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.01.003

  JCR (2022): 18.6. Ranking position in Computer Science, AI: 4/145 (Q1D1)
- David Guijo-Rubio, Víctor M. Vargas, **Javier Barbero-Gómez**, Jose V. Die and Pablo González-Moreno. 'Hackathon in Teaching: Applying Machine Learning to Life Sciences Tasks'. In: *International Joint Conference 15th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2022) 13th International Conference on EUropean Transnational Education (ICEUTE 2022)*. Lecture Notes in Networks and Systems. Springer Nature Switzerland, 2023, pp. 236–246. ISBN: 978-3-031-18409-3. DOI: 10.1007/978-3-031-18409-3_23

FUNDING

This PhD Thesis has been funded by:

- Projects TIN2017-85887-C2-1-P and PID2020-115454GB-C22 of the Spanish Ministry of Science, Innovation and Universities (MICINN), with FEDER funds.
- The hiring programme for the training of doctors (FPI) of the MICINN, reference PRE2018-085659, associated with the former two projects.
- Project TIN2017-90567-REDT of the MICINN, with FEDER funds.
- Project UCO-1261651 of the Ministry of Economic Transformation, Industry, Knowledge and University of the Andalusian Government, with FEDER funds.
- Project PS-2020-780 of the Ministry of Health and Family of the Andalusian Government.
- Proyecto PY20_00074 of the FEDER Operational Programme 2014-2020.

FINANCIACIÓN

Esta Tesis Doctoral ha sido financiada por:

- Los Proyectos TIN2017-85887-C2-1-P y PID2020-115454GB-C22 del Ministerio de Ciencia, Innovación y Universidades (MICINN), con fondos FEDER.
- El programa de ayudas para contratos predoctorales para la formación de doctores (FPI) del MICINN, referencia PRE2018-085659, asociado a los proyectos anteriores.
- El proyecto TIN2017-90567-REDT del MICINN, con fondos FEDER.
- El Proyecto UCO-1261651 de la Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía, con fondos FEDER.
- El Proyecto PS-2020-780 de la Consejería de Salud y Familia de la Junta de Andalucía.
- El Proyecto PY20_00074 del Programa Operativo FEDER 2014-2020.

Junta de Andalucía
Consejería de Transformación Económica, Industria, Conocimiento y Universidades

UNIVERSIDAD
DE
CÓRDOBA

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the ROC Curve
BI-RADS	Breast Imaging Reporting and Data System
CAD	computer-aided diagnosis
CAM	Class Activation Map
CCR	Correct Classification Rate
CLM	Cumulative Link Model
CNN	Convolutional Neural Network
CT	Computed Tomography
DDSM	Digital Database for Screening Mammography

DR	Diabetic Retinopathy
ECOC	Error Correcting Output Codes
ELU	Exponential Linear Unit
GAN	Generative Adversarial Network
GAP	global average pooling
GMP	global maximum pooling
IBA	Information Bottleneck for Attribution
LReLU	Leaky Rectified Linear Unit
MAE	Mean Absolute Error
ML	Machine Learning
MLP	multi-layer perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
OBD	Ordinal Binary Decomposition
OGO-SP	Ordinal Graph Oversampling via Shortest Paths
OvR	One vs. Rest
PD	Parkinson's disease
PET	positron emission tomography
POM	Proportional Odds Model
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ReLU	Rectified Linear Unit
RoI	Region of Interest
SGD	stochastic gradient descent
SMOTE	Synthetic Minority Oversampling Technique
SPECT	single-photon emission computed tomography
SVM	Support Vector Machine
SVOREX	Support Vector for Ordinal Regression with Explicit constraints
SVORIM	Support Vector for Ordinal Regression with Implicit constraints
XAI	Explainable AI

Part I

INTRODUCTION

1 INTRODUCTION

1.1 MACHINE LEARNING

The field of Machine Learning (ML) is difficult to categorize. It lies on the intersection of Artificial Intelligence (AI) and statistics, in some cases even philosophy, biology and cognitive science [65]. Its aim is the creation of computational models that learn to perform tasks from experience instead of being explicitly programmed [82].

Even though its humble beginning in the 1950s, ML has recently revolutionized a lot of fields, specially since several milestone achievements were accomplished during the 2010s: competitive strategy game playing, image and video object recognition, protein structure prediction, speech recognition, language translation...

The now seemingly unlimited potential impact of ML on solving complex problems relies on its ability to make data-driven decisions instead of depending on sometimes unreliable heuristics. Generally speaking, ML models learn from a set of examples called a *dataset*, and each of these examples is comprised of a set of characterising *features*.

The types of tasks that can be solved using ML are varied and ever-expanding. These can be divided up into three main types: supervised learning, unsupervised learning and reinforcement learning.

1.1.1 *Supervised learning*

As an example, suppose that we want to predict the final yield of a certain crop plantation based on the total amount of rainfall that it received during the growing period. This is a prime example of a task that can be tackled using supervised learning.

In the *supervised learning* paradigm the examples in the dataset are labelled, meaning that an external source (generally an expert in the problem domain) has determined the value (namely, the *ground truth*) of a special variable that is commonly referred to as the *response* or *dependent variable*. In our example, a group of farmers could register the total amount of rainfall (learning examples) and then report the final yield of that season (ground truth value of the response variable).

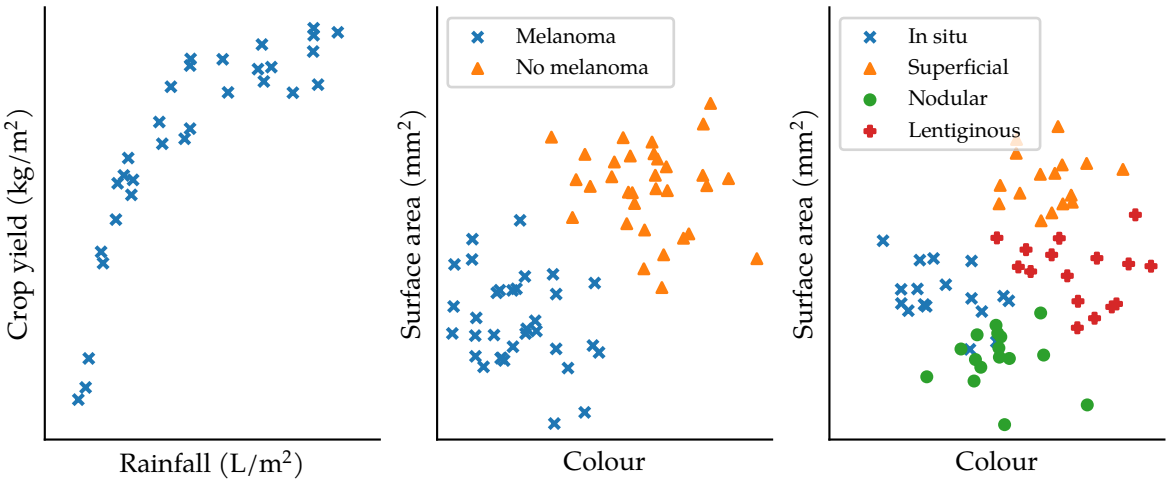
The goal will be to build a model that is able to predict as accurately as possible the value of the response variable using only the input features of an arbitrary example (present or not in the original dataset). The process of using the data for building and optimizing the model that achieves this goal is called *training*, and the data used for this process is the *train dataset*. After this process is completed, we need to check that the model is able to *generalize* well, i. e. to correctly classify examples which it has not been trained on. For this, we use a separate set of data called the *test dataset*. Coming back to our example, we hope to use the data gathered by farmers over different seasons so that we may be able to predict the final yield beforehand in the future.

In mathematical terms, given a dataset $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$ where each $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a training example drawn from a specific distribution $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$, and each $y_i \in \mathcal{Y}$ corresponds to the ground truth value for the response variable, the prediction model

is a function $f : \mathcal{X} \to \mathcal{Y}$ mapping input examples to a predicted response $\hat{y} = f(\mathbf{x})$. The objective of the training process is to obtain a prediction model f such that it aligns with the real distribution of the response variable $\mathcal{D}_{\mathcal{Y}}$ as close as possible.

When the response variable can take a continuous value the task is referred to as *regression*. This is the case for our previous example, where the predicted variable can take any positive real value ($\mathcal{Y} \subset \mathbb{R}^+$). An illustration of this can be seen in Fig. 1.1a.

Suppose now a different example: a clinic is trying to determine if a skin lesion in a patient is a melanoma based on the colour and surface area by using previous cases as reference. In this situation, the training examples have two different features (colour and surface area) and, more importantly, the response variable can only take two values: melanoma or no melanoma. When the response variable can only take a discrete finite number of values the task is referred to as *classification* and the possible values of the response are called *classes*. When only two different classes are considered like in our current example (that is, $\mathcal{Y} = \{C_+, C_-\}$) it is referred to as *binary classification*, whereas when there are $Q > 2$ possible classes ($\mathcal{Y} = \{C_1, \dots C_Q\}$), say differentiating between different types of disease, it is referred to as *multiclass classification*. Illustrations for these kinds of tasks can be seen in Figs. 1.1b and 1.1c, respectively.



(a) Regression (b) Binary classification (c) Multiclass classification

Figure 1.1: Illustration of the different tasks in the supervised learning paradigm

1.2 ORDINAL REGRESSION

Regarding multiclass classification tasks, there are certain situations where a natural order exists between the different classes. For instance, consider a situation where a medical professional tries to determine the stage of a degenerative disease for a specific patient and considers the following possible diagnoses: healthy, mild, moderate and severe. This situation is similar to regression in that there exists an ordering of the possible values of the response, but no specific continuous value can be assigned to each patient. It is also similar to classification in that only a finite discrete number of values are possible, but regular classification treats all classes as equally dissimilar, whereas in this case a classification error of two stages is more severe than an off-by-one-stage error.

These kinds of tasks in the frontier between regression and classification belong to the field of *ordinal regression* (also referred to as *ordinal classification* in the literature) [38]. Starting from the multiclass classification framework where the range of the response variable is defined as $\mathcal{Y} = \{C_1, \dots C_Q\}$ for a Q-class problem, ordinal-domain tasks present an ordering relationship $\prec$ between the classes such that $C_1 \prec C_2 \prec \cdots \prec C_Q$. That is, $C_i \prec C_j \; \forall i < j$.

Ordinal regression is specially relevant in ML application fields like medical diagnosis [58], age estimation [13], quality assessment [101], weather forecasting [27] and many more.

A plethora of approaches are available for tackling ordinal regression tasks [38], organized in the taxonomy in Fig. 1.2.

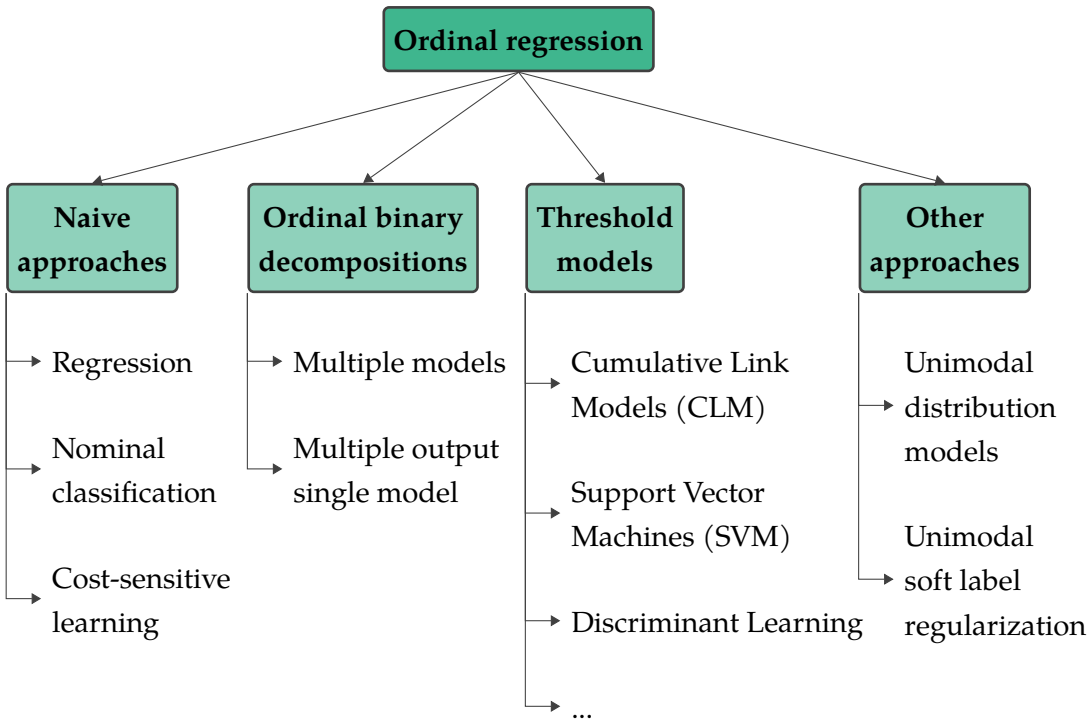


Figure 1.2: A taxonomy of ordinal regression methods based on the work by [38].

1.2.1 *Naive approaches*

Because ordinal regression exists in the border between regression and classification, the simplest approach is to treat an ordinal task just as regular regression or nominal classification.

If treated as regression, a real valued label is needed in order to substitute the original ordinal label [54] through some transformation $h : \mathcal{Y} \to \mathbb{R}$, e. g. $h(C_q) = q$. This mapping should hopefully reflect the original label order. This, however, requires making assumptions about the distance between class labels which are unknowable without prior information.

Treating it as nominal classification avoids making this assumption and is a widespread approach, but in the process it also avoids incorporating the order relation $\prec$ between the class labels, which may hinder performance as explored in Section 1.2.5.

A compromise can be made by simply introducing a misclassification error proportional to label distance in the target metric during the training process: this is known as cost-sensitive learning. Even so, the magnitude of the error is rarely known a priori.

Ordinal Binary Decompositions

OrderedPartitions	*OneVsNext*	*OneVsFollowers*	*OneVsPrevious*
$\begin{pmatrix} - & - & - & - \\ + & - & - & - \\ + & + & - & - \\ + & + & + & - \\ + & + & + & + \end{pmatrix}$	$\begin{pmatrix} - & 0 & 0 & 0 \\ + & - & 0 & 0 \\ 0 & + & - & 0 \\ 0 & 0 & + & - \\ 0 & 0 & 0 & + \end{pmatrix}$	$\begin{pmatrix} - & 0 & 0 & 0 \\ + & - & 0 & 0 \\ + & + & - & 0 \\ + & + & + & - \\ + & + & + & + \end{pmatrix}$	$\begin{pmatrix} + & + & + & + \\ + & + & + & - \\ + & + & - & 0 \\ + & - & 0 & 0 \\ - & 0 & 0 & 0 \end{pmatrix}$

Table 1.1: Types of OBD for an ordinal 5-class regression task. Each column represents each of the four binary sub-problems and each row represents the role of each class in each sub-problem: + means it is treated as the positive class, − the negative class and 0 means it is not considered.

1.2.2 *Ordinal binary decomposition*

While an ordinal regression task may be difficult to deal with all at once, the order relationship between the class labels can be used to decompose the original problem into a set of simpler binary sub-problems. This approach is known as Ordinal Binary Decomposition (OBD).

There are many ways to perform this decomposition (as illustrated in Table 1.1) with their own pros and cons. For example, the *OrderedPartitions* scheme allows the training process for each sub-problem to use the whole training dataset, as it considers every class as a valid option in all cases, but in exchange the sub-problems to be solved are more complex. On the other hand, the *OneVsNext* approach only considers the available samples for two contiguous classes for each sub-problem, reducing the number of available examples for training but also reducing the complexity of the problem that needs to be solved.

This new set of binary problems can, in turn, be solved using a separate model for each one (an *ensemble*) or a single model with multiple outputs.

1.2.3 *Threshold models*

It is common to assume the existence of a latent continuous variable y^* as the basis of the ordinal response y. Threshold models arise under this assumption: instead of learning the ordinal response directly, a function $f : \mathcal{X} \to \mathbb{R}$ trying to approximate the latent variable is learned along with a set of ordered thresholds $\mathbf{b} = (b_1, \dots, b_{Q-1}) \in \mathbb{R}^{Q-1}$, $b_i < b_j \forall i < j$ representing intervals in the range of f such that:

$$\hat{y} = C_q \iff b_{q-1} < f(\mathbf{x}) < b_q, \tag{1.1}$$

assuming that $b_0 = -\infty$ and $b_Q = \infty$.

While this type of model is related to the naive regression approach, in this case distances between labels are not assumed a priori but rather learned trough the training process, making them more flexible.

1.2.3.1 *Cumulative Link Models*

Cumulative Link Models (CLMs) are a family of ordinal regression models using a set of binary classification rules based on the same latent variable approximation $f(\mathbf{x})$ in order

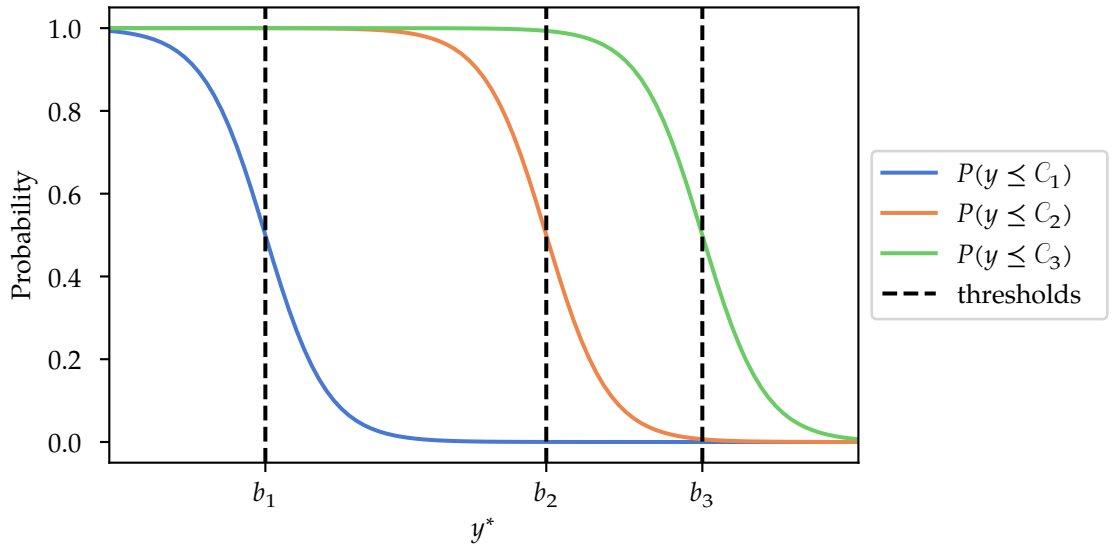


Figure 1.3: Interaction between the latent variable y^* and the thresholds b_i in a 4-class CLM

to predict probabilities of belonging to groups of contiguous classes $P(y \preceq C_q \mid \mathbf{x})$. These probabilities can then be easily transformed to single class probabilities $P(y = C_q \mid \mathbf{x}) = P(y \preceq C_q \mid \mathbf{x}) - P(y \preceq C_{q-1} \mid \mathbf{x})$.

One such member of the CLM family is the Proportional Odds Model (POM) [64]. This model extends the concept of logistic regression where $P(y = C_+) = 1/\big(1 + \exp\big(b - \mathbf{w}^{\mathrm{T}}\mathbf{x}\big)\big)$ by setting cumulative probabilities to be:

$$\mathrm{P}(y \preceq C_q \mid \mathbf{x}) = \frac{1}{1 + \exp\big(b_q - \mathbf{w}^{\mathrm{T}}\mathbf{x}\big)} \quad \forall 1 \leq q < Q. \tag{1.2}$$

In this context, the latent variable y^* is approximated through a linear model $\mathbf{w}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{w}$ are the parameters of the model. The interaction between the latent variable and the thresholds is illustrated in Fig. 1.3.

The idea behind the POM can be adapted to work with any other regression model by substituting the linear discriminant $\mathbf{w}^{\mathrm{T}}\mathbf{x}$ in Eq. (1.2). For example, in [99] the linear model is substituted by a Convolutional Neural Network (CNN) model with a single neuron as output, which allows both considering non-linear relationships as well as applying this method to work on structured inputs such as images.

1.2.4 *Unimodal distribution models*

So far, no assumptions have been made about the distribution of the response variable y, namely $\mathcal{D}_y$. However, in the framework of ordinal regression some general assumptions could be made.

Let's use the previous example of an ordinal regression task where we are trying to determine the stage of a degenerative disease. If the most probable scenario is the one corresponding to a moderate level, the next most probable one should be one of the contiguous affectation levels, namely mild or severe, and the likelihood should decrease as the distance between the class labels grows.

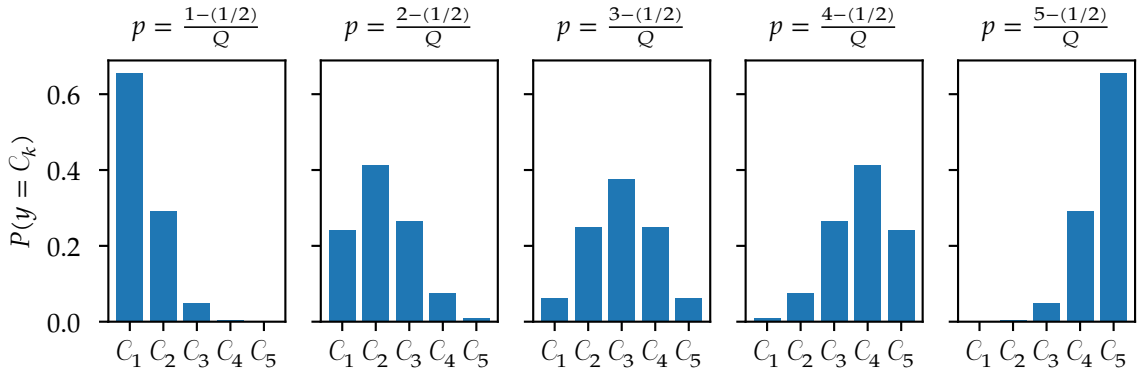


Figure 1.4: Example binomial distribution output $B(n, p)$ for a 5-class ordinal regression task. The value of parameter n is set to $Q - 1 = 4$. The ML model estimates the value of parameter p.

In mathematical terms, if the highest probability is assigned to C_q, i. e.:

$$C_q = \arg\max_{C_k \in \mathcal{Y}} \mathrm{P}(y = C_k \mid \mathbf{x}), \tag{1.3}$$

then the probability should monotonically increase from C_1 to C_q and monotonically decrease from C_q to C_Q, in other words, the distribution should be *unimodal*.

Unimodality can be enforced directly in-model through a parametric approach: the authors of [73] assume a specific unimodal distribution, such as the binomial or Poisson distributions. The goal of the ML model is then to estimate the parameters of the distribution that achieve a best fit with the training data. An example is shown in Fig. 1.4.

On the other hand, the non-parametric approach [72] does not impose a hard constrain on the shape of the distribution but instead introduces a penalty term in the optimization goal that favours the presence of a single mode in the output of the model. Both these methodologies have been explored using Artificial Neural Network (ANN) [72, 73] and Support Vector Machine (SVM) [74] models as the backbone.

Note that there is a certain level of intersection with CLMs, as their output is also enforced to be unimodal due to the strict ordering of the latent variable thresholds.

A more recent approach related to the work in this thesis is *soft labelling regularization*, where the target output of the model favoured by the optimization goal follows a predefined unimodal distribution such as the beta [100] or triangular [98]. This will be expanded in Chapter 6.

1.2.5 *Performance metrics*

A relevant and varied set of performance metrics is crucial in order to validate the well functioning of an ML model. Because ordinal regression is contained inside multiclass classification, traditional metrics of model performance like the following ones are still relevant.

Below are presented the most relevant metrics for this work. Metrics which are supposed to be maximized are indicated with (↑), and those to be minimized (i. e. errors) are indicated with (↓).

1.2.5.1 *Binary classification performance metrics*

THE BINARY CONFUSION MATRIX The confusion matrix is useful for representing all correct and incorrect classifications of an ML model and can be used to define all sorts of metrics.

In the binary case it consists of a 2×2 matrix, as seen in Table 1.2. All elements in the first row have a positive ground truth label whereas elements in the second row all have a negative ground truth label. As for the columns, elements of the first column have been assigned a positive predicted class label while elements of the second column have been assigned a negative label. Thus, all elements on the diagonal have been correctly classified and all elements not in the diagonal have been misclassified. The sum of all elements is equal to the total number of samples in the dataset N.

		Predicted class ($\hat{y}$)	
		Positive (C_+)	Negative (C_-)
True class (y)	Positive (C_+)	True Positives (TP)	False Negative (FN)
	Negative (C_-)	False Positive (FP)	True Negatives (TN)

Table 1.2: Binary classification confusion matrix

ACCURACY ($\uparrow$) It measures the ratio of samples classified correctly over the total number of samples. It is a value between 0 and 1, often expressed as a percentage. It can be defined using the indicator function $\mathbb{1}\{\cdot\}$, which is equal to 1 when its argument condition is true and 0 otherwise, or using the confusion matrix:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\hat{y}_i = y_i\} = \frac{\text{TP} + \text{TN}}{N}. \tag{1.4}$$

SENSITIVITY ($\uparrow$) The ratio of true positive samples that are correctly classified, also called the True Positive Rate or TPR:

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{1.5}$$

SPECIFICITY ($\uparrow$) The ratio of true negative samples that are correctly classified, also called the True Negative Rate or TNR:

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{1.6}$$

Its opposite metric is the False Positive Rate or $\text{FPR} = 1 - \text{TNR}$.

AREA UNDER THE ROC CURVE (AUC) ($\uparrow$) The Receiver Operating Characteristic (ROC) curve is a plot of the performance of a binary classifier as its discriminant threshold is increased or decreased. More precisely, it represents the TPR against the FPR. By its very nature, this is a monotonically increasing curve. In an ideal classifier, a high TPR is achievable with low FPR, i. e. the area under the curve would be close to 1. A classifier no better than

random guessing would show an area under the curve close to 0.5. An example of a ROC curve can be seen in Fig. 1.5.

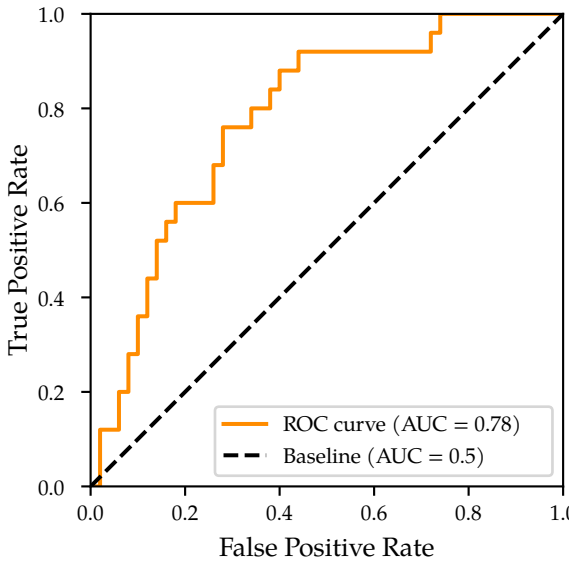


Figure 1.5: Example of a ROC curve

1.2.5.2 *Multiclass classification metrics*

THE MULTICLASS CONFUSION MATRIX The binary confusion matrix can be extended for the multiclass case, in which each row and column represent each ground truth and predicted class label, respectively, as is illustrated in Table 1.3. As with the binary case, the sum of all elements adds up to N and all of those in the diagonal have been correctly classified. The sum of the i-th row (the number of elements with true class C_i) is denoted as $n_{i,\bullet}$ and the sum of the j-th column (the number of elements predicted as C_j) is denoted as $n_{\bullet,j}$.

		Predicted class ($\hat{y}$)			
		C_1	C_2	...	C_Q
True class (y)	C_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,Q}$
	C_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,Q}$
	⋮	⋮	⋮	⋱	⋮
	C_Q	$n_{Q,1}$	$n_{Q,2}$	...	$n_{Q,Q}$

Table 1.3: Multiclass classification confusion matrix

CORRECT CLASSIFICATION RATE (CCR) (↑) CCR is the parallel of Accuracy for multiclass classification and is defined accordingly:

$$CCR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\hat{y}_i = y_i\} = \frac{1}{N}\sum_{q=1}^{Q} n_{q,q}\,. \tag{1.7}$$

PER-CLASS SENSITIVITY (↑) In multiclass classification, sensitivity can be thought of as a per-class metric S_q when looking at only the subset of samples which have a specific ground truth label C_q. Monitoring the minimum ($MinS$) or geometric average (GMS) can help identify when one of the classes is neglected by the model:

$$S_q = \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = \hat{y}_i = C_q\}}{\sum_{i=1}^{N} \mathbb{1}\{y_i = C_q\}} = \frac{n_{q,q}}{n_{q,\bullet}}, \tag{1.8}$$

$$MinS = \min_{1 \le q \le Q} S_q, \tag{1.9}$$

$$GMS = \sqrt[Q]{\prod_{q=1}^{Q} S_q}. \tag{1.10}$$

PER-CLASS SPECIFICITY (↑) In the same manner, per-class specificity Sp_q can also be monitored, as well as its minimum ($MinSp$) and geometric mean ($GMSp$) to make sure that no one class is being overly assigned:

$$Sp_q = \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i \neq C_q \wedge \hat{y}_i \neq C_q\}}{\sum_{i=1}^{N} \mathbb{1}\{y_i \neq C_q\}} = \frac{N - n_{q,\bullet} - n_{\bullet,q} + n_{q,q}}{N - n_{q,\bullet}} = 1 - \frac{n_{\bullet,q} - n_{q,q}}{N - n_{q,\bullet}}, \tag{1.11}$$

$$MinSp = \min_{1 \le q \le Q} Sp_q, \tag{1.12}$$

$$GMSp = \sqrt[Q]{\prod_{q=1}^{Q} Sp_q}. \tag{1.13}$$

AVERAGE AREA UNDER THE ROC CURVE ($AvAUC$) (↑) The ROC curve can only be obtained in a binary classification setting. However, if a certain class label q is considered as the positive class and all the rest are considered as the negative class (a scheme known as One vs. Rest or OvR) a different curve can be obtained for each class and the average area under them can be computed.

1.2.5.3 *Ordinal performance metrics*

When tackling ordinal regression tasks, traditional performance metrics fail to consider the different magnitudes of different misclassification errors. A misclassification of n classes $\hat{y} = C_{q \pm n}$ over the true class label $y = C_q$ should always be of less importance than a misclassification of $n + 1$ classes $\hat{y} = C_{q \pm (n+1)}$. The following metrics take this into account.

MEAN ABSOLUTE ERROR (MAE) ($\downarrow$) An error measure taken from regular regression. It is the mean absolute difference between the integer rank of the ground truth label and the predicted label. If we define the integer rank of an ordinal label as $O(C_q) = q$, then:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O(\hat{y}_i) - O(y_i)| = \frac{1}{N}\sum_{q=1}^{Q}\sum_{k=1}^{Q}|q-k|\,n_{q,k}. \tag{1.14}$$

MAE can also be defined as a per-class metric MAE_q, so that the average ($AvMAE$) and the maximum ($MaxMAE$) can be observed:

$$MAE_q = \frac{\sum_{i=1}^{N}\mathbb{1}\{y_i = C_q\}\,|O(\hat{y}_i) - O(y_i)|}{\sum_{i=1}^{N}\mathbb{1}\{y_i = C_q\}} = \frac{1}{n_{q,\bullet}}\sum_{k=1}^{Q}|q-k|\,n_{q,k}, \tag{1.15}$$

$$AvMAE = \frac{1}{Q}\sum_{q=1}^{Q}MAE_q, \tag{1.16}$$

$$MaxMAE = \max_{1\leq q\leq Q} MAE_q. \tag{1.17}$$

ROOT MEAN SQUARED ERROR ($RMSE$) ($\downarrow$) Similar to *MAE*, *RMSE* measures the squared difference of integer ranks instead of the absolute difference. The square root is then applied to maintain the original variable's units:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O(\hat{y}_i) - O(y_i))^2}. \tag{1.18}$$

COHEN'S WEIGHTED KAPPA COEFFICIENT (κ) ($\uparrow$) Proposed as a binary agreement metric [20], it was extended to admit different disagreement errors [21]. This coefficient measures the rating agreement between two different scorers (e. g. the ground truth label and an ML model output) based on a predetermined disagreement penalty and is bound between -1 and 1:

$$\kappa = 1 - \frac{\sum_{i=1}^{Q}\sum_{j=1}^{Q} w_{i,j}\,p_{i,j}}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} w_{i,j}\,e_{i,j}}, \tag{1.19}$$

$$p_{i,j} = n_{i,j}, \tag{1.20}$$

$$e_{i,j} = \frac{n_{i,\bullet}\cdot n_{\bullet,j}}{N}, \tag{1.21}$$

where $p_{i,j}$ is the observed agreement, $e_{i,j}$ is the expected agreement due to chance and $w_{i,j}$ is the disagreement cost when $y = C_i$ and $\hat{y} = C_j$. Throughout this work, the quadratic weighting is always considered, where $w_{i,j} = (i-j)^2$.

SPEARMAN'S RANK CORRELATION COEFFICIENT (r_s) ($\uparrow$) This metric measures the monotonicity of the relationship between two variables. It will be high when observations have a similar (or identical for a correlation of 1) rank, and low when observations have a dissimilar

(or fully opposed for a correlation of -1) rank between the two variables, which makes it suitable to assess ordinal regression performance. It is defined as:

$$r_s = \frac{\mathrm{Cov}(O(\hat{y}), O(y))}{\sigma_{O(\hat{y})}\sigma_{O(y)}}, \tag{1.22}$$

where $\mathrm{Cov}(O(\hat{y}), O(y))$ is the covariance between the predicted label integer rank and the ground truth label integer rank and $\sigma_{O(\hat{y})}$ and $\sigma_{O(y)}$ are the standard deviation of the predicted label integer rank and the ground truth label integer rank, respectively.

KENDALL'S RANK CORRELATION COEFFICIENT (τ_b) ($\uparrow$) Another rank correlation metric related to r_s proposed in [50]. It is based on the notion of observation pairs of two variables (in this case, the ground truth and predicted class labels) $(y_1, \hat{y}_1), (y_2, \hat{y}_2), \dots, (y_N, \hat{y}_N)$. Two different pairs $(y_i, \hat{y}_i)$ and $(y_j, \hat{y}_j)$ are concordant if either both $y_i < y_j$ and $\hat{y}_i < \hat{y}_j$ or $y_i > y_j$ and $\hat{y}_i > \hat{y}_j$, otherwise they are considered discordant. In this thesis a variant of the original coefficient that accounts for ties where $y_i = y_j$ or $\hat{y}_i = \hat{y}_j$ is used(τ_b) [51], which is specially important in a multiclass classification task:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(N(N-1)/2 - t_y)(N(N-1)/2 - t_{\hat{y}})}}, \tag{1.23}$$

$$n_c = \sum_{i=1}^{Q}\sum_{k=i+1}^{Q}\sum_{j=1}^{Q}\sum_{l=j+1}^{Q} n_{i,j}n_{k,l}\,, \tag{1.24}$$

$$n_d = \sum_{i=1}^{Q}\sum_{k=i+1}^{Q}\sum_{j=1}^{Q}\sum_{l=1}^{j-1} n_{i,j}n_{k,l}\,, \tag{1.25}$$

$$t_y = \frac{1}{2}\sum_{i=1}^{Q} n_{i,\bullet}(n_{i,\bullet} - 1), \tag{1.26}$$

$$t_{\hat{y}} = \frac{1}{2}\sum_{j=1}^{Q} n_{\bullet,j}(n_{\bullet,j} - 1), \tag{1.27}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, t_y is the number of ground truth label ties and $t_{\hat{y}}$ is the number of predicted label ties. Its value is bounded by $-1 \leq \tau_b \leq 1$ and it is expected to be zero when the ranks are independent and equal to 1 or -1 if the correlation is perfect or perfectly inverse, respectively.

1.3 DATA IMBALANCE AND DATA AUGMENTATION

It is often the case that the distribution of the class labels in the training dataset for a supervised learning task is severely skewed, i. e. very few examples exist of certain classes (called *minority classes*) while having an excess of examples of others (called *majority classes*). Moreover, these minority classes are more often than not the ones where misclassification errors are the costliest. This is a problem specially prevalent on medical domain tasks, where sick patients are usually the rarest and, of course, the ones in need of more attention.

This situation causes a lot of standard supervised learning algorithms to yield a lopsided performance between the different classes, in some cases even completely disregarding the minority classes with null sensitivity [39]. In this regard, global performance metrics fail to consider this aspect, and new ones are required to watch out for this effect, namely extreme metrics like *MinS*, *MinSp* and *MaxMAE* as well as class average metrics like *GMS*, *GMSp*, *AvMAE* and *AvAUC*, all of them defined in Section 1.2.5.

In order to overcome this obstacle, different types of approaches have been proposed, from introducing error cost weighting into the training process to methods selecting the minimum necessary samples to define class boundaries. However, the most popular are sampling methods, which modify the training dataset directly to provide a balanced class distribution.

In this work, we focus on *data augmentation* techniques (also known as *synthetic sampling*), which try to identify characterising features of the examples in the minority classes in order to create new synthetic but plausible examples to include in the training dataset with the goal of enhancing the model's generalization performance.

1.4 ARTIFICIAL FEEDFORWARD NEURAL NETWORKS

Among the many mathematical models that have been proposed to solve different regression and classification tasks, Artificial Neural Networks (ANNs) have been one of the most important in the latest decades of ML. The same way as other early proposals, this is a model inspired by biological learning, i. e. different theories of how learning occurs in the brain. This tie to neuroscience is why they are called *neural*, although general ML research is usually more concerned with its mathematical properties and abilities rather than its accuracy mimicking real neurological systems [46].

One of the most basic forms of an ANN, the multi-layer perceptron (MLP), is a function $f(\mathbf{x}; \mathbf{w}, \mathbf{b}) = \hat{y}$ mapping samples to a predicted label according to a set of weight and bias parameters, $\mathbf{w}$ and $\mathbf{b}$, respectively. The computation is performed in stages called *layers*. Each layer is composed of a certain number of *units* or *neurons* which are each connected to the neurons in the next layer forming a *network*, in such manner that information flows from one layer to the next without feedback connections, reason why they are called *feedforward*. The most common type of MLP assumes that all neurons in a layer are connected to all the neurons in the previous layer, i. e. they are *fully connected*. An illustration of the information flow in an MLP can be seen in Fig. 1.6.

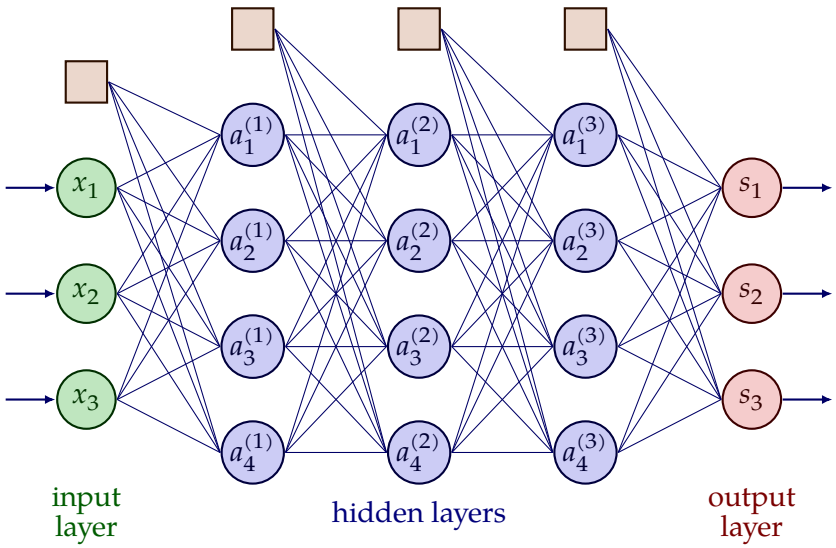


Figure 1.6: Example of a fully connected MLP architecture with 3 inputs, 3 hidden layers with 4 neurons each and 3 outputs. Circular nodes represent neurons and square nodes represent bias terms in the computation.

Each of these forward computations consists of a matrix multiplication with a bias term, producing an affine transformation, followed by a non-linear operation called *activation function*. This process is illustrated in Fig. 1.7. The selection of these functions is of great importance in the construction of the model [97], and their non-linearity is essential in giving the model its ability to approximate arbitrarily complex functions [56].

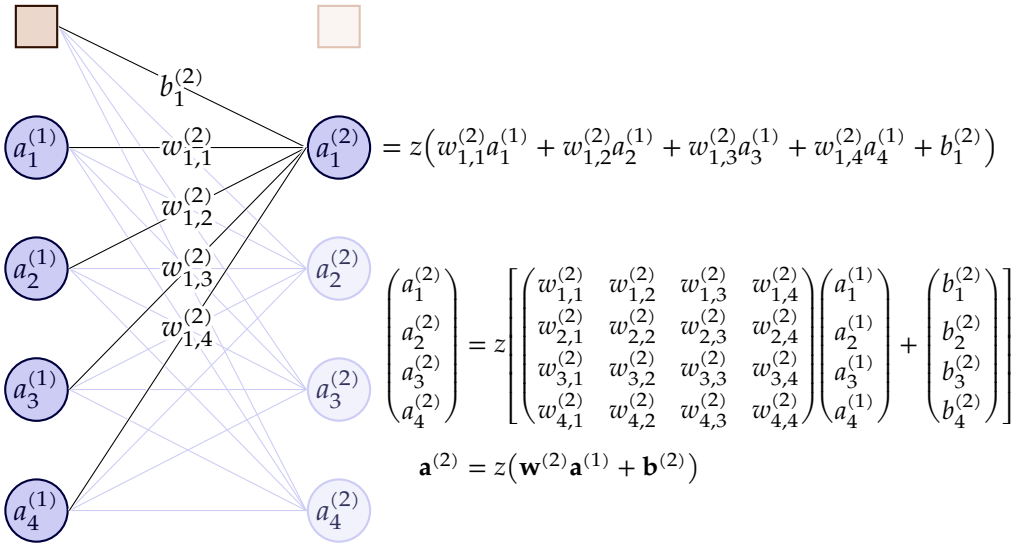


Figure 1.7: Computation of a single neuron activation $a_1^{(2)}$ (above) and of the whole second layer activations (below, both in detail and in vector notation) in the architecture presented in Fig. 1.6, using z as the activation function.

1.4.1 *Activation functions*

A great variety of activation functions are available suited to different contexts inside an ANN model depending on their domain, range, behaviour, shape, derivatives, etc.

The sigmoid function

The sigmoid function (also known as the logistic function), commonly denoted as $\sigma(x)$ maps any real value into the $(0,1)$ interval ($\sigma : \mathbb{R} \to (0,1)$). It is defined as:

$$\sigma(x) = \frac{1}{1+\exp(-x)}. \tag{1.28}$$

It maps large negative values into values close to 0 and large positive values into values close to 1, with gradient tending towards 0 in the extremes and maximum gradient of 0.25 when $x = 0$. It can be conceptualized as a smoother version of the 'threshold' activation $\mathbb{1}\{x > 0\}$ which is differentiable. Although it can create a problem of *vanishing gradients* in the hidden layers during training of an ANN, it is useful for *producing probabilities from unbounded scores*.

The Rectified Linear Unit (ReLU)

The Rectified Linear Unit (ReLU) is a piecewise-defined function that solves the problem of the vanishing gradients of the sigmoid function. It is defined as:

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1.29}$$

Note how the derivative is always 1 for $x > 0$. Even though it is not differentiable at $x = 0$, this does not affect its performance.

However, this introduces a new problem: its derivative is 0 for all negative values of the input. This may cause too many neurons in the network to have a null gradient and thus, hinder the training process. To fix this, an alternative called the Leaky Rectified Linear Unit (LReLU) is used, which adds a small slope in the negative region through a parameter $\alpha > 0$ (set before the training process):

$$\text{LReLU}(x) = \begin{cases} x, & \text{if } x > 0, \\ \alpha x, & \text{otherwise.} \end{cases} \tag{1.30}$$

This parameter is often set to values significantly lower than 1 like $\alpha = 10^{-2}$.

The LReLU, although simple, increases the variance of its output by virtue of having an arbitrarily large negative output. To reduce this complexity, the Exponential Linear Unit (ELU) uses a saturated negative part defined by an exponential function:

$$\text{ELU}(x) = \begin{cases} x, & \text{if } x > 0, \\ \alpha(e^x - 1), & \text{otherwise.} \end{cases} \tag{1.31}$$

Another alternative is the softplus function (s_+), a smoothed version of ReLU. It removes the discontinuity at $x = 0$, has non-zero gradients in the negative input region and also has a derivative approaching 1 as x increases, giving it robustness against the vanishing gradients problem:

$$\text{s}_+(x) = \ln(1 + e^x). \tag{1.32}$$

Numerous other alternatives to the ReLU exist with learnable parameters (e. g. the Parametric ReLU) or even with some added randomness (e. g. the Randomized LReLU) [97].

The softmax function

In classification tasks, the output stage of a network is often trying to approximate a conditional discrete probability distribution $\text{P}(y = C_q \mid \mathbf{x})$. Per the basic probability rules, it is desirable in these cases that the total sum $\sum_{C_q \in \mathcal{Y}} \text{P}(y = C_q \mid \mathbf{x})$ is equal to 1. This can be achieved by using the softmax function as the activation of the output layer. In this case, the activation of each output neuron s_i is dependent on each other neuron in the same layer:

$$\text{softmax}(s_i) = \frac{\exp(s_i)}{\sum_{j=1}^{Q} \exp(s_j)}. \tag{1.33}$$

In this way, the network could be though of as approximating a kind of 'log probabilities' usually called *scores*, which are then transformed into probabilities through the softmax function. As its name suggests, the highest probability will be given to the neuron with highest score and so on.

1.4.2 *ANN model training*

Ideally, one would want a way to optimize the values of the parameters $\boldsymbol{\theta} = (\mathbf{w}, \mathbf{b})$ (that is, the values of each layer's weights and biases) in order to approximate as closely as possible the real distribution of data $\mathcal{D}$. This can be conceptualized as minimizing a *risk* function J (i. e. the classification error) over the parameters of the model to be minimized:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f(\mathbf{x};\boldsymbol{\theta}), y)], \tag{1.34}$$

where $\mathcal{L}(f(\mathbf{x};\boldsymbol{\theta}), y)$ is the *loss function* for a single example $\mathbf{x}$ given its true label y. For a classification task this could be, for example, the '0-1 error': $\mathcal{L}_{01}(f(\mathbf{x};\boldsymbol{\theta}), y) = \mathbb{1}\{y_i = f(\mathbf{x}_i \mid \boldsymbol{\theta})\}$.

In ML however, the real distribution of the data is unknown, and only a sample D is available. Instead of minimizing the general risk, the *empirical risk* $\hat{J}$ can be used as a proxy for optimization:

$$\hat{J}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}_i,y_i)\in D}[\mathcal{L}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i)] = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i). \tag{1.35}$$

The problem with minimizing the empirical risk is that it is prone to *overfitting*, that is, adjusting the parameters to fit the specifics of sample D at the cost of increasing the general risk J.

In practice, empirical risk minimization is not used with ANN models, but rather a similar approach called *gradient descent* [105]. Gradient descent is performed in steps: the parameters $\boldsymbol{\theta}$ of the model are updated at each step following the negative direction of the gradient of the empirical risk. The magnitude of these steps is controlled through a training hyperparameter known as the *learning rate* η:

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \eta\nabla\hat{J}(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta\frac{1}{N}\sum_{i=1}^{N}\nabla\mathcal{L}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i). \tag{1.36}$$

Note how gradient descent requires the cost function $\mathcal{L}$ to be differentiable, which a typical error function like the previously mentioned $\mathcal{L}_{01}$ is not. Thus, a new differentiable surrogate loss function aligned with the real goal is defined, such as the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(f(\mathbf{x};\boldsymbol{\theta}), y) = \sum_{q=1}^{Q} -\log(P(y = C_q \mid \mathbf{x}))\mathbb{1}\{y = C_q\}. \tag{1.37}$$

For large datasets, however, Eq. (1.36) is a fairly expensive computation. The process can be broken down into smaller samples from the training dataset, called *minibatches*, taking the average over only that small sample. This method is known as *stochastic gradient descent* (SGD).

Unlike general optimization where the process stops when a local minimum is achieved, when training an ANN a different stopping criterion is used: typically, a subset of the training

set is reserved as a *validation set* and a certain metric like CCR is monitored, stopping when overfitting is detected.

There exist numerous variants of SGD that add elements like adaptive learning rates for individual parameters (like AdaGrad [28] and RMSProp [42]) or momentum terms (like Adam [52]).

Computing the loss gradient is the core operation in SGD methods. In order to do it efficiently, the feedforward nature of ANN models can be exploited by the use of the *back-propagation algorithm* [46].

1.5 CONVOLUTIONAL NEURAL NETWORKS

When dealing with structured grid-like data, traditional MLP models fall short of performance and efficiency. This includes time-series (1D temporal data, e. g. an electrocardiogram), images (2D grids of pixels, e. g. conventional radiography) and volumetric scans (3D grids of voxels, e. g. a CT scan) data. General matrix multiplication considers all interactions between all inputs, hindering the training process due to inefficiency.

Convolution is an operation of two arguments, an *input* I and a *kernel* K, resulting in a *feature map* S, denoted by $S = K * I$. Although borrowed from statistics, it takes on a new meaning in an ANN context: both the inputs and outputs are assumed to be multidimensional arrays. For a 2D convolution operation between an input of size $H \times W$ and C channels and a kernel of size $h \times w$, the operation for a specific pixel (i, j) can be defined as:

$$S(i,j) = (K * I)(i,j) = \sum_{c=1}^{C} \sum_{a=1}^{h} \sum_{b=1}^{w} I(i+a-1, j+b-1, c) K(a,b,c). \tag{1.38}$$

This process can be thought of as sliding the kernel as a window over the input, multiplying each overlapping entry and adding up the result to compute the corresponding value of the feature map. A visualization of this can be seen in Fig. 1.8.

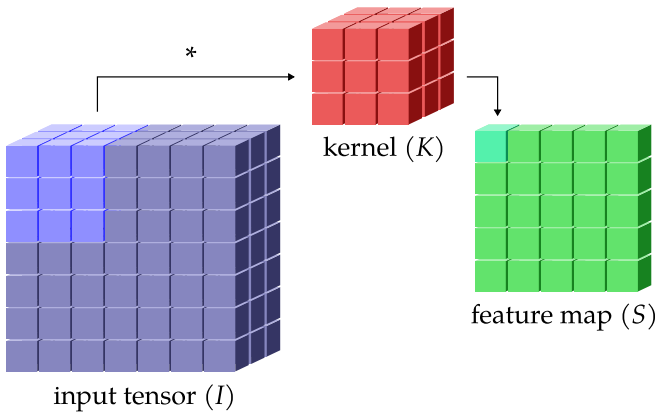


Figure 1.8: Example of a convolution operation. The highlighted part of the input tensor I ($7 \times 7 \times 3$) is multiplied element-wise with the elements of kernel K ($3 \times 3 \times 3$) and the resulting sum corresponds to the highlighted pixel of feature map S ($5 \times 5 \times 1$). This is repeated for every possible offset of K over I to obtain the complete S.

Convolution can be used at certain stages of an ANN instead of matrix multiplication. These kinds of models are called Convolutional Neural Networks (CNNs), and it will be the goal of an ML algorithm to optimize the values of the feature kernels as parameters.

They present several advantages when dealing with structured data like the type mentioned previously:

- They can detect *sparse interactions* in the input, i. e. interactions between spatially or temporally close values of the input, and build up to higher order interactions by composing several interactions at different layers instead of considering all interactions at once.
- *Parameters are shared* between computations at every location of the image, so a single set of parameters can be considered for all locations.
- They are also *equivariant* to certain transformations, mainly translation of features: if an object is translated in space or time in the input, the resulting feature map will experience the same translation.

Convolution operations are usually performed in the initial stages of a CNN model. At a certain point, the spatial structure of the information is dropped, often through what is called a *global pooling* operation by which all the elements in each feature map are condensed into a single scalar (e. g. by taking the average, called global average pooling or GAP, or the maximum, called global maximum pooling or GMP). After that, one or more fully connected layers are usually present.

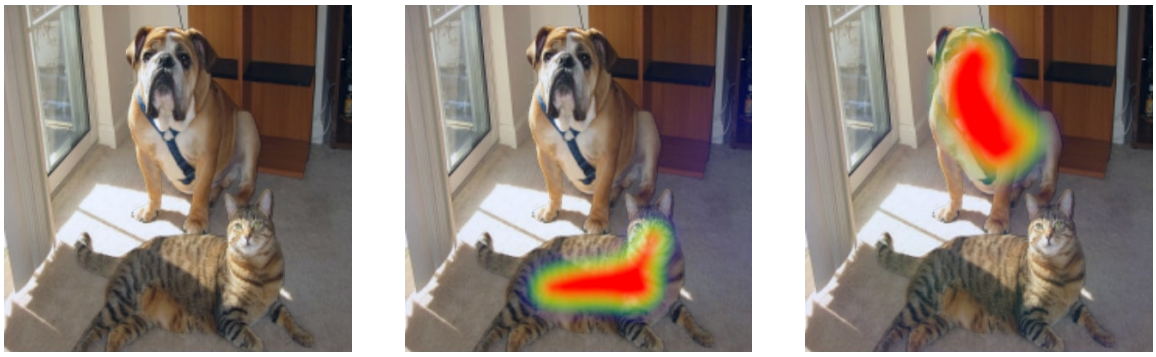
1.6 EXPLANATION METHODS

ANN models function as a black-box: deciphering which parts of the input are really relevant to the final decision at the output after a long series of intricate operations seems insurmountable. In some contexts, obtaining such explanation is a desirable or even necessary step. For example, one could leverage this information in order to debug a certain implementation of a model. Also, in critical situations such as medical diagnostic, the rationale behind a decision, like the finding of a specific lesion, could be even more important than the decision itself. This is the problem of *feature attribution*.

Methods that try to solve this problem are referred to as *explanation methods* and are also known as saliency methods or attribution methods. Given an input sample $\mathbf{x}$, a target class $y = C_q$, a CNN model, and the computation of the output of the model when given $\mathbf{x}$ as input, the result of an explanation method will be a array E_q with a single channel and the same size as $\mathbf{x}$, called the *explanation map*. Locations where E_q is close to 0 are deemed irrelevant and those close to 1 are considered important to the model decision for class C_q. An example is shown in Fig. 1.9.

1.7 MEDICAL IMAGING

In the realm of medical image analysis, the selection and understanding of diverse imaging modalities play a pivotal role in the success and applicability of ML algorithms. The intricate nature of medical data demands a nuanced comprehension of the distinctive characteristics inherent in various imaging modalities, each offering a unique perspective into the underlying physiological or pathological processes. This section serves as a foundational exploration into the landscape of medical image modalities and their essential characteristics, justifying its inclusion within the broader context of ML applications.



(a) Input image (b) Explanation for the class *cat* (c) Explanation for the class *dog*

Figure 1.9: Example of explanation maps for a CNN classifying animals in photos. The explanation map has been superimposed over the image: values close to 0 are transparent, while values close to 1 are highlighted in red. Image taken from the COCO dataset: https://cocodataset.org/#explore?id=114269

Medical professionals of every discipline use imaging techniques everyday for diagnosis, treatment, intervention and research tasks. Imaging allows them to observe internal structures of the body which are otherwise hidden.

Different cases and goals call for particular image modalities which let the clinician observe specific anatomical structures, lesions and bodily processes.

By delving into the intricacies of these modalities we aim to establish an understanding of the challenges and opportunities posed by different medical imaging sources for the subsequent development and optimization of ML models tailored for effective diagnostic, prognostic, and therapeutic applications in the domain of medical image analysis.

1.7.1 *Projectional imaging vs. Tomography*

Most traditional imaging techniques fall into the category of *projectional imaging*, in which radiation is captured onto a flat medium (usually film or a digital sensor) to form a 2D image.

However, situations arise where projections alone are not suitable to observe the object of interest. Computed Tomography (CT) is a technique that enables clinicians to obtain full 3D reconstructions of anatomical parts by capturing sequences of 2D slices of a set thickness which can then be stacked. The acquisition of said slices from a set of projections is possible through the use of reconstruction algorithms.

Note that in some of the imaging techniques described below both projections and tomographies are applicable.

1.7.2 *Photography*

The most basic type of medical image is simply photography: the capture of visible light images for later analysis, which is performed nowadays in digital form almost exclusively. Simple consumer grade cameras are suitable for tasks like posture or gait analysis, although specialized equipment is also used for specific images like *fundus photography* (capturing an image of the inner surface of the back of the eye), as shown in Fig. 1.10.

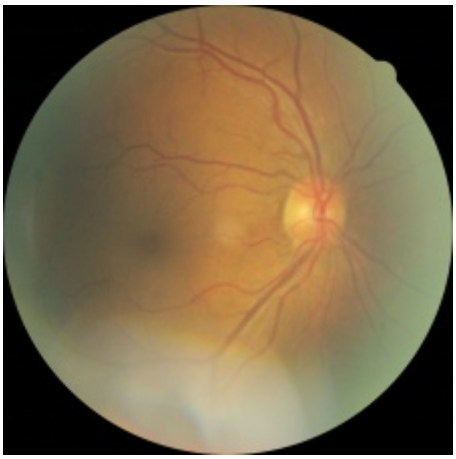


Figure 1.10: Example of fundus photography. Source: Kaggle 'Diabetic Retinopathy Detection' challenge: https://visualsonline.cancer.gov/details.cfm?imageid=2510

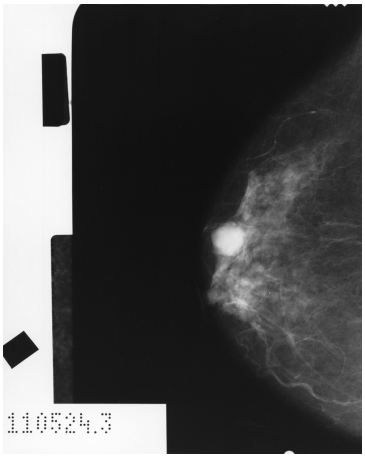


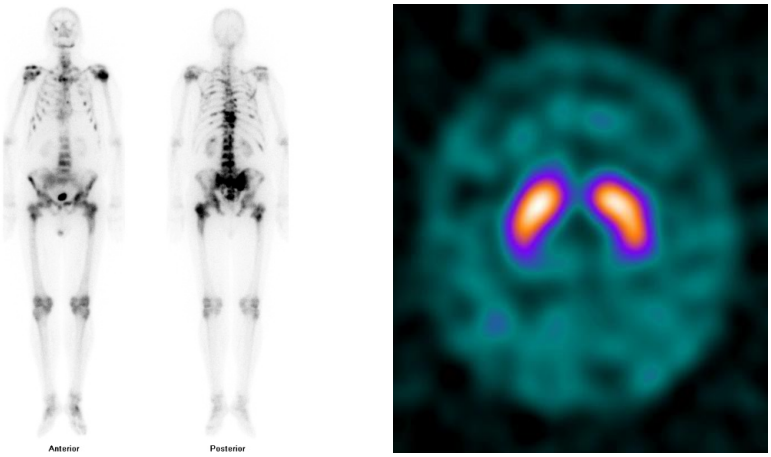
Figure 1.11: Film radiography of a breast, known as a *mammogram*, showing a colloid carcinoma. Source: American College of Radiology: https://visualsonline.cancer.gov/details.cfm?imageid=2510

1.7.3 *Radiography*

X-rays (from 10 nm to 10 pm of wavelength), gamma rays (< 10 pm) and other types of ionizing radiation exist in a different part of the light spectrum for which the external parts of the body are transparent. This allows radiologists to observe the internal parts without invasive procedures.

The most common type of radiography is *projectional radiography*, where parts of the body are exposed to a source of high-energy radiation and the resulting shadow is captured in film or by a digital sensor. Different types of substances block varying amounts of radiation, resulting in an image where properties like tissue density can be observed. An example can be seen in Fig. 1.11.

Radiography may also be used in CT to obtain 3D volumes from a set of slices.



(a) Whole body bone scintigraphy

(b) Brain SPECT

Figure 1.12: Examples of nuclear imaging. Figure (a) courtesy of Carlos Eduardo Anselmi, Radiopaedia.org, rID: 12294. Used under a CC BY-NC-SA 3.0 license. Figure (b) from the novel Parkinson's dataset presented in Section 4.3.1

1.7.4 *Magnetic Resonance*

Magnetic Resonance Imaging (MRI) techniques use the power of large magnets (usually in the range of 1.5 T to 3 T) to excite the nuclei of hydrogen atoms of water molecules present in the body. An electromagnetic pulse in the radio frequency causes these nuclei to disarrange and later realign with the magnetic field, producing their own radio waves in the process, which are then detected and reconstructed into an image. This type of imaging is used exclusively in CT.

MRI offers high contrast in areas of soft tissue while avoiding the use of ionizing radiation, but poses some limitations like greater patient discomfort due to noise and capture time and incompatibility with metal implants when compared with other techniques.

1.7.5 *Nuclear Imaging*

In nuclear imaging, special radioactive substances known as radioligands are administered safely to a patient with the goal of observing certain physiological processes. The isotopes present in these substances are more likely to be absorbed by biologically active tissue in the patient's body, like points of bone fracture or tumours. Unlike other types of imaging where radiation comes from outside the body, the radiation given off by these substances originating from the body can then be captured by special sensors.

One of the most common types of nuclear image is *scintigraphy* (Fig. 1.12a), where special gamma radiation cameras can capture 2D images. Whole body scintigraphy images are used regularly in conjunction with technetium 99m methylene diphosphonate (^{99m}Tc-MDP) in order to detect bone lesions like some types of bone metastasis in cancer patients.

Single-photon emission computed tomography (SPECT) (Fig. 1.12b) is a 3D imaging technique that uses gamma cameras in combination with CT techniques. This allows the clinician to observe levels of biological activity in specific 3D areas, e. g. dopaminergic

activity in the brain using ioflupane as a radioligand (containing ^{123}I) for the early diagnosis of Parkinson's disease (PD).

2 THESIS OVERVIEW

2.1 MOTIVATION AND CHALLENGES

ML techniques developed in the last decade have made it possible to apply the available computational power to tasks previously inaccessible to computer science: deep learning has been extensively applied to 2D image recognition tasks of all kinds (classification, segmentation, pose estimation, etc.) as well as to time series (1D). There is potential in applying these techniques to 3D or *volumetric* images, although their use is not as widespread as in the 2D case.

Computational models can also serve as a support tool for the physician or medical team, being able to provide additional decision support or to detect inconsistencies and/or potentially complex cases that may have been overlooked. Of particular interest is this type of analysis for medical imaging to automate tasks such as anomaly detection [96] and diagnosis [12].

Regarding ordinal regression tasks, traditional ML methods have already been successfully applied and adapted to this framework such as logistic regression, SVMs or ANNs [18]. However, deep learning models have hardly been explored for ordinal regression tasks and have great potential to improve classification performance. This kind of models, even though very powerful, function as a 'black box': their internal workings are not readily apparent. Explanation methods exist which are able to highlight relevant parts of the input for the decision process, but these do not consider ordinal class information.

Finally, a very common problem in real applications especially relevant in biomedical settings is the imbalance of class examples. The very nature of these problems means that there is a much smaller sample size of some classes compared to others (for example, far fewer sick patients than healthy ones, or far fewer cases in the intermediate stage of a disease than at the end). Deep learning methods are sensitive to this imbalance and require a large number of examples to generalise their predictions [48]. This is why the development of data augmentation methods is required to circumvent these obstacles and thus extract as much potential as possible from the available data.

Thus, a gap is identified in the vicinity of ordinal regression. There exists potential in designing deep learning models and explanation methods that incorporate ordinal information from the ground up in their formulation and training process. Such innovations may be able to improve the performance of both existing and new automatic tasks in the medical domain, especially regarding medical imaging. This thesis aims to address this gap by providing novel solutions and methodologies.

2.2 OBJECTIVES

The main objectives of this thesis can be summarized in the following points:

1. To develop new deep learning CNN architectures capable of dealing with ordinal response variables natively in order to improve their ordinal performance.

2. To propose new data augmentation techniques to tackle class imbalance problems in ordinal regression tasks which take into account ordinal information.

3. To apply these innovations to solve real medical 2D and 3D image diagnosis problems as well as related biomedical ordinal regression problems and study the potential performance improvements.

4. To propose new explanation methods that incorporate ordinal information in the explanation map generation process so as to improve the detection of relevant input features.

2.3 THESIS STRUCTURE

The rest of this thesis is organized as follows:

- Chapter 3: NATIVE ORDINAL REPRESENTATIONS FOR CNNS

  Introduction to the limitations of traditional CNN architectures in handling ordinal regression tasks and the development of the Ordinal Binary Decomposition (OBD) model with an Error Correcting Output Codes (ECOC) scheme, aiming to improve classification performance for ordinal tasks.

- Chapter 4: COMPUTER-AIDED DIAGNOSIS FOR PARKINSON'S DISEASE: AN APPLICATION OF ORDINAL DATA AUGMENTATION FOR 3D IMAGE IMBALANCED DATASETS

  Exploration of the application of the OBD model to diagnose neurological damage in Parkinson's disease using volumetric brain scans. Study of the challenges of working with 3D images, low sample size, and class imbalance, introducing a native 3D CNN architecture and an ordinal data augmentation procedure to address these issues.

- Chapter 5: EXAMINING THE DECISION PROCESS OF ORDINAL CNNS

  Recognition of the interpretability challenges in CNN models, especially in the context of ordinal regression tasks. Validation of existing explanation methods on ordinal regression, introducing two novel ordinal-specific adaptations, GradOBD-CAM and OIBA, aiming to provide insights into the decision-making processes of ordinal CNN models. Proposal of a visual explanation evaluation procedure for assessing ordinal performance.

- Chapter 6: ADDITIONAL WORKS

  Discussion of several additional works developed in parallel to this thesis.

- Chapter 7: CONCLUSIONS AND FUTURE WORK

  Summary of overall contributions, evaluation of the achievement of the proposed objectives and foundation for future work.

Part II

CONTRIBUTIONS

3

NATIVE ORDINAL REPRESENTATIONS FOR CNNS

Traditional ML methodologies involving Convolutional Neural Network (CNN) models have shown an outstanding performance in nominal classification tasks where the input information consists of images. As of today, CNNs are able to achieve 90 % accuracy on the ImageNet-1K dataset [11, 24].

However, nominal CNN architectures have been generally designed in a similar fashion, suited to classification problems that deal with class labels without any domain-defined relations. The presence of an order relationship between them, as is the case for ordinal regression tasks, is therefore ignored during training and evaluation of the model.

Exploiting this information has the potential to improve the performance of these models regarding more relevant metrics to the specific domain of the problem, i. e. ordinal metrics that treat misclassification errors differently from one another according to the order relationship of the labels.

This chapter is dedicated to the *development of a novel output architecture* for CNNs based on the idea of Ordinal Binary Decomposition (OBD) which is compatible with existing proposals as a nearly drop-in replacement. This architecture is paired with a *matching class assignment rule* based on the Error Correcting Output Codes (ECOC) scheme. We aim to prove that this approach is capable of improving the classification performance for ordinal tasks.

ASSOCIATED PUBLICATION: **Javier Barbero-Gómez**, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Error-Correcting Output Codes in the Framework of Deep Ordinal Classification'. In: *Neural Processing Letters* (12th May 2022). DOI: 10.1007/s11063-022-10824-7.

JCR (2022): 3.1. Ranking position in Computer Science, AI: 85/145 (Q3)

3.1 RELATED WORK

Previous research in the field of ordinal regression has primarily focused on handling unstructured input data that lacked spatial or temporal relationships between inputs. These early approaches include using conventional regression techniques with rounding applied to the outputs [54] or incorporating label distance as a cost penalty [53]. However, their performance is often constrained due to the unequal underlying distances between labels.

To address this limitation, Cumulative Link Models (CLMs) emerged as a promising solution. CLMs, such as the Proportional Odds Model (POM) [64] and the gologit model [107], not only learn a latent continuous variable but also derive a set of thresholds for each rank. This advancement improves their ability to handle ordinal data effectively. Additionally, adaptations of Support Vector Machines (SVMs), such as Support Vector for Ordinal Regression with Implicit constraints (SVORIM) and Support Vector for Ordinal Regression with Explicit constraints (SVOREX) [19], introduce ordinal constraints into the model optimization process, further enhancing their suitability for ordinal regression tasks.

Another notable approach known as OBD aims to break down the original ordinal problem into a series of binary subproblems. Examples of OBD methods include the cascade linear utility model [108], where distinct models address each binary subproblem, and neural networks with multiple outputs, each dedicated to a binary subproblem [18, 25]. However, OBD methods face challenges in combining the individual outputs to make a final decision.

These existing approaches are not directly applicable to structured data like 2D images, where domain-specific feature extraction remains essential. CNNs emerged as a valuable tool for automatically extracting learned features from structured data in classification tasks. Nevertheless, CNNs have a tendency to overfit due to their large number of parameters, resulting in suboptimal generalization performance. Researchers have explored various techniques to mitigate this issue, including traditional methods like L_2 regularization and dropout, as well as more recent strategies such as multi-stage implicit regularization [111] and network path pruning [110].

Adapting CNNs to handle ordinal information is a recent research direction that requires further exploration. Some initial efforts [32, 99] have incorporated CLMs as activation functions for a single output neuron of a CNN. In [68], a CNN architecture was proposed for addressing the OBD version of age estimation, with a straightforward approach to combining binary outputs to obtain a rank. [59] introduced an alternative methodology tailored for small datasets, relying on triplets of samples and majority voting. Lastly, [17] built upon the work presented in [68] by constraining the maximum binary error for each output, resulting in performance improvements. This area of research still warrants extensive investigation and development.

3.2 GOALS

For this chapter, our goals are centred around objective number 1 from Section 2.2. More precisely, they are:

- Testing the different ways in which CNNs can be adapted to work with ordinal data.
- Proposing a novel output scheme for solving ordinal regression tasks with CNNs.
- Testing the hypothesis that an ordinal method is able to outperform a nominal classification technique.
- Testing the hypothesis that our proposed method offers improvements over other ordinal approaches.

3.3 BASE NOMINAL CNN METHODOLOGY

The general framework for nominal multiclass classification is explained in Section 1.1.1. In short, this task consists in assigning a class label from a discrete finite set $\mathcal{Y} = \{C_1, \dots, C_Q\}$ to a given sample $\mathbf{x}$ from the population. In the nominal framework no predetermined relation is assumed between the class labels.

Where CNNs excel is in their image classification capabilities. As explained in Section 1.5 they are able to natively capture the spatial relation between nearby pixels, which are more strongly associated than distant ones.

The specifics of each CNN architecture are wildly different, but most of them follow the same general premise:

- In the first phase of the network several concatenated *convolution* (and maybe *pooling*) operations are carried out starting from the input image.
- The resulting *mapped features* (sometimes summarized by some global pooling operation) are processed by one or more fully connected layers.
- In the end, an *output layer* with as many neurons as class labels is used. These *class scores* are then transformed to *class probabilities* $\mathrm{P}(y = C_q \mid \mathbf{x})$ using the softmax activation function (Section 1.4.1).

In order to train this model, categorical cross-entropy is chosen as the loss function to be minimized, as defined in Eq. (1.37).

During evaluation and deployment of the model, the predicted class label $\hat{y}$ assigned to sample $\mathbf{x}$ is the one that maximizes the output probability of the model:

$$\hat{y} = \underset{C_q \in \mathcal{Y}}{\arg\max}\, \mathrm{P}(y = C_q \mid \mathbf{x}). \tag{3.1}$$

3.4 ADAPTING CNNS FOR ORDINAL REGRESSION TASKS

Ordinal regression tasks (Section 1.2) differ from nominal multiclass classification tasks in that an ordering relationship $\prec$ is assumed between the class labels: $C_1 \prec C_2 \prec \cdots \prec C_Q$, that is, $i < j \iff C_i \prec C_j$. In this situation not all misclassification errors are considered equal, e. g. assigning a predicted label $\hat{y} = C_{q\pm1}$ neighbouring the true target label $y = C_q$ is an error of less magnitude than assigning one two labels away $\hat{y} = C_{q\pm2}$.

Several avenues are available for introducing this information into CNN models:

1. Using a loss function that penalizes different misclassification errors according to the ordering relationship of the class labels.
2. Altering only the output phase of an existing architecture to accommodate an ordinal structure, leaving the rest of the network as-is.
3. Additionally to the previous point, possibly modifying the predicted label decision rule.

Next are shown two different examples from the literature as well as a novel one presented in this thesis.

3.4.1 *An ordinal loss function: Quadratic Weighted Kappa*

As discussed previously, categorical cross-entropy treats every misclassification equally. One naive approach to ordinal regression is substituting this for an order-sensitive loss function.

The authors of [23] adapt the Cohen's weighted kappa coefficient defined in Eq. (1.19) as a loss function that can be used for training a CNN model. For this, they first express the original formulation of κ using output probabilities instead of predicted labels:

$$\hat{\kappa} = 1 - \frac{\sum_{i=1}^{N} \sum_{q=1}^{Q} w_{O(y_i),q}\, \mathrm{P}(y_i = C_q \mid \mathbf{x}_i)}{\sum_{c=1}^{Q} \frac{N_c}{N} \sum_{q=1}^{Q} \left(w_{c,q} \sum_{i=1}^{N} \mathrm{P}(y_i = C_q \mid \mathbf{x}_i) \right)}, \tag{3.2}$$

where N_c is the number of samples with class label C_c in the dataset. Then, the kappa loss $\mathcal{L}_\kappa$ is defined as:

$$\mathcal{L}_\kappa = \log(1 - \hat{\kappa}), \text{ where } \mathcal{L}_\kappa \in (-\infty, 2]. \tag{3.3}$$

A model trained in this way can use the same label assignment scheme as the one from Eq. (3.1).

3.4.2 *The Cumulative Link Model*

As explained in Section 1.2.3.1, Cumulative Link Models use an approximation of a latent continuous variable along with a set of thresholds in order to predict the probability of the true class label of a sample being in a set of contiguous groups.

Whereas the approximation of the latent variable performed by the POM is a linear model, as shown in Eq. (1.2), the authors of [99] substitute it by a CNN model with a single neuron output denoted by $f(\mathbf{x})$:

$$\mathrm{P}(y \preceq C_q \mid \mathbf{x}) = \frac{1}{1 + \exp\big(b_q - f(\mathbf{x})\big)} \ \forall q,\ 1 \leq q < Q. \tag{3.4}$$

In order to use the arg max label assignment scheme from Eq. (3.1), equality probabilities need to be computed from these in the following manner:

$$\mathrm{P}(y = C_q \mid \mathbf{x}) = \begin{cases} \mathrm{P}(y \preceq C_1 \mid \mathbf{x}), & \text{if } q = 1, \\ \mathrm{P}(y \preceq C_q \mid \mathbf{x}) - \mathrm{P}(y \preceq C_{q-1} \mid \mathbf{x}), & \text{if } 1 < q < Q, \\ 1 - \mathrm{P}(y \preceq C_{Q-1} \mid \mathbf{x}), & \text{if } q = Q. \end{cases} \tag{3.5}$$

Once these equality probabilities are obtained, the model can be trained using cross-entropy as the loss function.

3.4.3 *A novel solution: Ordinal Binary Decomposition for CNNs*

For our ordinal approach, we decompose the original Q-class ordinal problem into $Q - 1$ binary decision problems, a strategy referred to as Ordinal Binary Decomposition (OBD). Each problem q involves determining whether $y \succ C_q$ conditioned on sample $\mathbf{x}$ ($1 \leq q < Q$), following the *OrderedPartitions* scheme from Table 1.1.

To adapt the model's outputs for this approach, we replace the final fully-connected block with $Q - 1$ separate fully-connected blocks. Each block contains a single output unit with sigmoid activation. Each of these $Q - 1$ output units aims to predict the probability $\mathrm{P}(y \succ C_q \mid \mathbf{x})$. This modification results in the creation of $Q - 1$ distinct models that share convolutional feature extraction parameters and are trained simultaneously.

3.4.3.1 *Error Correcting Output Codes as an output consensus method*

In the case of OBD models, the output is a vector $\mathbf{p} = (p_1, p_2, \ldots, p_{Q-1})$ of cumulative probabilities $p_q = \mathrm{P}(y \succ C_q \mid \mathbf{x})$. Therefore, the decision rule involves combining multiple outputs. However, these probabilities may not adhere to basic probability properties, such

as $\mathrm{P}(y \succ C_q) \geq \mathrm{P}(y \succ C_{q+1})$ and $\sum_{q=1}^{Q} \mathrm{P}(y = C_q) = 1$. Consequently, Eq. (3.5) cannot be applied as in the case of CLMs.

To address this issue, we employ a stable approach based on the ECOC framework. This approach considers the ideal output vector $\mathbf{v}(C_q)$ for each class C_q, defined as:

$$\mathbf{v}(C_q) = (c_1, \dots, c_{Q-1}), \tag{3.6}$$

$$c_j = \mathbb{1}\{C_q \succ C_j\}, \tag{3.7}$$

i. e. a vector with ones in positions corresponding to classes lower than C_q in the ordinal scale and zeros everywhere else. This results in the ideal output vector for a sample $\mathbf{x}$ with label $y = C_q$ being:

$$\mathbf{v}(C_q) = (c_1, \dots, c_{q-1}, c_q, \dots, c_{Q-1}) = (1, \dots, 1, 0, \dots, 0), \tag{3.8}$$

where, for instance, in a 4-class ordinal problem with labels C_1, C_2, C_3, and C_4, the ideal outputs are $\mathbf{v}(C_1) = (0,0,0)$, $\mathbf{v}(C_2) = (1,0,0)$, $\mathbf{v}(C_3) = (1,1,0)$, and $\mathbf{v}(C_4) = (1,1,1)$.

The decision rule aims to identify the ideal vector that minimizes the distance from the output vector $\mathbf{p} = (p_1, p_2, \dots, p_{Q-1})$ computed by the CNN model:

$$\hat{y} = \underset{C_1 \preceq C_q \preceq C_Q}{\arg\min} \left\| \mathbf{p} - \mathbf{v}(C_q) \right\|_2, \tag{3.9}$$

where $\|\cdot\|_2$ represents the L_2 norm. This choice of distance metric aligns with the loss function used in the optimization process.

As an example, consider a 4-class ordinal problem. For a given sample $\mathbf{x}$, if the model's output is a 3-dimensional vector $\mathbf{p} = (0.8,\ 0.3,\ 0.2)$, the distances to each ideal class vector would be calculated as follows:

$$\begin{aligned}
\|\mathbf{p} - \mathbf{v}(C_1)\|_2 &= \|(0.8,\ 0.3,\ 0.2) - (0,\ 0,\ 0)\|_2 = 0.77, \\
\|\mathbf{p} - \mathbf{v}(C_2)\|_2 &= \|(0.8,\ 0.3,\ 0.2) - (1,\ 0,\ 0)\|_2 = 0.17, \\
\|\mathbf{p} - \mathbf{v}(C_3)\|_2 &= \|(0.8,\ 0.3,\ 0.2) - (1,\ 1,\ 0)\|_2 = 0.57, \\
\|\mathbf{p} - \mathbf{v}(C_4)\|_2 &= \|(0.8,\ 0.3,\ 0.2) - (1,\ 1,\ 1)\|_2 = 1.17.
\end{aligned} \tag{3.10}$$

This process is illustrated in Figure 3.1. The vector closest to $\mathbf{p}$ is $\mathbf{v}(C_2)$, indicating that sample $\mathbf{x}$ would be assigned the class label $\hat{y} = C_2$.

For the OBD methodology, we replace categorical cross-entropy with the Mean Squared Error (*MSE*) loss, as it better aligns with the distance function used in the ECOC decision rule [2]:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{Q-1} \left(\mathbb{1}\{y_i \succ C_q\} - \mathrm{P}(y_i \succ C_q \mid \mathbf{x}_i) \right)^2. \tag{3.11}$$

An illustration comparing the four methodologies presented in this section is shown in Fig. 3.2.

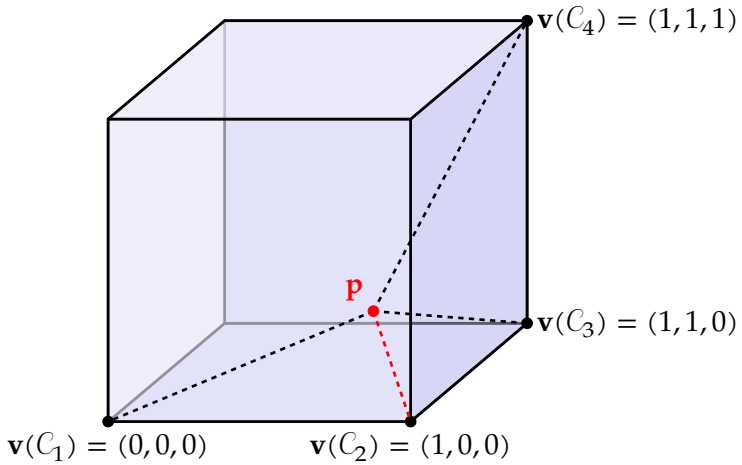


Figure 3.1: Visualization of the model output vector $\mathbf{p}$ (red dot) for sample $\mathbf{x}$ and its distances to the ideal class vectors (dashed lines) in a 3D graphical representation. Each dimension corresponds to one of the three model outputs. The closest vector is $\mathbf{v}(C_2)$ (marked in red), leading to the assignment of label C_2 for sample $\mathbf{x}$.

3.5 EXPERIMENT DESIGN

3.5.1 *Datasets*

To test and compare the four methodologies described previously, four different image ordinal regression tasks have been selected. The datasets for these tasks present a high level of class imbalance as an additional challenge. A sample of each class from all of them can be seen in Figs. 3.3 to 3.6 and the distribution of the class samples can be seen in Fig. 3.7.

3.5.1.1 *'Adience' face age estimation*

A set of 17 702 photos of people scraped from the web and pre-aligned to fit their face, categorized into 8 different age groups [29] of increasing value: 0 to 2 years, 4 to 6 years, 8 to 13 years, 15 to 20 years, 25 to 32 years, 38 to 43 years, 48 to 53 years, and 60 years and up.

3.5.1.2 *CBIS-DDSM*

A database of 2620 scanned film mammography studies curated from the larger Digital Database for Screening Mammography (DDSM) and each one assigned a Breast Imaging Reporting and Data System (BI-RADS) assessment [55] by a trained mammographer. The assessment is done on a scale from 0 to 5 according to the standard for a total of 6 classes (there are no cases of class 6 as there is no biopsy information). For this dataset, before being resized, all images were cropped into a square centred around the Region of Interest (RoI) of the lesion.

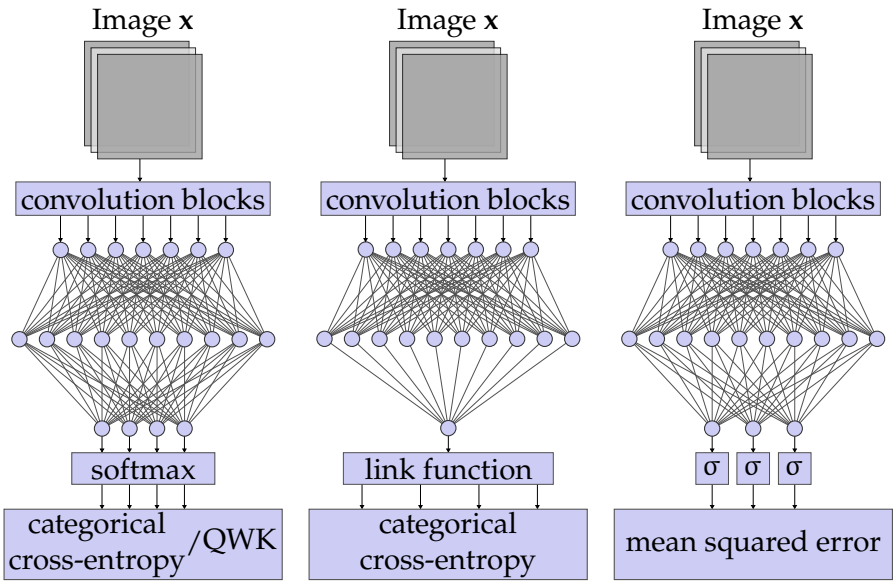


Figure 3.2: Visual comparison of the four methodologies. From left to right: the baseline nominal architecture (using both $\mathcal{L}_{CE}$ and $\mathcal{L}_{\kappa}$ as the loss function), CLM, and our proposed approach, OBD.

3.5.1.3 *Diabetic Retinopathy diagnosis*

A collection of 53 569 high-resolution retina images rated by a clinician on the presence of Diabetic Retinopathy (DR), an eye disease present in a large proportion of diabetes patients, on a scale from 0 (no DR) to 4 (proliferative DR) for a total of 5 classes [1].

3.5.1.4 *Herlev Pap Smear Dataset*

917 images of single Pap smear cells classified by doctors and technicians into 7 different classes, 3 of them normal from different parts of the cervix (242 images in total) and 4 of them abnormal in different stages of dysplasia (675 images in total) [47]. These are condensed into 4 ordinal classes, following the Bethesda System standard [67].

3.5.2 *Validation scheme*

The following four methodologies are compared:

- A baseline architecture using $\mathcal{L}_{CE}$ as the loss function for training described in Section 3.3, referred to as 'Nominal'.
- The same architecture, but using $\mathcal{L}_{\kappa}$ as the loss function for training described in Section 3.4.1, referred to as 'QWK'.
- The CLM approach described in Section 3.4.2, referred to as 'CLM'.
- The OBD approach using the ECOC decision rule described in Section 3.4.3, referred to as 'OBD'.

1 https://www.kaggle.com/c/diabetic-retinopathy-detection/data

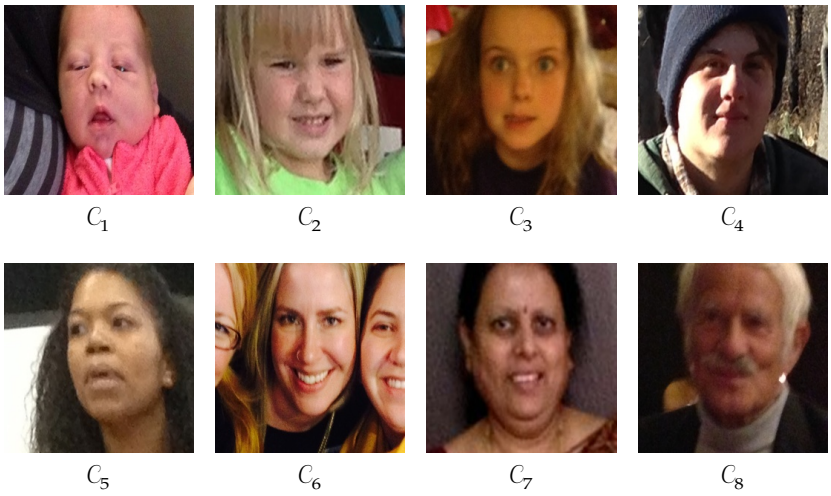


Figure 3.3: Sample image from each class of the Adience dataset

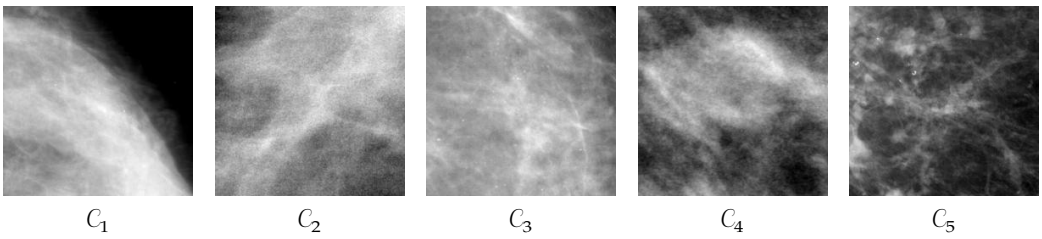


Figure 3.4: Sample image from each class of the CBIS-DDSM dataset

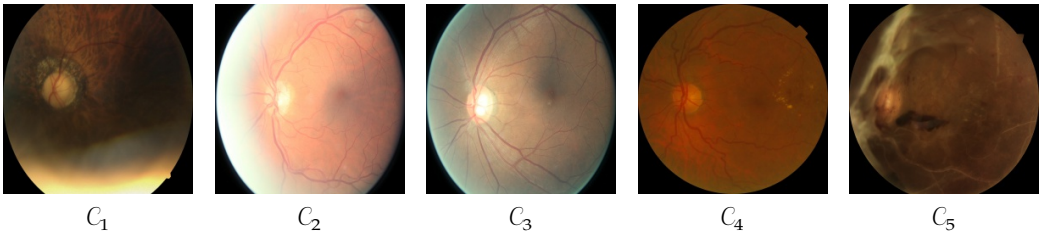


Figure 3.5: Sample image from each class of the Retinopathy dataset

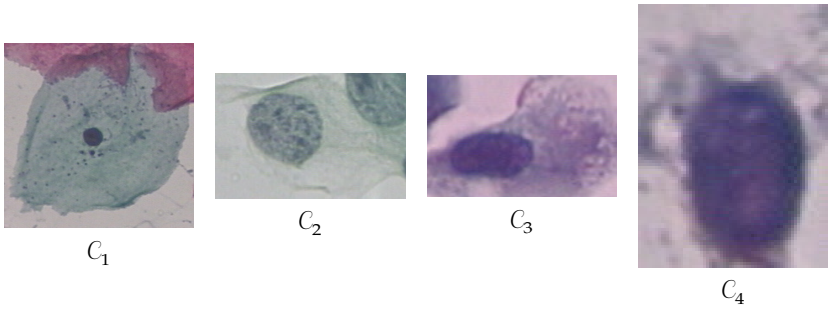


Figure 3.6: Sample image from each class of the Herlev dataset

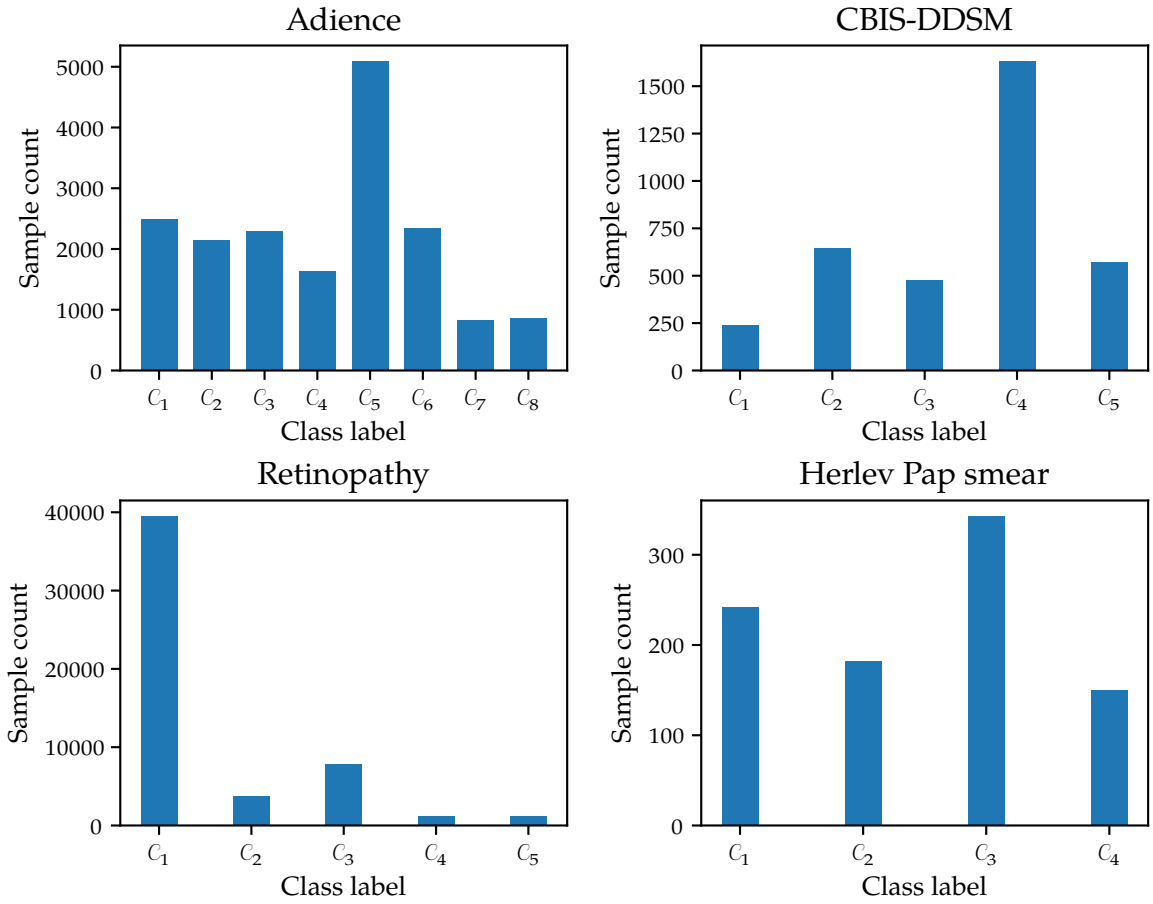


Figure 3.7: Distribution of class labels in the datasets

All of these methodologies are tested with four different architectures: VGG11 [87], ResNet18 [40], MobileNetV3 [44] and ShuffleNetV2 [60], which yields a total of sixteen different experiments for each of the four datasets.

In order to obtain a statistically significant result to test our hypothesis, each experiment is repeated 30 times on 30 different holdout splits of the original dataset, where 80 % of samples are used for training, 10 % are used for validation and early stopping and 10 % are used for model evaluation. This split is performed in a stratified fashion, preserving the original proportion of the class labels of the original dataset in the subsets.

3.5.3 *Training scheme*

In all experiments, weights are initialized randomly using the He initialization scheme [41]. They are then adjusted using the Adam optimization method [52] with a learning rate of $\eta = 10^{-4}$.

In the case of VGG11, both dropout ($p = 0.5$) and L_2 regularization (with a weight of 5×10^{-4}) are applied only in the fully connected layers as in the original paper [87]. For ResNet18, batch normalization is applied after every convolution operation and L_2 penalty (with a weight of 10^{-4}) is added to all mappings [40]. The number of trainable parameters for each model is available on Table 3.1.

Architecture	Adience	CBIS-DDSM	Retinopathy	Herlev
VGG11	132 871 344 (4084 MiB)	132 868 341 (4084 MiB)	132 868 341 (4084 MiB)	132 867 340 (4084 MiB)
ResNet18	11 697 520 (2729 MiB)	11 694 517 (2729 MiB)	11 694 517 (2729 MiB)	11 693 516 (2729 MiB)
MobileNetV3	5 491 040 (4838 MiB)	5 488 037 (4838 MiB)	5 488 037 (4838 MiB)	5 487 036 (4838 MiB)
ShuffleNetV2	7 402 004 (4114 MiB)	7 399 001 (4114 MiB)	7 399 001 (4114 MiB)	7 398 000 (4114 MiB)

Table 3.1: Number of trainable parameters and total memory size of the trained models for each architecture and dataset

In order to help overcome the class imbalance, class weighting is applied to the loss function based on the proportion of training samples for each class. The weight of a sample with class label C_q is determined as:

$$w_q = \frac{\exp(-CN_q)}{\sum_{c=1}^{Q} \exp(-CN_c)}, \tag{3.12}$$

where N_q is the number of training samples with class label C_q and $C = 3 \times 10^{-5}$ is a constant that has been tuned manually based on validation performance.

Model weights are updated in batches of 72 training samples and loss performance is monitored on both training and validation for a maximum of 200 full epochs. If validation performance does not increase for 20 epochs, training is halted and the best performing parameters over the validation set are restored.

3.5.4 *Performance metrics*

The traditional performance metric in classification tasks is the Correct Classification Rate (*CCR*). However, given that all four datasets present a very high class imbalance, *CCR* is not a representative measure of model performance: for example, in the case of the Retinopathy dataset, a dummy classifier that always assign the majority class label C_1 would obtain a *CCR* of 73 %.

In order to monitor global per-class performance, metrics such as the Average Area Under the ROC Curve (*AvAUC*) and minimum sensitivity (*MinS*) will also be included.

Also, for ordinal regression problems, rank agreement metrics including the Root Mean Squared Error (*RMSE*), Spearman's rank correlation coefficient (r_s) and the Quadratic Weighted Cohen's Kappa (κ) have been selected as well for evaluation.

All of these metrics are fully described in Section 1.2.5.

3.6 RESULTS

The average of the training curves over all 30 repetitions is shown in Figs. 3.8 to 3.11. Note how the QWK methodology fails to converge when used in conjunction with the VGG11 architecture: the high depth of this architecture makes the gradients disappear in the

back-propagation phase of training. All the other architectures tested implement residual paths into the network, allowing them to avoid this problem [102]. Note how the OBD methodology does not alter the depth of the CNN model, so it will never cause this problem by itself.

In addition, average training times for the different models are reported in Fig. 3.12. It can be seen that the CLM methodology takes the longest to converge.

The average experimental results for each experiment are shown in Tables 3.2 to 3.5.

For the Adience dataset (Table 3.2), the OBD methodology is able to outperform the other three with respect to *CCR*, *AvAUC* and *MinS*, with a close second place in *RMSE*, r_s and κ behind CLM.

For the CBIS-DDSM dataset (Table 3.3), OBD is ahead regarding *RMSE* and second regarding all other metrics but *MinS* behind both the Nominal and CLM approach.

As for the Retinopathy dataset (Table 3.4), OBD is able to outperform the rest in *MinS*, *RMSE* and κ, being second in all other except *CCR*.

Finally, for the Herlev dataset (Table 3.5), even though close to the rest, it only achieves the best mean results in *RMSE*.

In general, it can be observed that OBD rarely fails catastrophically regarding sensitivity, which the other two ordinal methodologies (QWK and CLM) do. This is reflected in the amount of cases where the *MinS* drops to zero, meaning that in every repetition of the experiment at least one of the classes was ignored. Even when OBD is not the clear winner in average, it is usually never far behind. To formalize this observation further statistical analysis is needed.

3.6.1 *Statistical analysis*

To determine the statistical significance of the mean differences observed for each classifier, each architecture and each dataset, we have carried out a parametric Analysis of Variance (ANOVA) test [34, 35] for each of the evaluated metrics. The three factors considered for the experimental design are: (i) the database (Adience, CBIS-DDSM, Retinopathy and Herlev), (ii) the CNN network architecture (VGG11, ResNet18, MobileNetV3 and ShuffleNetV2) and (iii) the methodology (Nominal, QWK, CLM and OBD).

For each combination of these three factors we have repeated the experiment 30 times with different data splits and weight initialization seeds. We have tested, using the Kolmogorov-Smirnov test [63] for all metrics mentioned in Section 3.5.4, whether the null hypothesis stating that the results are drawn from a normal distribution cannot be rejected (for a significance level of $\alpha = 0.05$). This is true for all metrics except *MinS*, namely the *CCR*, *AvAUC*, *RMSE*, r_s and κ. Only these metrics will be considered for the subsequent analysis, given that ANOVA is a parametric test and can only be applied to normally distributed variables.

After this, ANOVA is performed for these five metrics, the results of which can be seen in Tables 3.6 to 3.10. According to this analysis, for all normally distributed metrics there exist significant differences in the mean value (for a significance level of $\alpha = 0.05$) concerning the three individual factors of dataset, architecture and methodology (all p-values < 0.001). Then, we also found significant interactions between all the pairs of factors (all p-values < 0.001) and between all three single factors (all p-values < 0.001). This shows that:

1. the impact of the architecture and the methodology varies across datasets,

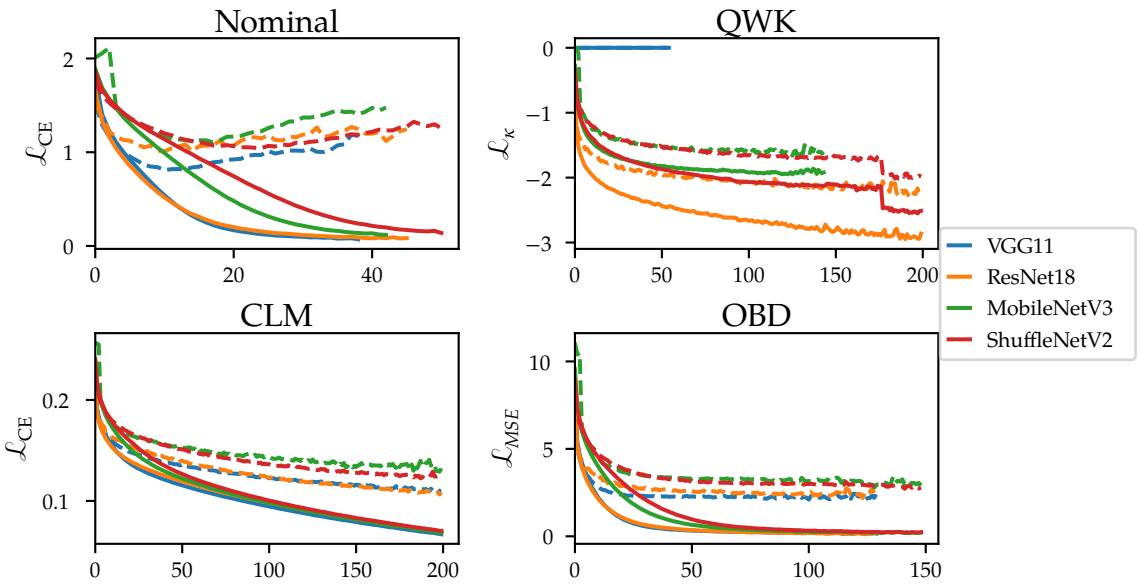


Figure 3.8: Average training curves for each model and methodology when applied to the Adience dataset. Train loss shown as solid lines and validation loss as dashed lines. The average is computed over all executions which reach the corresponding iteration

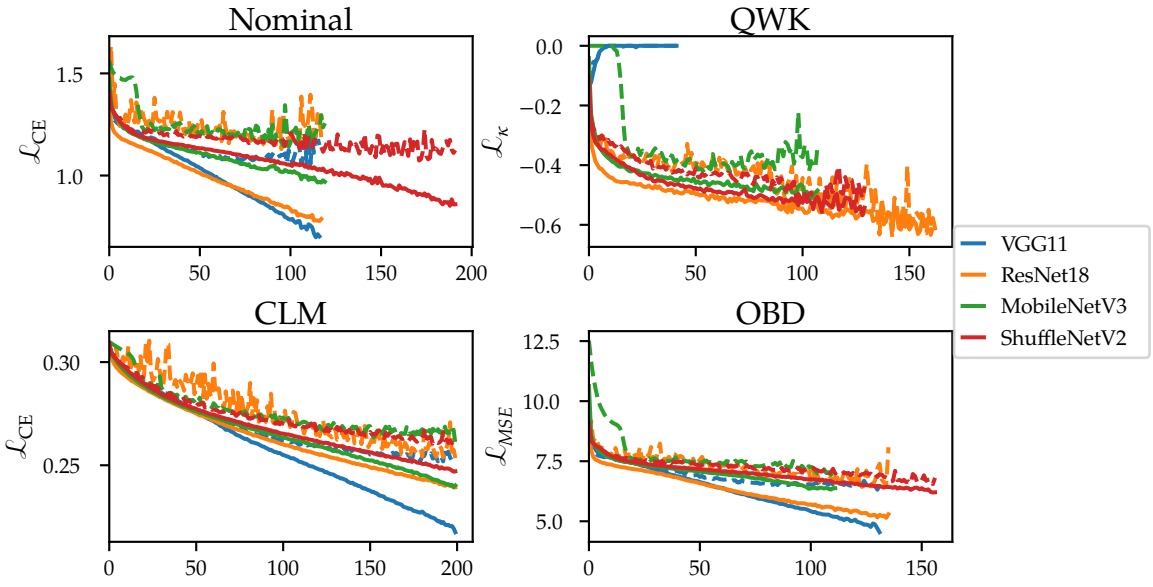


Figure 3.9: Average training curves for each model and methodology when applied to the CBIS-DDSM dataset. Train loss shown as solid lines and validation loss as dashed lines. The average is computed over all executions which reach the corresponding iteration

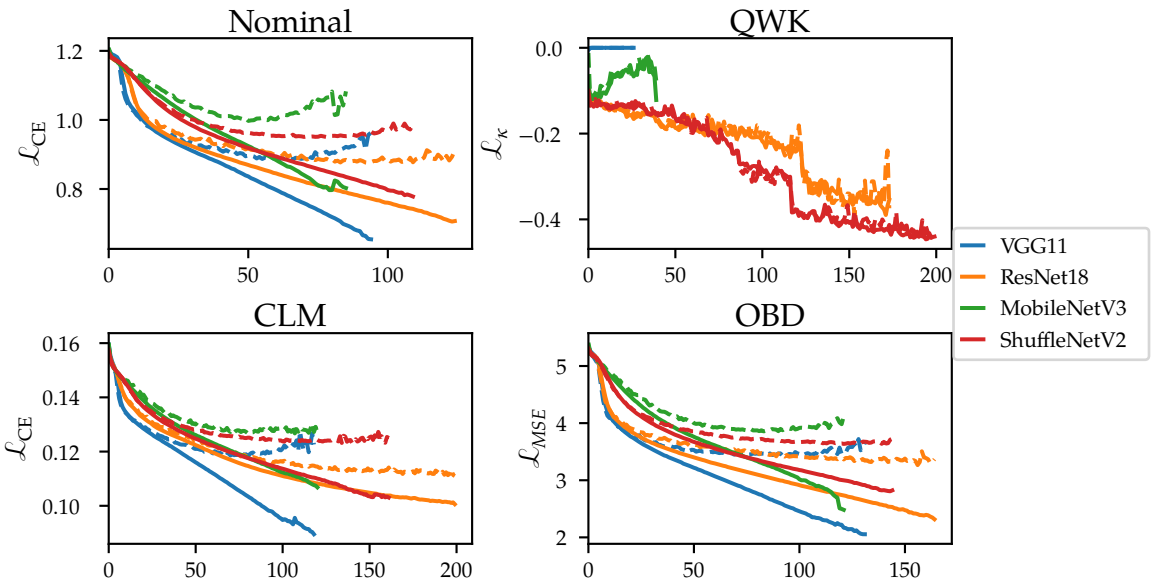


Figure 3.10: Average training curves for each model and methodology when applied to the Diabetic Retinopathy dataset. Train loss shown as solid lines and validation loss as dashed lines. The average is computed over all executions which reach the corresponding iteration

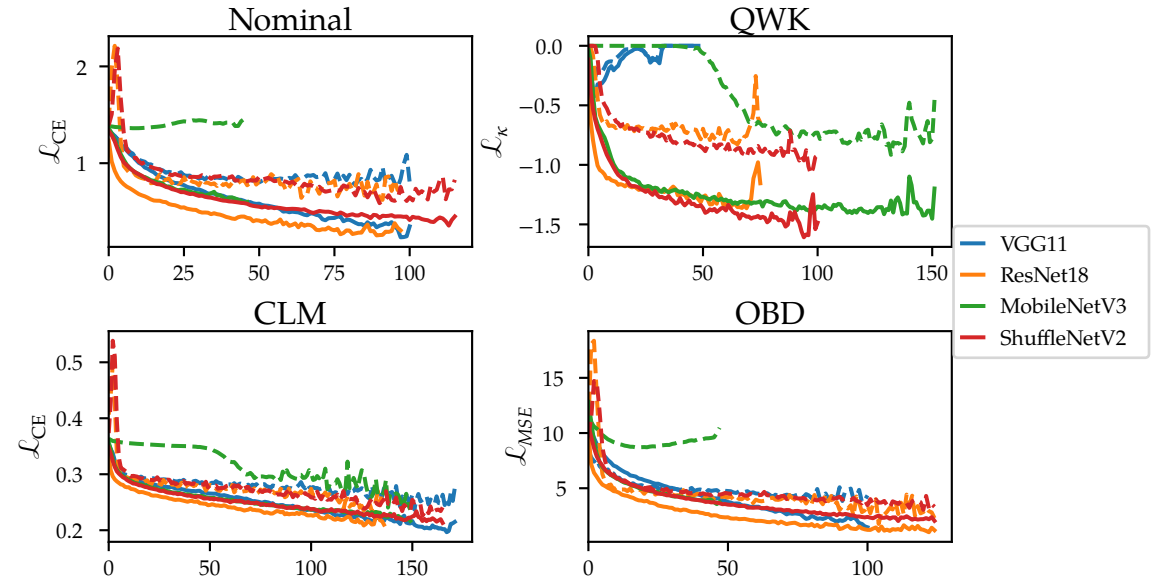


Figure 3.11: Average training curves for each model and methodology when applied to the Herlev dataset. Train loss shown as solid lines and validation loss as dashed lines. The average is computed over all executions which reach the corresponding iteration

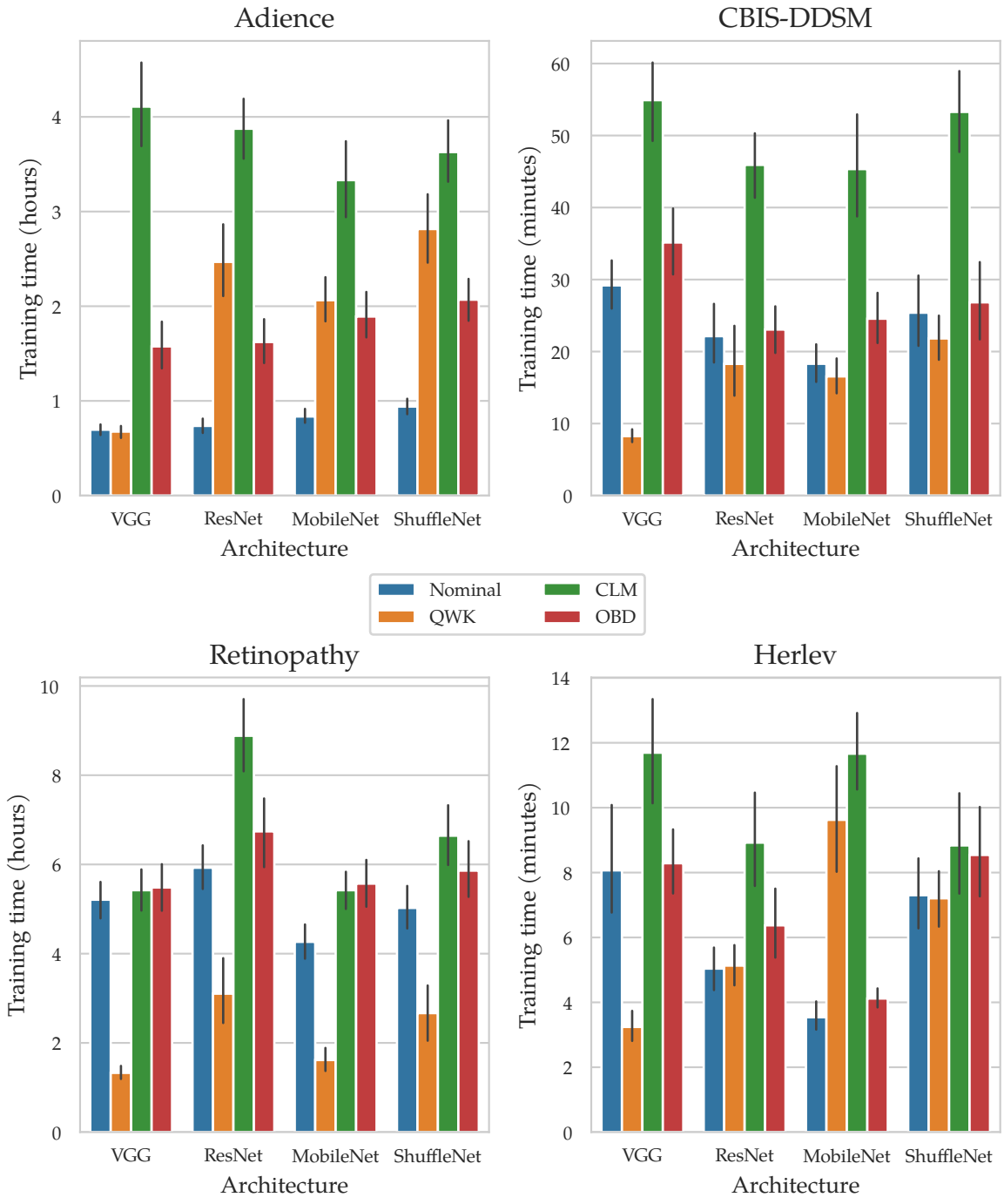


Figure 3.12: Average training time per dataset, architecture and methodology. Error bars indicate the 95 % confidence interval.

	Nominal		QWK		CLM		OBD	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
CCR (↑)	0.6808	0.0464	0.4210	0.1991	*0.7215*	0.0492	**0.7427**	0.0338
AvAUC (↑)	0.9331	0.0180	0.6877	0.1137	*0.9351*	0.0156	**0.9377**	0.0138
MinS (↑)	*0.3239*	0.1159	0.0000	0.0000	0.1221	0.1577	**0.4865**	0.0779
RMSE (↓)	0.9911	0.1175	1.5984	1.1095	**0.7676**	0.1006	*0.8159*	0.0883
r_s (↑)	0.8605	0.0314	0.6436	0.3699	**0.9128**	0.0230	*0.9025*	0.0209
κ (↑)	0.8690	0.0308	0.6440	0.3726	**0.9200**	0.0217	*0.9110*	0.0197

Table 3.2: Mean results for the Adience dataset. Metrics to maximize are marked with (↑) and metrics to minimize with (↓). Best and second best results are highlighted in bold and italics, respectively

	Nominal		QWK		CLM		OBD	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
CCR (↑)	**0.5672**	0.0246	0.3394	0.1099	0.5411	0.0315	*0.5558*	0.0239
AvAUC (↑)	**0.7568**	0.0345	0.5613	0.0402	0.6847	0.0346	*0.6987*	0.0400
MinS (↑)	**0.0106**	0.0275	*0.0065*	0.0182	0.0000	0.0000	0.0003	0.0038
RMSE (↓)	1.1572	0.0435	1.4035	0.3038	*1.1558*	0.0459	**1.1035**	0.0415
r_s (↑)	0.4180	0.0574	0.3035	0.0931	**0.4533**	0.0530	*0.4311*	0.0559
κ (↑)	0.3733	0.0563	0.3111	0.1073	**0.4121**	0.0543	*0.3837*	0.0595

Table 3.3: Mean results for the CBIS-DDSM dataset. Metrics to maximize are marked with (↑) and metrics to minimize with (↓). Best and second best results are highlighted in bold and italics, respectively

	Nominal		QWK		CLM		OBD	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
CCR (↑)	**0.7638**	0.0156	0.4216	0.1864	*0.7638*	0.0152	0.7238	0.0198
AvAUC (↑)	**0.8152**	0.0230	0.5105	0.0206	0.8053	0.0208	*0.8085*	0.0154
MinS (↑)	*0.0094*	0.0098	0.0000	0.0000	0.0000	0.0000	**0.1370**	0.0270
RMSE (↓)	0.8857	0.0548	1.0957	0.1803	*0.8591*	0.0478	**0.8396**	0.0392
r_s (↑)	0.5021	0.0555	0.0756	0.0585	**0.5134**	0.0490	*0.5064*	0.0386
κ (↑)	0.5636	0.0594	0.0620	0.0491	*0.5754*	0.0502	**0.5929**	0.0402

Table 3.4: Mean results for the Retinopathy dataset. Metrics to maximize are marked with (↑) and metrics to minimize with (↓). Best and second best results are highlighted in bold and italics, respectively

2. the architecture significantly affects performance,
3. the effect of the methodology is affected by the CNN architecture (that is, some architectures are better suited for each methodology), and
4. the methodology alone affects the performance.

Given the ANOVA results identify a significant difference with respect to the Methodology factor, we now analyse the magnitude of these differences. We perform a post-hoc Tukey's honestly significant difference test [94] on each of these metrics. This test groups the methodologies into groups of similar performance, where each group is significantly different than the rest. The test results are shown in Table 3.11 and graphically in Fig. 3.13.

From these results it can be observed that the OBD methodology is always present in the group with significantly best results, that is, it is never outperformed significantly by any of the other three methodologies regarding any of the studied metrics. It shows similar performance to the Nominal methodology in *CCR*, *AvAUC*, r_s and κ and it also show similar performance to the CLM methodology in *AvAUC*, r_s and κ. In this last case, it is important to note that it achieves this level of performance while reducing the convergence time with respect to the CLM methodology. Furthermore, it significantly outperforms every other methodology with regard to *RMSE*.

3.7 CONCLUSIONS

This chapter introduced a novel ordinal CNN architecture based on Ordinal Binary Decomposition along with a decision scheme using Error Correcting Output Codes. The results demonstrate that this approach obtains similar and significantly better results compared to a purely nominal method and two other existing ordinal techniques, particularly in highly imbalanced scenarios, such as medical and web-scraped datasets. Notably, the proposed OBD methodology enhances the *RMSE* performance without compromising any of the other studied metrics. Importantly, this methodology is easily adaptable for various ordinal tasks.

While the tested architectures represent well-established and high-performing models, the approach remains versatile, accommodating the integration of different and more innovative models. This provides a versatile tool for classification tasks that leverage ordinal information. Furthermore, these modifications do not inflate the number of network parameters, memory consumption, or significantly extend training time.

This versatility will allow applying this methodology to a wildly different input type like 3D images in the following chapter, as the required modifications primarily affect the latter stages of the network, allowing for arbitrary input shapes. Addressing class imbalance challenges requires further investigation. Improved class balancing approaches, beyond simple loss weighting, could be employed, including data augmentation methods that account for ordinal information to enhance model performance.

	Nominal		QWK		CLM		OBD	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
CCR (↑)	**0.6028**	0.1366	0.5054	0.0933	0.4047	0.0401	*0.6013*	0.1367
AvAUC (↑)	*0.7932*	0.1753	0.7348	0.1028	**0.8416**	0.0350	0.7921	0.1482
MinS (↑)	**0.3844**	0.2392	0.2590	0.1905	0.0000	0.0000	*0.3585*	0.2396
RMSE (↓)	*0.9111*	0.1823	0.9549	0.2186	1.1566	0.1185	**0.8595**	0.2071
r_s (↑)	0.5186	0.3073	*0.5719*	0.1937	**0.6331**	0.0743	0.5441	0.3215
κ (↑)	0.5196	0.3080	*0.5629*	0.1973	**0.5866**	0.0769	0.5448	0.3221

Table 3.5: Mean results for the Herlev Pap smear dataset. Metrics to maximize are marked with (↑) and metrics to minimize with (↓). Best and second best results are highlighted in bold and italics, respectively

Source	Sum of Squares	Degrees of freedom	F-ratio	Sig. level
Dataset (D)	9.79	3	1385.33	<0.001
Architecture (A)	1.37	3	193.54	<0.001
Methodology (M)	17.70	3	2505.40	<0.001
D.A Interaction	2.70	9	127.28	<0.001
D.M Interaction	7.59	9	358.24	<0.001
A.M Interaction	1.82	9	85.92	<0.001
D.A.M Interaction	6.75	27	106.13	<0.001
Residual	4.37	1856		

Table 3.6: ANOVA III table for the *CCR* results

Source	Sum of Squares	Degrees of freedom	F-ratio	Sig. level
Dataset (D)	10.22	3	4413.39	<0.001
Architecture (A)	1.70	3	734.01	<0.001
Methodology (M)	13.50	3	5832.93	<0.001
D.A Interaction	2.17	9	312.51	<0.001
D.M Interaction	3.20	9	460.41	<0.001
A.M Interaction	2.17	9	313.11	<0.001
D.A.M Interaction	2.69	27	129.03	<0.001
Residual	1.43	1856		

Table 3.7: ANOVA III table for the *AvAUC* results

Source	Sum of Squares	Degrees of freedom	F-ratio	Sig. level
Dataset (D)	22.25	3	390.27	<0.001
Architecture (A)	14.12	3	247.62	<0.001
Methodology (M)	35.49	3	622.46	<0.001
D.A Interaction	15.64	9	91.43	<0.001
D.M Interaction	34.82	9	203.60	<0.001
A.M Interaction	57.49	9	336.13	<0.001
D.A.M Interaction	60.50	27	117.90	<0.001
Residual	35.27	1856		

Table 3.8: ANOVA III table for the *RMSE* results

Source	Sum of Squares	Degrees of freedom	F-ratio	Sig. level
Dataset (D)	59.21	3	6584.14	<0.001
Architecture (A)	6.12	3	680.42	<0.001
Methodology (M)	15.23	3	1694.04	<0.001
D.A Interaction	10.58	9	392.27	<0.001
D.M Interaction	9.76	9	361.70	<0.001
A.M Interaction	13.66	9	506.41	<0.001
D.A.M Interaction	12.62	27	155.87	<0.001
Residual	5.56	1856		

Table 3.9: ANOVA III table for the r_s results

Source	Sum of Squares	Degrees of freedom	F-ratio	Sig. level
Dataset (D)	59.76	3	6744.61	<0.001
Architecture (A)	6.24	3	704.17	<0.001
Methodology (M)	16.21	3	1829.51	<0.001
D.A Interaction	10.86	9	408.55	<0.001
D.M Interaction	14.75	9	555.08	<0.001
A.M Interaction	14.56	9	547.65	<0.001
D.A.M Interaction	12.30	27	154.22	<0.001
Residual	5.48	1856		

Table 3.10: ANOVA III table for the κ results

CCR

Methodology	Subsets		
	1	2	3
QWK	0.422		
CLM		0.608	
Nominal			**0.654**
OBD			**0.656**
p-values	1	1	0.994

AvAUC

Methodology	Subsets	
	1	2
QWK	0.624	
OBD		**0.809**
CLM		**0.817**
Nominal		**0.825**
p-values	1	0.142

RMSE

Methodology	Subsets		
	1	2	3
QWK	1.263		
Nominal		0.986	
CLM		0.985	
OBD			**0.905**
p-values	1	1	1

r_s

Methodology	Subsets		
	1	2	3
QWK	0.399		
Nominal		0.575	
OBD		0.596	**0.596**
CLM			**0.628**
p-values	1	0.545	0.185

κ

Methodology	Subsets	
	1	2
QWK	0.395	
Nominal		**0.581**
OBD		**0.608**
CLM		**0.624**
p-values	1	0.051

Table 3.11: Results of the Tukey's HSD test for all tested metrics. Methodologies are grouped such that the elements within a subset are not significantly different, while the differences between each group are significant. The first subset contains the worst methodologies, while the last subset groups the best methodologies. The best performing subset is highlighted in bold

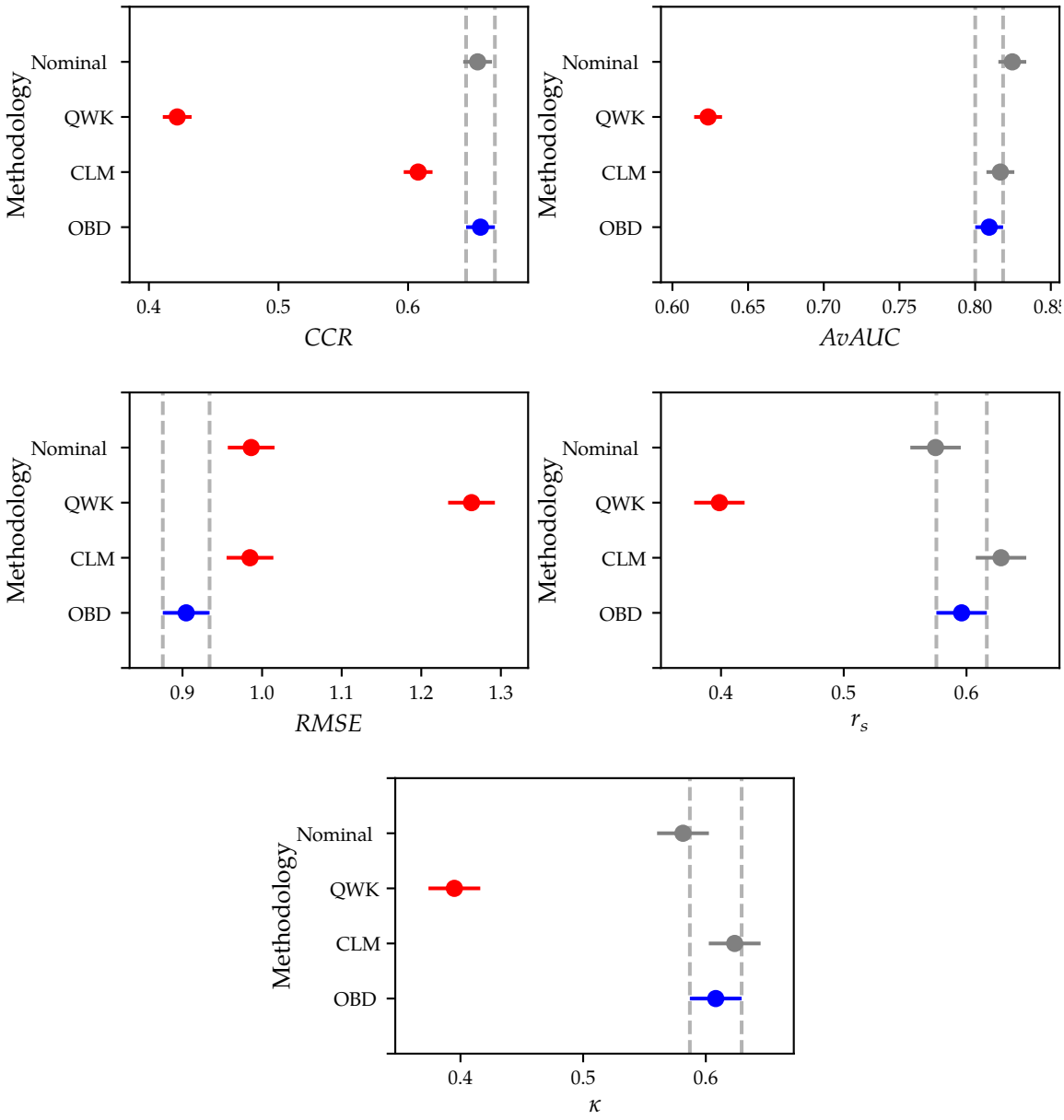


Figure 3.13: Confidence intervals for the Tukey's HSD test for al tested metrics. OBD methodology highlighted in blue, all other overlapping methodologies in grey and the rest in red. All overlapping methodologies form groups corresponding to those in Table 3.11.

4

COMPUTER-AIDED DIAGNOSIS FOR PARKINSON'S DISEASE: AN APPLICATION OF ORDINAL DATA AUGMENTATION FOR 3D IMAGE IMBALANCED DATASETS

Having proposed in the previous chapter an output architecture for CNNs suited to ordinal regression tasks, we proceed to tackle a real problem using a novel dataset.

We have had the opportunity to work alongside the Nuclear Medicine unit from the Hospital Universitario 'Reina Sofía', where doctors deal with the task of evaluating the neurological damage of Parkinson's disease (PD) patients from a volumetric scan of the brain. The evaluation consists on assigning an ordinal label in a discrete scale, which maps perfectly to the ordinal regression framework.

However, this task presents a unique series of challenges that have to be overcome, namely (a) the volumetric (that is, 3D) nature of the images, (b) a relatively low sample size and (c) an acute class representation imbalance.

This chapter is dedicated to the development of a comprehensive methodology including a *native 3D CNN architecture* and a *data augmentation procedure* capable of tackling this task. Two versions of this methodology are presented: one where the ordinal information is ignored as well as another one where it is *exploited in both the network output and the data augmentation method*. In this last regard, several different configurations are considered to find the best performing one. The nominal scheme is then compared to the ordinal approach in order to show the better performance of the latter.

ASSOCIATED PUBLICATION: **Javier Barbero-Gómez**, Pedro-Antonio Gutiérrez, Víctor-Manuel Vargas, Juan-Antonio Vallejo-Casas and César Hervás-Martínez. 'An Ordinal CNN Approach for the Assessment of Neurological Damage in Parkinson's Disease Patients'. In: *Expert Systems with Applications* 182 (15th Nov. 2021), p. 115271. DOI: 10.1016/j.eswa.2021.115271.

JCR (2021): 8.655. Ranking position in Computer Science, AI: 21/145 (Q1)

4.1 RELATED WORK

The diagnosis of PD has been studied extensively in the literature, with some already existing work on applying ML techniques.

PD is a neurodegenerative disorder that primarily affects the nervous system and manifests with motor-related symptoms such as tremors, gait disturbances, slowness, and walking difficulties. Additionally, patients may experience symptoms related to sleep, emotions, and sensory functions. The societal cost of PD increases as the disease progresses, with a substantial portion allocated to patient care and nursing home expenses. In the UK alone, the annual cost has been estimated to range from £ 445 million to £ 3.3 billion [33].

Assessing the severity of neurological damage in PD patients is critical for appropriate treatment. Administering an excessively high dose of levodopa, the most common medication for PD, can exacerbate symptoms in the long term [93]. Physicians assess patients' motor capabilities through observations [61] and employ imaging techniques such as Mag-

netic Resonance Imaging (MRI) [61] and nuclear tomography methods like single-photon emission computed tomography (SPECT) or positron emission tomography (PET) [3].

In recent years, there has been a growing interest in applying ML techniques to such medical images, eliminating the need for prior assumptions about RoI or relevant areas for the given task. These methods autonomously determine where and what to examine in images, relying solely on data previously labelled by medical professionals. Popular methods address binary classification tasks (e. g. healthy control or disease) [103] or nominal classification with multiple disease categories [1].

Several datasets related to PD are accessible online for research purposes, such as the Parkinson's Progression Markers Initiative (PPMI)[1] and the LRRK2 Cohort Consortium (LCC)[2]. All of the available datasets deal only with binary or nominal diagnostic labels.

4.1.1 *Data augmentation*

In classification tasks, especially in the medical domain, dealing with imbalanced data is a common challenge, given the prevalence of healthy cases compared to the relatively rare occurrence of diseases. Moreover, the process of gathering and accurately labelling medical data is often costly and time-consuming. In such scenarios, the use of data augmentation techniques becomes essential to enhance the performance of ML models.

One of the fundamental strategies for augmenting spatial data such as medical images, involves operations like image translation, rotation, flipping, and cropping [70]. The selection of specific augmentation techniques depends on the nature of the task at hand. For instance, tasks like object detection, including anomaly or lesion detection, can benefit from the utilization of cropped RoI as augmented samples [76].

While classic techniques like Synthetic Minority Oversampling (SMOTE) [16, 78] perform well with low-dimensional data, more advanced techniques such as Autoencoders [43] and Generative Adversarial Networks (GANs) [37] are capable of harnessing convolutional operations, improving performance and efficiency with spatial data. However, it's worth noting that these advanced techniques require a substantial amount of training data and tuning efforts to avoid challenges like mode collapse.

To meet the demand for larger datasets, recent efforts have introduced more sophisticated data augmentation methods for medical data. For instance, the authors of [81] combine GANs with Markov Random Field models to augment 3D functional MRI data from multiple subjects, resulting in enhanced multiclass classification performance.

Additionally, data augmentation techniques have evolved to consider the ordinal information within class labels to improve the generation of synthetic data. One family of such methods, presented by [71], are the Ordinal Graph-based Oversampling (OGO) methods. These techniques involve constructing a graph that captures the latent manifold structure in the data by leveraging ordinal information in the labels. Subsequently, the edges of this graph are employed to generate synthetic samples, akin to the principles of SMOTE.

Furthermore, established techniques like SMOTE and OGO can be adapted to work with spatial data. This adaptation involves initially training a CNN model to learn a projection from high-dimensional data to a lower-dimensional space. Subsequently, traditional data augmentation methods can be applied to the resulting low-dimensional data.

1 https://www.michaeljfox.org/news/parkinsons-progression-markers-initiative-ppmi
2 https://www.michaeljfox.org/news/lrrk2-cohort-consortium

4.2 GOALS

To the best of our knowledge, prior research has not explored the application of ML methods to assess the severity of brain damage from a patient's brain SPECT 3D image. These methods have the potential to assist healthcare professionals in diagnosing and treating conditions like PD and other parkinsonisms through computer-aided diagnosis (CAD) systems. Moreover, they can contribute to alleviating the public health costs associated with these diseases.

The primary objectives of this chapter relate to objectives 2 and 3 from Section 2.2. They are the following:

1. Investigating the potential enhancement in classification performance by leveraging ordinal label information.
2. Adapting classical data augmentation and class balancing techniques for spatial 3D data.
3. Evaluating the methodologies developed in points 1 and 2 using a novel and extensive database of SPECT images obtained from the Hospital Universitario 'Reina Sofía' (Córdoba, Spain).
4. Examining the possibility of refining the data augmentation methodology presented by [71] by implementing a more appropriate probability distribution for generating synthetic samples at class boundaries.

4.3 TASK DESCRIPTION

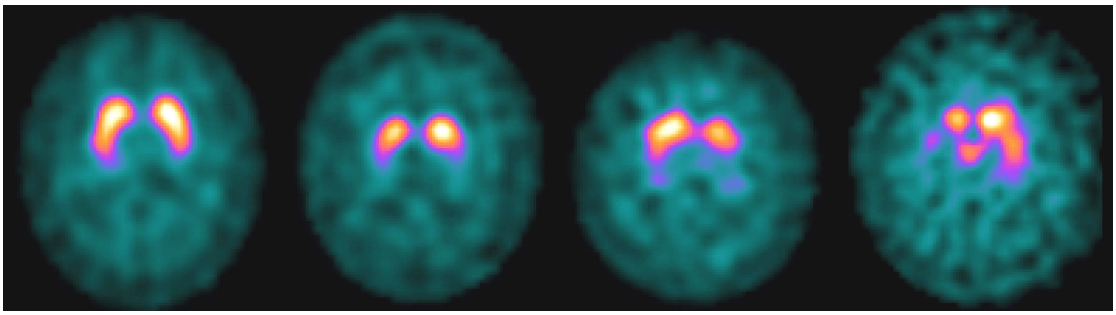
The task at hand is the evaluation of the level of neurological damage in PD patients using SPECT imaging of the brain. This technique uses a specific radiopharmaceutical called *radioligand* that binds to a specific targeted receptor. The radiation emitted by this chemical can then be detected and an image can be formed, highlighting regions with high activity of the receptor. In the case of PD, doctors are interested in the dopaminergic activity inside the brain as one of the main markers. Ioflupane (^{123}I), commercially known as DaTscan, is a neuroimaging drug frequently used to evaluate dopaminergic activity in the nigrostriatal dopaminergic pathway, especially in the early stages of the disease [22]. It is injected into the patient's bloodstream before conducting a SPECT scan to visualize the brain's dopaminergic activity. Assessing this damage often necessitates significant expertise.

4.3.1 *Dataset*

The data collected for this experimentation consists of 508 3D SPECT images of PD patients, all of them collected and labelled in a scale from 1 to 4 according to their level of neurological damage by experts in the Nuclear Medicine clinical unit at the Hospital Universitario 'Reina Sofía'.

Of these, 314 (61.8 %) are labelled as healthy patients (C_1), 42 (8.3 %) present a slight alteration (C_2), 52 (10.2 %) show a more advanced alteration (C_3) and 100 (19.7 %) show a severe alteration (C_4). Like it is common in medical diagnosis datasets, it presents an intense class imbalance, with more than 60 % of samples corresponding to healthy patients.

The gradual nature of the degenerative process makes this an ordinal regression task, which allows us to apply ordinal-specific techniques in order to exploit this information.



(a) C_1 (b) C_2 (c) C_3 (d) C_4

Figure 4.1: Transverse sections of example images from the dataset for each class.

To standardise the resolution, orientation and scale of the images, automatic linear registration is performed using the FMRIB's Linear Image Registration Tool (FLIRT) from the FMRIB Software Library (FSL) [88]. Specifically, all images were transformed to the MNI152 standard space [31] with a resolution of 2 mm using a T2 SPECT template, making the final images have a resolution of $91 \times 109 \times 91$ voxels.

A sample of this data can be seen in Fig. 4.1, including one image from each class.

4.4 MODEL ARCHITECTURE

The CNN model employed in this study features a modular architecture designed to effectively process brain SPECT 3D images. It comprises convolutional blocks composed of repeated layers which progressively reduce the image size while increasing the number of feature maps.

Each convolutional block is composed of a 3D convolution layer followed by a batch normalization layer. Parameters such as kernel size and stride for the convolution are subject to cross-validation during the training phase. The output from each block undergoes a LReLU activation function, chosen for its advantageous properties, including scale-invariance and 1-Lipschitz continuity [91, 97].

The low resolution feature maps generated by the convolutional blocks serve as inputs for a fully connected neuron layer. The number of neurons in this layer is another parameter subject to cross-validation during training, and it employs LReLU as the activation function. The final classification is computed by the output layer, and the model's weights are updated in the training phase using the Adam optimization algorithm [52] to align the outputs with the annotated labels.

Two distinct architectures, which are described in Chapter 3 will be tested: a conventional architecture designed for nominal classification (detailed in Section 3.3) and an ordinal architecture that takes into account the order of class labels (the OBD with ECOC consensus method as described in Section 3.4.3). Both architectures share the same convolutional structure, but they differ in their approach to computing the final output. An illustration of both can be seen in Fig. 4.2.

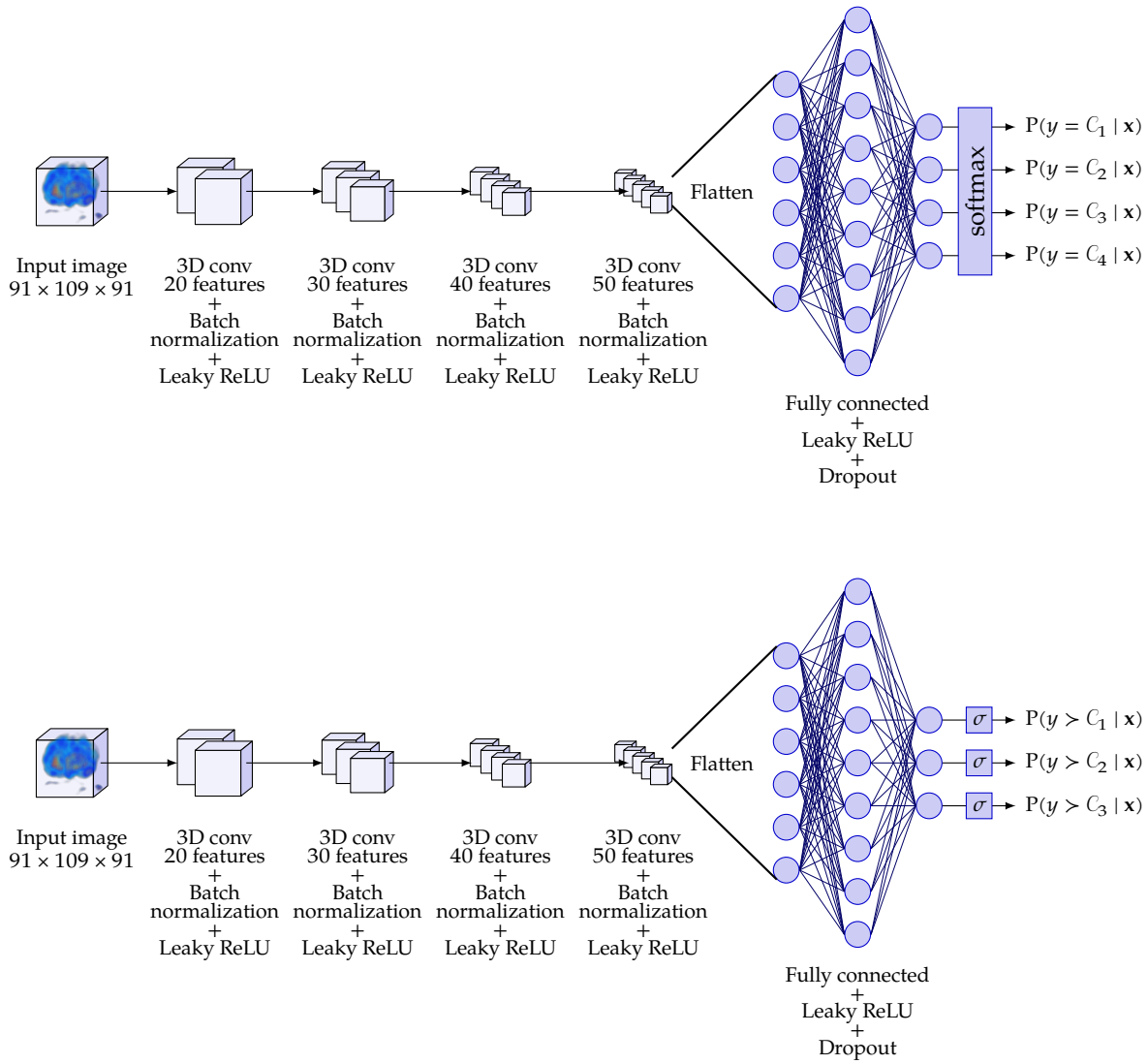


Figure 4.2: The two network architectures for classifying 3D brain SPECT images: conventional (above) and ordinal (below).

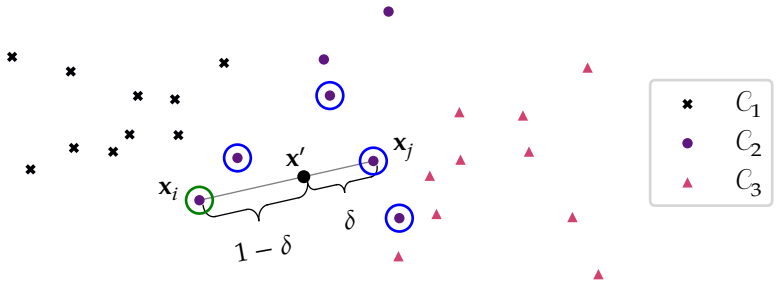


Figure 4.3: Example of the SMOTE procedure. First, $\mathbf{x}_i$ (highlighted in green) is selected, then $\mathbf{x}_j$ is selected from its neighbours (highlighted in blue) and finally $\mathbf{x}'$ (in black) is generated by interpolating between the two according to δ.

4.5 CLASS BALANCING

In scenarios characterized by class imbalance, such as medical diagnosis, where certain classes have significantly fewer samples than others, it is crucial to take specific measures during the training process to prevent biases that could compromise the model's ability to generalize effectively.

Various techniques, including class balancing, offer effective solutions to address this issue. These methods primarily aim to equalize the proportion of training samples from each class presented to the classifier during training.

4.5.1 *SMOTE*

The Synthetic Minority Oversampling technique, commonly known as SMOTE [16] is able to create new samples for the minority class (or any other class) by interpolating between similar real samples of said class directly in the feature space. It is particularly effective when dealing with datasets with a manageable number of features which take continuous values. Its procedure is as follows:

1. Let D_q be the subset of the original dataset D containing all the samples of class C_q to be augmented: $D_q = \{\mathbf{x}_i \mid (\mathbf{x}_i, y_i) \in D \wedge y_i = C_q\}$. Select a random sample $\mathbf{x}_i \in D_q$.
2. Based on some distance measure (e. g. the euclidean distance) select a second random sample $\mathbf{x}_j \in D_q$ from the same class in the neighbourhood of the k nearest samples to $\mathbf{x}_i$ of class C_q.
3. Draw a value δ from a uniform random distribution between 0 and 1: $\delta \sim U(0,1)$.
4. Generate a new sample $\mathbf{x}' = (1-\delta)\mathbf{x}_i + \delta\mathbf{x}_j$.
5. Add the new synthetic training example $(\mathbf{x}', C_q)$ to the augmented training dataset.

An illustration of this procedure can be seen in Fig. 4.3.

4.5.2 *OGO-SP*

While SMOTE is able to generate samples in the intra-class regions (that is, in the regions in between samples of the same class) it is unable to generate them in the inter-class or border regions (the regions of the space in-between classes). Moreover, because of this limitation, it cannot exploit the order relation between class labels present in an ordinal regression task.

The Ordinal Graph Oversampling via Shortest Paths (OGO-SP) [71] method aims to solve these limitations by defining a way to construct a graph which captures the relations between samples of neighbouring classes, considering the ordering information provided by the class labels.

When augmenting class $\mathbb{C}_q$, it consists on creating an undirected graph $G_q = (V_q, E_q)$ where V_q is the set of vertices corresponding to a subset of samples in dataset D and E_q is the set of edges representing neighbouring samples:

$$V_q \subseteq \{v_1, v_2, \dots, v_N\}, \tag{4.1}$$

$$E_q \subseteq \Big\{\{v_i, v_j\} \mid \forall v_i, v_j \in V_q, i \neq j\Big\}. \tag{4.2}$$

In order to construct G_q we first construct a different graph G'_q of the same form from other three sub-graphs $G_{q-1,q}$, $G_{q,q}$ and $G_{q,q+1}$:

$$G_{q-1,q}=(V_{q-1,q},E_{q-1,q}), \tag{4.3}$$

$$G_{q,q} \quad =(\quad V_{q,q},\quad E_{q,q}), \tag{4.4}$$

$$G_{q,q+1}=(V_{q,q+1},E_{q,q+1}), \tag{4.5}$$

$$\begin{aligned} G'_q = (V'_q, E'_q) = (V_{q-1,q} \cup V_{q,q} \cup V_{q,q+1}, \\ E_{q-1,q} \cup E_{q,q} \cup E_{q,q+1}). \end{aligned} \tag{4.6}$$

The edges of graph $G_{q-1,q}$ are determined by the intersection of two different sets obtained from a neighbourhood analysis based on the distance relation d:

$$\mathfrak{N}_d(D_a, D_b, k) = \Big\{\{v_i, v_j\} \mid (\mathbf{x}_i \in D_a) \wedge (\mathbf{x}_j \in D_b) \wedge \Big(\mathbf{x}_j \in nn_d(\mathbf{x}_i, D_b, k)\Big)\Big\}, \tag{4.7}$$

$$E_{q-1,q} = \mathfrak{N}_d\Big(D_{q-1}, D_q, k\Big) \cap \mathfrak{N}_d\Big(D_q, D_{q-1}, k\Big), \tag{4.8}$$

where $nn_d(\mathbf{x}_i, D_b, k)$ is the set of the k nearest neighbours of $\mathbf{x}_i$ in D_b and $\mathfrak{N}_d(D_a, D_b, k)$ is the k-neighbourhood of D_a with respect to D_b. The vertices $V_{q-1,q}$ are all those that appear in any edge from $E_{q-1,q}$:

$$V_{q-1,q} = \Big\{v_i \mid \exists e \in E_{q-1,q}\,,\, v_i \in e\Big\}. \tag{4.9}$$

Using the intersection of both neighbourhoods ensures that only the connecting regions of each class are considered. Parameter k controls how broad the region to consider is.

Graph $G_{q,q+1}$ is defined analogously:

$$E_{q,q+1} = \mathfrak{N}_d\Big(D_q, D_{q+1}, k\Big) \cap \mathfrak{N}_d\Big(D_{q+1}, D_q, k\Big), \tag{4.10}$$

$$V_{q,q+1} = \Big\{v_i \mid \exists e \in E_{q,q+1},\, v_i \in e\Big\}, \tag{4.11}$$

and finally, $G_{q,q}$ is simply defined as:

$$E_{q,q} = \mathfrak{N}_d\big(D_q, D_q, k\big), \tag{4.12}$$

$$V_{q,q} = \big\{v_i \mid \exists e \in E_{q,q}\,,\, v_i \in e\big\}. \tag{4.13}$$

For the case of the extreme classes (C_1 and C_Q), one of $G_{q-1,q}$ or $G_{q,q+1}$ will be empty and only the connecting region to the one adjacent class is considered.

Based on the ordinal regression hypothesis that the distance to adjacent classes is lower than the distance to non-adjacent classes, the final graph $G_q = (V_q, E_q)$ is constructed based on the previously constructed G'_q. Ideally, a distance-based manifold exists in the class labels such that D_q lies in the space between D_{q-1} and D_{q+1}. In reality, some outliers may exist in D_q that are not desirable in the over-sampling procedure. In order to identify the key samples which lie between the adjacent classes, the shortest paths between the samples of D_{q-1} and D_{q+1} are used to decide the edges present in the final graph G_q.

A path between two vertices v_{n_1} and v_{n_z} of the graph is defined as the sequence $P = \langle v_{n_1}, v_{n_2}, \dots, v_{n_z}\rangle$ such that any two consecutive vertices v_{n_i} and $v_{n_{i+1}}$ are connected by an edge: $\{v_{n_i}, v_{n_{i+1}}\} \in E'_q\ \forall 1 \leq i < z$. If a cost function $c : E'_q \to \mathbb{R}$ assigning a cost to every edge is defined, the shortest path P_{n_1,n_z} between any two vertices v_{n_1} and v_{n_z} is that which minimizes the total sum of the costs of the edges $\sum_{i=1}^{z-1} c\big(\{v_{n_i}, v_{n_{i+1}}\}\big)$. In this case, the cost function selected is the same as the distance d used for nn_d, which is the L_2 norm or euclidean distance:

$$c\big(\{v_{n_i}, v_{n_{i+1}}\}\big) = d(\mathbf{x}_{n_i}, \mathbf{x}_{n_{i+1}}) = \|\mathbf{x}_{n_i} - \mathbf{x}_{n_{i+1}}\|_2. \tag{4.14}$$

In order to find those patterns in D_q lying in the latent ordinal manifold, all the shortest paths between all the vertices in $V_{q-1,q}$ and all in $V_{q,q+1}$ will be computed using Dijkstra's algorithm [26], and only the edges contained in one or more of these paths will be included in E_q:

$$\begin{aligned} E_q = \{\{v_i, v_j\} \mid i \neq j,\, \exists v_\alpha \in V_{q-1,q},\, v_\omega \in V_{q,q+1} \\ \text{s. t. } (\langle v_i, v_j\rangle \in P_{\alpha,\omega}) \vee (\langle v_i, v_j\rangle \in P_{\omega,\alpha})\}, \end{aligned} \tag{4.15}$$

$$V_q = \big\{v_i \mid \exists e \in E_q\,,\, v_i \in e\big\}. \tag{4.16}$$

Note that, if q is any of the extreme classes, $V_{q,q}$ will have to be used instead of $V_{q-1,q}$ or $V_{q,q+1}$, depending on the case.

An example of the graph construction procedure is shown in Fig. 4.4.

Finally, new synthetic samples can be generated from G_q: in order to generate sample $(\mathbf{x}', C_q)$, a random edge $e = \{v_i, v_j\} \in E_q$ is selected so that $\mathbf{x}'$ lies in the segment between $\mathbf{x}_i$ and $\mathbf{x}_j$ the same way as in SMOTE:

$$\mathbf{x}' = (1 - \delta)\mathbf{x}_i + \delta\mathbf{x}_j \tag{4.17}$$

where δ is a random variable in the range $[0, 1]$. However, this time the distribution from where δ is sampled will depend on the selected edge e:

- If both $y_i = C_q$ and $y_j = C_q$ (i. e. e is an intra-class edge), then δ is sampled from a uniform distribution $U(0, 1)$ in the same manner as SMOTE.

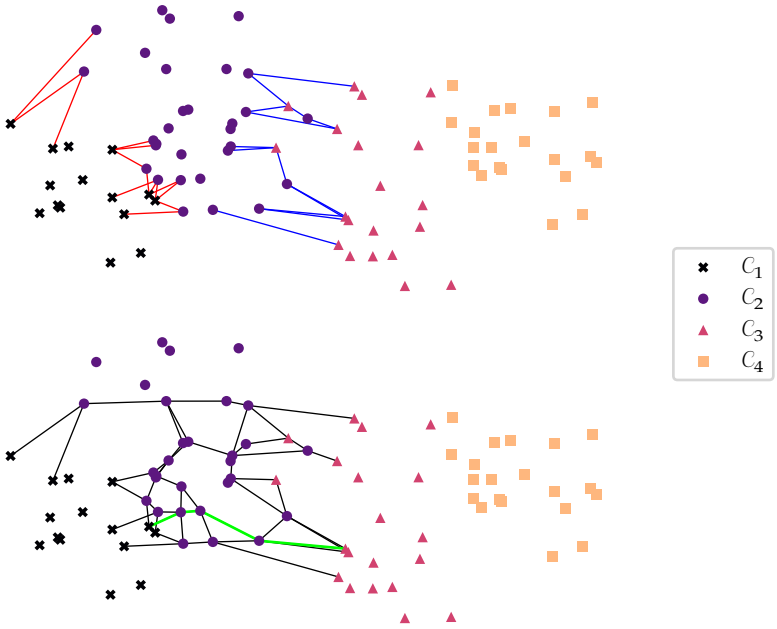


Figure 4.4: Example of the OGO-SP graph construction procedure. The markers represent samples of the dataset. The graph $G_2 = (V_2, E_2)$ corresponding to C_2 is constructed. The top diagram shows $E_{1,2}$ (in red) and $E_{2,3}$ (in blue). The bottom diagram shows the shortest path between two vertices in $V_{1,2}$ and $V_{2,3}$ (in green) and the edges of the final constructed graph E_2 (in black).

- If $y_i = C_q$ but $y_j \neq C_q$ (i. e. e is an inter-class edge), then δ is sampled from an asymmetrical distribution so that the new synthetic sample favours the augmented class but is able to capture the class transition phase. In the original OGO-SP paper δ follows a gamma distribution $\delta \sim \Gamma(k = 2, \theta = 0.15)$.

4.5.3 *Limitations of OGO-SP*

While the gamma distribution used in the inter-class region generation procedure presents the desired asymmetry, it is not without problems. This distribution is not bounded, i. e. it generates a δ in the range $\delta \in [0, \infty)$, meaning that $P(\delta > 1) > 0$. Because of this, samples could be generated outside the selected edge of the graph which would no longer constitute as interpolation.

To solve this, we propose using a more appropriate distribution like the beta distribution, controlled by two parameters $\alpha, \beta > 0$ [49].This distribution is bounded in the $[0, 1]$ interval and its parametrization allows it to be skewed towards any of the two extremes depending on the values of α and β, which the gamma distribution only allows for the lower bound of 0.

This distribution has been used to model the behaviour of finitely bounded variables in a variety of disciplines. In its standard form, its probability density function $f(x; \alpha, \beta)$ is defined as:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{4.18}$$

where $0 < x < 1$, $\alpha, \beta > 0$ and B is the beta function which acts as a normalization constant defined as:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{4.19}$$

Several basic properties can be derived from this definition:

- If $\alpha \geq 1$ then $f(0) = 0$. Similarly, if $\beta \geq 1$, then $f(1) = 0$.
- For both $\alpha, \beta > 1$, the mode is unique and equal to $\frac{\alpha-1}{\alpha+\beta-2}$.
- If $\alpha = \beta$ the distribution is symmetric. The special case where $\alpha = \beta = 1$ is equivalent to a uniform distribution.

This modified version of OGO-SP with beta inter-class distribution will be referred to as OGO-SP-β.

Based on the two distinct possibilities for the endpoints represented by $f(0)$ and $f(1)$, we can derive four unique asymmetric shapes for this distribution. However, we exclude one of these shapes, characterized by parameters $\alpha > 1$ and $\beta < 1$, as it tends to concentrate more of the probability mass on the neighbouring class side of the distribution. To ensure this, we impose a quantile constraint, requiring that $P(\delta < 0.5) = 0.75$. This constraint ensures that the majority of the probability mass is concentrated on the side corresponding to the augmented class.

The authors of [95] have demonstrated that employing two quantile constraints is sufficient to parametrize the beta distribution, and they have devised a numerical method to calculate the values of α and β that satisfy these constraints. To create the three distinct shapes, we employ three additional quantile constraints, each in combination with the previous one. Using the aforementioned method, we derive the values of α and β for each distribution:

(a) Beta distribution with $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.65) = 0.9 \rightarrow \delta \sim \text{Beta}(\alpha = 1.558, \beta = 2.827)$

(b) Beta distribution with $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.75) = 0.9 \rightarrow \delta \sim \text{Beta}(\alpha = 0.513, \beta = 1.186)$

(c) Beta distribution with $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.85) = 0.9 \rightarrow \delta \sim \text{Beta}(\alpha = 0.243, \beta = 0.642)$

These three configurations will be tested and compared to the original OGO-SP with gamma distribution. We expect the beta distribution will be a better candidate for synthetic sample generation for certain datasets, like the one presented in Section 4.3.1. Configuration (a) is similar to the original gamma distribution just for comparison, while configurations (b) and (c) of OGO-SP-β exploit the versatility of the beta distribution.

A visual representation of the probability density function's shape for all the proposed configurations can be observed in Fig. 4.5. It becomes evident that when $\alpha < 1$, the probability density function exhibits an infinite value at $\delta = 0$, as it allocates a greater portion of probability mass toward that extreme. Similarly, when $\beta < 1$ and $\delta = 1$, a comparable trend occurs. Consequently, this approach enables precise control over the probability of generating samples in the inter-class region while promoting the generation of samples near the augmented class region, relative to the neighbouring class.

In the context of this study, the dataset described in Section 4.3.1 exhibits significant class imbalance, making data augmentation a critical factor for enhancing the classification

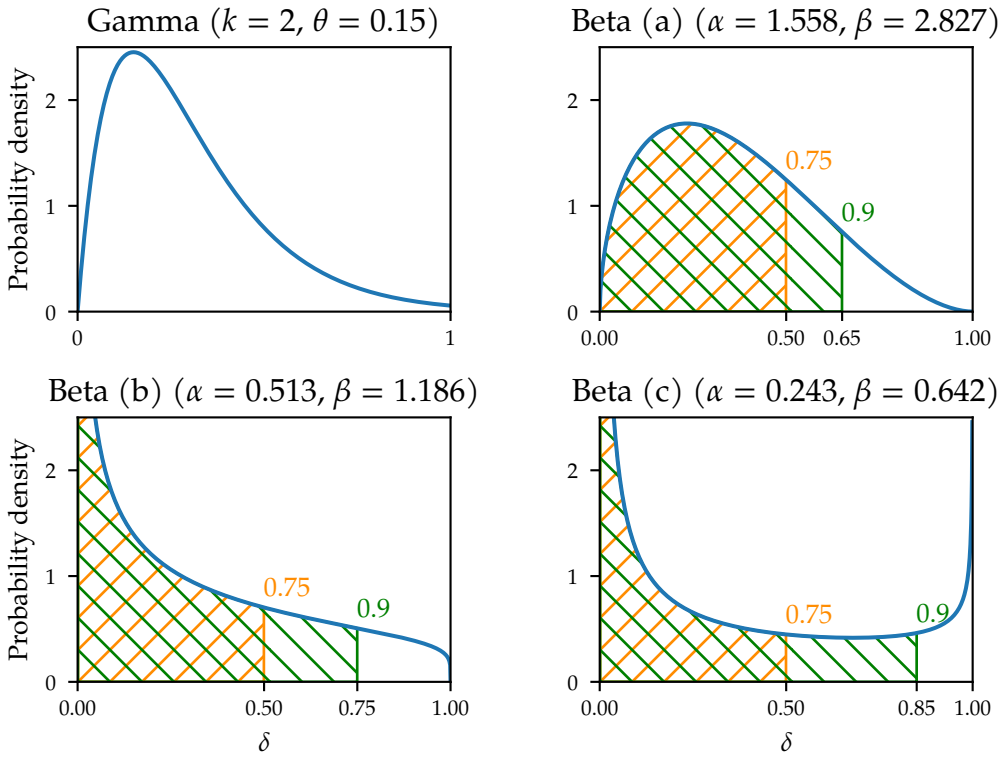


Figure 4.5: Shape of the probability density function for the four different distributions of δ: the original gamma distribution from [71] and the three proposed configurations of the beta distribution. The labelled shaded regions indicate the imposed quantile constraints.

model's performance. We contend that the OGO-SP-β algorithm is particularly well-suited for this scenario, primarily owing to the ordinal nature of the problem. The gradual progression of dopaminergic activity alteration implies that more pronounced damage is expected in the later stages of the disease, and vice versa. Additionally, given that the intermediate classes constitute the minority, the utilization of the beta distribution inherently promotes the generation of samples within these class regions.

4.5.4 *Application to spatial data*

Directly using techniques such as SMOTE or OGO-SP for spatial data, like images or 3D scans, is not a suitable approach. These methods fail to capture the complex positional variability of various objects within a scene or anatomical elements in a CT scan. Applying such techniques directly to the original image space leads to the generation of entirely unnatural and unsuitable synthetic samples, adversely affecting the model's generalization capabilities.

Conversely, the convolutional segment of a CNN strives to project data from the original space into a reduced-dimensional space, better suited for interpolation and the application of class-balancing methods such as SMOTE and its derivatives. Consequently, in this study, we propose a two-step training process for class balancing:

1. Initially, the entire network (comprising both the convolutional and fully connected components) is trained on the original dataset D.

2. Upon reaching the stopping criterion, the convolutional component of the network, g, is employed to project the original dataset D into a new space with reduced dimensionality, yielding D':

$$D' = \{(g(\mathbf{x}_i), y_i) \mid (\mathbf{x}_i, y_i) \in D\}, \tag{4.20}$$

$$g : \mathbb{R}^m \to \mathbb{R}^{m'}, \tag{4.21}$$

where m represents the original dimensionality of the data, and m' is the new, reduced dimensionality.

3. Subsequently, new synthetic samples for each class C_q are generated using SMOTE in the nominal case and OGO-SP/OGO-SP-β in the ordinal case. Let N_q denote the number of samples labelled as C_q in D'. The set of generated samples for C_q is referred to as D'_{+q}. The objective is to generate samples to balance the number of training samples for all classes, which means no synthetic sample will be generated for the majority class:

$$\left|D'_{+q}\right| = \left(\max_{1 \le k \le Q} N_k\right) - N_q. \tag{4.22}$$

4. The synthetic samples are then merged with the dataset D' to create the augmented dataset D'_+:

$$D'_+ = D' \cup \left(\bigcup_{q=1}^{Q} D'_{+q}\right).$$

5. Finally, training is resumed only for the fully connected component of the original model using D'_+, with the same stopping criterion.

4.6 EXPERIMENTATION

The four different methodologies previously described (one nominal and four ordinal for the different distributions of δ) will be compared against each other in order to evaluate the effect of using ordinal information in the learning process. This includes both the CNN architecture as well as the data augmentation procedure. More specifically, the nominal architecture will be paired with SMOTE and the ordinal architecture will be paired with OGO-SP and OGO-SP-β in its three different configurations proposed in Section 4.5.3.

4.6.1 *Experimental design*

A stratified 5-fold cross-validation process is conducted over the entire dataset. This process involves splitting the dataset into 5 roughly equal-sized subsets while ensuring that the class distribution is balanced within each fold. During each step of this cross-validation, one of the subsets is designated for testing, while the remaining subsets serve as training samples.

Within each of these 5-fold steps, the first phase entails model selection, wherein the algorithm's hyperparameters are fine-tuned. To accomplish this, three 90/10 holdout splits are produced from the training folds and all feasible combinations of the following parameter values are explored:

- Learning rate (η): $\{10^{-3}, 10^{-4}\}$.

- Hidden layer size (H): $\{2048, 4096\}$.
- Convolution kernel size (k): $\{3, 5\}$.
- Neighbourhood size for the data augmentation technique: $\{3, 5\}$.

The parameter combinations are assessed based on the mean *MAE* score across the three splits, and the optimal combination for each fold is selected for further evaluation.

Following the hyperparameter selection phase, the optimal parameter combination is used in the second phase for the final evaluation. The model is initialized with a different random seed and trained with a different 90/10 holdout split for train/validation 30 times. The trained models are then evaluated using the test fold.

This identical procedure is iterated for each of the data folds. An illustration of this process is depicted in Fig. 4.6.

In all cases a training batch size of 32 is used and the validation set is used to monitor the training process: when the loss over the validation set does not increase for 50 epochs training is stopped and the best performing weights are restored.

4.6.2 *Evaluation metrics*

While *CCR* is typically the predominant criterion in classification tasks, its relevance diminishes in cases of significant class imbalance [75]. In such scenarios, models that disregard minority classes can still yield high *CCR* scores, which is undesirable. Therefore, a focus on per-class sensitivity becomes imperative [83]. It's worth noting that while optimizing sensitivity, there may be a trade-off in terms of specificity, necessitating close attention.

Furthermore, *CCR* fails to consider the extent of deviation of each prediction from the ground truth. It is primarily designed for general multiclass classification problems, where all errors are equally penalized. However, in the context of ordinal regression, it is more desirable to prioritize a classification error of only 1 class over an error of 2 classes. Hence, rank difference metrics such as the Mean Absolute Error (*MAE*), Spearman's rank correlation coefficient (r_s), Kendall's rank correlation coefficient (τ_b) [14], and the weighted Cohen's kappa coefficient (κ) [10] prove to be more suitable for evaluating model performance. Additionally, in scenarios with high class imbalance, it's valuable to consider per-class *MAE*.

Additionally, when assessing per-class metrics, examining the extreme values becomes essential to ensure that performance improvements do not come at the cost of neglecting certain classes.

Therefore, the metrics monitored in the test subset for this study are:

- For general classification performance:
  - Correct Classification Rate (*CCR*).
- For ordinal regression performance:
  - Mean Absolute Error (*MAE*)
  - Quadratic weighted Cohen's kappa (κ).
  - Kendall's rank correlation coefficient (τ_b).
  - Spearman's rank correlation coefficient (r_s).
- For class balancing performance:

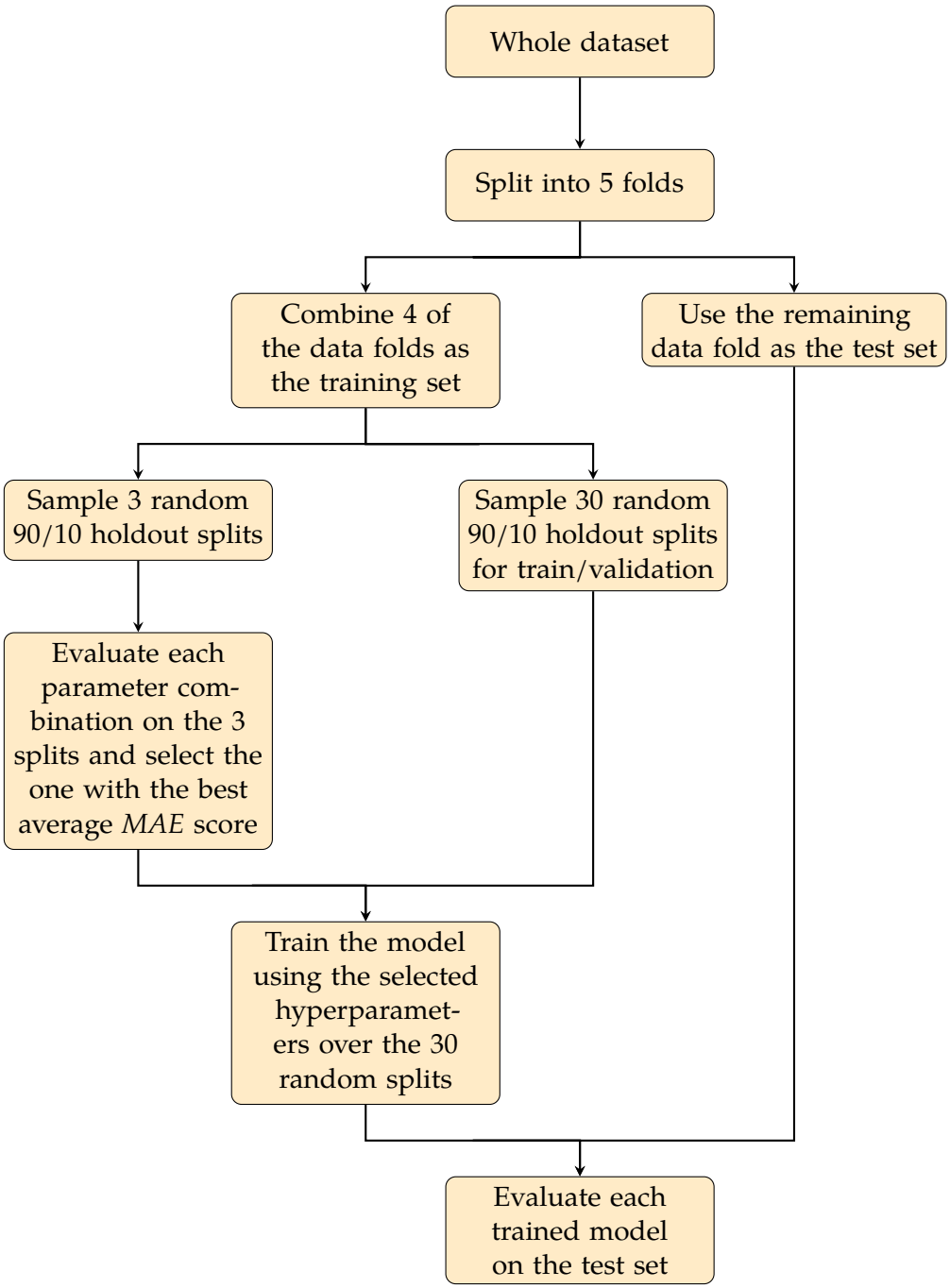


Figure 4.6: Cross-validation scheme used for the validation of hyperparameters and evaluation of the models

	Nominal		OGO-SP		OGO-SP-β (a)		OGO-SP-β (b)		OGO-SP-β (c)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
CCR (↑)	**0.7448**	0.0412	0.7121	0.0422	0.7048	0.0487	0.7108	0.0644	*0.7255*	0.0423
GMS (↑)	0.1927	0.2015	*0.4256*	0.1847	0.4134	0.1891	0.4239	0.1734	**0.4403**	0.1830
MinS (↑)	0.0679	0.0774	*0.2426*	0.1611	0.2297	0.1575	0.2235	0.1475	**0.2460**	0.1451
GMSp (↑)	0.8898	0.0179	*0.8979*	0.0138	0.8957	0.0146	0.8977	0.0229	**0.9009**	0.0170
MinSp (↑)	0.7691	0.0569	0.8276	0.0407	0.8223	0.0406	*0.8294*	0.0601	**0.8312**	0.0512
AvAUC (↑)	0.8486	0.0344	*0.8588*	0.0345	0.8553	0.0385	0.8567	0.0331	**0.8596**	0.0366
MAE (↓)	0.3826	0.0737	0.3738	0.0586	0.3791	0.0643	*0.3729*	0.0751	**0.3639**	0.0649
AvMAE (↓)	0.6803	0.1127	0.5671	0.1131	0.5668	0.1070	*0.5631*	0.1038	**0.5594**	0.1186
MaxMAE (↓)	1.1427	0.1328	*0.9043*	0.2148	**0.9021**	0.2005	0.9187	0.1800	0.9113	0.1896
τ_b (↑)	0.7119	0.0565	0.7323	0.0442	0.7282	0.0496	*0.7343*	0.0467	**0.7389**	0.0481
κ (↑)	0.7663	0.0602	0.7901	0.0534	0.7882	0.0539	**0.7926**	0.0561	*0.7910*	0.0571
r_s (↑)	0.7702	0.0600	0.7980	0.0516	0.7968	0.0527	**0.8011**	0.0512	*0.7986*	0.0548

Table 4.1: Summary of evaluation results. Best and second best results are highlighted in bold and italics, respectively

- Minimum sensitivity (*MinS*).
- Geometric mean of the sensitivities (*GMS*).
- Minimum specificity (*MinSp*).
- Geometric mean of the specificities (*GMSp*).
- Average *MAE* (*AvMAE*).
- Maximum *MAE* (*MaxMAE*).
- Average area under the ROC curve (*AvAUC*).

All of these metrics are defined in Section 1.2.5.

4.7 RESULTS

Table 4.1 includes a summary of all the results from the 150 executions of the five different methodologies, based on the different performance metrics introduced in the previous subsection.

From these results, it is evident that the (c) configuration of OGO-SP-β consistently outperforms all other configurations of the beta distribution across most metrics. Comparing OGO-SP-β with the nominal methodology and the original OGO-SP, it always outperforms both except for *CCR*, where the nominal methodology achieves the best results, with OGO-SP-β (c) securing the second position.

A close examination of the *GMS* and *MinS* metrics reveals that the nominal methodology fails to address the prevailing class imbalance. It tends to disregard the minority classes, essentially grouping most test samples into a couple of the majority classes, which significantly lowers the scores, often resulting in zeros across multiple evaluation splits.

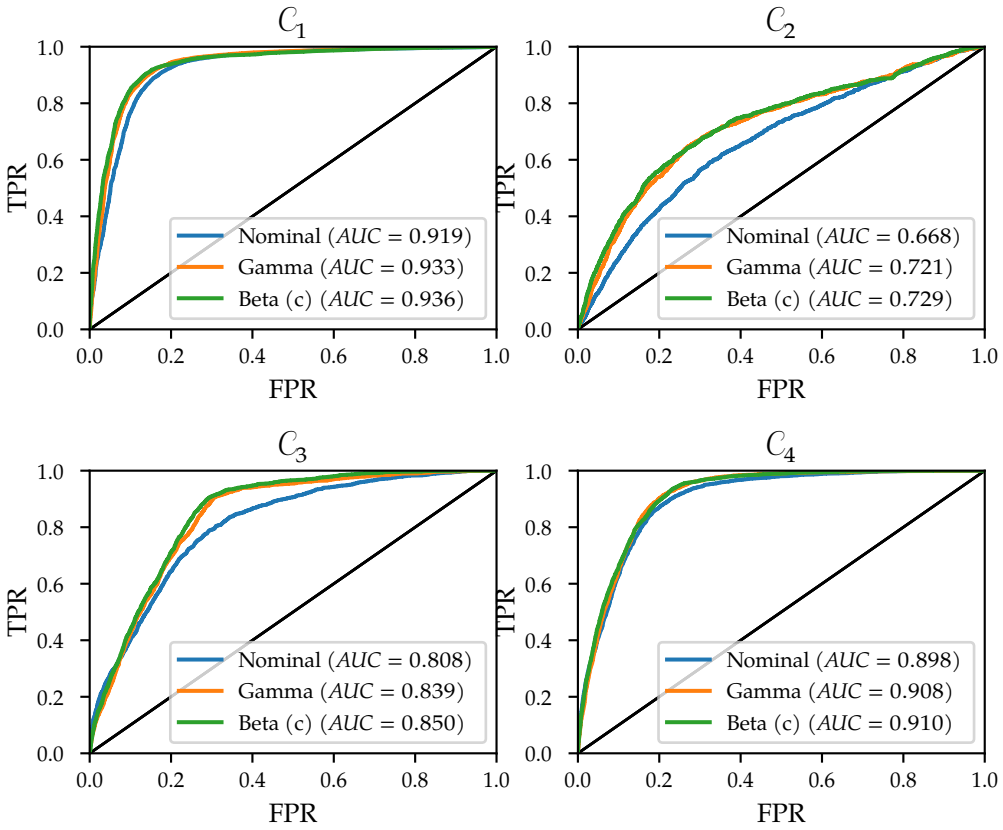


Figure 4.7: ROC obtained for each of the four classes by three of the evaluated methodologies. The curves are obtained according to the output scores of each model using a OvR scheme for each class

Comparing OGO-SP-β in its three configurations with the original OGO-SP, it is evident that OGO-SP-β consistently delivers better average performance across all splits, especially in terms of *MAE*.

Both ordinal methodologies clearly outperform the nominal approach. Figure 4.7 illustrates the corresponding ROC curves for the four classes in the problem, constructed in a OvR fashion. Both ordinal methodologies exhibit a significant advantage in the intermediate minority classes, particularly for C_2, which has the fewest samples.

To assess the significance of the performance difference introduced by OGO-SP-β compared to both the purely nominal methodology and the original OGO-SP approach with gamma inter-class distribution, we conduct a two-sided Wilcoxon signed-rank test across all metrics [106]. This non-parametric statistical test is well-suited for our scenario since the results cannot be assumed to follow a normal distribution. We formulate the null hypothesis that the performance rank difference for the compared methodologies is symmetrical about zero, with the alternative hypothesis suggesting an asymmetry. If the achieved p-value is less than $\alpha = 0.05$ the null hypothesis is rejected and the sign of the Wilcoxon statistic T (sum of signed ranks) will indicate which of the two methodologies is favoured.

The p-values for a two-tailed test for each metric are provided in Table 4.2.

	OGO-SP-β (c) vs. Nominal		OGO-SP-β (c) vs. OGO-SP	
	T	p-value	T	p-value
CCR	**−4481**	**<0.001**	**3482**	**0.001**
GMS	**9246**	**<0.001**	1174	0.271
MinS	**9647**	**<0.001**	493	0.643
GMSp	**7473**	**<0.001**	**2500**	**0.019**
MinSp	**9110**	**<0.001**	1155	0.278
AvAUC	**4767**	**<0.001**	911	0.393
MAE	**3549**	**<0.001**	**2148**	**0.044**
AvMAE	**10 531**	**<0.001**	1494	0.161
MaxMAE	**9495**	**<0.001**	−295	0.782
τ_b	**6123**	**<0.001**	1913	0.073
κ	**5485**	**<0.001**	479	0.653
r_s	**6577**	**<0.001**	451	0.672

Table 4.2: Two-tailed Wilcoxon signed rank test results. p-values less than $\alpha = 0.05$ have been highlighted in bold. A positive value of T means that the first methodology is favoured and a negative value means otherwise

Significant differences in performance consistently favour the ordinal methodology over the nominal one across all metrics except for *CCR*. As anticipated, the nominal methodology prioritizes improving *CCR* at the expense of other ordinal metrics, which are generally more relevant for ordinal-type classification problems. Additionally, the nominal approach tends to neglect minority classes, leading to notably poorer results in *GMS*, *MinS*, *GMSp*, *MinSp*, and *AvAUC*. Furthermore, OGO-SP-β exhibits significant improvements in *CCR*, *GMSp*, and *MAE* when compared to the original OGO-SP using the gamma distribution.

Even in cases where the statistical tests do not reveal significant differences, OGO-SP-β consistently demonstrates superior overall performance in ordinal metrics. The uncertainty observed in imbalance-sensitive metrics (*GMS*, *MinS*, *MinSp*, and *MaxMAE*) could be attributed to their inherent instability, resulting in a larger standard deviation and making it more challenging for the tests to reach conclusive outcomes.

4.8 COMPARISON TO STATE-OF-THE-ART

Given the absence of prior literature addressing the diagnosis of presynaptic damage stages in PD, certain binary metrics can be derived from the results for the sole purpose of comparison with existing studies. These previous works typically focus on the binary classification problem of distinguishing PD patients from healthy controls. Directly comparing metrics like *CCR* against the context of a binary classifier would be inappropriate and biased, as the multiclass metric is more demanding, requiring classification into four distinct classes instead of just two.

To facilitate this comparison, all labels corresponding to C_1 will be treated as the 'negative class' or healthy control label, while the remaining labels (C_2 through C_4) will be regarded as the 'positive class' or PD label. This classification scheme allows the extraction of a binary

confusion matrix and so accuracy, sensitivity and specificity can be obtained as discussed in Section 1.2.5.1.

These results will be compared to the following works from the literature:

- Rizzo et al. (2016) [79]: a meta-analysis of 20 different studies, all using different techniques, between 1988 and 2014.
- Fuente-Fernández (2012) [36]: an aggregation of 2 different studies both using SPECT imaging.
- Martinez-Murcia et al. (2017) [62]: a CNN approach for binary classification from SPECT imaging.
- Orozco-Arroyave et al. (2016) [69]: an application of radial base Support Vector Machines to running speech audio samples from patients.
- El Maachi, Bilodeau and Bouachir (2020) [30]: an application of neural networks to gait sensor data from patients.

It is important to highlight beforehand that the task addressed in these studies differs, as the various potential positive labels for neurological damage are not distinguished, significantly reducing the complexity. Despite this distinction and acknowledging the greater informativity of the proposed models, alongside variations in experimental setups and datasets, we can deduce from the results shown in Table 4.3 that the performance achieved by our proposals is competitive. This is particularly evident when aiming for a balance across the three binary metrics. In essence, the additional information provided by the proposed multi-class classifiers does not compromise performance in the binary task.

Method/work	Type	Data	Accuracy	Sensitivity	Specificity
Nominal	A	I	87.62 %	77.52 %	93.84 %
OGO-SP	A	I	87.30 %	86.74 %	87.64 %
OGO-SP-β (a)	A	I	86.94 %	86.67 %	87.09 %
OGO-SP-β (b)	A	I	87.14 %	87.10 %	87.16 %
OGO-SP-β (c)	A	I	88.19 %	86.21 %	89.38 %
Rizzo et al. (2016) [a]	H	C	86.4 %	86.6 %	84.7 %
Fuente-Fernández (2012) [b]	H	I	84 % to 98 %	98 %	67 % to 94 %
Martinez-Murcia et al. (2017)	A	I	95.5 %	96.2 %	-
Orozco-Arroyave et al. (2016) [c]	A	S	85 % to 97 %	76.7 % to 98 %	93.3 % to 96 %
El Maachi et al. (2020)	A	G	98.7 %	98.1 %	100 %

a Values are taken from the refined clinical diagnosis by experts
b Ranges represent the performance between early and established diagnosis, when provided.
c Ranges represent the worst and best performance across all three tested languages.

Table 4.3: Binary metric results for the five tested methodologies and five additional studies. Rows labelled as 'A' represent automatic methods, while rows labelled as 'H' pertain to studies involving human expert diagnosis. The Data column specifies the reference data employed for diagnosis: imaging (I), speech (S), gait (G), or a detailed clinical diagnosis (C)

4.9 CONCLUSIONS

In this chapter, experimental confirmation has been obtained regarding the enhancement of performance in a complex task, specifically the assessment of altered brain activity in individuals with PD, through the exploitation of ordinal information. This enhancement is attributed to various factors, including the model architecture, optimization objectives, and data augmentation techniques. Notably, an extension of the repertoire of models utilizing ordinal information is achieved, exemplified by a fully 3D CNN.

The class imbalance issue is effectively addressed by the proposed methodology, which simultaneously enhances ordinal performance metrics. Application potential is observed in existing ordinal regression tasks within the medical domain, where class imbalance is a prevalent challenge.

Furthermore, improved performance in both ordinal and nominal metrics, in comparison to the original OGO-SP, is demonstrated by the introduced OGO-SP-β ordinal augmentation algorithm.

From a more conventional binary diagnosis perspective, adequate performance is exhibited by our methodology, as evidenced in the comparison against expert diagnoses and other ML techniques.

5

EXAMINING THE DECISION PROCESS OF ORDINAL CNNS

Despite the remarkable success of CNN models for image classification, a notable challenge lies in their interpretability, hindering debugging, validation, and auditing processes. In response to this challenge, a variety of explanation methods have been devised by researchers to shed light on the internal mechanisms of CNN models. This is empowering to both developers of CNN models as a debugging tool and to final users for which more context is given on the model's decision. In the medical field, where high levels of accountability and transparency are necessary, this is especially crucial [92].

However, there exists a gap in research concerning the validity of these methods when applied to ordinal regression tasks. Having introduced an output architecture for CNNs tailored to ordinal regression tasks in the preceding chapters, our focus shifts to addressing the explainability problem inside the ordinal framework. The nature of the output variable in these types of tasks can be fundamentally different, necessitating more specially tailored methods.

This chapter is thus dedicated to the validation of existing explanation methods on ordinal regression tasks in addition to the development of two different ordinal-specific adaptations compatible with and specially adequate to the architectures presented in the previous chapters. In the process, a visual explanation evaluation procedure is proposed which can capture the level of ordinal performance achieved by each method.

ASSOCIATED PUBLICATION: **Javier Barbero-Gómez**, Ricardo Cruz, Jaime S. Cardoso, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Evaluating the Performance of Explanation Methods on Ordinal Regression CNN Models'. In: International Work-Conference on Artificial Neural Networks (IWANN 2023). Ponta Delgada, Portugal, June 2023.

5.1 EXPLAINABLE AI

Explainable AI (XAI), sometimes synonymous with Interpretable AI or Explainable Machine Learning (XML), entails ensuring that an AI system remains comprehensible and transparent to humans [8]. This field encompasses both the concept of maintaining human oversight over AI systems and the methodologies employed to achieve this transparency. The central objective is to clear up the rationale behind the decisions or predictions made by AI, mitigating the inherent 'black box' nature commonly associated with ML. In contrast to scenarios where the decision-making process is opaque, XAI strives to make AI systems more understandable, allowing humans, including the system's designers, to grasp the underlying reasoning.

5.2 EXPLANATION METHODS: PROBLEM DEFINITION

Explanation methods for CNNs are a set of techniques designed in the pursuit of more explainable CNN models. Their aim is to create an explanation map that identifies the significant regions of an input image utilized by the model for its classification decision C_q.

This map is denoted by a matrix $E_q \in \mathbb{R}^{H \times W}$, where H and W correspond to the height and width of the input image, respectively. Each entry in E_q assumes values between 0 and 1, where 0 signifies no relevance, and 1 signifies maximum relevance, for every pixel in the original image.

5.3 RELATED WORK

Various methods have been developed to explain the decisions made by CNNs. These approaches span from basic occlusion analysis [109] to more advanced techniques like back-propagation of gradients, which result in the generation of saliency maps [4, 66, 86, 89]. Class Activation Map (CAM) methods integrate these gradients with layer activations [85, 112], with Grad-CAM standing out as one of the widely embraced algorithms [85]. Additionally, the Information Bottleneck for Attribution (IBA) method [84] introduces a perturbation in the network, resulting in a bottleneck capable of evaluating the importance of each region for the final output. It is worth noting that the resulting explanation maps obtained through any of these methods are commonly employed for object localization tasks [45].

Beyond post-hoc explanations methods, more recent research lines explore the option of including the explanation generation inside the model architecture itself [77]. While potentially more powerful, this approach requires a substantial modification of existing models and a multi-objective training process which is difficult to fine-tune.

Despite the prevalence of these explanation methods, there is a noticeable gap in research regarding their applicability and validity for ordinal regression tasks.

5.4 GOALS

In the previous chapters, it was shown that ordinal-specific models were able to outperform nominal classifiers in relevant metrics. We now turn our attention towards explanation methods in the ordinal setting. The performance of existing explanation methods has not yet been evaluated when applied to ordinal tasks and models. It could be possible that incorporating order information in the explanation procedure may lead to better insights into the classification process. This is a motivation to try to devise specific ordinal explanation methods able to further exploit ordinality and improve the explanations obtained.

The specific goals of this chapter, encompassed by objective 4 from Section 2.2, can be summarized as follows:

- Proposing an evaluation procedure for explanation maps that takes into account ordinal regression performance.
- Proposing the Gradient OBD Class Activation Map (GradOBD-CAM) and Ordinal Information Bottleneck Attribution (OIBA) explanation methods as modifications to Grad-CAM and IBA, respectively, integrating ordinal information in the explanation map building process.
- Comparing the performance of existing explanation methods against these new proposals on ordinal regression tasks using the evaluation procedure.

5.5 NOMINAL EXPLANATION METHODS FOR CNN MODELS

The following explanation methods were designed within a general nominal classification framework, with no regard for the potential order relationship between the class labels.

5.5.1 *CAM*

An early approach to explanation methods for CNNs were Class Activation Maps (CAMs) [112]. This method is limited to CNN architectures with an output stage consisting of GAP followed by a fully connected layer. It is expected of said architectures that the activations on the stage prior to GAP correspond to specific visual patterns which may or may not contribute positively or negatively to each output class score. For this reason, the CAM method proposes generating the explanation map as a linear combination of these activations. Consequently, CAM and its derivatives are considered *activation-based* methods.

For a network with K feature maps right before GAP, let A^k be the activation of the k-th feature map and w_q^k be the weight of the contribution of A^k on the output activations for class C_q. CAM defines the explanation map E_q for class C_q to be:

$$E_q(i,j) = \text{ReLU}\left(\sum_{k=1}^{K} w_q^k A^k(i,j)\right), \tag{5.1}$$

where $E_q(i,j)$ and $A^k(i,j)$ are the pixel in the i-th row and j-th column of explanation map E_q and feature map A^k, respectively. The element-wise ReLU function, defined in Eq. (1.29), discards negative activations in order to highlight only positively correlated regions.

5.5.2 *Grad-CAM*

The architectural requirements of CAM mean that it is impossible to apply it to any network with a different output computation. Not only that, but the resolution of the explanation map is limited to the resolution of the last convolutional layer.

Grad-CAM [85] is designed as a generalization of CAM with the goal of overcoming these limitations. Now, any intermediate layer may be selected, and the weights of the linear combination are derived from the average gradient with respect to the class score.

Let the output score for class C_q be s_q and the selected intermediate layer A have a height and width of H_A and W_A, respectively. According to Grad-CAM, the explanation map E_q for class C_q is defined as:

$$E_q(i,j) = \text{ReLU}\left(\sum_{k=1}^{K} w_q^k A^k(i,j)\right), \tag{5.2}$$

$$w_q^k = \frac{1}{H_A \times W_A} \sum_{i=1}^{H_A} \sum_{j=1}^{W_A} \frac{\partial s_q}{\partial A^k(i,j)}. \tag{5.3}$$

5.5.2.1 *Grad-CAM++*

Improvements to the feature map weighting of Grad-CAM have been proposed in the literature, such as Grad-CAM++ [15]. Its authors pose that Grad-CAM is less effective when trying to localize multiple occurrences of the same object class or just the whole of an object in the input image. Thus, they introduce a per-pixel weighting scheme in the computation of w_q^k:

$$\alpha_q^k(i,j) = \frac{\left(\frac{\partial s_q}{\partial A^k(i,j)}\right)^2}{2\left(\frac{\partial s_q}{\partial A^k(i,j)}\right)^2 + \sum_{a=1}^{H_A}\sum_{b=1}^{W_A} A^k(a,b)\left(\frac{\partial s_q}{\partial A^k(i,j)}\right)^3}, \tag{5.4}$$

$$w_q^k = \sum_{i=1}^{H_A}\sum_{j=1}^{W_A} \alpha_q^k(i,j)\,\mathrm{ReLU}\left(\frac{\partial s_q}{\partial A^k(i,j)}\right). \tag{5.5}$$

5.5.2.2 *Score-CAM*

Score-CAM [104] drops the use of gradients entirely and assigns weights to each feature map based on an *increase of confidence* measure, i. e. the amount of influence that the regions with high activations in A^k really have over the class output score s_q.

To this end, first the baseline score $\bar{s}_q$ that the model assigns to a baseline empty image (all zeroes) is obtained. Then, after obtaining A^k and s_q, A^k is upsampled to the input image size $H \times W$, normalized between 0 and 1, and used as a mask over the original input image. The new output score for the masked input s'_q is obtained and the increase of confidence contributed by A^k is used as the weight w_q^k of the k-th feature map for the explanation map E_q:

$$w_q^k = s'_q - \bar{s}_q. \tag{5.6}$$

5.5.3 *IBA*

The IBA method, as proposed by [84], is a *perturbation-based* method, meaning that some kind of information is introduced in the computation in order to study its effects in the output of the model. However, unlike other perturbation methods that alter the information at the input of the model, it consists of injecting a perturbation amidst its information flow, creating a bottleneck in the network. This bottleneck helps evaluate the impact in the output of the regions from the input image.

To achieve this, it introduces a new random variable Z that maximizes the amount of information it shares with the output score of the target class s_q while minimizing the information it shares with the model input $\mathbf{x}$:

$$\max\ I[s_q; Z] - \beta I[\mathbf{x}; Z], \tag{5.7}$$

where I denotes the mutual information and β controls the trade-off between predicting the labels well and using little information of the input. $Z \in \mathbb{R}^{H_A \times W_A}$ acts as a substitute for the output of one of the intermediate layers A adding a certain noise $\epsilon \in \mathbb{R}^{H_A \times W_A}$:

$$Z = \lambda(\mathbf{x})A + (1 - \lambda(\mathbf{x}))\epsilon, \tag{5.8}$$

where $\lambda(\mathbf{x}) \in [0,1]^{H_A \times W_A}$ adjusts how much of the original signal is passed along.

In order to obtain a value of $\lambda(\mathbf{x})$ that aligns with the objective posed in Eq. (5.7), a loss function $\mathcal{L}_\lambda$ to optimize is designed. To estimate how much information from A is passed along in Z, mutual information is used:

$$I[A;Z] = \mathbb{E}_A[\mathrm{D_{KL}}[\mathrm{P}(Z \mid A) \,\|\, \mathrm{P}(Z)]], \tag{5.9}$$

where $\mathrm{P}(Z \mid A)$ and $\mathrm{P}(Z)$ are the respective probabilities, $\mathrm{D_{KL}}$ is the Kullback-Leibler divergence and $\mathbb{E}_A$ the expectation over A. This, however, is an unmanageable computation, so an approximation $Q(Z) = \mathcal{N}(\mu_A, \sigma_A)$ is made assuming that all dimensions of Z are distributed independently and normally, which overestimates the real value:

$$I[A;Z] = \mathbb{E}_A[\mathrm{D_{KL}}[\mathrm{P}(Z \mid A) \,\|\, Q(Z)]] - \mathrm{D_{KL}}[\mathrm{P}(Z) \,\|\, Q(Z)]. \tag{5.10}$$

Finally, the information loss function $\mathcal{L}_\mathrm{I}$ is defined as:

$$\mathcal{L}_\mathrm{I} = \mathbb{E}_A[\mathrm{D_{KL}}[\mathrm{P}(Z \mid A) \| Q(Z)]], \tag{5.11}$$

and the final loss function $\mathcal{L}_\lambda$ is defined as the combination of $\mathcal{L}_\mathrm{I}$ and the cross-entropy loss $\mathcal{L}_\mathrm{CE}$ as specified in Eq. (1.37):

$$\mathcal{L}_\lambda = \mathcal{L}_\mathrm{CE} + \beta \mathcal{L}_\mathrm{I}. \tag{5.12}$$

This can now be used to optimize $\lambda(\mathbf{x})$, parametrised as $\lambda(\mathbf{x}) = \sigma(\alpha(\mathbf{x}))$ (where $\alpha \in \mathbb{R}^{H_A \times W_A}$ and σ is the sigmoid function), by minimizing $\mathcal{L}_\lambda$ using any stochastic gradient descent algorithm such as the Adam back-propagation method [52].

Regions of the image with relevant information will present a λ value close to 1 and, conversely irrelevant parts will present a value close to 0. For this reason, the output explanation map E is just λ upsampled to the original input size.

A diagram of the bottleneck architecture of IBA can be seen in Fig. 5.1.

5.6 ORDINAL EXPLANATION METHODS FOR CNN MODELS

Although the effectiveness and performance of the techniques discussed in the preceding section have been assessed in the context of conventional nominal classification tasks, it is imperative to note that neither Grad-CAM nor IBA explicitly considers the inherent order relationship between class labels in ordinal regression tasks.

In this section, we introduce two novel explanation methods tailored for CNN OBD models. These methods, built upon Grad-CAM and IBA, respectively, are specifically designed to incorporate order information into their processes. The aim is to enhance the explanatory power of the resulting maps in the context of ordinal regression, acknowledging and leveraging the ordinal relationships among class labels for improved model interpretability.

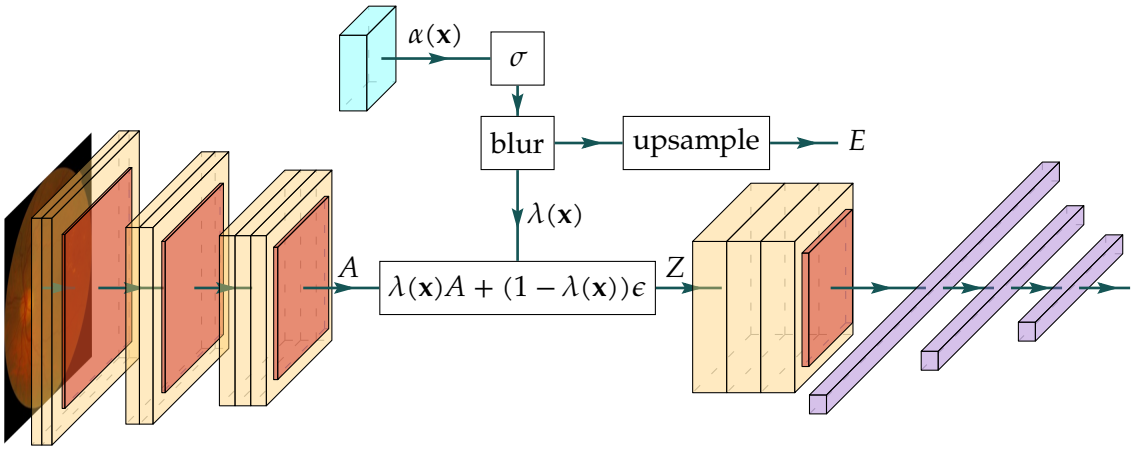


Figure 5.1: Bottleneck architecture of IBA

5.6.1 *GradOBD-CAM*

We present a novel explanation method derived from Grad-CAM, tailored to leverage gradient information concerning the activation of all output neurons p_q of the OBD model. In an OBD model designed for Q-class ordinal regression, there exist $Q-1$ output neurons, each corresponding to successive class threshold probabilities: $\mathrm{P}(y \succ C_1 \mid \mathbf{x})$, $\mathrm{P}(y \succ C_2 \mid \mathbf{x})$, ..., $\mathrm{P}(y \succ C_{Q-1} \mid \mathbf{x})$. Emphasizing feature maps contributing positively to the output probabilities $\mathrm{P}(y \succ C_q \mid \mathbf{x})$ for $y \succ C_q$ and de-emphasizing those contributing negatively for $y \preceq C_q$ is crucial. This objective introduces a new parameter ν_c^k, facilitating the incorporation of ordinality into the class activation map computation:

$$w_q^k = \sum_{c=1}^{Q-1} \nu_c^q \frac{1}{H_A \times W_A} \sum_i^{H_A} \sum_j^{W_A} \frac{\partial o_c}{\partial A^k(i,j)}, \tag{5.13}$$

$$\nu_c^q = \begin{cases} +1 \text{ if } c < q, \\ -1 \text{ if } c \geq q. \end{cases} \tag{5.14}$$

The introduction of ν_c^q ensures that the resulting class activation map aligns more naturally with the ordinal structure of the output and, consequently, adheres more closely to the principles of the OBD approach. Thus, we name this approach the Gradient OBD Class Activation Map method (GradOBD-CAM).

5.6.2 *OIBA*

Moreover, our proposed extension to the perturbation-based IBA method involves leveraging the ordinal loss inherent in the OBD model during the optimization of $E = \lambda(\mathbf{x})$. This enhancement, which we term Ordinal IBA (OIBA), aims to further refine the explanation map generation process.

The information bottleneck is comprehensively addressed by the information loss $\mathcal{L}_\mathrm{I}$ term as defined in Eq. (5.11). However, the cross-entropy loss $\mathcal{L}_\mathrm{CE}$ poses limitations in the context

of ordinal regression. To address this, we propose substituting the cross-entropy loss with the *MSE* loss $\mathcal{L}_{MSE}$, as outlined in Eq. (3.11):

$$\mathcal{L}_\lambda = \mathcal{L}_{MSE} + \beta \mathcal{L}_{\mathrm{I}}. \tag{5.15}$$

By introducing this modification, we enable the explanation map construction process to adeptly harness information from all outputs of the model in its native representation. This adjustment ensures that the Ordinal IBA method aligns more closely with the nuances of ordinal regression, offering a refined and tailored approach to generating explanation maps in the context of ordinal tasks.

5.7 EVALUATING THE PERFORMANCE OF EXPLANATION METHODS

To evaluate the influence of specific regions in an image on the classification decision, perturbation analysis is a common technique which involves occluding parts of the input images and observing the resulting changes in the model's outputs. For a given explanation denoted as E_i, the identified relevant regions are occluded and if the model's classification performance shows a pattern of decline, these regions are deemed impactful.

A method proposed by [15] provides a simple implementation of this concept. It multiplies the input $\mathbf{x}_i$ by the explanation E_i to obtain an occluded image $\tilde{\mathbf{x}}_i$:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \circ E_i, \tag{5.16}$$

where $\circ$ denotes element-wise multiplication. The average drop in the score of the target class C_q, denoted as $f^q(\mathbf{x})$, is calculated for instances where the score decreases:

$$\text{Average drop} = \frac{1}{N} \sum_i^N \frac{\max(0, f^q(\mathbf{x}_i) - f^q(\tilde{\mathbf{x}}_i))}{f^q(\mathbf{x}_i)}. \tag{5.17}$$

The goal is to minimize the average drop, as a good explanation should result in minimal score reduction. However, this metric does not consider the ordering information among class labels. Specifically, as the confidence in the target class decreases, it is preferable for the confidence in nearby classes to increase rather than distant ones, a factor not captured by the average drop metric.

An alternative method proposed by [84] involves dividing the explanation map into tiles (e. g. 8×8) and ranking them based on the total sum of relevance within each tile. The input image is then occluded tile by tile, starting from the most relevant and progressing to the least relevant. This generates the Most Relevant Features (MoRF) curve, plotting the target class score against the level of image degradation (i. e. the number of occluded tiles). A meaningful explanation map is expected to result in a sharp decrease in score at the beginning when the most relevant parts of the image are occluded. The same procedure in reverse order of relevance produces the Least Relevant Features (LeRF) curve, where the score should not significantly drop until the most relevant parts are occluded at the end. Normalizing the extremes of these curves between 0 and 1 and computing the signed area between them yields the degradation score. A substantial area between the MoRF and LeRF curves indicates a relevant explanation map, which makes this a *maximization score*. An illustrative example is presented in Fig. 5.2.

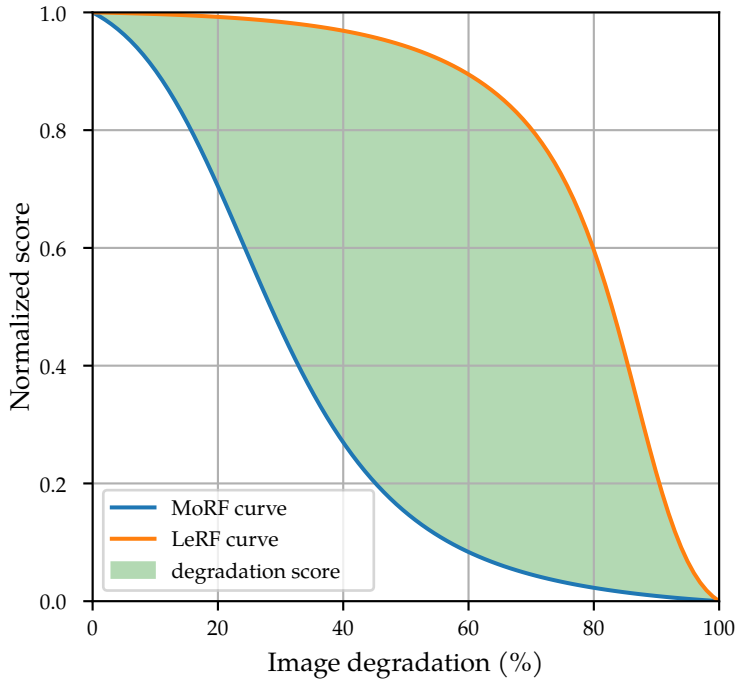


Figure 5.2: Example of MoRF and LeRF curves and the area between them, which constitutes the degradation score

Note how there may be undesirable instances where the MoRF curve raises above the LeRF curve. In cases like this, the area above the LeRF curve and below the MoRF curve is considered negative, reducing the degradation score. This sometimes leads to negative results.

This approach offers an advantage by enabling the study of the behaviour of any metric, not limited to the target score. Thus, we propose examining the degradation of the following classification performance metrics defined in Section 1.2.5, most of them specific to ordinal regression:

- For general classification performance:
  - Correct Classification Rate (*CCR*).
- For ordinal regression performance:
  - Mean Absolute Error (*MAE*).
  - Quadratic weighted Cohen's kappa (κ).
  - Spearman's rank correlation coefficient (r_s).
- For class balancing performance:
  - Average area under the ROC curve (*AvAUC*).

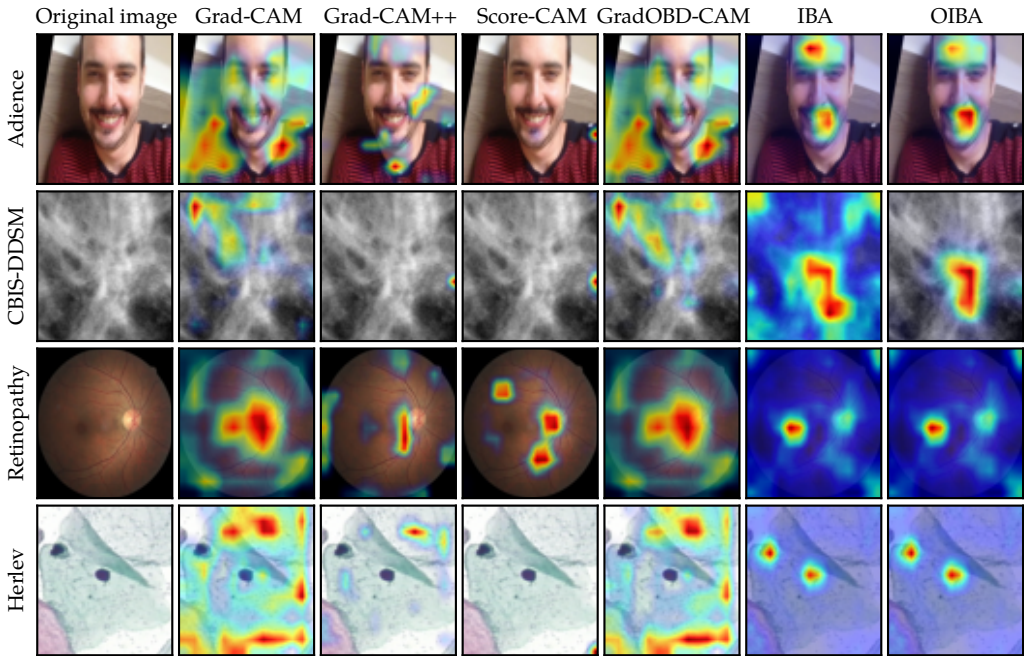


Figure 5.3: Example explanations maps for each dataset and explanation method

5.8 EXPERIMENT DESIGN

We compare the performance of the already existing methods described in Section 5.5 with the ordinal versions described in Section 5.6 when applied to an ordinal OBD model with ECOC decision rule, which was introduced in Section 3.4.3. We do this to evaluate the performance difference when assumptions about ordinality are introduced into the explanation method.

A ResNet34 CNN model, initially pre-trained on ImageNet-1K [24], serves as the foundation for this study. The model undergoes training on the four distinct ordinal regression datasets described in Section 3.5.1.

This is done under 60 distinct random initializations of the model parameters as well as a corresponding random 80/10/10 split for training, validation, and testing. The training employs the Adam back-propagation method [52], with batches of 64 samples, lasting a maximum of 200 epochs. Early stopping occurs when the validation set's loss fails to improve for 20 consecutive epochs. Subsequently, six explanation methods are applied: the three CAM methods, namely Grad-CAM [85], Grad-CAM++ [15] and Score-CAM [104], are compared against GradOBD-CAM (from Section 5.6.1) and IBA [84] is compared against OIBA (from Section 5.6.2). The degradation of various metrics is assessed following the guidelines in Section 5.7. These metrics are *CCR*, *AvAUC*, *MAE*, κ, and r_s.

5.9 RESULTS

In Table 5.1 mean results for each metric degradation and dataset are shown. Some example of the explanation maps are shown in Fig. 5.3.

The overall performance of GradOBD-CAM demonstrates superiority over Grad-CAM++ and Score-CAM, while also being comparable to or better than Grad-CAM across various datasets. Notably, the Adience dataset exhibits the most significant performance difference,

	CCR		*MAE*		κ		r_s		*AvAUC*	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Adience										
Grad-CAM	−0.0363	0.0504	−0.0350	0.0565	−0.0282	0.0476	−0.0232	0.0475	−0.0154	0.0374
Grad-CAM++	−0.1343	0.0456	−0.1202	0.0757	−0.0917	0.0495	−0.0750	0.0557	−0.1001	0.0343
Score-CAM	−0.1723	0.0449	−0.1580	0.0756	−0.1264	0.0508	−0.1095	0.0580	−0.1371	0.0352
GradOBD-CAM	**0.0273**	0.0471	**0.0088**	0.0478	**−0.0016**	0.0504	**0.0042**	0.0495	**0.0247**	0.0417
IBA	**0.1296**	0.0301	0.1548	0.0376	0.1653	0.0333	0.1611	0.0336	0.1410	0.0223
OIBA	0.1251	0.0281	**0.1614**	0.0363	**0.1767**	0.0322	**0.1722**	0.0330	**0.1474**	0.0230
CBIS-DDSM										
Grad-CAM	0.4269	0.4855	0.3902	1.6636	0.1957	0.1255	**0.1812**	0.1287	**0.1799**	0.1079
Grad-CAM++	−0.0197	0.1646	−0.0084	0.2074	0.0098	0.0575	0.0007	0.0578	−0.0069	0.0523
Score-CAM	−0.0734	0.1586	−0.0696	0.1919	−0.0118	0.0665	−0.0219	0.0628	−0.0367	0.0617
GradOBD-CAM	**0.4285**	0.4820	**1.7691**	7.3760	**0.1974**	0.1381	0.1780	0.1394	0.1709	0.1256
IBA	**1.9412**	3.9603	1.9521	4.9261	0.5361	0.1554	0.5485	0.1602	**0.5244**	0.1105
OIBA	1.7439	3.3878	**2.2170**	5.2769	**0.5499**	0.1385	**0.5577**	0.1501	0.4713	0.1039
Retinopathy										
Grad-CAM	1.0257	1.0351	0.5367	0.4947	0.2196	0.0967	**0.2899**	0.1382	0.2103	0.0912
Grad-CAM++	−0.1231	0.4273	−0.1043	0.2267	0.0478	0.0545	0.0522	0.0601	0.0357	0.0572
Score-CAM	0.0621	0.4485	0.0330	0.1957	0.0170	0.0486	0.0130	0.0373	−0.0059	0.0290
GradOBD-CAM	**1.0476**	1.0501	**0.5515**	0.5020	**0.2237**	0.0954	0.2898	0.1362	**0.2155**	0.0917
IBA	0.3918	0.8463	0.2165	0.3509	0.1660	0.0768	0.2303	0.0690	0.2172	0.0603
OIBA	**0.5082**	0.8553	**0.2829**	0.3619	**0.2191**	0.1013	**0.2668**	0.1310	**0.2915**	0.1162
Herlev										
Grad-CAM	0.1876	0.0951	0.1638	0.0902	0.1318	0.0730	0.1377	0.0753	0.2121	0.1309
Grad-CAM++	0.0162	0.0646	0.0133	0.0696	0.0130	0.0685	0.0143	0.0676	0.0026	0.0656
Score-CAM	−0.0056	0.0694	−0.0056	0.0716	−0.0087	0.0674	−0.0024	0.0673	−0.0083	0.0772
GradOBD-CAM	**0.2149**	0.0930	**0.1884**	0.0969	**0.1479**	0.0786	**0.1565**	0.0847	**0.2712**	0.1720
IBA	**0.1341**	0.0913	**0.1292**	0.0981	**0.1145**	0.0787	**0.1077**	0.0804	**0.1142**	0.1056
OIBA	0.1315	0.0984	0.1292	0.1055	0.1141	0.0845	0.1036	0.0859	0.1067	0.1099

Table 5.1: Summary of the results of the experimentation for each dataset. Columns correspond to the degradation of each metric, as explained in Section 5.7. Best results in each group of explanation methods for each dataset are highlighted in bold

although other datasets also display discernible variations. Concerning IBA methods, OIBA consistently outperforms IBA across all metrics in the Adience and Retinopathy datasets, with marginal or negligible differences in performance observed for the CBIS-DDSM and Herlev datasets.

It is also of note that some negative results are obtained. This could be explained by some of the studied datasets requiring more complex or abstract explanations which are more difficult to produce.

To validate these results rigorously, statistical hypothesis testing is conducted. A one-sided Wilcoxon signed-rank test is executed for each metric and dataset to assess whether the performance rank difference with GradOBD-CAM is either symmetrical about zero or skewed towards the other method (null hypothesis) or significantly skewed towards GradOBD-CAM (alternative hypothesis). Similarly, the same test is performed to compare OIBA with IBA. A standard significance level of $\alpha = 0.05$ is applied to all tests, and the outcomes are detailed in Table 5.2.

The test results reveal that GradOBD-CAM consistently exhibits superior median performance across all metrics and datasets compared to both Grad-CAM++ and Score-CAM. Compared to the original Grad-CAM, performance varies notably across datasets. GradOBD-CAM shows good performance in the Adience, Retinopathy and Herlev datasets, although it underperforms in CBIS-DDSM.

Concerning OIBA, it attains a higher median performance in two out of the four tested datasets (Adience and Retinopathy). In all other cases, the performance remains consistent across all metrics without significant drops.

5.10 CONCLUSIONS

This chapter introduced and evaluated two novel explanation methods, GradOBD-CAM and OIBA, tailored for interpreting ordinal regression tasks in the context of the OBD model. GradOBD-CAM adapts the Grad-CAM attribution technique to the ordinal regression scenario, leveraging the distinctive characteristics of the OBD model. On the other hand, OIBA incorporates the ordinal loss function native to the OBD model into the explanation process.

We proposed an explanation evaluation procedure focusing on the degradation score of ordinal metrics through the occlusion of regions deemed most/least relevant. Our findings underscore the superiority of GradOBD-CAM over all its CAM counterparts, demonstrating statistically significant improvements in three out of four datasets, presenting compelling evidence for its efficacy. OIBA, our second proposed method, consistently outperformed IBA across all ordinal metrics in two datasets, showcasing its effectiveness in enhancing interpretability.

	GradOBD-CAM vs. Grad-CAM		GradOBD-CAM vs. Grad-CAM++		GradOBD-CAM vs. Score-CAM		OIBA vs. IBA	
	T^+	p-value	T^+	p-value	T^+	p-value	T^+	p-value
Adience								
CCR	**1816**	**<0.001**	**1830**	**<0.001**	**1830**	**<0.001**	213	0.999
MAE	**1747**	**<0.001**	**1816**	**<0.001**	**1828**	**<0.001**	**1730**	**<0.001**
κ	**1691**	**<0.001**	**1806**	**<0.001**	**1826**	**<0.001**	**1797**	**<0.001**
r_s	**1667**	**<0.001**	**1775**	**<0.001**	**1819**	**<0.001**	**1749**	**<0.001**
AvAUC	**1736**	**<0.001**	**1830**	**<0.001**	**1830**	**<0.001**	**1627**	**<0.001**
CBIS-DDSM								
CCR	870	0.630	**1811**	**<0.001**	**1816**	**<0.001**	617	0.986
MAE	**1342**	**0.001**	**1810**	**<0.001**	**1814**	**<0.001**	703	0.941
κ	1007	0.249	**1791**	**<0.001**	**1789**	**<0.001**	1136	0.052
r_s	892	0.567	**1768**	**<0.001**	**1782**	**<0.001**	1065	0.135
AvAUC	773	0.852	**1816**	**<0.001**	**1822**	**<0.001**	244	0.999
Retinopathy								
CCR	**1525**	**<0.001**	**1828**	**<0.001**	**1766**	**<0.001**	**1451**	**<0.001**
MAE	**1670**	**<0.001**	**1825**	**<0.001**	**1762**	**<0.001**	**1684**	**<0.001**
κ	**1578**	**<0.001**	**1823**	**<0.001**	**1823**	**<0.001**	**1751**	**<0.001**
r_s	909	0.518	**1828**	**<0.001**	**1826**	**<0.001**	**1335**	**0.001**
AvAUC	**1448**	**<0.001**	**1827**	**<0.001**	**1828**	**<0.001**	**1760**	**<0.001**
Herlev								
CCR	**1520**	**<0.001**	**1830**	**<0.001**	**1830**	**<0.001**	836	0.720
MAE	**1568**	**<0.001**	**1828**	**<0.001**	**1830**	**<0.001**	892	0.567
κ	**1458**	**<0.001**	**1808**	**<0.001**	**1823**	**<0.001**	938	0.433
r_s	**1487**	**<0.001**	**1812**	**<0.001**	**1826**	**<0.001**	824	0.749
AvAUC	**1571**	**<0.001**	**1830**	**<0.001**	**1830**	**<0.001**	708	0.936

Table 5.2: One-sided Wilcoxon test results for each metric. In cases where the p-value is less than $\alpha = 0.05$ (highlighted in bold) the performance difference significantly favours the first method

6 ADDITIONAL WORKS

During the development of this thesis other related additional works were developed alongside several other authors. This chapter contains a brief summary of the contributions of each one.

6.1 ACTIVATION FUNCTIONS FOR CNNS

ASSOCIATED PUBLICATION: Víctor Manuel Vargas, Pedro Antonio Gutiérrez, **Javier Barbero-Gómez** and César Hervás-Martínez. 'Activation Functions for Convolutional Neural Networks: Proposals and Experimental Study'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021), pp. 1–11. DOI: 10.1109/TNNLS.2021.3105444.

As discussed in Section 1.5, CNNs are constructed by assembling various computations known as layers, primarily involving convolutions, pooling operations, and fully connected layers. To introduce non-linearity into the model, a transformation is applied to the output of these layers, known as activation functions.

The functions used in the preceding chapters along with some widely adopted ones, are summarized in Section 1.4.1. However, a diverse range of proposed alternatives exists in the literature. This work introduces two alternative activation functions based on the softplus, defined in Eq. (1.32), and the ELU, defined in Eq. (1.31). More specifically a parametrized version of the softplus and a linear combination of the softplus and ELU.

These alternatives are then compared against 19 other previously proposed functions, encompassing various variations of the ReLU and softplus functions. A classification of these functions is presented based on several aspects:

- Presence of negative activation ('leakiness').
- Presence of learnable parameters ('parametric').
- Presence of a random component ('randomized').

Drawing from the work of [91], a discussion is conducted on two desirable properties of activation functions: scale invariance and 1-Lipschitz continuity.

Following extensive experimentation across six different image classification tasks (with a focus on the ILSVRC dataset [80]) using two distinct CNN architectures, it is concluded that, apart from a few worst-performing ones, most exhibit similar performance overall. Notably, for the large-scale dataset, the two proposed functions demonstrate significantly superior performance compared to ReLU and ELU, the two most commonly used activation functions.

6.2 SOFT LABELLING BASED ON TRIANGULAR DISTRIBUTIONS

ASSOCIATED PUBLICATION: Víctor Manuel Vargas, Pedro Antonio Gutiérrez, **Javier Barbero-Gómez** and César Hervás-Martínez. 'Soft Labelling Based on Triangular Distributions

for Ordinal Classification'. In: *Information Fusion* 93 (1st May 2023), pp. 258–267. DOI: 10.1016/j.inffus.2023.01.003.

In complex real-world scenarios, such as medical images, the assignment of labels involves the input of multiple expert opinions [90]. In situations where only the combined outcome of opinions is accessible, without knowledge of the underlying probability distribution, multi-label classification methods become impractical.

To tackle this issue, a soft labelling approach, briefly introduced in Section 1.2.4 and named *unimodal regularization*, is proposed. This approach assumes that expert opinions for ordinal problems are distributed over an interval centred around the ground truth label. Previous works have explored the use of various discrete and continuous distributions, such as truncated Poisson, binomial [9], exponential function followed by softmax [57], and beta distribution [100], to generate soft labels. These methods aim to adjust the encoding of ordinal labels to improve loss function computation when only aggregated labels are available.

Building on the successes of earlier approaches in addressing ordinal problems, this study introduces a novel unimodal regularization technique that utilizes a combination of triangular distributions. The main advantage of this method lies in its simplicity, requiring adjustment of only one parameter that determines the upper limit of the error within adjacent classes. Additionally, the method outlines a procedure to compute all triangular distribution parameters based on this single adjustable parameter.

A series of experiments is designed to evaluate the performance of this approach on six different ordinal regression tasks, comparing it to other unimodal regularization techniques and a case with no regularization. The results of the experiments indicate that the use of the triangular distribution led to improved outcomes, surpassing both the baseline and other soft labelling methodologies. Furthermore, a statistical analysis confirmed the significance of this enhancement.

Part III

CONCLUSIONS

7 CONCLUSIONS AND FUTURE WORK

7.1 SUMMARY OF CONTRIBUTIONS

This thesis addressed the challenges and opportunities presented by ordinal regression tasks in the realm of deep learning, particularly focusing on CNNs. The contributions can be summarized as follows.

7.1.1 *Native ordinal representations for CNNs*

In Chapter 3, we proposed a novel output architecture for CNNs designed explicitly for ordinal regression tasks. The OBD model, coupled with a matching class assignment rule based on the ECOC scheme, demonstrated superior performance in comparison to traditional nominal methods and existing ordinal techniques across various datasets. Importantly, our methodology provided enhanced *RMSE* performance without compromising other metrics, highlighting its adaptability and efficiency.

7.1.2 *An application of ordinal regression techniques: computer-aided diagnosis for Parkinson's disease*

Chapter 4 extended our ordinal CNN architecture to address the unique challenges posed by 3D images in the context of diagnosing PD. The proposed methodology, incorporating a native 3D CNN architecture and data augmentation procedure, showcased significant improvements in the assessment of altered dopaminergic brain activity. The versatility of the methodology, applicable to different input types, positions it as a promising tool for ordinal tasks, especially in medical domains.

7.1.3 *Examining the decision process of ordinal CNNs*

Chapter 5 focused on the explainability of ordinal CNN models, a crucial aspect often overlooked in the literature. We introduced and validated two novel explanation methods, GradOBD-CAM and OIBA, designed specifically for ordinal regression tasks. Our results demonstrated the superiority of GradOBD-CAM over existing methods, offering a robust solution for explaining the decision-making process of ordinal CNN models.

7.2 ACHIEVEMENT OF PROPOSED GOALS

GOAL 1: DEVELOPMENT OF NEW CNN ARCHITECTURES FOR ORDINAL REGRESSION The primary goal of developing new CNN architectures capable of handling ordinal response variables natively has been successfully achieved. The introduction of the OBD model, designed specifically for ordinal regression tasks, showcased its effectiveness across various datasets. This novel architecture demonstrated improved ordinal performance metrics

without compromising other classification metrics, addressing the unique challenges presented by ordinal relationships between class labels.

GOAL 2: PROPOSAL OF NEW DATA AUGMENTATION TECHNIQUES The goal of proposing innovative data augmentation techniques to address class imbalance issues in ordinal regression tasks was realized through the introduction of the OGO-SP-β ordinal augmentation algorithm. This technique, which exploits ordinal information, demonstrated its effectiveness in mitigating class imbalance challenges. The adaptability and performance enhancements observed across different datasets underscore the success in achieving this goal.

GOAL 3: APPLICATION TO REAL MEDICAL IMAGE DIAGNOSIS PROBLEMS The application of the developed methodologies to real medical image diagnosis problems, including both 2D and 3D image datasets, has been a central focus of this research. The successful application of the proposed OBD model to diagnose neurological damage in Parkinson's disease using volumetric brain scans demonstrates the practical utility of the developed architectures. The methodology's effectiveness in addressing the challenges posed by medical image data reaffirms the accomplishment of this goal.

GOAL 4: PROPOSAL OF NEW EXPLANATION METHODS FOR ORDINAL INFORMATION The goal of proposing new explanation methods that incorporate ordinal information to improve the detection of relevant input features has been realized through the introduction of GradOBD-CAM and OIBA. These methods offer insights into the decision-making process of ordinal CNNs and outperform existing methods in various datasets. The successful development of these explanation methods attests to the achievement of this goal, enhancing the interpretability of ordinal regression tasks.

In conclusion, the proposed goals have been successfully achieved, demonstrating the effectiveness of the developed methodologies in addressing the unique challenges posed by ordinal regression tasks, particularly in medical image diagnosis problems. These accomplishments contribute to the advancement of deep learning applications in domains where ordinal relationships between classes play a crucial role.

7.3 OVERALL CONCLUSIONS

This thesis presents a comprehensive exploration of ordinal regression tasks within the deep learning paradigm, emphasizing the development of tailored methodologies and addressing the interpretability challenges. The proposed OBD model, along with its adaptations for 3D image datasets, has shown consistent improvements in performance across various metrics. Furthermore, the introduced explanation methods, GradOBD-CAM and OIBA, contribute significantly to understanding the decision process of ordinal CNNs, offering insights into their inner workings.

The successful adaptation of ordinal methodologies to diverse datasets and tasks underscores the versatility and generalizability of the proposed approaches. Future research can explore enhancements in addressing class imbalance challenges and further investigations into interpretability methods, potentially extending their application to other ordinal regression models.

In conclusion, this thesis not only contributes to the theoretical understanding of ordinal regression in deep learning but also provides practical solutions and methodologies that can find applicability across different domains. The achieved improvements in performance and interpretability pave the way for more robust and transparent applications of deep learning in ordinal regression scenarios.

7.4 FUTURE WORK

The future exploration of ordinal regression methodologies in deep learning opens avenues for refinement and extension. The following are promising opportunities opened up by the research carried out in this thesis.

7.4.1 *Ordinal CNN architectures*

Investigating the impact of different ECOC configurations on the ordinal regression performance could enhance the adaptability of the method across diverse datasets. Additionally, exploring the integration of ECOC with other ordinal regression architectures may contribute to a broader understanding of its applicability and effectiveness.

Furthermore, the inclusion of ordinal structure in the model architecture is not limited to the output stage. New computation layers can be designed to take into account the ordinal structure of the output, as well as the possible ordinality of input and latent features. These could guide the training process (e. g. a dropout procedure) or intrinsically alter the information flow along the network.

7.4.2 *Computer-aided diagnosis (CAD)*

New advancements in the automatic diagnosis of Parkinson's disease can be pursued through expanded data acquisition efforts. Collecting additional ordinal task datasets and exploring those publicly available for ordinal information could provide a more comprehensive understanding of the model's performance across diverse populations.

Additionally, there is promise in exploring 3D ordinal applications beyond the medical field, allowing for a broader assessment of the methodology's efficacy in varied contexts. This could open up the possibility of transfer learning (i. e. the use of pre-trained models as a starting point for a potential convergence and performance improvement) to medical 3D image analysis tasks, which is currently very limited.

Regarding the deployment of these models into comprehensive CAD systems, further performance improvements are needed to meet the reliability demands of the medical field. These improvements could potentially alleviate the workload on medical professionals and enhance the overall diagnostic quality.

7.4.3 *Explainability of ordinal models*

The development of improved explainability methods remains an ongoing area of research. Future work can focus on enhancing the robustness and generalizability of GradOBD-CAM and OIBA. Exploring ways to make these methods more adaptive to different ordinal regression architectures would contribute to their versatility. Additionally, investigating

the potential of GradOBD-CAM to promote ordinality within the latent feature space offers a promising direction, opening new possibilities for understanding how models capture ordinal relationships in their internal representations.

Furthermore, considering the inherent challenges associated with ordinal regression tasks, future research could delve into the exploration of novel explanation methods specifically tailored to address the nuances of ordinal data. This could involve the development of techniques that provide more insights into the decision-making process by including the explanation generation procedure into the model alongside the classification procedure, something known as an 'in-model' or 'intrinsic' approach. Additionally, the exploration of visualization tools and metrics tailored for ordinal regression could enhance the interpretability of models in this domain.

BIBLIOGRAPHY

[1] Kemal Akyol. 'Stacking Ensemble Based Deep Neural Networks Modeling for Effective Epileptic Seizure Detection'. In: *Expert Systems with Applications* 148 (15th June 2020), p. 113239. DOI: 10.1016/j.eswa.2020.113239.

[2] Erin L. Allwein, Robert E. Schapire and Yoram Singer. 'Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers'. In: *Journal of Machine Learning Research* 1 (2000), pp. 113–141.

[3] J. Arbizu et al. 'Functional neuroimaging in the diagnosis of patients with parkinsonism: Update and recommendations for clinical use'. In: *Revista Española de Medicina Nuclear e Imagen Molecular (English Edition)* 33.4 (2014), pp. 215–226. DOI: 10.1016/j.remnie.2014.05.002. pmid: 24731551.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller and Wojciech Samek. 'On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation'. In: *PLOS ONE* 10.7 (10th July 2015), e0130140. DOI: 10.1371/journal.pone.0130140.

[5] Javier Barbero-Gómez, Ricardo Cruz, Jaime S. Cardoso, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Evaluating the Performance of Explanation Methods on Ordinal Regression CNN Models'. In: International Work-Conference on Artificial Neural Networks (IWANN 2023). Ponta Delgada, Portugal, June 2023.

[6] Javier Barbero-Gómez, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Error-Correcting Output Codes in the Framework of Deep Ordinal Classification'. In: *Neural Processing Letters* (12th May 2022). DOI: 10.1007/s11063-022-10824-7.

[7] Javier Barbero-Gómez, Pedro-Antonio Gutiérrez, Víctor-Manuel Vargas, Juan-Antonio Vallejo-Casas and César Hervás-Martínez. 'An Ordinal CNN Approach for the Assessment of Neurological Damage in Parkinson's Disease Patients'. In: *Expert Systems with Applications* 182 (15th Nov. 2021), p. 115271. DOI: 10.1016/j.eswa.2021.115271.

[8] Alejandro Barredo Arrieta et al. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI'. In: *Information Fusion* 58 (1st June 2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[9] Christopher Beckham and Christopher Pal. 'Unimodal Probability Distributions for Deep Ordinal Classification'. In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 17th July 2017, pp. 411–419.

[10] Arie Ben-David. 'Comparison of Classification Accuracy Using Cohen's Weighted Kappa'. In: *Expert Systems with Applications* 34.2 (1st Feb. 2008), pp. 825–832. DOI: 10.1016/j.eswa.2006.10.022.

[11] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li and Xiangyu Zhang. *Reversible Column Networks*. Version 3. 1st Feb. 2023. arXiv: 2212.11696 [cs].

[12] J Camacho-Canamon, M Guiote Moreno, A Santos Buenos, E Rodriguez Caceres, E Carmona Asenjo, J Vallejo Casas, P A Gutierrez and C Hervas-Martinez. 'Image Classification of Synaptic Dopamine Transporters I-123-Ioflupane by Machine Learning Techniques'. In: *Proceedings of the 2017 Annual Congress of the European Association of Nuclear Medicine (EANM17)*. Annual Congress of the European Association of Nuclear Medicine. Vol. 44. Vienna, Austria: Springer, 2017, S285–S286. DOI: doi.org/10.1007/s00259-017-3822-1.

[13] Wenzhi Cao, Vahid Mirjalili and Sebastian Raschka. 'Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation'. In: *Pattern Recognition Letters* 140 (1st Dec. 2020), pp. 325–331. DOI: 10.1016/j.patrec.2020.11.008.

[14] Jaime S. Cardoso and Ricardo Sousa. 'Measuring the Performance of Ordinal Classification'. In: *International Journal of Pattern Recognition and Artificial Intelligence* 25.08 (1st Dec. 2011), pp. 1173–1195. DOI: 10.1142/S0218001411009093.

[15] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader and Vineeth N. Balasubramanian. 'Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks'. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097. arXiv: 1710.11063 [cs].

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. 'SMOTE: Synthetic Minority Over-sampling Technique'. In: *Journal of Artificial Intelligence Research* 16 (1st June 2002), pp. 321–357. DOI: 10.1613/jair.953.

[17] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le and Mike Rao. 'Using Ranking-CNN for Age Estimation'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5183–5192.

[18] Jianlin Cheng, Zheng Wang and Gianluca Pollastri. 'A Neural Network Approach to Ordinal Regression'. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, June 2008, pp. 1279–1284. ISBN: 978-1-4244-1820-6. DOI: 10.1109/IJCNN.2008.4633963.

[19] Wei Chu and S. Sathiya Keerthi. 'Support Vector Ordinal Regression'. In: *Neural Computation* 19.3 (1st Mar. 2007), pp. 792–815. DOI: 10.1162/neco.2007.19.3.792.

[20] Jacob Cohen. 'A Coefficient of Agreement for Nominal Scales'. In: *Educational and Psychological Measurement* 20.1 (1st Apr. 1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[21] Jacob Cohen. 'Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit'. In: *Psychological Bulletin* 70.4 (1968), pp. 213–220. DOI: 10.1037/h0026256.

[22] Jacques Darcourt, Jan Booij, Klaus Tatsch, Andrea Varrone, Thierry Vander Borght, Ozlem L. Kapucu, Kjell Någren, Flavio Nobili, Zuzana Walker and Koen Van Laere. 'EANM Procedure Guidelines for Brain Neurotransmission SPECT Using (123)I-labelled Dopamine Transporter Ligands, Version 2'. In: *European Journal of Nuclear Medicine and Molecular Imaging* 37.2 (Feb. 2010), pp. 443–450. DOI: 10.1007/s00259-009-1267-x. pmid: 19838702.

[23] Jordi de la Torre, Domenec Puig and Aida Valls. 'Weighted Kappa Loss Function for Multi-Class Classification of Ordinal Data in Deep Learning'. In: *Pattern Recognition Letters* 105 (2018), pp. 144–154. DOI: 10.1016/j.patrec.2017.05.018.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei. 'ImageNet: A Large-Scale Hierarchical Image Database'. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[25] Wan-Yu Deng, Qing-Hua Zheng, Shiguo Lian, Lin Chen and Xin Wang. 'Ordinal Extreme Learning Machine'. In: *Neurocomputing*. Artificial Brains 74.1 (1st Dec. 2010), pp. 447–456. DOI: 10.1016/j.neucom.2010.08.022.

[26] E. W. Dijkstra. 'A Note on Two Problems in Connexion with Graphs'. In: *Numerische Mathematik* 1.1 (1st Dec. 1959), pp. 269–271. DOI: 10.1007/BF01386390.

[27] M. Dorado-Moreno, P. A. Gutiérrez, L. Cornejo-Bueno, L. Prieto, S. Salcedo-Sanz and C. Hervás-Martínez. 'Ordinal Multi-class Architecture for Predicting Wind Power Ramp Events Based on Reservoir Computing'. In: *Neural Processing Letters* 52.1 (1st Aug. 2020), pp. 57–74. DOI: 10.1007/s11063-018-9922-5.

[28] John Duchi, Elad Hazan and Yoram Singer. 'Adaptive Subgradient Methods for Online Learning and Stochastic Optimization'. In: *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159.

[29] Eran Eidinger, Roee Enbar and Tal Hassner. 'Age and Gender Estimation of Unfiltered Faces'. In: *IEEE Transactions on Information Forensics and Security* 9.12 (Dec. 2014), pp. 2170–2179. DOI: 10.1109/TIFS.2014.2359646.

[30] Imanne El Maachi, Guillaume-Alexandre Bilodeau and Wassim Bouachir. 'Deep 1D-Convnet for Accurate Parkinson Disease Detection and Severity Prediction from Gait'. In: *Expert Systems with Applications* 143 (1st Apr. 2020), p. 113075. DOI: 10.1016/j.eswa.2019.113075.

[31] Alan C. Evans, Andrew L. Janke, D. Louis Collins and Sylvain Baillet. 'Brain Templates and Atlases'. In: *NeuroImage*. 20 YEARS OF fMRI 62.2 (15th Aug. 2012), pp. 911–922. DOI: 10.1016/j.neuroimage.2012.01.024.

[32] F. Fernández-Navarro. 'A Generalized Logistic Link Function for Cumulative Link Models in Ordinal Regression'. In: *Neural Processing Letters* 46.1 (1st Aug. 2017), pp. 251–269. DOI: 10.1007/s11063-017-9589-3.

[33] Leslie J. Findley. 'The Economic Impact of Parkinson's Disease'. In: *Parkinsonism & Related Disorders*. Parkinson's Disease: From Premotor Symptoms to New Treatment Strategies 13 (1st Sept. 2007), S8–S12. DOI: 10.1016/j.parkreldis.2007.06.003.

[34] R. A. Fisher. 'Theory of Statistical Estimation'. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 22.5 (July 1925), pp. 700–725. DOI: 10.1017/S0305004100009580.

[35] Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. 20th ed. Edinburgh: Oliver and Boyd, 1954.

[36] Raúl de la Fuente-Fernández. 'Role of DaTSCAN and Clinical Diagnosis in Parkinson Disease'. In: *Neurology* 78.10 (6th Mar. 2012), pp. 696–701. DOI: 10.1212/WNL.0b013e318248e520. pmid: 22323748.

[37] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. *Generative Adversarial Networks*. 10th June 2014. arXiv: 1406.2661.

[38] Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro and César Hervás-Martínez. 'Ordinal Regression Methods: Survey and Experimental Study'. In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (2016), pp. 127–146. DOI: 10.1109/TKDE.2015.2457911.

[39] Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ: Wiley, 2013. ISBN: 978-1-118-07462-6.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep Residual Learning for Image Recognition*. 10th Dec. 2015. arXiv: 1512.03385 [cs].

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 6th Feb. 2015. arXiv: 1502.01852 [cs].

[42] Geoffrey Hinton, Nitish Srivastava and Kevin Swersky. 'Neural Networks for Machine Learning, Lecture 6a: Overview of Mini-batch Gradient Descent'. Lecture slides.

[43] Geoffrey E. Hinton and Richard S. Zemel. 'Autoencoders, Minimum Description Length and Helmholtz Free Energy'. In: *6th International Conference on Neural Information Processing Systems*. NIPS'93. 29th Nov. 1993, pp. 3–10.

[44] Andrew Howard et al. *Searching for MobileNetV3*. 20th Nov. 2019. arXiv: 1905.02244 [cs].

[45] Wenjun Hui, Chuangchuang Tan, Guanghua Gu and Yao Zhao. 'Gradient-Based Refined Class Activation Map for Weakly Supervised Object Localization'. In: *Pattern Recognition* 128 (1st Aug. 2022), p. 108664. DOI: 10.1016/j.patcog.2022.108664.

[46] Ian Goodfellow, Yoshua Bengio and Aaron Courville. 'Deep Feedforward Networks'. In: *Deep Learning*. Cambridge, MA: MIT Press, 2017. ISBN: 978-0-262-03561-3.

[47] Jan Jantzen, Jonas Norup, George Dounias and Beth Bjerregaard. 'Pap-Smear Benchmark Data For Pattern Classification'. In: *Nature Inspired Smart Information Systems (NiSIS)* (1st Jan. 2005).

[48] Justin M. Johnson and Taghi M. Khoshgoftaar. 'Survey on Deep Learning with Class Imbalance'. In: *Journal of Big Data* 6.1 (19th Mar. 2019), p. 27. DOI: 10.1186/s40537-019-0192-5.

[49] E. S. Keeping. 'The Beta Distribution'. In: *Introduction to Statistical Inference*. Dover Publications, 1995, pp. 83–85. ISBN: 978-0-486-68502-1.

[50] M. G. Kendall. 'A New Measure of Rank Correlation'. In: *Biometrika* 30.1/2 (1938), pp. 81–93. DOI: 10.2307/2332226. JSTOR: 2332226.

[51] M. G. Kendall. 'The Treatment of Ties in Ranking Problems'. In: *Biometrika* 33.3 (1945), pp. 239–251. DOI: 10.2307/2332303. JSTOR: 2332303.

[52] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 29th Jan. 2017. arXiv: 1412.6980 [cs].

[53] Sotiris B. Kotsiantis and Panagiotis E. Pintelas. 'A Cost Sensitive Technique for Ordinal Classification Problems'. In: *Methods and Applications of Artificial Intelligence*. Lecture Notes in Computer Science. 2004, pp. 220–229. ISBN: 978-3-540-24674-9. DOI: 10.1007/978-3-540-24674-9_24.

[54] Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer and Michael de Groeve. 'Prediction of Ordinal Classes Using Regression Trees'. In: *Foundations of Intelligent Systems*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 426–434. ISBN: 978-3-540-39963-6. DOI: 10.1007/3-540-39963-1_45.

[55] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy and Daniel L. Rubin. 'A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research'. In: *Scientific Data* 4.1 (1 19th Dec. 2017), p. 170177. DOI: 10.1038/sdata.2017.177.

[56] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus and Shimon Schocken. 'Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function'. In: *Neural Networks* 6.6 (1st Jan. 1993), pp. 861–867. DOI: 10.1016/S0893-6080(05)80131-5.

[57] Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu and Jane You. 'Unimodal Regularized Neuron Stick-Breaking for Ordinal Classification'. In: *Neurocomputing* 388 (7th May 2020), pp. 34–44. DOI: 10.1016/j.neucom.2020.01.025.

[58] Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You and B. V. K Vijaya Kumar. 'Ordinal Regression with Neuron Stick-breaking for Medical Diagnosis'. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.

[59] Y. Liu, A.W.K. Kong and C.K. Goh. 'Deep Ordinal Regression Based on Data Relationship for Small Datasets'. In: IJCAI International Joint Conference on Artificial Intelligence. Vol. 0. 2017, pp. 2372–2378. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/330.

[60] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng and Jian Sun. *ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design*. 30th July 2018. arXiv: 1807.11164 [cs].

[61] Silvia Marino, Rosella Ciurleo, Giuseppe Di Lorenzo, Marina Barresi, Simona De Salvo, Sabrina Giacoppo, Alessia Bramanti, Pietro Lanzafame and Placido Bramanti. 'Magnetic Resonance Imaging Markers for Early Diagnosis of Parkinson's Disease'. In: *Neural Regeneration Research* 7.8 (15th Mar. 2012), pp. 611–619. DOI: 10.3969/j.issn.1673-5374.2012.08.009. pmid: 25745453.

[62] Francisco Jesús Martinez-Murcia, Andres Ortiz, Juan Manuel Górriz, Javier Ramírez, Fermin Segovia, Diego Salas-Gonzalez, Diego Castillo-Barnes and Ignacio A. Illán. 'A 3D Convolutional Neural Network Approach for the Diagnosis of Parkinson's Disease'. In: *Natural and Artificial Computation for Biomedicine and Neuroscience*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 324–333. ISBN: 978-3-319-59740-9. DOI: 10.1007/978-3-319-59740-9_32.

[63] Frank J. Massey. 'The Kolmogorov-Smirnov Test for Goodness of Fit'. In: *Journal of the American Statistical Association* 46.253 (1951), pp. 68–78. DOI: 10.2307/2280095. JSTOR: 2280095.

[64] Peter McCullagh. 'Regression Models for Ordinal Data'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 42.2 (1980), pp. 109–142. DOI: 10.1111/j.2517-6161.1980.tb01109.x. JSTOR: 2984952.

[65] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Series in Computer Science. New York: McGraw-Hill, 1997. 414 pp. ISBN: 978-0-07-042807-2.

[66] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek and Klaus-Robert Müller. 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition'. In: *Pattern Recognition* 65 (1st May 2017), pp. 211–222. DOI: 10.1016/j.patcog.2016.11.008.

[67] Ritu Nayar and David C. Wilbur. *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Springer, 13th Apr. 2015. 342 pp. ISBN: 978-3-319-11074-5. Google Books: JPJICAAAQBAJ.

[68] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao and Gang Hua. 'Ordinal Regression with Multiple Output CNN for Age Estimation'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2016, pp. 4920–4928. DOI: 10.1109/CVPR.2016.532.

[69] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz and E. Nöth. 'Automatic Detection of Parkinson's Disease in Running Speech Spoken in Three Different Languages'. In: *The Journal of the Acoustical Society of America* 139.1 (Jan. 2016), pp. 481–500. DOI: 10.1121/1.4939739. pmid: 26827042.

[70] Luis Perez and Jason Wang. *The Effectiveness of Data Augmentation in Image Classification Using Deep Learning*. 13th Dec. 2017. arXiv: 1712.04621.

[71] María Pérez-Ortiz, Pedro Antonio Gutiérrez, César Hervás-Martínez and Xin Yao. 'Graph-Based Approaches for Over-Sampling in the Context of Ordinal Regression'. In: *IEEE Transactions on Knowledge and Data Engineering* 27.5 (May 2015), pp. 1233–1245. DOI: 10.1109/TKDE.2014.2365780.

[72] Joaquim F. Pinto da Costa, Hugo Alonso and Jaime S. Cardoso. 'The Unimodal Model for the Classification of Ordinal Data'. In: *Neural Networks* 21.1 (1st Jan. 2008), pp. 78–91. DOI: 10.1016/j.neunet.2007.10.003.

[73] Joaquim F. Pinto da Costa and Jaime S. Cardoso. 'Classification of Ordinal Data Using Neural Networks'. In: *Machine Learning: ECML 2005*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, pp. 690–697. ISBN: 978-3-540-31692-3. DOI: 10.1007/11564096_70.

[74] Joaquim F. Pinto da Costa, Ricardo Sousa and Jaime S. Cardoso. 'An All-at-once Unimodal SVM Approach for Ordinal Classification'. In: *2010 Ninth International Conference on Machine Learning and Applications*. 2010 Ninth International Conference on Machine Learning and Applications. Dec. 2010, pp. 59–64. DOI: 10.1109/ICMLA.2010.16.

[75] Foster Provost and Tom Fawcett. 'Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions'. In: *Proceedings of the 3rd International Conference on Knowledgediscovery and Data Mining*. 1997, pp. 43–48.

[76] Alberto Rey, Bernardino Arcay and Alfonso Castro. 'A Hybrid CAD System for Lung Nodule Detection Using CT Studies Based in Soft Computing'. In: *Expert Systems with Applications* (20th Nov. 2020), p. 114259. DOI: 10.1016/j.eswa.2020.114259.

[77] Isabel Rio-Torto, Kelwin Fernandes and Luís F. Teixeira. 'Understanding the Decisions of CNNs: An in-Model Approach'. In: *Pattern Recognition Letters* 133 (1st May 2020), pp. 373–380. DOI: 10.1016/j.patrec.2020.04.004.

[78] William A. Rivera and Petros Xanthopoulos. 'A Priori Synthetic Over-Sampling Methods for Increasing Classification Sensitivity in Imbalanced Data Sets'. In: *Expert Systems with Applications* 66 (30th Dec. 2016), pp. 124–135. DOI: 10.1016/j.eswa.2016.09.010.

[79] Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana and Giancarlo Logroscino. 'Accuracy of Clinical Diagnosis of Parkinson Disease: A Systematic Review and Meta-Analysis'. In: *Neurology* 86.6 (9th Feb. 2016), pp. 566–576. DOI: 10.1212/WNL.0000000000002350. pmid: 26764028.

[80] Olga Russakovsky et al. 'ImageNet Large Scale Visual Recognition Challenge'. In: *International Journal of Computer Vision* 115.3 (1st Dec. 2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[81] Addisson Salazar, Luis Vergara and Gonzalo Safont. 'Generative Adversarial Networks and Markov Random Fields for Oversampling Very Small Training Sets'. In: *Expert Systems with Applications* 163 (1st Jan. 2021), p. 113819. DOI: 10.1016/j.eswa.2020.113819.

[82] A. L. Samuel. 'Some Studies in Machine Learning Using the Game of Checkers'. In: *IBM Journal of Research and Development* 3.3 (1st July 1959), pp. 210–229. DOI: 10.1147/rd.33.0210.

[83] Javier Sánchez-Monedero, Pedro A. Gutiérrez, F. Fernández-Navarro and C. Hervás-Martínez. 'Weighting Efficient Accuracy and Minimum Sensitivity for Evolving Multi-Class Classifiers'. In: *Neural Processing Letters* 34.2 (28th May 2011), p. 101. DOI: 10.1007/s11063-011-9186-9.

[84] Karl Schulz, Leon Sixt, Federico Tombari and Tim Landgraf. *Restricting the Flow: Information Bottlenecks for Attribution*. 25th May 2020. arXiv: 2001.00396 [cs, stat].

[85] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra. 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization'. In: *International Journal of Computer Vision* 128.2 (Feb. 2020), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. arXiv: 1610.02391 [cs].

[86] Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 19th Apr. 2014. arXiv: 1312.6034 [cs].

[87] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 10th Apr. 2015. arXiv: 1409.1556 [cs].

[88] Stephen M. Smith et al. 'Advances in Functional and Structural MR Image Analysis and Implementation as FSL'. In: *NeuroImage* 23.1 (2004), S208–219. DOI: 10.1016/j.neuroimage.2004.07.051. pmid: 15501092.

[89] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox and Martin Riedmiller. *Striving for Simplicity: The All Convolutional Net*. 13th Apr. 2015. arXiv: 1412.6806 [cs].

[90] M. Stone. 'The Opinion Pool'. In: *The Annals of Mathematical Statistics* 32.4 (Dec. 1961), pp. 1339–1342. DOI: 10.1214/aoms/1177704873.

[91] Taiji Suzuki. 'Fast Generalization Error Bound of Deep Learning from a Kernel Perspective'. In: *International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. 31st Mar. 2018, pp. 1397–1406.

[92] Erico Tjoa and Cuntai Guan. 'A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI'. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (Nov. 2021), pp. 4793–4813. DOI: 10.1109/TNNLS.2020.3027314.

[93] Claire L. Tomlinson, Rebecca Stowe, Smitaa Patel, Caroline Rick, Richard Gray and Carl E. Clarke. 'Systematic Review of Levodopa Dose Equivalency Reporting in Parkinson's Disease'. In: *Movement Disorders* 25.15 (2010), pp. 2649–2653. DOI: 10.1002/mds.23429.

[94] John W. Tukey. 'Comparing Individual Means in the Analysis of Variance'. In: *Biometrics* 5.2 (1949), pp. 99–114. DOI: 10.2307/3001913. JSTOR: 3001913.

[95] J. René Van Dorp and Thomas A. Mazzuchi. 'Solving for the Parameters of a Beta a Distribution under Two Quantile Constraints'. In: *Journal of Statistical Computation and Simulation* 67.2 (1st Sept. 2000), pp. 189–201. DOI: 10.1080/00949650008812041.

[96] Eva M. van Rikxoort, Bartjan de Hoop, Max A. Viergever, Mathias Prokop and Bram van Ginneken. 'Automatic Lung Segmentation from Thoracic Computed Tomography Scans Using a Hybrid Approach with Error Detection'. In: *Medical Physics* 36.7 (9th June 2009), pp. 2934–2947. DOI: 10.1118/1.3147146.

[97] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Javier Barbero-Gómez and César Hervás-Martínez. 'Activation Functions for Convolutional Neural Networks: Proposals and Experimental Study'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021), pp. 1–11. DOI: 10.1109/TNNLS.2021.3105444.

[98] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Javier Barbero-Gómez and César Hervás-Martínez. 'Soft Labelling Based on Triangular Distributions for Ordinal Classification'. In: *Information Fusion* 93 (1st May 2023), pp. 258–267. DOI: 10.1016/j.inffus.2023.01.003.

[99] Víctor Manuel Vargas, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Cumulative Link Models for Deep Ordinal Classification'. In: *Neurocomputing* 401 (11th Aug. 2020), pp. 48–58. DOI: 10.1016/j.neucom.2020.03.034.

[100] Víctor Manuel Vargas, Pedro Antonio Gutiérrez and César Hervás-Martínez. 'Unimodal Regularisation Based on Beta Distribution for Deep Ordinal Regression'. In: *Pattern Recognition* 122 (1st Feb. 2022), p. 108310. DOI: 10.1016/j.patcog.2021.108310.

[101] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Riccardo Rosati, Luca Romeo, Emanuele Frontoni and César Hervás-Martínez. 'Deep Learning Based Hierarchical Classifier for Weapon Stock Aesthetic Quality Control Assessment'. In: *Computers in Industry* 144 (1st Jan. 2023), p. 103786. DOI: 10.1016/j.compind.2022.103786.

[102] Andreas Veit, Michael Wilber and Serge Belongie. *Residual Networks Behave Like Ensembles of Relatively Shallow Networks*. 26th Oct. 2016. arXiv: 1605.06431 [cs].

[103] Marcos Vinícius dos Santos Ferreira, Antonio Oseas de Carvalho Filho, Alcilene Dalília de Sousa, Aristófanes Corrêa Silva and Marcelo Gattass. 'Convolutional Neural Network and Texture Descriptor-Based Automatic Detection and Diagnosis of Glaucoma'. In: *Expert Systems with Applications* 110 (15th Nov. 2018), pp. 250–263. DOI: 10.1016/j.eswa.2018.06.010.

[104] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel and Xia Hu. 'Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks'. 13th Apr. 2020. arXiv: 1910.01279 [cs].

[105] Paul J. Werbos. 'Applications of Advances in Nonlinear Sensitivity Analysis'. In: *System Modeling and Optimization*. Lecture Notes in Control and Information Sciences. Berlin, Heidelberg: Springer, 1982, pp. 762–770. ISBN: 978-3-540-39459-4. DOI: 10.1007/BFb0006203.

[106] Frank Wilcoxon. 'Individual Comparisons by Ranking Methods'. In: *Breakthroughs in Statistics: Methodology and Distribution*. Springer Series in Statistics. New York, NY: Springer, 1992, pp. 196–202. ISBN: 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_16.

[107] Richard Williams. 'Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables'. In: *Stata Journal* 6 (1st Feb. 2006), pp. 58–82. DOI: 10.1177/1536867X0600600104.

[108] Hong Wu, Hanqing Lu and Songde Ma. 'A Practical SVM-based Algorithm for Ordinal Regression in Image Retrieval'. In: *11th ACM International Conference on Multimedia*. MULTIMEDIA '03. Berkeley: Association for Computing Machinery, 2nd Nov. 2003, pp. 612–621. ISBN: 978-1-58113-722-4. DOI: 10.1145/957013.957144.

[109] Matthew D Zeiler and Rob Fergus. 'Visualizing and Understanding Convolutional Networks'. In: *Lecture Notes in Computer Science* (*Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*). Vol. 8689 LNCS. 12th Nov. 2014, pp. 818–833. ISBN: 978-3-319-10589-5. DOI: 10.1007/978-3-319-10590-1_53. arXiv: 1311.2901.

[110] Qinghe Zheng, Xinyu Tian, Mingqiang Yang, Yulin Wu and Huake Su. 'PAC-Bayesian Framework Based Drop-Path Method for 2D Discriminative Convolutional Network Pruning'. In: *Multidimensional Systems and Signal Processing* 31.3 (1st July 2020), pp. 793–827. DOI: 10.1007/s11045-019-00686-z.

[111] Qinghe Zheng, Mingqiang Yang, Jiajie Yang, Qingrui Zhang and Xinxin Zhang. 'Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process'. In: *IEEE Access* 6 (2018), pp. 15844–15869. DOI: 10.1109/ACCESS.2018.2810849.

[112] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba. *Learning Deep Features for Discriminative Localization*. 13th Dec. 2015. arXiv: 1512.04150 [cs].