UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Towards better selection and characterisation criteria for high-redshift radio galaxies using machine-assisted pattern recognition

"Documento Provisório"

Doutoramento em Física e Astrofísica

Rodrigo Alonso Carvajal Pizarro

Tese orientada por:

José Afonso

Israel Matute

Hugo G. Messias

Documento especialmente elaborado para a obtenção do grau de doutor

2024

This page intentionally left blank.

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Towards better selection and characterisation criteria for high-redshift radio galaxies using machine-assisted pattern recognition

Doutoramento em Física e Astrofísica

Rodrigo Alonso Carvajal Pizarro

Tese orientada por:

José Afonso

Israel Matute

Hugo G. Messias

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship PD/BD/150455/2019 (PhD:SPACE Doctoral Network PD/00040/2012).

Documento especialmente elaborado para a obtenção do grau de doutor

2024

This page intentionally left blank.

Acknowledgements

This work was supported by the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through research grants PTDC/FIS-AST/29245/2017, EXPL/FIS-AST/1085/2021 (DOI: 10.54499/EXPL/FIS-AST/1085/2021), UID/FIS/04434/2019, UIDB/04434/2020 (DOI: 10.54499/UIDB/04434/2020), and UIDP/04434/2020 (DOI: 10.54499/UIDP/04434/2020). The author also acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship PD/BD/150455/2019 (PhD:SPACE Doctoral Network PD/00040/2012) and POCH/FSE (EC).

This page intentionally left blank.

Resumo

A compreensão da formação e evolução das galáxias ao longo da história do Universo exige uma caracterização precisa das suas fontes de radiação. Por exemplo, diferentes componentes numa galáxia contribuem para o seu orçamento energético total. Para além disso, à medida que requamos aos primórdios do Universo, o conhecimento dos processos radiativos torna-se crucial para compreender como o Universo evoluiu de um ambiente neutro para um ambiente ionizado durante a época da reionização (EoR). Embora a formação estelar (SF) seja considerada o principal factor desta mudança, o impacto dos núcleos galácticos ativos (AGNs) continua a ser desconhecido. Recentemente, grandes esforços têm sido feitos a nível teórico e observacional para determinar a produção de radiação dos AGNs ao longo da história cósmica utilizando uma abordagem em múltiplos comprimentos de onda.

Os avanços atuais no estudo das fontes extragalácticas incluem a utilização de classificação espectroscópica, ferramentas de ajuste de distribuição de energia espectral (SED) a filtros de banda larga ou uma caracterização bi-dimensional mais simples de cores, entre outros métodos. Embora estas técnicas possam ser muito bem-sucedidas na separação da emissão de AGNs e SF, a sua aplicação em conjuntos de dados muito grandes tornou-se altamente dispendiosa. Grandes configurações computacionais são necessárias para o estudo das observações mais recentes de milhões de fontes. Deste modo, a comunidade desenvolveu métodos alternativos para este propósito que podem fornecer resultados de alta qualidade numa fração do tempo significativamente reduzida comparado com técnicas anteriores e com custos energéticos e ambientais reduzidos. Específicamente, os métodos de aprendizagem automática (ML) podem identificar ligações entre as propriedades de um número reduzido de amostras e generalizar esse conhecimento para o conjunto total de dados.

Esta tese apresenta uma nova ferramenta de ML que aproveita os enormes volumes de dados atualmente disponíveis para a seleção eficiente de AGNs selecionados no infravermelho (IR) e emissores de radiação rádio. Além disso, a nossa ferramenta é capaz de estimar o desvio para o vermelho (uma medição de distância) nestas fontes. A nossa abordagem consiste num processo de três etapas independentes, cada uma incluindo cinco algoritmos de ML: extreme gradient boosting (`XGBoost`), Category Boosting (`CatBoost`), random forest (`RF`), gradient boosting (GB), extra trees (`ET`). Para melhorar os resultados destes modelos, estes foram combinados

por meio de empilhamento generalizado, que seleciona um modelo principal (chamado meta-aprendiz) enquanto utiliza os resultados dos restantes modelos como novas características de treino. A primeira etapa do nosso processo de previsão classifica as fontes detectadas no IR entre AGNs e SFGs, a segunda etapa é capaz de determinar se um AGN pode ser detectado em frequências de rádio, e a etapa final foi concebida para estimar valores de desvio para o vermelho em AGNs previstos como detectáveis a frequências de rádio.

Os modelos foram treinados com dados fotométricos em comprimentos de onda ópticos, próximos ao infravermelho (NIR) e infravermelho médio (MIR) no campo Hobby-Eberly Telescope Dark Energy Experiment (HETDEX). Em particular, a fotometria dos catálogos CatWISE2020, Pan-STARRS, 2MASS, AllWISE e as cores formadas com suas magnitudes foram incluídas no treino. Testámos o nosso processo de previsão em dados do próprio campo HETDEX e em fontes do campo Stripe 82, conduzindo a resultados comparáveis às actuais técnicas tradicionais e baseadas em ML para a seleção de AGNs e a estimativa de valores de desvio para o vermelho. Para a classificação entre AGNs e SFGs, o nosso primeiro modelo é capaz de recuperar mais de 94 % dos AGNs conhecidos em ambos os campos. O segundo modelo é capaz de recuperar quase 60 % dos AGNs previamente detectados por rádio e o terceiro modelo pode estimar os desvios para o vermelho dos AGNs detectáveis por rádio com mais de 75 % da amostra conhecida a corresponder muito bem aos valores verdadeiros.

Através da aplicação do nosso processo de previsão em mais de 18 milhões de fontes detectadas por IR, produzimos 68 252 candidatos a AGN em rádio no campo HETDEX e 22 445 candidatos no campo Stripe 82. Em ambos os campos, os novos candidatos têm desvios para o vermelho fotométricos previstos até $z = 4{,}4$. A maior parte destes candidatos têm cores IR de AGNs de acordo com a literatura, e aqueles que não os possuem exigirão um exame mais aprofundado para compreender as suas propriedades.

A fim de extrair uma visão física da utilização da nossa ferramenta, aplicámos várias técnicas para a determinação dos observáveis que contêm a maior quantidade de informação para a estimativa das propriedades das galáxias na amostra. Em particular, as Explicações Aditivas de Shapley (SHAP) podem fornecer uma lista ordenada das propriedades que têm o maior impacto sobre as previsões das fontes. Em geral, as cores fotométricas do CatWISE2020 e Pan-STARRS são assinaladas como as medições mais relevantes, confirmando as expectativas anteriores. Para a primeira etapa do nosso processo (AGNs vs SFGs), a análise SHAP revela que a característica mais relevante é a cor (*W1* - *W2*) do CatWISE2020, enquanto a segunda

etapa (emissão no rádio) tem como principal factor a cor (*g* - *i*) do Pan-STARRS, e o modelo do desvio para o vermelho mostra a cor (*g* - *r*), do Pan-STARRS, como a sua propriedade mais relevante.

Com estes resultados, somos capazes de produzir um novo critério óptico-NIR de seleção de AGNs de alto desvio para o vermelho ($z > 4{,}0$) que pode ser usado como um atalho para a aplicação das nossas previsões. O nosso critério, que combina as cores (*W1* - *W2*) e (*g* - *r*), mostra resultados que estão em linha com os critérios tradicionais de cores IR e podem ultrapassar os seus resultados dependendo do critério de pontuação utilizado.

Depois do estudo inicial de teste da ferramenta nos campos HETDEX e Stripe 82, aplicámos a mesma ferramenta para a selecionar candidatos a AGNs e SFGs emissores de rádio na área da pesquisa piloto no Mapa Evolutivo do Universo (EMU-PS). Em primeiro lugar, derivámos funções de luminosidade no rádio (RLFs) dos AGNs e SFGs previstos e contextualizámo-las com trabalhos anteriores, obtendo resultados compatíveis com o conhecimento atual das fontes detectadas em rádio. Enquanto os trabalhos anteriores de RLFs calibram os seus resultados usando correções das próprias observações em rádio, nós conseguimos fazê-lo apenas usando a avaliação das previsões sem recurso a medições em rádio.

Para além da compatibilidade das nossas RLFs baseadas em ML com trabalhos anteriores, a nossa análise beneficia do grande número de candidatos produzidos. As incertezas das RLFs que obtivemos são extremamente reduzidas, colocando limites rigorosos na demografia da população emissora de rádio no Universo. Além disso, o aumento do número de fontes permite-nos vislumbrar outras características nas RLFs de alto desvio para o vermelho que podem indicar a existência de populações distintas de fontes brilhantes que ainda não foram totalmente avaliadas até à data.

Em segundo lugar, a partir da análise das probabilidades que os nossos modelos atribuem a cada fonte de ser um AGN detetável em rádio, mostramos que é possível usar as nossas previsões como um método alternativo ao convencional de identificação das galáxias emissoras de rádio noutros comprimentos de onda (isto é, identificação de contrapartes). De todas as fontes IR que rodeiam uma deteção de rádio, aquelas com as maiores probabilidades previstas serão as contrapartes mais prováveis para a fonte de rádio. Embora sejam necessárias análises mais profundas, a utilização do nosso processo de previsão tem o potencial de ser comparável a técnicas mais avançadas, uma vez que analisa a distribuição fotométrica completa das fontes.

Esta tese também discute possíveis melhorias no nosso processo de previsão e os seus

resultados. Dado que se baseia em técnicas de ML, a estratégia mais direta é o aumento do volume do conjunto de dados de treino. A inclusão de mais fontes e com medições em mais bandas de comprimento de onda e de diferentes estudos e instrumentos provou, no passado, melhorar a qualidade de qualquer modelo de ML. Por exemplo, a inclusão de medições de raios X (por exemplo, do telescópio XMM-Newton) ou ultravioleta (UV, por exemplo do telescópio GALEX) poderia aumentar as hipóteses de os nossos modelos reconhecerem a emissão de AGNs. Para além das propostas anteriores, a inclusão explícita de erros e incertezas de medição poderia ajudar os modelos a alargar a cobertura do espaço de parâmetros. Outras medidas adicionais estão relacionadas com os objetivos específicos do processo de previsão. Por exemplo, limitar o foco da ferramenta estritamente para fontes de alto desvio para o vermelho teria de incluir etapas adicionais para a separação de candidatos próximos e distantes. O estudo dos buracos negros supermassivos (SMBHs) poderá ter de incluir, no seu conjunto de treino, as propriedades de fontes com propriedades medidas.

Para além das alterações na configuração do processo de previsão e dos modelos que ele inclui, examinámos possíveis aplicações em dados de levantamentos futuros e em curso. Tendo em conta os exemplos apresentados no campo EMU-PS, uma aplicação natural do processo de previsão é nas fontes do levantamento EMU. Milhões de detecções de rádio estarão disponíveis para a seleção de AGNs nesta área e para sua análise. Além disso, nos próximos anos, o Square Kilometre Array (SKA) observará o céu a frequências de rádio com capacidades extraordinárias que podem fornecer centenas de milhões de detecções.

A necessidade de ferramentas rápidas e fiáveis para a avaliação das fontes é óbvia e o nosso processo de previsão apresenta-se como uma opção perfeita para estudá-las. A nossa ferramenta visa esclarecer o caminho para a compreensão da conexão entre as fontes de radiação nas galáxias e como elas, através das suas interações, moldaram o Universo como podemos vê-lo atualmente.

Palavras-chave: Núcleo Galáctico Ativo; Radio Galáxias; Classificação de Galáxias; Determinação do Desvio para o Vermelho; Aprendizagem Automática.

Abstract

Understanding how galaxies and their constituents, like active galactic nuclei (AGN), evolve and interact across cosmic timescales remains a key challenge in astrophysics, especially during the epoch of reionisation (EoR), when the early Universe transitioned from a neutral to an ionised state. Even though star formation (SF) is considered to be its main contributor, the impact of AGN remains elusive. Recently, huge efforts have been put into the determination of the AGN bolometric radiation output via their search using a multi-wavelength approach.

To address such challenge, this thesis presents a novel machine learning (ML) tool –a pipeline of models– for the efficient selection and redshift characterisation of radio-detectable AGN by using multi-wavelength photometry of sources detected in the infrared (IR). By analysing sources in a wide range of redshift values, this tool enables the exploration of the AGN-galaxy co-evolution across cosmic times. Applied to millions of sources in the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Spring and the Stripe 82 (S82) fields, our pipeline has identified almost 100 thousand radio-AGN candidates, with predicted redshift values up to 4.4.

Beyond classification, we can extract the key parameters that most significantly impact candidate selection in our pipeline. This investigation has led to the design of a new AGN colour-colour selection criterion, offering a useful, ML-based, tool for the community.

Furthermore, by extracting radio-AGN candidates from the Evolutionary Map of the Universe Pilot Survey (EMU-PS), we can generate radio luminosity functions (RLFs) with highly constrained uncertainties. Our results are compatible with current knowledge and hint at the existence of a distinct population of bright sources.

Finally, this thesis explores the potential application of our tool to future surveys like the Square Kilometre Array (SKA), that is expected to generate immense radio datasets. Our rapid and reliable method will be instrumental in separating AGN from star-forming galaxies (SFGs) while helping to unveil their interplay across the history of the Universe.

Keywords: Active Galactic Nuclei; Radio Galaxies; Galaxy classification; Redshift Determination; Machine Learning.

This page intentionally left blank.

Part of the work presented in this thesis is based on the datasets, models, results and analyses published in the following articles and repositories.

- R. Carvajal, I. Matute, J. Afonso, S. Amarantidis, D. Barbosa, P. Cunha, and A. Humphrey (Oct. 2021). 'Exploring New Redshift Indicators for Radio-Powerful AGN'. in: *Galaxies* 9.4, p. 86. arXiv: 2111.00778 [astro-ph.GA]. DOI: 10.3390/galaxies9040086
- R. Carvajal, I. Matute, J. Afonso, R. P. Norris, K. J. Luken, P. Sánchez-Sáez, P. A. C. Cunha, A. Humphrey, H. Messias, S. Amarantidis, D. Barbosa, H. A. Cruz, H. Miranda, A. Paulino-Afonso, and C. Pappalardo (Nov. 2023a). 'Selection of powerful radio galaxies with machine learning'. In: A&A 679, A101, A101. arXiv: 2309.11652 [astro-ph.GA]. DOI: 10.1051/0004-6361/202245770
- R. Carvajal, I. Matute, J. Afonso, R. P. Norris, K. J. Luken, P. Sánchez-Sáez, P. A. C. Cunha, A. Humphrey, H. Messias, S. Amarantidis, D. Barbosa, H. A. Cruz, H. Miranda, A. Paulino-Afonso, and C. Pappalardo (Dec. 2023b). *Selection of powerful radio galaxies with machine learning*. DOI: 10.5281/zenodo.10220009

This page intentionally left blank.

Contents

ACKNOWLEDGEMENTS iii

RESUMO v

ABSTRACT ix

LIST OF TABLES xvii

LIST OF FIGURES xix

LIST OF ACRONYMS xxiii

LIST OF SYMBOLS xxix

1 INTRODUCTION 1

1.1 AGN AND THEIR IMPACT ON THE EVOLUTION OF THE UNIVERSE . . . 1

1.1.1 AGN DETECTION METHODS 6

1.1.2 REDSHIFT DETERMINATION 15

1.2 CHALLENGES IN THE ANALYSIS OF ASTRONOMICAL DATA 18

1.2.1 COMPUTATIONAL COSTS 19

1.2.2 MISSING MEASUREMENTS 20

1.2.3 DATA HETEROGENEITY 21

1.2.4 COUNTERPART IDENTIFICATION 21

1.3 MACHINE-ASSISTED PATTERN DETECTION 23

1.3.1 TYPES OF MACHINE-ASSISTED ANALYSES 24

1.3.2 ENSEMBLE LEARNING 25

1.3.3 MODEL EXPLAINABILITY AND FEATURE IMPORTANCE 26

GLOBAL FEATURE IMPORTANCES 27

LOCAL FEATURE IMPORTANCES 28

1.4 THIS THESIS 28

2 DATASETS FOR TRAINING AND TESTING 33
2.1 HETDEX SPRING FIELD . . . 34
2.2 STRIPE 82 FIELD . . . 36
2.3 PHOTOMETRY MEASUREMENTS . . . 37
2.4 MISSING DATA TREATMENT . . . 40
2.5 ADDITIONAL FEATURES . . . 44
2.5.1 ENGINEERED FEATURES . . . 44
2.5.2 GROUND TRUTH FEATURES . . . 46
2.6 DATA RE-SCALING AND NORMALISATION . . . 46
2.7 DATA SPLITTING . . . 48
2.8 COMPUTATIONAL SET-UP . . . 51

3 MODELS' TRAINING AND PREDICTION OF RADIO-AGN CANDIDATES 53
3.1 PREDICTION METRICS . . . 53
3.1.1 CLASSIFICATION METRICS . . . 55
3.1.2 REGRESSION METRICS . . . 57
3.2 CLASSIFICATION THRESHOLDS . . . 60
3.3 CLASSIFICATION CALIBRATION . . . 60
3.3.1 CALIBRATION SCORES . . . 61
3.4 FEATURE SELECTION . . . 62
3.5 MODEL STACKING . . . 64
3.6 MODEL TRAINING . . . 65
3.6.1 HYPERPARAMETERS OPTIMISATION . . . 66
3.6.2 CALIBRATION OF MODELS . . . 67
3.6.3 THRESHOLD SELECTION . . . 70
3.6.4 USE OF COMPUTATIONAL RESOURCES . . . 71
3.7 PREDICTION OF RADIO-AGN CANDIDATES . . . 71
3.7.1 AGN-SFG CLASSIFICATION . . . 71
3.7.2 RADIO DETECTION CLASSIFICATION . . . 73
3.7.3 REDSHIFT PREDICTION . . . 75
3.7.4 PREDICTION FROM PIPELINE . . . 76
3.7.5 NO-SKILL CLASSIFICATION . . . 81

4 ANALYSIS OF PREDICTION METHOD AND RESULTS 83
4.1 COMPARISON WITH PREVIOUS WORKS 83
4.1.1 AGN DETECTION PREDICTION 83
4.1.2 RADIO DETECTION PREDICTION 90
4.1.3 REDSHIFT PREDICTION 91
4.2 INFLUENCE OF DATA IMPUTATION 94
4.3 GLOBAL FEATURE IMPORTANCES 96
4.4 LOCAL FEATURE IMPORTANCES 99

5 MACHINE-ASSISTED LEARNING: UNLOCKING HIDDEN KNOWLEDGE 111
5.1 COLOUR-COLOUR AGN SELECTION CRITERION 111
5.2 RADIO LUMINOSITY FUNCTION 116
5.3 RADIO COUNTERPART ASSESSMENT 145

6 FUTURE DEVELOPMENTS 151
6.1 SUBSET ANALYSIS FOR RADIO-AGN IDENTIFICATION 151
6.2 IMPROVEMENT OF PREDICTION PIPELINE 152
6.3 PIPELINE APPLICATION IN ADDITIONAL DATASETS 155
6.3.1 EVOLUTIONARY MAP OF THE UNIVERSE 155
6.3.2 SQUARE KILOMETRE ARRAY 156

SUMMARY 159

DATA AND SOFTWARE ACKNOWLEDGEMENTS 167

REFERENCES 171

APPENDICES 209

A LUMINOSITY FUNCTION FORMULATION 211

B SAMPLE OF PREDICTED RADIO-DETECTABLE AGN 217

C EXTENDED PREDICTION PIPELINE 225
C.1 TRAINING AND MODEL SELECTION 227
C.2 APPLICATION OF STACKED MODELS 230

This page intentionally left blank.

List of tables

2.1 Available bands in training set . . . 39
2.2 Density of detected sources in HETDEX and S82 fields . . . 40
2.3 CW catalogue cross matches in HETDEX and S82 fields . . . 45
2.4 Feature names for models . . . 47

3.1 Model selection for AGN-SFG classification . . . 65
3.2 Model selection for radio detection classification . . . 66
3.3 Model selection for redshift value prediction . . . 66
3.4 Hyperparameters for modified pipeline . . . 67
3.5 Metrics from AGN-SFG classification model . . . 72
3.6 Metrics from radio detection model . . . 74
3.7 Metrics from redshift prediction model . . . 75
3.8 Metrics from radio AGN pipeline . . . 78
3.9 Results of no-skill source selection . . . 82

4.1 Comparison AGN colour-colour selection criteria and pipeline . . . 85
4.2 Comparison previous AGN selection criteria and pipeline . . . 90
4.3 Comparison previous redshift estimation methods and pipeline . . . 93
4.4 Feature importances from individual models . . . 97
4.5 Feature importances from base models . . . 98
4.6 SHAP values from base models . . . 102
4.7 Mean absolute SHAP values for high-z sources . . . 105

5.1 Metrics from colour-colour AGN detection criteria . . . 115
5.2 Catalogue cross matches in EMU pilot survey . . . 120
5.3 Predicted sources per redshift bin . . . 132
5.4 Imputed predicted sources per redshift bin . . . 141

B.1 Table columns description . . . 218
B.2 Predicted properties test set . . . 219

B.3 Predicted properties S82 220
B.4 Predicted properties unlabelled sources in HETDEX 221
B.5 Predicted properties unlabelled sources in S82 222

C.1 Catalogue cross matches updated pipeline 227
C.2 Individual modified models for radio detection classification for AGN 228
C.3 Individual modified models for radio detection classification for SFGs 228
C.4 Individual modified models for redshift on radio AGN 229
C.5 Individual modified models for redshift on radio SFGs 229
C.6 Hyperparameters for modified pipeline 230
C.7 Metrics from SFG-AGN classification model 231
C.8 Metrics from radio detection prediction models in AGN and SFGs 231
C.9 Metrics from joint classification models in AGN and SFGs 232
C.10 Metrics from redshift prediction model 232
C.11 Results of modified no-skill source selection in HETDEX 233
C.12 Results of modified no-skill source selection in S82 233

List of figures

1.1 AGN unification scheme diagram . . . 3
1.2 Radio-far-infrared M82 spectrum . . . 4
1.3 Example BPT-VO diagrams . . . 8
1.4 Example WHAN diagram . . . 9
1.5 Example Spitzer colour-colour diagram . . . 10
1.6 Example WISE colour-colour diagram . . . 11
1.7 Million Quasar Catalog v7.4d source density . . . 15
1.8 Generalised stacking flowchart . . . 26
1.9 Flowchart prediction pipeline . . . 30

2.1 HETDEX area footprint . . . 35
2.2 S82 area footprint . . . 37
2.3 Histogram of non-imputed magnitudes in HETDEX . . . 42
2.4 Histogram of imputed magnitudes in HETDEX . . . 43
2.5 Magnitude depths . . . 44
2.6 Data pre-process flowchart . . . 49
2.7 HETDEX data flowchart . . . 50
2.8 S82 data flowchart . . . 50

3.1 Flowchart expanded prediction pipeline . . . 54
3.2 Example of confusion matrix . . . 55
3.3 Reliability curves for uncalibrated classifiers . . . 69
3.4 Reliability curves for calibrated classifiers . . . 69
3.5 Precision-Recall curves for calibrated models . . . 70
3.6 Application of AGN-SFG model to test subset . . . 73
3.7 Application of radio detection model to test subset . . . 74
3.8 Application of redshift model to test subset . . . 76
3.9 Confusion matrix radio-AGN prediction on test subset . . . 78
3.10 Confusion matrix radio-AGN prediction on labelled S82 sources . . . 79

3.11 Application of redshift model to predicted radio-detectable AGN in test subset . 80
3.12 Application of redshift model to predicted radio-AGN from labelled S82 sources 80
3.13 Predicted redshift distribution 81

4.1 (*W1* - *W2*) vs (*W2* - *W3*) colour-colour diagrams in HETDEX 86
4.2 (*r* - *i*) vs (*g* - *r*) colour-colour diagrams in HETDEX 88
4.3 Evolution of predicted values with number of observed bands 95
4.4 Decision plot for AGN-SFG classification 101
4.5 Decision plot for radio detection model 103
4.6 Decision plot for redshift prediction model 104
4.7 SHAP decision plots for base AGN-SFG algorithms 107
4.8 SHAP decision plots from base radio algorithms 108
4.9 SHAP decision plots from base *z* algorithms 109

5.1 (*W1* - *W2*) vs (*g* - *r*)colour-colour AGN diagram 114
5.2 Flowchart extended prediction pipeline 117
5.3 EMU-PS area footprint 118
5.4 Magnitude depths in EMU-PS 120
5.5 Predicted redshifts in EMU-PS 121
5.6 Flux distribution predicted sources in EMU-PS 122
5.7 Predicted 1.4 GHz luminosity vs redshift in EMU-PS 123
5.8 Predicted 1.4 GHz luminosity vs z in EMU-PS with fixed classes 125
5.9 Predicted redshift values in EMU-PS with fixed classes 126
5.10 Recall distribution for predicted sources in HETDEX 127
5.11 Precision distribution for predicted sources in HETDEX 128
5.12 Recall distribution for predicted sources in EMU-PS 129
5.13 Precision distribution for predicted sources in EMU-PS 130
5.14 Radio luminosity function in EMU-PS per redshift bin 133
5.15 Predicted probabilities as function of redshift in EMU-PS 136
5.16 Gridded RLF in EMU-PS 138
5.17 Predicted $L_{1.4\,\mathrm{GHz}}$ vs z in EMU-PS with fixed classes and imputed fluxes 140
5.18 Radio luminosity function in EMU-PS with all predicted sources 142
5.19 Gridded RLF in EMU-PS with imputed fluxes 143

5.20 Gridded RLF in EMU-PS with imputed fluxes (detail) 144
5.21 CW-EMU counterpart example 146
5.22 CW-EMU counterparts examples (continued) 148

B.1 CW-HETDEX prediction examples 223
B.2 CW-S82 prediction examples 224

C.1 Flowchart extended prediction pipeline 226

This page intentionally left blank.

List of acronyms

2M	Two Micron All Sky Survey
AGN	Active galactic nuclei
ALMA	Atacama large millimeter/submillimeter array
ASKAP	Australian Square Kilometre Array Pathfinder
AW	AllWISE
B18	Blecha et al. (2018)
BLR	Broad line region
BLRG	Broad line radio galaxy
BPT	Baldwin-Phillips-Terlevich
BPZ	Bayesian Photometric Redshifts
BS	Brier score
BSS	Brier skill score
C23	Carvajal et al. (2023a)
`CatBoost`	Category Boosting
COSMOS	Cosmic Evolution Survey
CPU	Central processing unit
CV	Cross-validation
CW	CatWISE2020
DES	Dark Energy Survey
DES-DR2	DES data release 2
DESI	Dark Energy Spectroscopic Instrument
DEVILS	D10 field of the Deep Extragalactic VIsible Legacy Survey
DL	Deep learning
DT	Decision tree
EAZY	Easy and Accurate $z_{\rm phot}$ from Yale
eCDFS	extended Chandra Deep Field South
eFEDS	EROSITA Final Equatorial Depth Survey
ELAIS-S1	European Large Area ISO Survey-South 1

EMU	Evolutionary Map of the Universe
EMU-PS	EMU pilot survey
EoR	Epoch of reionisation
eROSITA	Extended ROentgen Survey with an Imaging Telescope Array
ET	Extra trees
FIR	Far infrared
FIRST	Faint Images of the Radio Sky at Twenty-Centimeters
FN	False negative
FP	False positive
FRI	Fanaroff-Riley class I
FRII	Fanaroff-Riley class II
FSRQ	Flat spectrum radio quasar
FWHM	Full width at half maximum
GALEX	Galaxy Evolution Explorer
GAMA	Galaxy and Mass Assembly
GB	Gradient boosting
GBC	Gradient boosting classifier
GBR	Gradient boosting regressor
GP	Gaussian process
HERG	High excitation radio galaxy
HET	Hobby-Eberly Telescope
HETDEX	Hobby-Eberly Telescope Dark Energy Experiment
IC	Inverse Compton
IFU	Integral field unit
IGM	Inter-galactic medium
IR	Infrared
IRAC	Infrared Array Camera
ISO	Infrared Space Observatory
KDE	Kernel density estimation
KNN	K-nearest neighbours
ΛCDM	Λ cold dark matter
LePHARE	Photometric Analysis for Redshift Estimate

LERG	Low excitation radio galaxy
LF	Luminosity function
`LightGBM`	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LINER	Low-ionization nuclear emission-line region
LOFAR	Low Frequency Array
LoTSS	LOFAR Two-metre Sky Survey
LoTSS-DR1	LoTSS - data release 1
LoTSS-DR2	LoTSS - data release 2
`LR`	Linear regression
LSST	Legacy Survey of Space and Time
M12	Mateos et al. (2012)
M16	Mingo et al. (2016)
MAD	Median absolute deviation
MAE	Mean absolute error
MaNGA	Mapping Nearby Galaxies at the Apache Point Observatory
MCC	Matthews correlation coefficient
MCMC	Markov Chain Monte Carlo
MeerKAT	Meer-Karoo Array Telescope
MIGHTEE	MeerKAT International GHz Tiered Extragalactic Exploration
MIPS	Multiband Imaging Photometer
MIR	Mid infrared
ML	Machine learning
MLR	Maximum likelihood ratio
MQC	Million Quasar Catalog
MSE	Mean squared error
NED	NASA/IPAC Extragalactic Database
NELG	Narrow emission line galaxy
NEOWISE	Near-Earth Object *WISE*
NEP	North Ecliptic Pole
`NGBoost`	Natural gradient boosting
ngVLA	next-generation VLA

NIR	Near infrared
NLRG	Narrow line radio galaxy
NMAD	Normalised median absolute deviation
NRAO	National Radio Astronomy Observatory
NVSS	NRAO VLA Sky Survey
OVV	Optically violent variables
Pan-STARRS	Panoramic Survey Telescope and Rapid Response System
PDF	Probability density function
PR	Precision-recall
PS1	Pan-STARRS data release 1
PSF	Point-spread function
PyBDSF	Python Blob Detector and Source Finder
QSO	Quasi stellar object
Quaia G20.5	*Gaia*–unWISE Spectroscopic Quasar catalog
RACS	Rapid ASKAP Continuum Survey
RF	Random forest
RG	Radio galaxy
RL	Radio-loud
RLF	Radio luminosity function
RMSE	Root mean square error
RQ	Radio-quiet
RSD	Relative standard deviation
S12	Stern et al. (2012)
S82	Stripe 82
SDC	SKA data challenge
SDSS	Sloan Digital Sky Survey
SDSS-DR15	SDSS data release 15
SDSS-DR16	SDSS data release 16
SDSS-DR17	SDSS data release 17
SED	Spectral energy distribution
SF	Star formation
SFG	Star-forming galaxy

SFR	Star formation rate
SFRD	Star formation rate density
SHAP	SHapley Additive exPlanations
SKA	Square Kilometre Array
SMBH	Super-massive black hole
SSRQ	Steep spectrum radio quasar
`SVM`	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate
UV	Ultra violet
VEXAS	VISTA EXtension to Auxiliary Surveys
VEXAS-DR2	VEXAS data release 2
VISTA	Visible and Infrared Survey Telescope for Astronomy
VLA	Karl G. Jansky Very Large Array
VLAS82	VLA SDSS Stripe 82 Survey
VLASS	VLA Sky Survey
VLBI	Very long baseline interferometry
VO	Veilleux-Osterbrock
WISE	Wide-field Infrared Survey Explorer
`XGBoost`	Extreme gradient boosting

This page intentionally left blank.

List of symbols

$''$	Arcsecond (also arcsec)
$\mathbb{C}$	Source class
r	Distance to source
η	Outlier fraction
F_β	F-score
β	F-score parameter
F1	F-1 score
F	Flux
S_ν	Flux density
ν	Frequency
H_0	Hubble constant
$\mathcal{K}$	K-correction factor
L	Luminosity
$\hat{f}_{wa}$	Luminosity density function
D_L	Luminosity distance
ϕ	Luminosity function
$\hat{\phi}$	Luminosity function estimation
σ_{MAD}	MAD
σ_{NMAD}	NMAD
Ω_Λ	Dark matter density
Ω_m	Cosmological mass density
ρ	Pearson's correlation factor
L_ν	Power density
P	Probability
$P\,(\mathbb{C})$	Probability of belonging to class $\mathbb{C}$
$p(z)$	Probability density
Δz	Redshift difference
Δz^{N}	Normalised redshift difference

Δz_{Total}	Total redshift difference
$\Delta z^{\text{N}}_{\text{Total}}$	Total normalised redshift difference
$\mathbb{d}_{\text{eff}}$	Sample effective size
$\mathbb{d}$	Sample size
$\mathcal{P}$	Luminosity selection function
σ_z	Standard deviation
σ^{N}_z	Normalised standard deviation
α	Radio spectral index
V	Volume
z	Redshift
z_{phot}	Photometric redshift
$z_{\text{Predicted}}$	Predicted redshift
z_{spec}	Spectroscopic redshift
z_{True}	True redshift

1

Introduction

1.1 AGN and their impact on the evolution of the Universe

The epoch of reionisation (EoR), a critical period in cosmic history, attested the evolution of a neutral to an ionised Universe and the formation of the first large structures. Galaxies, through their emission and the emergence of super-massive black holes (SMBHs) in their central regions, are thought to have played a leading role in this transformation (e.g. Tripodi et al. 2022; Robertson 2022) which has had consequences until present times. For this reason, having a clear understanding of their birth, development, and connection with their environment throughout the history of the Universe becomes a prime goal in astrophysics.

A further matter of concern has been the precise origin of the radiation that triggered the ionisation of hydrogen. Among several processes (see, for instance, Katz et al. 2018, 2019), the two main options have been the star formation (SF) events (Fukugita and Kawasaki 1994; Haiman and Loeb 1997; Madau et al. 1999; Ciardi and Ferrara 2005; Sipple and Lidz 2024) or the emission from the active galactic nuclei (AGN) (Meiksin 2005; Faucher-Giguère et al. 2009; Haardt and Madau 2012; Madau and Haardt 2015). From most recent observations, models, and simulations, a growing consensus has made the emission from SF the main source of ionising radiation (Loeb and Furlanetto 2013; Mitra et al. 2018; Matsuoka et al. 2018; Kulkarni et al. 2019; Shen et al. 2020; Robertson 2022; Dayal et al. 2024) with all the remaining sources of ionising radiation playing a minor role.

Nonetheless, AGN and their emission have been subject of extensive study as a way to understand the processes taking place in the centre of galaxies and in which ways they could be connected to their host galaxies at all epochs (e.g. King and Pounds 2015; Hickox and Alexander 2018; Blandford et al. 2019). As such, AGN are instrumental in determining the nature, growth, and evolution of SMBHs as well as probing their surroundings (Padovani et al. 2017). Their

strong emission allows us, also, to study the vicinity of the galaxies by which they are hosted, namely, the inter-galactic medium (IGM) (e.g. Nicastro et al. 2017, 2018; Kovács et al. 2019; Fan et al. 2023). Additionally, the study of AGN can help understanding the overall evolution of large structures in the early Universe given their ubiquity and large energetic output (Krumpe et al. 2014; Porqueres et al. 2018; Magliocchetti 2022). By analysing their behaviour, AGN can be used as a tool to understand the overall evolution of the constituents of the Universe.

In order to use AGN for the analysis of their hosts and environments, detailed knowledge must exist on their structure and energetics. The most accepted model for the emission of AGN consists on the unified scheme (Antonucci 1993; Urry and Padovani 1995; Bicknell et al. 1997; Urry 2004), where differences in the populations of AGN are due to the observation angle and the presence of material in the surroundings of the central black hole. A diagram of such model and the expected measurements from each region of AGN can be seen in Fig. 1.1, where each quadrant of the figure presents a different observing angle with all available structures, measurements, and corresponding labelling from that perspective.

The inner regions of the galactic centre can host different structures, such as an accretion disk, broad-line regions, a central obscuring torus, a narrow-line region, a thin molecular disk, and central radio jets (Netzer 2015). Through different processes, these structures can radiate in different wavelengths that can be observed and analysed. Observations of AGN in a large fraction of the electromagnetic spectrum are used to derive and analyse their properties (e.g. Padovani et al. 2017). Emission in specific wavelengths can give information of phsyical processes fueling their radiation (Nour and Sriram 2023). X-ray emission is thought to be arise from the hot corona as inverse Compton (IC) radiation of optical and ultra violet (UV) photons (from the accretion disk) that can be directly radiated or reflected in the torus as well as from powerful jets in the form of continuum emission (Haardt and Maraschi 1991; Haardt and Maraschi 1993; Brandt and Alexander 2015). UV radiation is also thought to be originated in the accretion disk of AGN (Bahcall and Kozlovsky 1969; Shakura and Sunyaev 1973; Davidson and Netzer 1979), which also photo-ionises material in the broad line region (BLR). High-energy photons can also be reprocessed into infrared (IR) thermal emission by dust and give rise to some of the features seen in the IR wavelenghts of AGN spectral energy distributions (SEDs). Dust is located at all scales in the host galaxy with the closest concentration to the AGN found in the torus-like structure (Hickox and Alexander 2018; Lyu and Rieke 2022; U 2022).

AGN are not the only contributors to the UV, optical, and IR photons that are emitted

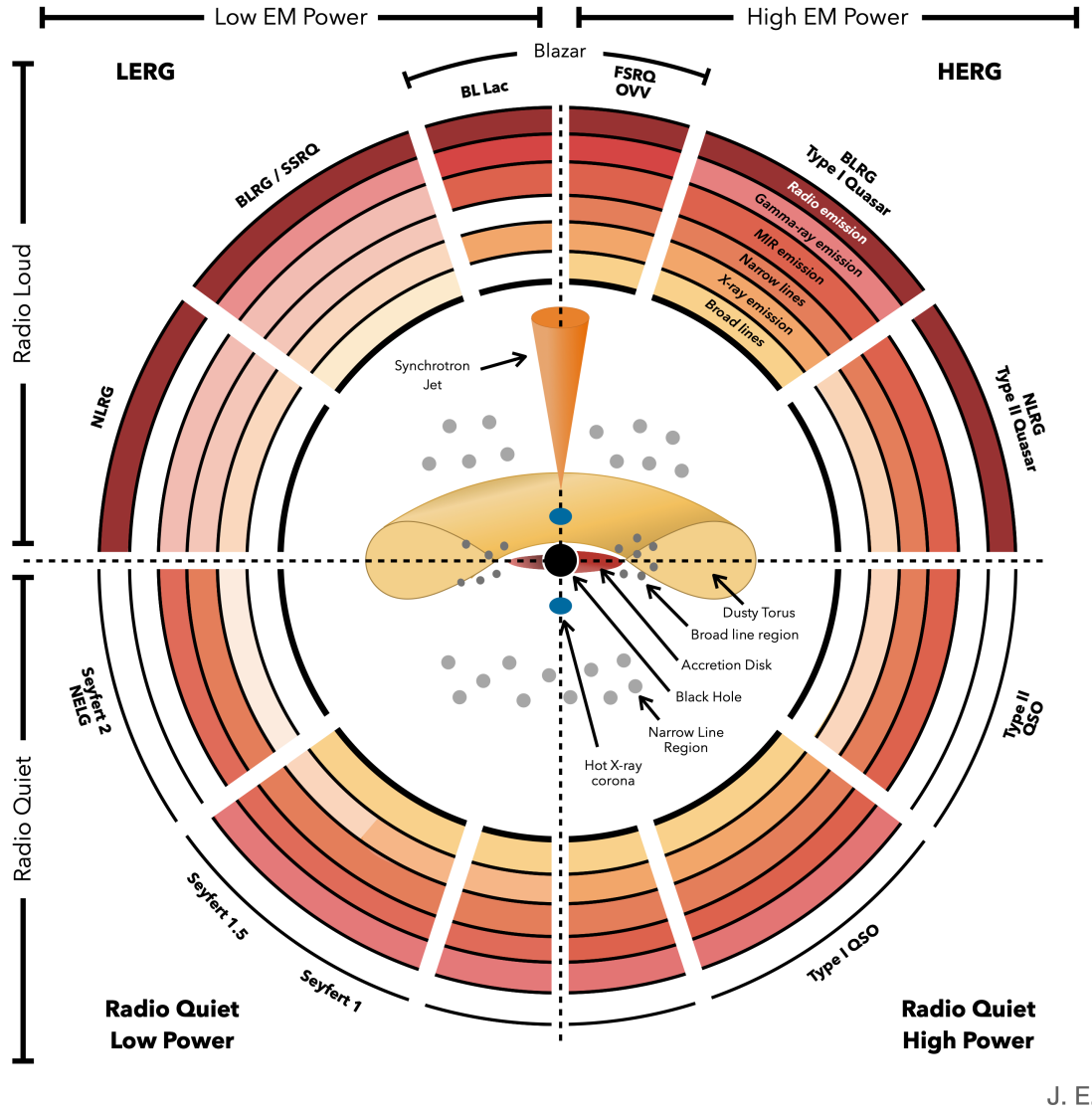


Figure 1.1: Schematic representation of the orientation-driven unified model of AGN. The type of object observed depends on the viewing angle, whether or not the AGN produces a significant jet (radio loud or radio quiet), and the rate of accretion onto the central SMBH (low or high electromagnetic power). The centre of the schematic shows their typical components. The upper left and upper right quadrants are commonly referred to as low excitation radio galaxys (LERGs) and high excitation radio galaxys (HERGs) respectively. Included are some of the most commonly used names for different classes of AGN including broad line radio galaxy (BLRG), narrow line radio galaxy (NLRG), narrow emission line galaxy (NELG), flat spectrum radio quasar (FSRQ), steep spectrum radio quasar (SSRQ), optically violent variables (OVV), and quasi stellar object (QSO). Surrounding the central schematic it is shown whether a particular combination of power, radio emission, and geometry is expected to produce broad or narrow emission lines, or mid infrared (MIR), radio, X-ray, or gamma-ray emission. The transparency of the colour in each ring corresponds to the increasing strength or prevalence of a particular emission type. Image and description credits: Thorne et al. (2022a), published under a CC BY 4.0 license.

from a galaxy. Light from stars (e.g. Dai et al. 2018; Bowler et al. 2021) and dust obscuration (Yan et al. 2023) can significantly dilute AGN signatures. The sensitivity and sky coverage on current X-ray facilities pose difficulties to the access to this privileged wavelength regime for AGN selection (see, for example, Mazzolari et al. 2024). Thus, obtaining direct estimates of AGN intrinsic properties turns out to be a challenging task.

On the other side of the electromagnetic spectrum, emission in the radio frequencies can trace either highly star-forming regions of their host galaxy (Magliocchetti 2022) or very powerful jets produced by the central engine (radio galaxy (RG); Heckman and Best 2014). Contrary to other wavelengths, radio observations present very low optical depth values (Hildebrand 1983), allowing the observation of objects that can be highly obscured in IR, optical, UV, or X-ray wavelengths (e.g. Chen et al. 2020; Pérez-Torres et al. 2021), making radio observations better suited for the search of such sources. An example of the emission of a bright galaxy in the radio-to-IR region of the spectrum is presented in Fig. 1.2. From the figure, it is possible to see that both radio and IR regions of the SED can be identified from the slopes of their emission.

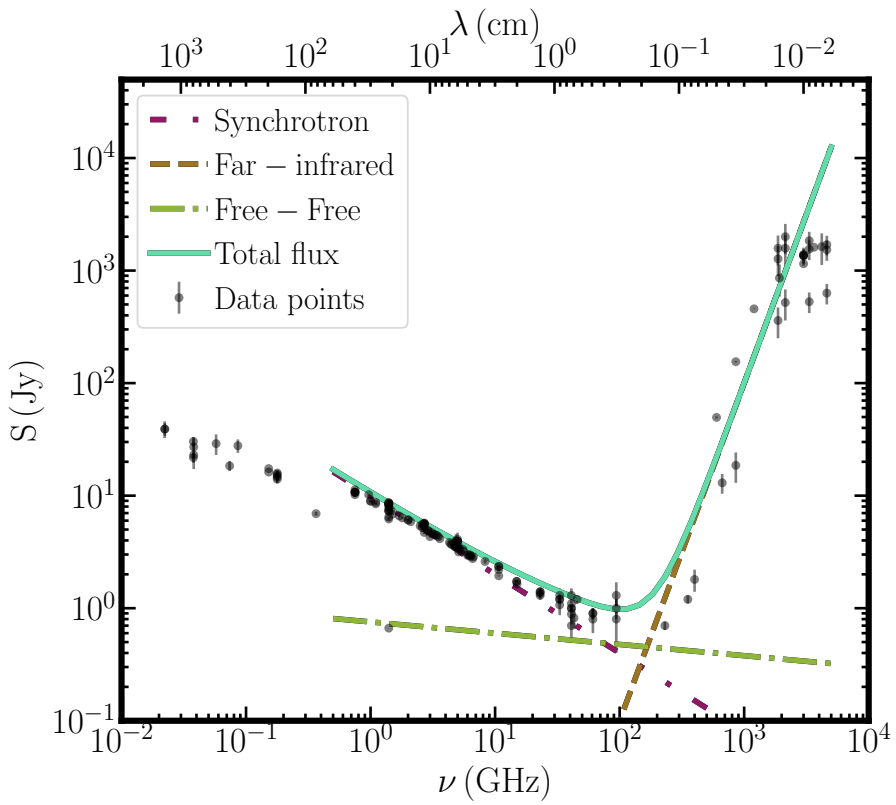


Figure 1.2: Archetypical radio-to-far infrared (FIR) spectrum of starburst galaxy M82 in the style of Fig. 1 from Condon (1992). Observed data points (in black, with error bars) have been retrieved from the NASA/IPAC Extragalactic Database (NED) website. Purple, double-dot-dashed line shows a synchrotron component. Green, dot-dashed line represents a thermal (free-free) emission profile, while brown, dashed line shows FIR emission. Aquamarine, continuous line represents the addition of the three mentioned components.

Besides very bright AGN, a fraction of star-forming galaxies (SFGs) have been discovered using radio bands (e.g. McGreer et al. 2006; Kuźmicz and Jamrozy 2021; Delhaize et al. 2021; Lal 2021). Nevertheless, most radio observations of AGN and SFGs have been the result of follow-up projects for already-known objects (Radcliffe et al. 2021b). Thus, most sources detected in radio frequencies need multi-wavelength observations for their confirmation and characterisation. With the advent of more powerful instruments and surveys, objects with dimmer radio emission could be detected. Some examples include the Square Kilometre Array (SKA; Braun et al. 2019), next-generation VLA (ngVLA; Selina et al. 2018; Selina et al. 2023), and the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE; Jarvis et al. 2016) survey that uses the Meer-Karoo Array Telescope (MeerKAT; Jonas and MeerKAT Team 2016; Camilo et al. 2018; Mauch et al. 2020). But as sensitivity levels are improved, emission from star formation can also be detected, making more difficult the distinction between emission from the AGN and their hosts (Rawlings 2003), adding more challenges to the identification of RGs.

Recently-developed, wider, and deeper radio instruments and surveys, such as the Faint Images of the Radio Sky at Twenty-Centimeters (FIRST; Helfand et al. 2015), the EMU pilot survey (EMU-PS; Norris et al. 2021), the VLA Sky Survey (VLASS; Lacy et al. 2020; Gordon et al. 2020), the LOFAR Two-metre Sky Survey (LoTSS; Shimwell et al. 2017), and very long baseline interferometry (VLBI; for instance, Falcke et al. 2000; Kim et al. 2020), have allowed detection of larger numbers of RGs (e.g. Singh et al. 2014; Williams et al. 2018; Capetti et al. 2020). But determination of some of their properties –e.g. redshift, spectral indices– might still take very long observation times with high sensitivity detectors in, occasionally, other wavelengths (An et al. 2020). These difficulties make, effectively, characterisation of RGs a costly endeavour.

All these surveys and instruments can deliver measurements of sources at different distances and angular resolutions. In order to analyse such observations, intrinsic properties are needed (i.e. not affected by observational constraints). For instance, instead of measured fluxes (F) and redshifts (z), objects can be compared through luminosities (L). Obtaining luminosities can take into account the path that light has taken from the source to the observer (Rybicki and Lightman 2008). From a statistical point of view, one of the most commonly used tools for the description of sources through their luminosities and all the assumptions and biases associated to them is that of luminosity functions (LFs). LFs can provide a robust measure of

the evolution of the density of sources in different time (z) and brightness (L) bins (e.g. Salpeter 1955; Schmidt 1968; Schechter 1976; Steidel et al. 1999). Additionally, the analysis of LFs can help constraining the onset of ionising photons available for the ionisation of atoms in the IGM.

The production of LFs with the use of only the observed sources in a given z and L bin can reproduce biases presented in the collected samples from their design and observational constraints. To quantify, and attempt to correct such biases, a selection function, $\mathcal{P}$, is used to correct the sources count in the calculation of the LF. The selection function, then, summarises the corrections that the distribution of sources must suffer in order to be as close as possible to our best guess of their real distribution. More detailed formulation and description of luminosities and LFs are presented in Appendix A.

To test the number of sources we observe in different wavelengths, redshifts, and luminosities, cosmological simulations based on our more recent model of galaxy evolution have been used to obtain an estimate of the number of AGN available to be observed with specific instruments and sensitivities (Habouzit et al. 2022). Some of these simulations have shown (e.g. Amarantidis et al. 2019; Bonaldi et al. 2019; Thomas et al. 2021) that the distribution of AGN and RG along redshift will lead, potentially, to the detection of a few hundreds of objects per square degree closer to the end of the EoR with deep radio observations –e.g. SKA, which is projected to have μJy point-source sensitivity levels (Prandoni and Seymour 2015)–.

These expectations of an statistically significant number of AGN and RG in the high-redshift Universe do not match completely with the most recent compilations (e.g. Inayoshi et al. 2020; Ross and Cross 2020; Fan et al. 2023), which show that more than 300 have been confirmed to exist at redshifts higher than $z = 6$ in the whole sky. This mismatch emphasises the need to detect and confirm the presence of more AGN than can match models and simulations.

1.1.1 AGN detection methods

The presence of an AGN can be confirmed (or hinted) in several ways depending on the observed, and desired, wavelengths. After their discovery and confirmation in optical and radio wavelengths (Seyfert 1943; Schmidt 1963; Matthews and Sandage 1963), one of the wavelengths used to confirm the nature of AGN, and the dust enshrouding them, was IR (for a historical review, see Sajina et al. 2022). Assuming that the activity in SMBHs and some components of their host galaxies are correlated (see, for instance, Magorrian et al. 1998; Ferrarese and Merritt 2000; Gebhardt et al. 2000; Häring and Rix 2004; Gültekin et al. 2009; Beifiori et al. 2012;

McConnell and Ma 2013; Kormendy and Ho 2013; Heckman and Best 2014; and references therein), and the current unified model for AGN (Urry and Padovani 1995; Bicknell et al. 1997; Urry 2004; Netzer 2015), most of the activity from the accretion in AGN will be obscured by a dusty torus surrounding the SMBH (e.g. Lacy and Sajina 2020) which will re-emit this energy into IR wavelengths. The peak of this activity will be correlated with that of the SF in the host galaxy, thus, increasing the fraction of obscured light observed in such systems (Madau and Dickinson 2014). In this way, the highest probability of detecting an AGN will be by observing in IR wavelengths.

As mentioned previously, X-ray is considered as an efficient way to confirm the presence of an AGN (e.g. Donley et al. 2005; Radcliffe et al. 2021a; Andonie et al. 2022). Based upon either their physical extension or the intensity of their emission, X-ray sources can be identified as AGN without large uncertainties (LSST Science Collaboration et al. 2009; Padovani et al. 2017; Maitra et al. 2019; Osorio-Clavijo et al. 2023). If an X-ray point source has a luminosity higher than $L_X \sim 10^{42}\,\mathrm{erg\,s^{-1}}$, it is highly likely to be an AGN (Stern 2015; Auge et al. 2023). Thus several sources have been detected in this way (e.g. Chen et al. 2017; Martocchia et al. 2017; Ricci et al. 2017; Goulding et al. 2018; Maitra et al. 2019; Lansbury et al. 2020; Coleman et al. 2022; Wasleske and Baldassare 2023).

Many traditional AGN detection methods make use of spectral or photometric observations of objects which, based upon several criteria, determine their nature or class (Padovani et al. 2017; Hickox and Alexander 2018; Pouliasis 2020; Chaves-Montero et al. 2017). In the case of spectroscopy, Optical and IR observations have been used to look for the presence of emission lines that might indicate activity from AGN in their spectra (Magliocchetti 2022). This method provides the best way to determine the presence of an AGN. One method derived from spectroscopic observations is the use of the Baldwin-Phillips-Terlevich (BPT; Baldwin, Phillips, and Terlevich 1981) diagram, which, with the modifications made by Veilleux and Osterbrock (1987: also called VO diagrams), has been used extensively to detect and diagnose AGN and the SMBH they host based on detected emission lines (e.g. Toba et al. 2014; Sartori et al. 2015; Latimer et al. 2021; Birchall et al. 2020; Ceccarelli et al. 2022). The BPT-VO diagram uses ratios of the intensity of optical emission lines [O III] $\lambda5007/\mathrm{H}\beta$, [N II] $\lambda6584/\mathrm{H}\alpha$, [S II] $\lambda\lambda6717,6731/\mathrm{H}\alpha$, and [O I] $\lambda6300/\mathrm{H}\alpha$ to determine the source of ionisation of the studied sources and separate them between SFGs and AGN. Further studies have used the BPT-VO diagrams but different thresholds to separate SFGs and AGN (e.g. Kewley et al. 2001; Kauffmann

et al. 2003; Kewley et al. 2006; Schawinski et al. 2007). Figure 1.3 shows an example of the application of the BPT-VO diagrams, with different boundaries, to separate between SFGs, AGN, low-ionization nuclear emission-line regions (LINERs; Heckman 1980), and composite galaxies (an intermediate state between the previous stages).

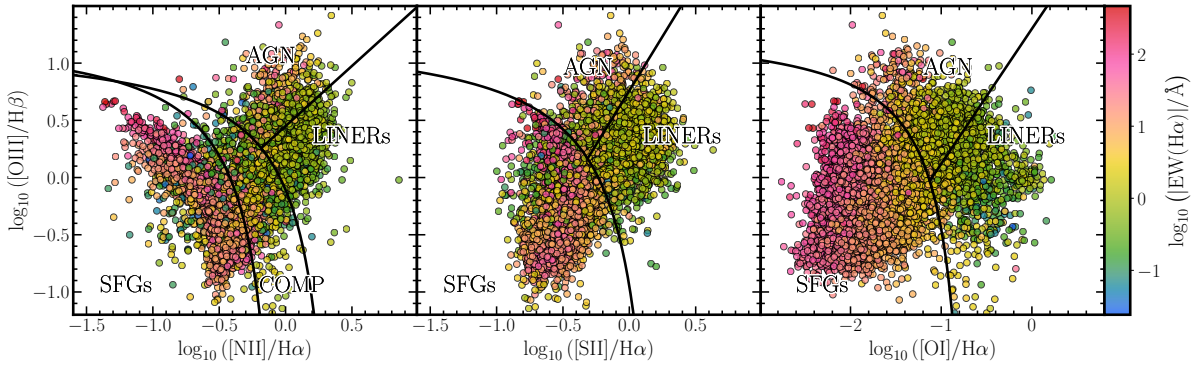


Figure 1.3: Example BPT-VO diagrams using data from sources in the Mapping Nearby Galaxies at the Apache Point Observatory (MaNGA) Pipe3D value added catalog data release 17 (Sánchez et al. 2018; Lacerda et al. 2022). Classification boundaries (as thick, black lines) from Kewley et al. (2001), Kauffmann et al. (2003), Kewley et al. (2006), and Schawinski et al. (2007). Points are coloured by the logarithm of the absolute value of the Hα equivalent width following the coding of the colorbar and error bars have been omitted for clarity. The plot in the leftmost panel can help separating SFGs and composite galaxies (COMP) from AGN/LINERs while the middle and rightmost panels can help classifying SFGs, AGN, and LINERs.

Additional diagrams have also been developed with the aim of using different combinations of emission lines. One remarkable example is the WHAN diagram (Cid Fernandes et al. 2010, 2011), which uses the information of only two emission lines, the equivalent width of Hα and the [N II] λ6584/Hα line ratio for AGN selection. An example of its application to a sample of sources is shown in Fig. 1.4, where a classification between SFGs, LINERs, and Seyfert galaxies (Osterbrock 1981) is possible.

In the case of photometry measurements, some of these methods involve the classification of sources using colours (i.e. differences in magnitudes) in different wavebands as a starting point. Usually, one method used to confirm the presence of AGN in a sample is using MIR or near infrared (NIR) colours. The most highly used data comes from photometric observations carried our with *Spitzer* (Werner et al. 2004) or the Wide-field Infrared Survey Explorer (*WISE*; Wright et al. 2010) given their space-borne configurations, which helps avoiding attenuation from our atmosphere. Both telescopes carry instruments able to observe in similar wavelengths. For instance, *Spitzer*'s Infrared Array Camera (IRAC) bands 1 and 2 and one of the bands of Multiband Imaging Photometer (MIPS) can gather observations at 3.6 µm, 4.5 µm, and 24 µm, while bands *W1*, *W2*, and *W4* of *WISE* are centred at 3.4 µm, 4.6 µm, and 22 µm (Antoniucci et al. 2014). Also, their magnitude limits are similar between *Spitzer* (16.9 mag, 15.9 mag, and 9.8 mag for

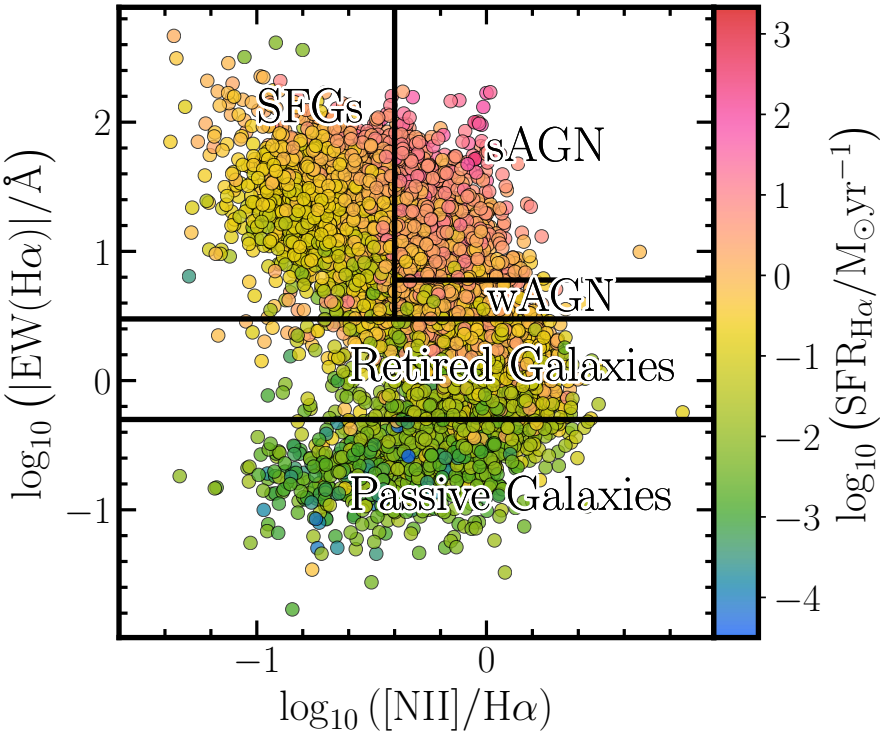


Figure 1.4: Example WHAN diagram using data from sources in the MaNGA Pipe3D value added catalog data release 17 (Sánchez et al. 2018; Lacerda et al. 2022). Classification boundaries in the Hα equivalent width (EW(Hα)) – [N II] λ6584/Hα space (in thick, black lines) obtained from Cid Fernandes et al. (2010, 2011) following Kauffmann et al. (2003) and Kewley et al. (2006) which allow the classification of sources between SFGs in the upper-left region of the plot, Seyfert galaxies (called strong AGN, sAGN) in the upper-right corner, and LINERs in the bottom-right corner, which have been sub-divided between weak AGN (wAGN) and retired galaxies. Passive galaxies have been placed in the bottom side of the plot. Points are coloured by the logarithm of the integrated star formation rate (SFR) derived from the Hα line following the coding of the colorbar and error bars have been omitted for clarity.

IRAC1, *IRAC2*, and MIPS; Antoniucci et al. 2014) and *WISE* (16.9 mag, 14 mag, and 9.4 mag for *W1*, *W2*, and *W4*; Cutri et al. 2012). Their differences start with their angular resolution, with *WISE* having a resolution two times worse than *Spitzer*, for the mentioned bands. Additionally, *Spitzer* is used, mostly, for pointed observations, while *WISE* is a survey instrument. The last difference makes *WISE* better suited for the genereration of large number of sources that might be ingested into the calculation of LFs. Nevertheless, both telescopes have been used extensively for the detection and characterisation of AGN and SFGs.

With observations from *Spitzer*, several schemes have been devised (e.g. Lacy et al. 2004; Donley et al. 2012). Based on the combination of measurements, different scales have been developed (e.g. Stern et al. 2005; Donley et al. 2012), which have been extensively used (e.g. Lacy et al. 2013; İkiz et al. 2020; Bonato et al. 2021; Lacy et al. 2021). An example of the application of such criteria is depicted in Fig. 1.5, in which sources can be selected as AGN from the use of four IRAC channels.

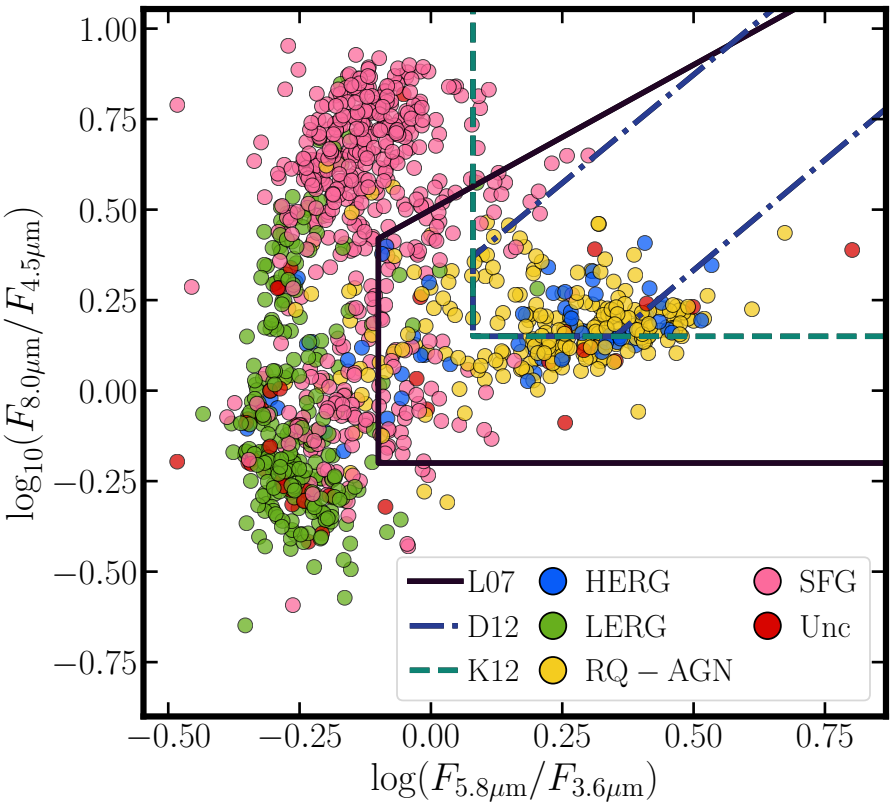


Figure 1.5: Example *Spitzer* colour-colour diagram for the selection of AGN from a sample of radio-detected sources obtained by Duncan et al. (2021) and Morabito et al. (2022). Point colours are related to the classification of the sources as given by Best et al. (2023), which can be HERGs, LERGs, radio-quiet (RQ) AGN, SFGs, or uncertain (Unc). On top of the points the AGN classification criteria from Lacy et al. (2007: L07), Donley et al. (2012: D12), and Kirkpatrick et al. (2012: K12) are plotted. Error bars have been omitted for clarity.

On the other side, combinations of *WISE* colours have been used to derive properties of

AGN and their host galaxies (e.g. Stern et al. 2012; Mateos et al. 2012; Assef et al. 2013; Toba et al. 2014; Menzel et al. 2016; Jarrett et al. 2017; Assef et al. 2018; Barrows et al. 2021). An example of its application is presented in Fig. 1.6, where sources can be separated between AGN, star-forming disks, intermediate disks and spheroidal galaxies. Additional colour criteria have been developed for the latest and future facilities and observations (e.g. Messias et al. 2012; Kirkpatrick et al. 2017; Langeroodi and Hjorth 2023; for JWST).

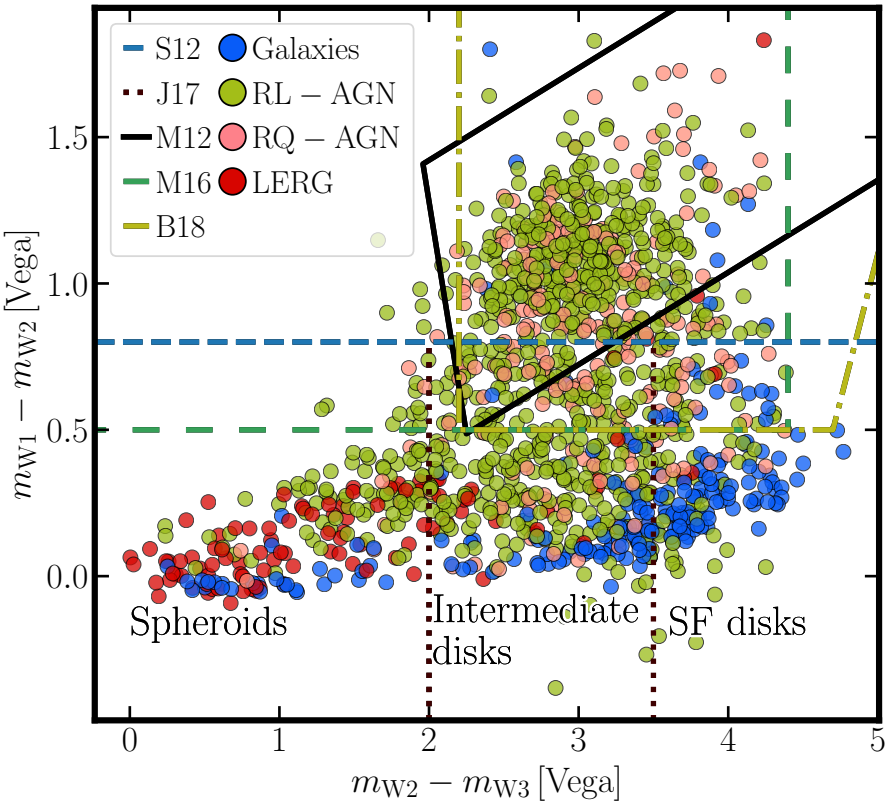


Figure 1.6: Example *WISE* colour-colour diagram for the selection of AGN from a set of sources obtained from the MIXR sample of AGN and SFGs Mingo et al. (2016: hereafter M16). Points show the classification according to Mingo et al. (2016), as galaxies (i.e. SFGs), radio-loud (RL) AGN, RQ AGN, and LERGs. On top of the points, the lines of several classifications from Stern et al. (2012: hereafter S12), Mateos et al. (2012: hereafter M12), Mingo et al. (2016: M16), Jarrett et al. (2017: hereafter J17), and Blecha et al. (2018: hereafter B18) are presented as well as the labelling of the galactic regions of the plot as made by Jarrett et al. (2017). Error bars have been omitted for clarity.

Other techniques to determine the presence of AGN are related to the use of SED fitting, proper motion measurements, variability, and morphology. For SED fitting, the available photometric measurements of an object are compared to a series of model templates (Pacifici et al. 2023). The models have been constructed using different combinations of properties –e.g. age, metallicity, contribution from different constituents, etc.–. Thus, the examined source will be assumed to have the properties from the model which fits the best. If one of the properties included in the selected template is an AGN, then the studied source will be assumed to be an AGN as well.

High quality astrometric measurements (e.g. the *Gaia* mission; Gaia Collaboration et al. 2016) have allowed using proper motions for the detection of AGN. In particular, the use of the extragalactic content (Gaia Collaboration et al. 2023a) of its data release 3 (DR3; Gaia Collaboration et al. 2023b) has helped to determine which sources have very small proper motions, which are indicative of large distances to a source and compatible with the presence of AGN or extragalactic sources in general (e.g. Storey-Fisher et al. 2024; Fu et al. 2024).

Another way of assessing AGN is through the use of photometric measurements in different epochs that allow one to also determine the variability scales of a source. AGN present continuum aperiodic variability in all their observed wavelengths in timescales from hours to years (Giveon et al. 1999). There is, for instance, evidence of correlation between the AGN variability of fluxes in X-ray, UV, optical, and NIR bands (Uttley et al. 2003; Arévalo et al. 2008, 2009; Breedt et al. 2009, 2010; McHardy et al. 2016; Troyer et al. 2016; Buisson et al. 2017; Suganuma et al. 2006; Koshida et al. 2009, 2014; Lira et al. 2011, 2015). This variability also depends on luminosity, wavelength, redshift, presence of radio or X-ray emission, and existence of broad-line systems (Vanden Berk et al. 2004). For these reasons, if particular variability patterns are found in multi-wavelength observations of a source, it can be classified as an AGN candidate.

When high spatial resolution observations are used, morphology can be a suitable tool to determine the presence of either an AGN or a SFG. It has been found that the presence of an AGN, even when not observed directly, can affect the morphological parameters of its host galaxy (Getachew-Woreta et al. 2022) regardless of their morphological classification. This effect is due to the flux from the AGN that impact some areas of the host galaxy and their properties. For the specific case of radio observations, the detection of a compact core or radio jets can strongly suggest the presence of an AGN (e.g. Richards et al. 2007).

As mentioned previously, most AGN have been first identified as such from observations in non-radio wavelengths and then their radio nature has been confirmed by direct observations (e.g. Glikman et al. 2023). Nevertheless, it is still possible to search for AGN, initially, in radio observations. Most radio-bright sources (i.e. with a flux, at 1 GHz higher than 1 mJy) will be RGs with non-thermal emission (Padovani 2016, 2017). With fainter radio sources, it is not possible to establish their nature directly as thermal emission starts to be more ubiquitous. Thus, further analyses are needed to classify radio sources. As with measurements in other wavelengths, it is possible to obtain radio colours (called and defined accordingly, in this context, spectral indices, α, Lisenfeld and Völk 2000), which might help determining whether

the emission from a detected source is produced by an AGN or not. In the context of radio measurements, spectral indices, between two frequencies (ν_a and ν_a), are defined as the value of the slope a power law fitted to the radio flux would have (e.g. Zajaček et al. 2019),

$$\alpha_{\nu_a,\nu_b} = \frac{\log\left(F_{\nu_a}/F_{\nu_b}\right)}{\log\left(\nu_a/\nu_b\right)}\,. \tag{1.1}$$

Bright AGN, for which most of their radio emission is understood to come from synchrotron processes, show spectral indices that are similar between them ($\alpha \sim -0.7$, under the assumption of an emission in the form $F_\nu \propto \nu^\alpha$, following, for instance, Ibar et al. 2010; Delhaize et al. 2017; Gürkan et al. 2019; Deka et al. 2024). In this way, it is possible to correlate the measured radio emission with the presence of an AGN (Condon 1992). This connection might be coupled with studies that show a slight correlation between radio spectral index and radio luminosity for AGN (e.g. Sabater et al. 2019). Besides their spectral indices, special care must be taken with sources that show low levels of radio emission. As explored by, for instance, Magliocchetti (2022), low-luminosity radio sources (as those currently detected by the latest radio observatories and surveys) can have, as the source of their emission, both the AGN they host or star formation events turning disentanglement of both sources of emission a difficult task. One simple approach to separate both populations, and to classify sources with faint radio luminosities, is using a single value for which all sources brighther that that might be labelled as AGN. Above a threshold in luminosity, it might be said that the radio emission detected in a source has been originated from the AGN. Several thresholds have been proposed using different approaches. Most of them have been devised for low-redshift regimes using the distributions of both AGN and SFGs (via LFs). For instance, one value used is $10^{25}\,\mathrm{W\,Hz^{-1}}$, above which sources can be considered, without large uncertainties, as radio-loud AGN (e.g. Williams and Röttgering 2015; Mo et al. 2020). Conversely, when the derived luminosities are close or below the threshold, the fraction of the emission budget that comes from star-formation events increases. Expanding on the idea of setting a threshold, it is possible to obtain a function for this limit that might depend on, for instance, redshift values. Given that the distributions of AGN and SFGs luminosities (i.e. LFs) have different behaviours, it is expected that the curves of both values will cross at some point (Magliocchetti 2022).

The detection, selection, and analysis methods presented in the previous paragraphs can also be applied in the opposite direction for the detection of radio emission in AGN. This process implies searching for radio detections and, afterwards, classifying them as AGN (or any other

kind of source). This procedure is based upon analysing the structure of the studied images and looking for features that might indicate the presence of an AGN (for instance, from their radio jets). Several tools and algorithms have been developed to detect sources. For instance, Python Blob Detector and Source Finder (PyBDSF; Mohan and Rafferty 2015), Blobcat (Hales et al. 2012a,b), and Aegean (Hancock et al. 2012; Hancock et al. 2018). In general, these tools look for islands of emission in images and, depending on the selected detection level, they can merge these structures and create larger objects that can be linked to astrophysical sources (not only AGN). Once these radio detections have been determined, they need to be cross-matched with counterparts in different wavelengths in order to apply further methods to classify their nature and estimate further properties. When the characteristics of the detected radio emission are clearly those of AGN (for instance, from their morphology), multi-wavelength associations are not needed for their confirmation (as done, for instance, with the human-associated tasks of radio galaxy zoo projects; Bowles et al. 2023; Hardcastle et al. 2023).

As already exposed, several techniques exist for the identification and classification of AGN. Given the focus of each of them on different properties of the observed source candidates, several works and techniques might identify the same source, independently, more than once. Thus, additional efforts are needed for the compilation of catalogues of confirmed AGN and the cleaning of duplicate sources across different techniques. Early efforts include the catalogues from de Veny et al. (1971) and Véron-Cetty and Véron (1984, 1985, 1987, 1989, 1991, 1993, 1996, 1998, 2000, 2001, 2003, 2006, 2010), with more than 150 000 sources listed in their latest revisions.

While the shape of the LF for local AGN has been fairly well determined (e.g. Sadler et al. 2002; Mauch and Sadler 2007; Smolčić et al. 2009b; Best and Heckman 2012), more studies are needed for the analysis of the number of AGN and its evolution in the earliest epochs of the Universe. Towards the goal of studying the evolution of the high-redshift Universe, newer compilations of AGN have focused on distant sources. That is the case, for example, of the catalogue by Perger et al. (2017), with sources at $z \geq 4.0$. Similarly, Ross and Cross (2020) compiled almost 500 QSOs at $z \geq 5.0$, Bosman (2022), created a list of $z \geq 5.7$ QSOs, while Inayoshi et al. (2020) and Fan et al. (2023) created lists of more than 200 $z \geq 6.0$ and more than 500 $z \geq 5.3$ AGN, respectively.

The emphasis on high-redshift sources has made that compilations of local-to-moderate redshift AGN have been updated very sparsely, making more difficult the study of large number

of diversely confirmed AGN. One of the few catalogues which keeps including local sources is the Million Quasar Catalog (MQC; Flesch 2015, 2019, 2021, 2023), which attempts to list all confirmed AGN and QSOs. In its last edition (`v8`; Flesch 2023), the MQC includes more than 900 000 sources up to redshifts higher than 7.0. A depiction of the spatial distribution of the sources in the version `7.4d` of the MQC, which is the version used in this work (cf. Sect. 2.5) and that compiles information for more than 1 100 000 AGN, is presented in Fig. 1.7. It can be seen that, from its sparse spatial distribution, a large number of individual catalogues and listings have been assembled, highlighting the breathtaking effort needed for creating such compilation. For instance, the sources from Sloan Digital Sky Survey (SDSS; York et al. 2000) and those in the Cosmic Evolution Survey (COSMOS; Scoville et al. 2007) field, which has been extensively observed in several wavelengths.

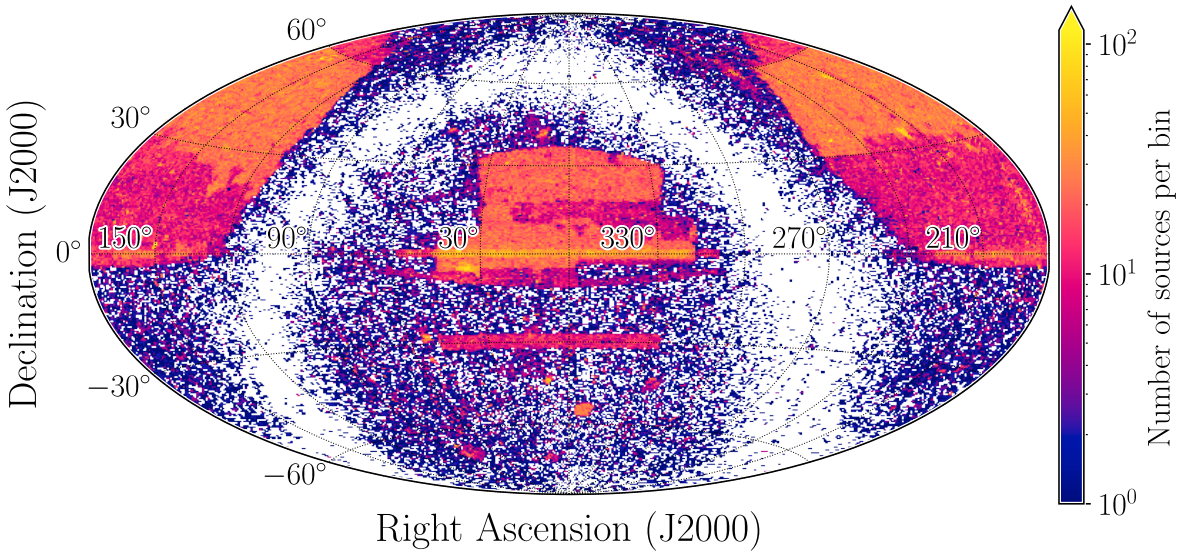


Figure 1.7: Million Quasar Catalog `v7.4d` source density in a Hammer-Aitoff projection (Snyder 1987, 1997) of the sky. Each coloured rectangle shows the number of AGN in that particular region of the sky as catalogued by the MQC following the code in the colorbar. Each overdensity, with respect to its surroundings, represents a different source catalogue or listing from which AGN have been drawn. More noticeable are, for instance, the areas covered by the SDSS observations.

1.1.2 Redshift determination

In order to determine a precise distribution of AGN across cosmic time, unambiguous redshift measurements are needed (e.g. Huterer et al. 2006; Tanaka et al. 2018; van den Busch et al. 2020; Naidoo et al. 2023). Spectroscopic redshifts, being the most precise measurements, can be determined for a large range of objects, from supernovae (e.g. Frederiksen et al. 2014; Baltay et al. 2021), to galaxies (e.g. Galametz et al. 2013; Le Fèvre et al. 2015), and AGN (e.g.

Rajagopal et al. 2021). Spectroscopic redshifts can be obtained by cross-correlation or fitting of the observed data and set of templates (Tonry and Davis 1979; Schuecker 1993; Glazebrook et al. 1998; Aihara et al. 2011; Machado et al. 2013) or by the direct detection and matching of powerful spectral features (Kurtz and Mink 1998; Stoughton et al. 2002; Garilli et al. 2010). However, their determination can take long and high-quality observations, which are not always available for all sources, rendering them not suited for large-sky catalogues (see, for instance, Silva et al. 2011; Pacifici et al. 2023).

Photometric redshifts are an option which comes from the use of photometry measurements and not explicit spectral features of an object (Salvato et al. 2019; Brescia et al. 2021; Newman and Gruen 2022). In general, they use observations that take less integration time than a comparable spectroscopic measurement and, thus, are used for large surveys that need measurements for large numbers of objects (e.g. Hoyle et al. 2018; Tanaka et al. 2018). They are also an option for faint sources.

Photometric redshift methods can deliver their estimates in the form of a probability density function (PDF). These functions can deliver a measure of the uncertainties that photometric redshifts might have. In general terms, photometric redshifts can be obtained using two different methods: template-based techniques and empirical relations.

Template-based methods come from the fitting of multi-wavelength photometry of a source to a model template (Baum 1957, 1962; Loh and Spillar 1986; Bolzonella et al. 2000; Newman and Gruen 2022; Pacifici et al. 2023). The models have been constructed using different combinations of properties –e.g. age, metallicity, contribution from different constituents, etc.–. Thus, the examined source will be assumed to have the properties from the model which fits the best. However, and depending upon the number and quality of the photometry measurements (e.g. photometric band width), these properties can have, sometimes, large uncertainties (Newman et al. 2015; Newman and Gruen 2022). Even though this method can use less precise values to determine a redshift, it can take a significative amount of time since it needs to contrast the measured SED to the full set of model templates (e.g. La Torre et al. 2024) and, when the number of available measurements is low, the quality of the estimation is largely degraded (e.g. Norris et al. 2019).

Using this method, redshift estimations can be obtained from, for instance, galaxies (e.g. Hernán-Caballero et al. 2021), and AGN (e.g. Ananna et al. 2017; Brescia et al. 2019). Some example tools that use template-based methods to retrieve photometric redshifts are Easy and

Accurate $z_{\rm phot}$ from Yale (EAZY; Brammer et al. 2008), Bayesian Photometric Redshifts (BPZ; Benítez 2000), and Photometric Analysis for Redshift Estimate (`LePHARE`; Arnouts et al. 1999; Ilbert et al. 2006).

For the case of empirical relations, the retrieval of photometric redshifts relies on the use of statistics and large sets of observables (e.g. fluxes and their uncertainties) to determine redshifts and correlations between them which can be used with future observations. Most of these empirical redshift determination methods are related to the use of machine learning (ML; Samuel 1959). These techniques will be further described in this work.

Finally, rough estimates of redshift can also be determined. Using differences among magnitudes –i.e. colours– it is possible to establish the redshift range in which a source is located. This technique –called drop-out (Steidel et al. 1996a,b)– is, by no means, precise, but can lead to further investigation of sources that are at relevant redshifts ranges for the researcher (with, for instance, the previously described photometric redshift methods). In this way, drop-outs are employed as a mean to generate candidates for pertinent redshift values. Given that it requires no more calculations than the comparison of some series of colours, it is highly efficient at generating rough redshifts of large samples. It has been, mainly, used to generate and study high-redshift sources or candidates that, otherwise, would not have enough information to produce a precise redshift value (e.g. Bouwens et al. 2020; Carvajal et al. 2020; Merlin et al. 2021; Uzgil et al. 2021; Champagne et al. 2023; Atek et al. 2023).

Since its first uses, this technique has allowed the detection of high-redshift galaxies (Steidel and Hamilton 1992; Steidel et al. 1996a) through the detection of sharp break in flux between broadband filters that sample the vicinities of the Lyman Break (at a rest-frame wavelength of 912 Å). The location of such break is a function of redshift allowing one to obtain a crude estimate of the redshift for the studied objects. Drop-outs have also been used for the selection of high-redshift QSOs, allowing, for instance, the first detection of a $z > 4.0$ QSO (Warren et al. 1987).

All the techniques and methods listed and described in this section highlight the ever-increasing number of different procedures for the detection, selection, and characterisation of extra-galactic sources. From their intrinsically different nature, all of them have their own advantages and pitfalls which need to be accounted for when compiling properties from a large sample of sources.

1.2 Challenges in the analysis of astronomical data

The progress of technology and methods used in astrophysics has been one of the main drivers for the advancement in our knowledge and understanding of the processes taking place in the Universe. But this undeniable improvement has brought some drawbacks that pose serious challenges that might hinder our ability of retrieving useful results from astronomical data. Most of these problems are rooted in the very large number of new and different observational efforts carried out throughout the years. This abundance of measurements can impact the processes that lead to new calculations and results since more resources and steps are needed to treat a large number of measurements consistently.

As more sources are needed to better constrain the properties of astronomical sources, new data sets have been compiled and published. Nowadays, multi-wavelength data are available for large fractions of the sky (e.g. Skrutskie et al. 2006; Wright et al. 2010; Gaia Collaboration et al. 2016; Chambers et al. 2016; Kollmeier et al. 2017; Abbott et al. 2018; Lacy et al. 2020). But this profusion of observations has come with new challenges with the most relevant being the volume of data. Lately, analysing all observations, one by one with traditional methods, has become unfeasible in terms of the time needed to fulfil the task (see, for instance, Brescia et al. 2021). This issue will become greater as future surveys and telescopes are put into service, with relevant examples being the SKA and the Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al. 2009; Ivezić et al. 2019).

Furthermore, over the last couple of decades, the observational capabilities of single instruments have been improved largely. It has become possible to retrieve measurements of very large areas of the sky without important variations in the observational properties (noise, calibration, etc.). The improvement in the overall properties of observations has made possible the production of surveys than can cover relevant fractions of the sky. Some examples include the FIRST survey, the Two Micron All Sky Survey (2MASS; Cutri et al. 2003a, 2003b; Skrutskie et al. 2006; Wright et al. 2010; hereafter 2M), VLASS, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016), the Galaxy Evolution Explorer (GALEX; Morrissey et al. 2007), and AllWISE (Cutri et al. 2013; hereafter AW). In the near future, they will be complemented by the LSST in the Vera C. Rubin Observatory, SKA, ngVLA, and *Euclid* (e.g. Euclid Collaboration et al. 2022, 2024), among others.

While being able to obtain information from more sources and regions of the sky is, by

itself, a very relevant improvement, such number of new measurements to analyse have brought some issues related to the treatment of very large datasets (for a review focused on the challenges of future radio surveys, see Norris 2017). Some of these obstacles are described in the following sub-sections.

1.2.1 Computational costs

Using very large surveys and catalogues for any sort of calculation involves, accordingly, very high computational costs. Recent observational catalogues might have up to billions of entries with several attributes each and images that can cover thousands of square degrees with very high angular resolution, reaching total sizes of tens of TB of data and tens of PB for future surveys (Mickaelian 2020; Zhang and Zhao 2015). Additionally, survey instruments will have data transfer rates well above the normal capabilities of a medium-sized server, reaching up to several TB/d rates (Enke et al. 2012). Dealing with such large datasets requires large amount of resources that are not completely available to everyone (Garofalo et al. 2017).

Additionally, most of the methods traditionally used for the detection, classification, and extraction of properties for astrophysical sources have not been developed, or updated, to be used with very large catalogues. For this reason, using them in the most recent catalogues and surveys can take restrictive running times that not even the most powerful computing facilities can deal with given their memory or central processing unit (CPU) usage (Mathews et al. 2023). Even if current methods are optimised for their use in large computational facilities, running times would still be prohibitively long. For instance, state-of-the-art SED methods can take between 50 CPU−h to 100 CPU−h for the analysis of a single galaxy (Leja et al. 2019; Gilda et al. 2021; Tacchella et al. 2022).

A further factor to consider is that of the energy expense of running such methods for long times in powerful machines. Excesive power consumption can impact negatively, first, in the economical costs of running calculations and, second, in the emission of greenhouse effect gases derived from the energy needed for computation. As a way to put these costs into number, it is possible to take the example of the LSST, which is expected to produce photometry for $\sim 10^{10}$ galaxies (Ivezić et al. 2019). Using traditional methods, analysing such number of sources would produce and estimate of 10^8 kg of CO_2, comparable to 200 d of continuous operation of a wide-body aircraft (Mathews et al. 2023). Effective reduction of CO_2 emissions from energy consumption can help alleviating the impact from the climate change (IPCC 2022)

and, while complete net-zero systems are expected to solve this issue, short-term reduction in the use energy spending are needed to help limiting global warming.

With a focus on ease of use and code readability, `Python` has become a standard language in most of recent astrophysical packages (Astropy Collaboration et al. 2022). Conversely, code written, purely, in `Python` tends to be one one of the least efficient in its energy and ecological impact (Pereira et al. 2017, 2021; Portegies Zwart 2020; Reya et al. 2023). Thus, and taking into account that the popularity of `Python` is not expected to decrease for the moment, it is needed to use techniques and code that can obtain results in shorter times than those available to date. One of the ways to improve the efficiency of `Python` programming is through the inclusion of faster lenguages for the most demanding sections of the algorithms as happens, for instance, with the libraries `Numpy` (Harris et al. 2020), `Scipy` (Virtanen et al. 2020), and `JAX` (Bradbury et al. 2018), which include lines of code written in `C`, `C++`, and `Fortran`, among others. With the advent of larger and more complex datasets, further steps are needed for the improvement of efficiency of algorithms and methods.

1.2.2 Missing measurements

As with any sort of pysical measurement, a fraction of observations might have issues that can render them unusable for any meaningful calculation (Rubin 1976; Josse and Reiter 2018). These problems include malfunction of detectors or incorrect cleaning of the data, among others. If different measurements (i.e. in different points in time) are to be combined, some sources might have been observed in one instance but not in the remaining ones. This might affect the study of time series or multi-wavelength, multi-instrument observations as this effect might increase the level of uncertainties. Furthermore, some analysis methods require all measurements to be available and, thus, the lack of one of them can render the full set of quantities from a source useless (Little and Rubin 2014) or with very relevant uncertainties and biases.

For the specific case of Astrophysics, an additional way of treating missing measurements is related to left-censored data, also called upper limits (Cohen 1957, 1961). If a known source is observed with an instrument, but no detection is made, it can be assumed that the emission of such source is below the detection limit of the measurement. This intuitive description has been formalised into a probability framework and included into the full treatment of astrophysical data (Feigelson and Nelson 1985; Isobe et al. 1986; Kashyap et al. 2010). While extensive work

has been formulated to extract the largest amount of information possible from upper limits, they still represent a source of nuisance for the calculation of properties.

1.2.3 Data heterogeneity

Another source of criticism for the use of ML methods is that the use of multi-wavelength observations of large areas of the sky can give rise to heterogeneity issues. Over time, many surveys and instruments gather data from many different areas in the sky and with very different sensitivities and observational properties. This makes applying ML techniques, and most of astronomical studies in general, a difficult task. ML modelling assumes, in general, that the data for all elements in the data set come from the same sources and have the same properties (Witten et al. 2011; Surana et al. 2020; Brescia et al. 2021) and thus steps have to be taken in order to obtain such properties when measurements of different quality are used.

One way to overcome this obstacle is generating observations of very large areas in the sky which can be analysed and compared with different data sets, thus covering a larger fraction of the available parameter space. As mentioned previously, in the following years, new facilities will be built and put into service delivering observations with similar qualities for large areas of the sky. This will allow the study of much more objects and sources in a statistical way without facing the downsides of inhomogeneous data.

1.2.4 Multi-wavelength counterpart identification

In the case of observations with different filters or different instruments, a new problem might arise. It involves the correct identification of the sources observed in each of the filters. Given that the emission in different wavelengths and different moments in time might come from separate components and processes in the studied objects, each observed instance can present a structure that does not match the others. For this reason, finding and matching counterparts for detected sources can be difficult. This problem is enhanced when observations in several bands, different instruments, and different sensitivity limits need to be combined as different point-spread functions (PSFs) have to be involved in calculations.

While optical, IR, and radio surveys have reached sub-arcsecond positional accuracy (e.g. Wright et al. 2010; Chambers et al. 2016; Shimwell et al. 2019), deep radio surveys have only recently obtained sufficiently high angular resolutions for very long-baseline observations

(e.g. Sweijen et al. 2022; Ye et al. 2023). This differences in resolution impact the correct identification of counterparts accross surveys and wavelengths, increasing uncertainties in the use of SED modelling.

There are a few approaches to find and link multi-wavelength, multi-instrument counterparts of sources. A straightforward way to find counterparts is through direct cross matching between catalogues. A search radius is defined and centred in the position of the source for which counterparts are needed. Then, all sources in the target catalogue located inside the circle of the previously defined radius are considered candidate counterparts. Depending on the conditions of the problem and the used catalogues, one of these candidates can be selected as the proper counterpart (e.g. the closest source either in angular or geometrical distance). Examples of the use of direct cross match of catalogues are Barbieri and Bertola (1972), Agüeros et al. (2005), Bianchi et al. (2007), Drake et al. (2014), Norris et al. (2021), and Storey-Fisher et al. (2024).

When postional errors, PSFs or synthesised beams are large enough to have several sources from the other catalogue inside it, direct cross matching cannot be used. Even if there is only one source within the search radius, it cannot be guaranteed that it corresponds to a counterpart. A more advanced approach is that of maximum likelihood ratio (MLR; Richter 1975; de Ruiter et al. 1977; Prestage and Peacock 1983; Wolstencroft et al. 1986; Sutherland and Saunders 1992). As defined by Sutherland and Saunders (1992), it looks for the sources that optimise the ratio of the likelihoods of being a genuine counterpart over that of being a background candidate. These likelihoods depend on the density of sources in both catalogues, their magnitude distributions, and their positional errors (e.g. Brusa et al. 2007). One advantage of this technique is that it can output the degree of reliability of each counterpart allowing the researcher to select, if needed, the most secure sources. Some examples of the application of the MLR method include Brusa et al. (XMM-COSMOS; 2007), Abdo et al. (Fermi; 2010), Xue et al. (Chandra Deep Field-South; 2011), LaMassa et al. (Stripe 82X; 2016), Marchesi et al. (Chandra COSMOS; 2016), Ananna et al. (Stripe 82X; 2017), Auge et al. (2023), Hardcastle et al. (LoTSS; 2023), and Whittam et al. (2024: MIGHTEE-COSMOS).

For the selection of counterparts among radio surveys, which can have very different spatial resolutions, a different approach can be used. Apart from the distances among candidate counterparts, the sizes of the sources extracted from an image (including whether they can be resolved or not) and their fitted Gaussians are taken into account. If the Gaussians have some degree of overlap, pairs of sources can be selected as counterparts. This approach has been

applied by, for instance, Böhme et al. (2023) with measurements in several radio surveys.

A fourth method is the Bayesian approach. Contrary to the previous techniques, it does not rely on the specific distribution of sources in the studied catalogues, using Bayesian priors to derive the most likely counterparts of the base catalogue. In this way, it does not suffer from being applied to small areas (Salvato et al. 2018; `NWAY`). This method was first introduced by Budavári and Szalay (2008) and it can be applied to the search for counterparts in simultaneous catalogues.

Additionally, ML-based methods can be used to derive the most likely counterpart of sources in catalogues. By using photometric information (or other properties) from sources detected in other wavelengths, it is possible to train a model and extract the probability of that source to have a counterpart in a new catalogue. One early example of such technique is the work by Rohde et al. (2005, 2006) where the authors used support vector machines (`SVM`s; Vapnik 1995; Cortes and Vapnik 1995), together with model calibration (see Sect. 3.3) in order to obtain a counterpart probability. More recently, Liu et al. (2019) used Gaussian process (GP; Rasmussen and Williams 2005) modelling to quantify the confidence of associations of Atacama large millimeter/submillimeter array (ALMA) detections in the COSMOS fields. Furthermore, Schneider et al. (2022) used `SVM`s to extract stellar counterparts of sources in the eROSITA Final Equatorial Depth Survey (eFEDS; Brunner et al. 2022) and Alegre et al. (2022) used a binary classifier to identify LoTSS sources that require visual inspection (rather than automated methods) for the reliable selection of optical and MIR counterparts.

1.3 Machine-assisted pattern detection

Taking into account all the issues that the analysis of large datasets might pose, new tools have been developed as a way to tackle them. For astrophysics, in particular, the existence of these major AGN detection, radio measurement, and redshift determination methods raises the need of new techniques which might be able to obtain these properties for large amounts of astrophysical sources with enough precision within a shorter amount of time.

Given that this is a problem suffered by several scientific and, even, non-scientific disciplines (e.g. business-related applications; Costa-Climent et al. 2023), large efforts have been put in order to solve it and many techniques have been developed to deal with the ever-increasing data volumes. New statistical and computer methods can analyse thousands or millions of ele-

ments and find relevant trends among their properties (Garofalo et al. 2017) within reasonable time frames. One branch of these techniques is able to, using previously-fed data, predict, with relevant confidence, the behaviour new data will have –i.e. the values of their properties–. This is what has been called ML.

In Astronomy, ML has been used in a wide range of subjects, such as redshift determination (e.g. Nakoneczny et al. 2021; Wenzl et al. 2021), morphological classification (e.g. Ma et al. 2019; Lukic et al. 2019; Mostert et al. 2021; Vardoulaki et al. 2021; Burhanudin et al. 2021), emission prediction (e.g. Dobbels and Baes 2021), anomaly detection (e.g. Baron and Poznanski 2017; Giles and Walkowicz 2019; Lochner and Bassett 2021; Storey-Fisher et al. 2021; Wagstaff et al. 2022), image reconstruction (e.g. Guglielmetti et al. 2022; Adam et al. 2023; Wilber et al. 2023), observations planning (e.g. Garcia-Piquer et al. 2017; Jia et al. 2023; Sravan et al. 2023), and more (Ball and Brunner 2010; Baron 2019; Sen et al. 2022; Huertas-Company and Lanusse 2023).

With ML, it is possible to use previously available measurements and extract useful trends and correlations that can suggest the behaviour of properties from future observations or simulations. ML models are, in general, only fed with measurements and not with physical assumptions (Desai and Strachan 2021) and they do not need to check the consistency of the predictions or results they provide. This can bring, as a consequence, that running times for this kind of algorithms might be less than typical physically-based codes (e.g. Buchner 2019; Mathews et al. 2023). One way to incorporate physical knowledge or assumptions into ML methods is through what has been called physics-informed ML (Miller et al. 2020; Karniadakis et al. 2021). Under this paradigm, algorithms can be modified to incorporate physical rules as priors or general conditions that analised observables must follow. For example, loss or target functions can be adjusted to reproduce border conditions or other physical principles.

1.3.1 Types of machine-assisted analyses

Concentrating our review on the application of ML, two main branches exist for the application of such techniques. The first of them, called supervised learning, deals with the idea that, for each set of measurements, there is a response value that, via modelling, we can predict with some degree of confidence (James et al. 2023). This definition implies that it is possible to determine, for a studied sample, values that could, otherwise, be measured. On the other side, unsupervised learning refers to the analysis of data that does not have an associated quantity.

One of the most popular applications of unsupervised learning is clustering of elements (e.g. Garcia-Dias et al. 2018; Reis et al. 2021; Mohale and Lochner 2024). Modelling data would imply separating them by how similar are their properties (or a combination of them).

Then, in the case of supervised learning, further divisions are possible. If the predicted variable (target) is a discrete quantity, this prediction is called a classification (e.g Saz Parkinson et al. 2016; Baron and Poznanski 2017; Ma et al. 2019; Lukic et al. 2019; Giles and Walkowicz 2019). Opposite to that, if the predicted target is continuous, the process is called regression (e.g. Vanzella et al. 2004; Nakoneczny et al. 2021).

1.3.2 Ensemble learning

Once a problem has been defined and ML techniques have been selected for its resolution, several techniques exist for the improvement and optimisation of their application and the results they provide. Given that a large fraction ML techniques are based in statistical tools, one straightforward way to improve them is by the enlargement of the analysed dataset. The inclusion of a larger number of elements can help covering a broader area of the space of parameters of the problem (including uncertainties and biases) and the likelihood of retrieving better answers is higher than with small samples.

Another option for the improvement of ML-based results is through the combination of several techniques and methods. In the context of ML, such combination involves the use of a number of, either different algorithms or several instances of the same tool. By design, each ML algorithm has been developed and tuned to work better with certain data conditions, that is, balance of target categories, ranges of base features, sparsity of values, etc. Such combination of ML algorithms is called ensemble learning. It involves the joint use of individual results from ML models, that have been trained to solve the same problem, into one larger model or rule that can deliver a final prediction (Schapire 1990; Breiman 1996; Freund and Schapire 1996). It has been shown that the combination of several models, and their predictions, can improve the overall prediction results (Opitz and Maclin 1999).

In order to combine several predictions into a final result, several options are available. The most used and also one of the earliest ways to merge all individual estimations consists of averaging each predicted value into the final prediction (e.g. Sollich and Krogh 1995). Such average can be obtained for either different models or several instances of the same model. This option is useful for both regression and classification tasks (using its output scores). For

the specific case of classification, a voting system can be implemented, where the majority of decisions of the base individual predictors is taken as the final predicted class. This method has been proven to work efficiently (Schapire et al. 1998) with a reduction in the test errors.

An alternative to combine the predicting power of different algorithms is the use of 'meta-learners' (Vanschoren 2019). Meta-learners use the properties or predictions from other algorithms (base learners) as additional information during their training stages. A simple implementation of this procedure (and a third method to combine individual predictions) is called Generalised Stacking (Wolpert 1992) which can be interpreted as the addition of priors to the model training stage. In this way, two levels of predictors are used. The first level includes all the individual models that are trained on the training set. The second level corresponds to a single model which is trained only on the outputs of the first-level models. An example flowchart of this process is presented in Fig. 1.8. Generalised stacking has been applied in several astrophysical problems. That is the case of Zitlau et al. (2016), Carvajal et al. (2021), Cunha and Humphrey (2022), Moya and López-Sastre (2022), Zammit and Adami (2023), and Euclid Collaboration et al. (2023a,b).

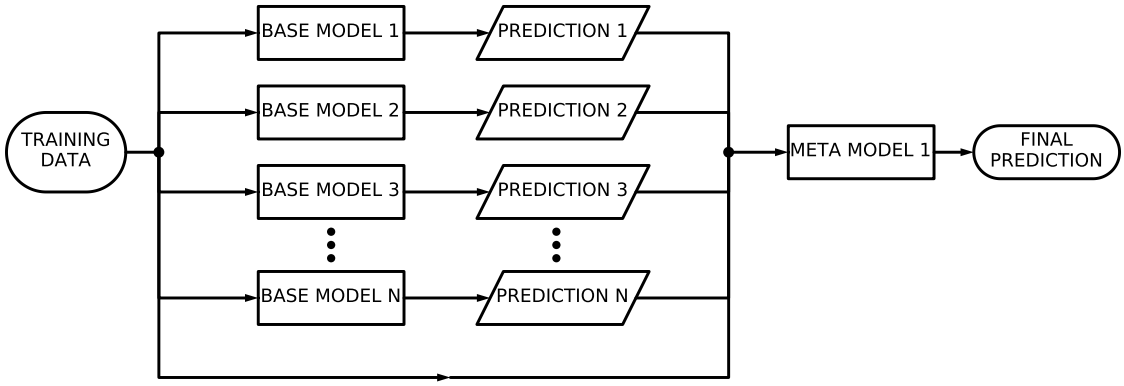


Figure 1.8: Illustrative diagram of generalised stacking. Training data is fed to each base model and to the final meta model. Output of base learners is also fed into the training of the meta learner.

1.3.3 Model explainability and feature importance

Despite the large number of applications it might have, ML has received important criticism related to the lack of interpretability –or explainability, as it is also called in ML jargon– of the their derived models, trends, and correlations (e.g. Linardatos et al. 2021; Gao and Guan 2023). Some of the most complex ML models, after taking a series of measurements and properties as input, deliver a prediction of a different property (or a set of them). But they cannot provide coefficients or an analytical expression, that might allow to find an equation

for future predictions (Goebel et al. 2018). An important counter-example of this fact is the use of Symbolic Regression (Gerwin 1974; Langley 1977, 1979; Langley et al. 1981; Langley and Zytkow 1990), which has been developed to extract explicit analytic expressions from the analysed data. Some examples of the use of symbolic regression in astrophysics are Cranmer et al. (2020), Villaescusa-Navarro et al. (2021), and Cranmer (2023). The lack of explainability implies that, for most ML models, it is not a simple task to determine which properties, and to what extent, help predict and interpret another attribute (e.g. Roscher et al. 2020a). This fact hinders our capability to understand the results in physical terms and extrapolate ML results into further ranges of properties.

Besides the development of symbolic regression, recent work has been done to overcome the low levels of explainability in ML models (Främling 2023). The most widely used assessment is done with feature importance (Casalicchio et al. 2019; Roscher et al. 2020b), both global and local (Saarela and Jauhiainen 2021). Feature importance analysis helps to know the relative weights that the measured properties have in the decision-making process (D'Isanto et al. 2018; Främling 2023). In this way, physical insight might be gained about the correlations or triggers between different properties (observed or derived) of the objects of study.

Global feature importances

Overall, mean or global feature importances can, usually, be retrieved from models that are based on decision trees (DTs) (e.g. random forests and boosting models, Breiman 2001, 2003). For each feature, the decrease in impurity (a term frequently used in the literature related to ML; Louppe et al. 2013) of the dataset is calculated for all the nodes of the tree in which that feature is used. Features with the highest impurity decrease will be more important for the model (Louppe et al. 2013). For some models that are not based on DT, feature importances can be obtained from the coefficients that the training process delivers for each feature. These coefficients are related to the level to which each quantity is scaled to obtain a final prediction (as in the coefficients from a polynomial regression). Insight into the decision-making of the pipeline can only rely on the specific weights of the original set of features (see Sect. 3.4). An additional example method of the retrieval of global feature importances is that of the creation and training of simple, linear surrogate models, which are completely explainable for the studied sample (e.g. Saarela and Jauhiainen 2021).

Local feature importances

As opposed to the global (or mean) assessment of feature importances derived from the decrease in impurity, local (i.e. source by source) information on the performance of such features can be obtained from, for instance, Shapley values. This is a method from coalitional game theory that tells us how to fairly distribute the dividends (the prediction in our case) among the features (Shapley 1953). The previous statement means that the relative influence of each property from the dataset can be derived for individual predictions in the decision made by the model (which is not the same as obtaining causal correlations between features and the target; Ma and Tourani 2020). Game theory based analyses, such as the Shapley and SHapley Additive exPlanations (SHAP) values, have also been used to understand the importance of features in astrophysics (e.g. Machado Poletti Valle et al. 2021; Carvajal et al. 2021; Dey et al. 2022; Anbajagane et al. 2022; Alegre et al. 2022; Carvajal et al. 2023a; Pearl et al. 2024).

A different approach to local feature importances is that of Local Interpretable Model-agnostic Explanations (LIME; Tulio Ribeiro et al. 2016). LIME can be used to explain individual predictions regardless of the type of model used. It is based on the idea that a fully explainable model can be created to mimic the result of a specific prediction. Then, this prediction can be perturbed by the removal of each of the involved features. The analysis of this perturbation will return the feature importance reported by LIME. In astrophysics, LIME has been used by, for instance, Ulmer-Moll et al. (2019) and Pasquato et al. (2024).

1.4 This thesis

As mentioned earlier in this text, most multi-wavelength measurements from extra-galactic sources convey information from AGN and their host galaxies. Thus, separating both components becomes a complex process given that different wavelengths will carry different fractions of information from each component of the observed sources.

Then, and as also presented in Sect. 1.1, one relevant exception to this behaviour are radio measurements. Given that radio light can escape the galaxy without major obscuration or absorption, this emission has been, historically, better suited for obtaining direct information from the central regions of bright AGN.

However, until recently, major radio observatories and surveys lacked the capabilities to resolve, meaningfully, the emission from distant (and sometimes, faint) AGN. Only with the

advent of recent facilities, we have the capabilities to better establish the radio nature of high-redshift sources. To add more complications, these new exquisite measurements are as sensitive as to capture emission from faint SF episodes as it happens with other wavelengths.

Consequently, we face a complex issue if we want to extract information from AGN (and in particular, high-redshift AGN). We are able to observe them directly in radio wavelengths but, so far, it is difficult to determine the exact origin of radio emission. On the other hand, the lack of strong observational connection between the AGN radio emission with their host galaxies makes it difficult to establish strong correlations or trends between their intensities and other galactic measurements (mostly, in non-radio wavelengths).

Attending, then, to the difficulties in relating radio emission from AGN with measurements of these sources in additional wavelengths, the main goal of this thesis is to explore and understand possible indicators of the radio emission in AGN from multi-wavelength, multi-instrument, measurements. Thus, we want to develop a process that, in particular, can take information from IR-detected sources (for which there is all-sky coverage with good sensitivity levels) and deliver an indication of whether these sources can correspond to AGN and, more specifically, to radio-detectable AGN or not.

Given the importance of the detection of AGN in early epochs of the Universe and the need to compare their intrinsic properties, we also aim to use the aforementioned machinery to derive estimates of photometric redshifts for the sources labelled as prospective radio-detectable AGN. In that way, the focus of our search can be put in the selection of sources as close as possible to the EoR. Or at least, in redshift ranges suitable for specific studies.

Having outlined the major issues that exist with the use of new astronomical datasets and surveys (cf. Sect. 1.2), the use of machine-assisted techniques (and in particular, ML, which aims at the use of available datasets to find relevant trends among their properties to estimate the behaviour of new, unseen data; Samuel 1959) becomes more relevant than ever before. The possibility of analysing very large datasets with reduced computational costs (in time and energy consumption, Sect. 1.2.1), and with minimal homogenisation procedures applied to them, is one of the main drivers behind the development of this work.

In Fig. 1.9, we present a flowchart of the prediction pipeline we propose for the generation of radio-detectable AGN candidates. We aim to start with a set of IR-detected sources with ancillary multi-band data that can be fed into the a first step that classifies between AGN and SFGs (i.e. not hosting an AGN). Given that we are interested in AGN, we use the predicted AGN

and feed them into the second step, which classifies AGN according to their radio detectability. After this step, we select the predicted AGN that have a high likelihood of being radio-detectable. A final step uses the predicted radio-detectable AGN and estimates a redshift value for them. Consequently, the prediction pipeline delivers a set of candidate radio-detectable AGN with a redshift estimate.

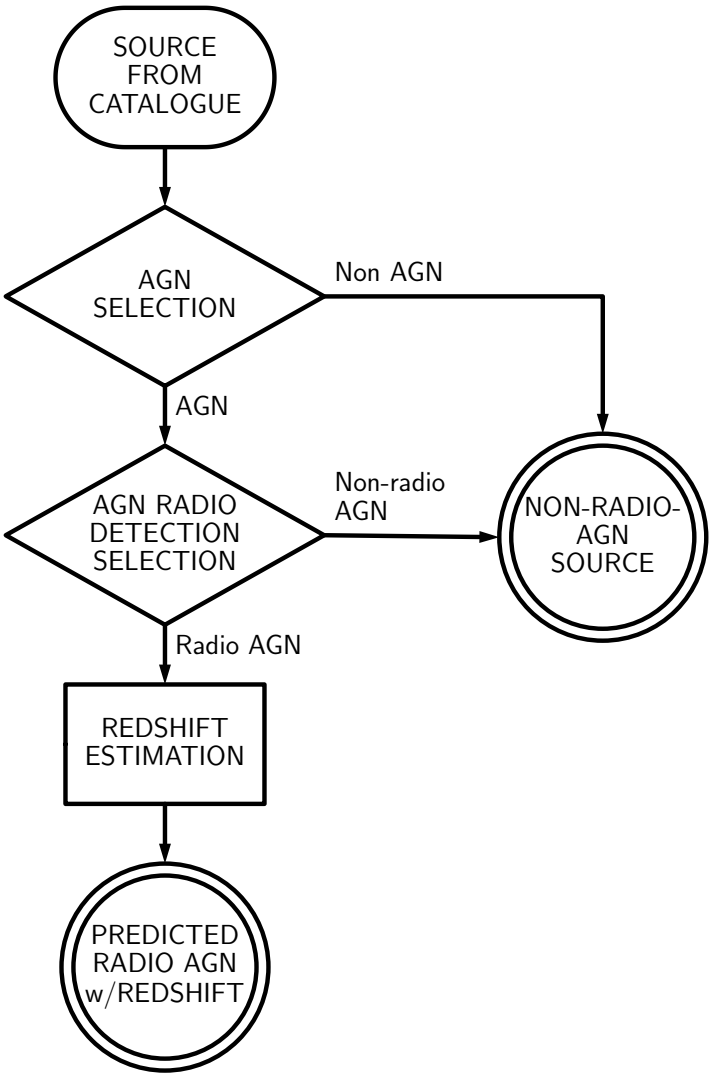


Figure 1.9: Flowchart representing the proposed prediction pipeline used to predict the presence of radio-detected AGN and their redshift values from IR-detected sources. Sources ingested to the pipeline have been already processed. Each step takes as input the results from the previous stage of the pipeline.

Taking into consideration the structure of the pipeline, the ML models in each step will be trained with a different sample of sources. The first step, classification between AGN and SFGs will be trained with all available sources that have been labelled previously as either AGN or SFGs. The second step, classification of radio detection in AGN, will be modelled only

with confirmed AGN (with or without radio detections). Finally, the third step of the pipeline, which estimates photometric redhsifts, will be trained with radio-detected AGN. Our focus on radio-detectable AGN is the basis for the omission of the remaining sources in the data sets.

Following the production of candidates for radio-detectable AGN, together with their redshift values, we aim to understand the predictions and how they relate to phyisical properties of the analysed sources. For this goal, we want to apply feature importance analyses to the prediction processes. These techniques can help understanding the inner correlations and trends that allow the creation of several selection rules and prediciton schemes for the creation of predicted sources and some of their properties.

Once the mechanisms leading to the prediction of radio-detectable AGN, and their redshift values, have been understood, we want to apply such indicators and the candidates derived from their use to the analysis and possible solution of different problems related to the observation, classification, and distribution of radio-detected AGN. The application of ML techniques can help creating large collections of candidate radio-AGN that might not have been available previously. The use of such sources might, then, contribute to the improvement of the answers for the questions previously mentioned.

Part of this thesis is based on the work and analyses presented by Carvajal et al. (2021) and Carvajal et al. (2023a). In the following chapters, we present the data sets (as well as the treatment applied to them) used for the generation of the models (Chapter 2), the result of the selection of models, their training, calibration, and the analyses of the use of the prediction pipeline on the selected data sets (Chapter 3). Futhermore, Chapter 4 describes the analysis of the predictions and the models themselves. Finally, Chapters 5 and 6 introduce extensions of the use of the results from the prediction pipeline in different contexts as well as future developments to be applied to the prediction pipeline and its individual steps. This thesis concludes with a summary, where final remarks and findings are outlined.

This page intentionally left blank.

2

Datasets for training and testing

In order to train all models efficiently and test them without fear of obtaining biased metrics, good quality and abundant data are needed. These requirements can be translated into selecting a field with coverage in several bands and by diverse instruments. Such variety can help the training of the models to cover a broad fraction of the parameter space. At the same time, a broader wavelength coverage improves the photometric sampling of the physical processes (and their emission) occuring in the AGN and their host galaxies. This coverage also needs to be deep enough to detect a large fraction of the sources to be studied in a given area.

Furthermore, these measurements need to be spread over a sufficiently large area as a way to avoid any biases from using sources that might be connected in some manner to the particular region of the sky that is under study (e.g. cosmic variance and large-scale structure; Driver and Robotham 2010). Additionally, to validate the predictions from the models, the selected field needs to have an adequate number of sources with accurate labels.

To improve the adaptability of our models to real-world variations in data quality (see Sect 1.2.3), we will test them using datasets with varying levels of heterogeneity. While we will minimise the impact of heterogeneity, we will not entirely eliminate it. This approach allows us to utilise a wider range of datasets during training and testing.

Since most AGN identifications rely on optical or IR measurements, we will use these same wavelength bands for both training and testing our models. Only the variations within each dataset will be incorporated into the models. However, for radio measurements, which are only used for detection prediction and not for deriving other properties, this requirement is relaxed. This change allows us to leverage radio surveys with different frequency measurements, angular resolutions, and depths.

Finally, and given that one of our goals is predicting radio detectability of sources, the chosen area for training must have deep and homogeneous radio coverage. If a shallow radio survey is used, the model will have access to only a small fraction of potential RGs detections, which might bias the predictions and their assessment. This deep radio coverage requirement

is extended to an additional testing region, although the specific radio measurements might differ. Still, both areas need to be at sufficiently similar levels for the correct evaluation of the predictions.

For all the aforementioned reasons, we have selected two sufficiently large areas for our study: the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX; Hill et al. 2008) Spring field (Gebhardt et al. 2021) as the training area, and the SDSS Stripe 82 (S82; Annis et al. 2014; Jiang et al. 2014) field, as testing region, as they are both covered to µJy levels. The observational coverage in these fields is detailed in the following sections.

2.1 HETDEX Spring field

As training field we selected the area of the HETDEX Spring field (Gebhardt et al. 2021). The HETDEX experiment aims at studying the the redshift distribution of Lyα emitting galaxies in the redshift range $1.88 < z < 3.5$ over an area of 540 deg^2 for a high-quality measurement of the Hubble expansion parameter and the angular distance diameter distance at $z \sim 2$ (Papovich et al. 2016). The design of the Hobby-Eberly Telescope (HET) integral field unit (IFU) does not allow full steerable operation leading to observations at a fixed elevation of 55°. Such constrain has led to the division of the HETDEX field into two distinct areas for optimisation of observation time. One high-declination region, the Spring field (390 deg^2), and the equatorial Fall field, covering 150 deg^2 (Gebhardt et al. 2021), which are expected to be fully observed with the HET by the end of 2024. For the correct identification and characterisation of emission lines in both HETDEX regions, ancillary data will be needed, improving the coverage of the studied areas. Davis et al. (2023) list all existing surveys initially incorporated in the HETDEX analysis.

The first-class capabilities of the HETDEX and the available observations over their regions have sparked interest from additional surveys and observatories. Notably, one of them, the Low Frequency Array (LOFAR; van Haarlem et al. 2013) observatory has covered the HETDEX Spring field with the first data release of the LoTSS, which is centred at right ascension 12h59m0s and declination 53°0′0″. The LoTSS - data release 1 (LoTSS-DR1; Shimwell et al. 2019) survey covers 424 deg^2 in the region of the HETDEX Spring field (see Fig. 2.1) with LOFAR 150 MHz observations that have a median sensitivity of 71 µJy/beam (65 %, 90 %, and 95 % of the HETDEX Spring field area has noise levels below 78 µJy/beam, 115 µJy/beam, and 147 µJy/beam,

respectively; Shimwell et al. 2019) and an angular resolution of 6″. The deep and homogeneous radio observations in this field can help the training stages to retrieve information from a large fraction of sources in the area. The LoTSS-DR1 has selected the HETDEX Spring field (hereafter, HETDEX field) because of its large contiguous area at high elevation combined with a very high overlap with the SDSS survey (Shimwell et al. 2017).

It has been highlighted that, given their configurations, LOFAR and HETDEX observations can be combined for the study of common goals. Both surveys can help tracing, for instance, the star formation rate density (SFRD), merger rate of galaxies, and AGN populations at their peak of activity ($1.9 \lesssim z \lesssim 3.5$; e.g. Jarvis and Rawlings 2000; Madau and Dickinson 2014; Conselice 2014; Rigby et al. 2015; Shimwell et al. 2017). The deep and spatially resolved observations of the LoTSS-DR1 and the rich multi-wavelenght coverage of the HETDEX field, have made it a perfect selection as a training area for our pipeline. In Sect. 2.3, we describe the coverage used in this work, namely from surveys produced with Pan-STARRS, 2M, and *WISE*.

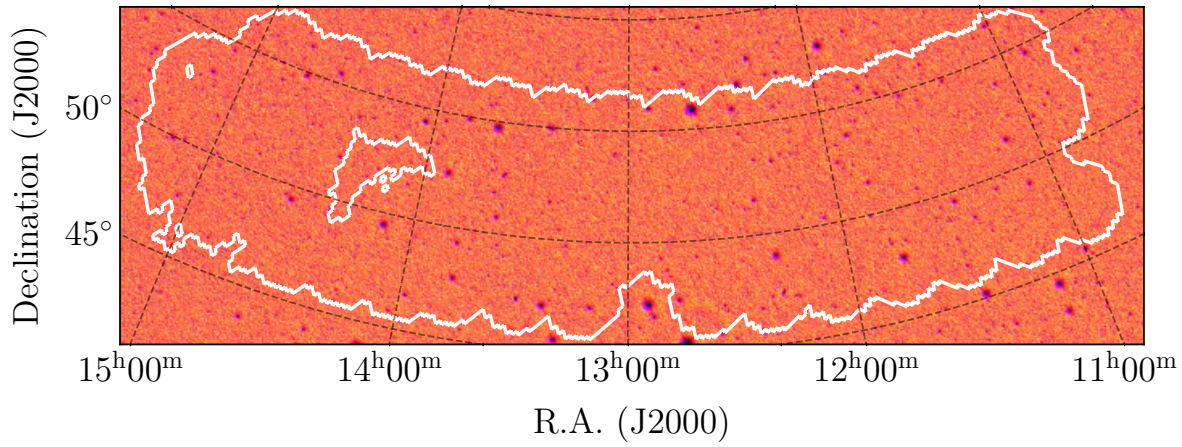


Figure 2.1: Footprint of the area used in the HETDEX field for this work. In the background, W1 image from the un*WISE* coadds (Lang 2014; Meisner et al. 2022). The white countours limit the area of the HETDEX LoTSS-DR1 field, covering 424 deg^2.

Our models and analyses have been fed with positions and integrated flux densities of the radio source catalogue from the LoTSS-DR1 survey. Shimwell et al. (2019) obtained a total of 325 694 sources (i.e. a source density of $\sim 770\,\mathrm{deg}^{-2}$) in the HETDEX area by running the software PyBDSF on each of their 58 observed mosaics. A $5\,\sigma$ peak detection threshold was defined and a $4\,\sigma$ threshold was established to determine the boundaries of the source islands. A point-source completeness of 90 % at 0.45 mJy has been estimated for the source detection process. Catalogues from each mosaic were combined and duplicated sources were combined, keeping the measurement closest to the corresponding mosaic centre. PyBDSF can output

source positions, peak brightness, integrated flux density, source sizes, and orientations along with their statistical uncertainties.

2.2 Stripe 82 field

In order to test the performance of the models when applied to different areas of the sky, and with different coverages from radio surveys, we have selected the SDSS S82 field. The S82 field has been defined as a region covering 8h in right ascension from 20h to 4h and 2°.5 in declination from −1°.25 to 1°.25 (York et al. 2000; Bramich et al. 2008). The S82 has been repeatedly imaged between 1998 and 2005 (Adelman-McCarthy et al. 2008; Quinn and Smith 2009) with the initial goals of finding variable objects and reaching co-added magnitude limits much better than the rest of the SDSS survey (York et al. 2000; Annis et al. 2014). This region was selected as it allows good observing conditions during the boreal autumn months, when most of SDSS fields are not accesible from the Apache Point Observatory (Kowalski et al. 2009).

The large sampling of SDSS observations on the S82 field, has sparked multi-wavelength coverage and analyses that have allowed a myriad of works based on sources from this field. In particular, we want to take advantage of the exquisite AGN catalogues covering the area (e.g. Fu et al. 2015; Glikman et al. 2018; Baldassare et al. 2018; Gross et al. 2023; Savić et al. 2023; LaMassa et al. 2024). In the S82 field, we have access to a robust sample of AGN selected with several techniques, which can be an excellent test bed for the selection from our pipeline.

For the S82 field, we collected data from the same surveys as with the HETDEX field (Pan-STARRS, 2M, and *WISE*, see Sect. 2.3) but with one important caveat: no LoTSS-DR1 data is available in the field and, thus, we gathered the radio information from the VLA SDSS Stripe 82 Survey (VLAS82; Hodge et al. 2011). VLAS82 covers an area of 92 deg^2, centred at right ascension 0h13m8s and declination 0°7′37″, with a median rms noise of 52 µJy/beam at 1.4 GHz. We have selected the S82 field (and, in particular, the area covered by VLAS82, see Fig. 2.2) from the fact that it presents deep radio observations but taken with a different instrument than LOFAR. This difference allows us to test the suitability of our models and procedures in conditions that are not exactly the same as those from the training.

One expected caveat is that, given the shallower nature of the radio observations in S82, the model might predict the radio detection of a source but it might be fainter than the limit from S82. Following part of the description of Sect. 1.1 and our focus on AGN, we can assume a synchrotron

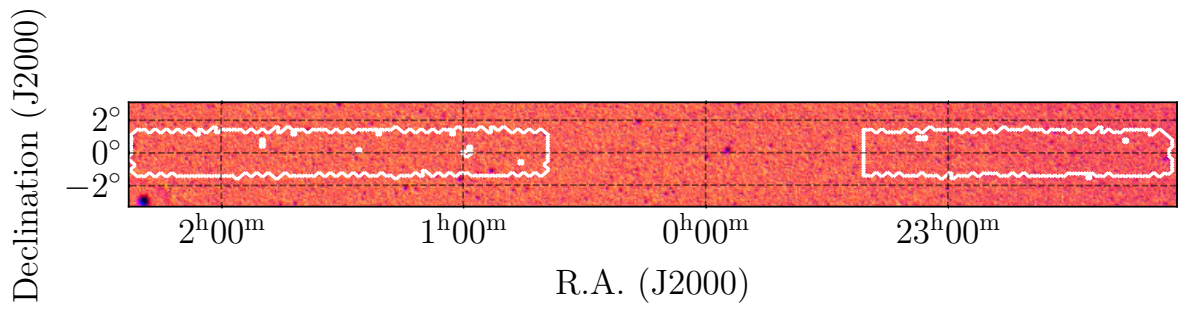


Figure 2.2: Footprint of the area used in the VLAS82 field for this work. In the background, W1 image from the un*WISE* coadds (Lang 2014; Meisner et al. 2022). The white countours limit the area of the VLAS82 field, which covers an area of 92 deg^2 split in two sections.

radio slope of $\alpha = -0.7$ (canonical for most radio source populations, e.g. Condon et al. 2002; Sabater et al. 2019). Thus, the 5 σ detection limit of LoTSS (355 µJy/beam) corresponds to $\approx$ 72 µJy/beam, which is below the 5 σ detection limit of VLAS82, 260 µJy/beam. Furthermore, if a typical distribution of AGN spectral indices between 150 MHz and 1.4 GHz is taken ($\langle\alpha\rangle = -0.78^{+0.33}_{-0.30}$ Calistro Rivera et al. 2017), the observations from LoTSS-DR1 will still remain below the 260 µJy/beam from VLAS82 ($60.2^{+67.3}_{-29.8}$ µJy/beam for LoTSS-DR1 at 1.4 GHz). This difference will be taken into account when comparing metrics between fields.

2.3 Photometry measurements

For the training of our pipeline, we have settled on the use of photometric coverage in three main wavelength ranges. Optical photometry can be of help for the selection of the large number of AGN identified using, for instance, spectral information in these wavelengths (e.g. York et al. 2000). In turn, NIR and MIR photometric measurements can help understanding the presence of AGN that have been heavily obscured, rendering an optical identification very unlikely (e.g. Lacy and Sajina 2020).

The base survey from which all the studied sources have been drawn is the CatWISE2020 (Marocco et al. 2021; hereafter CW). It lists MIR-detected elements selected from *WISE* and Near-Earth Object *WISE* (NEOWISE; Mainzer et al. 2011, 2014) over the entire sky at 3.4 µm and 4.6 µm (*W1* and *W2* bands, respectively). This catalogue includes sources detected at 5 σ in either of the used bands (i.e. *W1* $\sim$ 17.43 and *W2* $\sim$ 16.47 mag$_{\rm Vega}$ respectively). The HETDEX field contains 15 136 878 sources listed in CW. Conversely, in the S82 field, there are 3 590 306 of them. These values correspond to source densities of 35 700 deg^{-2} and 39 025 deg^{-2} for the HETDEX and the S82 fields, respectively, indicating similar conditions in

both fields.

It is important to note that there are MIR instruments that can deliver observations with better spatial resolution, potentially allowing better source identification. One relevant example is *Spitzer*, which can obtain photometry in 3.6 µm, 4.5 µm, 5.8 µm, and 8.0 µm with its instrument IRAC (i.e. very similar to the bands observed by *WISE*). However, we have selected *WISE* observations, and the CW survey in particular, because of their all-sky coverage, which allows us to obtain relatively homogeneous measurements in any region of the sky.

Applying a 1.″1 search radius criteria to match radio sources to Pan-STARRS data release 1 (Chambers et al. 2016; Flewelling et al. 2020; hereafter PS1) and *WISE* observations as it the smallest PSF size of the bands included in PS1 (Chambers et al. 2016), we have found multi-wavelength counterparts for CW sources in additional catalogues. As it matches the smallest scales in PS1, that search radius helps avoiding obtaining a large number of misclassifications and can alleviate the effect of source confusion. These catalogues include optical photometry in bands *g* (4866 Å), *r* (6215 Å), *i* (7545 Å), *z* (8679 Å), and *y* (9633 Å) from PS1, NIR bands *J* (1.235 µm), *H* (1.662 µm), and *Ks* (2.159 µm) from 2M, and the MIR bands *W3* (12 µm) and *W4* (22 µm) from AW[1]. Furthermore, the source density of the radio (LOFAR, Karl G. Jansky Very Large Array –VLA–) and 2M catalogues imply a low statistical (< 1 %) spurious counterpart association, this is not the case for PS1, where the source density is higher. For these reasons, and to mantain a statistically low spurious association between CW and PS1, we limited our search radius to 1.″1. A list of used bands, together with 5 σ limiting magnitudes and their origin instruments and surveys is shown in Table 2.1.

For the purposes of model training, observations in LoTSS-DR1 and VLAS82 are only used to determine if a source is radio detected. In particular, no check has been performed on whether a selected source is extended or not in any of the radio surveys. A single Boolean feature is created from the radio measurements (see Sect. 2.5) and no further analyses were performed regarding the detection levels that might be found in any of the fields.

Traditional ML models struggle to include measurement errors into their training processes (e.g. Jiang et al. 2021; Michelucci and Venturini 2023). Newer algorithms like GPs, natural gradient boosting (`NGBoost`; Duan et al. 2020), and conformal regression (Gammerman et al. 1998; Saunders et al. 1999; Vovk et al. 2022) can account for uncertainties and even provide them as outputs. GPs can model data by considering all possible relationships between data

[1]For the purposes of the analyses, and except when clearly stated otherwise, all photometric measurements were converted to AB magnitudes.

Table 2.1: Bands available, and selected, for model training in our dataset together with their limiting magnitudes.

Survey	Band (col. name)[a]	$5\,\sigma$ lim. mag. (AB mag)
Pan-STARRS (PS1)	g (gmag)	23.3
	r (rmag)	23.2
	i (imag)	23.1
	z (zmag)	22.3
	y (ymag)	21.4
2MASS (2M)	J (Jmag)	17.45
	H (Hmag)	17.24
	Ks (Kmag)	16.59
CatWISE2020 (CW)	$W1$ (W1mproPM)	20.13
	$W2$ (W2mproPM)	19.81
AllWISE (AW)	$W3$ (W3mag)	16.67
	$W4$ (W4mag)	14.62

[a] In parentheses are shown the names of the columns or features in our dataset that represent each band.

points, allowing for inherent uncertainty estimation. `NGBoost` is a recent algorithm designed for probabilistic prediction, meaning it can estimate the probability distribution of a target variable rather than just a single point prediction. Conformal regression, on the other hand, is a framework that can be applied to various algorithms to provide guarantees on the coverage of predictions (in the form of confidence intervals), even without assuming a specific data distribution. The astronomical community has also attempted to modify existing techniques to include uncertainties in their ML studies as shown by the work of Ball et al. (2008), Reis et al. (2019), and Shy et al. (2022).

Despite all the efforts, Euclid Collaboration et al. (2023b) have shown that, in specific cases, the inclusion of measurement errors might not improve the models and can be even detrimental to the prediction metrics. This degradation can likely be related to the fact that, by virtue of the large number of sources included in the training stages, the uncertainties are already encoded in the dataset as scatter. Thus, including explicit uncertainties might dilute the information from, in the case of astronomical data, the photometric measurements. The possibility of data degradation coupled with the desire to analyse the inner data correlations of the models, have induced us to rely on traditional ML algorithms and, therefore, discard the measurement errors of all bands and not include them in the training stages.

Finally, and as derived from the previous descriptions, the number of valid measurements

for each field and photometric band (i.e. all CW detections and their retrieved multi-wavelength counterparts in the selected fields) is shown in Fig. 2.3. Additionally, the associated source densities, obtained by dividing the number of valid measurements over the effective area of each field (Sects. 2.1 and 2.2), are shown in Table 2.2.

Table 2.2: Density of detected sources (in units of sources per square degree) per optical, NIR, and MIR band in each field (following band names from Table 2.1).

HETDEX Field							
Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})
g	6380.66	*z*	10 331.93	*H*	1335.55	*W2*	35700.18
r	9304.58	*y*	6735.97	*Ks*	1335.55	*W3*	14045.08
i	11 242.35	*J*	1335.55	*W1*	35 700.18	*W4*	14044.78
S82 Field							
Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})
g	8249.04	*z*	13 214.70	*H*	2330.92	*W2*	39025.05
r	12 962.35	*y*	9226.45	*Ks*	2330.92	*W3*	15393.12
i	14 507.01	*J*	2330.92	*W1*	39 025.01	*W4*	15472.75

2.4 Missing data treatment

In general, ML methods (and their underlying statistical methods, as introduced in Sect. 1.2.2) struggle with catalogues that have missing entries (Allison 2001; Josse et al. 2019). Several techniques have been devised to handle datasets that lack some of their entries. The simplest of them is list-wise deletion, which drops all observations that miss, at least, one measurement (Pepinsky 2018). This process is inneficient as it reduces the size of the parameter space from which the models can obtain information for its training. A second method is imputation. Imputation is the process of replacing non-available measurements with substitute values. In general, there is enough information in the remaining entries to derive a meaningful substitute value (Kalton and Kasprzyk 1982). In this way, typical quantities used for replacement are the mean of the remaining entries or a function of the other measurements of the same entry.

Depending on the origin of the substitute value, different categories of imputation exist. Using the definitions from Kalton and Kasprzyk (1982) and Chattopadhyay (2017), it is possible to separate them into multiple and single imputation. In turn, single imputation can be divided

into mean, random, regression, hot deck, and cold deck imputation. Single imputation replaces each missing entry with a single value, which could be the mean value of the available entries or any value (random) from the existent measurements. Also, regression imputation uses the full set of available data to derive a possible value for the missing entry. Hot and cold deck imputations replace the missing value with other instances of the data set (or additional data sets in cold deck imputation) that have the same values for the remaining measurements. While convenient, single imputation tends to introduce biases in the analysed sample as its variability is greatly reduced (Chattopadhyay 2017). Meanwhile, multiple imputation, and as initially proposed by Rubin (1987), creates a set of possible values (usually, based upon statistical arguments) which are included as new instances of the measured object.

Despite their potential issues, we have applied an ad-hoc variation of the most simple single imputation method to replace missing values and magnitudes fainter than $5\,\sigma$ limits with meaningful quantities that represent the lack of a measurement. We have opted for the inclusion of the same $5\,\sigma$ limiting magnitudes as the value to impute with. This method of imputation, with some variations, has been successfully applied and tested, recently, by Arsioli and Dedin (2020), Carvajal et al. (2021), Curran (2022), and Curran et al. (2022). Nevertheless, we acknowledge that more sophisticated imputation strategies could potentially improve the estimates produced by the models and reduce their biases (Tong et al. 2019).

In this way, observations from 12 non-radio bands were gathered (as listed in Table 2.1). The magnitude density distribution for the sample from the HETDEX and S82 fields, without any imputation, is shown in Fig. 2.3. After imputation, the distribution of magnitudes changes, as shown in Fig. 2.4. Each panel of the figure displays, in their upper-right corner, the number of sources which have a measurement above its $5\,\sigma$ limit in such band. Following the same argument of measurement errors, upper limit values and non detections have been replaced by the assumed missing value within their entries as described previously. Additionally, a representation of the observational $5\,\sigma$ limits of the bands and surveys used in this work is presented in Fig. 2.5.

In what concerns radio detection, it is worth noting the $5\,\sigma$ depth difference between VLAS82 and LoTSS-DR1 is $\sim 190\,\mu$Jy (260 µJy for VLAS82 and 72 µJy for LoTSS-DR1) for a typical synchrotron emitting source ($F_\nu \propto \nu^\alpha$ with $\alpha = -0.7$) at a frequency of 1.4 GHz, allowing the latter survey reach fainter sources (see Sect. 2.2).

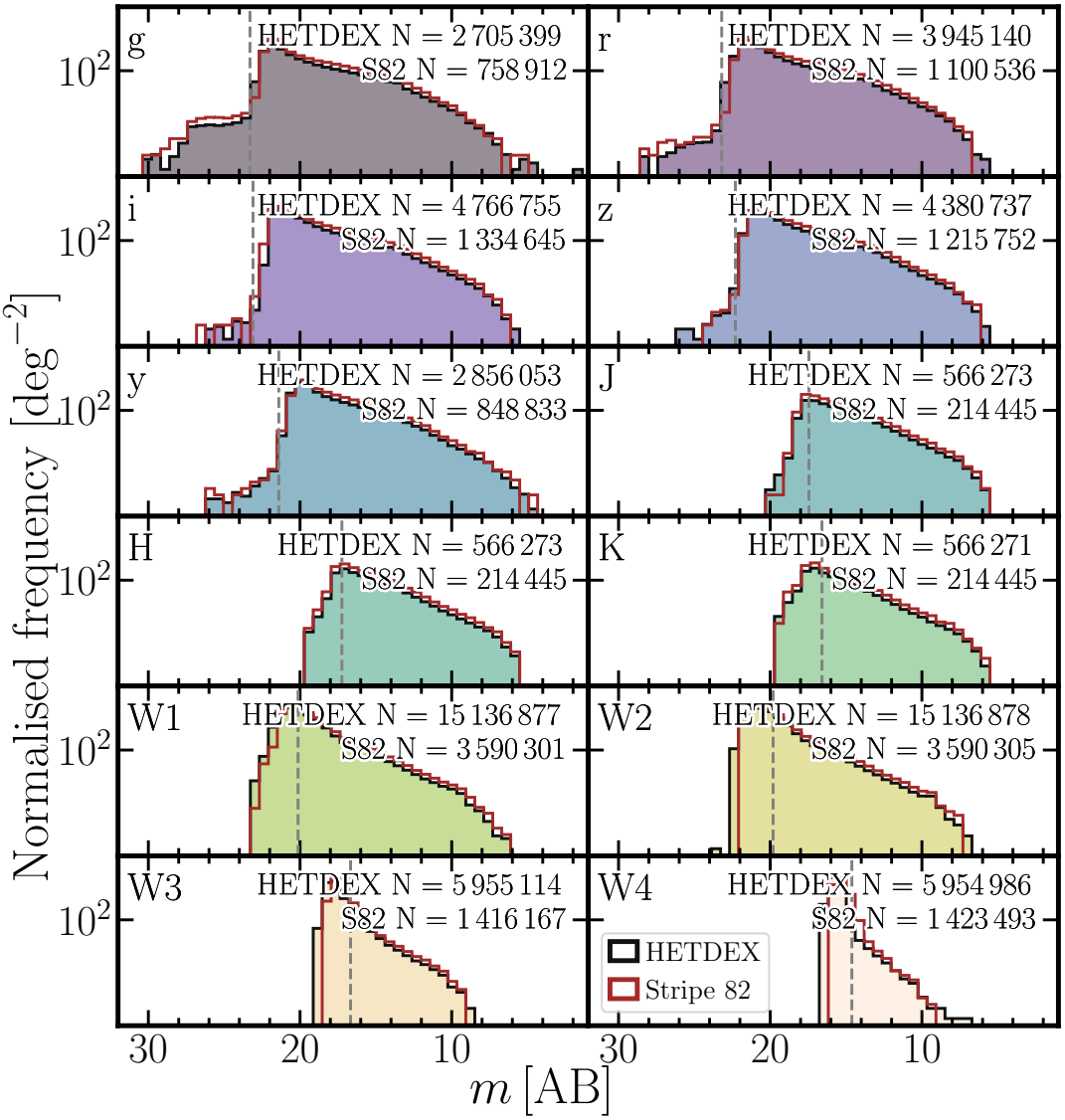


Figure 2.3: Histograms of collected, non-imputed, non-radio bands for HETDEX (coloured, background histograms) and S82 (empty, brown-outlined histograms) fields. Each panel shows the distribution of measured magnitudes of detected sources divided by the total area of the field. Dashed, vertical lines represent the 5 σ magnitude limit for each band. The number in the upper right corner of each panel shows the number of measured magnitudes included in their corresponding histogram.

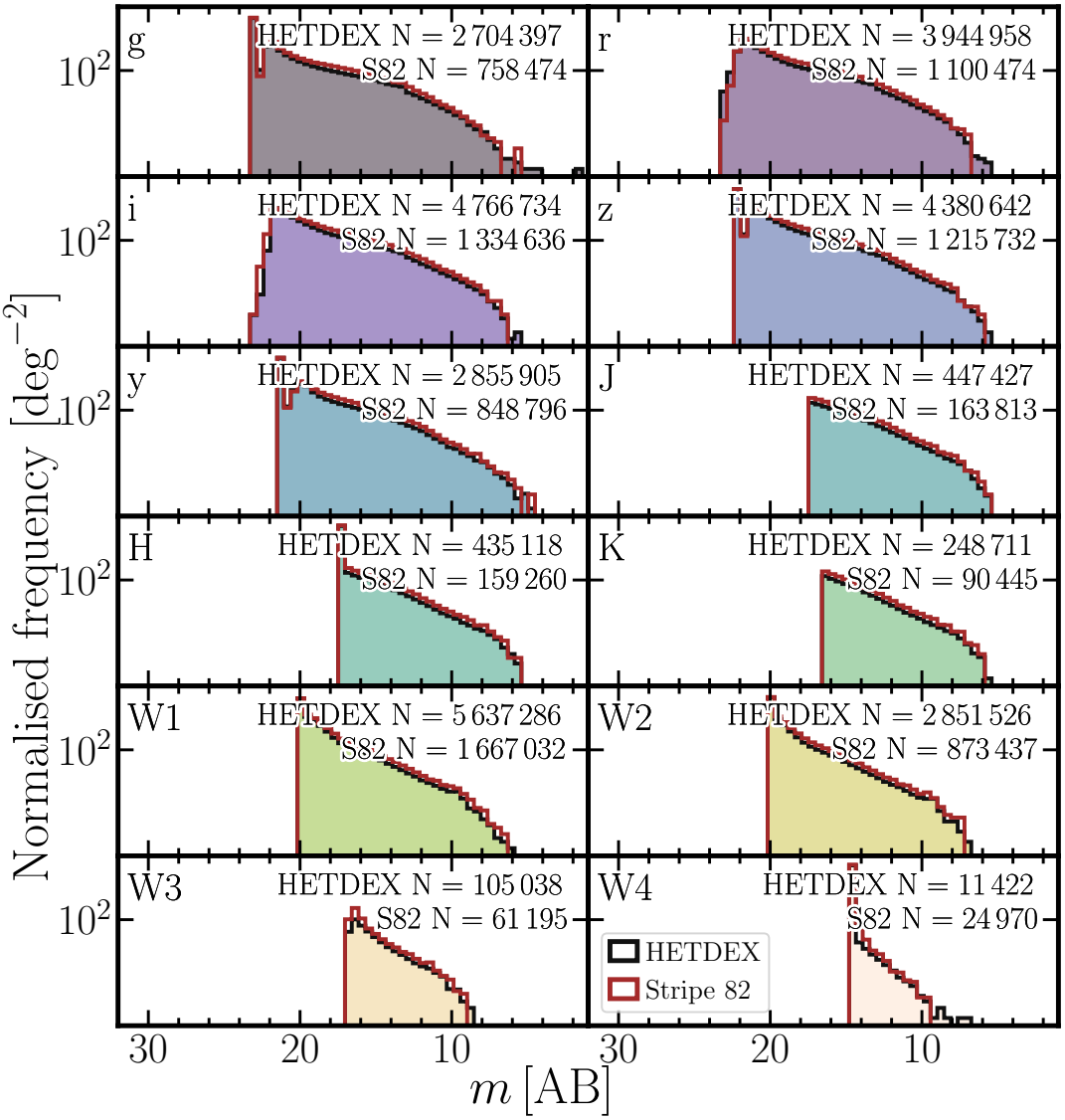


Figure 2.4: Histograms of collected, imputed, non-radio bands for HETDEX (coloured, background histograms) and S82 (empty, brown-outlined histograms) fields. Description as in Fig. 2.3. The number in the upper right corner of each panel shows the number of sources with magnitudes originally measured above the $5\,\sigma$ limit included in their corresponding histogram for each field.

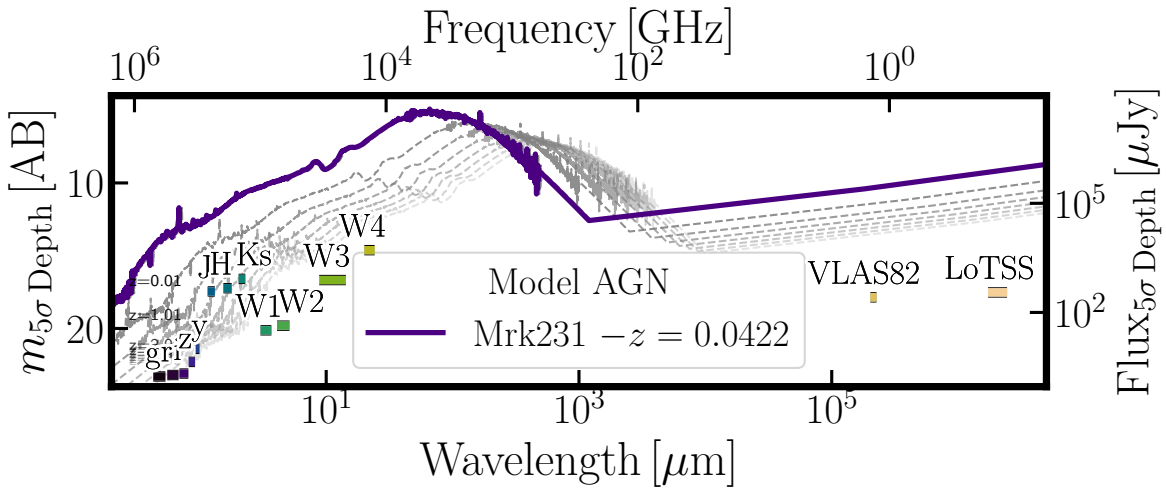


Figure 2.5: Flux and magnitude depths ($5\,\sigma$) from the surveys and bands used in this work. Limiting magnitudes and fluxes were obtained from the description of the surveys, as referenced in Sect. 2.3. In purple, rest-frame SED from Mrk231 (z = 0.0422, Brown et al. 2019) is displayed as an example AGN. Redshifted (z = 0.01, 1, 2, 3, 4, 5, 6, and 7) versions of this SED are shown in dashed grey lines.

2.5 Additional features

As mentioned in the previous sections, each magnitude corresponds to one feature that will be used in the training stages. In order to give the models more information to improve their training and to assess their results, we have generated more quantities as decribed below.

2.5.1 Engineered features

First, AGN labels and redshift information were obtained by cross-matching (with a 1″.1 search radius) the catalogue with the version `v7.4d` of the MQC (Flesch 2021), which lists information from more than 1 500 000 objects that have been classified as optical QSO, AGN, or Blazars. Sources listed in the MQC may have additional counterpart information, including radio or X-ray associations. For the purposes of this work, only sources with secure spectroscopic redshifts were used. The matching yielded 50 538 spectroscopically confirmed AGN in HETDEX and 17 743 confirmed AGN in S82. A depiction of the source density in the MQC `v7.4d` has been presented in Fig. 1.7.

Similarly, the sources in our parent catalogue (i.e. CW) were cross-matched with the SDSS data release 16 (SDSS-DR16; Ahumada et al. 2020). This cross-match was done solely to determine which sources have been spectroscopically classified as SFGs (`spClass == GALAXY`) and no magnitude measurements were extracted from it. For most of these SFGs, SDSS-DR16

lists a spectroscopic redshift value (for which no check on the quality of the measurement was made by us), which will be used in some stages of this work. In the HETDEX field, SDSS-DR16 provides 68 196 spectroscopically confirmed galaxies. In the S82 field, SDSS-DR16 classifies 4085 galaxies spectroscopically. Given that MQC has access to more AGN detection methods than SDSS (e.g. by radio and X-ray associations, and from legacy catalogues; Flesch 2015), when sources were identified as both SFGs (in SDSS-DR16) and AGN (in the MQC), a final label of AGN was given. A description of the number of sources in each field and the multi-wavelength counterparts found for them is presented in Table 2.3.

Table 2.3: Composition of initial CW catalogue and number of cross matches with additional surveys and catalogues.

Step	Survey	HETDEX	Stripe82
Base catalogue	CatWISE2020	15 136 878	3 590 306
Photometry cross-match	AllWISE	5 955 123	1 424 576
	Pan-STARRS	4 837 580	1 346 915
	2MASS	566 273	214 445
	LoTSS	187 573	...
	VLAS82	...	8747
Source identification	MQC (AGN)	50 538	17 743
	SDSS (SFGs)	68 196	4085

Then, colours from measured and imputed magnitudes were added as features. We created 66 of them, which correspond to all the available combinations of two magnitudes between the 12 selected bands, $\binom{12}{2}$. These colours are labelled in the form `X_Y` where `X` and `Y` are the respective filter labels. Depending on the stage of the training process, the number of used colours might be reduced.

An additional feature shows the number of non-radio bands in which a source has valid (i.e. non imputed) measurements. We have called it `band_num` and it has been produced by counting the number of valid values that each source showed before imputation. The values of this feature can range from 2 (source only detected in CW) up to 12 (source detected in every band of all selected surveys). This feature could be, very loosely, assimilated to the total flux a source can display. A higher `band_num` will imply that such source can be detected in more bands, suggesting that it has a higher flux (regardless of redshift). The use of features with counting or aggregation of elements in the studied dataset is well established in ML as has been presented by, for example, Zheng and Casari (2018), Duboue (2020), Sánchez-Sáez et al. (2021), and Euclid Collaboration et al. (2023b). These works have shown that, despite making

the dataset more complex, the inclusion of a new feature with the count of measurements can help the models to obtain better scores.

2.5.2 Ground truth features

To test whether or not a IR-detected source has been detected in any of the selected radio surveys we have used in Sect. 2.3, we created a feature, called `radio_detect`, which shows a Boolean flag. Its value is `True` (`1`) if we have a valid entry (i.e. a detection) in its corresponding radio catalogue. As such, this flag can only tell if a source can be detected with radio observations similar to the deepest survey from our set (i.e. LoTSS-DR1) and cannot give information of the existence, or not, of radio emission in general.

Lastly, we created an additional boolean feature, called `class`, which shows whether a source has been cross-matched with an element of the MQC or with a galaxy in SDSS-DR16. A value of zero (`0`) means that the source has been found in the SDSS-DR16 galaxy sample, a value of one (`1`) implies that the source has been identified by the MQC. Sources that have not been included neither as AGN nor as SFGs (i.e. unknown sources) have not given any value for `class` and have been left for the final prediction stages of our prediction pipeline. It is worth mentioning that a value of `0` in this flag does not mean directly that a source is not an AGN. It only implies that the studied source has not been listed in the MQC as a confirmed AGN. A list of the features created for this work and their representation in the code and in some of the figures is presented in Table 2.4.

It is worth noting that the fraction of labelled sources, both in the HETDEX and S82 fields, is very low when compared to the total size of the datasets (0.8 % for HETDEX and 0.6 % for S82). These fractions confirm that the problems exhibited in Chapter 1.2 are ubiquitous and complementary analyses tools (such as our proposed prediction pipeline) are needed to understand the nature of sources detected in existing catalogues.

2.6 Data re-scaling and normalisation

Attending to the intrinsic differences between ML algorithms, not all of them have the same performance when being trained with features that have absolute values spanning a wide range of values (i.e. several orders of magnitude). In particular, linear modelling of data might overrepresent features with larger absolute values when measuring distances between

Table 2.4: Names of columns or features used in the pipeline and what they represent.

Photometry measurements (magnitudes and fluxes)					
Code name	Feature	Code name	Feature	Code name	Feature
gmag	*g* (PS1)	ymag	*y* (PS1)	W1mproPM	*W1* (CW)
rmag	*r* (PS1)	Jmag	*J* (2M)	W1mproPM	*W2* (CW)
imag	*i* (PS1)	Hmag	*H* (2M)	W3mag	*W3* (AW)
zmag	*z* (PS1)	Kmag	*Ks* (2M)	W4mag	*W4* (AW)
Colours					
Sixty six colours from all combinations of non-radio magnitudes. (A sub-sample of them is shown.)					
g_r	(*g* - *r*) (PS1)	...	...	W2_W3	(*W2* (CW) - *W3* (AW))
g_i	(*g* - *i*) (PS1)	...	...	W2_W4	(*W2* (CW) - *W4* (AW))
g_z	(*g* - *z*) (PS1)	...	...	W3_W4	(*W3* - *W4* (AW))
Outputs of base models					
Code name	Feature	Code name	Feature	Code name	Feature
XGBoost	XGBoost	ET	Extra Trees	GBR	Gradient Boosting Regressor
CatBoost	CatBoost	GBC	Gradient Boosting Classifier		
RF	Random Forest				

Categorical flags	
Code name	Feature
band_num	Number of bands with measurements
Boolean flags	
Code name	Feature
class	AGN or SFG
band_num	Detection in, at least, one radio band.
Redshift	
Code name	Feature
z	Spectroscopic redshift

data points. For this reason, it is customary to re-scale the available non-discrete values for each feature to either be contained within the range $[\,0, 1\,]$ or to have similar distributions (e.g. Toth et al. 1993; Sola and Sevilla 1997) between features. We applied a version of the latter transformation to our features (not the targets) as to have a mean value of $\mu = 0$ and a standard deviation of $\sigma = 1$ for each feature (i.e. standardisation). Additionally, these new values were power-transformed to resemble a Gaussian distribution. This transformation helps the models avoid using the distribution of values as additional information for the training (Kamiran and Calders 2012). For this work, a Yeo-Johnson transformation (Yeo and Johnson 2000) was applied. A representation of the steps performed for the pre-processing of the data (both in HETDEX and S82 without the model training), and also described in Sects. 2.3, 2.4, and 2.5, is presented in Fig. 2.6. Categorical and Boolean features do not suffer the same problems of continuous values as they are constructed to have quantities with the same distance between them (e.g. the integer values in `band_num`). Thus, no transformation is applied to those features.

2.7 Data splitting

Given that we need to be able to compare the results from the training and the application of the ML models with values obtained independently (i.e. ground truth), we divided our dataset into labelled and unlabelled sources. Labelled sources are all elements of our catalogue that have been classified as either AGN or SFGs. Unlabelled sources are those which lack such classification and that will only be subject to the prediction of our models, not taking part in any training step. Labelled data will be used for training procedures and to properly assess the performance of the models.

Before any calculation or transformation is applied to the data from the HETDEX field, we split the labelled dataset into training, validation, calibration, and testing subsets. The early creation of these subsets helps avoid information leakage from the test subset into the models. Initially, 20 % of the dataset has been reserved as testing data. Of the remaining elements, 80 % of them (i.e. 64 % of the full sample) have been used for training, and the rest of the data has been divided equally between calibration and validation subsets (i.e. 10 % each, or 8 % of the full sample). In the case of data from S82, the only splitting is done between labelled and unlabelled sources, since the sources from S82 do not take part into the training of the models. The splitting process and the number of elements for each subset are shown in Figs. 2.7 and 2.8.

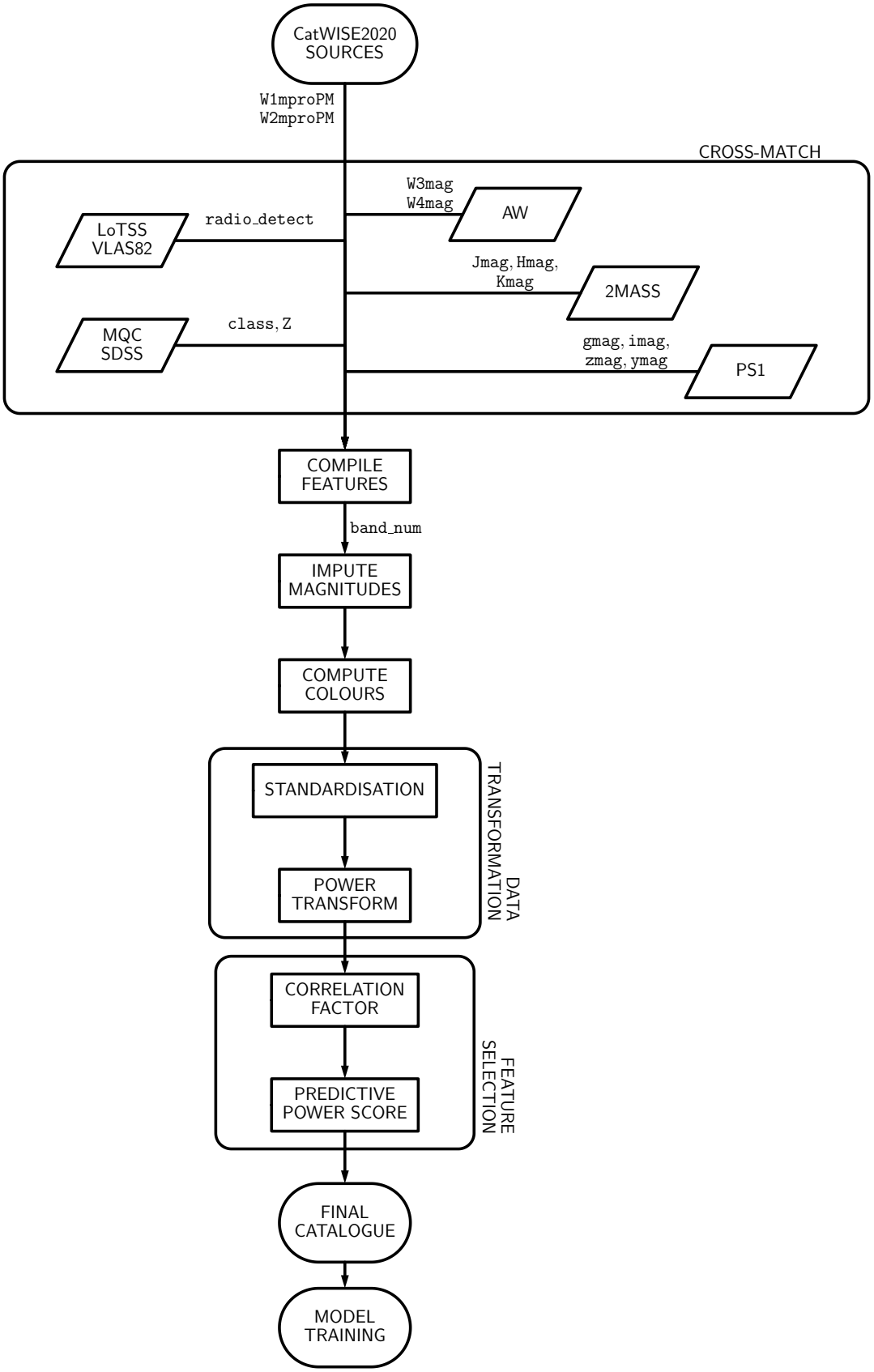


Figure 2.6: Flowchart with steps for data pre-processing. Labels in teletype font state the features produced from each step or data catalogue. Features obtained from photometric observations have been named following Table 2.1. Steps belonging to the same category have been grouped inside larger boxes.

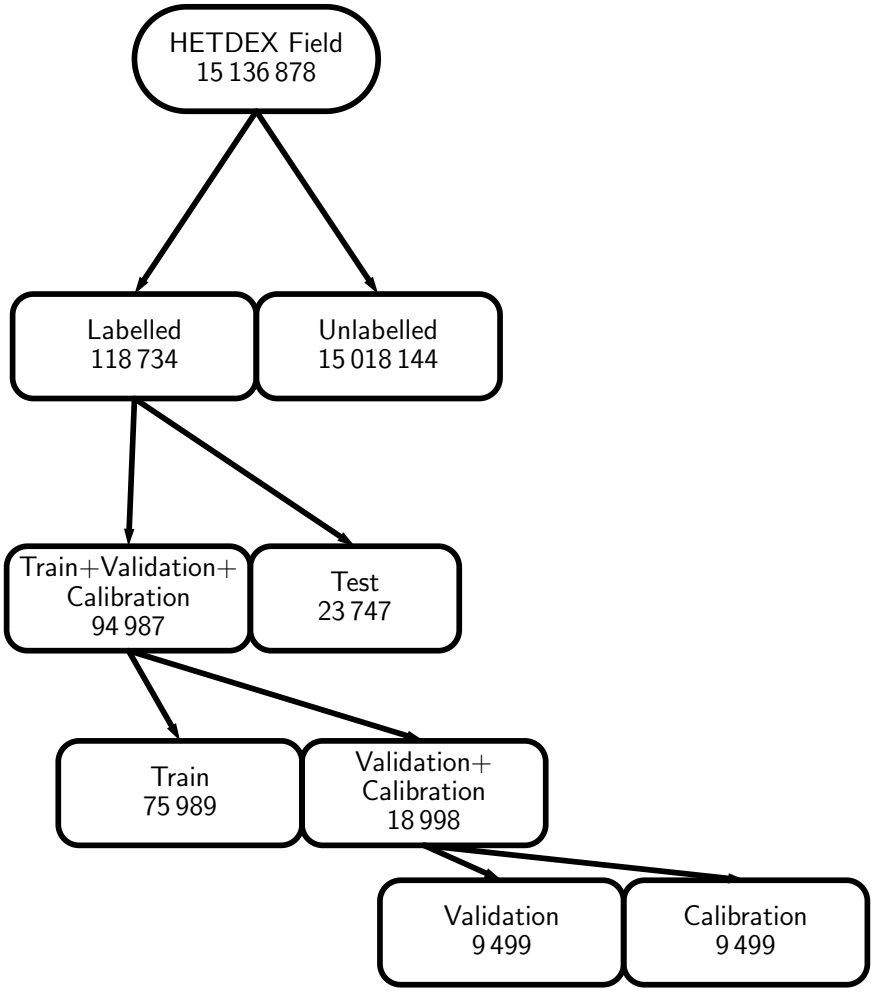


Figure 2.7: Composition of data from HETDEX (CW-detected sources) used for the different steps of this work.

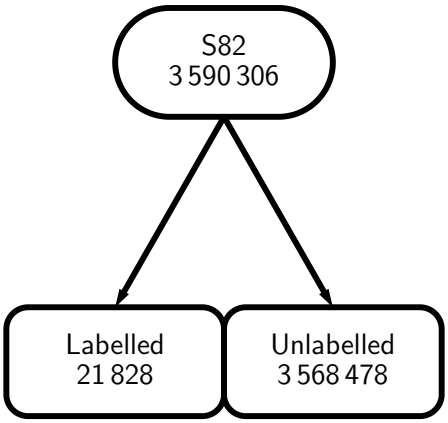


Figure 2.8: Composition of data from S82 (CW-detected sources) used for the different steps of this work.

Depending on the model to be trained, sources are selected from each of the subsets that have been already created. The training set will be used to select base and meta algorithms for each step and to optimise their parameters and hyperparameters. Parameters are those that can be derived from the direct analysis of the data, while hyperparameters are, typically, defined by the user and can be obtained from a grid of possible values. The inclusion of the validation subset helps in the parameter optimisation of the models as it is used after the selection of the stacking configuration offering an untouched dataset. The probability calibration of the trained model is performed over the calibration subset and, finally, the completed models are tested on the test subset and the labelled sources in S82. The use of these subsets will be expanded further in the text.

2.8 Computational set-up

The processes outlined in Chapter 3 (training and calibration, among other minor procedures) were run on a system consisting of a Dell PowerEdge R630 machine with an Intel® Xeon® E5-2650 v3 64 bit CPU at 2.30 GHz with 2×10 cores and 256 GiB of DDR4 ECC memory available. Our software was set in place on a Scientific Linux distribution running the kernel `3.10.0-1160.59.1.el7.x86_64`.

This page intentionally left blank.

3

Training of models and prediction of radio-AGN candidates

In this chapter, we present the results of the development of our prediction pipeline, its constituent models, and their final performance. Our pipeline consists of a series of sequential steps (as illustrated in Fig. 3.1, which expands upon the description of Fig. 1.9). We describe, for each step, the process of feature selection, algorithm optimisation, and the hyperparameter tuning for the mentioned steps. Additionally, we include a description of the computational resources used for the generation, training, and optimisation of the models. Prior to the pipeline description, we introduce the metrics used for the assessment of the models. Once the parameters and hyperparameters of the models (AGN-SFG classification, radio detection prediction, and redshift estimation) have been tuned and their scores have been calibrated, we tested their final performance through their application on the testing subset (cf. Sect. 2.7) and on data from the S82 field. The performance of the whole pipeline is also described and discussed in the final part of this chapter with an emphasis on the improvements made by the models when compared with a no-skill prediction.

3.1 Prediction metrics

Several methods exist to assess the results from machine-assisted methods. Most of them have been designed for supervised tasks. In general, predictions are compared with the original (or true) quantities and a new quantity, a metric, is derived to determine how good the prediction is (i.e. how close to the true values or classes). A set of metrics will be used to understand the reliability of the results and put them in context with results in the literature. Since our work includes the use of classification and regression models, we briefly discuss the appropriate metrics in the following sections.

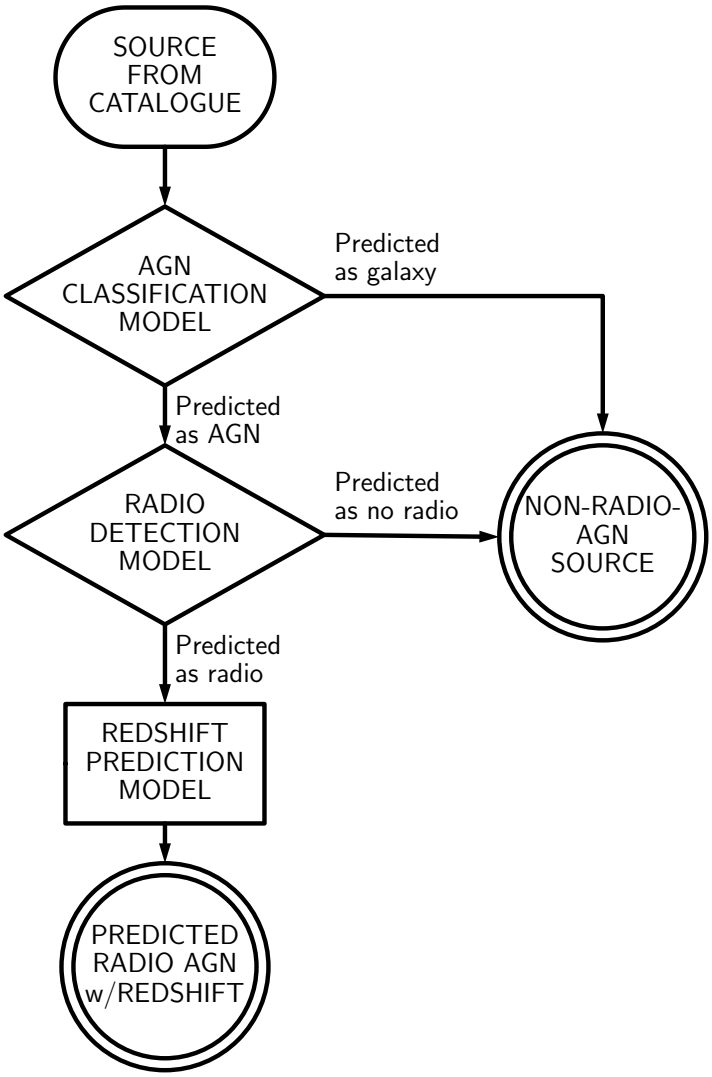


Figure 3.1: Flowchart representing the proposed prediction pipeline, with its ML-based models, used to predict the presence of radio-detected AGN and their redshift values from IR-selected sources. Diamonds represent classification models and rectangle, regression model. Double circles represent end states for the data in the pipeline.

3.1.1 Classification metrics

The main tool to assess the performance of classification methods is the confusion matrix (also known as contingency or error matrix; Mead and Meyer 1977; Hoffer and Fleming 1978; Card 1982; Congalton et al. 1983). It is a two-dimension (predicted vs true) matrix where the number of true and predicted classes are compared and results stored in cells with the rate of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). An example diagram showing the elements of a confusion matrix is shown ing Table 3.2. An ideal classifier would be represented by a diagonal matrix with no incorrectly predicted elements. As mentioned earlier in Sect. 1.4, we seek to maximise the number of positive-class sources that are recovered as such. Using the elements of the confusion matrix, this aim can be translated into the maximisation of TPs and, consequently, the minimisation of FNs.

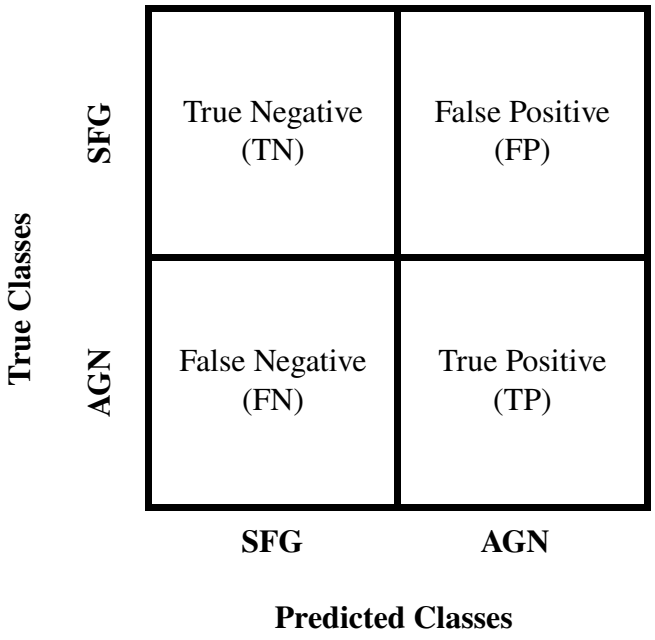


Figure 3.2: Diagram of confusion matrix for the classification of sources between AGN and SFGs. Columns represent classes predicted by the models, while rows correspond to the true categories as obtained from different methods.

From the elements of the confusion matrix, we can obtain additional metrics, such as the F1 and F_β scores (Dice 1945; Sørenson 1948; van Rijsbergen 1979), and the Matthews correlation coefficient (MCC; Yule 1912; Cramér 1946; Matthews 1975) which are better suited for unbalanced data (i.e. when the fraction of elements of one class is much higher than that of the other) as they take into account the behaviour and correlations among all elements of the confusion matrix. As such, the F1 coefficient is defined as:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \,. \tag{3.1}$$

F1 values can go from 0 (no prediction of positive instances) to 1 (perfect prediction of elements with positive labels). This definition assigns equal weight (importance) to both the number of FNs and FPs.

An extension to the F1 score, which adds a non-negative parameter, β, to increase the importance given to each one of them is the F-Score (F_β), defined as:

$$\mathrm{F}_\beta = \frac{(1+\beta^2) \times \mathrm{TP}}{(1+\beta^2) \times \mathrm{TP} + \beta^2 \times \mathrm{FN} + \mathrm{FP}} \,. \tag{3.2}$$

Using $\beta > 1$, more relevance is given to the optimisation of FNs. When $0 \leq \beta < 1$, the optimisation of FPs is more relevant. If $\beta = 1$, the initial definition of F1 is recovered. As with F1, F_β values can be in the range $[\,0, 1\,]$. Given that we seek to minimise the number of FNs detection, we adopt a conservative value of $\beta = 1.1$, giving more significance to their reduction without removing the aim for FPs. Also, this value is still close enough to $\beta = 1$, which will allow us to compare our scores to those produced in previous works.

MCC is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \,, \tag{3.3}$$

which includes also the information about the TN elements. MCC can range from -1 (total disagreement between true and predicted values) to $+1$ (perfect prediction) with 0 representing a prediction analogous to a random guess.

The Recall (also called Completeness, Sensitivity, or True Positive Rate –TPR–; Yerushalmy 1947) corresponds to the rate of relevant, or correct, elements that have been recovered by a process. Using the elements from the confusion matrix, it can be defined as:

$$\mathrm{Recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \,. \tag{3.4}$$

The true positive rate (TPR) can go from 0 to 1, with a value of 1 meaning that the model can recover all the true instances.

The last metric used is Precision (also known as Purity), which can be defined as the ratio between the number of correctly classified elements and the number of sources in the positive

class (AGN or radio detectable, depending on the step of the pipeline):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \,. \tag{3.5}$$

Precision can range from 0 to 1, where higher values show that more real positive instances of the studied set were retrieved as such by the model.

In order to establish a baseline from which the aforementioned metrics can be assessed, it is possible to obtain them in the case of a random, or no-skill prediction. Following, for instance, the derivations and notation from Poisot (2023), a no-skill confusion matrix has the form:

$$\begin{pmatrix} (1-p)^2 & p(1-p) \\ (1-p)p & p^2 \end{pmatrix}, \tag{3.6}$$

where p corresponds to the ratio between the elements of the positive class and the total number of elements involved in the prediction. From Eq. 3.6, no-skill versions of classification metrics (Eqs. 3.2–3.5) are then:

$$\text{F}_\beta^{\text{no-skill}} = p \,, \tag{3.7}$$

$$\text{MCC}^{\text{no-skill}} = 0 \,, \tag{3.8}$$

$$\text{Recall}^{\text{no-skill}} = p \,, \tag{3.9}$$

$$\text{Precision}^{\text{no-skill}} = p \,. \tag{3.10}$$

If the analysed data contains information about a particular classification, it is expected that the metrics from our prediction pipeline improve upon the no-skill values.

3.1.2 Regression metrics

The last step of our prediction pipeline, redshift prediction, corresponds to a regression task and, as such, different ways to evaluate are needed. Usually, regression tasks are assessed with the use of metrics such as mean squared error (MSE), root mean square error (RMSE), and mean absolute error (MAE). These metrics measure the deviation of the predicted value from the original quantity. If the original value is called y_{True} and its predicted version is $y_{\text{Predicted}}$, these three regression metrics can be defined as follows.

The MSE is

$$\mathrm{MSE}(y) = \frac{1}{\mathbb{d}} \sum_{i}^{\mathbb{d}} (y_{\mathrm{True}} - y_{\mathrm{Predicted}})^2 \,, \tag{3.11}$$

with $\mathbb{d}$ being the number of elements in the studied sample (i.e. its size). A direct modification of MSE appears when calculating its square root. Then, the root mean squared error is

$$\mathrm{RMSE}(y) = \sqrt{\frac{1}{\mathbb{d}} \sum_{i}^{\mathbb{d}} (y_{\mathrm{True}} - y_{\mathrm{Predicted}})^2} \,. \tag{3.12}$$

A third way to quantify the deviation of the predictions from the true values is through the MAE;

$$\mathrm{MAE}(y) = \frac{1}{\mathbb{d}} \sum_{i}^{\mathbb{d}} |y_{\mathrm{True}} - y_{\mathrm{Predicted}}| \,. \tag{3.13}$$

While MSE and RMSE are sensitive to large deviations in the predictions, the MAE has a linear behaviour with respect to the fluctuations in predicted quantities. In this way, the MAE suffers less from hallucinations (predictions that go against patterns established in the training) of models.

For the case of redshift value determination, the previous metrics are not fully able to assimilate its logarithmic behaviour. Thus, further modifications are needed in order to use suitable metrics. Namely, a factor must be included to take into account the fact that differences between low redshift values should be penalized more strongly than those at higher redshifts. It is possible to start with the difference between true (z_{True}) and predicted ($z_{\mathrm{Predicted}}$) redshift values,

$$\Delta z = z_{\mathrm{True}} - z_{\mathrm{Predicted}} \,, \tag{3.14}$$

and its normalised difference,

$$\Delta z^{\mathrm{N}} = \frac{z_{\mathrm{True}} - z_{\mathrm{Predicted}}}{1 + z_{\mathrm{True}}} \,. \tag{3.15}$$

If the comparison is made over a larger sample of elements, the bias of the redshift is used (Dahlen et al. 2013), with the median of the quantities instead of its mean to avoid the strong influence of extreme values:

$$\Delta z_{\rm Total} = \text{median}\left(z_{\rm True} - z_{\rm Predicted}\right) = \text{median}(\Delta z)\,, \tag{3.16}$$

$$\Delta z^{\rm N}_{\rm Total} = \text{median}\left(\frac{z_{\rm True} - z_{\rm Predicted}}{1 + z_{\rm True}}\right) = \text{median}(\Delta z^{\rm N})\,. \tag{3.17}$$

Using the definitions in Eqs. 3.14 and 3.15, four additional metrics can be calculated. These are the median absolute deviation (MAD) and the normalised median absolute deviation (NMAD, $\sigma_{\rm NMAD}$; Hoaglin et al. 1983; Ilbert et al. 2009), which have little sensitivity to outliers, as $\Delta z_{\rm Total}$ and $\Delta z^{\rm N}_{\rm Total}$ do (Salvato et al. 2011). Also, the standard deviation of the predictions, σ_z, and its normalised version, $\sigma_z^{\rm N}$ are typically used. They are defined as:

$$\sigma_{\rm MAD} = 1.48 \times \text{median}\left(|\Delta z|\right)\,, \tag{3.18}$$

$$\sigma_{\rm NMAD} = 1.48 \times \text{median}\left(\left|\Delta z^{\rm N}\right|\right)\,, \tag{3.19}$$

$$\sigma_z = \sqrt{\frac{1}{\mathbb{d}}\sum_i^{\mathbb{d}} (\Delta z)^2}\,, \tag{3.20}$$

$$\sigma_z^{\rm N} = \sqrt{\frac{1}{\mathbb{d}}\sum_i^{\mathbb{d}} \left(\Delta z^{\rm N}\right)^2}\,. \tag{3.21}$$

Additionally, the outlier fraction (η, as used in Dahlen et al. 2013; Lima et al. 2022) is considered, which is defined as the fraction sources with a predicted redshift difference ($\left|\Delta z^{\rm N}\right|$, Eq. 3.15) larger than a previously set value. Taking the results from Ilbert et al. (2009) and Hildebrandt et al. (2010), we have selected this threshold to be 0.15, leaving the definition of the outlier fraction as:

$$\eta = \frac{\#\left(\left|\Delta z^{\rm N}\right| > 0.15\right)}{\mathbb{d}}\,, \tag{3.22}$$

where # symbolises the number of sources fulfilling the described relation, and $\mathbb{d}$ corresponds to the size of the selected sample.

Most metrics presented in this section have been specially devised or adapted for the study of redshift values and their particular properties. Their use in ML algorithms can be, thus, considered a rudimental application of physics-informed ML (cf. Sect. 1.3).

3.2 Classification thresholds

Metrics presented in Sect. 3.1.1 work with a prediction of the status (class) of an element. However, most classifiers deliver a score rather than a definite class prediction. Scores, in the range $[0, 1]$ need to be translated into positive or negative classes. A threshold is defined to separate both states. By default, these models set a threshold at 0.5 in score (which we will call a naive threshold) but, in principle and given the characteristics of the problem, a different optimal threshold might be needed.

In our case, we want to optimise (increase) the number of recovered elements in each model (i.e. AGN or radio-detectable sources). This maximisation corresponds to obtaining thresholds that optimise the recall (Eq. 3.4). This can be done by decreasing the threshold by which a source is classified as a positive instances. Setting this threshold to its minimum, 0.0, would increase the recall (completeness). But every source would be predicted to be an AGN or detected on the radio regardless of their properties, defeating the purpose of the prediction pipeline. Thus, a different approach must be taken.

An optimised threshold can be obtained through the use of the statistical tool called precision-recall (PR) curve. Thresholds derived from the PR curves will be labelled as PR. PR curves can help to understand the behaviour of a classifier as a function of its threshold. In well-behaved models, both quantities, precision (Eq. 3.5) and recall, show an inverse correlation, and both depend on the selected threshold. Thus, they can be used to retrieve the score value for which both quantities are balanced. This optimisation is done by finding the threshold that maximises the F_β score (Eq. 3.2). This operation can be performed over the union of training and validation sets, which have been used to create and train each model.

3.3 Classification calibration

Classifiers deliver scores in the range $[\,0, 1\,]$, which could be, roughly, associated to the probability of a studied source being part of the relevant class (in our work, AGN or radio detectable). The classifier uses a threshold above which, any predicted element would be considered a positive instance.

With the exception of few algorithms (including the family of logistic regressions), scores from classifiers cannot be directly used as probabilities (Caruana and Niculescu-Mizil 2006).

As a consequence of this inability, such values cannot be compared from one type of model to some other and can not be combined to obtain a joint score. Therefore, in order to retrieve joint scores and treat them as probabilities, scores (and, by extension, the classifiers) need to be calibrated (Gilda et al. 2021). This calibration means that, when taking all predictions with a probability P of being of a class $\mathbb{C}$ (i.e. $P(\mathbb{C})$), a fraction P of them really belong to that class (e.g. Lichtenstein et al. 1982; Silva Filho et al. 2023).

Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation. Beta transformation functions have the general form

$$\mu_{beta}(S; a, b, c) = \frac{1}{1 + \frac{1}{\left(e^{c}\frac{S^{a}}{(1-S)^{b}}\right)}}, \tag{3.23}$$

with s being the score from the classifier and a, b, c, free parameters to be optimised. It allows one to re-distribute the scores of a classifier allowing them to get closer to the definition of probability (Kull et al. 2017a,b). Calibration steps in our workflow have been applied using the `Python` package `betacal`[1]. In the case of the radio detection model, the new scores have a wider range than the original, uncalibrated scores.

Calibration (or reliability) plots (Murphy and Winkler 1977; Wilks 1990; Niculescu-Mizil and Caruana 2005b) show how well calibrated the predicted scores of a classifier are by displaying the fraction of sources that are part of a given class as a function of the predicted probability. A perfectly calibrated classifier would have all its prediction lying in the $x = y$ line (Niculescu-Mizil and Caruana 2005a). The magnitude of the deviations from that line give information of the miscalibration a model has (see, for instance, Bröcker and Smith 2007; Van Calster et al. 2019).

3.3.1 Calibration scores

One of the most used analytical metrics to assess calibration of a model is the Brier score (BS; Brier 1950). It measures the mean square difference between the predicted probability of an element and its true class. If the total number of elements in the studied sample is $\mathbb{d}$, the BS can be written (for binary classification problems, as the ones studied in this work) as:

[1] `https://betacal.github.io`

$$\mathrm{BS} = \frac{1}{\mathbb{d}} \sum_{i}^{\mathbb{d}} (\mathbb{C} - \mathrm{class})^2 \,, \tag{3.24}$$

where $\mathbb{C}$ is the predicted class and `class` the true class of each of the elements in the sample (0 or 1). The BS can range between 0 and 1 with 0 representing a model that is completely reliable in its predictions. Additionally, the BS can be used to compare the reliability (or calibration) between a model and a reference using the Brier skill score (BSS; Glahn and Jorgensen 1970):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \,. \tag{3.25}$$

In our case, $\mathrm{BS}_{\mathrm{ref}}$ corresponds to the value calculated from the uncalibrated model. The BSS can take values between -1 and 1. The closer the BSS gets to 1, the more reliable the analysed model is. These values include the case where $\mathrm{BSS} \approx 0$, in which both models perform similarly in terms of calibration.

For our pipeline, after a model has been fully trained, a calibrated version of their scores will be obtained. With both of them, the BSS will be calculated and, if it is not considerably lower than 0, that calibrated transformation will be used as the final scores from the prediction.

3.4 Feature selection

ML algorithms, as most data analysis tools, require execution times which increase with the size of the datasets. In order to reduce training times without losing relevant information for the model, the most important features were selected at each step through a process called feature selection (e.g. Blum and Langley 1997; Kohavi, Ron and John 1997; Guyon and Elisseeff 2003). Feature selection can also help avoiding the inclusion of data that might add noise to the model predictions.

For each model, the process of feature selection begins with 79 base features (Table 2.4) and three targets (`class`, `LOFAR_detect`, and `Z`). Feature selection is run, independently, for each trained model (i.e. AGN-SFG classification, radio detection, and redshift predictions), delivering three different sets of features. To avoid redundancy, the process starts discarding features that have a high correlation with another property of the dataset. For discarding features, we calculated Pearson's correlation matrix (built with the Pearson's correlation factors, ρ, between features; Bravais 1844; Galton 1886; Pearson and Galton 1895) for the full train+validation

dataset only and selected the pairs of features that showed a correlation factor higher than $\rho = 0.75$, in absolute values. A value of $\rho = 0.75$ is a compromise between very stringent thresholds (e.g. $\rho = 0.5$) and more relaxed values (e.g. $\rho \approx 0.9$) (for an explanation on how to consider different correlation values, see, for instance Ratner 2009). From each pair, we discarded the feature with the lowest relative standard deviation (RSD; Johnson and Leone 1964), which is defined as the ratio between the standard deviation of a set and its mean value. A feature which covers a small portion of its probable values (i.e. low coverage of parameter space, and lower RSD) will give less information to a model than one with largely spread values. Thus, its elimination might not have a large impact the final model.

For our analysis, we opted for traditional ML models over deep learning (DL) techniques, which use complex artificial neural networks and representation learning for prediction tasks (e.g. LeCun et al. 1998, 2015). While DL excels in various tasks, traditional methods often prove more efficient for tabular data classification and regression, particularly tree-based models (e.g. Borisov et al. 2022; Shwartz-Ziv and Armon 2022).

For the AGN-SFG classifier, feature selection was applied in the train+validation subset with 85 488 confirmed elements (galaxies from SDSS-DR16 and AGN from MQC, i.e. `class == 0` or `class == 1`). After the selection procedure described in Sect. 3.4, 18 features were selected for training[2]: `band_num`, `W4mag`, `g_r`, `r_i`, `r_J`, `i_z`, `i_y`, `z_y`, `z_W2`, `y_J`, `y_W1`, `y_W2`, `J_H`, `H_K`, `H_W3`, `W1_W2`, `W1_W3`, and `W3_W4`. The target feature is `class`. In the case of the radio detection classifier, feature selection was applied to the train+validation subset, with 36 387 confirmed AGN. The target feature is `LOFAR_detect` and the base of 17 selected features are: `band_num`, `W4mag`, `g_r`, `g_i`, `r_i`, `r_z`, `i_z`, `z_y`, `z_W1`, `y_J`, `y_W1`, `J_H`, `H_K`, `K_W3`, `K_W4`, `W1_W2`, and `W2_W3`.

Finally, feature selection (cf. Sect. 3.4) was applied to the train+validation subset for the photometric redshift regressor, with 4612 sources, leading to the selection of 17 features. The target feature is `Z` (redshift) and the selected base features are `band_num`, `W4mag`, `g_r`, `g_W3`, `r_i`, `r_z`, `i_z`, `i_y`, `z_y`, `y_J`, `y_W1`, `J_H`, `H_K`, `K_W3`, `K_W4`, `W1_W2`, and `W2_W3`.

Is is important to highlight that, for all three models, a similar number of features was selected. Additionally, the base of adopted features contains quantities that are analogous among them. This fact might suggest that our full dataset might be reduced to a minimum number of features which carry the same information as the full sample of 77 quantities, independent of

[2]The order in which the features are presented here does not convey any preference from the models.

the studied target.

3.5 Model stacking

Base and meta learners (cf. Sect. 1.3.2) are selected based upon the metrics described in Sect. 3.1. We trained five algorithms with the training subset and calculated the metrics for all of them using a 10-fold cross-validation (CV) approach (e.g. Stone 1974; Allen 1974) over the same training subset, which separates the data into ten sets that are held out consecutively while models are trained on the remaining entries. For each metric, the learners are ranked (from 1 to 5, with 1 being the best possible value) and a mean value of the property or class to predict has been obtained from them. Out of the analysed algorithms, the one with the best overall performance (i.e. best mean rank) is selected to be the meta learner while the remaining four are used as base learners.

For the AGN-SFG and radio detection classification problems, we tested the following classification algorithms: random forests (`RFs`; Breiman 2001), gradient boosting classifier (`GBC`; Friedman 2001), Extra Trees (`ET`; Geurts et al. 2006), extreme gradient boosting (`XGBoost, v1.5.1`; Chen and Guestrin 2016), and Category Boosting (`CatBoost, v1.0.5`; Prokhorenkova et al. 2018; Dorogush et al. 2018). For the redshift prediction problem, we tested five regressors as well: `RF`, `ET`, `XGBoost`, `CatBoost`, and gradient boosting regressor (`GBR`; Friedman 2001). We have used the `Python` implementations of these algorithms and, in particular for `RF`, `ET`, `GBC`, and `GBR`, the versions offered by the package `scikit-learn`[3] (`v0.23.2`; Pedregosa et al. 2011). These algorithms were selected given that they offer tools to interpret the global and local influence of the input features in the training and predictions (cf. Sect. 1.3.3).

All the algorithms selected for this work fall into the broad family of tree-based models. Additionally, forest models (`RF` and `ET`) rely on a collection of decision trees to, after applying a majority vote, predict either a class or a continuum value. Each of these decision trees uses a different, randomly-selected subset of features to make a decision on the training set (Breiman 2001). Opposite to forests, gradient boosting models (`GBC`, `GBR`, `XGBoost`, and `CatBoost`) apply decision trees sequentially to improve the quality of the previous predictions (Friedman 2001, 2002).

[3] `https://scikit-learn.org`

3.6 Model training

The procedure described in Sect. 3.5 includes an initial fit of the selected algorithms to the training data (including the selected features) to optimise their parameters. The stacking step includes a new optimisation of the parameters of the meta-learner using 10-fold CV on the training data with the addition of the output from the base learners, which are treated as regular features (see last section of Table 2.4). Then, following Michailidis (2017), the hyperparameters of the stacked models are optimised over the training subset (a more detailed description of this step is presented in Sect. 3.6.1).

The final step involves a last parameter fitting instance but using, this time, the combined train+validation subset, which includes the output of the base algorithms, to ensure wider coverage of the parameter space and better-performing models. Consequently, only the testing set is available for assessing the quality of the predictions made by the models.

The results of model testing for the AGN-SFG classification are reported in Table 3.1. The `CatBoost` algorithm provides the best metric values (best mean rank) and is therefore selected as the meta-model. `XGBoost`, `RF`, `ET`, and `GBC` were used as base learners.

Table 3.1: Performance rating for models in the AGN-SFG classification using metrics defined in main text

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
`CatBoost`	95.70 ± 0.28	92.46 ± 0.48	95.45 ± 0.32	95.91 ± 0.37	1.00
`XGBoost`	95.67 ± 0.27	92.40 ± 0.48	95.41 ± 0.39	95.88 ± 0.34	2.00
`RF`	95.52 ± 0.36	92.14 ± 0.63	95.28 ± 0.46	95.71 ± 0.40	3.00
`ET`	95.40 ± 0.40	91.94 ± 0.69	95.13 ± 0.43	95.63 ± 0.47	4.00
`GBC`	95.26 ± 0.31	91.66 ± 0.54	94.82 ± 0.41	95.63 ± 0.35	5.00

[a] Metrics obtained using the default probability threshold of 0.5.
[b] Algorithms are sorted by decreasing mean rank values.
[c] For display purposes, all metrics have been multiplied by 100.
[d] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 3.5)

Training of the radio detection model was applied only to sources confirmed to be AGN (`class == 1`). The performance of the tested algorithms is shown in Table 3.2. In this case, `GBC` shows the best mean rank and is, therefore, selected as the meta-learner while the remaining four models (`XGBoost`, `CatBoost`, `RF`, and `ET`) become the base-learners.

The redshift value prediction model was applied to sources confirmed to be radio-detected

Table 3.2: Performance rating for models in the radio detection classification using metrics defined in main text

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
GBC	29.60 ± 1.66	31.31 ± 1.93	62.55 ± 3.95	20.66 ± 1.40	1.75
CatBoost	29.57 ± 1.62	30.56 ± 1.71	60.10 ± 2.85	20.85 ± 1.36	2.25
XGBoost	29.98 ± 2.29	29.81 ± 2.17	56.74 ± 2.93	21.61 ± 2.00	2.75
RF	29.16 ± 2.47	30.26 ± 2.65	60.03 ± 3.73	20.48 ± 1.96	3.75
ET	28.40 ± 1.27	29.73 ± 1.47	60.06 ± 2.85	19.80 ± 1.05	4.50

[a] Values and uncertainties as in Table 3.1.
[b] Algorithms sorted by decreasing mean rank values.

AGN (i.e. `class == 1` and `radio_detect == 1`). The tested algorithms performed as shown in Table 3.3. Based on their mean rank values, `RF`, `CatBoost`, `XGBoost`, and `GBR` were selected as base learners and `ET` was used as meta-learner.

Table 3.3: Performance rating of base models for redshift value prediction using metrics defined in main text

Model	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^N (×100)	η (×100)	Rank
ET	18.53 ± 1.03	8.42 ± 0.43	41.12 ± 4.16	18.65 ± 2.26	19.24 ± 1.16	1.8
RF	17.88 ± 1.41	7.95 ± 0.50	42.02 ± 5.28	19.38 ± 2.44	19.51 ± 1.98	2.0
CatBoost	21.71 ± 1.38	10.08 ± 0.47	40.35 ± 3.03	18.52 ± 1.39	21.93 ± 1.55	2.2
XGBoost	22.89 ± 1.05	10.84 ± 0.78	43.14 ± 3.99	19.62 ± 1.78	24.15 ± 1.84	4.0
GBR	27.73 ± 1.57	12.72 ± 0.74	44.82 ± 3.80	20.41 ± 1.67	28.67 ± 2.25	5.0

[a] Algorithms sorted by decreasing mean rank values.
[b] Uncertainties as in Table 3.1.

Is is worth noting that, while the use of the mean rank is helpful to select a meta learner, the proper differences in metric values between models are small. These similarities might imply that most algorithms (at least classifiers) can extract the same level of information from the data. For this reason, the use of ensemble learning is justified to help the algorithms extract even more information than what they can retrieve on their own.

3.6.1 Hyperparameters optimisation

After the selection of the meta learners of each prediction stage of our pipeline, the predicted values (scores for classifiers and redshift for the regressor) are incorporated to the feature set as new quantities to learn from. Thus, and as shown in Table 2.4, four new features are added per training instance.

In Table 3.4, we present the optimised hyperparameters from our meta-learners. For all three instances of modelling (AGN-SFG, radio detection, and redshift), hyperparameters were optimised using the `SkoptSearch` algorithm embedded in the package `tune-sklearn`[4] (`v0.4.1`; Head et al. 2021), which implements a Bayesian search in the hyperparameter space.

Table 3.4: Hyperparameters values for meta-learners in modified pipeline after tuning.

AGN-SFG model (`CatBoost`)			
Parameter	Value	Parameter	Value
`learning_rate`	0.0075	`random_strength`	0.1
`depth`	6	`l2_leaf_reg`	10
Radio detection model (`GradientBoosting`)			
Parameter	Value	Parameter	Value
`n_estimators`	187	`min_samples_leaf`	2
`learning_rate`	0.0560	`max_depth`	9
`subsample`	0.3387	`max_features`	0.5248
`min_samples_split`	5		
Redshift prediction model (`ET`)			
Parameter	Value	Parameter	Value
`n_estimators`	100	`criterion`	`mae`
`max_depth`	`None`	`min_samples_split`	2
`max_features`	`auto`	`min_samples_leaf`	1
`bootstrap`	`False`		

[a] This table shows the parameters which were subject to tuning.
[b] Remaining hyperparameters used their default values as defined by their developers.

3.6.2 Calibration of models

It has been shown by, for instance Niculescu-Mizil and Caruana (2005a), that gradient boosting models output poorly calibrated posterior probabilities. Thus, it is expected that our classification models (with boosting-based meta learners) will require probability calibration. In Fig. 3.3, we present the reliability curves for the uncalibrated classifiers (see Sect. 3.3). It can be seen that the scores for the AGN-SFG classifier are found clustered in a small range around 0.5. This behaviour might indicate an issue with the predictions. But, as presented in Table 3.1, all models used in the training show very high (i.e. satisfactory) metric values. This apparent contradiction might be explained by the fact that the sample used for training is

[4] https://github.com/ray-project/tune-sklearn

highly unbalanced, with most of the sources being labelled as galaxies. Thus, there might be a fraction of elements of both classes that share a significant region of the parameter space. When ML algorithms try to classify elements under these two circumstances, they tend to deliver predictions with very low certainties (see, for instance, Vuttipittayamongkol et al. 2021; Santos et al. 2022). This issue can be solved, among other techniques, with the use of probability calibration, which has been implemented in our pipeline.

The previously presented problem does not seem the exist, in the same fashion, in the classification of radio-detectable AGN. There, the distribution of prediction scores ranges from 0.0 up to ~ 0.8. In this case, and given the conditions of the problem of finding indicators of the detection of radio sources from optical and infrared measurements, the source of a lack of scores close to 1.0 can be related to the imposibility of the models of finding stronger connections between all measurements rather than, for example, problems with the balance of datasets.

In Fig. 3.4, we present the reliability curves for the calibrated versions of the classifiers. For the AGN-SFG classifier, the improvement is remarkable. Now, predicted probabilities are distributed in the range $[\,0, 1\,]$ and they follow closely the line of perfect calibration. In the case of the radio detectability prediction, the improvement is milder, with the new probabilities getting closer to the line of perfect probability calibration. Nevertheless, the new probabilities maintain the same distribution as the original scores (between 0.0 and ~ 0.8). This result implies that probability calibration cannot be used to solve limitations with the extraction of information from the available features.

From a numerical point of view, when obtaining the BSS values for both classification, the AGN-SFG classifier has a score of BSS $= -0.002$, demonstrating that no major changes were applied to the intrinsic distribution of scores. For the radio detection classifier, the score is BSS $= -0.434$. Even though the BSS value is slightly negative for the AGN-SFG classifier, we keep it since its range of values now can be compared and combined with additional probabilities. In the case of the radio detection classifier, the BSS shows a degradation of the calibration, but the calibrated model was kept since it provides, overall, better values for the remaining metrics. Additionally, and as mentioned previously, the use of calibrated probabilities allows one to combine them through consecutive models, which is our goal with the prediction of radio-detectable AGN.

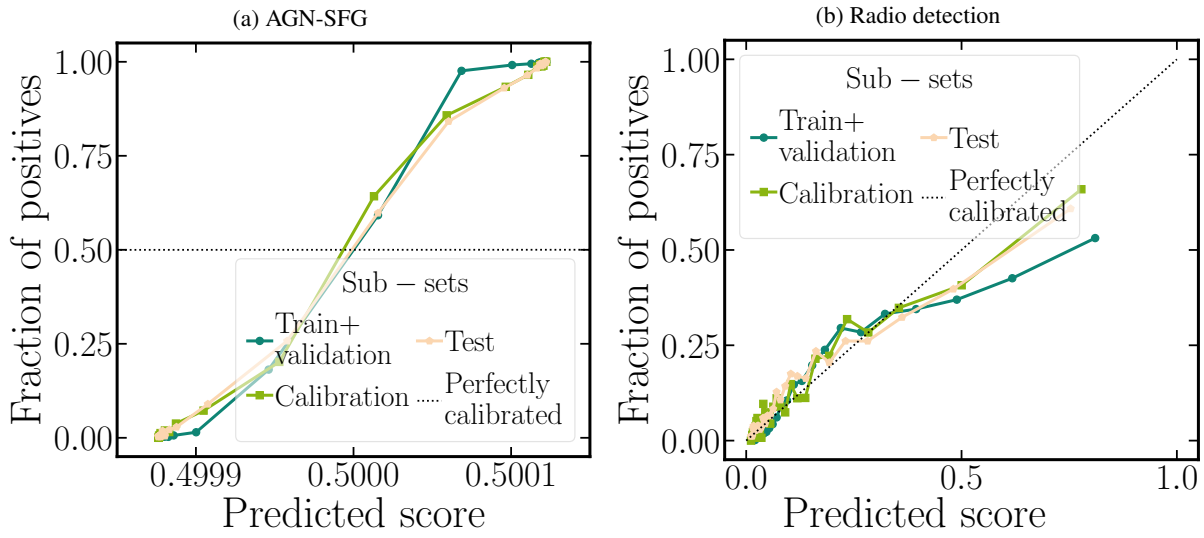


Figure 3.3: Reliability curves for uncalibrated classifiers. Each solid line represents the calibration curve for each subset in HETDEX field colour-coded following description of legend. Data has been binned and each bin (represented by the points) has the same number of elements per curve. Dashed, black line represents a perfectly calibrated model. (a) AGN-SFG classification model. (b) Radio detection model.

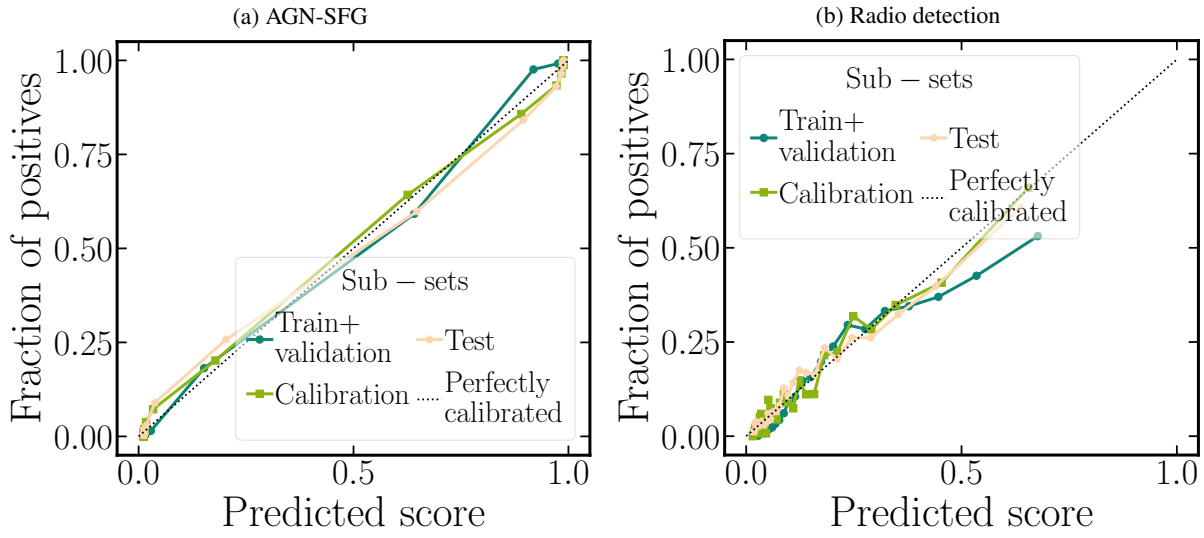


Figure 3.4: Reliability curves for calibrated classifiers. (a) AGN-SFG classification model. (b) Radio detection model. Details as in Fig. 3.3.

3.6.3 Threshold selection

PR curves (see Sect. 3.2) for all subsets used in our classification models are shown in Fig. 3.5. In the case of AGN-SFG classification, it can be seen (Fig. 3.5a) that the PR curve does not present any anormality (i.e. the curves are smooth and can reach close to the upper-right side of the plot). From the optimisation (i.e. maximisation) of the F_β score, the optimal threshold for the calibrated meta model is 0.34895. This value was used for the AGN-SFG model throughout this work. It might also be seen that the behaviour of the PR curve is consistent among subsets, an indication that the model has been properly trained and, probably, very low levels of over-fitting are present.

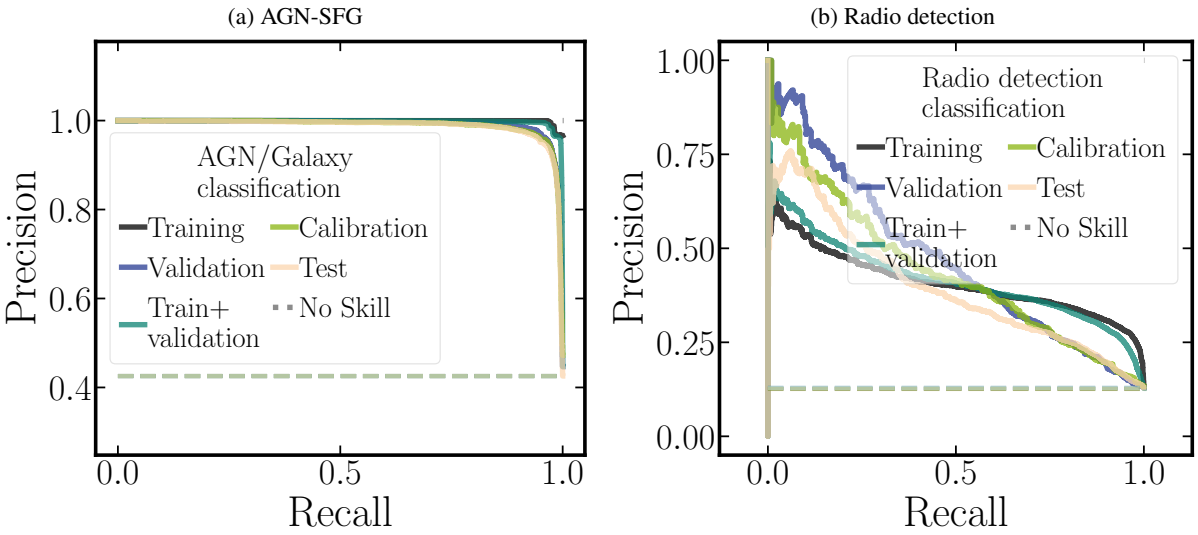


Figure 3.5: Precision-Recall curves for the (a) AGN-SFG and (b) radio detection classification models. Each solid line represents the PR curve for each subset in HETDEX field colour-coded following description of legend.

On the other side, the PR curves for the calibrated radio-detectability meta model (Fig. 3.5b) present a diffent behaviour, with noticeable variation among subsets. Such behaviour might be the expression of a lack of enough retrieval of information from the training set, making the model unsuited for a successful application in a diferent dataset. As most sources in this work have been classified as AGN or SFG by means of optical and IR information, our model does not have full access to data that can help with radio classification (see Sect. 1.4). Additionally, each curve shows high levels of irregularities, highlighting the possibility that the model is very sensitivity to, even small, changes of threshold. As mentioned in Sect. 3.2, PR curves underline the relation between precision and recall in a model. Figure 3.5b hints a highly extreme relation between both metrics, where no similar combination of values can be

found. Finally, from the maximisation of the F_β score, the optimal threshold for this model is found to be 0.20460.

3.6.4 Use of computational resources

We run our training and calibration steps with the resources described in Sect. 2.8. For a rough estimate of the CPU usage, our training scheme used 12 cores for slightly more than an hour in each step. Such values imply that these 12 CPU cores were used for more than three hours, meaning $\sim 40\,\mathrm{CPU-h}$ of computing time for training.

3.7 Prediction of radio-AGN candidates

Now, we focus on the application of the already trained, optimised, and calibrated models. We tested them on the HETDEX testing subset and on the sources in the S82 field. For the analysis of the predictions and their quality, we applied the models (independently and as part of the full prediction pipeline) in the HETDEX test subset and in the labelled sources in the S82 field. After its assessment, the prediction pipeline was implemented in the unlabelled sources in both studied fields (HETDEX and S82) for the generation of new radio-AGN candidates.

3.7.1 AGN-SFG classification

The results of the application of the stacked and calibrated model for the testing subset and the labelled sources in S82 are presented in the two firts blocks of Table 3.5 (HETDEX-test and S82-label). The metrics are shown for the use of two different thresholds, the naive value of 0.5 and the PR-derived value of 0.34895, which maximises the F_β score. The confusion matrix of the predictions (calculated on the testing dataset) is shown in Fig. 3.6.

Overall, the model is able to separate AGN from SFGs with a very high success rate (recall $\gtrsim 93\,\%$) regardless of the selected threshold. For the case of the test set, the MCC scores for the two analysed thresholds are in similar levels with each other and with the training levels (see Table 3.1). That can be taken as an indication of a good training process, in which no substantial over-fitting can be detected, in line with the results presented in Sect. 3.6.3 for the PR analysis. In general, values using the naive threshold are, when considering the uncertainty values, compatible with those obtained with the PR threshold. Since in both cases, the scores

Table 3.5: Resulting metrics of AGN-SFG classification model for the test subset and the labelled sources in S82 using two different threshold values. HETDEX and S82 pipeline results are described in Sect. 3.7.4.

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	95.37 ± 0.36	91.81 ± 0.67	97.47 ± 0.69	95.89 ± 2.27
	PR	95.42 ± 0.38	91.85 ± 0.70	94.49 ± 0.65	96.21 ± 0.43
S82-label	Naive	94.15 ± 0.44	70.54 ± 2.02	95.16 ± 0.41	93.33 ± 0.66
	PR	94.37 ± 0.36	70.67 ± 1.72	94.81 ± 0.40	94.01 ± 0.59
HETDEX-pipe	Naive	95.37 ± 0.36	91.81 ± 0.67	97.47 ± 0.69	95.89 ± 2.27
	PR	95.42 ± 0.38	91.85 ± 0.70	94.49 ± 0.65	96.21 ± 0.43
S82-pipe	Naive	94.15 ± 0.44	70.54 ± 2.02	95.16 ± 0.41	93.33 ± 0.66
	PR	94.37 ± 0.36	70.67 ± 1.72	94.81 ± 0.40	94.01 ± 0.59

[a] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 3.5)

are fairly high, it is possible that the model is, already, extracting most of the information from the dataset and, regardless of the treatment applied to the scores, predictions will mostly be correct. Thus, it is possible that the model does not need a very large effort (i.e. a large internal decision structure) to distinguish between classes.

One important caveat to the good results presented in Table 3.5 is that of the relatively lower MCC values obtained for the S82 field. Considering that, among the metrics used in this work, is the only one that incorporates the number of TNs (cf. Sect. 3.1.1), the MCC reflects the imbalance between classes in a dataset. As discussed in Sect. 2.5, the HETDEX and S82 fields have different fraction of sources (AGN and SFGs) among them and, thus, different values of MCC are expected, reflecting the difficulties that a model might have when applied to moderately different datasets.

A closer inspection to the confusion matrix in Fig. 3.6 shows that close to 4 % of the AGN from the MQC were discarded by our model. And less than 6 % of the predicted AGN are not labelled as such by the MQC. Additionally, it is possible to see the level of imbalance present in this dataset, where almost 60 % of the sources are, originally, labelled as SFGs. Despite this disparity, our model achieves high-quality metric values in both HETDEX and S82 fields, which creates a well-predicted dataset to be fed into the following step of the prediction pipeline.

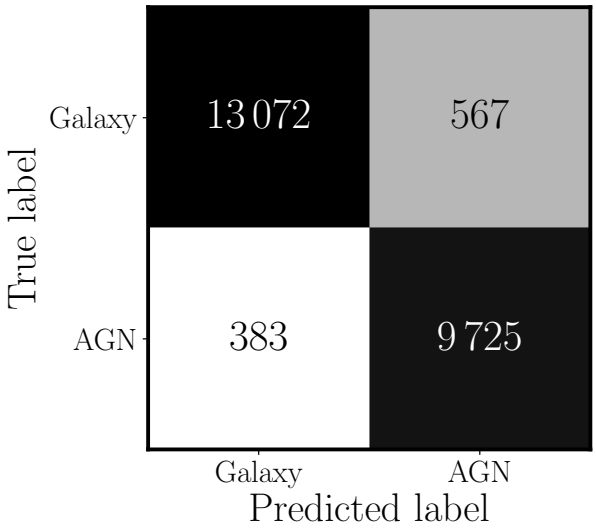


Figure 3.6: Confusion matrix from the results of application of AGN-SFG classification model to the HETDEX test subset. Rows represent the separation of sources according to their true (or original) classification. Columns divide the sample in their predicted classes. A description of confusion matrices, in general, is offered in Sect. 3.1.1 and, specifically, in Fig. 3.2.

3.7.2 Radio detection classification

The application of the stacked model for the prediction of the radio detection in the testing subset is summarised in the two first blocks of Table 3.6 (labelled as HETDEX-test and S82-label). Contrary to the results from the previous step of the prediction pipeline, the metrics for the radio detection prediction are somewhat lower. This difference might be a demonstration that the model is not able to extract an amount of information that is enough to discriminate between both classes. Additionally, it is important to stress the differences in depth between LoTSS-DR1 and VLAS82. As shown in Sect. 2.4, VLAS82 measurements are shallower than those from LoTSS-DR1, which might interfere with the correct assessment of predictions, as the number of FP and TP sources might be increased.

Another difference with the results from the AGN-SFG classifier is that there is a noticeable disparity between the metrics obtained from the use of the naive threshold and those from the application of the PR-based threshold. Even when factoring in the uncertainties, metrics in both cases remain clearly different. This difference emphasises the importance of the use of the PR curve for the extraction of optimised classification thresholds (see Fig. 3.5b), even though it is, also, an expression of the instability of such curve for this classifier (which, as shown in Sect. 3.6.3, is very sensitive to changes in threshold).

Similarly, the confusion matrix derived from the prediction results over the test sample is shown in Fig. 3.7. From it, the reasons for the low metrics can be seen more patently, with rather

Table 3.6: Resulting metrics of the radio detection model on the test subset and the labelled sources in S82 using two different threshold values. HETDEX and S82 pipeline results shown as part of the discussion in Sect. 3.7.4

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	24.87 ± 2.94	27.36 ± 3.46	60.61 ± 8.18	16.72 ± 2.31
	PR	42.88 ± 2.93	32.47 ± 3.49	35.28 ± 2.74	52.16 ± 3.59
S82-label	Naive	27.15 ± 2.28	23.36 ± 2.27	25.72 ± 1.91	28.47 ± 3.24
	PR	21.62 ± 1.20	19.37 ± 1.64	12.29 ± 0.73	58.16 ± 3.06
HETDEX-pipe	Naive	24.37 ± 3.53	26.93 ± 4.18	59.36 ± 7.17	16.38 ± 2.63
	PR	41.57 ± 4.16	31.67 ± 4.81	34.65 ± 3.24	49.80 ± 5.85
S82-pipe	Naive	26.52 ± 5.44	23.29 ± 5.73	25.71 ± 5.89	27.72 ± 5.21
	PR	20.19 ± 2.84	18.40 ± 4.07	11.45 ± 1.58	54.78 ± 8.44

[a] Values and uncertainties as in Table 3.5.

high values for FN and FP sources. Again, such values might find their roots in the possible inability of this model to extract all the available information to make a higher quality decision about the radio detectability of AGN.

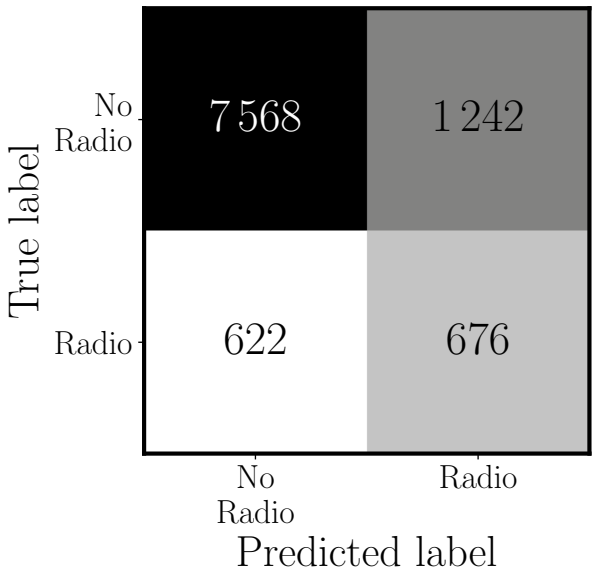


Figure 3.7: Confusion matrix from the results of application of radio-detection classification model for AGN to the HETDEX test subset. Description as in Fig. 3.6.

Another relevant property of the radio prediction confusion matrix is the that it reflects the high level of imbalance between both classes (with or without radio detection). There are almost six times more AGN without radio detection than with radio detection in the HETDEX field. For AGN in the S82 field, this ratio increases to 23 times. While the model has been trained to handle a large imbalance (as that present in the HETDEX field), it shows some difficulties when applied to sources in the very unbalanced S82 field.

3.7.3 Redshift prediction

In the case of redshift values prediction, the application of the stacked model over the testing subset and the labelled sources in the S82 field is summarised in the two first rows of Table 3.7 (named as HETDEX-test and S82-label).

These results show some degree of over-fitting, since the testing scores are a factor of two worse than those from the training subset (cf. Table 3.3). This trend is present in all used metrics. One reason for these differences might be due to the design of our prediction pipeline. As each step has been trained in a smaller dataset (and with sources that have to fulfill even more conditions than in the previous model), the possibilities that the redshift prediction model to extract information that can be successfully applied into independent datasets are inferior. Consequently, its application in the S82 field might (and does, as shown in Table 3.7) provide lower-quality results.

Table 3.7: Redshift prediction metrics for the test subset from HETDEX and S82 labelled sources as discussed in Sect. 3.7.4

Subset	$\sigma_{\rm MAD}$ ($\times 100$)	$\sigma_{\rm NMAD}$ ($\times 100$)	σ_z ($\times 100$)	$\sigma_z^{\rm N}$ ($\times 100$)	η ($\times 100$)
HETDEX-test	16.54 ± 2.55	7.27 ± 0.99	41.14 ± 9.97	20.56 ± 5.98	19.03 ± 3.35
S82-label	18.66 ± 2.26	9.28 ± 1.37	51.08 ± 11.62	24.69 ± 4.36	24.29 ± 4.68
HETDEX-pipe-Naive	8.11 ± 3.95	5.42 ± 2.19	32.00 ± 12.27	20.97 ± 9.69	19.01 ± 8.22
HETDEX-pipe-PR	15.86 ± 1.77	7.17 ± 0.81	37.80 ± 3.06	22.93 ± 2.73	18.91 ± 1.59
S82-pipe-Naive	15.17 ± 2.70	9.14 ± 1.23	43.05 ± 7.20	24.32 ± 5.00	24.09 ± 4.52
S82-pipe-PR	20.71 ± 1.23	9.84 ± 0.56	45.14 ± 4.42	26.14 ± 3.77	25.18 ± 2.26

[a] Values and uncertainties as in Table 3.5.

Likewise, the comparison between the original redshift values and those derived from the prediction results in the test subset is shown in Fig. 3.8. Apart from the good quality of predictions (as shown by the small fraction of sources in the outlier region, Eq. 3.22), an interesting effect can be noted. For sources with $z_{\rm True} \lesssim 1.0$, predicted values tend to be overestimated while, for $z_{\rm True}$ values above that limit, predictions do not show a preferred direction. These differences might hint issues with the prediction of radio-AGN redshifts at that range and with the quality of the data available for sources at different redshift ranges. Section 4.2 presents an analysis in this topic.

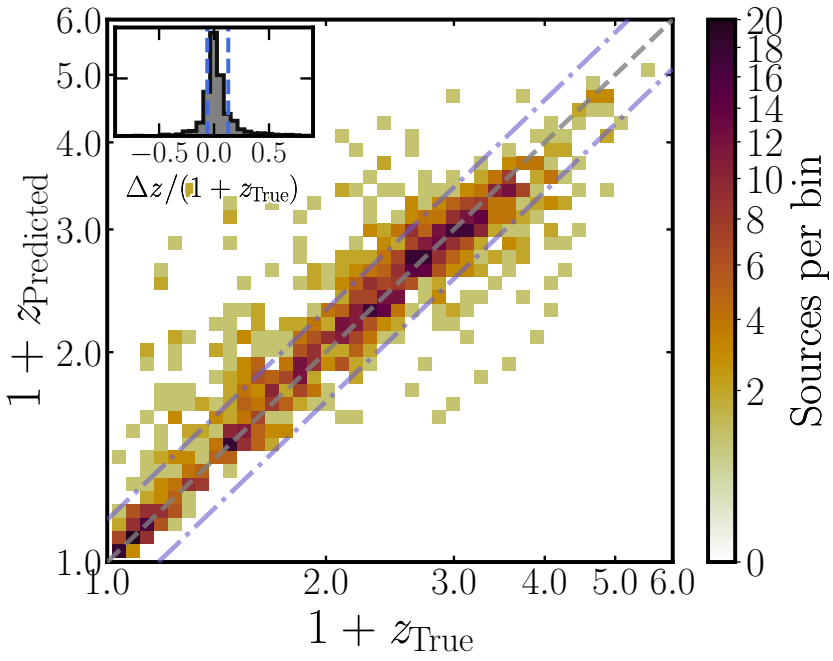


Figure 3.8: Two-dimensional histogram of comparison between original and predicted redshifts from the results of application of redshift prediction model to testing subset. Each point is coloured according to the source density in the bin following the colorbar. Grey, dashed line shows $x = y$ relation and purple, dot-dashed lines show the limits where outliers are defined (cf. Eqn. 3.22). Inset shows the distribution of Δz^{N} values from the points shown in main plot, with a $\langle \Delta z^{\mathrm{N}} \rangle = 0.0442$. Full scores of the presented subset can be found in Table 3.7 under the label 'HETDEX-test'.

3.7.4 Prediction from pipeline

The sequential combination of the models described in Sect. 3.6 defines the pipeline for the prediction of radio-detectable AGN and their redshift. As separate tasks, the pipeline was applied to the labelled sources in the HETDEX testing subset, to the labelled sources in S82, and to all unlabelled sources across both fields. We note that S82 provides a fully independent test of the pipeline as no data in this field was used for training the different models.

Metrics can be presented for each of the steps in their sequential application. That implies obtaining, first, metrics for the selection between AGN and SFGs. Then, metrics for the application of the radio detection model to sources predicted to be AGN (independent of their true class). And finally, metrics from the prediction of redshift values for predicted AGN that were also predicted to have radio detections. This sequence implies that uncertainties for any given step will be a convolution of the those from their own step and those of any previous step (if any). Such metrics are presented in Tables 3.5, 3.6, and 3.7 for the rows labelled as 'pipe'. In these cases, results are mixed when comparing them with the individual metrics (i.e. with labels 'test' or 'label'). It is important to note that, for the AGN-SFG classification, both analysed datasets, and their metrics, are the same since this is the first step of the pipeline.

The scores for the pipeline application of the radio-detection classification for the predicted AGN show (Table 3.6, 'pipe' subsets), a decrease in all the metrics when compared with individual application of steps ('label' subsets). But when considering the uncertainties, both values, from the individual application of the model and as part of the pipeline, are compatible. These results imply that no discernible change is found when using the pipeline. One possible explanation for this rubustness in the metrics might be related to the good performance of the first step of the pipeline (AGN-SFG classification). As most of the sources used in the second step of the pipeline have been correctly classified in the previous stage, no considerable change in the parameter space has been applied and the radio detection classifier would deliver results of the same quality.

In the case of the redshift prediction, the use of the uncertainties lead to full compatibility between the metrics in the individual models ('test' or 'label') and those that are part of the pipeline ('pipe-PR') and the large uncertainties are to be blamed for these results. Attending to the low (when compared to the AGN-SFG model) metrics of the radio detection model, it is expected that results have large uncertainties given the, apparent, inability of the model to extract fully meaningful data for the classification and application to different datasets from the training subset. Regardless of their compatibility, if uncertainties are omitted and only the mean metrics are considered, different evolutions are seen in each metric and subset. For the sources in the HETDEX field, all metrics improve when using the structure of the pipeline, except the value of $\sigma_z^{\rm N}$. When analysing the source in the S82 field, all metrics worsen except the value of σ_z. As defined in Sect. 3.1.2, both σ_z and $\sigma_z^{\rm N}$ have a structure that is similar to obtaining mean values of a sample. Opposite to that, $\sigma_{\rm MAD}$ and $\sigma_{\rm NMAD}$ use the median values of the sample. Thus, the latter metrics are less affected by outilers, which might be the reason behind their distinct behaviour.

This degradation might be, then, understood by the fact that the pipeline is composed of three sequential models. Each additional step is fed with sources classified by the previous algorithm. And some of these sources might not be similar, in terms of features, to those used for training, thus adding noise to the output of such model.

Additionally, both classification steps can be combined into one single stage (i.e. classification of radio-AGN), and metrics can be obtained accordingly. Their joint metrics are shown in Table 3.8 for the HETDEX test sample and the labelled sources in S82. Here, the differences between naive and PR-based results are varied, depending on the analysed score. Taking into

account that our classifiers have been optimised for the recall values, it is expected that this value presents the best evolution when applying the PR-based thresholds, even at the expense of the remaining metrics. In this way, an improvement of 25 % to 30 % can be verified for the recall.

Table 3.8: Results of application of radio AGN prediction pipeline to the labelled sources in the HETDEX and S82 fields

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	20.68 ± 3.17	24.93 ± 3.72	52.34 ± 6.56	13.79 ± 2.27
	PR	37.99 ± 2.59	33.66 ± 2.79	32.20 ± 2.72	44.61 ± 2.46
S82-label	Naive	24.08 ± 3.44	21.43 ± 3.53	25.44 ± 3.64	23.07 ± 3.72
	PR	19.42 ± 2.31	17.23 ± 3.08	11.33 ± 1.32	47.36 ± 6.22

[a] Values and uncertainties as in Table 3.5.

Figures 3.9 and 3.10 show the confusion matrices for the joint application of the radio-AGN prediction over the HETDEX and S82 fields, respectively. The most relevant trait of both matrices is the very high imbalance between the two analysed classes (radio-AGN and non-radio-AGN). Taking into account the areas of both fields (Sects.2.1 and 2.2), the density of both true and predicted radio-AGN is larger in the S82 field.

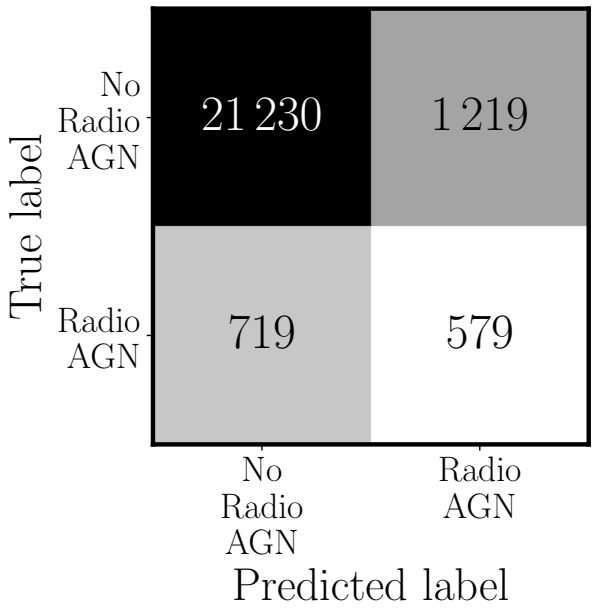


Figure 3.9: Combined confusion matrix from the full radio-AGN detectability prediction computed using the testing subset from HETDEX. Description as in Fig. 3.6.

On the other side, Figs. 3.11 and 3.12 show the contrast between the true and predicted redshifts for the sources predicted to be radio-AGN in the HETDEX and S82 fields. As with the results of the individual application of the model (Fig. 3.8), there are two distinct behaviours in

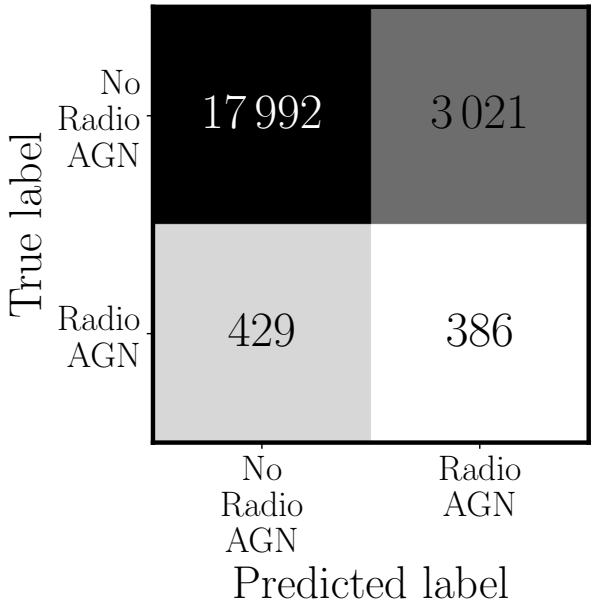


Figure 3.10: Combined confusion matrix from the full radio-AGN detectability prediction computed using the labelled sources from the S82 field. Description as in Fig. 3.6.

the studied sample at both sides of the $z_{\rm True} \sim 1.0$ value. The sources with $z_{\rm True} \lesssim 1.0$ will have their predicted redshift values equal or higher than the true value. Above that value, predictions can go in both directions, with a hint of general underprediction for both fields.

The application of the prediction pipeline to the 15 018 144 unlabelled sources from the HETDEX field led to 9 974 990 predicted AGN ($\sim$ 66 % of available sources), from which 68 252 ($\sim$ 0.5 % of the total number of sources and $\sim$ 0.7 % of predicted AGN) were predicted to be radio detectable. The pipeline predicts, as well, 2 073 997 AGN in the 3 568 478 unlabelled sources from S82 (close to a 58 %), being 22 445 ($\sim$ 0.6 % of the total number of available sources and $\sim$ 1.1 % of the predicted AGN) of the candidates to be detected in the radio (to the detection level of LoTSS-DR1). The pipeline outputs for a small sample of 20 highly probable predicted radio AGN are presented in Tables B.4 and B.5 for HETDEX and S82 respectively, while the full tables are available, from Carvajal et al. (2023b), at `https://zenodo.org/doi/10.5281/zenodo.10220008`. Additionally, Figs. B.1 and B.2 show radio cutouts from this selection of sources for a simple assessment of the prediction quality.

As a different graphical method to compare the predictions from our pipeline and those compiled previously in the literature, the normalised distribution of the predicted redshifts for radio-AGN in HETDEX and S82 is presented in Fig. 3.13. These histograms allow one to assess even predictions for sources that do not have a previous true value, as the analysis is focused on the distributions rather than the individual values.

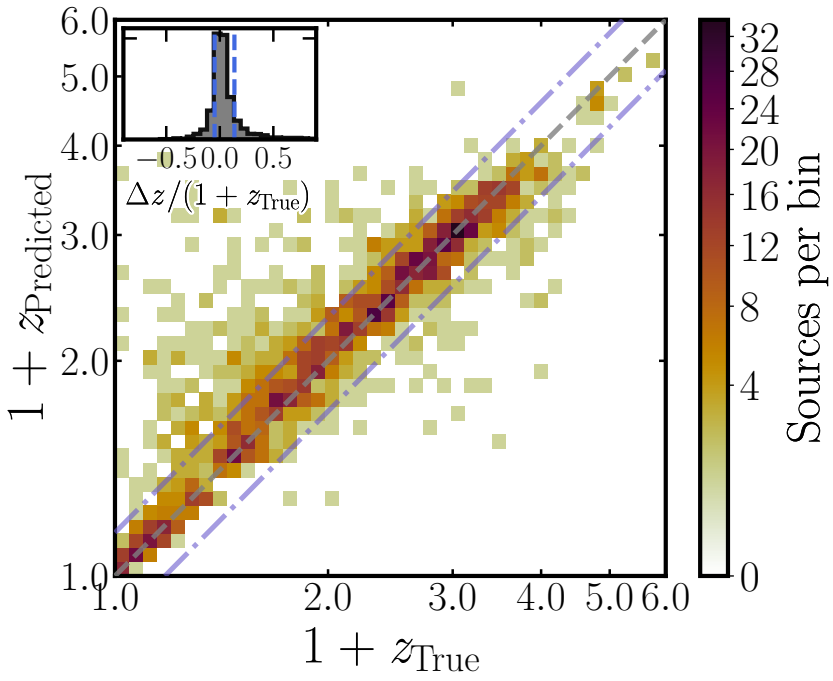


Figure 3.11: Two-dimensional histogram of comparison between original and predicted redshifts from the results of application of redshift prediction model to sources predicted to be radio-detectable AGN in the HETDEX testing subset. Details as in Fig. 3.8.

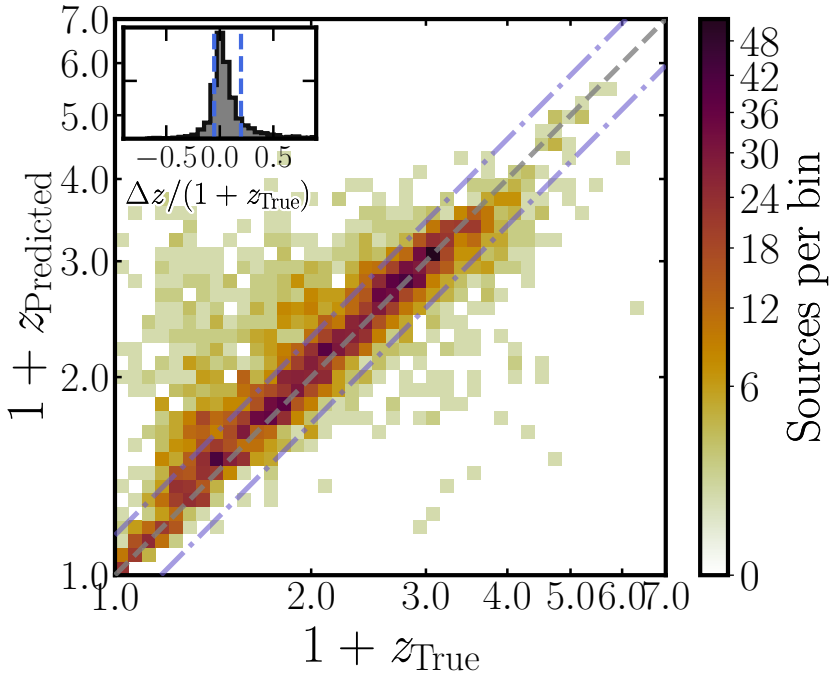


Figure 3.12: Two-dimensional histogram of comparison between original and predicted redshifts from the results of application of redshift prediction model to sources predicted to be radio-detectable AGN among labelled sources in the S82 field. Details as in Fig. 3.8.

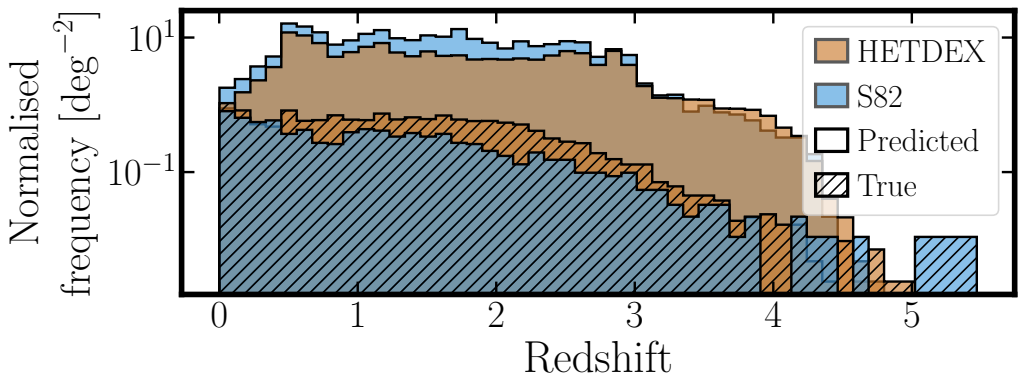


Figure 3.13: Redshift density distribution of the predicted radio-AGN within the unlabelled sources (clean histograms) in HETDEX (ochre histograms) and S82 (blue histograms) and true redshifts from labelled radio-AGN (dashed histograms).

Even though the ML models in our pipeline do not explicitly aim to reproduce the distribution of training targets when predicting in a dataset, they still reproduce biases and hidden correlations ingrained in the original data (e.g. Du et al. 2021). While data preprocessing (cf. Sect. 2.6) aims to solve these issues (Kamiran and Calders 2012), it is still possible to see, in Fig. 3.13, that the predicted redshifts, in both HETDEX and S82 fields, have very similar distributions.

A different, and diverging, approach to explain the distributions in Fig. 3.13 assumes that the similarity between predicted redshifts is an indicator that the application of the model, in two separate fields, does not affect the quality of the predictions. The model has already obtain all the necessary connections between the measurements in the training sample and it has become robust to changes in the features. In this way, rather than a disadvantage, their resemblance can be taken as a good symptom of solid models.

3.7.5 No-skill classification

As presented in Sect. 3.1.1, Eqs. 3.7 to 3.10 show the base results for a classification with no skill. Table 3.9 presents the scores generated by using this procedure in sources of the HETDEX and S82 fields. These values are the base from which any improvement, from a random selection, can be assessed.

Subsets and prediction modes displayed in Table 3.9 are the same as those exhibited in Tables 3.5, 3.6, and 3.8. For instance, in the test HETDEX sub-sample, $\sim 43\,\%$ of sources are labelled as AGN. From all AGN, $\sim 13\,\%$ of them have radio detections. These percentages can be summarised stating that $\sim 6\,\%$ of all sources in the test sub-sample are radio-detected AGN.

Table 3.9: Results of no-skill selection of sources in different stages of pipeline to the labelled sources in the HETDEX test subset and S82 fields

Subset	Prediction	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX	AGN-SFG	42.57	0.00	42.57	42.57
	Radio-detection (label)	12.84	0.00	12.84	12.84
	Radio AGN	5.47	0.00	5.47	5.47
S82	AGN-SFG	81.29	0.00	81.29	81.29
	Radio-detection (label)	4.59	0.00	4.59	4.59
	Radio AGN	3.73	0.00	3.73	3.73

Contrasting the results of Table 3.9 with those presented in Tables 3.5, 3.6, and 3.8 can give some insight into the level at which the trained models have learned the correlations and links between the used features and the target to predict. It is possible to see that all models present a very high improvement upon a random selection. With the exception of the AGN-SFG classification in the S82 field, all instances show improvements of $\gtrsim 40\,\%$ in the recall values. Such relevant changes can point to the fact that all models are, indeed, learning effectively about their tasks and the data underlying the classifications.

The case of the AGN-SFG classification in the S82 field presents an interesting view on the quality of the models. Even though S82 has a better base upon which improve for a classification between AGN and SFGs (81 % of S82 sources are AGN), the trained model does not take the metrics further than that in the HETDEX field. This behaviour might indicate that the model has already extracted all needed information and there is no room for over-fitting or other training issues.

4

Analysis of prediction method and results

Once the quality of the prediction pipeline, and the models that make part of it, has been established. And after the generation of new radio-AGN candidates in the HETDEX and S82 fields, these results need to be put in context and the impact of the training data in models analysed further.

This chapter attempts to understand if the creation and application of our prediction pipeline can be compared with the results of the application of different techniques for the selection of radio-detected AGN and the estimation of photometric redshift values. In particular, we aim to compare our results with those obtained by other teams using photometric measurements in different contexts and with different tools. Furthermore, we intend to explore the impact that the training (and testing) data can have on the behaviour and quality of the predictions. Towards this goal, we apply several ML-related analysis tools to each of our models as a way to understand which observed properties (and the physical processes linked to them) can drive the estimations and predictions.

4.1 Comparison with previous works

In this section, we provide a few examples of related published works as well as plausible explanations for observed discrepancies with our results when these are present. This comparison attempts to be representative of the literature on the subject but does not intends to be complete in any way.

4.1.1 AGN detection prediction

In order to understand the significance of our results and ways for future improvement, we separate the comparison with previous works in two parts. First, we present previously published results from traditional methodologies. In second place, we offer a comparison with ML methods.

As presented in Sect. 1.1.1, traditional AGN selection methods are based on the comparison of the measured SED photometry to a template library (Walcher et al. 2011). A recent example of its application is presented by Thorne et al. (2022b), where best-fit classifications were calculated for more than 700 000 galaxies in the D10 field of the Deep Extragalactic VIsible Legacy Survey (DEVILS; Davies et al. 2018a) and the Galaxy and Mass Assembly (GAMA; Driver et al. 2011; Liske et al. 2015) fields. Their 91 % recovery rate of AGN, selected through various means (X-ray measurements, narrow and broad emission lines, and MIR colours), is very much in line with our findings in S82, where our recall (completeness) reaches $\sim$ 94 %, and in the HETDEX test set, with a recall of $\sim$ 96 %, setting a baseline from which our results can be properly assessed.

Traditional methods also encompass the colour-based selection of AGN. While less precise, they provide access to a much larger base of candidates with a very low computational cost. We implemented some of the most common colour criteria on the data from S82. Of particular interest is the predicting power of the MIR colour selection due to its potential to detect hidden or heavily obscured AGN activity.

Based on *WISE* data, Stern et al. (2012: hereafter S12) proposed a threshold at (*W1* - *W2*) $\geq$ 0.8 to separate AGN from non-AGN using data from AGN in the COSMOS field. A more stringent criterion was developed by Mateos et al. (2012: hereafter M12), the AGN wedge, which can be defined by the sources located inside the region defined by the relations (*W1* - *W2*) $<$ 0.315 $\times$ (*W2* - *W3*) + 0.791, (*W1* - *W2*) $>$ 0.315 $\times$ (*W2* - *W3*) $-$ 0.222, and (*W1* - *W2*) $>$ $-$3.172 $\times$ (*W2* - *W3*) + 7.624. In order to define this wedge, they used data from X-ray selected AGN over an area of 44.43 deg^2 in the northern sky. Mingo et al. (2016: hereafter M16) cross-correlated data from *WISE* observations with X-ray and radio surveys creating a sample of SFGs and AGN in the northern sky. They developed individual relations to separate classes of SFGs and AGN in the (*W1* - *W2*), (*W2* - *W3*) space and, for AGN the criterion, the relation is (*W1* - *W2*) $\geq$ 0.5 and (*W2* - *W3*) $<$ 4.4.

More recently, Blecha et al. (2018: hereafter B18) analysed the quality of MIR colour selection methods for the identification of obscured AGN involved in mergers. Using hydrodynamic simulations for the evolution of AGN in galaxy mergers, they developed a selection criterion from *WISE* colours which is shown to be able to separate, with high reliability, starburst galaxies from AGN. The expressions have the form (*W1* - *W2*) $>$ 0.5, (*W2* - *W3*) $>$ 2.2, and (*W1* - *W2*) $>$ 2 $\times$ (*W2* - *W3*) $-$ 8.9.

The results from the application of these criteria to our samples in the testing subset and in the labelled sources of S82 field are summarised in Table 4.1 and a graphical representation of the boundaries they create in their respective parameter spaces is presented in Fig. 4.1. Table 4.1 shows that previous colour-colour criteria have been designed and calibrated to have very high precision (purity) values. Thus, most of the sources deemed to be AGN by them are, indeed, of such class. Our classifier, although tuned to maximise its recall (and F_β to a lesser extent), shows precision values compatible with those of such criteria. This result underlines the power of ML methods. They can be on a par with traditional colour-colour criteria and excel in additional metrics.

Table 4.1: Results of application of several AGN detection criteria to our testing subset and the labelled sources from the S82 field. Last row includes results from the first step of our prediction pipeline as presented in Table 3.5.

HETDEX test set				
Method[a]	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	86.10	78.78	93.98	80.51
M12	51.80	49.71	98.87	37.18
M16	67.21	61.30	97.48	53.48
B18	82.14	75.76	97.54	72.66
Our pipeline	95.42	91.85	94.49	96.21
S82 (labelled)				
Method[a]	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	83.59	45.47	93.93	76.62
M12	46.80	28.22	99.59	32.54
M16	64.69	37.76	98.80	50.32
B18	79.71	51.07	98.72	68.77
Our pipeline	94.37	70.67	94.81	94.01

[a] Naming codes for the used methods are described in the main text (cf. Sect. 4.1.1).

The four plots shown in Figure 4.1 are constructed as a confusion matrix, plotting in each quadrant the whole *WISE* population in the background and in colour contours the corresponding fraction of the testing set (TP, TN, FP, and FN, see descriptions of Fig. 3.6 and Sect. 3.1.1). As expected, our pipeline is able to separate with high confidence sources which are closer to the AGN or the SFG loci from previous works (M12, S12, M16, and B18 curves in the TP and TN quadrants) while sources in the FN and FP quadrant show a different situation. AGN predicted to be SFGs (FN, 1.6 % of true AGN for HETDEX, and 4.9 % for S82) are located in

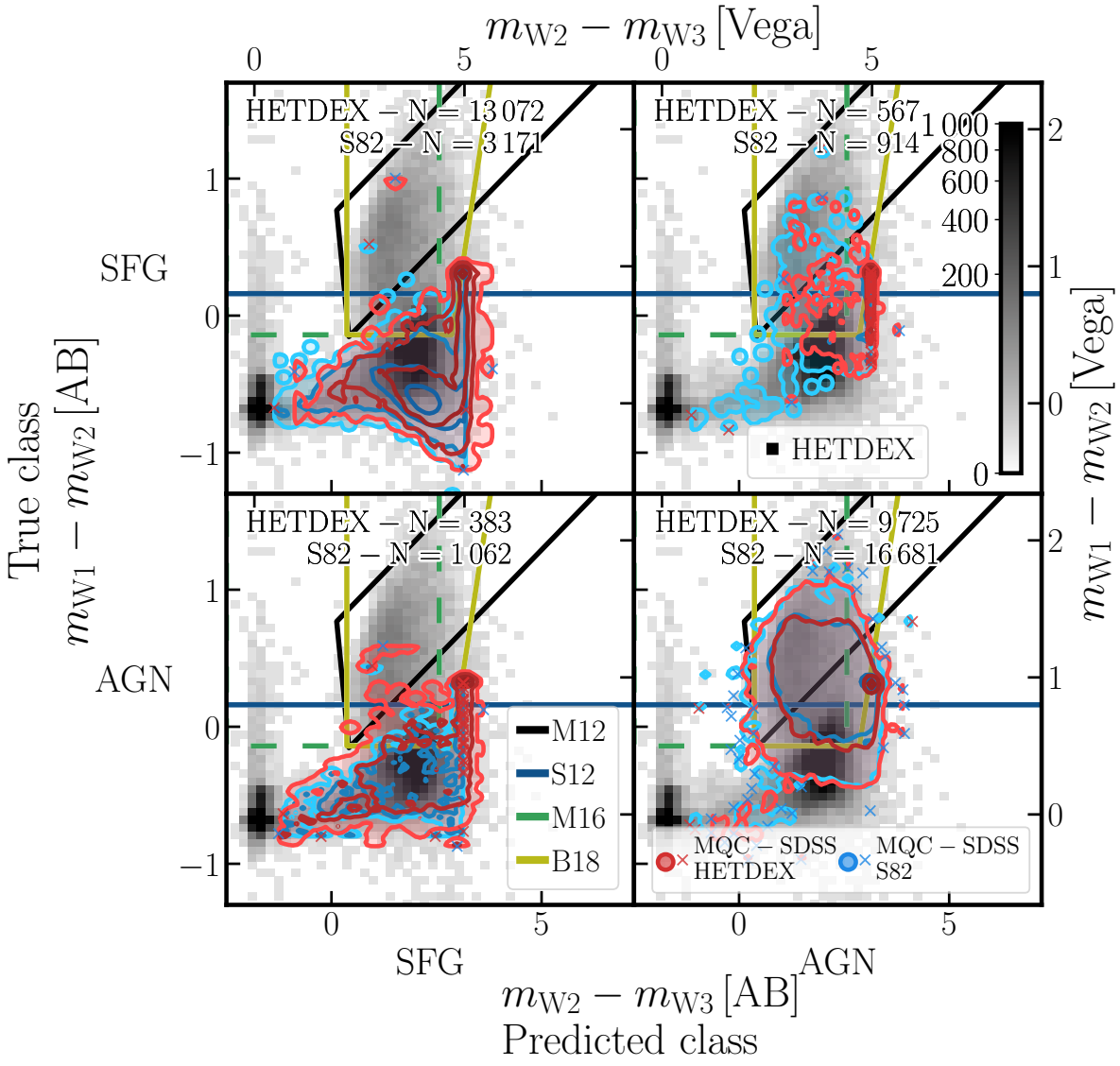


Figure 4.1: (*W1* - *W2*) vs (*W2* - *W3*) colour-colour diagrams for sources in the testing subset, from HETDEX, and labelled sources from S82 given their position in the AGN-SFG confusion matrix (see, for HETDEX, Fig. 3.9 and, for S82, Fig. 3.10). In the background, two-dimensional histogram of all CW-detected sources in the full HETDEX field sample is displayed. Colour of each square represents the number of sources in that position of parameter space, with darker squares having more sources (as defined in the colorbar of the upper-right panel). Contours represent distribution of sources for each of the aforementioned subsets at $1\,\sigma$, $2\,\sigma$, $3\,\sigma$, and $4\,\sigma$ levels (shades of blue, for testing set and shades of red for labelled S82 sources) and crosses show sources outside the selected contours (i.e. outliers, in the same colours as contours). Coloured, solid lines display limits from the criteria for the detection of AGN described in Sect. 4.1.1.

the SFG region of the colour-colour diagram, thus not presenting MIR properties of AGN. On the opposite corner of the plot, SFGs predicted to be AGN (FP, 2.4 % of sources for HETDEX, and 4.2 % for S82) cover the areas of AGN and SFGs uniformly but with apparent centroids located in the AGN region. FN sources might be sources that are identified as AGN by means not included in our feature set (e.g. X-ray, radio emission). Sources in the FP quadrant, alternatively, might be SFGs with extreme properties and colours similar to those of AGN. Thus, our model hints they might be AGN and that further analyses are needed for these sources.

The presence of sources in the off-diagonal elements of Fig. 4.1 motivates a deeper look into properties that might cause models to classify them differently from their original labels. While *WISE* colour distributions provided initial insights, further exploration with optical photometry can reveal additional factors.

Figure 4.2 shows the (r - i) vs (g - r) colour-colour diagram for our HETDEX test set sources and the S82-labeled sources. This colour combination is a well-established method for separating AGN and SFGs from stars (e.g. Richards et al. 2002, 2009; Schneider et al. 2010; Haggard et al. 2010). Its application has expanded from SDSS photometry to Pan-STARRS (e.g. Fu et al. 2024) and future LSST observations by Savić et al. (2023), who determined a region in the (r - i) vs (g - r) diagram for the selection of AGN (S23).

Focusing on PS1 photometry, we compare our results with those from Fu et al. (2024), who studied QSO candidates from *Gaia* data release 3 (Gaia Collaboration et al. 2023a,b). Their Fig. 3 includes the (r - i) vs (g - r) diagram, where clear separations exist between QSOs, SFGs, and the main stellar population. We observe a similar pattern in our data, with the background source distribution in Fig. 4.2 closely resembling the stellar distribution in Fu et al. (2024).

Crucially, the peaks for TP AGN and TN SFGs are distinct and well-separated. Notably, most FP sources lie within the TP AGN region, suggesting these sources might not have been previously classified as AGN despite existing signs in favour. Conversely, FN sources reside closer to the peak of the TN SFG distribution, consistent with Fig. 4.1.

For the case of ML-based models for AGN-SFG classification, several analyses have been published in recent years. An example of their application is provided in Clarke et al. (2020) where a RF model for the classification of stars, SFGs and AGN was trained using optical and NIR photometric data from more than 3 000 000 sources in the SDSS (DR15; Aguado et al. 2019) and *WISE* with associated spectroscopic observations. Close to 400 000 sources have a QSO spectroscopic label and, from the application of their model to a validation subset, they

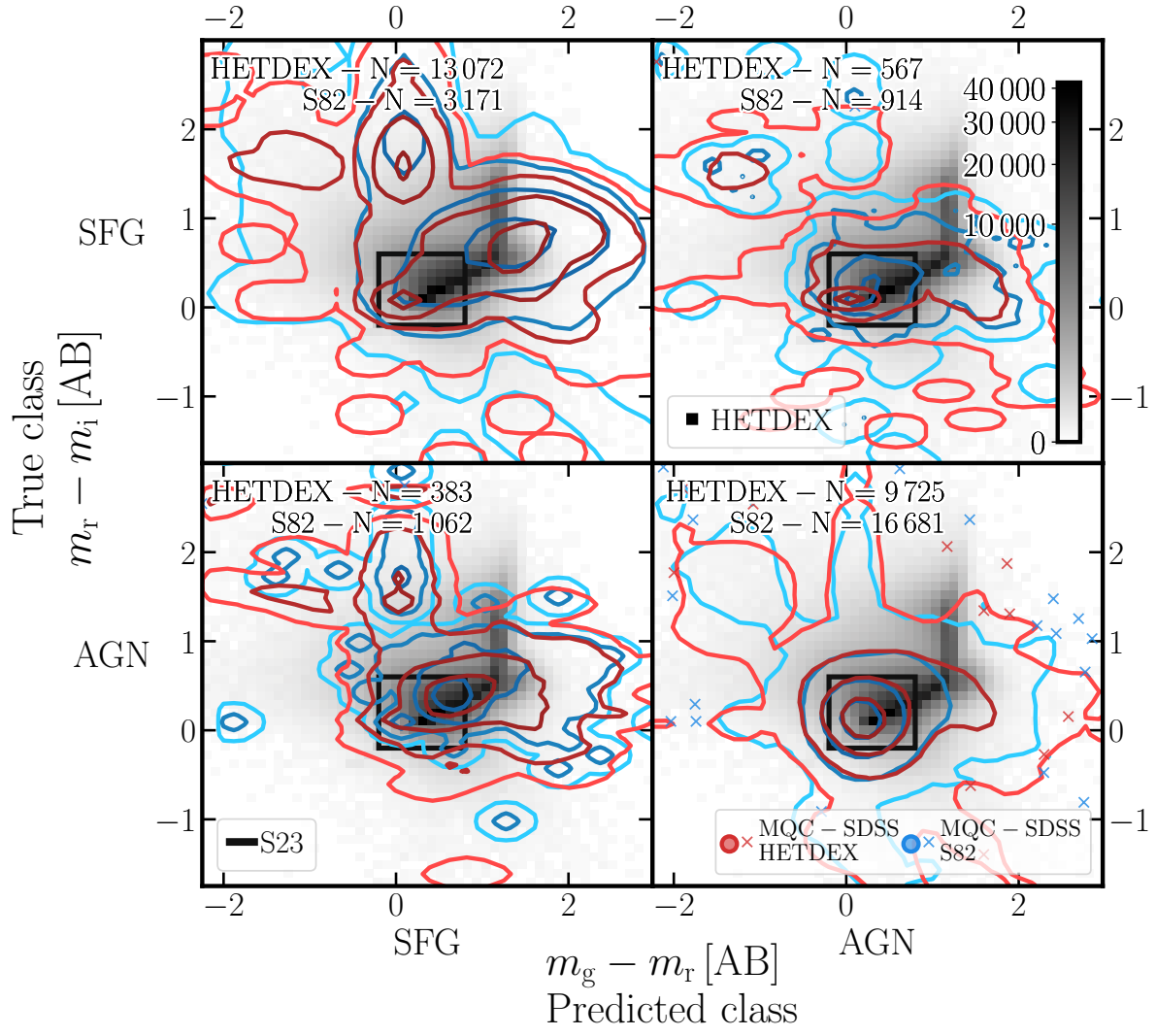


Figure 4.2: (*r* - *i*) vs (*g* - *r*) colour-colour diagrams for sources in the testing subset from HETDEX (blue) and labelled sources from S82 (red) given their position in the AGN-SFG confusion matrix. In the background, two-dimensional histogram of all CW-detected sources in the full HETDEX field sample following colourbar of upper-right panel. Contours represent distribution of sources for each of the aforementioned subsets at 1 σ, 2 σ, 3 σ, and 4 σ levels and crosses show sources outside the selected contours (i.e. outliers, in the same colours as contours). Black, solid lines display limits from the AGN selection criteria region described in Savić et al. (2023: S23).

obtain a recall (completeness) of 0.929 and F1-score of 0.943 for the QSO classification. These scores are of the same order as those obtained when applying our AGN-SFG model to the testing set (see Table 3.5). Thus, and despite using an order of magnitude fewer sources for the full training and validation process, our model can achieve equivalently good scores. Such difference in the number of training source can underline the fact that our models are able to extract, effectively, information for the connection between features and targets from a smaller dataset.

Expanding on the results from Clarke et al. (2020), Cunha and Humphrey (2022) built an ML pipeline, `SHEEP`[1], for the classification of sources into stars, SFGs and QSOs. In contrast to Clarke et al. (2020) or the pipeline described in this thesis, the first step in their analysis is the redshift prediction, which is used as part of the training features by the subsequent classifiers. They extracted *WISE* and SDSS data release 15 (SDSS-DR15; Aguado et al. 2019) photometric data for almost 3 500 000 sources classified as stars, SFGs or QSOs. The application of their pipeline to sources predicted to be QSO led to a recall of 0.960 and an F1 score of 0.967. The improved scores in their pipeline might be a consequence not only of the slightly larger pool of sources, but also the inclusion of the coordinates of the sources (right ascension, declination) and the predicted redshift values as features in the training that might hint hidden correlations between sources that are, for instance, located in similar regions of the sky (i.e. with comparable line-of-sight properties). Authors also associate that coordinates have high importance in their models as QSOs tend to be detected away from the galactic plane. Additionally, and as shown in their feature importance analysis, a large fraction of the predictions are driven by the colours constructed by the differences of magnitudes in the same bands but with different techniques (e.g. PSF and model magnitudes), allowing the inclusion of morphological information into their models. For instance, the difference of two equal-band magnitudes calculated with different apertures can hint how extended a source can be.

A test with a larger number of ML methods, but with a smaller pool of training sources, was performed by Poliszczuk et al. (2021). For training, they used optical and infrared data from close to 1500 sources (SFGs and AGN) located at the AKARI North Ecliptic Pole (NEP) Wide-field (Lee et al. 2009; Kim et al. 2012) covering a 5.4 $\deg^2$ area. They tested linear regression (`LR`), `SVM`, `RF`, `ET`, and `XGBoost` including the possibility of generalised stacking. In general, they obtained results with F1-scores between 0.60 and 0.70 and recall values in the range of

[1] https://github.com/pedro-acunha/SHEEP

50 % to 80 %. These values, lower than the works described here, can be fully understood given the small size of the training sample and the studied area. Once again, a larger photometric sample covers a wider range of the parameter space which significantly helps improving the metrics of any given model (e.g. Brescia et al. 2021).

Table 4.2 summarises the results of AGN selection from previous works (as described in the previous paragraphs) for completeness and includes metrics from the first step of our prediction pipeline for comparison.

Table 4.2: Reported results of application of several AGN detection criteria. Last row includes results from the first step of our prediction pipeline as presented in Table 3.5.

Reference[a]	Region/Survey	N	F1 (×100)	Precision (×100)	Recall (×100)
Thorne et al. (2022b)	DEVILS	494 084	...	...	41.5
	GAMA	233 762	...	...	28.7
Fu et al. (2024)	*Gaia*-DR3	6 649 162	99.85	99.85	99.85
Clarke et al. (2020)	SDSS-DR15	3 099 457	94.3	95.7	92.9
Cunha and Humphrey (2022)	SDSS-DR15	3 497 864	96.7	97.5	96.0
Poliszczuk et al. (2021)[c]	AKARI	1547	66.0	73.0	61.0
Our pipeline	HETDEX	118 734	95.42[b]	94.49	96.21

[a] Naming codes for the used methods are described in the main text (cf. Sect. 4.1.1).
[b] Our pipeline reports F_β instead of F1 score.
[c] Metrics from stacked classifier.

4.1.2 Radio detection prediction

We have not found in the literature any work attempting to estimate the likelihood of AGN being detected, in the radio, above a certain level and therefore this is, to the best of our knowledge, the first attempt at doing so. In the literature we do find several correlations between the radio emission (their flux or luminosity) of a source and that at other wavelengths (e.g. with IR emission, Helou et al. 1985; Condon 1992; and with X-rays, D'Amato et al. 2022) and substantial effort has been done towards classifying radio galaxies based upon their morphology (e.g. Aniyan and Thorat 2017; Wu et al. 2019; Fanaroff-Riley class I (FRI), Fanaroff-Riley class II (FRII), bent jets, etc.) and its connection to environment (Miley and De Breuck 2008; Magliocchetti 2022). None of these extensive works has directly focused on the *a priori* presence or absence of radio emission above a certain threshold. Therefore, the results presented here are the first attempt at such an effort.

The twofold increase in the success rate of the pipeline to identify radio emission in AGN ($\sim$ 44.61 % recall and $\sim$ 32.20 % precision; see Table 3.8) with the respect to a 'no-skill' or random ($\lessapprox$ 30 %) selection, provides the opportunity to understand what the model has learned from the data and, therefore, gain some insight into the nature or triggering mechanisms of the radio emission.

4.1.3 Redshift prediction

We compare our results to that of Ananna et al. (2017: hereafter Stripe 82X) where the authors analysed multi-wavelength data from more than 6100 X-ray detected AGN from the 31.3 deg^2 of the Stripe 82X survey. They obtained photometric redshifts for almost 6000 of these sources using the template-based fitting code `LePHARE`. Their results present a normalised median absolute deviation of $\sigma_{\rm NMAD} = 0.0602$ and an outlier fraction of $\eta = 13.69$ %, values which are similar to our results in HETDEX and S82 except for a better outlier fraction (as shown in Table 3.7, we obtain $\eta_{S82} = 25.18$ %, $\sigma_{\rm NMAD}^{\rm HETDEX} = 0.071$, and $\eta^{\rm HETDEX} = 18.9$ %). Such results reinforce the use of ML methods, and in particular our pipeline, for the estimation of photometric redshifts of radio-detectable AGN.

On the ML side, we compare our results to those produced by Carvajal et al. (2021) in S82, with $\sigma_{\rm NMAD} = 0.1197$ and $\eta = 29.72$ %, and find that our redshift prediction model improves by at least 25 % for any given metric. The source of improvement is probably many-fold. First, it might be related to the different sets of features used (colours vs ratios of magnitudes) and second, the more specific population of radio-AGN used to train our models. Carvajal et al. (2021) used a limited set of colours to train their model, while we have allowed the use of all available combinations of magnitudes (Sect. 2.5). Additionally, their redshift model was trained on all available AGN in HETDEX, while we have trained (and tested) it only with radio-detected AGN. Using a more constrained sample reduces the likelihood of handling sources that are too different in the parameter space.

Another example of the use of ML for AGN redshift prediction has been presented by Luken et al. (2019). They studied the use of the k-nearest neighbours (`KNN`; Cover and Hart 1967) algorithm, a non-parametric supervised learning approach, to derive redshift values for radio-detectable sources. They combined 1.4 GHz radio measurements, infrared, and optical photometry in the European Large Area ISO Survey-South 1 (ELAIS-S1; Oliver et al. 2000) and extended Chandra Deep Field South (eCDFS; Lehmer et al. 2005) fields, matching their

sensitivities and depths to the expected values in the Evolutionary Map of the Universe (EMU; Norris et al. 2011) field. From the different experiments they run, their resulting NMAD values are in the range $\sigma_{\rm NMAD} = 0.05$ to 0.06, and their outlier fraction can be found between $\eta = 7.35\,\%$ and $13.88\,\%$. As an extension to the previous results, Luken et al. (2022) analysed multi-wavelength data from radio-detected sources in the eCDFS and ELAIS-S1 fields. Using `KNN` and `RF` methods to predict the redshifts of more than 1300 RGs, they have developed regression methods that show NMAD values between $\sigma_{\rm NMAD} = 0.03$ and 0.06, $\sigma_z = 0.10$ to 0.19, and outlier fractions of $\eta = 5.85\,\%$ and $12.75\,\%$. Comparing with the scores from the use of PR-based thresholds (see Table 3.7), their $\sigma_{\rm NMAD}$ are compatible with our results. Nevertheless, our σ_z and η values are, approximately, two times theirs. It is worth reminding that the definitions of η and $\sigma_{\rm NMAD}$ make these metrics susceptible to extreme outlier values. The sample used by Luken et al. (2022) included spectroscopic confirmations and redshifts obtained with a single method (Yuan et al. 2015) and, thus, their sample is less prone to large photometric variation than ours. On the other hand, our dataset has been built with classifications from both SDSS and MQC, thus considering several detection and redshift determination methods (e.g. including low-spectral resolution measurements from *Gaia*).

In addition to the previous work, Norris et al. (2019) compared a number of methodologies, mostly related with ML but also `LePHARE`, for predicting redshift values for radio sources. They have used more than 45 photometric measurements (including 1.4 GHz fluxes) from different surveys in the COSMOS field. From several settings of features, sensitivities, and parameters, they retrieved ML-based redshift predictions with NMAD values between $\sigma_{\rm NMAD} = 0.054$ and $\sigma_{\rm NMAD} = 0.48$ and outlier fractions that range between $\eta = 7\,\%$ and $\eta = 80\,\%$. The broad span of obtained values might be due to the combinations of properties for each individual training set (including the use of radio or X-ray measurements, the selection depth, and others) and to the size of these sets, which was small for ML purposes (less than 400 sources). When compared to our metrics, the slightly better results for some of their configurations (e.g. using `LePHARE` with their full photometric set and ML-based methods using deep IR photomery) can be understood given the heavily populated photometric data available in COSMOS.

Specifically related to the HETDEX field, it is possible to compare our results to those from Duncan et al. (2019). They used a hybrid photometric redshift approach combining traditional template fitting redshift determination (Brammer et al. 2008; Brown et al. 2014; Salvato et al. 2011, 2018) and ML-based methods. In particular, they implemented a GP algorithm (GPz;

Almosallam et al. 2016b,a), which is able to model both the intrinsic noise and the uncertainties of the training features. Their redshift prediction analysis of AGN with a spectroscopic redshift detected in the LoTSS-DR1 (6811 sources) found a NMAD value of $\sigma_{\rm NMAD} = 0.102$ and an outlier fraction of $\eta = 26.6\,\%$. The differences between these results and those obtained from the application of our models (individually and as part of the prediction pipeline) might be due to the differences in the creation of the training sets. Duncan et al. (2019) used information from all available sources in the HETDEX field for training the redshift GP whilst our redshift model has been only trained on radio-detected AGN, giving it the opportunity to focus its parameter exploration only on these sources.

Finally, Cunha and Humphrey (2022) also produced photometric redshift predictions for almost 3 500 000 sources (stars, SFGs, and QSO) as part of their pipeline (see Sect. 4.1.1). They combined three algorithms for their predictions through ensemble learning: `XGBoost`, `CatBoost`, and Light Gradient Boosting Machine (`LightBGM`; Ke et al. 2017). This procedure leads to $\sigma_{\rm NMAD} = 0.018$, $\sigma_z^{\rm N} = 0.0124$, and $\eta = 2.65\,\%$ for all three classes of sources combined. As with previous examples, the differences with our results can be a consequence of the number of training samples, including galaxies and stars. Also, in the case of Cunha and Humphrey (2022), they applied an additional post-processing step to the redshift predictions attempting to predict and understand the appearance of catastrophic outliers.

Table 4.3 summarises the redshift estimation results from previous works (as described in the previous paragraphs) for completeness and includes metrics from the third step of our prediction pipeline for comparison.

Table 4.3: Reported results of application of several redshift estimation criteria. Last row includes results from the third step of our prediction pipeline as presented in Table 3.7.

Reference[a]	Region/Survey	N	$\sigma_{\rm NMAD}$ ($\times100$)	σ_z ($\times100$)	$\sigma_z^{\rm N}$ ($\times100$)	η ($\times100$)
Ananna et al. (2017)	Stripe82	6181	6.02	...	41.5	13.69
Carvajal et al. (2021)	Stripe82	2941	13.92	27.56	11.62	11.58
Luken et al. (2022)[b]	eCDFS + ELAIS-S1	4780	3.0	10.0	...	5.85
Norris et al. (2019)[b]	COSMOS	757	5.0	...	...	7.0
Duncan et al. (2019)	HETDEX	6811	10.2	...	...	26.6
Our pipeline	HETDEX	4612	9.28	51.08	24.69	24.29

[a] Naming codes for the used methods are described in the main text (cf. Sect. 4.1.1).
[b] Best value from each metric is presented.

4.2 Influence of data imputation

One effect process which might influence the training of the models and, consequently, the prediction for new sources is the imputation of missing values as described in Sect. 2.4 and the work by Curran (2022) and Curran et al. (2022). To better quantify the impact of imputed data we analysed the output of our models as a function of the number of imputed measurements per source.

In Fig. 4.3, we have plotted the distributions of predicted scores (for classification models) and predicted redshift values as a function of the number of measured bands (`band_num`) for each step of the pipeline as applied to sources predicted to be of each class in the test subset. The top panel of Fig. 4.3 shows the influence of the degree of imputation in the classification between AGN and SFGs. For most of the bins, probabilities for predicted SFGs are distributed close to 0.0, without any noticeable trend, implying a low impact of the imputation and robustness of the predictions. In the case of predicted AGN, the combination of low number of sources in some bins and high degree of imputation (`band_num` < 5) lead to low mean probabilities. Only from analysing the plot, it is possible to state that studied sources might have between 5 and 9 proper measurements (i.e. 3 to 7 missing entries) and, still, have reliable AGN predictions.

The case of radio detection classification is somewhat different. Given the number and distribution of sources per bin, it is not possible to extract any strong trend for the probabilities of radio-predicted sources. The absence of evolution with the number of observed bands is stronger for sources predicted to be devoid of radio detection. This lack of strong probabilities, combined with the previous analyses of the radio detection model, can be translated into the low metrics found for this classifier.

Finally, a stronger effect can be seen with the evolution of predicted redshift values for radio-detectable AGN. Despite the lower number of available sources, it is possible to recognise that sources with higher number of available measurements are predicted to have lower redshift values. Sources that are closer to us have, then, higher chances to be detected in a large number of bands. Thus, it is expected that our model predicts lower redshift values for the sources with the largest number of measurements in the field. Apart from being a data-based correlation, the behaviour seen between the number of measurements and the predicted redshift has its foundations in the physical connection between distance from a source and measured fluxes.

One interesting feature of all panels in Fig. 4.3 is the lack of sources with `band_num` $= 11$.

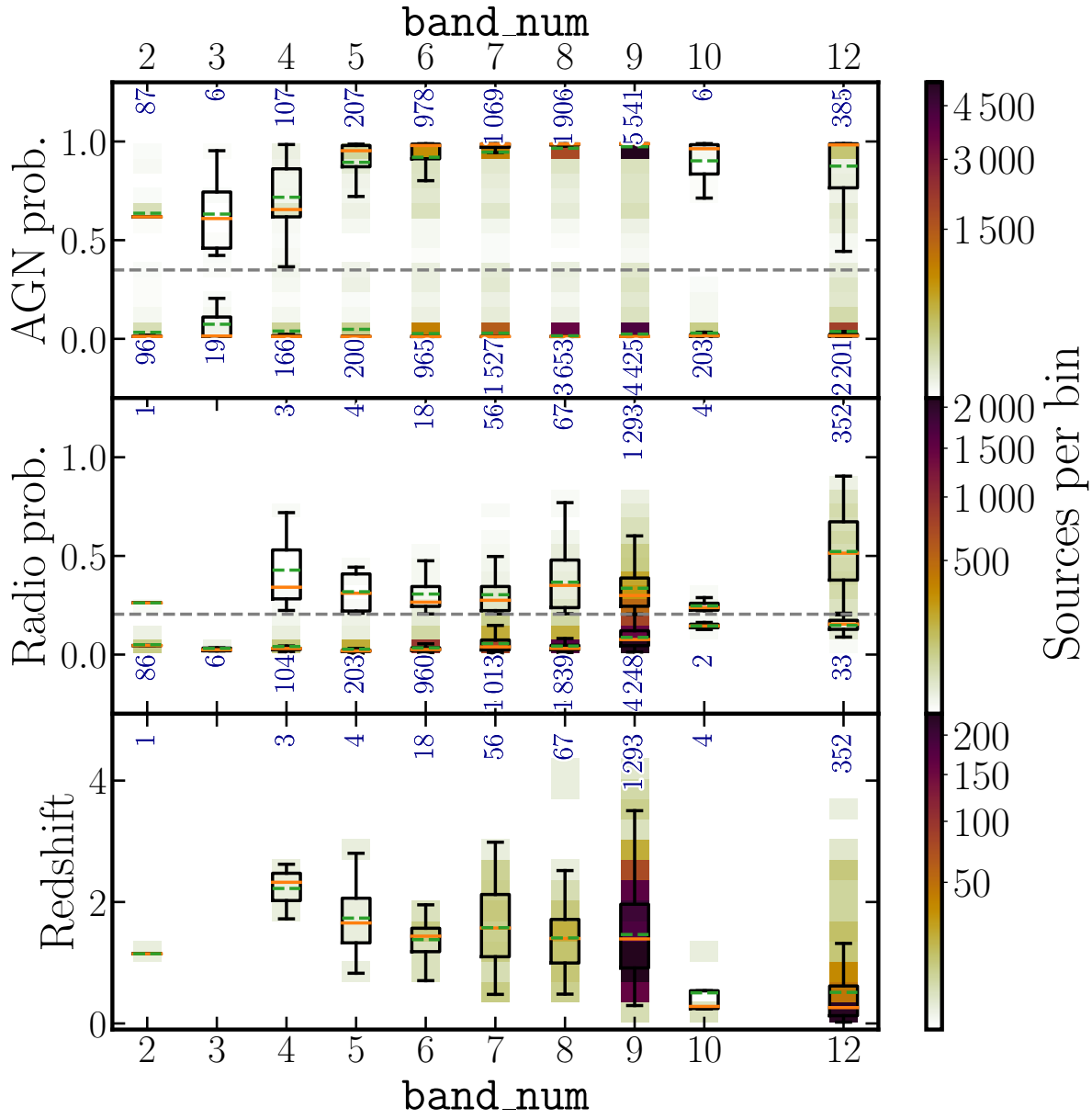


Figure 4.3: Evolution of predicted probabilities (top: probability to be AGN, middle: probability of AGN to be radio-detected) and redshift values for radio-detectable AGN (bottom panel) as function of number of observed bands for sources in test set. In top panel, sources have been divided between those predicted to be AGN and SFGs. In middle panel, sources are divided between predicted AGN that are predicted to be radio-detected and those predicted to not have radio detection. Background density plots (following colour coding in colorbars) show location of predicted values. Overlaid boxplots display main statistics for each number of measured bands. Black rectangles encompass sources in second and third quartiles. Error bars show population range from first to fourth quartiles. Orange lines represent median value of sample and dashed, green lines indicate their mean values. Dashed, grey lines show PR thresholds for AGN-SFG and radio detection classifications. Inset values in blue report the number of sources considered to create each set of statistics.

This effect is caused by the inclusion of measurements from 2M. As seen in Fig. 2.3, all three 2M bands have the highest (and almost the same) number of missing measurements. Thus, it is possible to infer that the inclusion of a measurement in one of the 2M bands will imply the addition, almost in every case, of the two remaining bands.

In consequence, Fig. 4.3 allows us to understand the influence of imputation over the predictions. The most highly affected quantity is the redshift, where large fractions of measured magnitudes are needed to obtain scores that are in line with previous results (cf. Sect. 4.1.3). The AGN-SFG and radio detection classifications show a mild influence of imputation in their results.

4.3 Global feature importances

Following the description in Sect. 1.3.3, it becomes of paramount importance to understand which features, and to what level, drive the predictions made by ML-based models. While good prediction metrics are the first filter for the selection of successful models, if the reasons behind their decisions (i.e. the used features) are not physically sound, their predictions cannot be considered as correct. For this reason, we have applied global and local feature importance tools to our models to retrieve the physical basis for their predictions.

Following the description of Sect. 1.3.3, all algorithms selected in this work (`RF`, `CatBoost`, `XGBoost`, `ET`, `GBR`, and `GBC`) belong to the DT class. Thus, it is possible to obtain their global feature importances, which are based on the mean decrease of impurity produced by each feature. Table 4.4 presents the ranked combined importances from the observables selected in each of the three sequential models that compose the pipeline. They have been combined using the importances from the meta-learner (as shown in Table 4.5) and that of base-learners. The derived importances will be dependent on the dataset used, including any imputation for the missing data, and the details of the models, that is, algorithms used and stacking procedure. We first notice in Table 4.4 that the order of the features is different for all three models. This difference reinforces the need, as stated in Chapter 4, of developing separate models for each of the prediction stages of this work that would evaluate the best feature weights for the related classification or regression task.

For the AGN-SFG classification model, it is very interesting to note that the most important features for the predicted probability of a source to be an AGN are the *WISE* colours (*W1* - *W2*)

Table 4.4: Relative importances (rescaled to add to 100) for observed features from the three models combined between meta and base models.

AGN-SFG (meta-model: CatBoost)					
Feature	Importance	Feature	Importance	Feature	Importance
W1_W2	68.945	H_K	1.715	z_W2	1.026
W1_W3	4.753	y_W1	1.659	z_y	0.722
g_r	4.040	y_W2	1.513	W3_W4	0.669
r_J	4.006	i_y	1.441	W4mag	0.558
r_i	3.780	i_z	1.366	H_W3	0.408
band_num	1.842	y_J	1.187	J_H	0.371
Radio detection (meta-model: GBC)					
Feature	Importance	Feature	Importance	Feature	Importance
W2_W3	9.609	y_W1	7.150	W4mag	4.759
y_J	8.102	g_r	7.123	K_W4	2.280
W1_W2	8.010	z_W1	7.076	J_H	1.283
g_i	7.446	r_z	6.981	H_K	1.030
K_W3	7.357	i_z	6.867	band_num	1.018
z_y	7.321	r_i	6.588		
Redshift prediction (meta-model: ET)					
Feature	Importance	Feature	Importance	Feature	Importance
y_W1	35.572	y_J	3.018	i_z	1.215
W1_W2	13.526	r_z	3.000	J_H	1.162
W2_W3	12.608	r_i	2.896	g_W3	1.000
band_number	6.358	z_y	2.827	K_W3	0.925
H_K	4.984	W4mag	2.784	K_W4	0.762
g_r	4.954	i_y	2.408		

[a] Relative feature importance values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model. A meaningful comparison can be done by contrasting the order in which features are sorted.

Table 4.5: Relative feature importances (rescaled to add to 100) for base algorithms in each prediction step.

AGN-SFG model (CatBoost)			
Feature	Importance	Feature	Importance
gbc	49.709	xgboost	14.046
et	19.403	rf	8.981
Remaining feature importances:			7.861

Radio detection model (GBC)			
Feature	Importance	Feature	Importance
rf	12.024	catboost	7.137
et	7.154	xgboost	6.604
Remaining importances:			67.081

Redshift prediction model (ET)			
Feature	Importance	Feature	Importance
xgboost	25.138	catboost	21.072
gbr	21.864	rf	13.709
Remaining importances:			18.217

[a] Relative feature importance values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model. A meaningful comparison can be done by contrasting the order in which features are sorted.

and (*W1* - *W3*). In particular, (*W1* - *W2*), which shows a very high importance fraction, is indeed one of the axes of the widely used *WISE* colour-colour selection, with the second axis being the (*W2* - *W3*) colour (cf. Sect 4.1.1). However, the *WISE W3* photometry is significantly less sensitive than *W1*, *W2* or PS1 (see Fig. 2.5) and a significant number of sources will be represented with upper limits in such plot (see Table 2.3).

Finally, the redshift prediction model shows that the final estimate is mostly driven by the results of the base learners, accounting for $\sim$ 82 % of the predicting power. The overall combined importance of features shows also in this case a strong dependence on several MIR colours of which (*y* - *W1*) and (*W1* - *W2*) are the most relevant ones (these colours have also been highlighted in previous ML redshift determinations, e.g. Kunsági-Máté et al. 2022). The model still relies, to a lesser extent, on a broad range of optical features needed to trace the broad range of redshift possibilities ($z \in [0, 6]$, the base of our training sample).

4.4 Local feature importances

As mentioned previously, and opposite to global feature importances, local feature importances are instrumental in understanding the properties that can drive individual predictions (or a subset of predictions). While global feature importances can give a mean view of the most relevant properties in a model, their local counterparts facilitate the analysis of specific areas of the parameter space. This is of relevance in the study of mispredictions, where one needs to understand the reasons for the incorrect output from the model and compare them with the drivers of the correcly predicted elements.

The combination of Shapley values with several other model explanation methods was used by Lundberg and Lee (2017) to create the SHAP values, which are a more computationally efficient way to calculate Shapley values for ML. In this work, SHAP values were calculated using the `Python` package `SHAP`[2] and, in particular, its module for tree-based predictors (Lundberg et al. 2020). To speed calculations up, the package `FastTreeSHAP`[3] (`v0.1.2`; Yang 2021) was also used, which allows the user to run multi-thread computations.

One graphical way to display these SHAP values is through the so-called decision plots. They can show how individual predictions are driven by the inclusion of each feature. Besides determining the most relevant properties that help the model make a decision, it is possible to

[2] https://github.com/slundberg/shap
[3] https://github.com/linkedin/fasttreeshap

detect sources that follow different prediction paths which could be, eventually and upon further examination, labelled as outliers.

For ease of visualisation purposes, we have produced example decision plots for a small subsample of sources. In this case, we have selected high-redshift ($z \geq 4.0$) spectroscopically classified AGN (i.e. they can be found in the MQC) in the HETDEX field (121 sources, regardless of them being part of any subset involved in the training or validation of the models). Given our interest on the earliest epochs of galactic evolution (cf. Chapter 1), the selection of high-redshift confirmed AGN can help us extract additional information on the features that can help our models make their decisions.

The decision plot of the AGN-SFG classification for the highest-redshift AGN subsample ($z_{\rm spec} \geq 4$) is shown in Fig. 4.4. The different features used by the meta-learner are stacked on the vertical axis with increasing weight and these final weight are summarised in Table 4.6. Similarly, SHAP decision plots for the radio-detection and redshift prediction are presented in Figs. 4.5 and 4.6, respectively. A thin trace of predictions (i.e. the set of lines as evolving through features) is a sign that the displayed sources have very similar values of the features involved (e.g. bottom features in Fig. 4.4). Once the lines flare out, the model shows that the studied sources occupy a wider region on the parameter space for these features and it is easier to differentiate them for classification (or regression) purposes.

As it can be seen, for the three models, base learners are amongst the features with the highest influence. This result raises the question of what observed properties drive these individual base predictions. In order to approach an answer to this question, Figs. 4.7, 4.8, and 4.9 show SHAP decision plots for all base learners used in this work which will be analysed further in this text. Additionally, and to be able to compare these results with the features importances from Sect. 4.3, we constructed Table 4.7, which displays the combined SHAP values of base and meta learners but, in this case, for the same 121 high-redshift confirmed AGN (with 29 of them detected by LoTSS). Table 4.7 shows, as Table 4.4, that the colour (*W1* - *W2*) is the most important discriminator between AGN and SFGs for this specific set of sources. The importance of the rest of the features is mixed: similar colours are located on the top spots (e.g. (*g* - *r*), (*W1* - *W3*) or (*r* - *i*)). When comparing these results with those obtained with global feature importances (Sect. 4.3), it is possible to see that both methods present (*W1* - *W2*) as the most relevant feature. However, the values of the importances vary by a factor of two. This difference might point to the fact that analysing only high-redshift sources can be harder for the

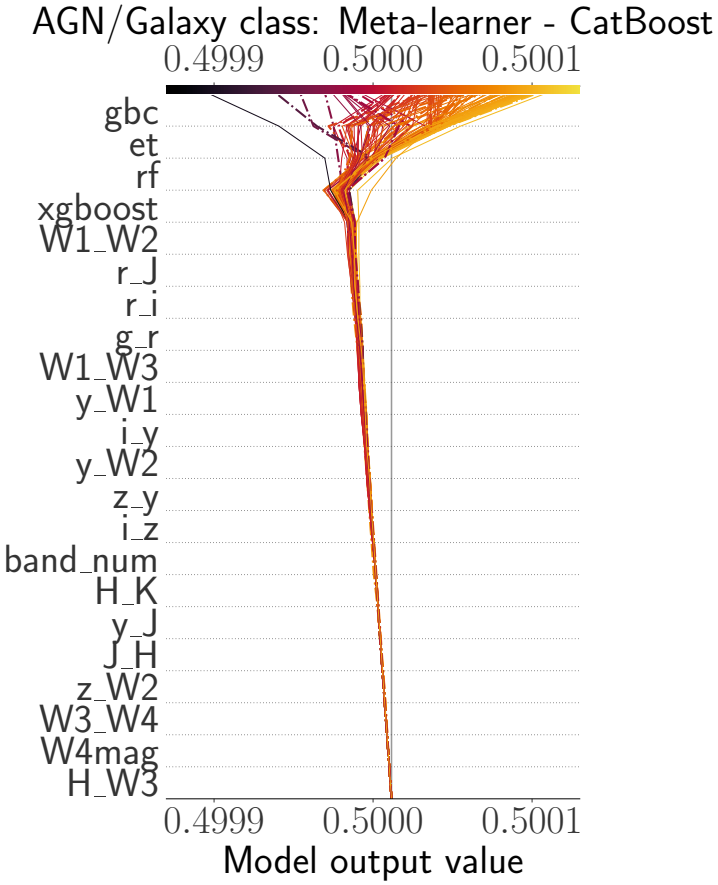


Figure 4.4: Decision plot from SHAP values for AGN-SFG classification from the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN in HETDEX. Horizontal axis represents the model's output with a starting value for each source centred on the selected naive threshold for classification. Vertical axis shows features used in the model sorted, from top to bottom, by decreasing mean absolute SHAP value. Each prediction is represented by a coloured line corresponding to its final predicted value as shown by the colorbar at the top. Moving from the bottom of the plot to the top, SHAP values for each feature are added to the previous value in order to highlight how each feature contributes to the overall prediction. Predictions for sources detected by LOFAR are highlighted with a dotted, dashed line.

Table 4.6: SHAP values (rescaled to add to 100) for base algorithms in each prediction step for observed features using 121 spectroscopically confirmed AGN at high redshift values ($z > 4$).

AGN-SFG model (CatBoost)			
Feature	SHAP value	Feature	SHAP value
gbc	36.250	rf	21.835
et	30.034	xgboost	7.198
	Remaining SHAP values:		4.683
Radio detection model (GBC)			
Feature	SHAP value	Feature	SHAP value
rf	11.423	catboost	5.696
xgboost	7.741	et	5.115
	Remaining SHAP values:		70.025
Redshift prediction model (ET)			
Feature	SHAP value	Feature	SHAP value
xgboost	41.191	gbr	13.106
catboost	20.297	rf	11.648
	Remaining SHAP values:		13.758

[a] SHAP values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model and with different data. A meaningful comparison can be done by contrasting the order in which features are sorted.

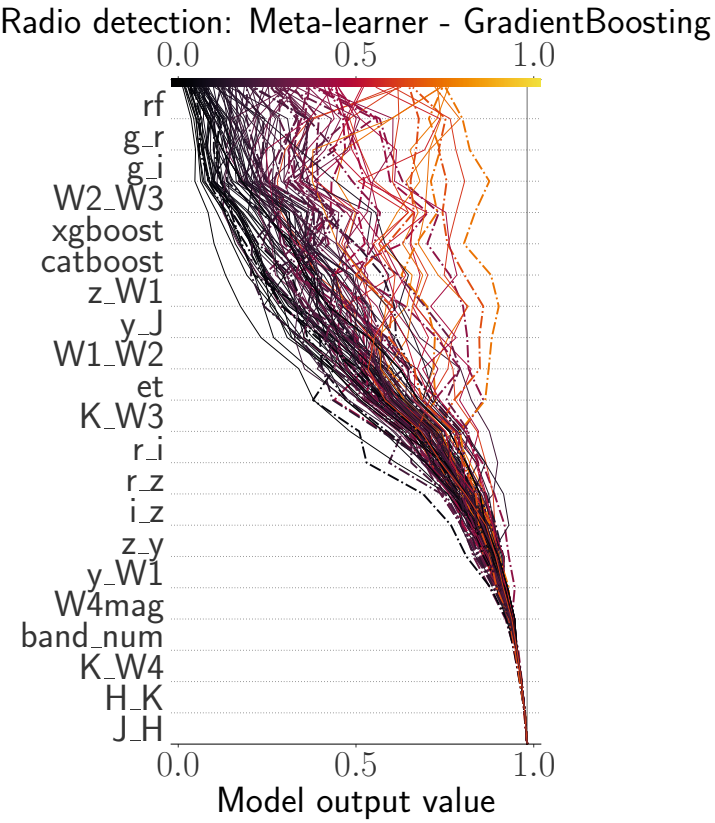


Figure 4.5: Decision plot from the SHAP values for all features from the radio detection model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 4.4.

model as it needs a higher input from the remaining features. One additional attribute of the meta AGN-SFG is the range in which the predictions are located. This issue is reflected in the narrow decision margin for the non-calibrated stacked model (see model output values –x-axis– close to ~ 0.5 in Fig. 4.6 and Sect. 3.6.2).

For the radio classification step of the pipeline, we find that features linked to those 121 high-z AGN perform at the same level as for the overall population (as shown by the global feature importances). As introduced in Sect. 3.7.2, radio-detection model shows difficulties when producing a classification based on the provided dataset. The improved metrics with respect to those obtained from the no-skill selection do indicate that the model has learned some connections between the data and the radio emission. The importances of the features have changed when compared to the overall population. Now, the colours (g - i) and ($W2$ - $W3$) are the most relevant for the model, whereas Table 4.4 presents ($W2$ - $W3$) and (y - Ks) as the most impactful quantities. If the radio emission observed from these sources were exclusively due to SF, this connection would imply SFR of several hundred $M_\odot\ yr^{-1}$. This explanation can not be completely ruled out from the model side but some contribution of radio emission from the AGN is expected.

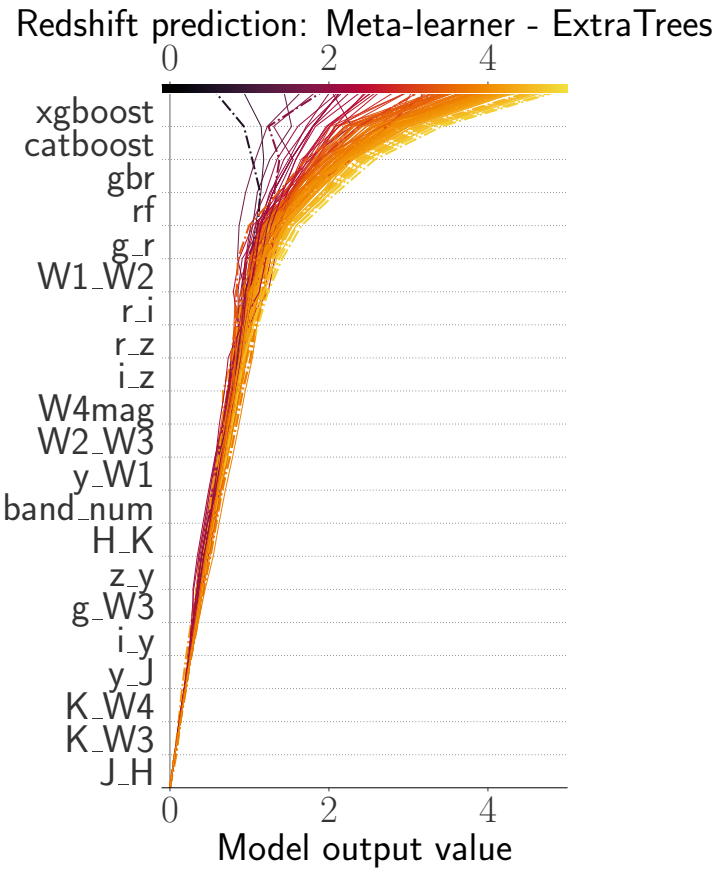


Figure 4.6: Decision plot from the SHAP values for all features from the redshift prediction model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 4.4.

Table 4.7: Combined and normalised (rescaled to add to 100) mean absolute SHAP values for observed features from the three models using 121 spectroscopically confirmed AGN at high redshift values ($z \geq 4$).

AGN-SFG model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
W1_W2	32.458	i_y	5.086	z_y	1.591
g_r	11.583	y_W1	4.639	H_W3	1.048
W1_W3	8.816	band_num	4.050	W4mag	0.514
r_i	7.457	y_W2	3.228	H_K	0.466
i_z	6.741	z_W2	2.348	W3_W4	0.466
r_J	6.613	y_J	1.718	J_H	0.178
Radio detection model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
g_i	14.120	z_W1	6.751	W4mag	2.691
W2_W3	13.201	r_i	5.577	band_num	2.661
g_r	12.955	r_z	5.161	K_W4	0.939
y_J	8.224	i_z	4.512	H_K	0.719
K_W3	7.441	z_y	4.121	J_H	0.190
W1_W2	6.874	y_W1	3.864		
Redshift prediction model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
g_r	32.594	z_y	3.557	W4mag	1.639
y_W1	20.770	y_J	3.010	g_W3	1.479
W2_W3	12.462	band_num	2.595	K_W3	0.853
W1_W2	5.692	i_y	2.381	K_W4	0.451
r_i	4.381	H_K	2.230	J_H	0.146
r_z	3.755	i_z	2.005		

For the redshift prediction model, Fig. 4.6 shows that, apart from the base models, (*W1* - *W2*) and (*g* - *r*), are the most relevant, individual features as they are immediately below the base learners in the decision plot. Focusing on Table 4.7, and when combininig resuls from each base model, (*g* - *r*) keeps its position as the most influential feature, followed, closely, by (*y* - *W1*) and (*W2* - *W3*).

Analysing the decision plots from each individual base model can give additional insight into the behaviour of single sources and how they relate with the rest of the selected sample (in this case, sources with $z \geq 4.0$). Base AGN-SFG classifiers (in Fig. 4.7) show a plethora of different behaviours. Initially, and in contrast with the decision plot from the meta learner, deviations from the bulk of the predictions start much lower in the list of features. This difference might imply that individual models need to rely on more observations than the meta model, which has that information encoded in each of the features from the base classifiers. Additionally, it is possible to see that some predictions change drastically when arriving to the last feature (in this case, (*W1* - *W2*)). Such changes highlight the importance of (*W1* - *W2*) as a decisive feature.

Similarly, Fig. 4.8 shows the decision plots for the individual radio detection base learners. In this case, most predictions arrive in the leftmost zone of the final scores, showing that base models struggle with deciding when a source could be detected in the radio.

For the redshift predictions in Fig. 4.9, most sources follow a similar prediction path. As the analysed sources have, originally, high redshift values, their predicted values tend to be in the same range. Such behaviour is an indication of the good predicting power of all base regressors.

While the application of the prediction pipeline and understanding its results and drivers are very relevant goals by themselves, it is possible to expand the knowledge brought by it. In this way, using the predicted probabilities, classes, and photometric redshift values delivered by each pipeline step, we can derive new tools for the analysis of AGN or obtain physical insight from the new set of predicted radio-AGN candidates. In the following chapter, we will attempt to use the information delivered by our models in different directions from that of the creation of source candidates.

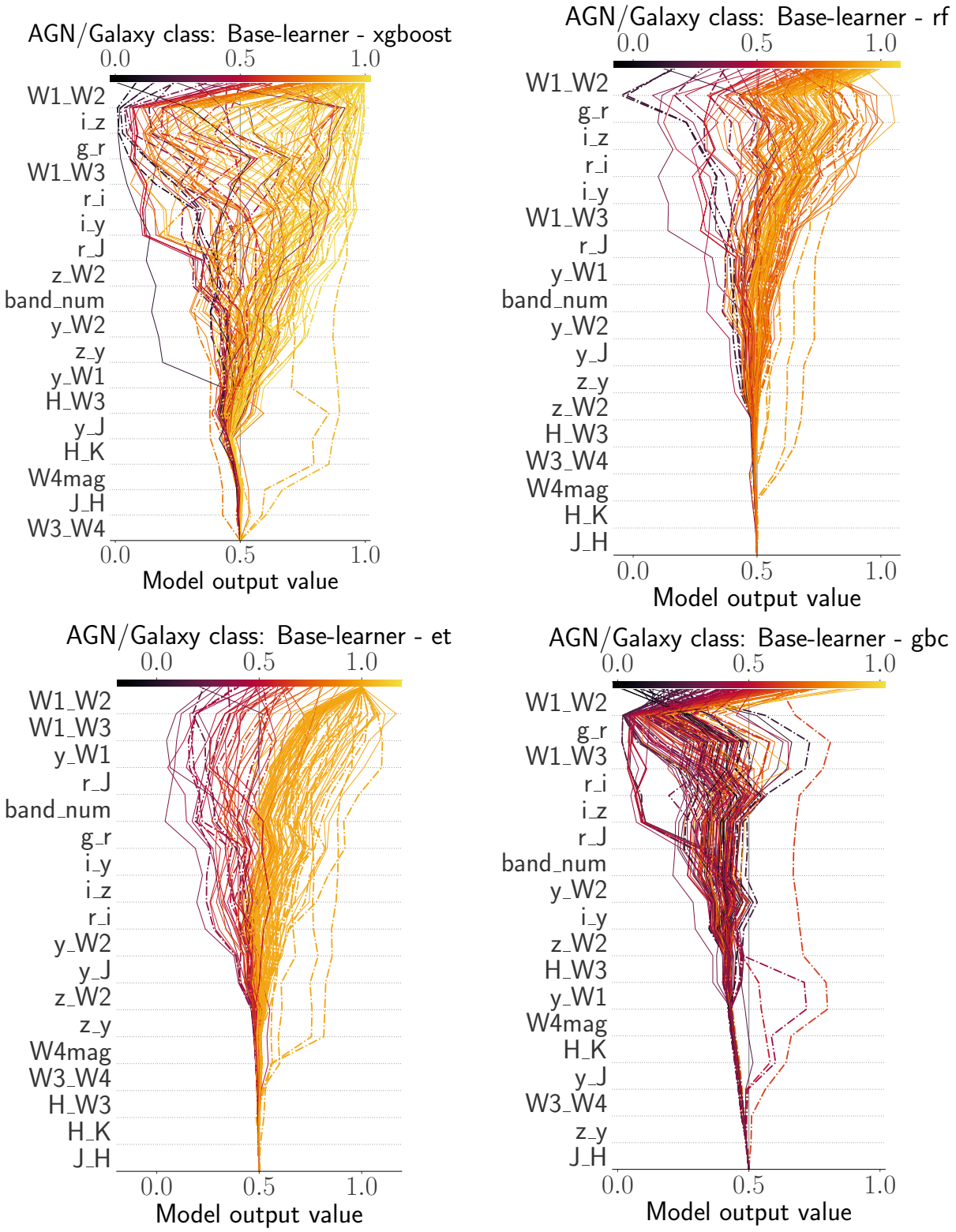


Figure 4.7: SHAP decision plots for base AGN-SFG algorithms. Details as described in Figs. 4.4. Starting point of predictions is the naive classification threshold. From left to right and from top to bottom, each panel shows the results from `XGBoost`, `RF`, `ET`, and `GBC`.

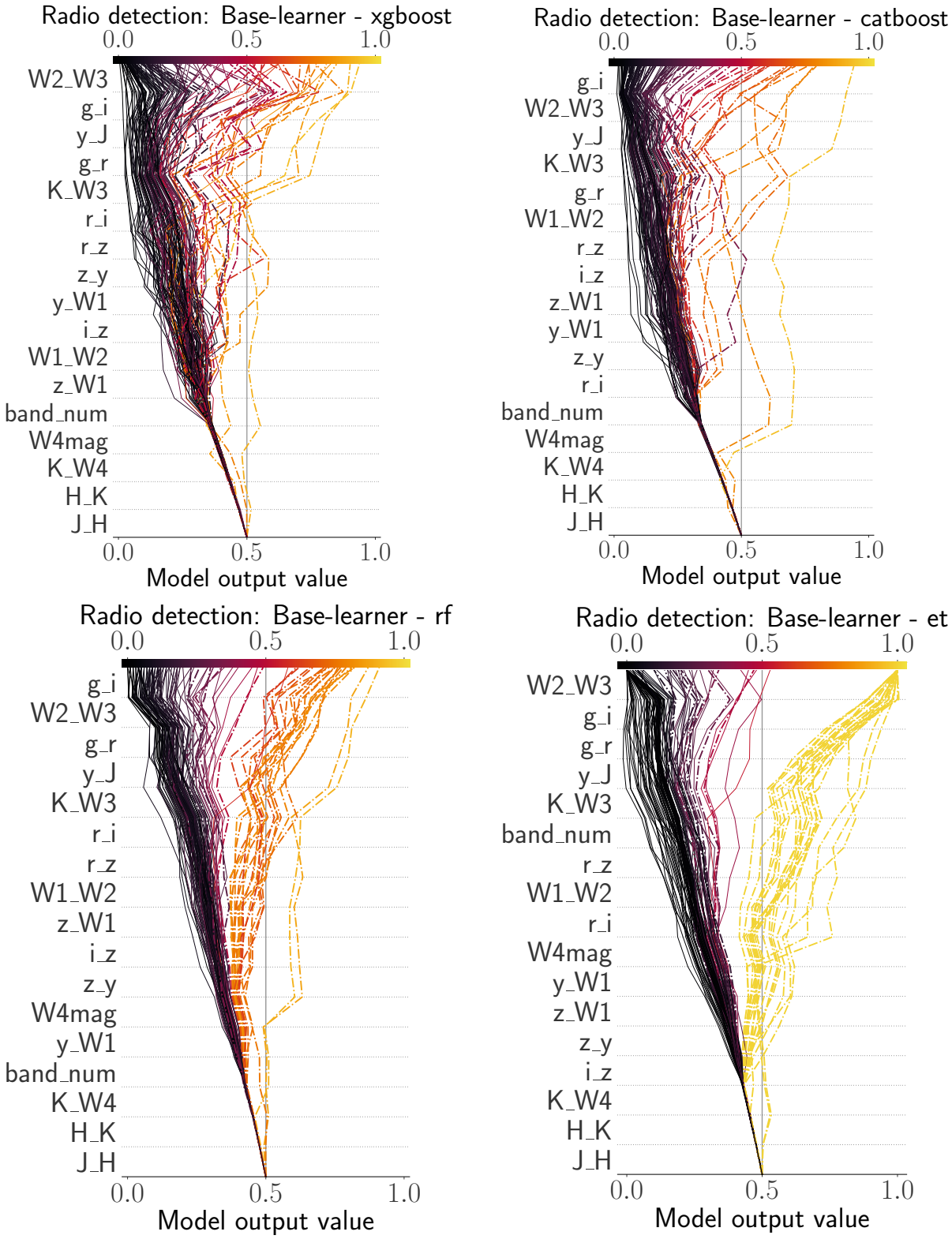


Figure 4.8: SHAP decision plots from base radio algorithms. Details as Figs. 4.4 and 4.7. Each panel with results for `XGBoost`, `CatBoost`, `RF`, and `ET`.

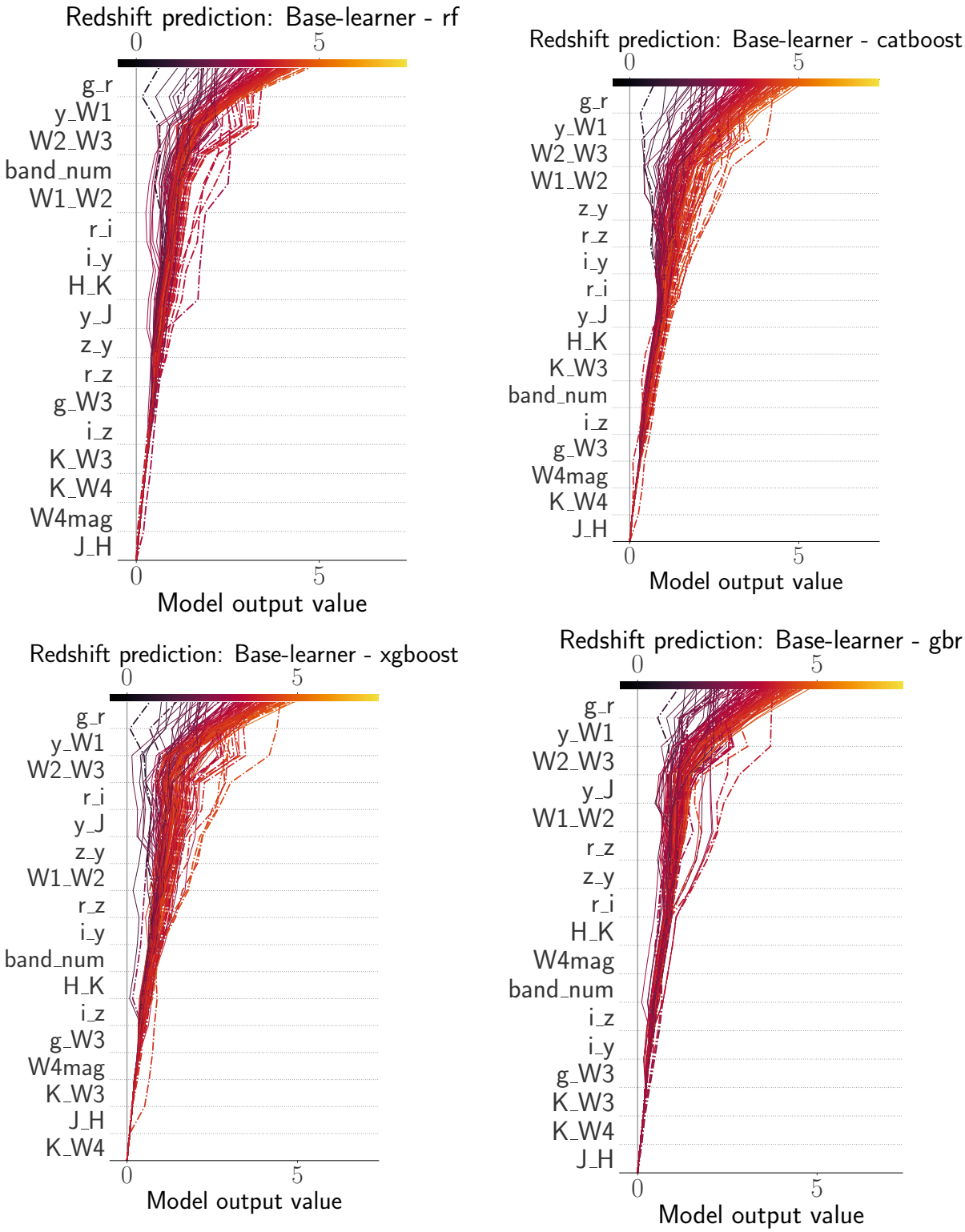


Figure 4.9: SHAP decision plots from base redshift algorithms. Details as in Fig 4.4. Each panel shows results for `ET`, `CatBoost`, `XGBoost`, and `GBR`.

This page intentionally left blank.

5

Machine-assisted learning from models: unlocking hidden knowledge

As shown in the previous chapters of this thesis, ML methods can be a powerful tool for astrophysical applications, but their potential extends beyond the predictions the models are trained for. Based on the fact that ML models and algorithms can uncover hidden connections within data, they expose valuable information about the studied sources. Such prospect couples with the modular design of our pipeline, where each model can analyse different aspects of the data, enhancing their insights. This chapter focuses on this exciting capability.

Here, we present exhibit the versatility of ML-based predictions through three compelling examples. First, we explore how the learning process of the models and their outputs can be used to define a new, and simple, AGN selection criterion. Next, we demonstrate how the outputs from the prediction pipeline can be directly applied to extract further knowledge from the distribution of radio-detectable AGN in different epochs in the form of a LF. Finally, we present the use of the results of the prediction pipeline for a different goal from its initial purpose: the assessment of multi-wavelength counterparts of radio detections.

This chapter opens the door to exciting new approaches in AGN research. The techniques examined here hold promise for unlocking hidden knowledge in large datasets that, otherwise, might remain unexplored. While only few examples are shown here, they show how ML predictions can be taken –and perhaps, should be taken– much beyond their original intended goals, providing substantial support to the analysis of large source populations.

5.1 Colour-colour AGN selection criterion

In Sect. 1.1.1, it was shown that a combination of photometric colours can be used to determine selection criteria for different subsets of AGN. Each one of these criteria is able to retrieve information on specific properties (e.g. redshift) or processes from AGN and their

hosts. Additionally, it is possible to extract details from the evolutionary state of some sources by observing their position in the colour-colour, or colour-magnitude diagrams.

Feature importances in Table 4.7 and the values presented in Fig. 2.3 suggest that the combination of particular colours might be highly efficient for the selection of AGN with performance comparable to existing methods (see Table 4.1) but for a much larger number of sources. While colour plots using *W3* offer access to 100 000 sources in our analysed datasets, 4 700 000 sources are available for colours based in optical bands (g, r, i or z) magnitudes (see Fig. 2.3 and Table 2.2). Thus, we want to explore the combination of colours that may be sensitive to specific AGN phenomena.

Leveraging the features with the highest SHAP values in the AGN-SFG classification, we tested this hypothesis and derived a selection criterion in the (g - r) vs (*W1* - *W2*) colour-colour parameter space as shown in Fig. 5.1 using the labelled sources in the test subset of the HETDEX field. These colours have the strongest potential to carry enough information to separate both populations. The results of the application of this criterion to the testing data and to the labelled sources in S82 are presented in the last row of Table 4.1. Replicating the creation of AGN wedges for their classification in MIR colours (e.g. M12, B18), we have defined the limits of an MIR-optical AGN wedge limited by the following expressions:

$$g - r > -0.76\,, \tag{5.1}$$

$$g - r < 1.80\,, \tag{5.2}$$

$$W1 - W2 > 0.227 \times (g - r) + 0.43\,, \tag{5.3}$$

where *W1*, *W2*, g, and r are Vega magnitudes. If using AB magnitudes, the expressions are:

$$g - r > -0.90\,, \tag{5.4}$$

$$g - r < 1.66\,, \tag{5.5}$$

$$W1 - W2 > 0.227 \times (g - r) - 0.18\,, \tag{5.6}$$

Our colour criterion, C23[1], provides homogeneous scores across the different metrics

[1]We call this criterion C23 as as it was first presented in Carvajal et al. (2023a).

with purity (precision) and completeness (recall) above 87 %. Avoiding the use of the longer *WISE* wavelength (*W3* and *W4*), the criteria can be applied to a much larger dataset (as seen in the number of missing values in Figs. 2.3 and 2.4). The boundaries of the MIR-optical wedge have been drawn to contain as many AGN as possible at the same time that the inclusion of the largest number of SFGs is avoided (i.e. maximising the AGN completeness while that of SFGs is minimised). In this way, our wedge recovers 92 % of the known AGN in the HETDEX test set (value equivalent to the recall, or completeness, of the selection method). From the application of the MIR-optical criterion, 88 % of the AGN in the S82 field are recovered.

A further visual analysis of the centroids of the distributions of AGN and SFGs in Fig. 5.1 shows that their separation is clearer in the (*W1* - *W2*) axis than in the (*g* - *r*) space. Such difference is linked to the local feature importances of Table 4.7, where the SHAP value of the first colour is almost three times that of the second. From a physical standpoint, the location of both centroids follows previous works. For optical colours, AGN tend to be bluer than SFGs (e.g. Ibata et al. 2017), while for MIR colours, it is known that AGN tend to be redder than SFGs (cf. Fig. 4.1), the latter having being shown in previous works (see, for instance, the description by Radcliffe et al. 2021a).

Even though each colour has been used separately for AGN diagnostics (e.g. Obrić et al. 2006; Assef et al. 2013; Yan et al. 2013; Secrest et al. 2015; Gatica et al. 2024), prior to the development of our pipeline, the combination of (*W1* - *W2*) and (*g* - *r*) colours has not been systematically explored for AGN-SFG separation.

One fairly recent example of the use of such colour combination is presented by Zeraatgari et al. (2024). They analysed the application of ML classfication for the separation of stars, AGN, emission-line galaxies, and normal galaxies using SDSS data release 17 (SDSS-DR17; Abdurro'uf et al. 2022) (bands *u*, *g*, *r*, *i*, and *z*) and *WISE* measurements (in bands *W1* and *W2*) of close to 1 500 000 sources in the SDSS-DR17 survey area. As part of their preparatory analysis, a (*W1* - *W2*) vs (*g* - *r*) colour-colour diagram (among others) was used to assess the usefulness of separate features to divide sources into classes. Their analysis concluded that combining several features can capture better the intricacies of each class for their separation. Notably, their global feature importance analysis also selected the colours (*W1* - *W2*) and (*g* - *r*) as one of the strongest feature combinations for the separation of sources into the four classes.

A different approach for the analysis of the use of MIR and optical colours was done by Daoutis et al. (2023). They developed an ML-based diagnostic tool for the selection and

classification of AGN and SFGs. Daoutis et al. (2023) used data from AW (bands *W1*, *W2*, and *W3*) and SDSS (bands *u*, *g*, and *r*) for over 40 000 sources within the SDSS footprint and created different models with combinations of colours from both samples. They concluded that their model is able to successfuly distinguish between extreme MIR starburst galaxies and obscured AGN by using only three colours: (*W1* - *W2*), (*W2* - *W3*), and (*g* - *r*).

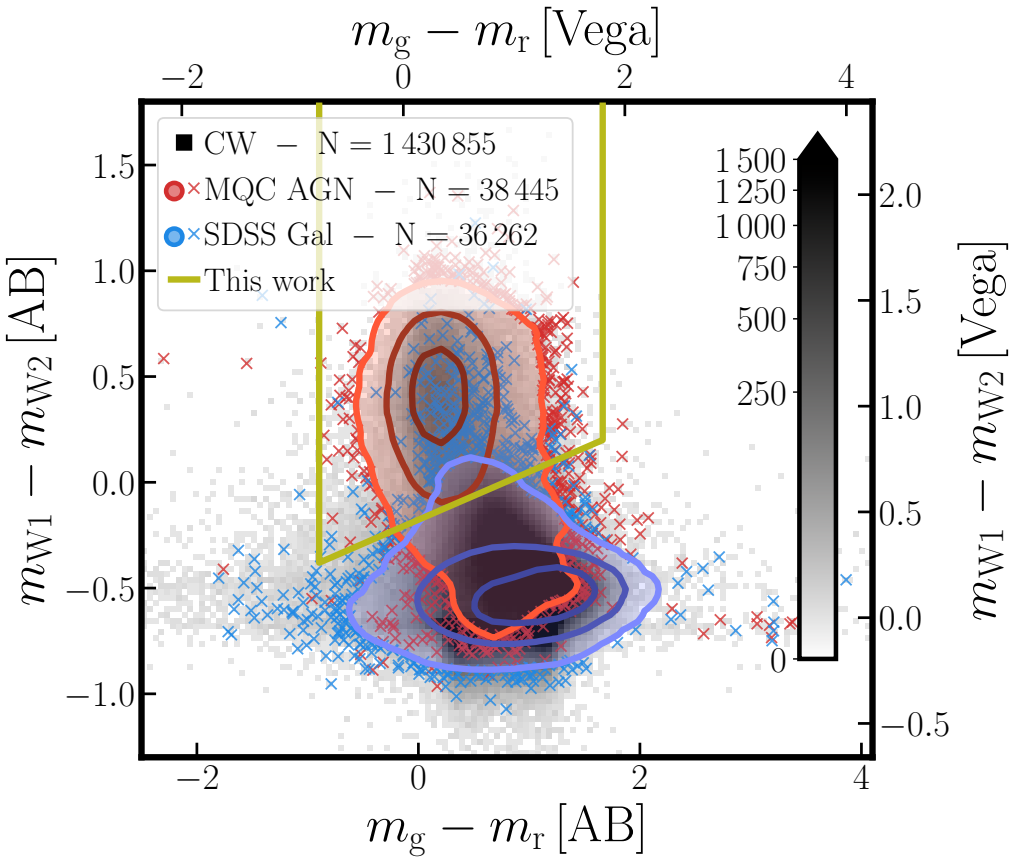


Figure 5.1: AGN classification colour-colour plot in the HETDEX field using CW (*W1* - *W2*) and PS1 (*g* - *r*) passbands. In the background, grey-scale, two-dimensional histogram includes all CW detected and non-imputed sources following colour-coding of colourbar. Red contours highlight the density distribution of the AGN in the MQC and blue contours show the density distribution for the SFGs from SDSS-DR16. Contours are located at 1 σ, 2 σ, and 3 σ levels and crosses show sources outside the selected contours (i.e. outliers, in the same colours as contours). Ochre line limits the selection criterion determined from the application of our prediction pipeline (C23, see Eqs. 5.4, 5.5, and 5.6).

Table 5.1 compares the performance of our new colour-colour criterion (C23) with previous (*W1* - *W2*) vs (*W2* - *W3*) AGN selection criteria (presented in Sect. 4.1.1 and Table 4.1) and our prediction pipeline. In both the HETDEX and S82 fields, C23 consistently recovers more AGN than previous methods, with differences ranging from 11 % to 55 %. This contrast highlights that the isolated use of MIR colours might be insufficient for efficient recovery of AGN and optical colours are needed to complement their selection. Additionally, the small differences between the metrics from using our full prediction pipeline and C23 suggests that, as shown in the feature importance analysis, most of the information that separates AGN from SFGs is

captured by the selected colours.

Table 5.1: Results of application of several colour-colour AGN diagnostics criteria to our testing subset and the labelled sources from the S82 field. Same as Table 4.1 but including colour-colour criteria from this work.

HETDEX test set				
Method[a]	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	86.10	78.78	93.98	80.51
M12	51.80	49.71	98.87	37.18
M16	67.21	61.30	97.48	53.48
B18	82.14	75.76	97.54	72.66
Our pipeline	95.42	91.85	94.49	96.21
C23[b]	92.71	87.64	94.00	91.67
S82 (labelled sources)				
Method[a]	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	83.59	45.47	93.93	76.62
M12	46.80	28.22	99.59	32.54
M16	64.69	37.76	98.80	50.32
B18	79.71	51.07	98.72	68.77
Our pipeline	94.37	70.67	94.81	94.01
C23[b]	90.63	58.53	94.15	87.91

[a] Naming codes for the used methods are described in the main text (cf. Sect. 4.1.1).
[b] Last row of each sub-table corresponds to the colour-colour criterion derived in this work.

Nevertheless, and understanding that the application of our newly derived C23 diagnostics criterion (or any other colour-colour criterion) can be a straightforward method for the separation between SFGs and AGN, we advocate for the use of the full prediction pipeline for the selection of (IR-detected) radio-AGN. Tables 4.1 and 5.1 show that the analysis of a full set of multi-wavelength information, in the form of our prediction pipeline, can deliver, overall, better results than those given by traditional colour-colour criteria in almost all the selected indicators.

Despite the increasing availability of large-area, multi-wavelength data from various surveys, not all bands offer the same level of detail (e.g. poorer spatial resolution and limited sensitivity). Future surveys aim to address this limitation. In the meantime, targeted color analysis, such as colour-colour diagnostics, can offer advantages.

Small regions with rich photometric coverage can be used to train full ML models for specific tasks, such as AGN redshift determination, separation of AGN from SFGs, and identification of members of galaxy clusters. The analysis of such models can identify key features

that can be translated into colour-colour selection criteria. These criteria can then be applied to vast regions with sparser data, allowing for efficient analysis across large regions of the sky.

5.2 Radio luminosity function

A full study of the evolution of the distribution of sources can be done with the use of LFs (cf. Sect. 1.1). As defined by Salpeter (1955), LFs represent the density of sources within a specific luminosity (or, equivalently, absolute magnitude) range, allowing us to quantify the distribution of radio-emitting AGN. This section leverages our established ML-based prediction pipeline to create a large-scale dataset of predicted AGN and SFGs. This larger dataset enables the derivation of a radio luminosity function (RLF) with superior statistical robustness due to the significantly increased number of sources compared to traditional methods that rely on smaller, observationally limited datasets.

In order to achieve such goal, we will utilise a modified version of our prediction pipeline to derive the RLF for these objects. We upgraded the data process and pipeline with two relevant changes and re-run, as a consequence, the full training sequence. The first change is related to the cross-match of the radio information for the training set. We increased the search radius used to find counterparts from 1″.1 (cf. Sect. 2.3) to 6″ to better match the spatial resolution of LoTSS-DR1. With this change, we ensure the selection of radio detections for a larger fraction of sources at the expense of possible misidentifications of radio counterparts. Furthermore, to gain a broader understanding of the radio luminosity distribution, we incorporated an additional branch into the prediction pipeline.

Taking the description of the pipeline of Figs. 1.9 and 3.1 as a base, we have extended our pipeline by including a branch linked to sources predicted as SFGs (i.e. not as AGN) by the first model. Thus, these sources will be subject to a prediction of their radio detectability and, those predicted as being radio detectable, will have their photometric redshift values estimated. The inclusion of a new branch is related to the need to compare the radio luminosity distributions between AGN and regular SFGs (see Chapter 1). The new configuration of our prediction pipeline is illustrated in Fig. 5.2, while a detailed description of the modified datasets and the models produced with them is presented in Appendix C.

In order to study the distribution of luminosities of AGN as derived from our prediction pipeline, and as part of the efforts to study the behaviour in areas that will be subject to future

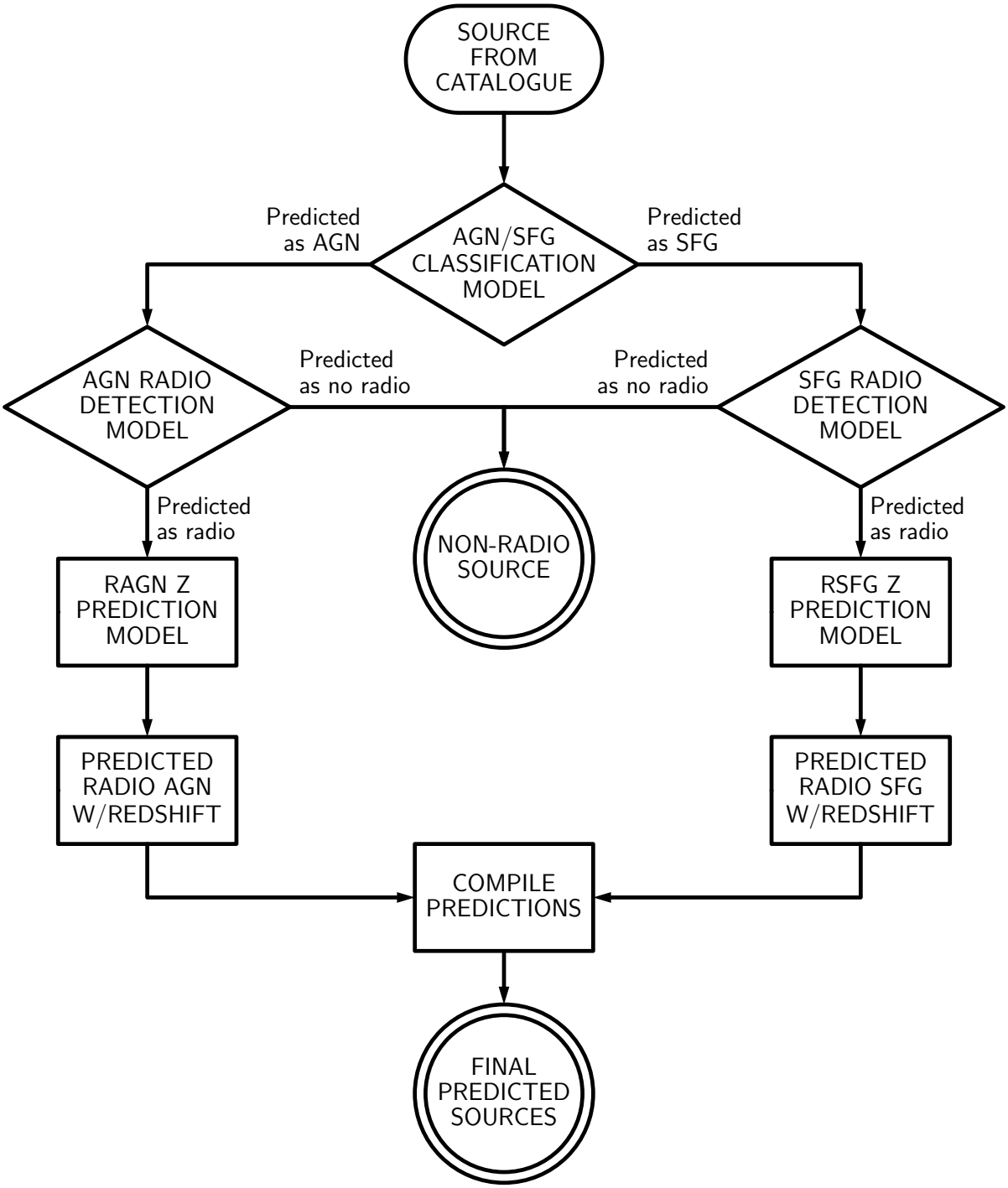


Figure 5.2: Flowchart representing the extended prediction pipeline used to predict the presence of radio-detected AGN and SFGs and their redshift values from IR-detected sources. Diamonds represent classification models and rectangles, regression model and intermediate data-collection steps. Double circles represent end states for the data in the pipeline.

radio surveys, we have selected the area of the EMU-PS as a test field, which is part of the EMU survey, a precursor of SKA. The EMU-PS catalogue[2] has radio information, at 944 MHz, from 178 921 compact sources in an area of 270 deg^2 in the southern sky (see Fig. 5.3 for a footprint of the area of EMU-PS) with a depth of 25 µJy/beam to 30 µJy/beam rms and a spatial resolution of 18″ (Norris et al. 2021). If we assume a synchrotron radio slope of $\alpha = -0.7$ (e.g. Sabater et al. 2019), the 5 σ detection limit of LoTSS (355 µJy/beam) at the frequency of EMU-PS would be $\approx$ 95 µJy/beam, which is below the 5 σ detection limit of EMU-PS, 125 µJy/beam. Thus, and assuming pure synchrotron emission, if LoTSS had observed the EMU-PS area, some of its fainter detections might have not been caught by the EMU-PS catalogue.

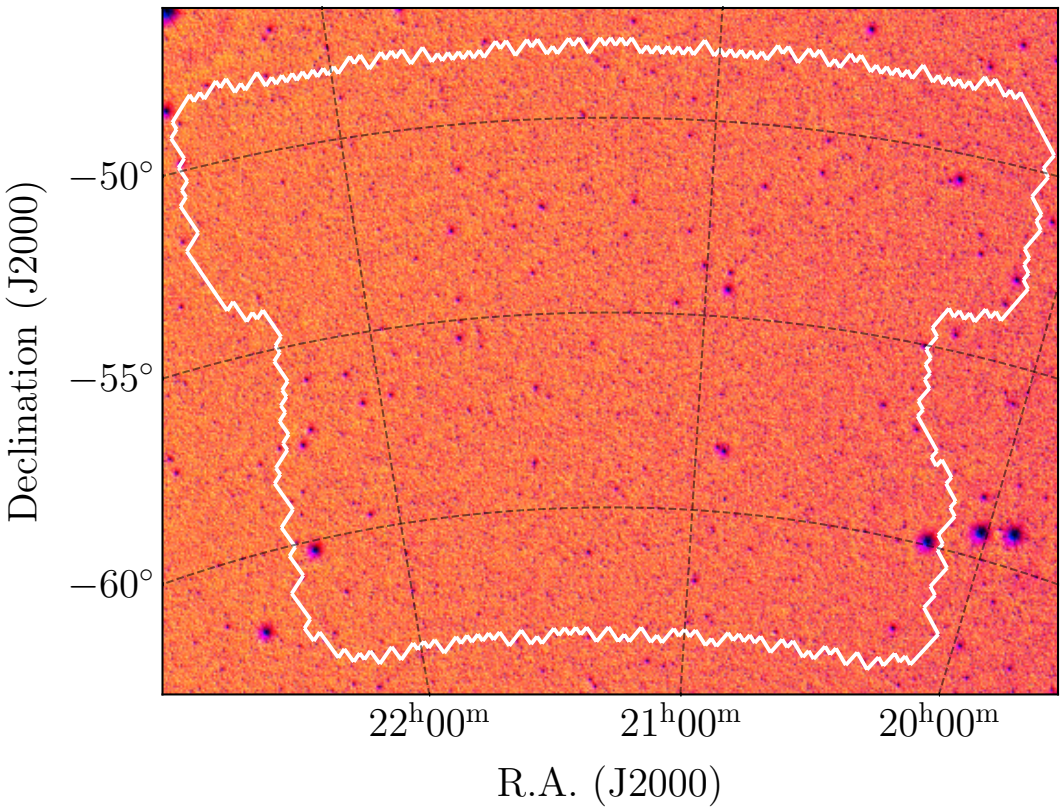


Figure 5.3: Footprint of the area of the EMU-PS field. In the background, *W1* image from the un*WISE* coadds (Lang 2014; Meisner et al. 2022). The white contours limit the area of the EMU-PS field, which covers 270 deg^2.

For the purposes of this exercise, and as presented in Sect. 2.3, we have collected measurements in the EMU-PS area to apply the prediction pipeline. Thus, we have started with the selection of CW-detected sources in the selected area, finding 10 355 457 detections and successive cross-matches were performed using a search radius of 1.″1. One exception to this distance was used for the cross-match with the EMU-PS catalogue itself, where a 10″ radius

[2]EMU-PS data can be obtained from https://doi.org/10.25919/exq5-t894

was used instead (exceeding the 6″ search radius used for the selection of radio counterparts in the training of the models), which is the maximum search radius used by Norris et al. (2021) to find CW counterparts for their radio detections. As quoted by Norris et al. (2021), the false cross-identification rate for a 10″ radius is ∼ 60 %, while a radius of 3″ leads to an ∼ 8 % false identification rate. The reason for increased radio search radius is twofold: a larger distance is similar to the size of the restoring beam (18″) and the need for testing the effectiveness of the use of our models for the assessment of radio counterparts. Additionally, we aim to apply corrections for misidentifications expecting to counteract the large false cross-identification rate. As aforementioned, an additional branch of two models has been included in the prediction pipeline. One that can predict the radio detectability of SFGs (i.e. not AGN) and a second model that can predict photometric redshift values for radio-detected SFGs.

Given that PS1 does not cover the EMU-PS area, a different survey was selected to obtain optical measurements of the studied sources. We retrieved the optical data from the DES data release 2 (DES-DR2; Abbott et al. 2021), which uses the same filters as PS1 (i.e. *g*, *r*, *i*, *z*, and *y*). We have used the same search radius, 1″.1, to obtain the counterparts for the CW-detected sources.

In contrast to the identification of sources in HETDEX and S82 fields and due to their different positions in the sky, the EMU-PS catalogue was cross-matched with alternative catalogues for the association of known AGN and SFGs. In the case of AGN, a more recent version of MQC (v8; Flesch 2023) was used[3] as well as QSO identifications from the spectroscopic sample of DES-DR2 (Yang and Shen 2023) and the *Gaia*–unWISE Spectroscopic Quasar catalog (Quaia G20.5; Storey-Fisher et al. 2024), which is based upon observations from *Gaia* data release 3 (DR3) extragalactic content (Gaia Collaboration et al. 2023a) and the *unWISE* reprocessing (Lang 2014; Meisner et al. 2019) of the *WISE* data. For the SFGs present in the EMU-PS field, and from the lack of SDSS measurements, we have included the identifications from the spectroscopic catalogue in the final VEXAS data release 2 (VEXAS-DR2; Khramtsov et al. 2021). In this way, the EMU-PS field harbours 12 649 known AGN and 1806 known SFGs, from which 2375 and 870, respectively, have a counterpart in the EMU-PS catalogue. As in the S82 field, the ratio of the number of AGN and SFGs in the EMU-PS field is the opposite to what can be found in the HETDEX field. Having a larger number of AGN than of SFGs reflects the

[3]For this exercise, and as done in the main model training, we classify a source as AGN or SFG only if it has a redshift value associated to it. This requirement allows us to fully compare the state of a source with the estimations from the prediction pipeline.

efforts made in each field to catalogue each class of sources and not the underlying density of sources of the region.

In order to have the largest possible sample of redshift measurements, we included those provided by MQC v8, Quaia G20.5, and spectroscopic redshifts from the Dark Energy Spectroscopic Instrument (DESI) imaging surveys (Dey et al. 2019; Zou et al. 2019), which are contained in the full EMU-PS catalogue. A summary of the number of sources and counterparts found in all different catalogues and surveys can be seen in Table 5.2 and a visual representation of the depths of the selected bands for this work is presented in Fig. 5.4.

Table 5.2: Composition of initial catalogue (sources detected by CW) and number of cross matches with additional surveys and catalogues in the area of the EMU-PS

Step	Survey	Number of sources
Base catalogue	CatWISE2020	10 355 457
Photometry cross-match	DES-DR2	7 091 485
	AllWISE	4 066 594
	2MASS	932 926
	EMU-PS (10″)	170 702
Source identification	MQC v8 AGN	471
	Quaia G20.5 (AGN)	12 491
	DES-DR2 (AGN)	46
	(Total AGN)	12 649
	VEXAS Spec V2 (SFGs)	1806

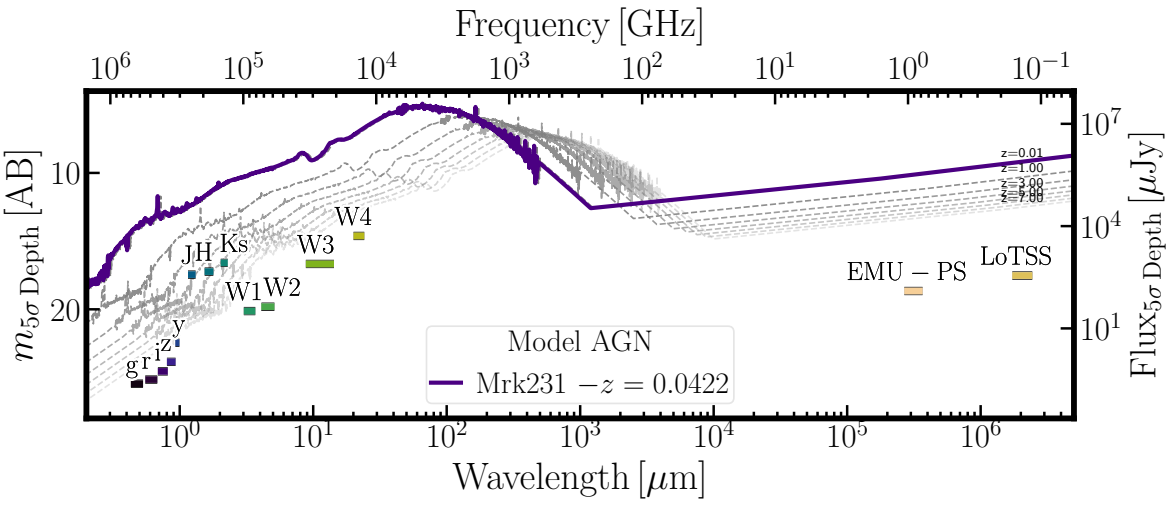


Figure 5.4: Flux and magnitude depths ($5\,\sigma$) from the surveys and bands covering the EMU-PS area. Limiting magnitudes and fluxes were obtained from the description of the surveys, as referenced in the main text. In purple, rest-frame SED from Mrk231 ($z = 0.0422$, Brown et al. 2019) is displayed as an example AGN. Redshifted ($z = 0.01,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,$ and 7) versions of this SED are shown in dashed grey lines.

The application of the modified prediction pipeline to the 10 355 457 IR-detected sources

in the EMU-PS area creates 92 113 candidates to be radio-detectable AGN and 128 249 to be radio-detectable SFGs. Among those radio-detectable candidates, 37 711 have a radio counterpart in the EMU-PS catalogue (15 484 predicted AGN and 22 227 predicted SFGs), that is, a measured radio flux. These numbers imply a 3878 % and 12 825 % (i.e. close to three and fourteen thousand times) increment of sources for radio-detected AGN and radio-detected SFGs respectively from the spectroscopically confirmed sources. The distribution of predicted photometric redshifts of both samples is depicted in Fig 5.5.

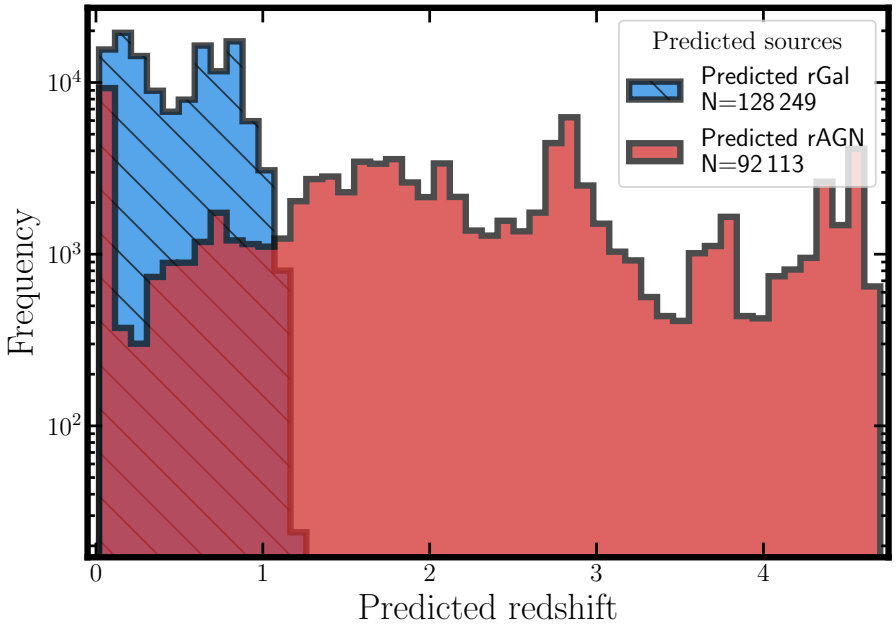


Figure 5.5: Distribution of predicted photometric redshift values for, in hatched blue, predicted radio-detectable SFGs and, in red, predicted radio-detectable AGN in the area of the EMU-PS catalogue.

It can be seen in Fig 5.5 that the distribution of radio-detectable SFGs is concentrated between redshift 0 to 1.2. This behaviour corresponds to the original distribution of SFGs used for training in the HETDEX field and thus, to the parameter space coverage of such sources. As expected, the model trained with them can only associate new sources to the values in that region of the space of parameters. For the same reason, the distribution of predicted redshifts for radio-detectable AGN spans a larger range, similar to the values of the training in the HETDEX field.

For the sources in the EMU-PS catalogue with confirmed counterparts in the CW survey (170 702, as indicated in Table 5.2), we can analyse the performance of our prediction pipeline. From the 170 702 CW-EMU sources, 36 324 have been predicted to be AGN and 134 376 to be SFGs. Among those, 15 484 elements have been predicted to be radio-AGN and 22 227 to be radio-SFGs. Thus, close to 80 % of EMU-PS sources have been predicted not to have radio

detections.

Most works on RLFs have used luminosity values at 1.4 GHz (e.g. Mauch and Sadler 2007; Simpson et al. 2012; McAlpine et al. 2013; Šlaus et al. 2020) as it can trace neutral hydrogen, H I, which can help in the determination of, among other quantities, the properties of the large-scale structure (e.g. Scott and Rees 1990). For meaningful comparison with previous studies, we will calculate EMU luminosities at that frequency. Rest-frame 1.4 GHz radio luminosities can be obtained using Eq. A.4 with EMU-PS fluxes (their distribution shown in Fig 5.6) and predicted photometric redshifts. We assumed a radio spectral index of $\alpha = -0.7$ (see, for instance, Simpson et al. 2012; Magliocchetti et al. 2014; Šlaus et al. 2020; Mandal et al. 2021; van der Vlugt et al. 2022; Lyu et al. 2022). In order to obtain the luminosity distances for the sources, we have adopted a flat Λ cold dark matter (ΛCDM) cosmology, with $\Omega_m = 0.31$, $\Omega_\Lambda = 0.69$, and $H_0 = 67.7\,\mathrm{km\,s^{-1}Mpc^{-1}}$, as presented by the Planck Collaboration et al. (2020)[4]. The distribution of such luminosities, as a function of predicted photometric redshift, is presented for both samples, predicted AGN and SFGs, in Fig. 5.7.

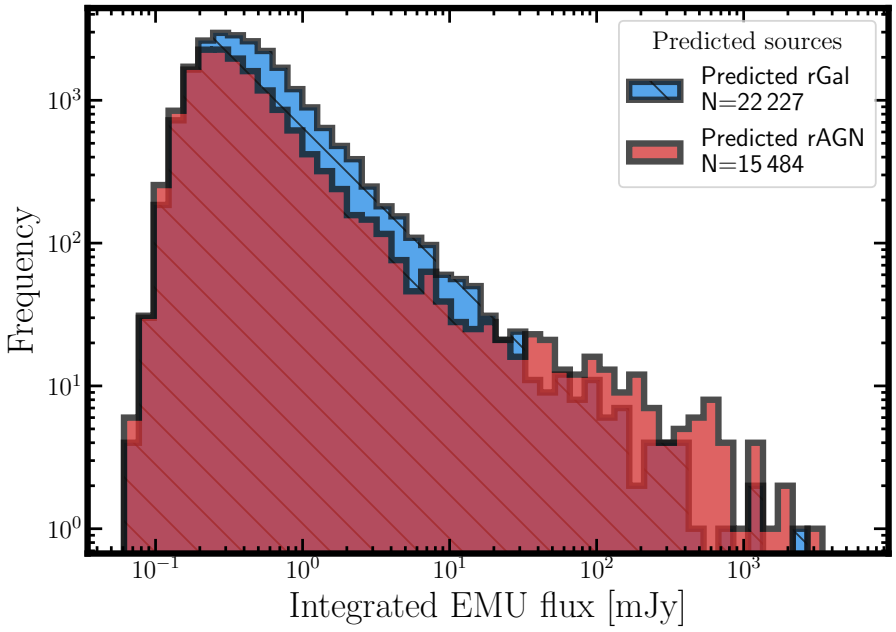


Figure 5.6: Distribution of EMU integrated fluxes (at 944 MHz) for, in hatched blue, predicted radio-detectable SFGs and, in red, predicted radio-detectable AGN in the area of the EMU-PS catalogue.

Taking into consideration the intrinsic uncertainties of the original pipeline and the changes introduced to the models in the extended prediction pipeline, the luminosity distributions of Fig. 5.7 appear to be relatively homogeneous. Two potential issues can be, nevertheless, identified. First, a fraction of luminosities are located below the 5 σ detection limit of EMU-PS.

[4]Values taken from their Table 2 under the column TT, TE, EE + lowE + lensing + BAO.

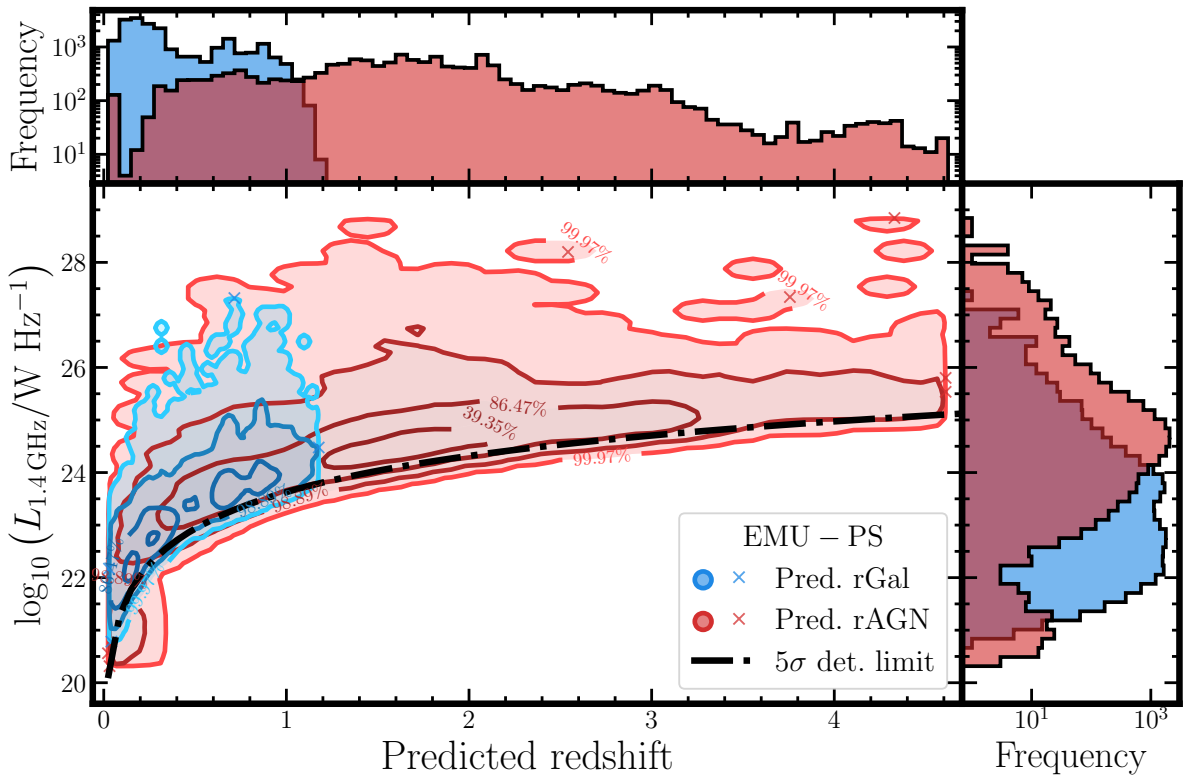


Figure 5.7: Distribution of predicted 1.4 GHz luminosities vs predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable SFGs (blue) in the area of the EMU-PS catalogue. In the top panel, the histograms of the predicted redshifts for both populations are presented. In the right-hand side of the figure, the distributions of predicted 1.4 GHz luminosities are displayed following the color code of the legend. In the central panel, two contour plots represent the joint distribution of 1.4 GHz luminosities and predicted photometric redshifts of both predicted radio-detectable AGN and radio-detectable SFGs. Contours represent the $1\,\sigma$, $2\,\sigma$, $3\,\sigma$, and $4\,\sigma$ two-dimensional levels of the distribution (39.35 %, 86.47 %, 98.89 %, and 99.97 %, respectively, of the corresponding sample) and crosses show sources outside the selected contours (i.e. outliers). Black, dashed line traces the luminosities associated to 25 m Jy, the $5\,\sigma$ detection limit of the EMU-PS catalogue.

The presence of these populations (both SFGs and AGN) can be explained by two interrelated factors. One of them is that the detection limit presented in Fig. 5.7 has been calculated with the mean detection depth of the survey (25 μJy/beam) and the other issue is that the detection limit used the fixed spectral index $\alpha = -0.7$, while the detected sources have been analysed with their own spectral indices.

Recalling the overview given in Sect. 1.1.1, an additional issue is related to sources that are predicted to be SFGs, but have radio luminosities that are too bright to be explained, only, by SF episodes. Following, for instance, the results from Magliocchetti et al. (2014), which have been based upon the work of Magliocchetti et al. (2002), Mauch and Sadler (2007), and McAlpine et al. (2013), a threshold, $L_{\rm cross}$, can be defined as a function of the redshift values of the sources. The expression can be written as follows:

$$\log_{10}\left(L_{\rm cross}\right) = \log_{10}\left(L_{0,\rm cross}\right) + z\,, \tag{5.7}$$

in which $L_{0,\rm cross} = 5.01 \times 10^{21}\,\mathrm{W\,Hz^{-1}\,sr^{-1}}$ for $z \leq 1.8$, and $L_{0,\rm cross} = 3.16 \times 10^{23}\,\mathrm{W\,Hz^{-1}\,sr^{-1}}$ for $z > 1.8$. After the addition of the angular factor 4π, these values can be expressed as $L_{0,\rm cross} = 6.3 \times 10^{22}\,\mathrm{W\,Hz^{-1}}$ for $z \leq 1.8$ and $L_{0,\rm cross} = 3.97 \times 10^{24}\,\mathrm{W\,Hz^{-1}}$ for $z > 1.8$. As a way to reduce this potential problem, we decided to use the thresholds of Eq. 5.7 and its description in our EMU-PS sample to alleviate the existence of too-bright radio SFGs. Thus, all predicted radio SFGs that presented a 1.4 GHz luminosity above the mentioned limit were re-labelled as predicted AGN and the corresponding branch of the prediction pipeline is applied to them. That is, their probability of being radio-detected AGN is obtained and, for those predicted to be radio-AGN, a new photometric redshift value is predicted. The modified distributions of luminosities and redshifts of the re-labelled AGN and SFGs sets are presented in Fig. 5.8 with 24 041 and 13 275 predicted and EMU-PS-detected radio-AGN and radio-SFGs respectively.

To further expand on the change of label and re-introduction to the prediction pipeline, but in the AGN branch, of some sources originally predicted as radio-SFGs, Fig. 5.9 presents the distribution of predicted redshifts before and after the re-labelling. There are 8952 predicted SFGs with high 1.4 GHz luminosities (a 40 % of the initial number of sources) that have been re-assigned as AGN. And 8557 of these sources (which are equivalent to a 55 % of the initial number of radio-AGN) are predicted to be detected in the radio and, therefore, have a new redshift value estimated. Most of the removed SFGs are located in the $z_{\rm Predicted} \gtrsim 0.5$ range, and are then translated into $0.5 \gtrsim z_{\rm Predicted} \gtrsim 1.5$ AGN. Thus, the impact of this correction is

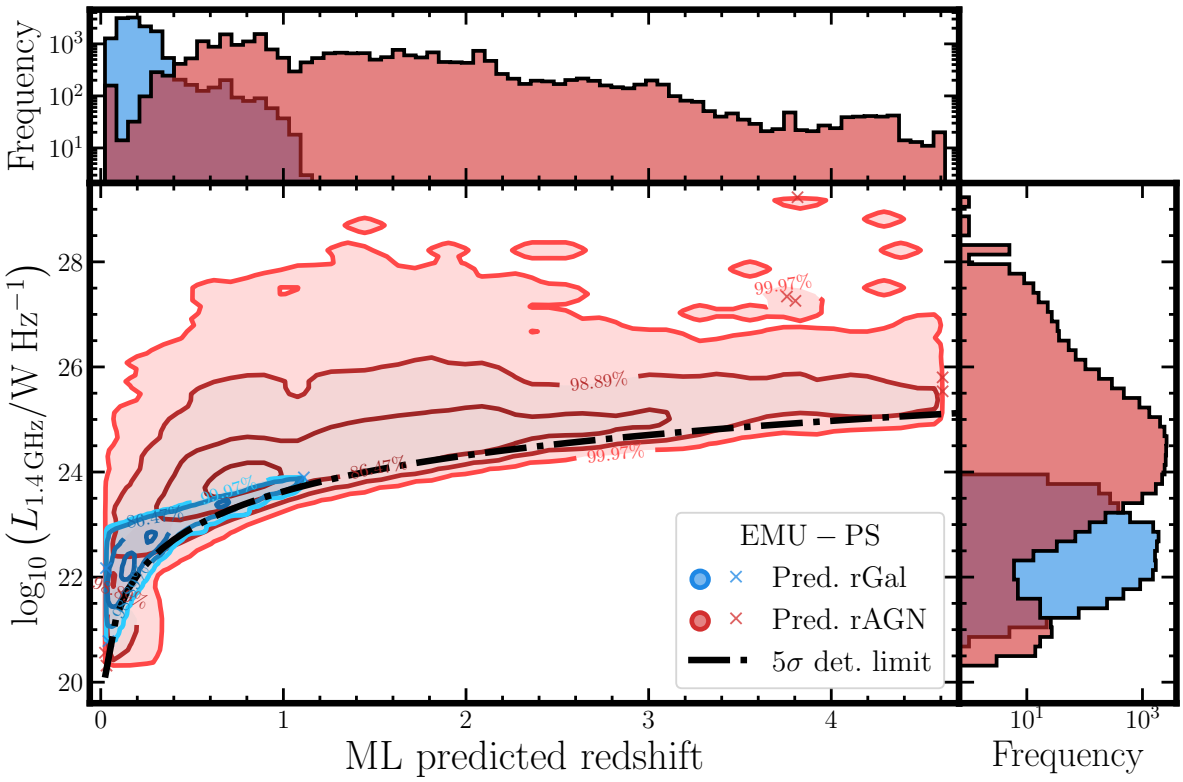


Figure 5.8: Distribution of predicted 1.4 GHz luminosities vs predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable SFGs (blue) in the area of the EMU-PS catalogue. SFGs with high 1.4 GHz luminosities have been re-labelled as AGN, following the prescriptions by Magliocchetti et al. (2014) and Magliocchetti (2022). Description as in Fig. 5.7.

limited to a fraction of the full distribution of predicted photometric redshifts.

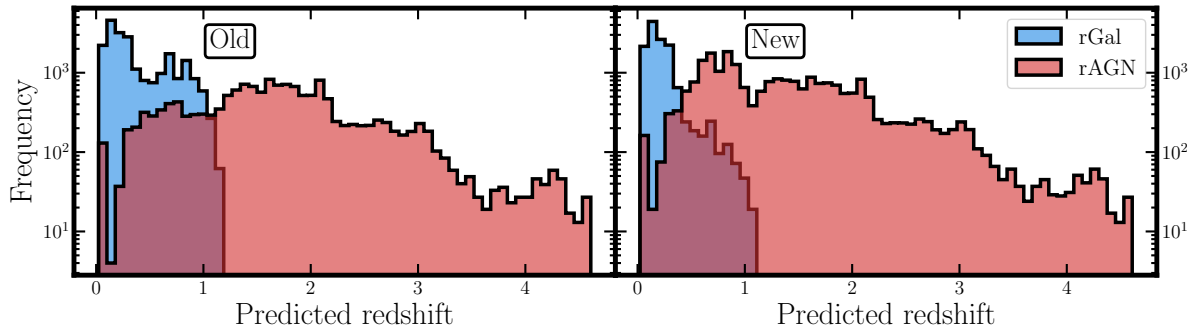


Figure 5.9: Histograms with distributions of predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable SFGs (blue) in the area of the EMU-PS catalogue before (left panel) and after (right panel) re-labelling 1.4 GHz-bright predicted SFGs as AGN.

In order to obtain RLFs for our predicted radio-detected sources, and from the description in Sect. 1.1.1, we implemented a version of the determination method presented by Page and Carrera (2000), which is called a binned estimator. This method, that is described in detail in Appendix A, is constructed in order to obtain an average of the LF over L and z bins (as described by Alqasim and Page 2023).

Taking advantage of the use of ML predictions, it is possible to obtain a correction, as a function of redshift and radio luminosity, for both completeness (recall) and purity (precision) of the radio-AGN and radio-SFGs in our selected sample that can be added to the selection function, $\mathcal{P}(z, L)$. In order to obtain these values, we took all known radio-detected sources (i.e. AGN and SFGs) in the HETDEX field as well as their predicted class, redshift values, and estimated 1.4 GHz luminosities (assuming a spectral index $\alpha = -0.7$; Sabater et al. 2019). For each element in the $\left(z_{\text{Predicted}}, \log_{10}\left(L_{1.4\,\text{GHz}}\right)\right)$ plane, all the sources located within a dimensionless distance of 1 are selected (i.e. their nearest neighbours). For these subsets of known radio sources, the class recall (cf. Eq. 3.4) and precision (cf. Eq. 3.5) are calculated independently. Their two-dimensional distributions of values are presented in Figs. 5.10 and 5.11 as a function of redshift and radio luminosity.

Then, each of the predicted sources from the EMU-PS catalogue is placed in the same plane of their corresponding class among the HETDEX sources. The ten closest HETDEX sources, with a Euclidean distance, are selected and their recall, or precision, values are averaged with the inverse of their distance to the EMU-PS source as weights. This averaged value is assigned to the predicted source as their estimated recall or precision. The two-dimensional distributions of such values are presented in Figs. 5.12 and 5.13. This method for the calculation of LF

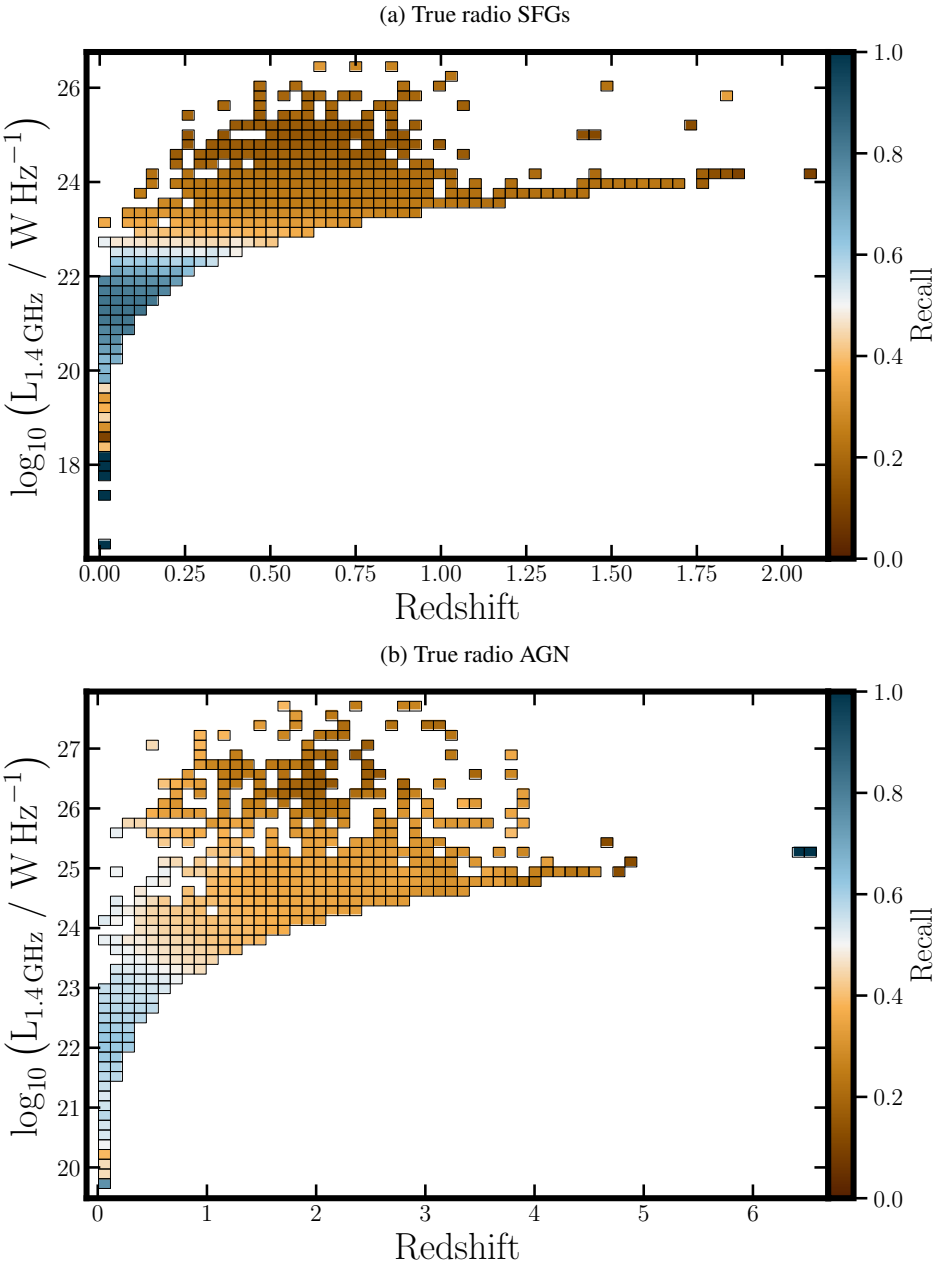


Figure 5.10: Two-dimensional binned distribution of true 1.4 GHz luminosity vs redshift for (a) radio SFGs and (b) radio AGN in the HETDEX catalogue. Sources in the 1.4 GHz luminosity-redshift plane have been divided into equal-sized bins and bins have been coloured according to the mean value of the estimated recall values of the sources in the bin and following each individual colourbar.

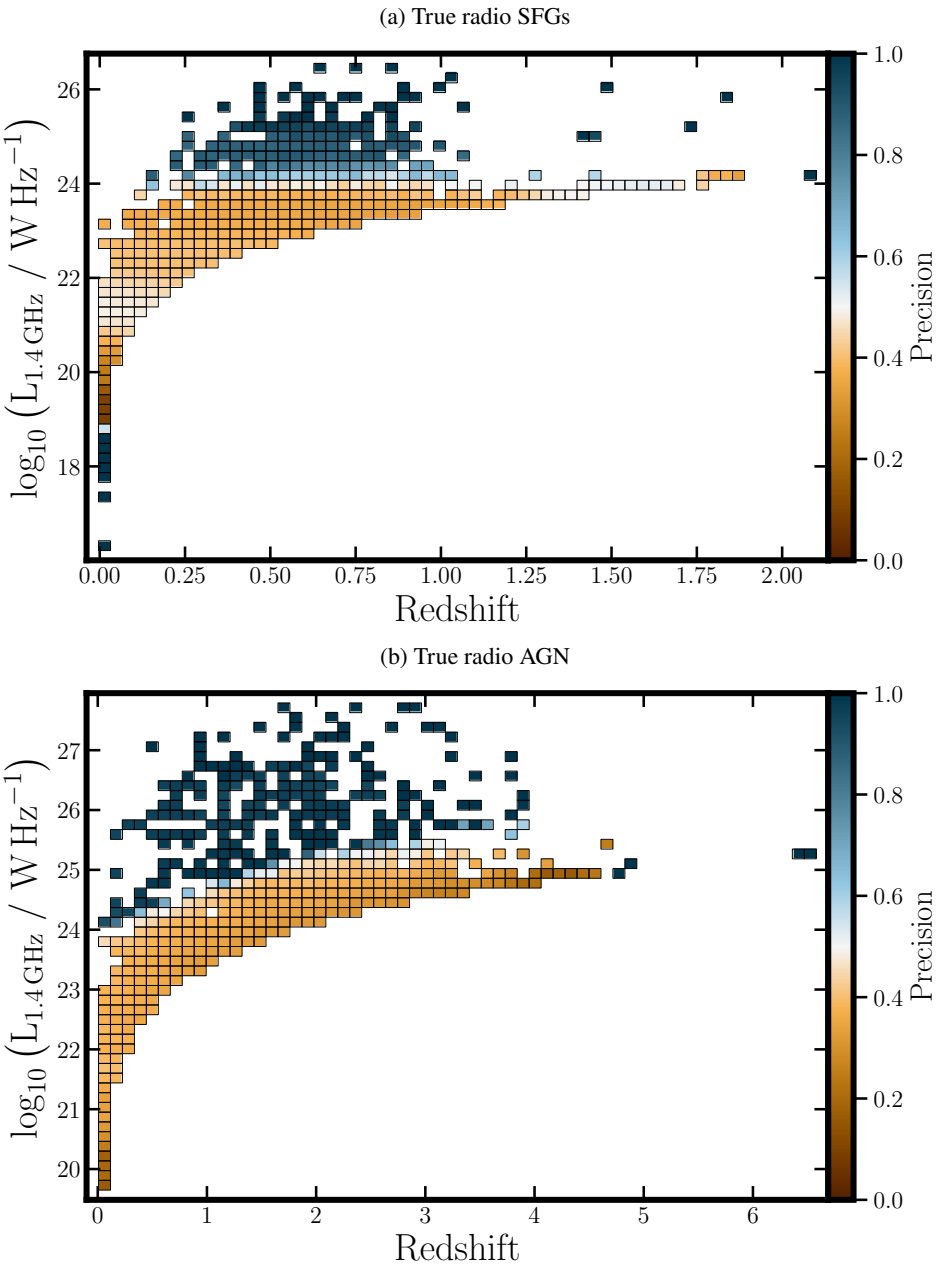


Figure 5.11: Two-dimensional binned distribution of true 1.4 GHz luminosity vs redshift for (a) radio SFGs and (b) radio AGN in the HETDEX catalogue. Following the description of Fig. 5.10, bins are coloured according to the mean estimated precision value of the sources included in each bin and following each individual colourbar.

correction factors takes, one step further, what has been usually the procedure of binning the (z, L) plane into regions that contain a reasonable amount of sources for the calculation of, for instance, recall values (e.g. Richards et al. 2006; Cameron and Driver 2007; Šlaus et al. 2020; van der Vlugt et al. 2022). In our case, and making use of the large number of AGN and SFGs produced by our pipeline, we can obtain an estimate of recall and precision for each individual source rather than for groups of sources.

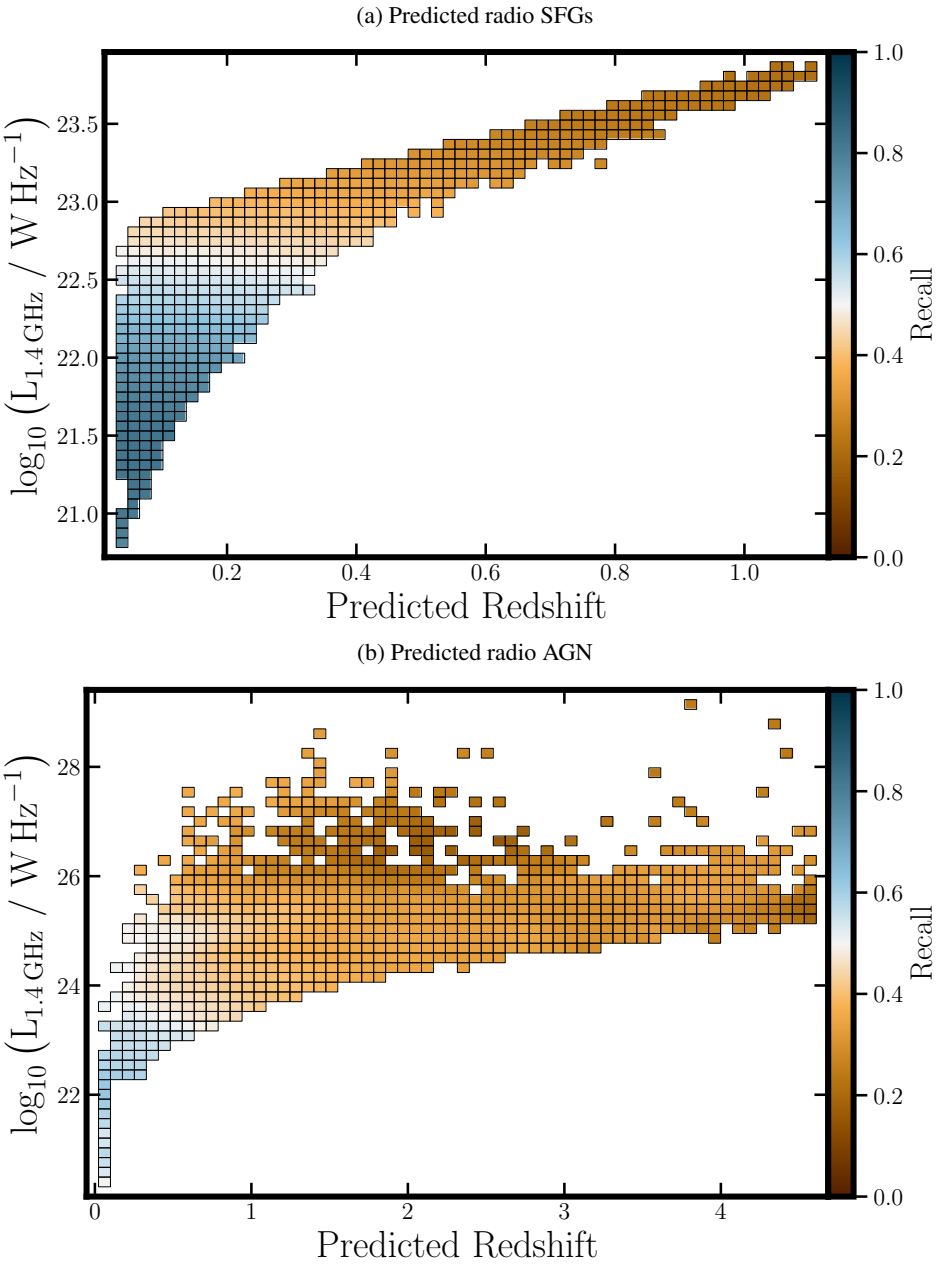


Figure 5.12: Two-dimensional binned distribution of predicted 1.4 GHz luminosity vs predicted redshift for predicted (a) radio SFGs and (b) radio AGN in the EMU-PS catalogue. Following the description of Fig. 5.10, bins are coloured according to the mean estimated recall value of the sources included in each bin and following each individual colourbar.

Having obtained the completeness and purity values for the predicted sources among the

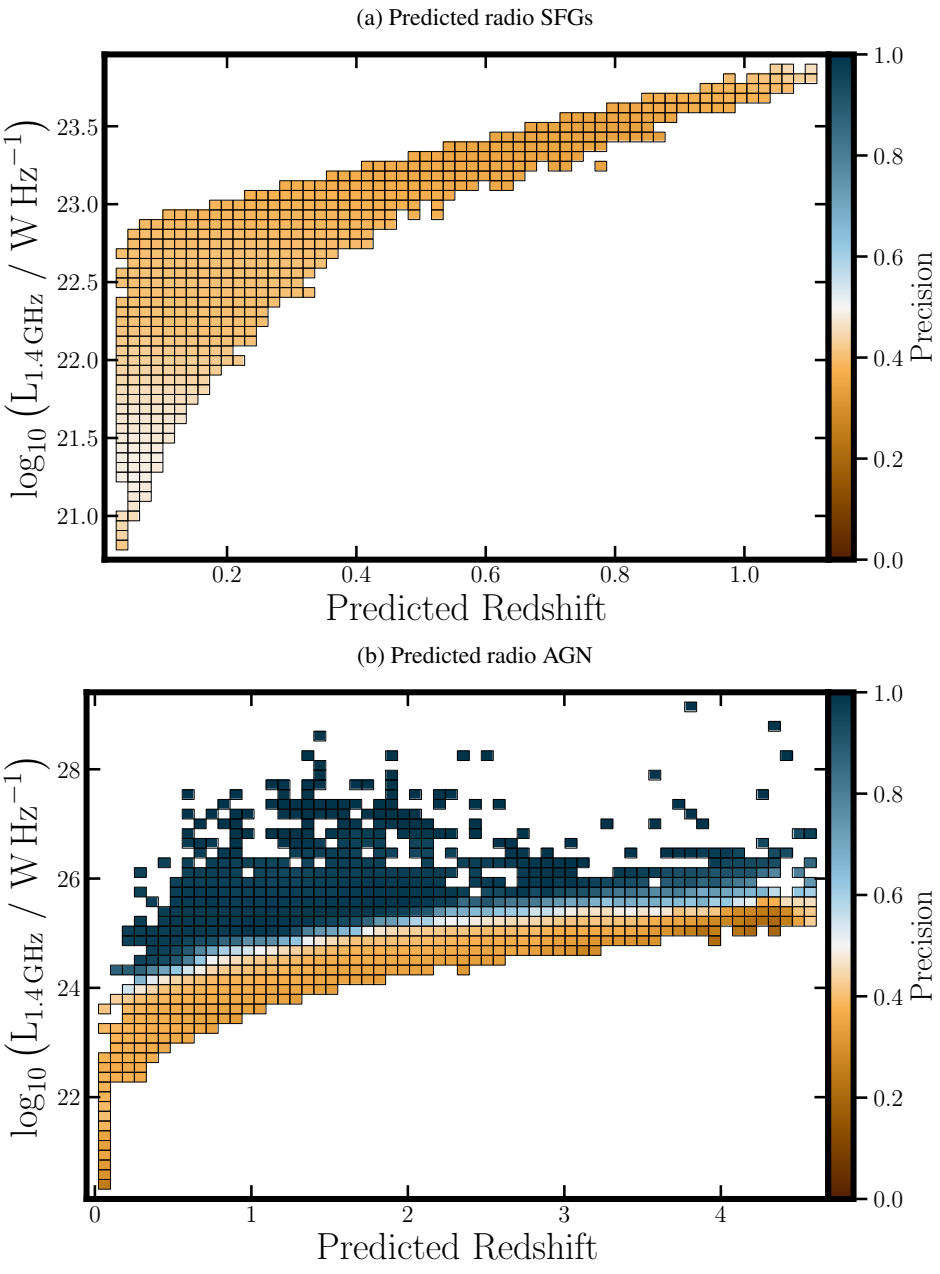


Figure 5.13: Two-dimensional binned distribution of redicted 1.4 GHz luminosity vs predicted redshift for predicted (a) radio SFGs and (b) radio AGN in the EMU-PS catalogue. Following the description of Fig. 5.10, bins are coloured according to the mean estimated precision value of the sources included in each bin and following each individual colourbar.

EMU-PS sources, it is possible to generate an initial definition for the selection function. First, and from the definitions in Sect. 3.1.1, the recall (completeness) indicates the fraction of sources that a prediction is missing given that they are present. Conversely, the purity (reliability) is related to, given the predicted elements, how many of them are not part of the desired class. Any correction based on those quantities will use the completeness as a reducing factor and the purity as an increasing quantity. In this way, the selection function for each source of our prediction, $\mathcal{P}(z_i, L_i)$, can be written as:

$$\mathcal{P}(z_i, L_i) = \frac{\text{Recall}_{\text{ML}}\,(z_i, L_i)}{\text{Precision}_{\text{ML}}\,(z_i, L_i)}\,, \tag{5.8}$$

with the ML-based recall (completeness) and precision (purity) ranging between 0 and 1. Then, the selection function values can be any non-negative real quantity. It can be seen, then, as the recall decreases, the selection function follows the same behaviour, while a decrease of the precision will imply an increment of the selection function values. Note that additional factors might be included in the definition of the selection function. For instance, some of them involve a correction for the resolution bias[5] (e.g. Prandoni et al. 2001, 2018; Mandal et al. 2021) and a correction for the Eddington bias[6] (Eddington 1913, 1940). For our project, and since it is more a proof of concept rather than a full analysis, only a ML-based correction (i.e. precision and recall) will be applied to the studied sources.

Taking into account the distribution of predicted redshifts from our set of predicted radio-detected AGN and the search for a distribution of sources that has a smooth number evolution across redshifts and radio luminosities (as done, for example, with radio-detected SFGs by van der Vlugt et al. 2022), 12 redshift bins were created: $(\,0.01, 0.08\,]$, $(\,0.08, 0.25\,]$, $(\,0.25, 0.35\,]$, $(\,0.35, 0.5\,]$, $(\,0.5, 0.7\,]$, $(\,0.7, 1.0\,]$, $(\,1.0, 1.3\,]$, $(\,1.3, 1.6\,]$, $(\,1.6, 2.0\,]$, $(\,2.0, 2.5\,]$, $(\,2.5, 3.2\,]$, and $(\,3.2, 4.8\,]$. In particular, we seek to have a statistically significant number of sources ($\gtrsim 100$) in each bin. Radio-AGN are distributed in all redshift bins while, in the case of radio-detected SFGs, the first seven redshift bins contain sources of this type (i.e. up to $z = 1.30$). The number of sources in each redshift bin is shown in Table 5.3.

We will obtain LF values from the use of binned estimations in each predicted redshift range with a chosen bin of size $\Delta \log_{10} L = 0.3$ following, for instance, Ross et al. (2013),

[5]Resolution bias occurs because extended sources are more easily hidden by random noise fluctuations compared to smaller sources. These fluctuations can lead to surveys underestimating the number of extended objects present.

[6]Eddington bias arises from errors in measuring faint objects. These errors can make them appear brighter than they truly are, leading to an overestimation of faint objects in high-luminosity bins of the LF.

Table 5.3: Number of predicted radio-detectable sources (AGN and SFGs) in the EMU-PS area by predicted resdhift bin. For radio-AGN and radio-SFGs, first column shows the absolute number of predicted sources, while the second column presents the count weighted by selection function values (which are used for the calculation of the RLFs as presented in Eq. 5.8).

	Radio AGN		Radio SFGs	
	Absolute count	Weighted count	Absolute count	Weighted count
z bin				
$0.01 < z \leq 0.08$	157	146.41	1009	1018.85
$0.08 < z \leq 0.25$	92	81.12	7916	7940.72
$0.25 < z \leq 0.35$	374	432.52	2763	3143.86
$0.35 < z \leq 0.50$	930	1148.91	670	959.26
$0.50 < z \leq 0.70$	2997	3791.67	502	754.65
$0.70 < z \leq 1.00$	5424	7464.52	376	592.63
$1.00 < z \leq 1.30$	2395	3753.98	39	72.43
$1.30 < z \leq 1.60$	2857	4640.50	...	...
$1.60 < z \leq 2.00$	3634	6180.56	...	...
$2.00 < z \leq 2.50$	2574	4132.91	...	...
$2.50 < z \leq 3.20$	1847	2742.60	...	...
$3.20 < z \leq 4.80$	760	1152.16	...	...
Total	24 041	35 667.88	13 275	14 482.41

Kondapally et al. (2022), Yuan et al. (2022), and Alqasim and Page (2023). Luminosity bins are defined starting from the faintest source available in the subset. Binned LFs for each redshift interval are then displayed in Fig. 5.14. Before establishing any comparison with previously derived RLFs, it is important to note that the values derived by us should be considered as a lower limit for a full RLF. The sources considered for this work are selected, as described earlier in the text, by their detection in the CW catalogue. That means that our parent sample is not complete as all our predicted sources need to be IR-selected. This construction does not include sources (either AGN or SFGs) that can be detected in radio bands (in particular, in EMU-PS) but are fainter than the CW 5 σ limit. Out of the 178 921 detections in the EMU-PS catalogue, 8219 do not have a counterpart in CW (i.e. close to a 5 % of them). For this reason, our RLF should be called, more precisely, IR-selected RLF. Nevertheless, we will keep the usual RLF nomenclature for the sake of simplicity and easiness of comparison with previous works.

Figure 5.14 depicts the values of the binned RLF calculated for predicted radio-AGN, radio-SFGs, and the total radio-predicted population in the EMU-PS field with EMU 944 MHz detections with their corresponding corrections. Along these values, RLFs from previous works have been included (AGN, SFGs, and total values; Mauch and Sadler 2007; M07; AGN; Pracy et al. 2016; P16; AGN; Šlaus et al. 2024; S24; SFGs; van der Vlugt et al. 2022; VdV22). Mauch and Sadler (2007) computed local RLFs using 7824 radio sources from the 1.4 GHz NRAO

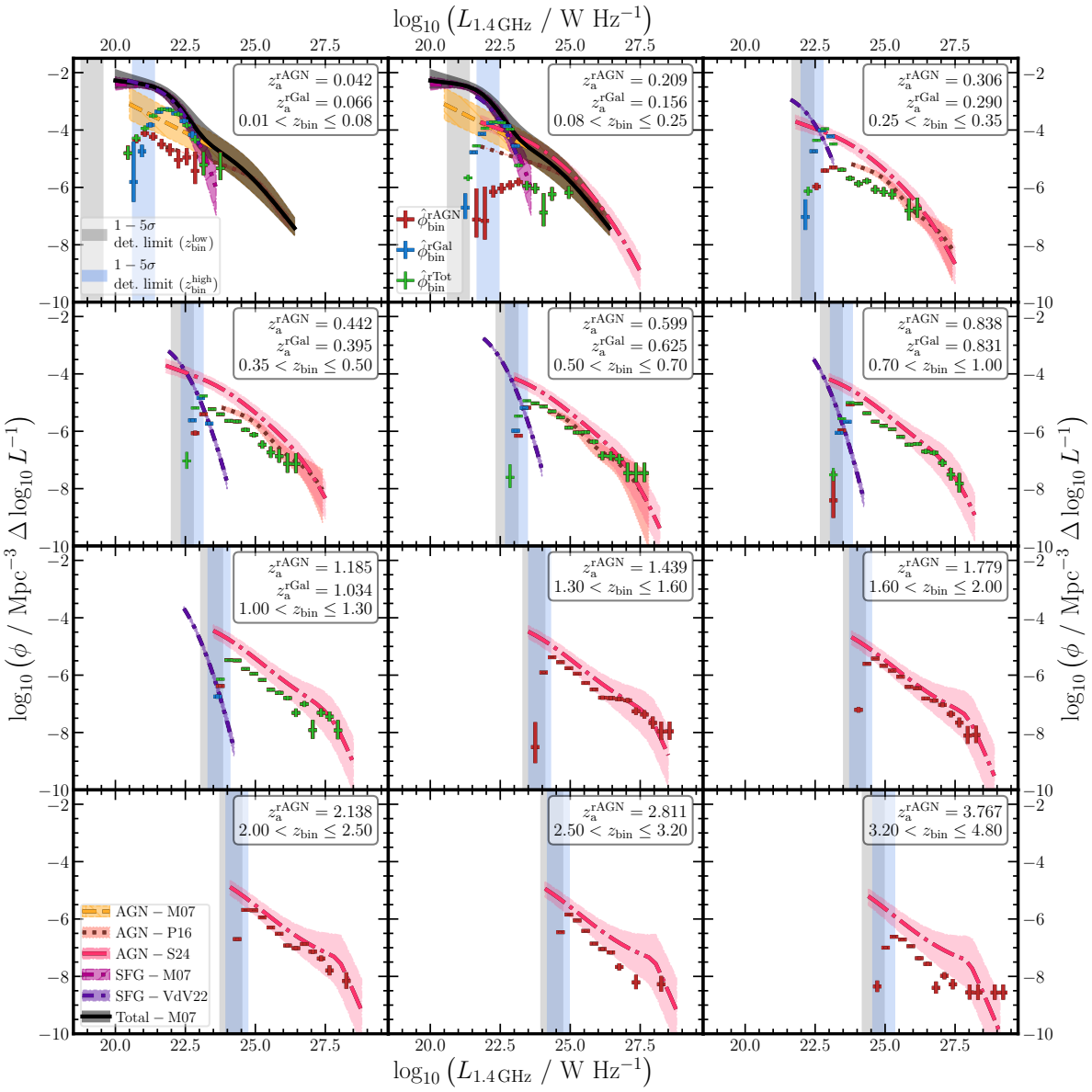


Figure 5.14: 1.4 GHz RLF in EMU-PS binned by predicted z values. Red symbols (and associated $1\,\sigma$ error bars) indicate the values of our predicted AGN RLF. Similarly, blue symbols show SFG RLF and green crosses, RLF for the full predicted sample. Horizontal uncertainties correspond to $\Delta \log L = 0.3$. Vertical uncertainties have been obtained following Clopper and Pearson (1934) and Gehrels (1986) with the inclusion of the cosmic variance following the recipes by Trenti and Stiavelli (2008). Solid lines and shadowed regions show previous RLF determinations and their associated $1\,\sigma$ uncertainties (AGN, SFGs, and total values; Mauch and Sadler 2007; M07; AGN; Pracy et al. 2016; P16; SFGs; van der Vlugt et al. 2022; VdV22) and $2\,\sigma$ uncertainties (AGN; Šlaus et al. 2024; S24). Grey and blue regions show $1\,\sigma$ to $5\,\sigma$ detection levels from EMU-PS measurements calculated using the lower (grey) and upper (blue) limits of each redshift bin. Upper-right corner of each panel shows median z of radio-AGN and radio-SFG, and z bin.

VLA Sky Survey (NVSS; Condon et al. 1998), covering an area of 7076 deg^2 in a redshift range $0.003 < z < 0.3$ (with a median redshift $z \approx 0.043$), which have been classified into AGN and SFGs. Pracy et al. (2016) have compiled a sample of 5026 $0.005 < z < 0.75$ optically detected and confirmed AGN in a ~ 900 deg^2 area in the FIRST survey. Also, van der Vlugt et al. (2022) have constructed 1.4 GHz RLFs for 1290 SFGs in the COSMOS-XS survey (van der Vlugt et al. 2021) with redshifts in the range $0.1 < z < 4.6$. Finally, (Šlaus et al. 2024) compiled radio measurements of 5446 AGN from different surveys of diverse depths and positions in the sky (combining almost 65 000 deg^2 of individual observations) up to $z \approx 3.0$ to develop an evolving RLF.

Only common redshift bins between previous works and our own RLF estimates are shown in Fig. 5.14, which shows that at the very last luminosity bin of every redshift range the RLF presents either a noticeable rise or a stabilisation. This effect has been discussed in previous works. For instance, Miyaji et al. (2001), Cara and Lister (2008), Croom et al. (2009), Yuan and Wang (2013), and Palanque-Delabrouille et al. (2016) have noted that the method by Page and Carrera (2000: binned LF), might cause pronounced biases near the flux cutoff of the analysed sample. The choice of the specific point in the $\log(L)$ bin can seriously impact in the final estimated value. Thus, any conclusion based on those bins has to be taken with care.

All redshift intervals show, for the faintest luminosities, a clear decrease in the values of the RLFs. Such drop is due to the effect of low sample completeness in those luminosity bins. As we have not corrected our measurements for the completeness of the measurements in the EMU-PS survey, it is expected to see such behaviour in the faint end of the RLFs. For the case of LoTSS-DR1, Fig. 14 from Shimwell et al. (2019) presents the radio-detection completeness values for the LOFAR observations in the HETDEX field, with a 90 % completeness at 0.45 mJy. Such values have not been included in our selection function as only ML-based corrections have been considered in this exercise.

In the closest redshift bin ($0.01 < z_{\rm Predicted} < 0.08$), and besides the low-luminosity drop mentioned earlier, our binned RLFs follow one of the curves from previous works: SFGs follow the curve from M07 and AGN, the curve from P16. Nevertheless, some issues are apparent. First, the AGN LF has a sharp jump at $\log\left(L_{1.4\,\rm GHz}/\rm W\,Hz^{-1}\right) \approx 21.5$. This jump implies that the density of local, low-luminosity radio-detectable AGN has an abrupt change. Such change is not seen in previous works, raising the question of its origin. Figures 5.7 and 5.8 clearly show an abnormal high number of AGN in the first redshift bin which, according to typical

radio-AGN redshift distributions should not be present (e.g. Simpson et al. 2012; Pracy et al. 2016; Šlaus et al. 2020; Ceraj et al. 2020; Bonato et al. 2021). One possible explanation for such overdensity is related to the (probably low) predicting power of our pipeline at low redshift values. To support this interpretation, Fig. 5.5 shows that the redshift bin with the largest number of predicted radio-AGN is the lowest one. The number of sources it has is compatible with the bins for predicted radio-SFGs. Thus, there is a strong likelihood of that a fraction of these low-redshift predicted radio-AGN are, radio-SFGs instead. As a way to test this possibility, Fig. 5.15 shows the two-dimensional distributions of the output probabilities from each step of the prediction pipeline as a function of the predicted redshift of all sources used in the calculation of the RLFs.

As expected, the second and fourth panels of Fig. 5.15 show a large number of predicted radio-AGN at low predicted redshift values. Prominently, a large fraction of them have low probabilities (i.e. $P(\mathrm{AGN}) \lesssim 0.7$) whereas, for the rest of the redshifts, their distribution is more homogeneous. From the point of view of the pipeline, if a source is predicted to be an AGN, but with a low certainty, it will probably will be predicted to have very low redshift. Conversely, and as also shown by the last panel in Fig. 5.15, the prediction of the radio detectability of such sources is not as strongly affected by this effect as the AGN classification.

The distribution of predicted probabilities shows that low-redshift radio-detectable AGN, in general, have not been classified as certainly as expected. And thus, the inclusion of such sources in the calculation of the RLF cannot be used to extract conclusions. For this reason, only the the higher redshift bins will be considered for analysis.

Returning to the analysis of Fig. 5.14, from its second redshift bin onwards, RLFs do not show such unexpected features. Nevertheless, one appealing detail is noticeable in the redshift bin $0.08 < z \leq 0.25$, where radio-AGN do not follow the distributions from previous works. From a numerical perspective, such difference arrives from the corrections (selection function) applied to the number of sources in the bin, which cannot enhance their distribution as to match previously obtained studies. That level of corrections is, most likely, due to the fact that sources in the LoTSS-DR1 area that are predicted to be in that redshift bin exhibit high metric values. Bearing in mind that our selection function only includes corrections from the application of ML models, and not from the observational properties of the studied surveys, there is no prior expectation for the values of the RLF to change drastically.

Another point to underline from our AGN RLFs can be found in the bins between $z = 0.25$

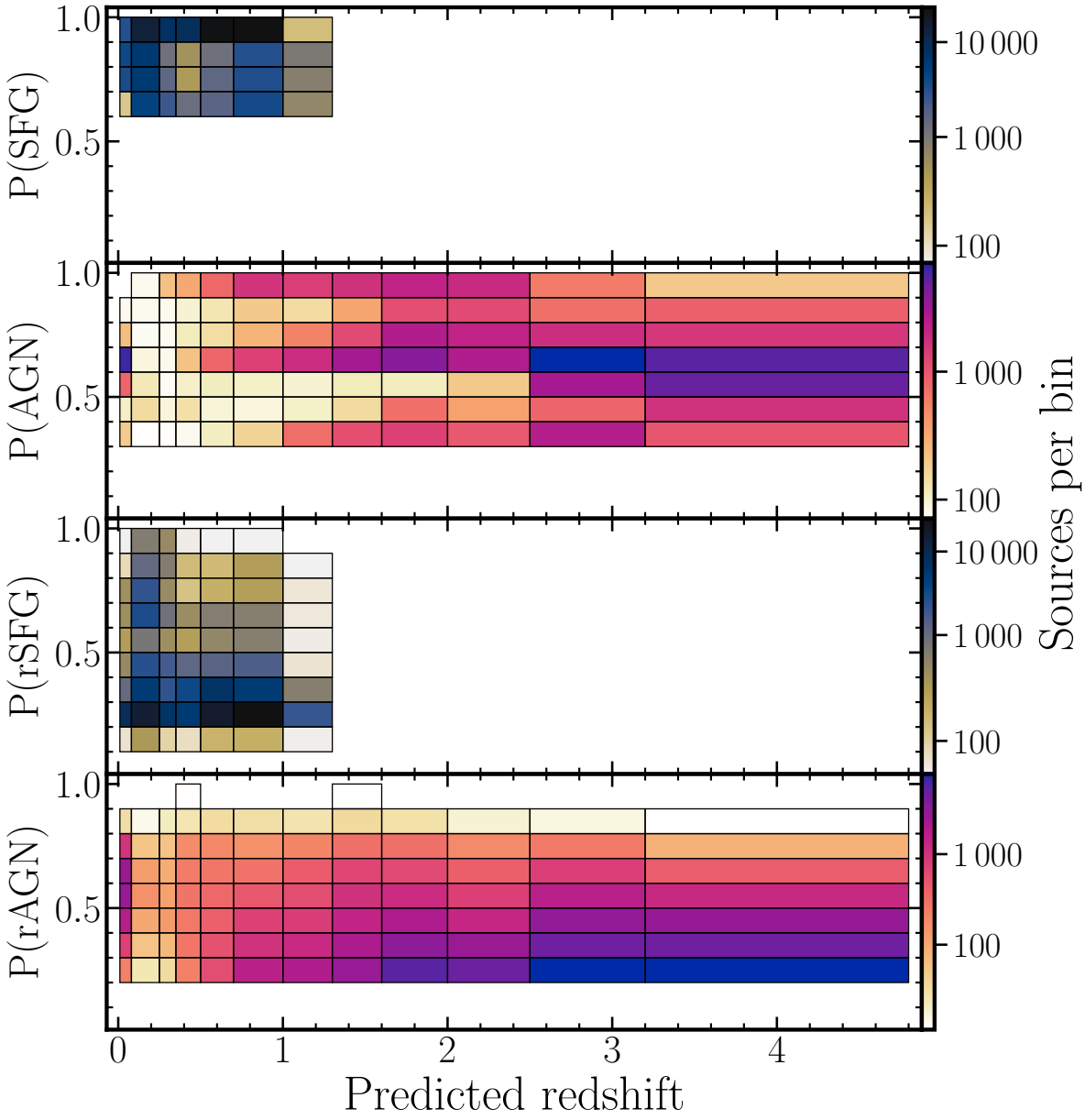


Figure 5.15: Two-dimensional histograms for the distribution of predicted probabilities (of each step in our modified prediction pipeline) as a function of estimated redshift for sources deemed to be radio-detected AGN and radio-detected SFGs and included in the calculation of the RLF. The darker the colour of the bin, the highest the number of sources with that combination of probability and predicted redshift, following the coding of the colorbars. From top to bottom, first panel shows the probabilities of a source to be SFG. Second panel presents the probability of a source to be an AGN. In the third panel, the probabilites of SFGs to be radio detected are included. Last panel shows the probability that a source predicted to be AGN can be detected in the radio. Bins in redshift correspond to the values defined in Table 5.3 and vertical bins are intervals of 0.1 in probability.

and $z = 2.5$, at luminosities in the region of 25 dex to 27.5 dex in units of W Hz^{-1}, depending on the redshift value. In these ranges, the AGN LF has a weak bump of less than 0.3 dex in source volumetric density units when compared to the canonical power-law distribution of the LF at high luminosities. All these features, when including uncertainties, are still present. One possible explanation for this elevation that our sample is influenced by the existence of an additional high-luminosity sub-population that has a stronger presence at high redshift values. Previous RLF determinations have already given hints on the existence of such population. The first of them is from Pracy et al. (2016), who separated the study of radio AGN into LERGs and HERGs. Their Fig. 10 shows the RLFs for LERGs and HERGs at $z < 0.75$. For their $z > 0.5$ bin, both fitted and observed LFs cross at around $\log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \sim 25.5$, being consistent with the presence of the bump in the corresponding bin of our sample. Furthermore, Willott et al. (2001) analysed radio sources at redshifts $z \lesssim 4$ and fitted a number of models to their density distributions. They assumed a two-population sample, with low-luminosity sources (FRIs and low-excitation FRIIs) and high-luminosity subset (mostly, FRII QSOs with strong emission lines). In all the realisations of their models at $z > 0.5$, the total RLFs show a break that starts at $\log\left(L_{151\,\mathrm{MHz}}/\mathrm{W\,Hz^{-1}}\right) \sim 26.1$ (which, using $\alpha = -0.7$, corresponds to $\log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \sim 26.5$). Similar results are shown by Best and Heckman (2012), Best et al. (2014), Ceraj et al. (2018), Williams et al. (2018), Butler et al. (2019), and Kondapally et al. (2022), who separated their AGN population into two subsets and by Smolčić et al. (2017), Slob et al. (2022), and Šlaus et al. (2024) who analysed their full AGN dataset. In particular, Šlaus et al. (2024) using both Bayesian and binned methods, find a bump, depending on the redshift, located close to 27.5 dex to 28 dex in W Hz^{-1} units. An additional consequence of the bump detected by Šlaus et al. (2024) is a flattening of their fitted RLFs at high redshifts and luminosities. Our RLFs between $z = 0.25$ and $z = 2.5$ can resemble that effect, where their brighter end tend to become flatter (beyond the effect on the brightest bin described earlier in this text).

A different point of attention regarding the analysis of RLFs is that of their evolution (or lack thereof) as a function of redshift. One option for the display of our RLF as a function of redshift is through Fig. 5.16. In this way, rather and focusing on the individual obtained values for the RLFs, the emphasis is on their overall variation in time.

Figure 5.16a shows the evolution of the radio-SFGs while Fig. 5.16b does it for sources predicted to be radio-AGN. For the radio-SFGs, and given their narrow distribution, a qualitative

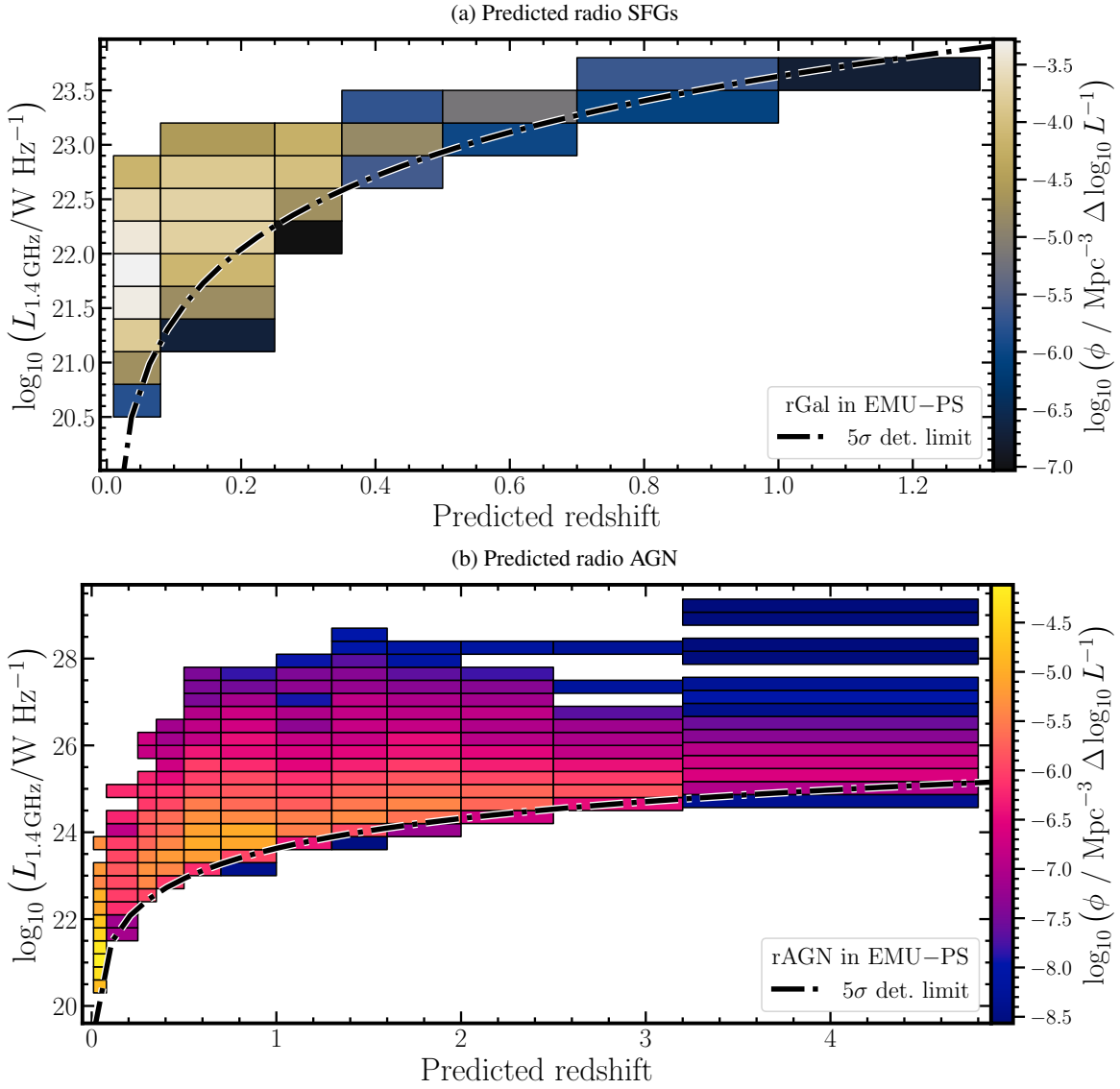


Figure 5.16: Predicted 1.4 GHz luminosity vs predicted redshift for predicted (a) radio SFGs and (b) radio AGN in the EMU-PS catalogue. Luminosity and redshift bins as in Fig. 5.14. Bins are coloured according to their estimated RLF values and following each individual colourbar. Dot-dashed black line represents 5 σ detection limit in EMU-PS.

analysis on their RLF evolution is possible in the range $22.5 \lesssim \log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \lesssim 23.5$. In the case of radio-AGN, Fig. 5.16b can be analysed in the range $24 \lesssim \log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \lesssim 28$. For both RLFs, a smooth evolution in redhift can be identified. Although further investigation is necessary for the correct quantification of such variations, it is possible to estimate that our RLFs decrease with redshift as much as ~ 1.5 dex in the range $0.5 \lesssim z \lesssim 4$.

In order to extract more information from our RLFs, a modification can be implemented in the previous calculations. If all the sources that are predicted to be radio detectable but do not have a counterpart in the EMU-PS catalogue are assumed to have a low flux (below the detection limit of EMU-PS), then the number of available sources for the determination of the RLF can clearly increase. As mentioned previously in this section, and using a spectral index value of $\alpha = -0.7$ for flux conversion, the detection limit of LoTSS-DR1 is lower than that of EMU-PS. Thus, it is possible to assign (as an imputed value) an EMU flux for these sources without an entry in the EMU-PS catalogue. From the lack of knowledge of the behaviour of the missing fluxes, we have assumed a Uniform, $F \sim \mathcal{U}\left(5\,\sigma^{944\,\mathrm{MHz}}_{\text{LoTSS-DR1}}, 5\,\sigma_{\text{EMU-PS}}\right)$, distribution to allocate the imputed values. In this way, and after the application of the prediction correction of Eq. 5.7, our analysis now has 100 670 and 119 297 predicted and EMU-PS-measured radio-AGN and radio-SFGs, respectively. Their distribution in the $(z, L_{1.4\,\mathrm{GHz}})$ plane, once their overly bright predicted radio SFGs have been re-labelled as AGN, is shown in Fig. 5.17. As expected, a large number of sources are concentrated below and around the detection limit for EMU-PS.

Assuming a flux for the predicted sources missing in the EMU-PS catalogue can be considered as an additional correction applied to our predictions. Without it, the mere verification of a detection in EMU-PS could have been associated to a correction by purity, where the total number of elements is reduced by their detectability. This detection validation affects, mostly, the faint end of the RLF, decreasing the number of low-luminosity sources. The flux imputation allows one to discard such pseudo-correction and only focus on the selection function based upon values from LoTSS-DR1.

With the new extended available radio measurements, it is possible to construct updated RLFs following the steps described earlier in this text. Using the same bin widths and corrections, their values and shapes are presented in Table 5.4, Fig. 5.18 as well as in Fig 5.19 for a condensed view. The new imputed fluxes affect, as expected, the faintest luminosity bins in each redshift range with minor changes at higher values (due to the re-arrangement of the luminosity bins). Upon inspection of Fig. 5.18, it is possible to see that fainter luminosity bins now follow closer

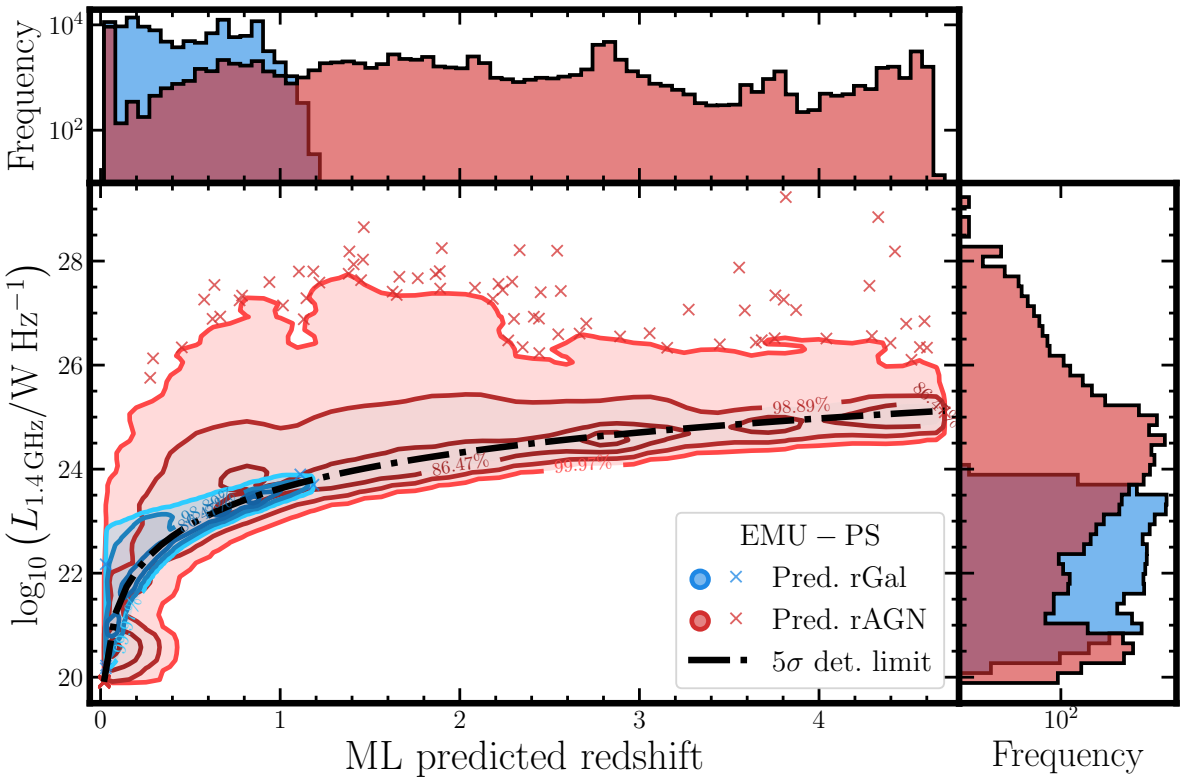


Figure 5.17: Distribution of predicted 1.4 GHz luminosities vs predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable SFGs (blue) in the area of the EMU-PS catalogue. SFGs with high 1.4 GHz luminosities have been re-labelled as AGN, following the prescriptions by Magliocchetti et al. (2014) and Magliocchetti (2022). Fluxes have been imputed for sources predicted to be radio detectable but without an entry in the EMU-PS catalogue. Description as in Fig. 5.7.

previous RLFs than those shown in Fig. 5.14. This effect is more pronounced, in low-redshift bins, for SFG RLFs than for those from AGN.

Table 5.4: Number of predicted and imputed radio-detectable sources (AGN and SFGs) in the EMU-PS area by predicted resdhift bin. For radio-AGN and radio-SFGs, first column shows the absolute number of predicted sources, while the second column presents the count weighted by selection function values (which are used for the calculation of the RLFs as presented in Eq. 5.8).

	Radio AGN		Radio SFGs	
	Absolute count	Weighted count	Absolute count	Weighted count
z bin				
$0.01 < z \leq 0.08$	9215	10 815.17	10 929	20 405.57
$0.08 < z \leq 0.25$	594	787.59	29 634	34 974.14
$0.25 < z \leq 0.35$	692	865.24	12 454	14 317.58
$0.35 < z \leq 0.50$	1739	2508.82	9856	13 653.18
$0.50 < z \leq 0.70$	4364	6237.56	24 265	36 592.20
$0.70 < z \leq 1.00$	8311	12 582.75	29 905	46 483.17
$1.00 < z \leq 1.30$	5963	10 318.04	2254	3751.77
$1.30 < z \leq 1.60$	8699	15 241.22	...	...
$1.60 < z \leq 2.00$	13 264	23 762.48	...	...
$2.00 < z \leq 2.50$	10 421	18 552.96	...	...
$2.50 < z \leq 3.20$	19 306	28 423.16	...	...
$3.20 < z \leq 4.80$	18 102	19 233.61	...	...
Total	100 670	149 328.58	119 297	170 177.61

Most of the features present in Fig. 5.14 are still visible in Fig. 5.18. For instance, the very high number of predicted radio-AGN in the first redshift bin, the distance between our AGN-RLF and previous works in the second redshift bin ($0.08 < z \leq 0.25$). Differences between both figures start with the sizes of error bars. The larger number of sources available for lower luminosities make the uncertainties, from both the number of elements and cosmic variance, smaller.

Another difference can be seen in the $0.25 < z \leq 0.35$ redshift bin. The previous version of the AGN RLF showed a hint for the bump present in other redshift ranges. However, the bin in Fig. 5.18 does not show such bump as before. Thus, we cannot extract firm conclusions on this regard for it. Nevertheless, we can now claim that the bump in the AGN RLF is present in the range $0.35 < z \leq 2.5$.

With the corrected RLFs, it is possible to comment on their very good agreement with RLFs from previous works. After taking into account that the redshift values used for our estimates and for the plotting of previous RLFs are slightly different, our estimates are (for non-local sources) fully compatible with those early RLFs for both radio-AGN and radio-SFGs.

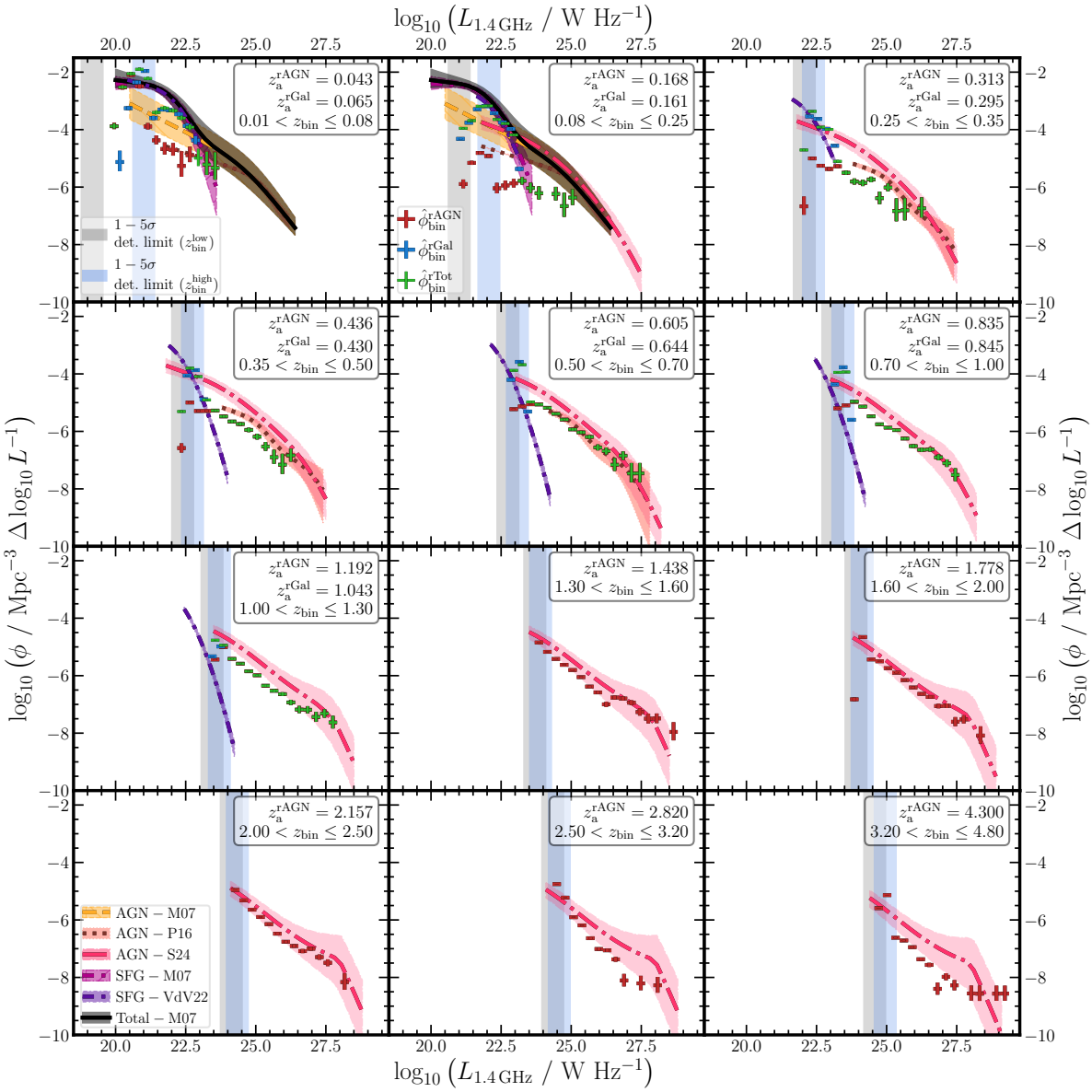
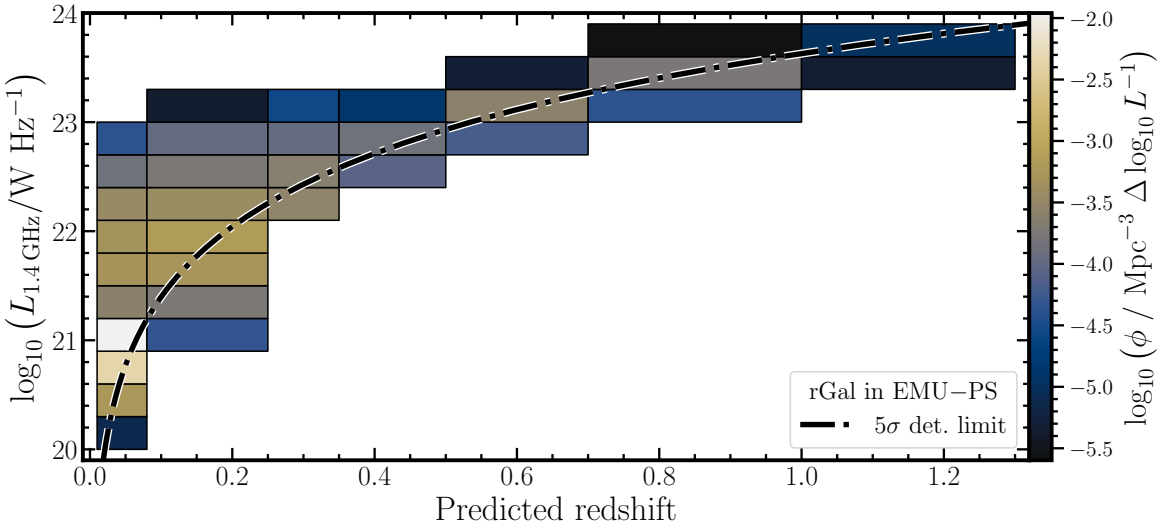


Figure 5.18: RLF (at 1.4 GHz) in EMU-PS binned by predicted z values using all sources predicted to be radio-AGN and radio-SFGs with assumed fluxes when missing. Figure description as in Fig. 5.14.

(a) Predicted radio SFGs



(b) Predicted radio AGN

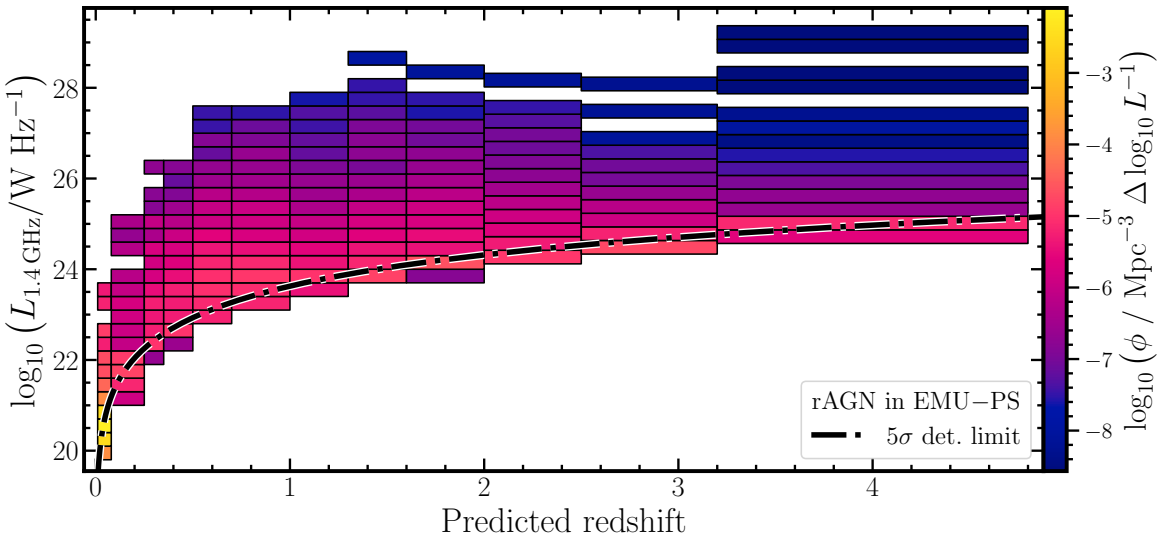


Figure 5.19: Predicted 1.4 GHz luminosity vs predicted redshift for predicted (a) radio SFGs and (b) radio AGN in the EMU-PS catalogue with imputed fluxes. Luminosity and redshift bins as in Fig. 5.14. Bins are coloured according to their estimated RLF values and following each individual colourbar. Dot-dashed black line represents 5 σ detection limit in EMU-PS.

From Fig. 5.19, it is possible to study the evolution of our RLF estimates as a function of redshift. For both AGN and SFGs, the change is smooth for similar luminosity ranges. In the case of radio-AGN, and if only the bins at $z > 0.35$ are selected, a detailed view of their evolution is shown in Fig. 5.20.

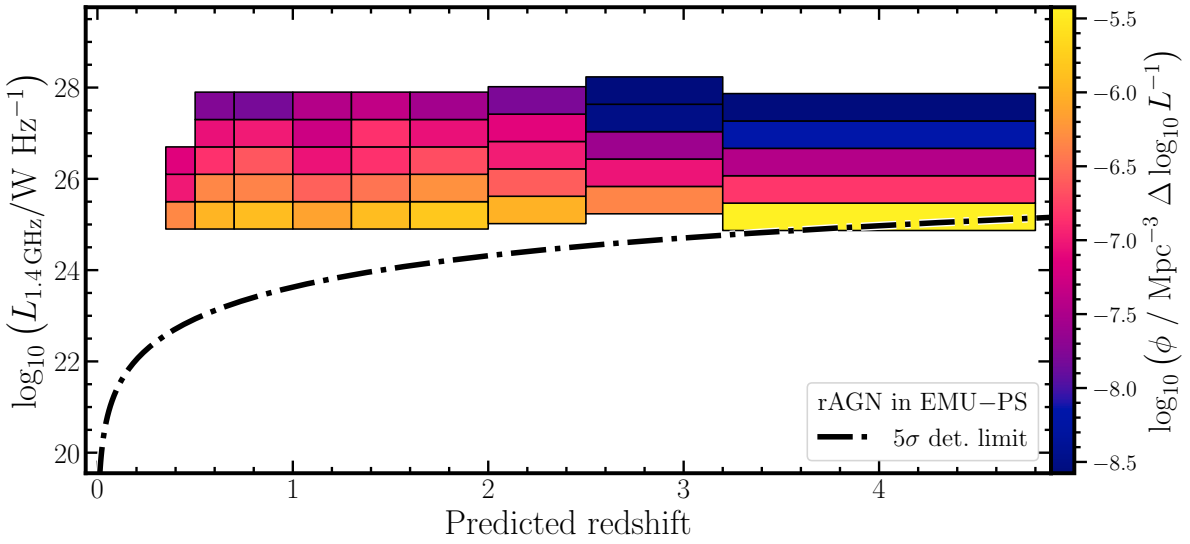


Figure 5.20: Predicted 1.4 GHz luminosity vs predicted redshift for predicted radio AGN in the EMU-PS catalogue with imputed fluxes at $z > 0.35$ and in the luminosity range $25 \lesssim \log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \lesssim 28$. Details as in Fig. 5.19. For displaying purposes, luminosity bins have been increased by a factor of two (i.e. $\Delta \log_{10} L = 0.6$).

Now, and as before, focusing in the range $25 \lesssim \log\left(L_{1.4\,\mathrm{GHz}}/\mathrm{W\,Hz^{-1}}\right) \lesssim 28$, the horizontal change is still close to 1.5 dex between $z \sim 0.5$ and $z \sim 4$. But now, it is possible to see that, in the region of $0.5 \lesssim z \lesssim 2$, there is a peak in the values of the RLF. This peak is, for instance, consistent with what Pracy et al. (2016) has found in their AGN sample ($z < 0.5$ sources). For a more general evolution scheme, Rigby et al. (2015) and Yuan et al. (2017) have suggested that the peak in RLFs evolves with luminosity, with lower luminosities presenting a maximum at $z \sim 1$.

All these estimates for the RLFs can also be contrasted to the values obtained from simulations. For radio-AGN in particular, they can be compared with those from Amarantidis et al. (2019), who compiled number density estimates from several cosmological simulations. For their intermediate-redshift range ($0.8 < z < 1.0$), their simulation-based estimates are compatible with our corresponding bin ($0.7 < z \leq 1.0$). This compatibility takes into consideration that our estimates are based on IR detections while those from Amarantidis et al. (2019) consider the full radio population derived from the simulations.

These previous results further reinforce the capabilities of our prediction pipeline and the feasibility of the results which allow us to obtain RLF estimates that are highly compatible with

the current calculations from very diverse methods and surveys.

5.3 Radio counterpart assessment

One of the main issues of multi-wavelength studies of astrophysical sources is the identification of counterparts in observations from different instruments and bands (i.e. multi-wavelength counterparts, cf. Sect. 1.2.4). Accurate positioning of sources is crucial to allow further studies (e.g. spectroscopic targeting).

In particular, the radio astronomical community has forecast that the use of ML will be instrumental to find radio counterparts of large catalogues (Lazio et al. 2014). Nevertheless, and as stated in Sect. 4.1.2, the number of works using ML for predicting the radio detectability of sources (problem comparable to finding radio counterparts of confirmed sources) is, to date, very limited. One recent example of the use of ML techniques, combined with traditional methods, is that by Mostert et al. (2024), who applied a machine-based pipeline for the selection and counterpart finding of sources in the LoTSS - data release 2 (LoTSS-DR2; Shimwell et al. 2022).

The lack of such studies prompts us to investigate the potential use of the developed ML pipeline to quantify the probability of a detection to have a counterpart in a different photometric catalogue. In order to assess the idea, we will utilise the radio-detection classification model described in Sect. 5.2 and Appendix C (which is a modification of the model presented in Chapters 4 and 6). Following the discussion of Rohde et al. (2006), we can make direct use of the output probabilities given by the models since they have been calibrated and their distribution is well behaved.

The pipeline described in Sect. 5.2 and Appendix C was applied to the IR-detected sources in the EMU-PS area. In order to test the radio counterparts, we selected the sources, regardless of their initial classification, that were predicted to be radio-detectable AGN. In particular, we focused on the sources that presented high probabilities to be of such class (i.e. probability of being AGN higher than $P(\mathbb{C}) = 0.7$ and probability to have radio detection higher than $P(\mathbb{C}) = 0.8$). Then, we plotted these predicted radio-detectable sources on top of the map of EMU-PS together with all other IR-detections in the surrounding region. Some example sources from this selection are shown in Figs. 5.21 to 5.22i.

It is worth noting that, in each image, all IR-detected sources have been plotted, regardless

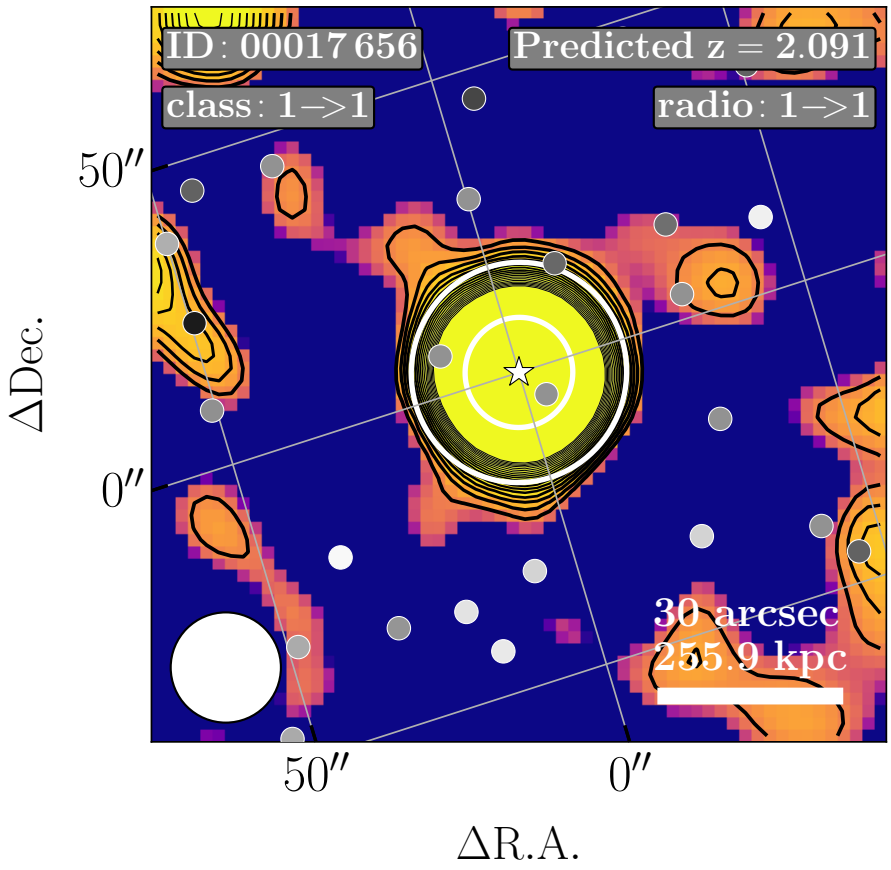


Figure 5.21: Postage stamp of CW-detected source ID `000017656` in the EMU-PS area (located in the centre of the image). All small circles show the position of a CW detection in the displayed region without a radio counterpart, while all stars are CW detections with a confirmed radio detection. Face colours of stars and small circles correlate with the predicted probability of such CW detections to have a radio source associated to them, where a brighter hue represents a higher probability. For displaying purposes, all emission below a $1\,\sigma$ detection limit has been set to zero (0) and black contours show emission at $2\,\sigma$ to $20\,\sigma$ levels. Concentric white circles in the centre of the image limit two regions with a radius of 1 and 2 times the full width at half maximum (FWHM) ($18\,''$) centred in the selected source. Large white circle in bottom left corner represents the EMU-PS synthesised beam size ($18\,''$) and the gray horizontal lines in the bottom right corner denote a $30\,''$ scale in physical units using the predicted redshift. In the top left corner of figure, the identification number of the source is written, as well as its confirmed and predicted AGN (i.e. `class`) states connected by an arrow, where `0` represents a SFGs, `1` represents an AGN, and -1, a source without prior identification. In the top right corner of the figure, the predicted redshift of the source is included, as well as its confirmed and predicted radio detection (i.e. `radio_detect`) status (also connected by an arrow), with `0` representing a source without radio detection (confirmed or predicted) and 1 stands for radio detection.

of their initial or predicted class. The inclusion of the full sample (AGN and SFGs) implies that the radio prediction model was applied to all candidates and thus, more uncertainties have been included in the results from its application. These uncertainties do not hinder the analysis regarding radio counterparts.

A first example is shown in Fig. 5.21. It presents a radio-detected AGN (ID `000017656`) that has been predicted to be in the same category (i.e. radio-detectable AGN, an accurate prediction). The star in the middle of the field shows that the source has been detected in the EMU-PS data, as evident by the bright region in the background image. There are two additional CW-detected sources within one synthesised beam of distance from the selected target. They present a darker face colour, indicating that their probability of having a radio counterpart is lower than that of `000017656`. Thus, the use of the pipeline with the selected source shows that it can be inferred that the large radio source in the background (and only associated by means of distance) can be the counterpart of `000017656` with a high likelihood. What Fig. 5.21 presents, then, is the expected output of our prediction pipeline, that is, a radio-AGN predicted to be such.

The same behaviour can be seen in Figs. 5.22a, 5.22b, 5.22c, 5.22d, and 5.22e. In particular, Figs. 5.22a and 5.22b show two confirmed radio-detected AGN that have been correctly predicted surrounded by several CW sources with lower probability of having radio counterparts. Interestingly, the radio source in Fig. 5.22b resembles the emission of a central source with two lobular arms, as radio AGN with a bent jet. Then, Figs. 5.22c, 5.22d, and 5.22e depict EMU-PS confirmed detections that do not have an associated AGN or SFGs confirmation. Nevertheless, they have been predicted as AGN with high confidence. As previously, these predictions have the highest likelihood of being detected in the radio in their surroundings (within two FWHM from their position). It is important to note that the emission depicted in Fig. 5.22d might not appear, in terms of shape, to that of an AGN or a similar source. Even though the radio prediction and the cross-match are in agreement, further studies are needed to confirm the nature of such source.

A different scenario is shown in Figs. 5.22f, 5.22g, 5.22h, 5.22i, 5.22j, and 5.22k, where all CW-detected sources could not be associated (with a direct cross-match) to a source in the EMU-PS catalogue. It is important to note that all emission shown in the postage stamps is above a $1\,\sigma$ limit, while the detections listed in the EMU-PS catalogue are above the $5\,\sigma$ limit (Norris et al. 2021). Thus, it is possible that the catalogue misses faint sources that are picked and predicted by our pipeline.

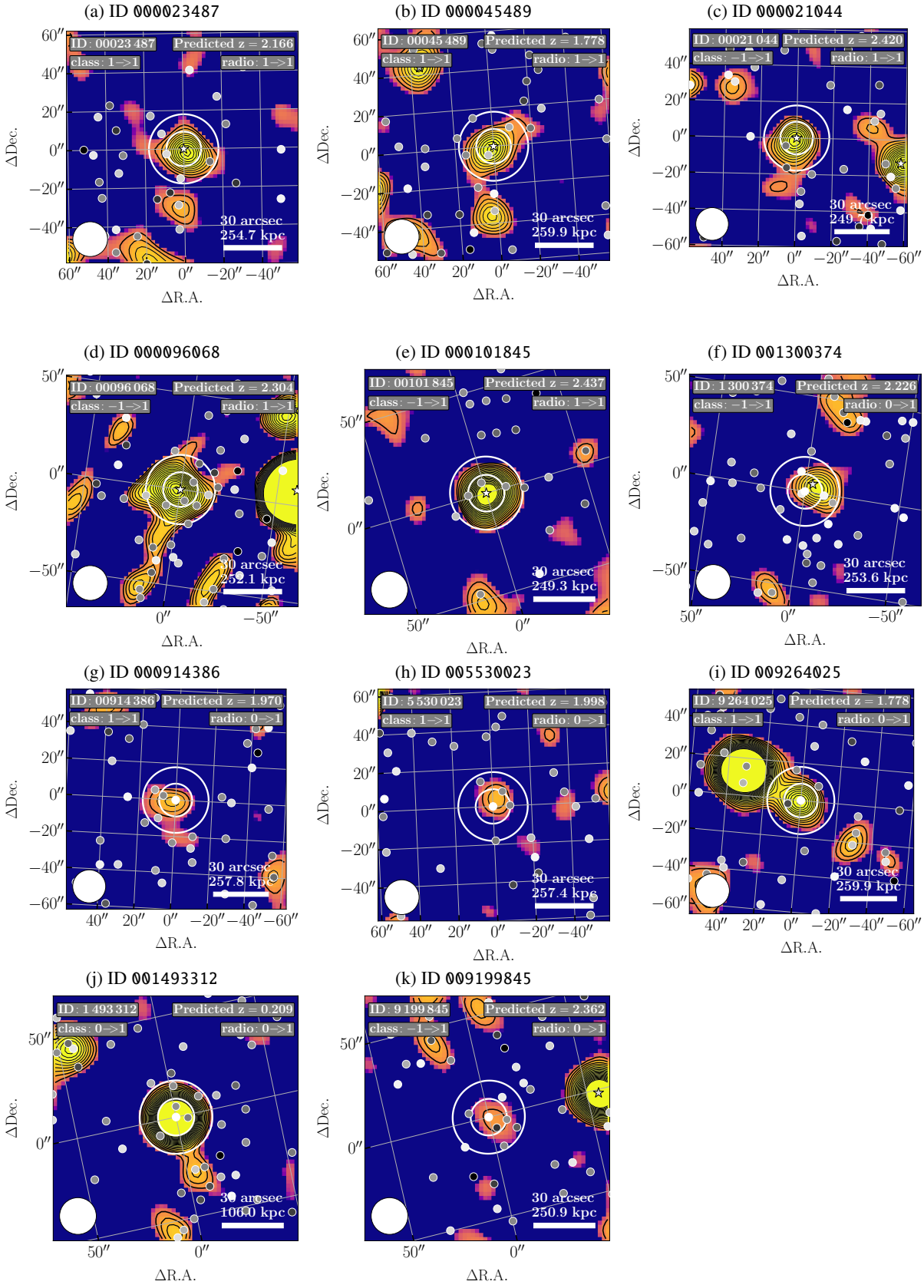


Figure 5.22: Postage stamps of CW-detected sources IDs 000023487, 000045489, 000021044, 000096068, 000101845, 001300374, 000914386, 005530023, 009264025, 001493312, and 009199845 in EMU-PS area. Description as in Fig. 5.21.

Figure 5.22f presents an example worth noting. Source ID `01300374`, which does not present any prior classification, shows the highest probability of radio detection in its surroundings, but there is an additional source that has been labelled as the counterpart to the background radio source. This source, shown as a star with a darker face colour, has obtained a lower probability to be detected in the radio bands. Therefore, our detection pipeline is contesting the radio counterpart assignation by distance. Additional measurements are needed to determine the most likely counterpart.

Sources ID `000914386` and ID `005530023`, depicted in Figs. 5.22g and 5.22h, have been originally labelled as an AGN without any radio detection. Our pipeline has assigned to them a high probability to have a radio counterpart over their neighbouring sources. At the same time, their background images show a bright source in the same position. Thus, we have predicted that these source can have a radio counterpart (which might have been previously catalogued should EMU-PS have used a lower detection threshold). A similar situation can be seen in Fig. 5.22i with source ID `009264025`. Its main difference is that it is located in what might be a more radio-populated region that could present some overlapping sources.

Finally, two sources without an original AGN label are presented. Source ID `001493312` (Fig. 5.22j), with a prior label as SFG, is located in a very IR-crowded region and it has been given, by our prediction pipeline, the highest radio-detection likelihood over their neighbours. In the background EMU-PS image, it is possible to see bright radio emission from a point-like source that has a bright appendix. Thus, we have been able to predict the existence of radio emission that was missed by the EMU-PS catalogue. Our last example source is ID `009199845` in Fig. 5.22k. Without any prior label (in either `class` or `radio_detect`), it has been predicted to be a radio-detectable AGN among all its neighbouring detections. The EMU-PS image shows a relatively bright source that might be associated to our prediction.

The method presented here can also be applied in the calculation of RLFs of Sect. 5.2. As shown previously, selecting the closest radio source to an IR-detected source might not always be the most efficient way to find its radio counterpart. Such inefficiency can lead to inaccurate RLF calculations, introducing uncertainties that might be difficult to assess and correct.

Instead of selecting the closest source as radio counterpart, as done in Norris et al. (2021), all IR dectections in a small search radius (e.g. the size of the synthesised beam) around a radio detection can be selected. Then, the IR source with the highest probability of being detected in radio frequencies is chosen as the counterpart. Additionally, a minimum probability threshold

can also be set to ensure only highly likely counterpart candidates are selected.

Using our prediction pipeline for the selection of radio counterparts has associated benefits. First, it goes beyond using distance as a selection criterion, using all available photometric information to make a more robust decision. Second, it provides an uncertainty measurement for the counterpart selection. The probability assigned by the model indicates the confidence level in the chosen counterpart.

All these examples show that our prediction pipeline can be used, without any further modification, to understand the location and distribution of radio counterparts of IR-detected sources. By virtue of its training with very deep radio observations (i.e. with LoTSS-DR1), our pipeline can also help finding sources that might have been overlooked or discarded by other radio surveys and catalogues, as was the case with the EMU-PS catalogue.

6

Future developments

With the experience gained during this thesis and the constant development and exploitation of ML strategies in the literature, this chapter focuses on different avenues where the pipeline results and characterisation can be improved, and also on near-future applications of our pipeline with significant scientific impact. We list here some steps that can be taken in order to expand our knowledge on either the prediction pipeline or the results obtained from its use. Additionally, we present some surveys and instruments in which their data can be subject to the application of our prediction pipeline.

6.1 Subset analysis for radio-AGN identification

Expanding on the analysis of Chapter 5, the behaviour of each model can be studied to understand radio-detectable AGN selection within the pipeline. In particular, it is possible to analyse targeted source subsets and explore the relationships between input features and model outputs. Visualising output values (probabilities for classification models and redshift for our regression step) as function of input features will identify regions of the space of parameters where each model has stronger predictive power. These combined analyses can provide a comprehensive understanding of the factors driving radio-detectable AGN selection within the data.

Inspired by the work by D'Isanto et al. (2018), and as an extension of the exploration of the models, local feature importances tools (Sect. 1.3.3) can be used to understand how the impact of each feature evolves with redshift for radio-detectable AGN. By considering the physical processes behind AGN emission in each band and their colours, we can understand what type of sources the models are likely to predict as radio-detectable AGN within each redshift range.

Ultimately, these in-depth analyses, as informed by both techniques, will lead to a deeper understanding of the physical and data-driven properties that define AGN (or any other source type) within the pipeline.

6.2 Improvement of prediction pipeline

Once the impact of the features included in each of the models of the prediction pipeline have in the output values is assessed, additional changes can be implemented to improve the quality of the predictions. One class of modifications is related to the inclusion of new categories of features. Apart from magnitudes and colours, and depending on the used catalogues and datasets, information about, for instance, morphology of the sources can be included. Letting the model know about the angular sizes, stellarity indices, ellipticities of the detections in some of the used wavelengths could improve, for instance, the metrics on the redshift estimation. Another proxy for morphology and sizes of sources can be given by the use of magnitudes calculated through different methods. For instance, using different apertures, magnitudes can cover wider or smaller areas and their combination can inform about the physical extension of galaxies and the AGN they might host. Such approach has been taken by, for example, Daoutis et al. (2023) and Zeraatgari et al. (2024) to classify galactic and extra-galactic sources according to their optical and NIR photometry.

Another way to include photometric information into a ML model is through the direct use of images (Dieleman et al. 2015). Such approach involves, in a very basic format, using small image cutouts for each source in each of the observed bands and surveys and training models on them. In this way, it is possible to extract information on the shape (or lack thereof) of the sources without the extraction of intermediate proxy values as magnitudes or fluxes (e.g. Smith and Geach 2023). A further advantage of this approach is that it does not need to apply any kind of correction or imputation to missing detections. As long as the studied sources are covered by observations, no assumptions are needed on their detection levels (if any detection is present in an image). However, one relevant drawback of using image cutouts is the large volume of computational resources needed for their analysis. Image-based ML techniques tend to be more expensive than tabular-based models as they use more parameters and features for their analyses (e.g. Domínguez Sánchez et al. 2018) but they can obtain better results than traditional ML algorithms and can deal with a larger variety of astrophysical problems (Yang et al. 2023). Other disadvantages of using images to feed ML algorithms include that, depending in the variety and quality of observations, different PSFs and noise properties would be combined. Additionally, models would need to be trained on separating potentially overlapping sources, increasing the sources of uncertainties. If the goal is interpreting the estimations and predictions

from the models, image-based analyses are more complex than tabular models, leading to more expensive feature analysis implementations.

One important aspect that should be adressed in future iterations of our pipeline is that of measurement uncertainties. As traditional statistical methods, our models should be able to incorporate, in some way, measurement errors for their training. Such effort has been part of the work from both the ML and Astrophysical communities, with relevant examples in the works from Reis et al. (2019), Shy et al. (2022), and Rodrigues et al. (2023) who have, respectively, modified the ML models, run several times a model with perturbed versions of the training data, and tried to extract noise patterns from runs of ML models. Additionally, it is expected to retrieve predictions with an associated uncertainty or confidence interval. While such values are natural for a calibrated classifier (such as those used in this work for the classification between SFGs and AGN and the prediction of radio detectability), the output from regressors do not usually have such associated quantities. Some relevant examples of the search for output regression uncertainties are given by Duan et al. (2020), with their algorithm `NGBoost`[1] and Boström (2022) with `crepes`[2], which implements conformal regression methods (Gammerman et al. 1998; Saunders et al. 1999; Vovk et al. 2022) for the prediction of quantities. In Astrophysics, `NGBoost` has been used by, for example Gilda et al. (2021) for the enhancement of SED modelling, and conformal regression has been used by Yong and Ong (2023) and Gilda (2024) for the quantification of black hole masses and the improvement of SED fitting, respectively.

Since one of the major goals of the community is to explore and understand the physical conditions present in the EoR (cf. Chapter 1), it is natural to focus the training and analyis of our prediction pipeline in sources at high redshift values. One way to achieve such goal is through the training, only, on high-redshift AGN and SFGs. Applying, directly, models trained in that way would introduce severe biases as they would not know how to interpret and handle inputs from sources at moderate and low redshift values. For this reason, an additional step should be implemented in the pipeline, which should aim to classify between low and high-redshift sources before the proper redshift estimation step.

Keeping the same emphasis on high-redshift detections, a redshift prediction model would, naturally, have less sources for its training potentially biasing it towards the covered paramenter space. For this reason, it might be useful to include all sorts of AGN in the operation. This change implies using both radio-detected and non-radio detected AGN for the expansion of the

[1] https://stanfordmlgroup.github.io/projects/ngboost/
[2] https://github.com/henrikbostrom/crepes

space of parameters. As it has been shown in Sect. 4.1.3, several redshift-prediction models have been trained with a large variety of AGN giving satisfactory results such as those by Norris et al. (2019), Duncan et al. (2019), and Cunha and Humphrey (2022).

An option for the improvement of the results of our prediction pipeline is using larger amounts of data for its training and calibration steps. Keeping similar qualities in the used observations can be achieved by using training data from the area of the recently released LoTSS-DR2. It covers 5635 $\deg^2$ in the northern sky with 144 MHz observations at a median rms sensitivity of 83 µJy/beam with a 6″ resolution. LoTSS-DR2 has catalogued 4 396 228 radio sources, an increase of more than one order of magnitude from their first data release (325 694 catalogued sources; Shimwell et al. 2019). Arnaudova et al. (2024) hava found 265 578 spectroscopically identified SDSS QSOs in a fraction of the LoTSS-DR2 area. Thus, the number of training sources can be greatly increased without a significant loss of radio sensitivity.

If the research goals are different (i.e. besides the determination of radio-detectable sources at specific redshift ranges), alternative data gathering approaches could be used. For instance, instead of using CW as a base catalogue, other surveys and datasets can be used and combined for the creation of the list of sources to study. Taking into account their expected high-quality observations, catalogues from LSST, *Euclid*, or *Gaia* could be utilised for source selection and training as they are planned to cover large fractions of the sky. In the case of *Euclid*, its measurements could also be used to replace 2M photometry as it is expected to be deeper and cover a more than 14 000 $\deg^2$ of the sky (Euclid Collaboration et al. 2024).

Again, and depending on the research purpose, it might be possible to include variability data for the training of the models in the prediction pipeline. As it has been presented in Sect. 1.1.1, observing a source in different moments can give hints of its possible classification. From the datasets we have used for our pipeline, PS1 contains observations spanning almost five years (Chambers et al. 2016). Thus, this temporal information could be included in the training stages. Additionally, the future observations of LSST will allow the exploration of time evolution of a vey large number of sources, as it has been already shown, for instance, in their LSST AGN data challenge (Savić et al. 2023).

A completely different training approach, which might help solving issues with the sampling of the space of parameters is that of using rest-frame information for creating the models. Following loosely part of the methods from, for instance, Gilda et al. (2021) and Gilda (2024), galaxy formation simulations can be utilised for the extraction of SEDs of sources with

very well-known properties and classifications. Then, each of these SEDs can be redshifted to any desired value and transmission functions can be then, applied to them for the extraction of realistic photometric measurements. These simulated photometric observations can be fed to the training stages of our models. Using such methods would render our prediction pipeline into a ML-based SED-fitting tool which could output a number of estimated galaxy or AGN properties.

Apart from the steps already implemented in our prediction pipeline, it is possible to include additional stages for the estimation of different properties of the created radio-AGN candidates (e.g stellar, or black hole masses of galaxies as well as SFRs). Our pipeline provides a general class (i.e. either AGN or SFG), radio detectability, and a redshift estimate.

6.3 Pipeline application in additional datasets

In order to obtain a successful and useful pipeline, we need to show its capabilities by applying it to as many datasets, fields, and surveys as possible for the creation of radio-AGN candidates. This thesis has shown that our prediction pipeline has the power to produce prospective radio-AGN (as well as radio-SFGs) along with their redshift values. While it is possible to use the pipeline in any region of the sky with coverage on the same bands used for training, our interest is centred on future and in-progress radio surveys. Here, we present two examples of surveys in which the application of our prediction pipeline might enhance the quality of their observations and help improve our knowledge of radio-detected AGN and SFGs.

6.3.1 Evolutionary Map of the Universe

Independently of the degree of knowledge one can have on the inner works of our prediction pipeline and the ways it could be improved or modified, predictions can still be obtained from its application. Even though our prediction pipeline (or, at least, the steps used to create it) can be applied to a large array of situations and surveys for the prediction of a wide range and sources and their properties, our focus is in the radio wavelengths.

Starting with Sects. 5.2 and 5.3, which show possible applications of the prediction pipeline on data from the EMU-PS, more studies can be performed on the area covered by EMU. The full products from the EMU survey, which aims to cover all the southern sky up to a declination +30° with an rms of $\sim 10\,\mu$Jy/beam at 1.3 GHz (Norris et al. 2011), can be subject to our prediction

pipeline and as a way to obtain a large number of radio-AGN candidates. The main barrier to achieve such goal is related to the existence of coverage from deep and homogeneous optical surveys. Our training data incorporates measurements from PS1, which is not fully available in the southern hemisphere. Thus, for the EMU-PS data and as shown in Sect. 5.2, measurements from the Dark Energy Survey (DES; Abbott et al. 2018) can be incorporated for the creation of the base photometry to be fed to our pipeline.

Moreover, the results from Sects. 5.2 and 5.3 have shown that it is possible to classify sources in the EMU-PS field with a set of models trained in a mildly different dataset. Thus, applying the general improvements mentioned in the previous section can be enough for the improvement of the results in this field and specific changes are not essential.

6.3.2 Square Kilometre Array

Ultimately, our prediction pipeline can be applied to the full area covered by the future SKA in order to accelerate the detection of radio-AGN. In its first stage, SKA is expected to detect more than 500 million sources with a $5\,\sigma$ continuum detection limit, at 1.4 GHz, of 10 µJy/beam (Norris et al. 2015; Braun et al. 2019). Our prediction pipeline (or a modified version of it) could be used to pre-select and assess radio-detectable AGN and SFG candidates.

Given their sensitivity, future SKA observations will detect very faint sources that have a strong fraction of their radio emission coming from SFR-related processes. It is expected to have $5\,\sigma$ continuum detection limits between 130 µJy/beam at 110 MHz and 12 µJy/beam at 12.5 GHz (Braun et al. 2019). Thus, further analyses will be needed to separate these sources from AGN-dominated galaxies. The application of our prediction pipeline can help accelerating that differentiation using its output classification probabilities as an indication of the origin of their radio emission. On top of that, using the sources predicted by our pipeline as priors for probable AGN or SFGs can allow SKA detections below its projected limits.

As with the EMU observations, it will be needed to apply some modifications to the prediction pipeline in order to have access to the largest possible fraction of data from the SKA measurements. Implementing some of the changes described in Sect. 6.2, such as using photometry from surveys available in the southern hemisphere might increase the area available for the application of our models. Nevertheless, it has been shown (see Sect. 5.2 and Appendix C) that minor changes of base photometry do not have a very strong impact on the capabilities of our pipeline.

Once a list of radio-AGN candidates has been created for the observed area of the SKA, confirmation of the predictions will be needed. Taking into account the very large volume of data that will be available, optimised processes are needed to access SKA detections. One way to do it is through the application of methods developed for the SKA data challenges (SDCs). In particular, the SDC-1 (Bonaldi et al. 2021), where several methods were developed for the detection and characterisation of radio sources in SKA observations. Catalogues produced from these methods can be contrasted with our predictions for relatively fast confirmation of radio-detected sources. Additionally, and in further stages of our pipeline, the output from these SDCs (catalogues of radio detections together with some of their properties) can be used for the training of our models.

This page intentionally left blank.

Summary

By exploiting the strength of ML-based techniques, this thesis presents the development, application, and analysis of a novel machine-assisted tool for the efficient selection and basic characterisation of radio-detectable AGN from photometric measurements of a sample of IR-detected sources in large-area surveys. Beyond the immediate goal of selecting such candidates, this work aims to obtain clues for understanding the role of AGN in the overall evolution of galaxies throughout the history of the Universe. Crucially, such understanding includes investigating the triggering mechanisms of radio emission in AGN.

Initially, this thesis describes the context for our study. With the need, by the astronomical community, to have a clear perception of the processes that have led to the formation, evolution, and interaction of galaxies, a great number of studies have tried to make sense of observations, simulations, and theory. In order to have a clear view of the origin of galaxies, their investigation across wide redshift ranges has turned crucial. It is apparent that by studying galaxies over different epochs, we could obtain clues on the assembly of the first galaxies and how they got to harbour SMBHs in their centres.

Typically, the study of extra-galactic sources has been possible through a multi-wavelength analysis of their properties and their correlations. Different wavelenghts can give access to different processes in both the central SMBHs and the galaxies that host them.

A prevailing theory among astronomers suggests that all central engines in galaxies are part of the same family of phenomena, AGN. The current unified scheme of AGN proposes that the different kinds populations of AGN are just expressions of the same type of object but observed with different viewing angles that mirror specific processes in the closest regions of the AGN. Different angles and inclinations can trigger emission in several wavelengths as central radiation travels through different structures in the central regions and throughout the galaxy and its surroundings.

Said interacions allow the study of, not only the central engine, but also the different regions of the host galaxy. Such opportunity can become a drawback when the focus is the analysis of the AGN radiation as it can get reflected, re-emitted, or obscured by the intervening medium or mixed with radiation from galactic processes. Historically, bright radio radiation has not suffered such problems as it is not strongly obscured by the galactic medium. This

advantage has been taken profusely through the observation and analysis of bright AGN in radio frequencies.

Recently, such advantage has been diluted by the operation of very sensitive radio observatories which can probe into intermediate and low radio brightness regimes. These sources can have an important fraction of their radio emission coming from SF events in the host galaxies obstructing studies of AGN emission. While such new observations have opened a new window to study SF in distant objects, investigations of AGN have been forced to find alternative ways to separate radio emission from both the central engine and its host.

Fortunately, advancements in observational techniques extend beyond radio astronomy. Across the electromagnetic spectrum, highly sensitive instruments are generating vast datasets of unprecedented depth. These datasets, containing orders of magnitude more objects than previously possible, allow researchers to analyse a wider range of phenomena with much greater detail, including disentangling radio emission from SF events and AGN with the help of these new, multi-wavelength datasets.

After the description of AGN and new observational facilities, we have digged into the complications that such studies have traditionally presented and those that have arrived in the last years. The main group of issues are related to the lack of (mostly computational) tools for the timely analysis of the very large datasets that have been generated lately and those expected to be produced in the coming years. In conjunction with such issues, the community has tried to address them by the use of statistical tools that can analyse very large data volumes in relatively short times and without an excessive use of computing power.

One subset of such tools are those based upon ML methods. They can, from the analysis of very large datasets, extract trends and correlations between their properties that can be used for tasks as diverse as the interpolation or extrapolation of new values, the classification of elements in the datasets, or their dimensionality reduction, among others. In this thesis, we aim at using ML-based methods for the retrieval of radio-detectable AGN candidates. In the process of getting such candidates, we also have the goal of extracting the correlations that such methods have obtained from their data analysis. With that additional knowledge, it might be possible to expand our understanding of the presence of AGN in galaxies and their history.

This is the point in which our work can be fully stated. We set ourselves to develop a prediction pipeline which, from the ingestion of multi-wavelength photometry of IR-detected sources, can deliver a list of highly probable radio-detectable AGN together with an estimate of

their photometric redshift. Most importantly, we have also described a series of methodologies to understand the driving properties of the different decisions made by each of the steps of our pipeline. For the generation (training) of such pipeline, we have used information from 94 987 $z \lesssim 5$ spectroscopically identified IR-detected sources in the HETDEX field and created stacked models with them to be applied in the same region and in the area of the S82 field.

The HETDEX field, as covered by the LoTSS-DR1 survey, is an area of 424 deg^2 observed with the LOFAR at 150 MHz with a median sensitivity of 71 μJy/beam. The 92 deg^2 area of the S82 region covered by the VLAS82 survey has VLA measurements with a median rms noise of 52 μJy/beam at 1.4 GHz. Assuming a synchrotron radio slope of $\alpha = -0.7$, the observations of VLAS82 are shallower than those from the LoTSS-DR1 for observations of non-thermal phenomena, such as radio AGN. In these areas, MIR detections (from the CW survey) were selected as the starting point of the data collection. These sources were cross-matched with optical photometry (from the PS1 survey), IR (from 2M and AW), and the radio (LoTSS-DR1 or VLAS82, depending on the area).

These models were applied, sequentially, to 15 018 144 IR detections without a classification in the HETDEX Spring field, arriving to the creation of 68 252 radio AGN candidates with their corresponding predicted redshift values. Additionally, we applied the models to 3 568 478 unlabelled IR detections in the S82 field, obtaining 22 445 new radio AGN candidates with their predicted redshift values (up to values of $z \lesssim 4.4$ in both datasets).

To assess the quality of our estimations we applied the models in our prediction pipeline to 9499 and 21 828 confirmed IR-detected sources in both the HETDEX and S82 fields, repectively. We have, then, applied a number of analyses on the models to understand the influence of the observed properties on the predictions and their confidence levels. In particular, the use of game theory analysis (SHAP values) gives the opportunity to extract the influence that the feature set has for each individual prediction.

From the application of the prediction pipeline on labelled and unlabelled sources and the analysis of the predictions and the models themselves, the following conclusions can be drawn.

- Generalised stacking (the combination of different models for the prediction of one single property or quantity) is a useful procedure which collects results from individual ML algorithms into a single model that can outperform each of the individual models, while preventing the inclusion of biases from individual algorithms. Proper selection of models and input features, together with detailed probability and threshold calibration maximises

the target metrics of the final model.

- Classification between AGN and SFGs derived from our model is, including uncertainties, in line with previous traditional and ML-based works. The first step of our pipeline is able to retrieve a high fraction of previously-classified AGN from HETDEX with a recall = (96.21 ± 0.43) % and a precision = (94.49 ± 0.65) %. From the S82 field, we can obtain a recall = (94.01 ± 0.59) % and a precision = (94.81 ± 0.40) %. For reference, the base, no-skill classification between AGN and SFGs gives a recall = 42.57 % and a precision = 42.57 % for the sources in the HETDEX field. In the case of the sources in the S82 field, we obtain a recall = 81.29 % and a precision = 81.29 %. Such improvement in our results implies that the features used for training contain already enough information to provide highly reliable classification of sources.

- Radio detection classification for predicted AGN has proven to be highly demanding in terms of data needed for creating the models. Thanks to the use of the techniques shown in this work (i.e. feature creation and selection, generalised stacking, probability calibration, and threshold optimisation), we are able to retrieve previously-known radio-detectable AGN in the HETDEX field with a recall = (52.16 ± 3.59) % and a precision = (35.28 ± 2.74) %. In the S82 field, we can obtain a recall = (58.16 ± 3.06) % and a precision = (12.29 ± 0.73) %. These rates improve significantly upon a purely random selection, with a recall = 12.84 % and a precision = 12.84 % for the HETDEX field, and a recall = 4.59 % and a precision = 4.59 % for the sources in the S82 field. Focusing on the recall (completeness), these improvements correspond to, roughly, 4 times better for the HETDEX field and 13 times better for S82, showing the power of ML methods for obtaining new RG candidates.

- When combining both predictions (classification between AGN and SFGs with radio detection prediction), the effects, uncertainties, and success rates merge together. In this way, and for the sources in the HETDEX field, the joint prediction of radio-detectable AGN has, as metrics, a recall = (44.61 ± 2.46) % and a precision = (32.20 ± 2.72) %. Sources in the S82 field, subject to the same procedure, have a recall = (47.36 ± 6.22) % and a precision = (11.33 ± 1.32) %. No-skill versions of these predictions give a recall = 5.47 % and a precision = 5.47 % for HETDEX and a recall = 3.73 % and a precision = 3.73 % for sources in the S82 field. For sources in both fields, the improvement upon a random

selection is unmistakable, highlighting the power of our pipeline and the inclusion of the selected features.

- The prediction of redshift values for sources classified to be radio-detectable AGN can deliver results that are in line with works that use either traditional or ML methods. For predicted radio-AGN in the HETDEX field, we obtain a NMAD of $\sigma_{\rm NMAD} = (7.17 \pm 0.81)$ % and an outiler fraction of $\eta = (18.91 \pm 1.59)$ %, while for the S82 field these values are $\sigma_{\rm NMAD} = (9.84 \pm 0.56)$ % and $\eta = (25.18 \pm 2.26)$ %.

- Our models (classification and regression) can be applied to areas of the sky which have different radio coverage from that used for training without a strong degradation of the prediction results (minor issues could appear in the vicinity of the detection limits of the training set and in the application measurements). This feature can lead to the use of our pipeline over very distinct datasets (in radio and multi-wavelength coverage) expecting to recover the sources predicted to be radio-detectable AGN with a high probability.

After the quality of the prediction from our pipeline has been assessed and established, we used its models for the expansion of our knowledge on some of the properties from radio-detectable AGN. In this way, and from the analysis of feature importances, we were able to derive a novel colour-colour diagnostic criterion for the selection of AGN. From the combination of (g - r) and ($W1$ - $W2$) colours (from PS1, and CW, respectively) and the boundaries defined in Eqs. 5.1, 5.2, and 5.3, it is possible to discern, with a high success rate, between AGN and SFGs in the studied samples. The metrics from the newly derived colour-colour criterion, C23, are in line with the scores from traditional IR-based selection criteria with the exception of the recall, highlighting the fact that C23 can provide more candidates than previous criteria.

While we did not compare, directly, the application of our pipeline with the use of BPT, VO, and WHAN diagrams for the selection of AGN, our method shows an advantage on the observational requirements. Our prediction pipeline can be used with photometry and does not need deep and expensive spectroscopic measurements for its application. It is even possible to have missing measurements and, still, obtain reliable extra-galactic source classifications through the use of imputation. The same conclusion applies when comparing our method with the use of SED fitting techniques. Our pipeline has the potential to classify a very large number of sources using a fraction of the time needed by typical SED fitting tools and without relying on SED models.

Additionally, and as a way to estimate the density of radio-detected AGN and SFGs, we utilised the predicted candidates in the field of the EMU-PS survey to build RLFs in different redshift bins. The EMU-PS survey is aimed at testing the capabilities of the in-progress EMU survey. Over an area of 270 deg^2 in the southern sky, it has a 25 μJy/beam to 30 μJy/beam rms depth at 944 MHz with a spatial resolution of $18''$. We are able, by applying corrections drawn from the use of the prediction pipeline on the EMU-PS area, to obtain RLFs that match with the results from the literature. Our calculations, from the large number of predicted sources, present small uncertainties, constraining the estimated densities much strongly than previous works. In order to derive meaningful values for AGN RLFs, we expanded our prediction pipeline by including an additional branch for the analysis of predicted SFGs. This new section takes the sources predicted to be SFGs and classify them between sources that could, or not, have radio detections. Those SFGs predicted to be radio-detectable sources are subject to a final estimation of their photometric redshift values.

Another possible use of the outputs from our prediction pipeline was tested in this work. We analysed the probabilities given by the models to the studied IR-detected sources (the likelihood to be AGN and radio detectable) and plotted them in the EMU-PS maps. In this way, it is possible to estimate the IR source that is the most likely counterpart of EMU-PS detections. Displaying some examples, we were able to show that the counterparts intially assigned for EMU-PS are not always the most likely match. Having this by-product from the prediction pipeline has the potential of speeding up the spatial correlations of sources in different wavelengths based upon the statistical analysis of their photometry.

Future developments of the pipeline will concentrate on minimising the existent biases in the training sample as well as in increasing the coverage of the parameter space. We also plan to generalise the pipeline to make it useful for non-radio or galaxy-related research communities (as described in Chapter 6 and Appendix C). These developments include, for instance, the capability to carry out the full analysis for the galactic or stellar populations. In other words, models to separate sources between AGN, SFGs, and stars which can determine if any of these sources can be detected in the radio and to predict redshift values for SFGs and non-radio AGN or distances for stars.

In order to increase the parameter space of our training sets, we plan to include information from radio surveys with different characteristics. Namely, shallower, but with larger area, and less extended but with deeper multi-wavelength data. Similarly, the inclusion of FIR, X-ray, and

multi-frequency radio measurements makes part of our efforts to improve detections, not only in radio, but in additional wavelengths. Such increase of the parameter space should also be linked to the addition of measurement uncertainties (both as input and output properties), which might add a sense of the confidence it is possible to have in the models.

With the next generation of observatories already producing source catalogues with an order of magnitude better sensitivity over large areas of the sky than previously, such as the Rapid ASKAP Continuum Survey (RACS; McConnell et al. 2020), EMU, and MIGHTEE, the need to understand the fraction of those radio detections related to AGN and determine counterparts across wavelengths is more necessary than ever. Although we developed our prediction pipeline as a tool to better understand the aforementioned issues, we foresee additional possibilities in which the pipeline can be of great use. One of these possibilities involves the use of the pipeline to assist with the selection of radio-detectable AGN within any set of observations. This application might turn particularly valuable in recent surveys carried out with MeerKAT or the future SKA and ngVLA where the population at the faintest sources will be dominated by SFGs. This change needs to use the corresponding data in the training set in order to extract the largest possible amount of information from the catalogues.

As increasingly larger simulations and observations unveil a great number of unanswered questions in the co-evolution of galaxies, SMBHs, and their environment, our work emerges as a crucial link in the chain of tools designed to exploit the unprecedented volumes of data available for the study and characterisation of galaxies and AGN. With a focus on the earliest sources in the Universe, and coupled with very high-quality radio observations, our new prediction pipeline can be on a par with current state-of-the-art methods for the selection of radio-detected AGN and galaxies. It is only to be expected that the number and quality of observations of the extra-galactic sky will increase very rapidly in the coming years and we believe that our work can pave the way for unprecedented discoveries in the radio sky.

This page intentionally left blank.

Data and software acknowledgements

This work made use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

LOFAR data products were provided by the LOFAR Surveys Key Science project (LSKSP[3]) and were derived from observations with the International LOFAR Telescope (ILT). LOFAR (van Haarlem et al. 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and which are collectively operated by the ILT foundation under a joint scientific policy. The efforts of the LSKSP have benefited from funding from the European Research Council, NOVA, NWO, CNRS-INSU, the SURF Co-operative, the UK Science and Technology Funding Council and the Jülich Supercomputing Centre.

The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eötvös Loránd University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

This thesis makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.

This work made use of public data from the Sloan Digital Sky Survey, Data Release 16. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance

[3] https://lofar-surveys.org/

Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This scientific work uses data obtained from Inyarrimanha Ilgari Bundara / the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

Part of this work is based on data obtained from the ESO Science Archive Facility with DOI: 10.18727/archive/56.

This project used public archival data from the Dark Energy Survey (DES). Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions

in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the OzDES Membership Consortium, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University. Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

This project makes use of the MaNGA-Pipe3D dataproducts. We thank the IA-UNAM MaNGA team for creating this catalogue, and the Conacyt Project CB-285080 for supporting them.

This research has made use of NASA's Astrophysics Data System, TOPCAT[4] (Taylor 2005), `JupyterLab`[5] (`v4.2.4`; Kluyver et al. 2016; for Jupyter Notebooks), Aladin sky atlas (`v11.0.24`; Bonnarel et al. 2000), CDS, Strasbourg Astronomical Observatory, France, the VizieR catalogue access tool (Ochsenbein et al. 2000), CDS, Strasbourg Astronomical Observatory, France (DOI: 10.26093/cds/vizier), the CDS cross-match service (Boch et al. 2012; Pineau et al. 2020), Strasbourg Astronomical Observatory, France, and the SIMBAD database (Wenger et al. 2000), CDS, Strasbourg Astronomical Observatory, France. This research has also made use of `hips2fits`,[6] (Boch et al. 2020) a service provided by CDS.

[4] http://www.star.bris.ac.uk/~mbt/topcat/
[5] https://github.com/jupyterlab/jupyterlab
[6] https://alasky.cds.unistra.fr/hips-image-services/hips2fits

This research has made use of the Spanish Virtual Observatory[7] (Rodrigo and Solano 2020; Rodrigo et al. 2012) project funded by MCIN/AEI/10.13039/501100011033/ through grant PID2020-112949GB-I00

This work made extensive use of the `Python` packages `NumPy`[8] (`v1.20.3`; Harris et al. 2020), `SciPy`[9] (`v1.5.3`; Virtanen et al. 2020), `PyCaret`[10] (`v2.3.10`; Ali 2020), `scikit-learn` (`v0.23.2`; Pedregosa et al. 2011), `pandas`[11] (`v1.4.2`; McKinney 2010), `Astropy`[12], a community-developed core `Python` package for Astronomy (`v5.0`; Astropy Collaboration et al. 2013, 2018, 2022), `Matplotlib` (`v3.5.1`; Hunter 2007), `betacal`[13] (`v1.1.0`; Kull et al. 2017a,b), `CMasher`[14] (`v1.6.3`; van der Velden 2020), `Colorcet`[15] (`v3.1.0`; Glasbey et al. 2007; Kovesi 2015), `faiss`[16] (`v1.7.2`; Johnson et al. 2019), `skyproj`[17] (`v1.2.2`), `PyCCL`[18] (`v2.6.1`; Chisari et al. 2019), `schemdraw`[19] (`v0.14`), and `mocpy`[20] (`v0.13.1`; Baumann et al. 2023).

[7] https://svo.cab.inta-csic.es
[8] https://numpy.org
[9] https://scipy.org
[10] https://pycaret.org
[11] https://pandas.pydata.org
[12] https://www.astropy.org
[13] https://betacal.github.io
[14] https://github.com/1313e/CMasher
[15] https://colorcet.holoviz.org
[16] https://faiss.ai
[17] https://github.com/LSSTDESC/skyproj
[18] https://github.com/LSSTDESC/CCL
[19] https://github.com/cdelker/schemdraw
[20] https://github.com/cds-astro/mocpy/

References

Abbott, T. M. C., Abdalla, F. B., et al. (Dec. 2018). 'The Dark Energy Survey: Data Release 1'. In: ApJS 239.2, 18, p. 18. DOI: 10.3847/1538-4365/aae9f0 (cit. on pp. 18, 156).

Abbott, T. M. C., Adamów, M., et al. (Aug. 2021). 'The Dark Energy Survey Data Release 2'. In: ApJS 255.2, 20, p. 20. DOI: 10.3847/1538-4365/ac00b3 (cit. on p. 119).

Abdo, A. A., Ackermann, M., et al. (May 2010). 'The First Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope'. In: ApJ 715.1, pp. 429–457. DOI: 10.1088/0004-637X/715/1/429 (cit. on p. 22).

Abdurro'uf, Accetta, K., et al. (Apr. 2022). 'The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data'. In: ApJS 259.2, 35, p. 35. DOI: 10.3847/1538-4365/ac4414 (cit. on p. 113).

Abramson, I. S. (1982). 'On Bandwidth Variation in Kernel Estimates-A Square Root Law'. In: *The Annals of Statistics* 10.4, pp. 1217–1223. ISSN: 00905364. URL: http://www.jstor.org/stable/2240724 (visited on 20/01/2024) (cit. on p. 213).

Adam, A., Perreault-Levasseur, L., et al. (July 2023). 'Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems Using Recurrent Inference Machines'. In: ApJ 951.1, 6, p. 6. DOI: 10.3847/1538-4357/accf84 (cit. on p. 24).

Adelman-McCarthy, J. K., Agüeros, M. A., et al. (Apr. 2008). 'The Sixth Data Release of the Sloan Digital Sky Survey'. In: ApJS 175.2, pp. 297–313. DOI: 10.1086/524984 (cit. on p. 36).

Aguado, D. S., Ahumada, R., et al. (Feb. 2019). 'The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library'. In: ApJS 240.2, 23, p. 23. DOI: 10.3847/1538-4365/aaf651 (cit. on pp. 87, 89).

Agüeros, M. A., Ivezić, Ž., et al. (Sept. 2005). 'The Ultraviolet, Optical, and Infrared Properties of Sloan Digital Sky Survey Sources Detected by GALEX'. In: AJ 130.3, pp. 1022–1036. DOI: 10.1086/432160 (cit. on p. 22).

Ahumada, R., Prieto, C. A., et al. (July 2020). 'The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra'. In: ApJS 249.1, 3, p. 3. DOI: 10.3847/1538-4365/ab929e (cit. on p. 44).

Aihara, H., Allende Prieto, C., et al. (Apr. 2011). 'The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III'. In: ApJS 193.2, 29, p. 29. DOI: 10.1088/0067-0049/193/2/29 (cit. on p. 16).

Alegre, L., Sabater, J., et al. (Nov. 2022). 'A machine-learning classifier for LOFAR radio galaxy cross-matching techniques'. In: MNRAS 516.4, pp. 4716–4738. DOI: 10.1093/mnras/stac1888 (cit. on pp. 23, 28).

Ali, M. (Apr. 2020). *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 2.3. URL: https://www.pycaret.org (cit. on p. 170).

Allen, D. M. (1974). 'The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction'. In: *Technometrics* 16.1, pp. 125–127. DOI: 10.1080/00401706.1974.10489157 (cit. on p. 64).

Allison, P. (2001). *Missing Data*. Quantitative Applications in the Social Sciences. SAGE Publications. ISBN: 9781452207902 (cit. on p. 40).

Almosallam, I. A., Jarvis, M. J., and Roberts, S. J. (Oct. 2016a). 'GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts'. In: MNRAS 462.1, pp. 726–739. DOI: 10.1093/mnras/stw1618 (cit. on p. 93).

Almosallam, I. A., Lindsay, S. N., et al. (Jan. 2016b). 'A sparse Gaussian process framework for photometric redshift estimation'. In: MNRAS 455.3, pp. 2387–2401. DOI: 10.1093/mnras/stv2425 (cit. on p. 92).

Alqasim, A. and Page, M. J. (Apr. 2023). 'A new method to determine X-ray luminosity functions of AGN and their evolution with redshift'. In: MNRAS 520.3, pp. 3827–3846. DOI: 10.1093/mnras/stad007 (cit. on pp. 126, 131, 132).

Amarantidis, S., Afonso, J., et al. (May 2019). 'The first supermassive black holes: indications from models for future observations'. In: MNRAS 485.2, pp. 2694–2709. DOI: 10.1093/mnras/stz551 (cit. on pp. 6, 144).

An, T., Zhang, Y., and Frey, S. (Sept. 2020). 'A method for checking high-redshift identification of radio AGNs'. In: MNRAS 497.2, pp. 2260–2264. DOI: 10.1093/mnras/staa2132 (cit. on p. 5).

Ananna, T. T., Salvato, M., et al. (Nov. 2017). 'AGN Populations in Large-volume X-Ray Surveys: Photometric Redshifts and Population Types Found in the Stripe 82X Survey'. In: ApJ 850.1, 66, p. 66. DOI: 10.3847/1538-4357/aa937d (cit. on pp. 16, 22, 91, 93).

Anbajagane, D., Evrard, A. E., and Farahi, A. (Jan. 2022). 'Baryonic imprints on DM haloes: population statistics from dwarf galaxies to galaxy clusters'. In: MNRAS 509.3, pp. 3441–3461. DOI: 10.1093/mnras/stab3177 (cit. on p. 28).

Andonie, C., Alexander, D. M., et al. (Dec. 2022). 'A panchromatic view of infrared quasars: excess star formation and radio emission in the most heavily obscured systems'. In: MNRAS 517.2, pp. 2577–2598. DOI: 10.1093/mnras/stac2800 (cit. on p. 7).

Aniyan, A. K. and Thorat, K. (June 2017). 'Classifying Radio Galaxies with the Convolutional Neural Network'. In: ApJS 230.2, 20, p. 20. DOI: 10.3847/1538-4365/aa7333 (cit. on p. 90).

Annis, J., Soares-Santos, M., et al. (Oct. 2014). 'The Sloan Digital Sky Survey Coadd: 275 deg^2 of Deep Sloan Digital Sky Survey Imaging on Stripe 82'. In: ApJ 794.2, 120, p. 120. DOI: 10.1088/0004-637X/794/2/120 (cit. on pp. 34, 36).

Antoniucci, S., Giannini, T., et al. (Feb. 2014). 'On the Mid-infrared Variability of Candidate Eruptive Variables (EXors): A Comparison between Spitzer and WISE Data'. In: ApJ 782.1, 51, p. 51. DOI: 10.1088/0004-637X/782/1/51 (cit. on pp. 8, 10).

Antonucci, R. (Jan. 1993). 'Unified models for active galactic nuclei and quasars.' In: ARA&A 31, pp. 473–521. DOI: 10.1146/annurev.aa.31.090193.002353 (cit. on p. 2).

Arévalo, P., Uttley, P., et al. (Sept. 2008). 'Correlated X-ray/optical variability in the quasar MR2251-178'. In: MNRAS 389.3, pp. 1479–1488. DOI: 10.1111/j.1365-2966.2008.13719.x (cit. on p. 12).

Arévalo, P., Uttley, P., et al. (Aug. 2009). 'Correlation and time delays of the X-ray and optical emission of the Seyfert Galaxy NGC 3783'. In: MNRAS 397.4, pp. 2004–2014. DOI: 10.1111/j.1365-2966.2009.15110.x (cit. on p. 12).

Arnaudova, M. I., Smith, D. J. B., et al. (Mar. 2024). 'Exploring the radio loudness of SDSS quasars with spectral stacking'. In: MNRAS 528.3, pp. 4547–4567. DOI: 10.1093/mnras/stae233 (cit. on p. 154).

Arnouts, S., Cristiani, S., et al. (Dec. 1999). 'Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North'. In: MNRAS 310.2, pp. 540–556. DOI: 10.1046/j.1365-8711.1999.02978.x (cit. on p. 17).

Arsioli, B. and Dedin, P. (Oct. 2020). 'Machine learning applied to multifrequency data in astrophysics: blazar classification'. In: MNRAS 498.2, pp. 1750–1764. DOI: 10.1093/mnras/staa2449 (cit. on p. 41).

Assef, R. J., Stern, D., et al. (July 2013). 'Mid-infrared Selection of Active Galactic Nuclei with the Wide-field Infrared Survey Explorer. II. Properties of WISE-selected Active Galactic Nuclei in the NDWFS Boötes Field'. In: ApJ 772.1, 26, p. 26. DOI: 10.1088/0004-637X/772/1/26 (cit. on pp. 11, 113).

Assef, R. J., Stern, D., et al. (Feb. 2018). 'The WISE AGN Catalog'. In: ApJS 234.2, 23, p. 23. DOI: 10.3847/1538-4365/aaa00a (cit. on p. 11).

Astropy Collaboration, Price-Whelan, A. M., et al. (Sept. 2018). 'The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package'. In: AJ 156.3, 123, p. 123. DOI: 10.3847/1538-3881/aabc4f (cit. on p. 170).

Astropy Collaboration, Price-Whelan, A. M., et al. (Aug. 2022). 'The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package'. In: ApJ 935.2, 167, p. 167. DOI: 10.3847/1538-4357/ac7c74 (cit. on pp. 20, 170).

Astropy Collaboration, Robitaille, T. P., et al. (Oct. 2013). 'Astropy: A community Python package for astronomy'. In: A&A 558, A33, A33. DOI: 10.1051/0004-6361/201322068 (cit. on p. 170).

Atek, H., Chemerynska, I., et al. (Oct. 2023). 'JWST UNCOVER: discovery of z > 9 galaxy candidates behind the lensing cluster Abell 2744'. In: MNRAS 524.4, pp. 5486–5496. DOI: 10.1093/mnras/stad1998 (cit. on p. 17).

Auge, C., Sanders, D., et al. (Nov. 2023). 'The Accretion History of AGN: The Spectral Energy Distributions of X-Ray-luminous Active Galactic Nuclei'. In: ApJ 957.1, 19, p. 19. DOI: 10.3847/1538-4357/acf21a (cit. on pp. 7, 22).

Avni, Y. and Bahcall, J. N. (Feb. 1980). 'On the simultaneous analysis of several complete samples. The V/Vmax and Ve/Va variables, with applications to quasars.' In: ApJ 235, pp. 694–716. DOI: 10.1086/157673 (cit. on p. 212).

Bahcall, J. N. and Kozlovsky, B.-Z. (Mar. 1969). 'Some Models of the Emission-Line Region of 3c 273'. In: ApJ 155, p. 1077. DOI: 10.1086/149935 (cit. on p. 2).

Baldassare, V. F., Geha, M., and Greene, J. (Dec. 2018). 'Identifying AGNs in Low-mass Galaxies via Long-term Optical Variability'. In: ApJ 868.2, 152, p. 152. DOI: 10.3847/1538-4357/aae6cf (cit. on p. 36).

Baldwin, J. A., Phillips, M. M., and Terlevich, R. (Feb. 1981). 'Classification parameters for the emission-line spectra of extragalactic objects.' In: PASP 93, pp. 5–19. DOI: 10.1086/130766 (cit. on p. 7).

Ball, N. M. and Brunner, R. J. (Jan. 2010). 'Data Mining and Machine Learning in Astronomy'. In: *International Journal of Modern Physics D* 19.7, pp. 1049–1106. DOI: 10.1142/S0218271810017160 (cit. on p. 24).

Ball, N. M., Brunner, R. J., et al. (Aug. 2008). 'Robust Machine Learning Applied to Astronomical Data Sets. III. Probabilistic Photometric Redshifts for Galaxies and Quasars in the SDSS and GALEX'. In: ApJ 683.1, pp. 12–21. DOI: 10.1086/589646 (cit. on p. 39).

Baltay, C., Grossman, L., et al. (Apr. 2021). 'Low-redshift Type Ia Supernova from the LSQ/LCO Collaboration'. In: PASP 133.1022, 044002, p. 044002. DOI: 10.1088/1538-3873/abd417 (cit. on p. 15).

Barbieri, C. and Bertola, F. (Jan. 1972). 'Identification of 5C4 radio sources.' In: MNRAS 156, pp. 399–409. DOI: 10.1093/mnras/156.4.399 (cit. on p. 22).

Baron, D. (Apr. 2019). 'Machine Learning in Astronomy: a practical overview'. In: *arXiv e-prints*, arXiv:1904.07248, arXiv:1904.07248 (cit. on p. 24).

Baron, D. and Poznanski, D. (Mar. 2017). 'The weirdest SDSS galaxies: results from an outlier detection algorithm'. In: MNRAS 465.4, pp. 4530–4555. DOI: 10.1093/mnras/stw3021 (cit. on pp. 24, 25).

Barrows, R. S., Comerford, J. M., et al. (Dec. 2021). 'A Catalog of Host Galaxies for WISE-selected AGN: Connecting Host Properties with Nuclear Activity and Identifying Contaminants'. In: ApJ 922.2, 179, p. 179. DOI: 10.3847/1538-4357/ac1352 (cit. on p. 11).

Baum, W. A. (Feb. 1957). 'Photoelectric determinations of redshifts beyond 0.2 c.' In: AJ 62, pp. 6–7. DOI: 10.1086/107433 (cit. on p. 16).

— (Jan. 1962). 'Photoelectric Magnitudes and Red-Shifts'. In: *Problems of Extra-Galactic Research*. Ed. by G. C. McVittie. Vol. 15, p. 390 (cit. on p. 16).

Baumann, M., Marchand, M., et al. (Dec. 2023). *cds-astro/mocpy: Release v0.13.1*. Version v0.13.1. DOI: 10.5281/zenodo.10257390 (cit. on p. 170).

Beifiori, A., Courteau, S., et al. (Jan. 2012). 'On the correlations between galaxy properties and supermassive black hole mass'. In: MNRAS 419.3, pp. 2497–2528. DOI: 10.1111/j.1365-2966.2011.19903.x (cit. on p. 6).

Benítez, N. (June 2000). 'Bayesian Photometric Redshift Estimation'. In: ApJ 536.2, pp. 571–583. DOI: 10.1086/308947 (cit. on p. 17).

Best, P. N. and Heckman, T. M. (Apr. 2012). 'On the fundamental dichotomy in the local radio-AGN population: accretion, evolution and host galaxy properties'. In: MNRAS 421.2, pp. 1569–1582. DOI: 10.1111/j.1365-2966.2012.20414.x (cit. on pp. 14, 137).

Best, P. N., Ker, L. M., et al. (Nov. 2014). 'The cosmic evolution of radio-AGN feedback to z = 1'. In: MNRAS 445.1, pp. 955–969. DOI: 10.1093/mnras/stu1776 (cit. on p. 137).

Best, P. N., Kondapally, R., et al. (Aug. 2023). 'The LOFAR Two-metre Sky Survey: Deep Fields data release 1. V. Survey description, source classifications, and host galaxy properties'. In: MNRAS 523.2, pp. 1729–1755. DOI: 10.1093/mnras/stad1308 (cit. on p. 10).

Bianchi, L., Rodriguez-Merino, L., et al. (Dec. 2007). 'Statistical Properties of the GALEX-SDSS Matched Source Catalogs, and Classification of the UV Sources'. In: ApJS 173.2, pp. 659–672. DOI: 10.1086/516648 (cit. on p. 22).

Bicknell, G. V., Dopita, M. A., and O'Dea, C. P. O. (Aug. 1997). 'Unification of the Radio and Optical Properties of Gigahertz Peak Spectrum and Compact Steep-Spectrum Radio Sources'. In: ApJ 485.1, pp. 112–124. DOI: 10.1086/304400 (cit. on pp. 2, 7).

Birchall, K. L., Watson, M. G., and Aird, J. (Feb. 2020). 'X-ray detected AGN in SDSS dwarf galaxies'. In: MNRAS 492.2, pp. 2268–2284. DOI: 10.1093/mnras/staa040 (cit. on p. 7).

Blandford, R., Meier, D., and Readhead, A. (Aug. 2019). 'Relativistic Jets from Active Galactic Nuclei'. In: ARA&A 57, pp. 467–509. DOI: 10.1146/annurev-astro-081817-051948 (cit. on p. 1).

Blecha, L., Snyder, G. F., et al. (Aug. 2018). 'The power of infrared AGN selection in mergers: a theoretical study'. In: MNRAS 478.3, pp. 3056–3071. DOI: 10.1093/mnras/sty1274 (cit. on pp. xxiii, 11, 84).

Blum, A. L. and Langley, P. (1997). 'Selection of relevant features and examples in machine learning'. In: *Artificial Intelligence* 97.1. Relevance, pp. 245–271. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(97)00063-5 (cit. on p. 62).

Boch, T., Fernique, P., et al. (Jan. 2020). 'HiPS2FITS: Fast Generation of FITS Cutouts From HiPS Image Datasets'. In: *Astronomical Data Analysis Software and Systems XXIX*. Ed. by R. Pizzo, E. R. Deul, et al. Vol. 527. Astronomical Society of the Pacific Conference Series, p. 121 (cit. on p. 169).

Boch, T., Pineau, F., and Derriere, S. (Sept. 2012). 'The CDS Cross-Match Service'. In: *Astronomical Data Analysis Software and Systems XXI*. Ed. by P. Ballester, D. Egret, and N. P. F. Lorente. Vol. 461. Astronomical Society of the Pacific Conference Series, p. 291 (cit. on p. 169).

Böhme, L., Schwarz, D. J., et al. (June 2023). 'Matching LOFAR sources across radio bands'. In: A&A 674, A189, A189. DOI: 10.1051/0004-6361/202245669 (cit. on p. 23).

Bolzonella, M., Miralles, J. .-., and Pelló, R. (Nov. 2000). 'Photometric redshifts based on standard SED fitting procedures'. In: A&A 363, pp. 476–492. DOI: 10.48550/arXiv.astro-ph/0003380 (cit. on p. 16).

Bonaldi, A., An, T., et al. (Jan. 2021). 'Square Kilometre Array Science Data Challenge 1: analysis and results'. In: MNRAS 500.3, pp. 3821–3837. DOI: 10.1093/mnras/staa3023 (cit. on p. 157).

Bonaldi, A., Bonato, M., et al. (Jan. 2019). 'The Tiered Radio Extragalactic Continuum Simulation (T-RECS)'. In: MNRAS 482.1, pp. 2–19. DOI: 10.1093/mnras/sty2603 (cit. on p. 6).

Bonato, M., Prandoni, I., et al. (Jan. 2021). 'New constraints on the 1.4 GHz source number counts and luminosity functions in the Lockman Hole field'. In: MNRAS 500.1, pp. 22–33. DOI: 10.1093/mnras/staa3218 (cit. on pp. 10, 135).

Bonnarel, F., Fernique, P., et al. (Apr. 2000). 'The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources'. In: A&AS 143, pp. 33–40. DOI: 10.1051/aas:2000331 (cit. on p. 169).

Borisov, V., Leemann, T., et al. (2022). 'Deep Neural Networks and Tabular Data: A Survey'. In: *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21. DOI: 10.1109/TNNLS.2022.3229161 (cit. on p. 63).

Bosman, S. E. I. (Feb. 2022). *All z>5.7 quasars currently known*. Version 1.15. Zenodo. DOI: 10.5281/zenodo.6039724 (cit. on p. 14).

Boström, H. (2022). 'crepes: a Python Package for Generating Conformal Regressors and Predictive Systems'. In: *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction and Applications*. Ed. by U. Johansson, H. Boström, et al. Vol. 179. Proceedings of Machine Learning Research. PMLR (cit. on p. 153).

Bouwens, R., González-López, J., et al. (Oct. 2020). 'The ALMA Spectroscopic Survey Large Program: The Infrared Excess of z = 1.5-10 UV-selected Galaxies and the Implied High-redshift Star Formation History'. In: ApJ 902.2, 112, p. 112. DOI: 10.3847/1538-4357/abb830 (cit. on p. 17).

Bowler, R. A. A., Adams, N. J., et al. (Mar. 2021). 'The rapid transition from star formation to AGN-dominated rest-frame ultraviolet light at z ≃ 4'. In: MNRAS 502.1, pp. 662–677. DOI: 10.1093/mnras/stab038 (cit. on p. 4).

Bowles, M., Tang, H., et al. (June 2023). 'Radio galaxy zoo EMU: towards a semantic radio galaxy morphology taxonomy'. In: MNRAS 522.2, pp. 2584–2600. DOI: 10.1093/mnras/stad1021 (cit. on p. 14).

Bradbury, J., Frostig, R., et al. (2018). *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. URL: http://github.com/google/jax (cit. on p. 20).

Bramich, D. M., Vidrih, S., et al. (May 2008). 'Light and motion in SDSS Stripe 82: the catalogues'. In: MNRAS 386.2, pp. 887–902. DOI: 10.1111/j.1365-2966.2008.13053.x (cit. on p. 36).

Brammer, G. B., van Dokkum, P. G., and Coppi, P. (Oct. 2008). 'EAZY: A Fast, Public Photometric Redshift Code'. In: ApJ 686.2, pp. 1503–1513. DOI: 10.1086/591786 (cit. on pp. 17, 92).

Brandt, W. N. and Alexander, D. M. (Jan. 2015). 'Cosmic X-ray surveys of distant active galaxies. The demographics, physics, and ecology of growing supermassive black holes'. In: A&A Rev. 23, 1, p. 1. DOI: 10.1007/s00159-014-0081-z (cit. on p. 2).

Braun, R., Bonaldi, A., et al. (Dec. 2019). 'Anticipated Performance of the Square Kilometre Array – Phase 1 (SKA1)'. In: *arXiv e-prints*, arXiv:1912.12699, arXiv:1912.12699. DOI: 10.48550/arXiv.1912.12699 (cit. on pp. 5, 156).

Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale Paris (cit. on p. 62).

Breedt, E., Arévalo, P., et al. (Mar. 2009). 'Long-term optical and X-ray variability of the Seyfert galaxy Markarian 79'. In: MNRAS 394.1, pp. 427–437. DOI: 10.1111/j.1365-2966.2008.14302.x (cit. on p. 12).

Breedt, E., McHardy, I. M., et al. (Apr. 2010). 'Twelve years of X-ray and optical variability in the Seyfert galaxy NGC 4051'. In: MNRAS 403.2, pp. 605–619. DOI: 10.1111/j.1365-2966.2009.16146.x (cit. on p. 12).

Breiman, L. (Aug. 1996). 'Bagging predictors'. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655 (cit. on p. 25).

— (Oct. 2001). 'Random Forests'. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on pp. 27, 64).

— (2003). 'Manual on setting up, using, and understanding random forests v4. 0'. In: *Statistics Department University of California Berkeley, CA, USA* (cit. on p. 27).

Brescia, M., Cavuoti, S., et al. (2021). 'Photometric Redshifts With Machine Learning, Lights and Shadows on a Complex Data Science Use Case'. In: *Frontiers in Astronomy and Space Sciences* 8, p. 70. ISSN: 2296-987X. DOI: 10.3389/fspas.2021.658229 (cit. on pp. 16, 18, 21, 90).

Brescia, M., Salvato, M., et al. (Oct. 2019). 'Photometric redshifts for X-ray-selected active galactic nuclei in the eROSITA era'. In: MNRAS 489.1, pp. 663–680. DOI: 10.1093/mnras/stz2159 (cit. on p. 16).

Brier, G. W. (1950). 'Verification of Forecasts Expressed in Terms of Probability'. In: *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2 (cit. on p. 61).

Bröcker, J. and Smith, L. A. (2007). 'Increasing the Reliability of Reliability Diagrams'. In: *Weather and Forecasting* 22.3, pp. 651–661. DOI: 10.1175/WAF993.1 (cit. on p. 61).

Brown, M. J. I., Duncan, K. J., et al. (Nov. 2019). 'The spectral energy distributions of active galactic nuclei'. In: MNRAS 489.3, pp. 3351–3367. DOI: 10.1093/mnras/stz2324 (cit. on pp. 44, 120).

Brown, M. J. I., Webster, R. L., and Boyle, B. J. (May 2001). 'The Evolution of Radio Galaxies at Intermediate Redshift'. In: AJ 121.5, pp. 2381–2391. DOI: 10.1086/320410 (cit. on p. 214).

Brown, M. J. I., Moustakas, J., et al. (June 2014). 'An Atlas of Galaxy Spectral Energy Distributions from the Ultraviolet to the Mid-infrared'. In: ApJS 212.2, 18, p. 18. DOI: 10.1088/0067-0049/212/2/18 (cit. on p. 92).

Brunner, H., Liu, T., et al. (May 2022). 'The eROSITA Final Equatorial Depth Survey (eFEDS). X-ray catalogue'. In: A&A 661, A1, A1. DOI: 10.1051/0004-6361/202141266 (cit. on p. 23).

Brusa, M., Zamorani, G., et al. (Sept. 2007). 'The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources'. In: ApJS 172.1, pp. 353–367. DOI: 10.1086/516575 (cit. on p. 22).

Buchner, J. (Oct. 2019). 'Collaborative Nested Sampling: Big Data versus Complex Physical Models'. In: PASP 131.1004, p. 108005. DOI: 10.1088/1538-3873/aae7fc (cit. on p. 24).

Budavári, T. and Szalay, A. S. (May 2008). 'Probabilistic Cross-Identification of Astronomical Sources'. In: ApJ 679.1, pp. 301–309. DOI: 10.1086/587156 (cit. on p. 23).

Buisson, D. J. K., Lohfink, A. M., et al. (Jan. 2017). 'Ultraviolet and X-ray variability of active galactic nuclei with Swift'. In: MNRAS 464.3, pp. 3194–3218. DOI: 10.1093/mnras/stw2486 (cit. on p. 12).

Burhanudin, U. F., Maund, J. R., et al. (Aug. 2021). 'Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data'. In: MNRAS 505.3, pp. 4345–4361. DOI: 10.1093/mnras/stab1545 (cit. on p. 24).

Butler, A., Huynh, M., et al. (May 2019). 'The XXL Survey. XXXVI. Evolution and black hole feedback of high-excitation and low-excitation radio galaxies in XXL-S'. In: A&A 625, A111, A111. DOI: 10.1051/0004-6361/201834581 (cit. on p. 137).

Calistro Rivera, G., Williams, W. L., et al. (Aug. 2017). 'The LOFAR window on star-forming galaxies and AGNs - curved radio SEDs and IR-radio correlation at 0<z<2.5'. In: MNRAS 469.3, pp. 3468–3488. DOI: 10.1093/mnras/stx1040 (cit. on p. 37).

Cameron, E. and Driver, S. P. (May 2007). 'The galaxy luminosity-size relation and selection biases in the Hubble Ultra Deep Field'. In: MNRAS 377.2, pp. 523–534. DOI: 10.1111/j.1365-2966.2007.11507.x (cit. on p. 129).

Camilo, F., Scholz, P., et al. (Apr. 2018). 'Revival of the Magnetar PSR J1622-4950: Observations with MeerKAT, Parkes, XMM-Newton, Swift, Chandra, and NuSTAR'. In: ApJ 856.2, 180, p. 180. DOI: 10.3847/1538-4357/aab35a (cit. on p. 5).

Capetti, A., Brienza, M., et al. (Oct. 2020). 'The LOFAR view of FR 0 radio galaxies'. In: A&A 642, A107, A107. DOI: 10.1051/0004-6361/202038671 (cit. on p. 5).

Cara, M. and Lister, M. L. (Oct. 2008). 'Avoiding Spurious Breaks in Binned Luminosity Functions'. In: ApJ 686.1, pp. 148–154. DOI: 10.1086/590902 (cit. on p. 134).

Card, D. H. (1982). 'Using known map category marginal frequencies to improve estimates of thematic map accuracy'. In: *Photogrammetric Engineering and Remote Sensing* 48.3, pp. 431–439 (cit. on p. 55).

Caruana, R. and Niculescu-Mizil, A. (2006). 'An Empirical Comparison of Supervised Learning Algorithms'. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 161–168. ISBN: 1595933832. DOI: 10.1145/1143844.1143865 (cit. on p. 60).

Carvajal, R., Bauer, F. E., et al. (Jan. 2020). 'The ALMA Frontier Fields Survey. V. ALMA Stacking of Lyman-Break Galaxies in Abell 2744, Abell 370, Abell S1063, MACSJ0416.1-2403 and MACSJ1149.5+2223'. In: A&A 633, A160, A160. DOI: 10.1051/0004-6361/201936260 (cit. on p. 17).

Carvajal, R., Matute, I., et al. (Nov. 2023a). 'Selection of powerful radio galaxies with machine learning'. In: A&A 679, A101, A101. DOI: 10.1051/0004-6361/202245770 (cit. on pp. xi, xxiii, 28, 31, 112).

— (Dec. 2023b). *Selection of powerful radio galaxies with machine learning*. DOI: 10.5281/zenodo.10220009 (cit. on pp. xi, 79, 217).

Carvajal, R., Matute, I., et al. (Oct. 2021). 'Exploring New Redshift Indicators for Radio-Powerful AGN'. In: *Galaxies* 9.4, p. 86. DOI: 10.3390/galaxies9040086 (cit. on pp. xi, 26, 28, 31, 41, 91, 93).

Casalicchio, G., Molnar, C., and Bischl, B. (2019). 'Visualizing the Feature Importance for Black Box Models'. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by M. Berlingerio, F. Bonchi, et al. Cham: Springer International Publishing, pp. 655–670. ISBN: 978-3-030-10925-7. DOI: 10.1007/978-3-030-10925-7_40 (cit. on p. 27).

Ceccarelli, L., Duplancic, F., and Garcia Lambas, D. (Jan. 2022). 'The impact of void environment on AGN'. In: MNRAS 509.2, pp. 1805–1819. DOI: 10.1093/mnras/stab2902 (cit. on p. 7).

Ceraj, L., Smolčić, V., et al. (Dec. 2018). 'The VLA-COSMOS 3 GHz Large Project: Star formation properties and radio luminosity functions of AGN with moderate-to-high radiative luminosities out to z~ 6'. In: A&A 620, A192, A192. DOI: 10.1051/0004-6361/201833935 (cit. on p. 137).

Ceraj, L., Smolčić, V., et al. (Oct. 2020). 'The XXL Survey. XLIII. The quasar radio loudness dichotomy exposed via radio luminosity functions obtained by combining results from COSMOS and XXL-S X-ray selected quasars'. In: A&A 642, A125, A125. DOI: 10.1051/0004-6361/201936776 (cit. on p. 135).

Chambers, K. C., Magnier, E. A., et al. (Dec. 2016). 'The Pan-STARRS1 Surveys'. In: *arXiv e-prints*, arXiv:1612.05560 (cit. on pp. 18, 21, 38, 154).

Champagne, J. B., Casey, C. M., et al. (Aug. 2023). 'A Mixture of LBG Overdensities in the Fields of Three 6 < z < 7 Quasars: Implications for the Robustness of Photometric Selection'. In: ApJ 952.2, 99, p. 99. DOI: 10.3847/1538-4357/acda8d (cit. on p. 17).

Chattopadhyay, A. K. (2017). 'Incomplete Data in Astrostatistics'. In: *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, pp. 1–12. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat07942 (cit. on pp. 40, 41).

Chaves-Montero, J., Bonoli, S., et al. (Dec. 2017). 'ELDAR, a new method to identify AGN in multi-filter surveys: the ALHAMBRA test case'. In: MNRAS 472.2, pp. 2085–2106. DOI: 10.1093/mnras/stx2054 (cit. on p. 7).

Chen, C. T. J., Brandt, W. N., et al. (Mar. 2017). 'Hard X-Ray-selected AGNs in Low-mass Galaxies from the NuSTAR Serendipitous Survey'. In: ApJ 837.1, 48, p. 48. DOI: 10.3847/1538-4357/aa5d5b (cit. on p. 7).

Chen, H., Garrett, M. A., et al. (June 2020). 'Searching for obscured AGN in z ∼ 2 submillimetre galaxies'. In: A&A 638, A113, A113. DOI: 10.1051/0004-6361/201937162 (cit. on p. 4).

Chen, T. and Guestrin, C. (2016). 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785 (cit. on p. 64).

Chisari, N. E., Alonso, D., et al. (May 2019). 'Core Cosmology Library: Precision Cosmological Predictions for LSST'. In: ApJS 242.1, 2, p. 2. DOI: 10.3847/1538-4365/ab1658 (cit. on p. 170).

Ciardi, B. and Ferrara, A. (Feb. 2005). 'The First Cosmic Structures and Their Effects'. In: Space Sci. Rev. 116.3-4, pp. 625–705. DOI: 10.1007/s11214-005-3592-0 (cit. on p. 1).

Cid Fernandes, R., Stasińska, G., et al. (Apr. 2010). 'Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS'. In: MNRAS 403.2, pp. 1036–1053. DOI: 10.1111/j.1365-2966.2009.16185.x (cit. on pp. 8, 9).

Cid Fernandes, R., Stasińska, G., et al. (May 2011). 'A comprehensive classification of galaxies in the Sloan Digital Sky Survey: how to tell true from fake AGN?' In: MNRAS 413.3, pp. 1687–1699. DOI: 10.1111/j.1365-2966.2011.18244.x (cit. on pp. 8, 9).

Clarke, A. O., Scaife, A. M. M., et al. (July 2020). 'Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra'. In: A&A 639, A84, A84. DOI: 10.1051/0004-6361/201936770 (cit. on pp. 87, 89, 90).

Clopper, C. J. and Pearson, E. S. (Dec. 1934). 'The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial'. In: *Biometrika* 26.4, pp. 404–413. ISSN: 0006-3444. DOI: 10.1093/biomet/26.4.404 (cit. on p. 133).

Cochrane, R. K., Kondapally, R., et al. (Aug. 2023). 'The LOFAR Two-metre Sky Survey: the radio view of the cosmic star formation history'. In: MNRAS 523.4, pp. 6082–6102. DOI: 10.1093/mnras/stad1602 (cit. on p. 211).

Cohen, A. C. (1957). 'On the Solution of Estimating Equations for Truncated and Censored Samples from Normal Populations'. In: *Biometrika* 44.1/2, pp. 225–236. ISSN: 00063444. DOI: 10.2307/2333256 (cit. on p. 20).

— (1961). 'Tables for Maximum Likelihood Estimates: Singly Truncated and Singly Censored Samples'. In: *Technometrics* 3.4, pp. 535–541. DOI: 10.1080/00401706.1961.10489973 (cit. on p. 20).

Coleman, B., Kirkpatrick, A., et al. (Sept. 2022). 'Accretion history of AGN: Estimating the host galaxy properties in X-ray luminous AGN from z = 0-3'. In: MNRAS 515.1, pp. 82–98. DOI: 10.1093/mnras/stac1679 (cit. on p. 7).

Condon, J. J. (Sept. 1984). 'Cosmological evolution of radio sources found at 1.4 GHz'. In: ApJ 284, pp. 44–53. DOI: 10.1086/162382 (cit. on p. 214).

— (Mar. 1989). 'The 1.4 GHz Luminosity Function and Its Evolution'. In: ApJ 338, p. 13. DOI: 10.1086/167176 (cit. on p. 214).

— (Jan. 1992). 'Radio emission from normal galaxies.' In: ARA&A 30, pp. 575–611. DOI: 10.1146/annurev.aa.30.090192.003043 (cit. on pp. 4, 13, 90).

Condon, J. J., Cotton, W. D., and Broderick, J. J. (Aug. 2002). 'Radio Sources and Star Formation in the Local Universe'. In: AJ 124.2, pp. 675–689. DOI: 10.1086/341650 (cit. on pp. 37, 211, 214, 215).

Condon, J. J., Cotton, W. D., et al. (May 1998). 'The NRAO VLA Sky Survey'. In: AJ 115.5, pp. 1693–1716. DOI: 10.1086/300337 (cit. on p. 134).

Congalton, R. G., Oderwald, R. G., and Mead, R. A. (1983). 'Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques'. In: *Photogrammetric engineering and remote sensing* 49.12, pp. 1671–1678 (cit. on p. 55).

Conselice, C. J. (Aug. 2014). 'The Evolution of Galaxy Structure Over Cosmic Time'. In: ARA&A 52, pp. 291–337. DOI: 10.1146/annurev-astro-081913-040037 (cit. on p. 35).

Cortes, C. and Vapnik, V. (Sept. 1995). 'Support-vector networks'. In: *Machine Learning* 20.3, pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018 (cit. on p. 23).

Costa-Climent, R., Haftor, D. M., and Staniewski, M. W. (2023). 'Using machine learning to create and capture value in the business models of small and medium-sized enterprises'. In: *International Journal of Information Management* 73, p. 102637. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2023.102637 (cit. on p. 23).

Cover, T. and Hart, P. (1967). 'Nearest neighbor pattern classification'. In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27. DOI: 10.1109/TIT.1967.1053964 (cit. on p. 91).

Cramér, H. (1946). *Mathematical methods of statistics*. English. Princeton University Press Princeton, xvi, 575 p. (Cit. on p. 55).

Cranmer, M. (May 2023). 'Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl'. In: *arXiv e-prints*, arXiv:2305.01582, arXiv:2305.01582. DOI: 10.48550/arXiv.2305.01582 (cit. on p. 27).

Cranmer, M., Sanchez-Gonzalez, A., et al. (2020). 'Discovering symbolic models from deep learning with inductive biases'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546. DOI: 10.5555/3495724.3497186 (cit. on p. 27).

Croom, S. M., Richards, G. T., et al. (Nov. 2009). 'The 2dF-SDSS LRG and QSO survey: the QSO luminosity function at 0.4 < z < 2.6'. In: MNRAS 399.4, pp. 1755–1772. DOI: 10.1111/j.1365-2966.2009.15398.x (cit. on p. 134).

Cunha, P. A. C. and Humphrey, A. (Oct. 2022). 'Photometric redshift-aided classification using ensemble learning'. In: A&A 666, A87, A87. DOI: 10.1051/0004-6361/202243135 (cit. on pp. 26, 89, 90, 93, 154).

Curran, S. J. (May 2022). 'Quasar photometric redshifts from incomplete data using deep learning'. In: MNRAS 512.2, pp. 2099–2109. DOI: 10.1093/mnras/stac660 (cit. on pp. 41, 94).

Curran, S. J., Moss, J. P., and Perrott, Y. C. (July 2022). 'Redshifts of radio sources in the Million Quasars Catalogue from machine learning'. In: MNRAS 514.1, pp. 1–19. DOI: 10.1093/mnras/stac1333 (cit. on pp. 41, 94).

Cutri, R. M., Skrutskie, M. F., et al. (2003a). *2MASS All Sky Catalog of point sources.* (Cit. on p. 18).

— (June 2003b). 'VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)'. In: *VizieR Online Data Catalog*, II/246, pp. II/246 (cit. on p. 18).

Cutri, R. M., Wright, E. L., et al. (Mar. 2012). *Explanatory Supplement to the WISE All-Sky Data Release Products*. Explanatory Supplement to the WISE All-Sky Data Release Products (cit. on p. 10).

Cutri, R. M., Wright, E. L., et al. (Nov. 2013). *Explanatory Supplement to the AllWISE Data Release Products* (cit. on p. 18).

D'Amato, Q., Prandoni, I., et al. (Dec. 2022). 'A deep 1.4 GHz survey of the J1030 equatorial field: A new window on radio source populations across cosmic time'. In: A&A 668, A133, A133. DOI: 10.1051/0004-6361/202244452 (cit. on p. 90).

D'Isanto, A., Cavuoti, S., et al. (Aug. 2018). 'Return of the features. Efficient feature selection and interpretation for photometric redshifts'. In: A&A 616, A97, A97. DOI: 10.1051/0004-6361/201833103 (cit. on pp. 27, 151).

Dahlen, T., Mobasher, B., et al. (Oct. 2013). 'A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation'. In: ApJ 775.2, 93, p. 93. DOI: 10.1088/0004-637X/775/2/93 (cit. on pp. 58, 59).

Dai, Y. S., Wilkes, B. J., et al. (Aug. 2018). 'Is there a relationship between AGN and star formationin IR-bright AGNs?' In: MNRAS 478.3, pp. 4238–4254. DOI: 10.1093/mnras/sty1341 (cit. on p. 4).

Daoutis, C., Kyritsis, E., et al. (Nov. 2023). 'A versatile classification tool for galactic activity using optical and infrared colors'. In: A&A 679, A76, A76. DOI: 10.1051/0004-6361/202347016 (cit. on pp. 113, 114, 152).

Davidson, K. and Netzer, H. (Oct. 1979). 'The emission lines of quasars and similar objects'. In: *Reviews of Modern Physics* 51.4, pp. 715–766. DOI: 10.1103/RevModPhys.51.715 (cit. on p. 2).

Davies, L. J. M., Robotham, A. S. G., et al. (Oct. 2018a). 'Deep Extragalactic VIsible Legacy Survey (DEVILS): motivation,design, and target catalogue'. In: MNRAS 480.1, pp. 768–799. DOI: 10.1093/mnras/sty1553 (cit. on p. 84).

Davies, T. M., Marshall, J. C., and Hazelton, M. L. (2018b). 'Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk'. In: *Statistics in Medicine* 37.7, pp. 1191–1221. DOI: 10.1002/sim.7577 (cit. on p. 213).

Davis, D., Gebhardt, K., et al. (Apr. 2023). 'The HETDEX Survey Emission-line Exploration and Source Classification'. In: ApJ 946.2, 86, p. 86. DOI: 10.3847/1538-4357/acb0ca (cit. on p. 34).

Dayal, P., Volonteri, M., et al. (Jan. 2024). 'UNCOVERing the contribution of black holes to reionization in the JWST era'. In: *arXiv e-prints*, arXiv:2401.11242, arXiv:2401.11242. DOI: 10.48550/arXiv.2401.11242 (cit. on p. 1).

de Ruiter, H. R., Willis, A. G., and Arp, H. C. (May 1977). 'A Westerbork 1415 MHz survey of background radio sources. II. Optical identifications with deep IIIa-J plates.' In: A&AS 28, pp. 211–293 (cit. on p. 22).

de Veny, J. B., Osborn, W. H., and Janes, K. (Oct. 1971). 'A Catalogue of Quasars'. In: PASP 83.495, p. 611. DOI: 10.1086/129187 (cit. on p. 14).

Deka, P. P., Gupta, N., et al. (Feb. 2024). 'The MeerKAT Absorption Line Survey (MALS) Data Release. I. Stokes I Image Catalogs at 1–1.4 GHz'. In: ApJS 270.2, 33, p. 33. DOI: 10.3847/1538-4365/acf7b9 (cit. on p. 13).

Delhaize, J., Heywood, I., et al. (Mar. 2021). 'MIGHTEE: are giant radio galaxies more common than we thought?' In: MNRAS 501.3, pp. 3833–3845. DOI: 10.1093/mnras/staa3837 (cit. on p. 5).

Delhaize, J., Smolčić, V., et al. (June 2017). 'The VLA-COSMOS 3 GHz Large Project: The infrared-radio correlation of star-forming galaxies and AGN to $z \lesssim 6$'. In: A&A 602, A4, A4. DOI: 10.1051/0004-6361/201629430 (cit. on pp. 13, 212).

Desai, S. and Strachan, A. (June 2021). 'Parsimonious neural networks learn interpretable physical laws'. In: *Scientific Reports* 11.1, p. 12761. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92278-w (cit. on p. 24).

Dey, A., Schlegel, D. J., et al. (May 2019). 'Overview of the DESI Legacy Imaging Surveys'. In: AJ 157.5, 168, p. 168. DOI: 10.3847/1538-3881/ab089d (cit. on p. 120).

Dey, B., Andrews, B. H., et al. (Oct. 2022). 'Photometric redshifts from SDSS images with an interpretable deep capsule network'. In: MNRAS 515.4, pp. 5285–5305. DOI: 10.1093/mnras/stac2105 (cit. on p. 28).

Dice, L. R. (1945). 'Measures of the Amount of Ecologic Association Between Species'. In: *Ecology* 26.3, pp. 297–302. ISSN: 00129658, 19399170. DOI: 10.2307/1932409 (cit. on p. 55).

Dieleman, S., Willett, K. W., and Dambre, J. (June 2015). 'Rotation-invariant convolutional neural networks for galaxy morphology prediction'. In: MNRAS 450.2, pp. 1441–1459. DOI: 10.1093/mnras/stv632 (cit. on p. 152).

Dobbels, W. and Baes, M. (Nov. 2021). 'Predicting far-infrared maps of galaxies via machine learning techniques'. In: A&A 655, A34, A34. DOI: 10.1051/0004-6361/202142084 (cit. on p. 24).

Domínguez Sánchez, H., Huertas-Company, M., et al. (Feb. 2018). 'Improving galaxy morphologies for SDSS with Deep Learning'. In: MNRAS 476.3, pp. 3661–3676. DOI: 10.1093/mnras/sty338 (cit. on p. 152).

Donley, J. L., Koekemoer, A. M., et al. (Apr. 2012). 'Identifying Luminous Active Galactic Nuclei in Deep Surveys: Revised IRAC Selection Criteria'. In: ApJ 748.2, 142, p. 142. DOI: 10.1088/0004-637X/748/2/142 (cit. on p. 10).

Donley, J. L., Rieke, G. H., et al. (Nov. 2005). 'Unveiling a Population of AGNs Not Detected in X-Rays'. In: ApJ 634.1, pp. 169–182. DOI: 10.1086/491668 (cit. on p. 7).

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). 'CatBoost: gradient boosting with categorical features support'. In: *CoRR* abs/1810.11363. URL: http://arxiv.org/abs/1810.11363 (cit. on p. 64).

Drake, A. J., Graham, M. J., et al. (July 2014). 'The Catalina Surveys Periodic Variable Star Catalog'. In: ApJS 213.1, 9, p. 9. DOI: 10.1088/0067-0049/213/1/9 (cit. on p. 22).

Driver, S. P., Hill, D. T., et al. (May 2011). 'Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release'. In: MNRAS 413.2, pp. 971–995. DOI: 10.1111/j.1365-2966.2010.18188.x (cit. on p. 84).

Driver, S. P. and Robotham, A. S. G. (Oct. 2010). 'Quantifying cosmic variance'. In: MNRAS 407.4, pp. 2131–2140. DOI: 10.1111/j.1365-2966.2010.17028.x (cit. on p. 33).

Du, M., Yang, F., et al. (2021). 'Fairness in Deep Learning: A Computational Perspective'. In: *IEEE Intelligent Systems* 36.4, pp. 25–34. DOI: 10.1109/MIS.2020.3000681 (cit. on p. 81).

Duan, T., Anand, A., et al. (July 2020). 'NGBoost: Natural Gradient Boosting for Probabilistic Prediction'. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. Daumé III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2690–2700. DOI: 10.5555/3524938.3525190 (cit. on pp. 38, 153).

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press. ISBN: 9781108709385 (cit. on p. 45).

Duncan, K. J., Kondapally, R., et al. (Apr. 2021). 'The LOFAR Two-meter Sky Survey: Deep Fields Data Release 1. IV. Photometric redshifts and stellar masses'. In: A&A 648, A4, A4. DOI: 10.1051/0004-6361/202038809 (cit. on p. 10).

Duncan, K. J., Sabater, J., et al. (Feb. 2019). 'The LOFAR Two-metre Sky Survey. IV. First Data Release: Photometric redshifts and rest-frame magnitudes'. In: A&A 622, A3, A3. DOI: 10.1051/0004-6361/201833562 (cit. on pp. 92, 93, 154).

Dunlop, J. S. and Peacock, J. A. (Nov. 1990). 'The redshift cut-off in the luminosity function of radio galaxies and quasars.' In: MNRAS 247, p. 19 (cit. on p. 214).

Eddington, A. S. (Mar. 1913). 'On a formula for correcting statistics for the effects of a known error of observation'. In: MNRAS 73, pp. 359–360. DOI: 10.1093/mnras/73.5.359 (cit. on p. 131).

— (Mar. 1940). 'The correction of statistics for accidental error'. In: MNRAS 100, p. 354. DOI: 10.1093/mnras/100.5.354 (cit. on p. 131).

Ellis, R. S., Colless, M., et al. (May 1996). 'Autofib Redshift Survey - I. Evolution of the galaxy luminosity function'. In: MNRAS 280.1, pp. 235–251. DOI: 10.1093/mnras/280.1.235 (cit. on p. 212).

Enke, H., Partl, A., et al. (Nov. 2012). 'Handling Big Data in Astronomy and Astrophysics: Rich Structured Queries on Replicated Cloud Data with XtreemFS'. In: *Datenbank-Spektrum* 12.3, pp. 173–181. ISSN: 1610-1995. DOI: 10.1007/s13222-012-0099-1 (cit. on p. 19).

Euclid Collaboration, Bisigello, L., et al. (Apr. 2023a). 'Euclid preparation - XXIII. Derivation of galaxy physical properties with deep machine learning using mock fluxes and H-band images'. In: MNRAS 520.3, pp. 3529–3548. DOI: 10.1093/mnras/stac3810 (cit. on p. 26).

Euclid Collaboration, Humphrey, A., et al. (Mar. 2023b). 'Euclid preparation. XXII. Selection of quiescent galaxies from mock photometry using machine learning'. In: A&A 671, A99, A99. DOI: 10.1051/0004-6361/202244307 (cit. on pp. 26, 39, 45).

Euclid Collaboration, Mellier, Y., et al. (May 2024). 'Euclid. I. Overview of the Euclid mission'. In: *arXiv e-prints*, arXiv:2405.13491, arXiv:2405.13491. DOI: 10.48550/arXiv.2405.13491 (cit. on pp. 18, 154).

Euclid Collaboration, Scaramella, R., et al. (June 2022). 'Euclid preparation. I. The Euclid Wide Survey'. In: A&A 662, A112, A112. DOI: 10.1051/0004-6361/202141938 (cit. on p. 18).

Falcke, H., Nagar, N. M., et al. (Oct. 2000). 'Radio Sources in Low-Luminosity Active Galactic Nuclei. II. Very Long Baseline Interferometry Detections of Compact Radio Cores and Jets in a Sample of LINERs'. In: ApJ 542.1, pp. 197–200. DOI: 10.1086/309543 (cit. on p. 5).

Fan, X., Banados, E., and Simcoe, R. A. (2023). 'Quasars and the Intergalactic Medium at Cosmic Dawn'. In: ARA&A 61. DOI: 10.1146/annurev-astro-052920-102455 (cit. on pp. 2, 6, 14).

Faucher-Giguère, C.-A., Lidz, A., et al. (Oct. 2009). 'A New Calculation of the Ionizing Background Spectrum and the Effects of He II Reionization'. In: ApJ 703.2, pp. 1416–1443. DOI: 10.1088/0004-637X/703/2/1416 (cit. on p. 1).

Feigelson, E. D. and Nelson, P. I. (June 1985). 'Statistical methods for astronomical data with upper limits. I. Univariate distributions.' In: ApJ 293, pp. 192–206. DOI: 10.1086/163225 (cit. on p. 20).

Felten, J. E. (Aug. 1976). 'On Schmidt's Vm estimator and other estimators of luminosity functions.' In: ApJ 207, pp. 700–709. DOI: 10.1086/154538 (cit. on p. 212).

Ferrarese, L. and Merritt, D. (Aug. 2000). 'A Fundamental Relation between Supermassive Black Holes and Their Host Galaxies'. In: ApJ 539.1, pp. L9–L12. DOI: 10.1086/312838 (cit. on p. 6).

Flesch, E. W. (Mar. 2015). 'The Half Million Quasars (HMQ) Catalogue'. In: PASA 32, e010, e010. DOI: 10.1017/pasa.2015.10 (cit. on pp. 15, 45).

— (Dec. 2019). 'The Million Quasars (Milliquas) Catalogue, v6.4'. In: *arXiv e-prints*, arXiv:1912.05614, arXiv:1912.05614. DOI: 10.48550/arXiv.1912.05614 (cit. on p. 15).

— (May 2021). 'The Million Quasars (Milliquas) v7.2 Catalogue, now with VLASS associations. The inclusion of SDSS-DR16Q quasars is detailed'. In: *arXiv e-prints*, arXiv:2105.12985, arXiv:2105.12985. DOI: 10.48550/arXiv.2105.12985 (cit. on pp. 15, 44).

— (Dec. 2023). 'The Million Quasars (Milliquas) Catalogue, v8'. In: *The Open Journal of Astrophysics* 6, 49, p. 49. DOI: 10.21105/astro.2308.01505 (cit. on pp. 15, 119).

Flewelling, H. A., Magnier, E. A., et al. (Nov. 2020). 'The Pan-STARRS1 Database and Data Products'. In: ApJS 251.1, 7, p. 7. DOI: 10.3847/1538-4365/abb82d (cit. on p. 38).

Främling, K. (2023). 'Feature Importance versus Feature Influence and What It Signifies for Explainable AI'. In: *Explainable Artificial Intelligence*. Ed. by L. Longo. Cham: Springer Nature Switzerland, pp. 241–259. ISBN: 978-3-031-44064-9. DOI: 10.1007/978-3-031-44064-9_14 (cit. on p. 27).

Frederiksen, T. F., Graur, O., et al. (Mar. 2014). 'Spectroscopic identification of a redshift 1.55 supernova host galaxy from the Subaru Deep Field Supernova Survey'. In: A&A 563, A140, A140. DOI: 10.1051/0004-6361/201321795 (cit. on p. 15).

Freund, Y. and Schapire, R. E. (1996). 'Experiments with a New Boosting Algorithm'. In: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. ICML'96. Bari, Italy: Morgan Kaufmann Publishers Inc., pp. 148–156. ISBN: 1558604197. DOI: 10.5555/3091696.3091715 (cit. on p. 25).

Friedman, J. H. (2001). 'Greedy function approximation: A gradient boosting machine.' In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451 (cit. on p. 64).

— (2002). 'Stochastic gradient boosting'. In: *Computational Statistics & Data Analysis* 38.4. Nonlinear Methods and Data Mining, pp. 367–378. ISSN: 0167-9473. DOI: 10.1016/S0167-9473(01)00065-2 (cit. on p. 64).

Fu, H., Myers, A. D., et al. (Jan. 2015). 'Radio-selected Binary Active Galactic Nuclei from the Very Large Array Stripe 82 Survey'. In: ApJ 799.1, 72, p. 72. DOI: 10.1088/0004-637X/799/1/72 (cit. on p. 36).

Fu, Y., Wu, X.-B., et al. (Apr. 2024). 'CatNorth: An Improved Gaia DR3 Quasar Candidate Catalog with Pan-STARRS1 and CatWISE'. In: ApJS 271.2, 54, p. 54. DOI: 10.3847/1538-4365/ad2ae6 (cit. on pp. 12, 87, 90).

Fukugita, M. and Kawasaki, M. (Aug. 1994). 'Reionization during Hierarchical Clustering in a Universe Dominated by Cold Dark Matter'. In: MNRAS 269, p. 563. DOI: 10.1093/mnras/269.3.563 (cit. on p. 1).

Gaia Collaboration, Bailer-Jones, C. A. L., et al. (June 2023a). 'Gaia Data Release 3. The extragalactic content'. In: A&A 674, A41, A41. DOI: 10.1051/0004-6361/202243232 (cit. on pp. 12, 87, 119).

Gaia Collaboration, Prusti, T., et al. (Nov. 2016). 'The Gaia mission'. In: A&A 595, A1, A1. DOI: 10.1051/0004-6361/201629272 (cit. on pp. 12, 18).

Gaia Collaboration, Vallenari, A., et al. (June 2023b). 'Gaia Data Release 3. Summary of the content and survey properties'. In: A&A 674, A1, A1. DOI: 10.1051/0004-6361/202243940 (cit. on pp. 12, 87).

Galametz, A., Grazian, A., et al. (June 2013). 'CANDELS Multiwavelength Catalogs: Source Identification and Photometry in the CANDELS UKIDSS Ultra-deep Survey Field'. In: ApJS 206.2, 10, p. 10. DOI: 10.1088/0067-0049/206/2/10 (cit. on p. 15).

Galton, F. (1886). 'Regression Towards Mediocrity in Hereditary Stature.' In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15, pp. 246–263. ISSN: 09595295. URL: http://www.jstor.org/stable/2841583 (visited on 29/02/2024) (cit. on p. 62).

Gammerman, A., Vovk, V., and Vapnik, V. (July 1998). 'Learning by transduction'. In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., pp. 148–155. ISBN: 155860555X. DOI: 10.5555/2074094.2074112 (cit. on pp. 38, 153).

Gao, L. and Guan, L. (Oct. 2023). 'Interpretability of Machine Learning: Recent Advances and Future Prospects'. In: *IEEE MultiMedia* 30.04, pp. 105–118. ISSN: 1941-0166. DOI: 10.1109/MMUL.2023.3272513 (cit. on p. 26).

Garcia-Dias, R., Allende Prieto, C., et al. (May 2018). 'Machine learning in APOGEE. Unsupervised spectral classification with K-means'. In: A&A 612, A98, A98. DOI: 10.1051/0004-6361/201732134 (cit. on p. 25).

Garcia-Piquer, A., Morales, J. C., et al. (Aug. 2017). 'Efficient scheduling of astronomical observations. Application to the CARMENES radial-velocity survey'. In: A&A 604, A87, A87. DOI: 10.1051/0004-6361/201628577 (cit. on p. 24).

Garilli, B., Fumana, M., et al. (July 2010). 'EZ: A Tool For Automatic Redshift Measurement'. In: PASP 122.893, p. 827. DOI: 10.1086/654903 (cit. on p. 16).

Garofalo, M., Botta, A., and Ventre, G. (June 2017). 'Astrophysics and Big Data: Challenges, Methods, and Tools'. In: *Astroinformatics*. Ed. by M. Brescia, S. G. Djorgovski, et al. Vol. 325, pp. 345–348. DOI: 10.1017/S1743921316012813 (cit. on pp. 19, 24).

Gatica, C., Demarco, R., et al. (Jan. 2024). 'The AGN fraction in high-redshift protocluster candidates selected by Planck and Herschel'. In: MNRAS 527.2, pp. 3006–3017. DOI: 10.1093/mnras/stad3404 (cit. on p. 113).

Gebhardt, K., Bender, R., et al. (Aug. 2000). 'A Relationship between Nuclear Black Hole Mass and Galaxy Velocity Dispersion'. In: ApJ 539.1, pp. L13–L16. DOI: 10.1086/312840 (cit. on p. 6).

Gebhardt, K., Mentuch Cooper, E., et al. (Dec. 2021). 'The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Survey Design, Reductions, and Detections'. In: ApJ 923.2, 217, p. 217. DOI: 10.3847/1538-4357/ac2e03 (cit. on p. 34).

Gehrels, N. (Apr. 1986). 'Confidence Limits for Small Numbers of Events in Astrophysical Data'. In: ApJ 303, p. 336. DOI: 10.1086/164079 (cit. on p. 133).

Gerwin, D. (1974). 'Information processing, data inferences, and scientific generalization'. In: *Behavioral Science* 19.5, pp. 314–325. DOI: 10.1002/bs.3830190504 (cit. on p. 27).

Getachew-Woreta, T., Pović, M., et al. (July 2022). 'Effect of AGN on the morphological properties of their host galaxies in the local Universe'. In: MNRAS 514.1, pp. 607–620. DOI: 10.1093/mnras/stac851 (cit. on p. 12).

Geurts, P., Ernst, D., and Wehenkel, L. (Apr. 2006). 'Extremely randomized trees'. In: *Machine Learning* 63.1, pp. 3–42. ISSN: 1573-0565. DOI: 10.1007/s10994-006-6226-1 (cit. on p. 64).

Gilda, S. (Feb. 2024). 'Beyond mirkwood: Enhancing SED Modeling with Conformal Predictions'. In: *Astronomy* 3.1, pp. 14–20. DOI: 10.3390/astronomy3010002 (cit. on pp. 153, 154).

Gilda, S., Lower, S., and Narayanan, D. (July 2021). 'MIRKWOOD: Fast and Accurate SED Modeling Using Machine Learning'. In: ApJ 916.1, 43, p. 43. DOI: 10.3847/1538-4357/ac0058 (cit. on pp. 19, 61, 153, 154).

Giles, D. and Walkowicz, L. (Mar. 2019). 'Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection'. In: MNRAS 484.1, pp. 834–849. DOI: 10.1093/mnras/sty3461 (cit. on pp. 24, 25).

Giveon, U., Maoz, D., et al. (July 1999). 'Long-term optical variability properties of the Palomar-Green quasars'. In: MNRAS 306.3, pp. 637–654. DOI: 10.1046/j.1365-8711.1999.02556.x (cit. on p. 12).

Glahn, H. R. and Jorgensen, D. L. (1970). 'Climatological Aspects of the Brier p-score'. In: *Monthly Weather Review* 98.2, pp. 136–141. DOI: 10.1175/1520-0493(1970)098<0136:CAOTBP>2.3.CO;2 (cit. on p. 62).

Glasbey, C., van der Heijden, G., et al. (June 2007). 'Colour displays for categorical images'. In: *Color Research & Application* 32.4, pp. 304–309. DOI: 10.1002/col.20327 (cit. on p. 170).

Glazebrook, K., Offer, A. R., and Deeley, K. (Jan. 1998). 'Automatic Redshift Determination by Use of Principal Component Analysis. I. Fundamentals'. In: ApJ 492.1, pp. 98–109. DOI: 10.1086/305039 (cit. on p. 16).

Glikman, E., Lacy, M., et al. (July 2018). 'Luminous WISE-selected Obscured, Unobscured, and Red Quasars in Stripe 82'. In: ApJ 861.1, 37, p. 37. DOI: 10.3847/1538-4357/aac5d8 (cit. on p. 36).

Glikman, E., Langgin, R., et al. (July 2023). 'A Candidate Dual QSO at Cosmic Noon'. In: ApJ 951.1, L18, p. L18. DOI: 10.3847/2041-8213/acda2f (cit. on p. 12).

Goebel, R., Chander, A., et al. (2018). 'Explainable ai: the new 42?' In: *International cross-domain conference for machine learning and knowledge extraction*. Springer. Springer International Publishing, pp. 295–303. ISBN: 978-3-319-99740-7. DOI: 10.1007/978-3-319-99740-7_21 (cit. on p. 27).

Gordon, Y. A., Boyce, M. M., et al. (Oct. 2020). 'A Catalog of Very Large Array Sky Survey Epoch 1 Quick Look Components, Sources, and Host Identifications'. In: *Research Notes of the American Astronomical Society* 4.10, 175, p. 175. DOI: 10.3847/2515-5172/abbe23 (cit. on p. 5).

Goulding, A. D., Zakamska, N. L., et al. (Mar. 2018). 'High-redshift Extremely Red Quasars in X-Rays'. In: ApJ 856.1, 4, p. 4. DOI: 10.3847/1538-4357/aab040 (cit. on p. 7).

Gross, A. C., Fu, H., et al. (Mar. 2023). 'Testing the Radio-selection Method of Dual Active Galactic Nuclei in the Stripe 82 Field'. In: ApJ 945.1, 73, p. 73. DOI: 10.3847/1538-4357/acb646 (cit. on p. 36).

Guglielmetti, F., Arras, P., et al. (Dec. 2022). 'Bayesian and Machine Learning Methods in the Big Data Era for Astronomical Imaging'. In: *Physical Sciences Forum*. Vol. 5. Physical Sciences Forum, 50, p. 50. DOI: 10.3390/psf2022005050 (cit. on p. 24).

Gültekin, K., Richstone, D. O., et al. (June 2009). 'The M-σ and M-L Relations in Galactic Bulges, and Determinations of Their Intrinsic Scatter'. In: ApJ 698.1, pp. 198–221. DOI: 10.1088/0004-637X/698/1/198 (cit. on p. 6).

Gupta, N., Pannella, M., et al. (May 2020). 'Constraining radio mode feedback in galaxy clusters with the cluster radio AGNs properties to z ∼ 1'. In: MNRAS 494.2, pp. 1705–1723. DOI: 10.1093/mnras/staa832 (cit. on p. 215).

Gupta, N., Saro, A., et al. (May 2017). 'High-frequency cluster radio galaxies: luminosity functions and implications for SZE-selected cluster samples'. In: MNRAS 467.3, pp. 3737–3750. DOI: 10.1093/mnras/stx095 (cit. on p. 215).

Gürkan, G., Hardcastle, M. J., et al. (Feb. 2019). 'LoTSS/HETDEX: Optical quasars. I. Low-frequency radio properties of optically selected quasars'. In: A&A 622, A11, A11. DOI: 10.1051/0004-6361/201833892 (cit. on p. 13).

Guyon, I. and Elisseeff, A. (Mar. 2003). 'An introduction to variable and feature selection'. In: *J. Mach. Learn. Res.* 3.null, pp. 1157–1182. ISSN: 1532-4435. DOI: 10.5555/944919.944968 (cit. on p. 62).

Haardt, F. and Maraschi, L. (Oct. 1991). 'A Two-Phase Model for the X-Ray Emission from Seyfert Galaxies'. In: ApJ 380, p. L51. DOI: 10.1086/186171 (cit. on p. 2).

Haardt, F. and Madau, P. (Feb. 2012). 'Radiative Transfer in a Clumpy Universe. IV. New Synthesis Models of the Cosmic UV/X-Ray Background'. In: ApJ 746.2, 125, p. 125. DOI: 10.1088/0004-637X/746/2/125 (cit. on p. 1).

Haardt, F. and Maraschi, L. (Aug. 1993). 'X-Ray Spectra from Two-Phase Accretion Disks'. In: ApJ 413, p. 507. DOI: 10.1086/173020 (cit. on p. 2).

Habouzit, M., Onoue, M., et al. (Apr. 2022). 'Co-evolution of massive black holes and their host galaxies at high redshift: discrepancies from six cosmological simulations and the key role of JWST'. In: MNRAS 511.3, pp. 3751–3767. DOI: 10.1093/mnras/stac225 (cit. on p. 6).

Haggard, D., Green, P. J., et al. (Nov. 2010). 'The Field X-ray AGN Fraction to z = 0.7 from the Chandra Multiwavelength Project and the Sloan Digital Sky Survey'. In: ApJ 723.2, pp. 1447–1468. DOI: 10.1088/0004-637X/723/2/1447 (cit. on p. 87).

Haiman, Z. and Loeb, A. (July 1997). 'Signatures of Stellar Reionization of the Universe'. In: ApJ 483.1, pp. 21–37. DOI: 10.1086/304238 (cit. on p. 1).

Hales, C. A., Murphy, T., et al. (Aug. 2012a). *BLOBCAT: Software to Catalog Blobs*. Astrophysics Source Code Library, record ascl:1208.009 (cit. on p. 14).

— (Sept. 2012b). 'BLOBCAT: software to catalogue flood-filled blobs in radio images of total intensity and linear polarization'. In: MNRAS 425.2, pp. 979–996. DOI: 10.1111/j.1365-2966.2012.21373.x (cit. on p. 14).

Hancock, P. J., Murphy, T., et al. (May 2012). 'Compact continuum source finding for next generation radio surveys'. In: MNRAS 422.2, pp. 1812–1824. DOI: 10.1111/j.1365-2966.2012.20768.x (cit. on p. 14).

Hancock, P. J., Trott, C. M., and Hurley-Walker, N. (Mar. 2018). 'Source Finding in the Era of the SKA (Precursors): Aegean 2.0'. In: PASA 35, e011, e011. DOI: 10.1017/pasa.2018.3 (cit. on p. 14).

Hardcastle, M. J., Horton, M. A., et al. (Oct. 2023). 'The LOFAR Two-Metre Sky Survey. VI. Optical identifications for the second data release'. In: A&A 678, A151, A151. DOI: 10.1051/0004-6361/202347333 (cit. on pp. 14, 22).

Häring, N. and Rix, H.-W. (Apr. 2004). 'On the Black Hole Mass-Bulge Mass Relation'. In: ApJ 604.2, pp. L89–L92. DOI: 10.1086/383567 (cit. on p. 6).

Harris, C. R., Millman, K. J., et al. (Sept. 2020). 'Array programming with NumPy'. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2 (cit. on pp. 20, 170).

Head, T., Kumar, M., et al. (Oct. 2021). *scikit-optimize/scikit-optimize*. Version v0.9.0. DOI: 10.5281/zenodo.5565057 (cit. on p. 67).

Heckman, T. M. (July 1980). 'An Optical and Radio Survey of the Nuclei of Bright Galaxies - Activity in the Normal Galactic Nuclei'. In: A&A 87, p. 152 (cit. on p. 8).

Heckman, T. M. and Best, P. N. (Aug. 2014). 'The Coevolution of Galaxies and Supermassive Black Holes: Insights from Surveys of the Contemporary Universe'. In: ARA&A 52, pp. 589–660. DOI: 10.1146/annurev-astro-081913-035722 (cit. on pp. 4, 7).

Helfand, D. J., White, R. L., and Becker, R. H. (Mar. 2015). 'The Last of FIRST: The Final Catalog and Source Identifications'. In: ApJ 801.1, 26, p. 26. DOI: 10.1088/0004-637X/801/1/26 (cit. on p. 5).

Helou, G., Soifer, B. T., and Rowan-Robinson, M. (Nov. 1985). 'Thermal infrared and nonthermal radio : remarkable correlation in disks of galaxies.' In: ApJ 298, pp. L7–L11. DOI: 10.1086/184556 (cit. on p. 90).

Hernán-Caballero, A., Varela, J., et al. (Oct. 2021). 'The miniJPAS survey: Photometric redshift catalogue'. In: A&A 654, A101, A101. DOI: 10.1051/0004-6361/202141236 (cit. on p. 16).

Hickox, R. C. and Alexander, D. M. (Sept. 2018). 'Obscured Active Galactic Nuclei'. In: ARA&A 56, pp. 625–671. DOI: 10.1146/annurev-astro-081817-051803 (cit. on pp. 1, 2, 7).

Hildebrand, R. H. (Sept. 1983). 'The determination of cloud masses and dust characteristics from submillimetre thermal emission.' In: QJRAS 24, pp. 267–282 (cit. on p. 4).

Hildebrandt, H., Arnouts, S., et al. (Nov. 2010). 'PHAT: PHoto-z Accuracy Testing'. In: A&A 523, A31, A31. DOI: 10.1051/0004-6361/201014885 (cit. on p. 59).

Hill, G. J., Gebhardt, K., et al. (Oct. 2008). 'The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX): Description and Early Pilot Survey Results'. In: *Panoramic Views of Galaxy Formation and Evolution*.

Ed. by T. Kodama, T. Yamada, and K. Aoki. Vol. 399. Astronomical Society of the Pacific Conference Series, p. 115 (cit. on p. 34).

Hoaglin, D., Mosteller, F., et al. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley Series in Probability and Statistics: Probability and Statistics Section Series. John Wiley & Sons. ISBN: 9780471097778 (cit. on p. 59).

Hodge, J. A., Becker, R. H., et al. (July 2011). 'High-resolution Very Large Array Imaging of Sloan Digital Sky Survey Stripe 82 at 1.4 GHz'. In: AJ 142.1, 3, p. 3. DOI: 10.1088/0004-6256/142/1/3 (cit. on p. 36).

Hoffer, R. M. and Fleming, M. D. (1978). 'Mapping vegetative cover by computer-aided analysis of satellite data'. In: *USDA Forest Service Gen. Tech. Rep. RM-55*, pp. 227–237 (cit. on p. 55).

Hogg, D. W. (May 1999). 'Distance measures in cosmology'. In: *arXiv e-prints*, astro-ph/9905116, astro–ph/9905116. DOI: 10.48550/arXiv.astro-ph/9905116 (cit. on p. 211).

Hoyle, B., Gruen, D., et al. (July 2018). 'Dark Energy Survey Year 1 Results: redshift distributions of the weak-lensing source galaxies'. In: MNRAS 478.1, pp. 592–610. DOI: 10.1093/mnras/sty957 (cit. on p. 16).

Huertas-Company, M. and Lanusse, F. (Jan. 2023). 'The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys'. In: PASA 40, e001, e001. DOI: 10.1017/pasa.2022.55 (cit. on p. 24).

Hunter, J. D. (2007). 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55 (cit. on p. 170).

Huterer, D., Takada, M., et al. (Feb. 2006). 'Systematic errors in future weak-lensing surveys: requirements and prospects for self-calibration'. In: MNRAS 366.1, pp. 101–114. DOI: 10.1111/j.1365-2966.2005.09782.x (cit. on p. 15).

Ibar, E., Ivison, R. J., et al. (Jan. 2010). 'Deep multi-frequency radio imaging in the Lockman Hole - II. The spectral index of submillimetre galaxies'. In: MNRAS 401.1, pp. L53–L57. DOI: 10.1111/j.1745-3933.2009.00786.x (cit. on p. 13).

Ibata, R. A., McConnachie, A., et al. (Oct. 2017). 'The Canada-France Imaging Survey: First Results from the u-Band Component'. In: ApJ 848.2, 128, p. 128. DOI: 10.3847/1538-4357/aa855c (cit. on p. 113).

İkiz, T., Peletier, R. F., et al. (Aug. 2020). 'Infrared-detected AGNs in the local Universe'. In: A&A 640, A68, A68. DOI: 10.1051/0004-6361/201935971 (cit. on p. 10).

Ilbert, O., Arnouts, S., et al. (Oct. 2006). 'Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey'. In: A&A 457.3, pp. 841–856. DOI: 10.1051/0004-6361:20065138 (cit. on p. 17).

Ilbert, O., Capak, P., et al. (Jan. 2009). 'Cosmos Photometric Redshifts with 30-Bands for 2-deg^2'. In: ApJ 690.2, pp. 1236–1249. DOI: 10.1088/0004-637X/690/2/1236 (cit. on p. 59).

Inayoshi, K., Visbal, E., and Haiman, Z. (Aug. 2020). 'The Assembly of the First Massive Black Holes'. In: ARA&A 58, pp. 27–97. DOI: 10.1146/annurev-astro-120419-014455 (cit. on pp. 6, 14).

IPCC (2022). *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by P. Shukla, J. Skea, et al. Cambridge, UK and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157926 (cit. on p. 19).

Isobe, T., Feigelson, E. D., and Nelson, P. I. (July 1986). 'Statistical Methods for Astronomical Data with Upper Limits. II. Correlation and Regression'. In: ApJ 306, p. 490. DOI: 10.1086/164359 (cit. on p. 20).

Ivezić, Ž., Kahn, S. M., et al. (Mar. 2019). 'LSST: From Science Drivers to Reference Design and Anticipated Data Products'. In: ApJ 873.2, 111, p. 111. DOI: 10.3847/1538-4357/ab042c (cit. on pp. 18, 19).

James, G., Witten, D., et al. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer International Publishing. ISBN: 9783031387470 (cit. on p. 24).

Jarrett, T. H., Cluver, M. E., et al. (Feb. 2017). 'Galaxy and Mass Assembly (GAMA): Exploring the WISE Web in G12'. In: ApJ 836.2, 182, p. 182. DOI: 10.3847/1538-4357/836/2/182 (cit. on p. 11).

Jarvis, M., Taylor, R., et al. (Jan. 2016). 'The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey'. In: *MeerKAT Science: On the Pathway to the SKA*, 6, p. 6. DOI: 10.22323/1.277.0006 (cit. on p. 5).

Jarvis, M. J. and Rawlings, S. (Nov. 2000). 'On the redshift cut-off for flat-spectrum radio sources'. In: MNRAS 319.1, pp. 121–136. DOI: 10.1046/j.1365-8711.2000.03801.x (cit. on p. 35).

Jia, P., Jia, Q., et al. (June 2023). 'Observation Strategy Optimization for Distributed Telescope Arrays with Deep Reinforcement Learning'. In: AJ 165.6, 233, p. 233. DOI: 10.3847/1538-3881/accceb (cit. on p. 24).

Jiang, L., Fan, X., et al. (July 2014). 'The Sloan Digital Sky Survey Stripe 82 Imaging Data: Depth-optimized Co-adds over 300 deg^2 in Five Filters'. In: ApJS 213.1, 12, p. 12. DOI: 10.1088/0067-0049/213/1/12 (cit. on p. 34).

Jiang, T., Gradus, J. L., et al. (Feb. 2021). 'Addressing Measurement Error in Random Forests Using Quantitative Bias Analysis'. In: *American Journal of Epidemiology* 190.9, pp. 1830–1840. ISSN: 0002-9262. DOI: 10.1093/aje/kwab010 (cit. on p. 38).

Johnson, J., Douze, M., and Jégou, H. (2019). 'Billion-scale similarity search with GPUs'. In: *IEEE Transactions on Big Data* 7.3, pp. 535–547 (cit. on p. 170).

Johnson, N. and Leone, F. (1964). *Statistics and Experimental Design in Engineering and the Physical Sciences*. Vol. 2. Wiley, p. 125. ISBN: 9780471444893 (cit. on p. 63).

Jonas, J. and MeerKAT Team (Jan. 2016). 'The MeerKAT Radio Telescope'. In: *MeerKAT Science: On the Pathway to the SKA*, 1, p. 1. DOI: 10.22323/1.277.0001 (cit. on p. 5).

Josse, J., Prost, N., et al. (Feb. 2019). 'On the consistency of supervised learning with missing values'. In: *arXiv e-prints*, arXiv:1902.06931, arXiv:1902.06931. DOI: 10.48550/arXiv.1902.06931 (cit. on p. 40).

Josse, J. and Reiter, J. P. (2018). 'Introduction to the Special Section on Missing Data'. In: *Statistical Science* 33.2, pp. 139–141. DOI: 10.1214/18-STS332IN (cit. on p. 20).

Kafka, P. (Jan. 1967). 'How to count Quasars'. In: Nature 213.5074, pp. 346–350. DOI: 10.1038/213346a0 (cit. on p. 212).

Kalton, G. and Kasprzyk, D. (1982). 'Imputing for missing survey responses'. In: *Proceedings of the section on survey research methods, American Statistical Association*. Vol. 22. American Statistical Association Cincinnati, p. 31 (cit. on p. 40).

Kamiran, F. and Calders, T. (Oct. 2012). 'Data preprocessing techniques for classification without discrimination'. In: *Knowledge and Information Systems* 33.1, pp. 1–33. ISSN: 0219-3116. DOI: 10.1007/s10115-011-0463-8 (cit. on pp. 48, 81).

Karniadakis, G. E., Kevrekidis, I. G., et al. (June 2021). 'Physics-informed machine learning'. In: *Nature Reviews Physics* 3.6, pp. 422–440. ISSN: 2522-5820. DOI: 10.1038/s42254-021-00314-5 (cit. on p. 24).

Kashyap, V. L., van Dyk, D. A., et al. (Aug. 2010). 'On Computing Upper Limits to Source Intensities'. In: ApJ 719.1, pp. 900–914. DOI: 10.1088/0004-637X/719/1/900 (cit. on p. 20).

Katz, H., Kimm, T., et al. (Aug. 2018). 'A Census of the LyC photons that form the UV background during reionization'. In: MNRAS 478.4, pp. 4986–5005. DOI: 10.1093/mnras/sty1225 (cit. on p. 1).

— (Feb. 2019). 'Tracing the sources of reionization in cosmological radiation hydrodynamics simulations'. In: MNRAS 483.1, pp. 1029–1041. DOI: 10.1093/mnras/sty3154 (cit. on p. 1).

Kauffmann, G., Heckman, T. M., et al. (Dec. 2003). 'The host galaxies of active galactic nuclei'. In: MNRAS 346.4, pp. 1055–1077. DOI: 10.1111/j.1365-2966.2003.07154.x (cit. on pp. 7–9).

Ke, G., Meng, Q., et al. (2017). 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (cit. on p. 93).

Kewley, L. J., Dopita, M. A., et al. (July 2001). 'Theoretical Modeling of Starburst Galaxies'. In: ApJ 556.1, pp. 121–140. DOI: 10.1086/321545 (cit. on pp. 7, 8).

Kewley, L. J., Groves, B., et al. (Nov. 2006). 'The host galaxies and classification of active galactic nuclei'. In: MNRAS 372.3, pp. 961–976. DOI: 10.1111/j.1365-2966.2006.10859.x (cit. on pp. 8, 9).

Khramtsov, V., Spiniello, C., et al. (July 2021). 'VEXAS: VISTA EXtension to Auxiliary Surveys. Data Release 2: Machine-learning based classification of sources in the Southern Hemisphere'. In: A&A 651, A69, A69. DOI: 10.1051/0004-6361/202040131 (cit. on p. 119).

Kim, J.-Y., Krichbaum, T. P., et al. (Aug. 2020). 'Event Horizon Telescope imaging of the archetypal blazar 3C 279 at an extreme 20 microarcsecond resolution'. In: A&A 640, A69, A69. DOI: 10.1051/0004-6361/202037493 (cit. on p. 5).

Kim, S. J., Lee, H. M., et al. (Dec. 2012). 'The North Ecliptic Pole Wide survey of AKARI: a near- and mid-infrared source catalog'. In: A&A 548, A29, A29. DOI: 10.1051/0004-6361/201219105 (cit. on p. 89).

King, A. and Pounds, K. (Aug. 2015). 'Powerful Outflows and Feedback from Active Galactic Nuclei'. In: ARA&A 53, pp. 115–154. DOI: 10.1146/annurev-astro-082214-122316 (cit. on p. 1).

Kirkpatrick, A., Alberts, S., et al. (Nov. 2017). 'The AGN-Star Formation Connection: Future Prospects with JWST'. In: ApJ 849.2, 111, p. 111. DOI: 10.3847/1538-4357/aa911d (cit. on p. 11).

Kirkpatrick, A., Pope, A., et al. (Nov. 2012). 'GOODS-Herschel: Impact of Active Galactic Nuclei and Star Formation Activity on Infrared Spectral Energy Distributions at High Redshift'. In: ApJ 759.2, 139, p. 139. DOI: 10.1088/0004-637X/759/2/139 (cit. on p. 10).

Kluyver, T., Ragan-Kelley, B., et al. (2016). 'Jupyter Notebooks – a publishing format for reproducible computational workflows'. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87–90 (cit. on p. 169).

Kohavi, Ron and John, G. H. (1997). 'Wrappers for feature subset selection'. In: *Artificial Intelligence* 97.1. Relevance, pp. 273–324. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(97)00043-X (cit. on p. 62).

Kollmeier, J. A., Zasowski, G., et al. (Nov. 2017). 'SDSS-V: Pioneering Panoptic Spectroscopy'. In: *arXiv e-prints*, arXiv:1711.03234, arXiv:1711.03234. DOI: 10.48550/arXiv.1711.03234 (cit. on p. 18).

Kondapally, R., Best, P. N., et al. (July 2022). 'Cosmic evolution of low-excitation radio galaxies in the LOFAR two-metre sky survey deep fields'. In: MNRAS 513.3, pp. 3742–3767. DOI: 10.1093/mnras/stac1128 (cit. on pp. 131, 137).

Kormendy, J. and Ho, L. C. (Aug. 2013). 'Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies'. In: ARA&A 51.1, pp. 511–653. DOI: 10.1146/annurev-astro-082708-101811 (cit. on p. 7).

Koshida, S., Minezaki, T., et al. (June 2014). 'Reverberation Measurements of the Inner Radius of the Dust Torus in 17 Seyfert Galaxies'. In: ApJ 788.2, 159, p. 159. DOI: 10.1088/0004-637X/788/2/159 (cit. on p. 12).

Koshida, S., Yoshii, Y., et al. (Aug. 2009). 'Variation of Inner Radius of Dust Torus in NGC4151'. In: ApJ 700.2, pp. L109–L113. DOI: 10.1088/0004-637X/700/2/L109 (cit. on p. 12).

Kovács, O. E., Bogdán, Á., et al. (Feb. 2019). 'Detection of the Missing Baryons toward the Sightline of H1821+643'. In: ApJ 872.1, 83, p. 83. DOI: 10.3847/1538-4357/aaef78 (cit. on p. 2).

Kovesi, P. (Sept. 2015). 'Good Colour Maps: How to Design Them'. In: *arXiv e-prints*, arXiv:1509.03700, arXiv:1509.03700. DOI: 10.48550/arXiv.1509.03700 (cit. on p. 170).

Kowalski, A. F., Hawley, S. L., et al. (Aug. 2009). 'M Dwarfs in Sloan Digital Sky Survey Stripe 82: Photometric Light Curves and Flare Rate Analysis'. In: AJ 138.2, pp. 633–648. DOI: 10.1088/0004-6256/138/2/633 (cit. on p. 36).

Krumpe, M., Miyaji, T., and Coil, A. L. (Nov. 2014). 'Clustering Measurements of broad-line AGNs: Review and Future'. In: *Multifrequency Behaviour of High Energy Cosmic Sources*, pp. 71–78. DOI: 10.14311/APP.2014.01.0071 (cit. on p. 2).

Kulkarni, G., Worseck, G., and Hennawi, J. F. (Sept. 2019). 'Evolution of the AGN UV luminosity function from redshift 7.5'. In: MNRAS 488.1, pp. 1035–1065. DOI: 10.1093/mnras/stz1493 (cit. on p. 1).

Kull, M., Filho, T. M. S., and Flach, P. (2017a). 'Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration'. In: *Electronic Journal of Statistics* 11.2, pp. 5052–5080. DOI: 10.1214/17-EJS1338SI (cit. on pp. 61, 170).

Kull, M., Filho, T. S., and Flach, P. (Apr. 2017b). 'Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 623–631. URL: https://proceedings.mlr.press/v54/kull17a.html (cit. on pp. 61, 170).

Kunsági-Máté, S., Beck, R., et al. (Oct. 2022). 'Photometric redshifts for quasars from WISE-PS1-STRM'. In: MNRAS 516.2, pp. 2662–2670. DOI: 10.1093/mnras/stac2411 (cit. on p. 99).

Kurtz, M. J. and Mink, D. J. (Aug. 1998). 'RVSAO 2.0: Digital Redshifts and Radial Velocities'. In: PASP 110.750, pp. 934–977. DOI: 10.1086/316207 (cit. on p. 16).

Kuźmicz, A. and Jamrozy, M. (Mar. 2021). 'Giant Radio Quasars: Sample and Basic Properties'. In: ApJS 253.1, 25, p. 25. DOI: 10.3847/1538-4365/abd483 (cit. on p. 5).

La Torre, V., Sajina, A., et al. (June 2024). 'Estimating Galaxy Parameters with Self-organizing Maps and the Effect of Missing Data'. In: AJ 167.6, 261, p. 261. DOI: 10.3847/1538-3881/ad3821 (cit. on p. 16).

Lacerda, E. A. D., Sánchez, S. F., et al. (Nov. 2022). 'pyFIT3D and pyPipe3D - The new version of the integral field spectroscopy data analysis pipeline'. In: New A 97, 101895, p. 101895. DOI: 10.1016/j.newast.2022.101895 (cit. on pp. 8, 9).

Lacy, M., Baum, S. A., et al. (Mar. 2020). 'The Karl G. Jansky Very Large Array Sky Survey (VLASS). Science Case and Survey Design'. In: PASP 132.1009, 035001, p. 035001. DOI: 10.1088/1538-3873/ab63eb (cit. on pp. 5, 18).

Lacy, M., Petric, A. O., et al. (Jan. 2007). 'Optical Spectroscopy and X-Ray Detections of a Sample of Quasars and Active Galactic Nuclei Selected in the Mid-Infrared from Two Spitzer Space Telescope Wide-Area Surveys'. In: AJ 133.1, pp. 186–205. DOI: 10.1086/509617 (cit. on p. 10).

Lacy, M., Ridgway, S. E., et al. (Oct. 2013). 'The Spitzer Mid-infrared Active Galactic Nucleus Survey. I. Optical and Near-infrared Spectroscopy of Obscured Candidates and Normal Active Galactic Nuclei Selected in the Mid-infrared'. In: ApJS 208.2, 24, p. 24. DOI: 10.1088/0067-0049/208/2/24 (cit. on p. 10).

Lacy, M., Storrie-Lombardi, L. J., et al. (Sept. 2004). 'Obscured and Unobscured Active Galactic Nuclei in the Spitzer Space Telescope First Look Survey'. In: ApJS 154.1, pp. 166–169. DOI: 10.1086/422816 (cit. on p. 10).

Lacy, M., Surace, J. A., et al. (Feb. 2021). 'A Spitzer survey of Deep Drilling Fields to be targeted by the Vera C. Rubin Observatory Legacy Survey of Space and Time'. In: MNRAS 501.1, pp. 892–910. DOI: 10.1093/mnras/staa3714 (cit. on p. 10).

Lacy, M. and Sajina, A. (Apr. 2020). 'Active galactic nuclei as seen by the Spitzer Space Telescope'. In: *Nature Astronomy* 4, pp. 352–363. DOI: 10.1038/s41550-020-1071-x (cit. on pp. 7, 37).

Lal, D. V. (July 2021). 'The Discovery of a Remnant Radio Galaxy in A2065 Using GMRT'. In: ApJ 915.2, 126, p. 126. DOI: 10.3847/1538-4357/ac042d (cit. on p. 5).

LaMassa, S. M., Peca, A., et al. (Mar. 2024). 'Stripe 82X Data Release 3: Multiwavelength Catalog with New Spectroscopic Redshifts and Black Hole Masses'. In: *arXiv e-prints*, arXiv:2403.20160, arXiv:2403.20160. DOI: 10.48550/arXiv.2403.20160 (cit. on p. 36).

LaMassa, S. M., Urry, C. M., et al. (Feb. 2016). 'The 31 Deg^2 Release of the Stripe 82 X-Ray Survey: The Point Source Catalog'. In: ApJ 817.2, 172, p. 172. DOI: 10.3847/0004-637X/817/2/172 (cit. on p. 22).

Lang, D. (May 2014). 'unWISE: Unblurred Coadds of the WISE Imaging'. In: AJ 147.5, 108, p. 108. DOI: 10.1088/0004-6256/147/5/108 (cit. on pp. 35, 37, 118, 119).

Langeroodi, D. and Hjorth, J. (Apr. 2023). 'PAH Emission from Star-forming Galaxies in JWST Mid-infrared Imaging of the Lensing Cluster SMACS J0723.3-7327'. In: ApJ 946.2, L40, p. L40. DOI: 10.3847/2041-8213/acc1e0 (cit. on p. 11).

Langley, P. (1977). 'BACON: A Production System That Discovers Empirical Laws'. In: *International Joint Conference on Artificial Intelligence*. URL: https://api.semanticscholar.org/CorpusID:2320342 (cit. on p. 27).

— (1979). 'Rediscovering Physics with BACON.3'. In: *Proceedings of the 6th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'79. Tokyo, Japan: Morgan Kaufmann Publishers Inc., pp. 505–507. ISBN: 0934613478. DOI: 10.5555/1624861.1624976 (cit. on p. 27).

Langley, P., Bradshaw, G. L., and Simon, H. A. (1981). 'BAC0N.5: The Discovery of Conservation Laws'. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'81. Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., pp. 121–126. DOI: 10.5555/1623156.1623181 (cit. on p. 27).

Langley, P. and Zytkow, J. M. (1990). 'Data-Driven Approaches to Empirical Discovery'. In: *Machine Learning: Paradigms and Methods*. USA: Elsevier North-Holland, Inc., pp. 283–312. ISBN: 0262530880. DOI: 10.5555/87755.87762 (cit. on p. 27).

Lansbury, G. B., Banerji, M., et al. (Jan. 2020). 'X-ray observations of luminous dusty quasars at z > 2'. In: MNRAS 495.3, pp. 2652–2663. DOI: 10.1093/mnras/staa1220 (cit. on p. 7).

Latimer, C. J., Reines, A. E., et al. (June 2021). 'A Chandra and HST View of WISE-selected AGN Candidates in Dwarf Galaxies'. In: ApJ 914.2, 133, p. 133. DOI: 10.3847/1538-4357/abfe0c (cit. on p. 7).

Lazio, J. W., Kimball, A., et al. (Feb. 2014). 'Radio Astronomy in LSST Era'. In: PASP 126.936, p. 196. DOI: 10.1086/675262 (cit. on p. 145).

Le Fèvre, O., Tasca, L. A. M., et al. (Apr. 2015). 'The VIMOS Ultra-Deep Survey: ∼10 000 galaxies with spectroscopic redshifts to study galaxy assembly at early epochs 2 < z ≃ 6'. In: A&A 576, A79, A79. DOI: 10.1051/0004-6361/201423829 (cit. on p. 15).

LeCun, Y., Bengio, Y., and Hinton, G. (May 2015). 'Deep learning'. In: Nature 521.7553, pp. 436–444. DOI: 10.1038/nature14539 (cit. on p. 63).

LeCun, Y., Bottou, L., et al. (1998). 'Gradient-based learning applied to document recognition'. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791 (cit. on p. 63).

Lee, H. M., Kim, S. J., et al. (Feb. 2009). 'North Ecliptic Pole Wide Field Survey of AKARI: Survey Strategy and Data Characteristics'. In: PASJ 61, p. 375. DOI: 10.1093/pasj/61.2.375 (cit. on p. 89).

Lehmer, B. D., Brandt, W. N., et al. (Nov. 2005). 'The Extended Chandra Deep Field-South Survey: Chandra Point-Source Catalogs'. In: ApJS 161.1, pp. 21–40. DOI: 10.1086/444590 (cit. on p. 91).

Leja, J., Johnson, B. D., et al. (June 2019). 'An Older, More Quiescent Universe from Panchromatic SED Fitting of the 3D-HST Survey'. In: ApJ 877.2, 140, p. 140. DOI: 10.3847/1538-4357/ab1d5a (cit. on p. 19).

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). 'Calibration of probabilities: The state of the art to 1980'. In: *Judgment under Uncertainty: Heuristics and Biases*. Ed. by D. Kahneman, P. Slovic, and A. Tversky. Cambridge University Press, pp. 306–334. DOI: 10.1017/CBO9780511809477.023 (cit. on p. 61).

Lima, E. V. R., Sodré, L., et al. (Jan. 2022). 'Photometric redshifts for the S-PLUS Survey: Is machine learning up to the task?' In: *Astronomy and Computing* 38, 100510, p. 100510. DOI: 10.1016/j.ascom.2021.100510 (cit. on p. 59).

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). 'Explainable AI: A Review of Machine Learning Interpretability Methods'. In: *Entropy* 23.1. ISSN: 1099-4300. DOI: 10.3390/e23010018 (cit. on p. 26).

Lira, P., Arévalo, P., et al. (Aug. 2011). 'Optical and near-IR long-term monitoring of NGC 3783 and MR 2251-178: evidence for variable near-IR emission from thin accretion discs'. In: MNRAS 415.2, pp. 1290–1303. DOI: 10.1111/j.1365-2966.2011.18774.x (cit. on p. 12).

Lira, P., Arévalo, P., et al. (Nov. 2015). 'Long-term monitoring of the archetype Seyfert galaxy MCG-6-30-15: X-ray, optical and near-IR variability of the corona, disc and torus'. In: MNRAS 454.1, pp. 368–379. DOI: 10.1093/mnras/stv1945 (cit. on p. 12).

Lisenfeld, U. and Völk, H. J. (Feb. 2000). 'On the radio spectral index of galaxies'. In: A&A 354, pp. 423–430 (cit. on p. 12).

Liske, J., Baldry, I. K., et al. (Sept. 2015). 'Galaxy And Mass Assembly (GAMA): end of survey report and data release 2'. In: MNRAS 452.2, pp. 2087–2126. DOI: 10.1093/mnras/stv1436 (cit. on p. 84).

Little, R. and Rubin, D. (2014). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9781118625880 (cit. on p. 20).

Liu, D., Lang, P., et al. (Oct. 2019). 'Automated Mining of the ALMA Archive in the COSMOS Field (A^3COSMOS). I. Robust ALMA Continuum Photometry Catalogs and Stellar Mass and Star Formation Properties for ~700 Galaxies at z = 0.5-6'. In: ApJS 244.2, 40, p. 40. DOI: 10.3847/1538-4365/ab42da (cit. on p. 23).

Lochner, M. and Bassett, B. A. (July 2021). 'ASTRONOMALY: Personalised active anomaly detection in astronomical data'. In: *Astronomy and Computing* 36, 100481, p. 100481. DOI: 10.1016/j.ascom.2021.100481 (cit. on p. 24).

Loeb, A. and Furlanetto, S. R. (2013). *The First Galaxies in the Universe*. Princeton Series in Astrophysics. Princeton University Press. ISBN: 9780691144924 (cit. on p. 1).

Loh, E. D. and Spillar, E. J. (Apr. 1986). 'Photometric Redshifts of Galaxies'. In: ApJ 303, p. 154. DOI: 10.1086/164062 (cit. on p. 16).

Louppe, G., Wehenkel, L., et al. (2013). 'Understanding variable importances in forests of randomized trees'. In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, et al. Vol. 26. Curran Associates, Inc., pp. 431–439. DOI: 10.5555/2999611.2999660 (cit. on p. 27).

LSST Science Collaboration, Abell, P. A., et al. (Dec. 2009). 'LSST Science Book, Version 2.0'. In: *arXiv e-prints*, arXiv:0912.0201, arXiv:0912.0201. DOI: 10.48550/arXiv.0912.0201 (cit. on pp. 7, 18).

Luken, K. J., Norris, R. P., et al. (Apr. 2022). 'Estimating galaxy redshift in radio-selected datasets using machine learning'. In: *Astronomy and Computing* 39, 100557, p. 100557. DOI: 10.1016/j.ascom.2022.100557 (cit. on pp. 92, 93).

Luken, K. J., Norris, R. P., and Park, L. A. F. (Oct. 2019). 'Preliminary Results of Using k-Nearest Neighbors Regression to Estimate the Redshift of Radio-selected Data Sets'. In: PASP 131.1004, p. 108003. DOI: 10.1088/1538-3873/aaea17 (cit. on p. 91).

Lukic, V., Brüggen, M., et al. (Aug. 2019). 'Morphological classification of radio galaxies: capsule networks versus convolutional neural networks'. In: MNRAS 487.2, pp. 1729–1744. DOI: 10.1093/mnras/stz1289 (cit. on pp. 24, 25).

Lundberg, S. M. and Lee, S.-I. (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (cit. on p. 99).

Lundberg, S. M., Erion, G., et al. (2020). 'From local explanations to global understanding with explainable AI for trees'. In: *Nature Machine Intelligence* 2.1, pp. 2522–5839. DOI: 10.1038/s42256-019-0138-9 (cit. on p. 99).

Lyu, J., Alberts, S., et al. (Dec. 2022). 'AGN Selection and Demographics in GOODS-S/HUDF from X-Ray to Radio'. In: ApJ 941.2, 191, p. 191. DOI: 10.3847/1538-4357/ac9e5d (cit. on p. 122).

Lyu, J. and Rieke, G. (May 2022). 'Infrared Spectral Energy Distribution and Variability of Active Galactic Nuclei: Clues to the Structure of Circumnuclear Material'. In: *Universe* 8.6, 304, p. 304. DOI: 10.3390/universe8060304 (cit. on p. 2).

Ma, S. and Tourani, R. (Aug. 2020). 'Predictive and Causal Implications of using Shapley Value for Model Interpretation'. In: *Proceedings of the 2020 KDD Workshop on Causal Discovery*. Vol. 127. Proceedings of Machine Learning Research. PMLR, pp. 23–38. URL: https://proceedings.mlr.press/v127/ma20a.html (cit. on p. 28).

Ma, Z., Xu, H., et al. (Feb. 2019). 'A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample'. In: ApJS 240.2, 34, p. 34. DOI: 10.3847/1538-4365/aaf9a2 (cit. on pp. 24, 25).

Maccacaro, T., della Ceca, R., et al. (June 1991). 'The Properties of X-Ray-selected Active Galactic Nuclei. I. Luminosity Function, Cosmological Evolution, and Contribution to the Diffuse X-Ray Background'. In: ApJ 374, p. 117. DOI: 10.1086/170102 (cit. on p. 212).

Machado, D. P., Leonard, A., et al. (Dec. 2013). 'Darth Fader: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise'. In: A&A 560, A83, A83. DOI: 10.1051/0004-6361/201219857 (cit. on p. 16).

Machado Poletti Valle, L. F., Avestruz, C., et al. (Oct. 2021). 'SHAPing the gas: understanding gas shapes in dark matter haloes with interpretable machine learning'. In: MNRAS 507.1, pp. 1468–1484. DOI: 10.1093/mnras/stab2252 (cit. on p. 28).

Machalski, J. and Godlowski, W. (Aug. 2000). '1.4 GHz luminosity function of galaxies in the Las Campanas redshift survey and its evolution'. In: A&A 360, pp. 463–471 (cit. on p. 215).

Madau, P. and Dickinson, M. (Aug. 2014). 'Cosmic Star-Formation History'. In: ARA&A 52, pp. 415–486. DOI: 10.1146/annurev-astro-081811-125615 (cit. on pp. 7, 35).

Madau, P. and Haardt, F. (Nov. 2015). 'Cosmic Reionization after Planck: Could Quasars Do It All?' In: ApJ 813.1, L8, p. L8. DOI: 10.1088/2041-8205/813/1/L8 (cit. on p. 1).

Madau, P., Haardt, F., and Rees, M. J. (Apr. 1999). 'Radiative Transfer in a Clumpy Universe. III. The Nature of Cosmological Ionizing Sources'. In: ApJ 514.2, pp. 648–659. DOI: 10.1086/306975 (cit. on p. 1).

Magliocchetti, M., Lutz, D., et al. (July 2014). 'The PEP survey: infrared properties of radio-selected AGN'. In: MNRAS 442.1, pp. 682–693. DOI: 10.1093/mnras/stu863 (cit. on pp. 122, 124, 125, 140).

Magliocchetti, M. (Dec. 2022). 'Hosts and environments: a (large-scale) radio history of AGN and star-forming galaxies'. In: A&A Rev. 30.1, 6, p. 6. DOI: 10.1007/s00159-022-00142-1 (cit. on pp. 2, 4, 7, 13, 90, 125, 140).

Magliocchetti, M., Maddox, S. J., et al. (June 2002). 'The 2dF Galaxy Redshift Survey: the population of nearby radio galaxies at the 1-mJy level'. In: MNRAS 333.1, pp. 100–120. DOI: 10.1046/j.1365-8711.2002.05386.x (cit. on p. 124).

Magorrian, J., Tremaine, S., et al. (June 1998). 'The Demography of Massive Dark Objects in Galaxy Centers'. In: AJ 115.6, pp. 2285–2305. DOI: 10.1086/300353 (cit. on p. 6).

Mainzer, A., Bauer, J., et al. (Apr. 2011). 'Preliminary Results from NEOWISE: An Enhancement to the Wide-field Infrared Survey Explorer for Solar System Science'. In: ApJ 731.1, 53, p. 53. DOI: 10.1088/0004-637X/731/1/53 (cit. on p. 37).

Mainzer, A., Bauer, J., et al. (Sept. 2014). 'Initial Performance of the NEOWISE Reactivation Mission'. In: ApJ 792.1, 30, p. 30. DOI: 10.1088/0004-637X/792/1/30 (cit. on p. 37).

Maitra, C., Haberl, F., et al. (Feb. 2019). 'Identification of AGN in the XMM-Newton X-ray survey of the SMC'. In: A&A 622, A29, A29. DOI: 10.1051/0004-6361/201833663 (cit. on p. 7).

Mandal, S., Prandoni, I., et al. (Apr. 2021). 'Extremely deep 150 MHz source counts from the LoTSS Deep Fields'. In: A&A 648, A5, A5. DOI: 10.1051/0004-6361/202039998 (cit. on pp. 122, 131).

Marchesi, S., Civano, F., et al. (Jan. 2016). 'The Chandra COSMOS Legacy survey: optical/IR identifications'. In: ApJ 817.1, 34, p. 34. DOI: 10.3847/0004-637X/817/1/34 (cit. on p. 22).

Marocco, F., Eisenhardt, P. R. M., et al. (Mar. 2021). 'The CatWISE2020 Catalog'. In: ApJS 253.1, 8, p. 8. DOI: 10.3847/1538-4365/abd805 (cit. on p. 37).

Martocchia, S., Piconcelli, E., et al. (Dec. 2017). 'The WISSH quasars project. III. X-ray properties of hyper-luminous quasars'. In: A&A 608, A51, A51. DOI: 10.1051/0004-6361/201731314 (cit. on p. 7).

Mateos, S., Alonso-Herrero, A., et al. (Nov. 2012). 'Using the Bright Ultrahard XMM-Newton survey to define an IR selection of luminous AGN based on WISE colours'. In: MNRAS 426.4, pp. 3271–3281. DOI: 10.1111/j.1365-2966.2012.21843.x (cit. on pp. xxv, 11, 84).

Mathews, E. P., Leja, J., et al. (Sept. 2023). 'As Simple as Possible but No Simpler: Optimizing the Performance of Neural Net Emulators for Galaxy SED Fitting'. In: ApJ 954.2, 132, p. 132. DOI: 10.3847/1538-4357/ace720 (cit. on pp. 19, 24).

Matsuoka, Y., Strauss, M. A., et al. (Dec. 2018). 'Subaru High-z Exploration of Low-luminosity Quasars (SHELLQs). V. Quasar Luminosity Function and Contribution to Cosmic Reionization at z = 6'. In: ApJ 869.2, 150, p. 150. DOI: 10.3847/1538-4357/aaee7a (cit. on p. 1).

Matthews, B. W. (1975). 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme'. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2, pp. 442–451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9 (cit. on p. 55).

Matthews, T. A. and Sandage, A. R. (July 1963). 'Optical Identification of 3C 48, 3C 196, and 3C 286 with Stellar Objects.' In: ApJ 138, p. 30. DOI: 10.1086/147615 (cit. on p. 6).

Mauch, T., Cotton, W. D., et al. (Jan. 2020). 'The 1.28 GHz MeerKAT DEEP2 Image'. In: ApJ 888.2, 61, p. 61. DOI: 10.3847/1538-4357/ab5d2d (cit. on p. 5).

Mauch, T. and Sadler, E. M. (Mar. 2007). 'Radio sources in the 6dFGS: local luminosity functions at 1.4 GHz for star-forming galaxies and radio-loud AGN'. In: MNRAS 375.3, pp. 931–950. DOI: 10.1111/j.1365-2966.2006.11353.x (cit. on pp. 14, 122, 124, 132, 133, 214, 215).

Mazzolari, G., Gilli, R., et al. (Jan. 2024). 'Heavily Obscured AGN detection: a Radio vs X-ray challenge'. In: *arXiv e-prints*, arXiv:2402.00109, arXiv:2402.00109. DOI: 10.48550/arXiv.2402.00109 (cit. on p. 4).

McAlpine, K., Jarvis, M. J., and Bonfield, D. G. (Dec. 2013). 'Evolution of faint radio sources in the VIDEO-XMM3 field'. In: MNRAS 436.2, pp. 1084–1095. DOI: 10.1093/mnras/stt1638 (cit. on pp. 122, 124).

McConnell, D., Hale, C. L., et al. (Nov. 2020). 'The Rapid ASKAP Continuum Survey I: Design and first results'. In: PASA 37, e048, e048. DOI: 10.1017/pasa.2020.41 (cit. on p. 165).

McConnell, N. J. and Ma, C.-P. (Feb. 2013). 'Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties'. In: ApJ 764.2, 184, p. 184. DOI: 10.1088/0004-637X/764/2/184 (cit. on p. 6).

McGreer, I. D., Becker, R. H., et al. (Nov. 2006). 'Discovery of a z = 6.1 Radio-Loud Quasar in the NOAO Deep Wide Field Survey'. In: ApJ 652.1, pp. 157–162. DOI: 10.1086/507767 (cit. on p. 5).

McHardy, I. M., Connolly, S. D., et al. (May 2016). 'The origin of UV-optical variability in AGN and test of disc models: XMM-Newton and ground-based observations of NGC 4395'. In: *Astronomische Nachrichten* 337.4-5, p. 500. DOI: 10.1002/asna.201612337 (cit. on p. 12).

McKinney, W. (2010). 'Data Structures for Statistical Computing in Python'. In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a (cit. on p. 170).

Mead, R. A. and Meyer, M. P. (1977). 'Landsat digital data application to forest vegetation and land use classification in Minnesota'. In: *LARS Symposia*, p. 220 (cit. on p. 55).

Meiksin, A. (Jan. 2005). 'Constraints on the ionization sources of the high-redshift intergalactic medium'. In: MNRAS 356.2, pp. 596–606. DOI: 10.1111/j.1365-2966.2004.08481.x (cit. on p. 1).

Meisner, A. M., Lang, D., et al. (Dec. 2019). 'unWISE Coadds: The Five-year Data Set'. In: PASP 131.1006, p. 124504. DOI: 10.1088/1538-3873/ab3df4 (cit. on p. 119).

Meisner, A. M., Lang, D., et al. (Sept. 2022). '9-yr Deep Sky unWISE Coadds'. In: *Research Notes of the American Astronomical Society* 6.9, 188, p. 188. DOI: 10.3847/2515-5172/ac913e (cit. on pp. 35, 37, 118).

Menzel, M. .-., Merloni, A., et al. (Mar. 2016). 'A spectroscopic survey of X-ray-selected AGNs in the northern XMM-XXL field'. In: MNRAS 457.1, pp. 110–132. DOI: 10.1093/mnras/stv2749 (cit. on p. 11).

Merlin, E., Castellano, M., et al. (May 2021). 'The ASTRODEEP-GS43 catalogue: New photometry and redshifts for the CANDELS GOODS-South field'. In: A&A 649, A22, A22. DOI: 10.1051/0004-6361/202140310 (cit. on p. 17).

Messias, H., Afonso, J., et al. (Aug. 2012). 'A New Infrared Color Criterion for the Selection of 0 < z < 7 AGNs: Application to Deep Fields and Implications for JWST Surveys'. In: ApJ 754.2, 120, p. 120. DOI: 10.1088/0004-637X/754/2/120 (cit. on p. 11).

Michailidis, M. (2017). 'Investigating machine learning methods in recommender systems'. PhD thesis. University College London, UK. URL: https://discovery.ucl.ac.uk/id/eprint/10031000 (cit. on p. 65).

Michelucci, U. and Venturini, F. (2023). 'New metric formulas that include measurement errors in machine learning for natural sciences'. In: *Expert Systems with Applications* 224, p. 120013. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120013 (cit. on p. 38).

Mickaelian, A. M. (Dec. 2020). 'Big Data in Astronomy: Surveys, Catalogs, Databases and Archives'. In: *Communications of the Byurakan Astrophysical Observatory* 67, pp. 159–180. DOI: 10.52526/25792776-2020.67.2-159 (cit. on p. 19).

Miley, G. and De Breuck, C. (Feb. 2008). 'Distant radio galaxies and their environments'. In: A&A Rev. 15.2, pp. 67–144. DOI: 10.1007/s00159-007-0008-z (cit. on p. 90).

Miller, S. T., Lindner, J. F., et al. (2020). 'The scaling of physics-informed machine learning with data and dimensions'. In: *Chaos, Solitons & Fractals: X* 5, p. 100046. ISSN: 2590-0544. DOI: 10.1016/j.csfx.2020.100046 (cit. on p. 24).

Mingo, B., Watson, M. G., et al. (Nov. 2016). 'The MIXR sample: AGN activity versus star formation across the cross-correlation of WISE, 3XMM, and FIRST/NVSS'. In: MNRAS 462.3, pp. 2631–2667. DOI: 10.1093/mnras/stw1826 (cit. on pp. xxv, 11, 84).

Mitra, S., Choudhury, T. R., and Ferrara, A. (Jan. 2018). 'Cosmic reionization after Planck II: contribution from quasars'. In: MNRAS 473.1, pp. 1416–1425. DOI: 10.1093/mnras/stx2443 (cit. on p. 1).

Miyaji, T., Hasinger, G., and Schmidt, M. (Apr. 2001). 'Soft X-ray AGN luminosity function from ROSAT surveys. II. Table of the binned soft X-ray luminosity function'. In: A&A 369, pp. 49–56. DOI: 10.1051/0004-6361:20010102 (cit. on p. 134).

Mo, W., Gonzalez, A., et al. (Oct. 2020). 'The Massive and Distant Clusters of WISE Survey. VIII. Radio Activity in Massive Galaxy Clusters'. In: ApJ 901.2, 131, p. 131. DOI: 10.3847/1538-4357/abb08d (cit. on p. 13).

Mohale, K. and Lochner, M. (May 2024). 'Enabling unsupervised discovery in astronomical images through self-supervised representations'. In: MNRAS 530.1, pp. 1274–1295. DOI: 10.1093/mnras/stae926 (cit. on p. 25).

Mohan, N. and Rafferty, D. (Feb. 2015). *PyBDSF: Python Blob Detection and Source Finder*. Astrophysics Source Code Library, record ascl:1502.007 (cit. on p. 14).

Morabito, L. K., Sweijen, F., et al. (Oct. 2022). 'Identifying active galactic nuclei via brightness temperature with sub-arcsecond international LOFAR telescope observations'. In: MNRAS 515.4, pp. 5758–5774. DOI: 10.1093/mnras/stac2129 (cit. on p. 10).

Morrissey, P., Conrow, T., et al. (Dec. 2007). 'The Calibration and Data Products of GALEX'. In: ApJS 173.2, pp. 682–697. DOI: 10.1086/520512 (cit. on p. 18).

Mostert, R. I. J., Duncan, K. J., et al. (Jan. 2021). 'Unveiling the rarest morphologies of the LOFAR Two-metre Sky Survey radio source population with self-organised maps'. In: A&A 645, A89, A89. DOI: 10.1051/0004-6361/202038500 (cit. on p. 24).

Mostert, R. I. J., Oei, M. S. S. L., et al. (Apr. 2024). 'Constraining the giant radio galaxy population with machine learning and Bayesian inference'. In: *arXiv e-prints*, arXiv:2405.00232, arXiv:2405.00232. DOI: 10.1051/0004-6361/202348897 (cit. on p. 145).

Moya, A. and López-Sastre, R. J. (July 2022). 'Stellar mass and radius estimation using artificial intelligence'. In: A&A 663, A112, A112. DOI: 10.1051/0004-6361/202142930 (cit. on p. 26).

Murphy, A. H. and Winkler, R. L. (1977). 'Reliability of Subjective Probability Forecasts of Precipitation and Temperature'. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 26.1, pp. 41–47. DOI: https://doi.org/10.2307/2346866 (cit. on p. 61).

Naidoo, K., Johnston, H., et al. (Feb. 2023). 'Euclid: Calibrating photometric redshifts with spectroscopic cross-correlations'. In: A&A 670, A149, A149. DOI: 10.1051/0004-6361/202244795 (cit. on p. 15).

Nakoneczny, S. J., Bilicki, M., et al. (May 2021). 'Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4'. In: A&A 649, A81, A81. DOI: 10.1051/0004-6361/202039684 (cit. on pp. 24, 25).

Netzer, H. (Aug. 2015). 'Revisiting the Unified Model of Active Galactic Nuclei'. In: ARA&A 53, pp. 365–408. DOI: 10.1146/annurev-astro-082214-122302 (cit. on pp. 2, 7).

Newman, J. A., Abate, A., et al. (Mar. 2015). 'Spectroscopic needs for imaging dark energy experiments'. In: *Astroparticle Physics* 63, pp. 81–100. DOI: 10.1016/j.astropartphys.2014.06.007 (cit. on p. 16).

Newman, J. A. and Gruen, D. (Aug. 2022). 'Photometric Redshifts for Next-Generation Surveys'. In: ARA&A 60, pp. 363–414. DOI: 10.1146/annurev-astro-032122-014611 (cit. on p. 16).

Nicastro, F., Kaastra, J., et al. (June 2018). 'Observations of the missing baryons in the warm-hot intergalactic medium'. In: Nature 558.7710, pp. 406–409. DOI: 10.1038/s41586-018-0204-1 (cit. on p. 2).

Nicastro, F., Krongold, Y., et al. (Mar. 2017). 'A decade of warm hot intergalactic medium searches: Where do we stand and where do we go?' In: *Astronomische Nachrichten* 338.281, pp. 281–286. DOI: 10.1002/asna.201713343 (cit. on p. 2).

Niculescu-Mizil, A. and Caruana, R. (2005a). 'Obtaining calibrated probabilities from boosting'. In: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. UAI'05. Edinburgh, Scotland: AUAI Press, pp. 413–420. ISBN: 0974903914. DOI: 10.5555/3020336.3020388 (cit. on pp. 61, 67).

— (2005b). 'Predicting Good Probabilities with Supervised Learning'. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. Bonn, Germany: Association for Computing Machinery, pp. 625–632. ISBN: 1595931805. DOI: 10.1145/1102351.1102430 (cit. on p. 61).

Norris, R., Basu, K., et al. (Apr. 2015). 'The SKA Mid-frequency All-sky Continuum Survey: Discovering the unexpected and transforming radio-astronomy'. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, 86, p. 86. DOI: 10.22323/1.215.0086 (cit. on p. 156).

Norris, R. P. (June 2017). 'Astroinformatics Challenges from Next-generation Radio Continuum Surveys'. In: *Astroinformatics*. Ed. by M. Brescia, S. G. Djorgovski, et al. Vol. 325, pp. 103–113. DOI: 10.1017/S1743921316012825 (cit. on p. 19).

Norris, R. P., Hopkins, A. M., et al. (Aug. 2011). 'EMU: Evolutionary Map of the Universe'. In: PASA 28.3, pp. 215–248. DOI: 10.1071/AS11021 (cit. on pp. 92, 155).

Norris, R. P., Marvil, J., et al. (Sept. 2021). 'The Evolutionary Map of the Universe pilot survey'. In: PASA 38, e046, e046. DOI: 10.1017/pasa.2021.42 (cit. on pp. 5, 22, 118, 119, 147, 149).

Norris, R. P., Salvato, M., et al. (Oct. 2019). 'A Comparison of Photometric Redshift Techniques for Large Radio Surveys'. In: PASP 131.1004, p. 108004. DOI: 10.1088/1538-3873/ab0f7b (cit. on pp. 16, 92, 93, 154).

Nour, D. and Sriram, K. (Jan. 2023). 'Association of optical, ultraviolet, and soft X-ray excess emissions in AGNs'. In: MNRAS 518.4, pp. 5705–5717. DOI: 10.1093/mnras/stac3505 (cit. on p. 2).

Obrić, M., Ivezić, Ž., et al. (Aug. 2006). 'Panchromatic properties of 99000 galaxies detected by SDSS, and (some by) ROSAT, GALEX, 2MASS, IRAS, GB6, FIRST, NVSS and WENSS surveys'. In: MNRAS 370.4, pp. 1677–1698. DOI: 10.1111/j.1365-2966.2006.10675.x (cit. on p. 113).

Ochsenbein, F., Bauer, P., and Marcout, J. (Apr. 2000). 'The VizieR database of astronomical catalogues'. In: A&AS 143, pp. 23–32. DOI: 10.1051/aas:2000169 (cit. on p. 169).

Oke, J. B. and Sandage, A. (Oct. 1968). 'Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies'. In: ApJ 154, p. 21. DOI: 10.1086/149737 (cit. on p. 211).

Oliver, S., Rowan-Robinson, M., et al. (Aug. 2000). 'The European Large Area ISO Survey - I. Goals, definition and observations'. In: MNRAS 316.4, pp. 749–767. DOI: 10.1046/j.1365-8711.2000.03550.x (cit. on p. 91).

Opitz, D. and Maclin, R. (July 1999). 'Popular Ensemble Methods: An Empirical Study'. In: *J. Artif. Int. Res.* 11.1, pp. 169–198. ISSN: 1076-9757. DOI: 10.5555/3013545.3013549 (cit. on p. 25).

Osorio-Clavijo, N., Gonzalez-Martín, O., et al. (July 2023). 'AGNs in the CALIFA survey: X-ray detection of nuclear sources'. In: MNRAS 522.4, pp. 5788–5804. DOI: 10.1093/mnras/stad1262 (cit. on p. 7).

Osterbrock, D. E. (Oct. 1981). 'Seyfert galaxies with weak broad H alpha emission lines'. In: ApJ 249, pp. 462–470. DOI: 10.1086/159306 (cit. on p. 8).

Pacifici, C., Iyer, K. G., et al. (Feb. 2023). 'The Art of Measuring Physical Parameters in Galaxies: A Critical Assessment of Spectral Energy Distribution Fitting Techniques'. In: ApJ 944.2, 141, p. 141. DOI: 10.3847/1538-4357/acacff (cit. on pp. 11, 16).

Padovani, P., Alexander, D. M., et al. (Aug. 2017). 'Active galactic nuclei: what's in a name?' In: A&A Rev. 25.1, 2, p. 2. DOI: 10.1007/s00159-017-0102-9 (cit. on pp. 1, 2, 7).

Padovani, P. (Sept. 2016). 'The faint radio sky: radio astronomy becomes mainstream'. In: A&A Rev. 24.1, 13, p. 13. DOI: 10.1007/s00159-016-0098-6 (cit. on p. 12).

— (Nov. 2017). 'Active Galactic Nuclei at all wavelengths and from all angles'. In: *Frontiers in Astronomy and Space Sciences* 4, 35, p. 35. DOI: 10.3389/fspas.2017.00035 (cit. on p. 12).

Page, M. J. and Carrera, F. J. (Jan. 2000). 'An improved method of constructing binned luminosity functions'. In: MNRAS 311.2, pp. 433–440. DOI: 10.1046/j.1365-8711.2000.03105.x (cit. on pp. 126, 134, 212, 213).

Palanque-Delabrouille, N., Magneville, C., et al. (Mar. 2016). 'The extended Baryon Oscillation Spectroscopic Survey: Variability selection and quasar luminosity function'. In: A&A 587, A41, A41. DOI: 10.1051/0004-6361/201527392 (cit. on p. 134).

Papovich, C., Shipley, H. V., et al. (June 2016). 'The Spitzer-HETDEX Exploratory Large-area Survey'. In: ApJS 224.2, 28, p. 28. DOI: 10.3847/0067-0049/224/2/28 (cit. on p. 34).

Pasquato, M., Trevisan, P., et al. (Apr. 2024). 'Interpretable Machine Learning for Finding Intermediate-mass Black Holes'. In: ApJ 965.1, 89, p. 89. DOI: 10.3847/1538-4357/ad2261 (cit. on p. 28).

Pearl, A. N., Zentner, A. R., et al. (Mar. 2024). 'The DESI One-percent Survey: Evidence for Assembly Bias from Low-redshift Counts-in-cylinders Measurements'. In: ApJ 963.2, 116, p. 116. DOI: 10.3847/1538-4357/ad1ffd (cit. on p. 28).

Pearson, K. and Galton, F. (1895). 'VII. Note on regression and inheritance in the case of two parents'. In: *Proceedings of the Royal Society of London* 58.347-352, pp. 240–242. DOI: 10.1098/rspl.1895.0041 (cit. on p. 62).

Pedregosa, F., Varoquaux, G., et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830. DOI: 10.5555/1953048.2078195 (cit. on pp. 64, 170).

Pepinsky, T. B. (2018). 'A Note on Listwise Deletion versus Multiple Imputation'. In: *Political Analysis* 26.4, pp. 480–488. DOI: 10.1017/pan.2018.18 (cit. on p. 40).

Pereira, R., Couto, M., et al. (2017). 'Energy efficiency across programming languages: how do energy, time, and memory relate?' In: *Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering*. SLE 2017. Vancouver, BC, Canada: Association for Computing Machinery, pp. 256–267. ISBN: 9781450355254. DOI: 10.1145/3136014.3136031 (cit. on p. 20).

— (2021). 'Ranking programming languages by energy efficiency'. In: *Science of Computer Programming* 205, p. 102609. ISSN: 0167-6423. DOI: https://doi.org/10.1016/j.scico.2021.102609 (cit. on p. 20).

Pérez-Torres, M., Mattila, S., et al. (Dec. 2021). 'Star formation and nuclear activity in luminous infrared galaxies: an infrared through radio review'. In: A&A Rev. 29.1, 2, p. 2. DOI: 10.1007/s00159-020-00128-x (cit. on p. 4).

Perger, K., Frey, S., et al. (Aug. 2017). 'A catalogue of active galactic nuclei from the first 1.5 Gyr of the Universe'. In: *Frontiers in Astronomy and Space Sciences* 4, 9, p. 9. DOI: 10.3389/fspas.2017.00009 (cit. on p. 14).

Pineau, F.-X., Boch, T., et al. (Apr. 2020). 'The CDS Cross-match Service: Key Figures, Internals and Future Plans'. In: *Astronomical Data Analysis Software and Systems XXVII*. Ed. by P. Ballester, J. Ibsen, et al. Vol. 522. Astronomical Society of the Pacific Conference Series, p. 125 (cit. on p. 169).

Planck Collaboration, Aghanim, N., et al. (Sept. 2020). 'Planck 2018 results. VI. Cosmological parameters'. In: A&A 641, A6, A6. DOI: 10.1051/0004-6361/201833910 (cit. on p. 122).

Poisot, T. (2023). 'Guidelines for the prediction of species interactions through binary classification'. In: *Methods in Ecology and Evolution* 14.5, pp. 1333–1345. DOI: 10.1111/2041-210X.14071 (cit. on p. 57).

Poliszczuk, A., Pollo, A., et al. (July 2021). 'Active galactic nuclei catalog from the AKARI NEP-Wide field'. In: A&A 651, A108, A108. DOI: 10.1051/0004-6361/202040219 (cit. on pp. 89, 90).

Porqueres, N., Jasche, J., et al. (Apr. 2018). 'Imprints of the large-scale structure on AGN formation and evolution'. In: A&A 612, A31, A31. DOI: 10.1051/0004-6361/201732141 (cit. on p. 2).

Portegies Zwart, S. (Sept. 2020). 'The ecological impact of high-performance computing in astrophysics'. In: *Nature Astronomy* 4, pp. 819–822. DOI: 10.1038/s41550-020-1208-y (cit. on p. 20).

Pouliasis, E. (Feb. 2020). 'Identification of Active Galactic Nuclei through different selection techniques'. PhD thesis. IAASARS, National Observatory of Athens. DOI: 10.12681/eadd/47201 (cit. on p. 7).

Pracy, M. B., Ching, J. H. Y., et al. (July 2016). 'GAMA/WiggleZ: the 1.4 GHz radio luminosity functions of high- and low-excitation radio galaxies and their redshift evolution to z = 0.75'. In: MNRAS 460.1, pp. 2–17. DOI: 10.1093/mnras/stw910 (cit. on pp. 132–135, 137, 144).

Prandoni, I., Gregorini, L., et al. (Jan. 2001). 'The ATESP radio survey. III. Source counts'. In: A&A 365, pp. 392–399. DOI: 10.1051/0004-6361:20000142 (cit. on p. 131).

Prandoni, I., Guglielmino, G., et al. (Dec. 2018). 'The Lockman Hole Project: new constraints on the sub-mJy source counts from a wide-area 1.4 GHz mosaic'. In: MNRAS 481.4, pp. 4548–4565. DOI: 10.1093/mnras/sty2521 (cit. on p. 131).

Prandoni, I. and Seymour, N. (Apr. 2015). 'Revealing the Physics and Evolution of Galaxies and Galaxy Clusters with SKA Continuum Surveys'. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, 67, p. 67 (cit. on p. 6).

Prestage, R. M. and Peacock, J. A. (July 1983). 'Optical identifications of Parkes radio sources using UK Schmidt plates.' In: MNRAS 204, pp. 355–364. DOI: 10.1093/mnras/204.2.355 (cit. on p. 22).

Prokhorenkova, L., Gusev, G., et al. (2018). 'CatBoost: unbiased boosting with categorical features'. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, et al. Vol. 31. Curran Associates, Inc., pp. 6639–6649. DOI: 10.5555/3327757.3327770 (cit. on p. 64).

Quinn, D. P. and Smith, M. C. (Dec. 2009). 'A strip search for new very wide halo binaries'. In: MNRAS 400.4, pp. 2128–2134. DOI: 10.1111/j.1365-2966.2009.15607.x (cit. on p. 36).

Radcliffe, J. F., Barthel, P. D., et al. (May 2021a). 'Nowhere to hide: Radio-faint AGN in the GOODS-N field. II. Multi-wavelength AGN selection techniques and host galaxy properties'. In: A&A 649, A27, A27. DOI: 10.1051/0004-6361/202038591 (cit. on pp. 7, 113).

Radcliffe, J. F., Barthel, P. D., et al. (May 2021b). 'The radio emission from active galactic nuclei'. In: A&A 649, L9, p. L9. DOI: 10.1051/0004-6361/202140791 (cit. on p. 5).

Radcliffe, J. F., Garrett, M. A., et al. (Nov. 2018). 'Nowhere to Hide: Radio-faint AGN in GOODS-N field. I. Initial catalogue and radio properties'. In: A&A 619, A48, A48. DOI: 10.1051/0004-6361/201833399 (cit. on p. 211).

Rajagopal, M., Marchesi, S., et al. (June 2021). 'Identifying the 3FHL Catalog. V. Results of the CTIO-COSMOS Optical Spectroscopy Campaign 2019'. In: ApJS 254.2, 26, p. 26. DOI: 10.3847/1538-4365/abf656 (cit. on p. 15).

Rasmussen, C. E. and Williams, C. K. I. (Nov. 2005). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001 (cit. on p. 23).

Ratner, B. (June 2009). 'The correlation coefficient: Its values range between +1/-1, or do they?' In: *Journal of Targeting, Measurement and Analysis for Marketing* 17.2, pp. 139–142. ISSN: 1479-1862. DOI: 10.1057/jt.2009.5 (cit. on p. 63).

Rawlings, S. (Sept. 2003). 'High-redshift radio galaxies: at the crossroads'. In: New A Rev. 47.4-5, pp. 397–404. DOI: 10.1016/S1387-6473(03)00056-3 (cit. on p. 5).

Reis, I., Rotman, M., et al. (2021). 'Effectively using unsupervised machine learning in next generation astronomical surveys'. In: *Astronomy and Computing* 34, p. 100437. ISSN: 2213-1337. DOI: 10.1016/j.ascom.2020.100437 (cit. on p. 25).

Reis, I., Baron, D., and Shahaf, S. (Jan. 2019). 'Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets'. In: AJ 157.1, 16, p. 16. DOI: 10.3847/1538-3881/aaf101 (cit. on pp. 39, 153).

Reya, N. F., Ahmed, A., et al. (2023). 'GreenPy: evaluating application-level energy efficiency in Python for green computing'. In: *Ann. Emerg. Technol. Comput.(AETiC)* 7.3, pp. 93–110. DOI: 10.33166/aetic.2023.03.005 (cit. on p. 20).

Ricci, C., Trakhtenbrot, B., et al. (Dec. 2017). 'BAT AGN Spectroscopic Survey. V. X-Ray Properties of the Swift/BAT 70-month AGN Catalog'. In: ApJS 233.2, 17, p. 17. DOI: 10.3847/1538-4365/aa96ad (cit. on p. 7).

Richards, A. M. S., Muxlow, T. W. B., et al. (Sept. 2007). 'Using VO tools to investigate distant radio starbursts hosting obscured AGN in the HDF(N) region'. In: A&A 472.3, pp. 805–822. DOI: 10.1051/0004-6361:20077598 (cit. on p. 12).

Richards, G. T., Fan, X., et al. (June 2002). 'Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample'. In: AJ 123.6, pp. 2945–2975. DOI: 10.1086/340187 (cit. on p. 87).

Richards, G. T., Myers, A. D., et al. (Jan. 2009). 'Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ~1,000,000 Quasars from Data Release 6'. In: ApJS 180.1, pp. 67–83. DOI: 10.1088/0067-0049/180/1/67 (cit. on p. 87).

Richards, G. T., Strauss, M. A., et al. (June 2006). 'The Sloan Digital Sky Survey Quasar Survey: Quasar Luminosity Function from Data Release 3'. In: AJ 131.6, pp. 2766–2787. DOI: 10.1086/503559 (cit. on p. 129).

Richter, G. A. (Jan. 1975). 'Search for Optical Identifications in the 5C3 Radio Survey. II. Statistical Treatment and Results'. In: *Astronomische Nachrichten* 296.2, p. 65. DOI: 10.1002/asna.19752960203 (cit. on p. 22).

Rigby, E. E., Argyle, J., et al. (Sept. 2015). 'Cosmic downsizing of powerful radio galaxies to low radio luminosities'. In: A&A 581, A96, A96. DOI: 10.1051/0004-6361/201526475 (cit. on pp. 35, 144).

Robertson, B. E. (Aug. 2022). 'Galaxy Formation and Reionization: Key Unknowns and Expected Breakthroughs by the James Webb Space Telescope'. In: ARA&A 60, pp. 121–158. DOI: 10.1146/annurev-astro-120221-044656 (cit. on p. 1).

Rodrigo, C. and Solano, E. (July 2020). 'The SVO Filter Profile Service'. In: *XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society*, 182, p. 182 (cit. on p. 170).

Rodrigo, C., Solano, E., and Bayo, A. (Oct. 2012). *SVO Filter Profile Service Version 1.0*. IVOA Working Draft 15 October 2012. DOI: 10.5479/ADS/bib/2012ivoa.rept.1015R (cit. on p. 170).

Rodrigues, N. V. N., Raul Abramo, L., and Hirata, N. S. T. (Dec. 2023). 'The information of attribute uncertainties: what convolutional neural networks can learn about errors in input data'. In: *Machine Learning: Science and Technology* 4.4, 045019, p. 045019. DOI: 10.1088/2632-2153/ad0285 (cit. on p. 153).

Rohde, D. J., Drinkwater, M. J., et al. (June 2005). 'Applying machine learning to catalogue matching in astrophysics'. In: MNRAS 360.1, pp. 69–75. DOI: 10.1111/j.1365-2966.2005.08930.x (cit. on p. 23).

Rohde, D. J., Gallagher, M. R., et al. (June 2006). 'Matching of catalogues by probabilistic pattern classification'. In: MNRAS 369.1, pp. 2–14. DOI: 10.1111/j.1365-2966.2006.10304.x (cit. on pp. 23, 145).

Roscher, R., Bohn, B., et al. (2020a). 'Explainable Machine Learning for Scientific Insights and Discoveries'. In: *IEEE Access* 8, pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199 (cit. on p. 27).

— (2020b). 'Explainable Machine Learning for Scientific Insights and Discoveries'. In: *IEEE Access* 8, pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199 (cit. on p. 27).

Ross, N. P. and Cross, N. J. G. (May 2020). 'The near and mid-infrared photometric properties of known redshift z ≥ 5 quasars'. In: MNRAS 494.1, pp. 789–803. DOI: 10.1093/mnras/staa544 (cit. on pp. 6, 14).

Ross, N. P., McGreer, I. D., et al. (Aug. 2013). 'The SDSS-III Baryon Oscillation Spectroscopic Survey: The Quasar Luminosity Function from Data Release Nine'. In: ApJ 773.1, 14, p. 14. DOI: 10.1088/0004-637X/773/1/14 (cit. on p. 131).

Rowan-Robinson, M. (Jan. 1968). 'The determination of the evolutionary properties of quasars by means of the luminosity-volume test'. In: MNRAS 138, p. 445. DOI: 10.1093/mnras/138.4.445 (cit. on p. 212).

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471087052 (cit. on p. 41).

Rubin, D. B. (Dec. 1976). 'Inference and missing data'. In: *Biometrika* 63.3, pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581 (cit. on p. 20).

Rybicki, G. B. and Lightman, A. P. (2008). *Radiative Processes in Astrophysics*. Physics textbook. Wiley. ISBN: 9783527618187 (cit. on p. 5).

Saarela, M. and Jauhiainen, S. (Feb. 2021). 'Comparison of feature importance measures as explanations for classification models'. In: *SN Applied Sciences* 3.2, p. 272. ISSN: 2523-3971. DOI: 10.1007/s42452-021-04148-9 (cit. on p. 27).

Sabater, J., Best, P. N., et al. (Feb. 2019). 'The LoTSS view of radio AGN in the local Universe. The most massive galaxies are always switched on'. In: A&A 622, A17, A17. DOI: 10.1051/0004-6361/201833883 (cit. on pp. 13, 37, 118, 126).

Sadler, E. M., Jackson, C. A., et al. (Jan. 2002). 'Radio sources in the 2dF Galaxy Redshift Survey - II. Local radio luminosity functions for AGN and star-forming galaxies at 1.4 GHz'. In: MNRAS 329.1, pp. 227–245. DOI: 10.1046/j.1365-8711.2002.04998.x (cit. on p. 14).

Sajina, A., Lacy, M., and Pope, A. (June 2022). 'The Past and Future of Mid-Infrared Studies of AGN'. In: *Universe* 8.7, p. 356. DOI: 10.3390/universe8070356 (cit. on p. 6).

Salpeter, E. E. (Jan. 1955). 'The Luminosity Function and Stellar Evolution.' In: ApJ 121, p. 161. DOI: 10.1086/145971 (cit. on pp. 6, 116).

Salvato, M., Buchner, J., et al. (Feb. 2018). 'Finding counterparts for all-sky X-ray surveys with NWAY: a Bayesian algorithm for cross-matching multiple catalogues'. In: MNRAS 473.4, pp. 4937–4955. DOI: 10.1093/mnras/stx2651 (cit. on pp. 23, 92).

Salvato, M., Ilbert, O., et al. (Dec. 2011). 'Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples'. In: ApJ 742.2, 61, p. 61. DOI: 10.1088/0004-637X/742/2/61 (cit. on pp. 59, 92).

Salvato, M., Ilbert, O., and Hoyle, B. (June 2019). 'The many flavours of photometric redshifts'. In: *Nature Astronomy* 3, pp. 212–222. DOI: 10.1038/s41550-018-0478-0 (cit. on p. 16).

Samuel, A. L. (1959). 'Some Studies in Machine Learning Using the Game of Checkers'. In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210 (cit. on pp. 17, 29).

Sánchez, S. F., Avila-Reese, V., et al. (Apr. 2018). 'SDSS IV MaNGA - Properties of AGN Host Galaxies'. In: Rev. Mexicana Astron. Astrofis. 54, pp. 217–260. DOI: 10.48550/arXiv.1709.05438 (cit. on pp. 8, 9).

Sánchez-Sáez, P., Reyes, I., et al. (Mar. 2021). 'Alert Classification for the ALeRCE Broker System: The Light Curve Classifier'. In: AJ 161.3, 141, p. 141. DOI: 10.3847/1538-3881/abd5c1 (cit. on p. 45).

Sandage, A., Tammann, G. A., and Yahil, A. (Sept. 1979). 'The velocity field of bright nearby galaxies. I. The variation of mean absolute magnitude with redshift for galaxies in a magnitude-limited sample.' In: ApJ 232, pp. 352–364. DOI: 10.1086/157295 (cit. on p. 215).

Sandage, A. (Sept. 1962). 'The Change of Redshift and Apparent Luminosity of Galaxies due to the Deceleration of Selected Expanding Universes.' In: ApJ 136, p. 319. DOI: 10.1086/147385 (cit. on p. 211).

Santos, M. S., Abreu, P. H., et al. (Dec. 2022). 'On the joint-effect of class imbalance and overlap: a critical review'. In: *Artificial Intelligence Review* 55.8, pp. 6207–6275. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10150-3 (cit. on p. 68).

Sartori, L. F., Schawinski, K., et al. (Dec. 2015). 'The search for active black holes in nearby low-mass galaxies using optical and mid-IR data'. In: MNRAS 454.4, pp. 3722–3742. DOI: 10.1093/mnras/stv2238 (cit. on p. 7).

Saunders, C., Gammerman, A., and Vovk, V. (1999). 'Transduction with confidence and credibility'. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'99. Stockholm, Sweden: Morgan Kaufmann Publishers Inc., pp. 722–726. DOI: 10.5555/1624312.1624322 (cit. on pp. 38, 153).

Saunders, W., Rowan-Robinson, M., et al. (Jan. 1990). 'The 60-mu.m and far-infrared luminosity functions of IRAS galaxies.' In: MNRAS 242, pp. 318–337. DOI: 10.1093/mnras/242.3.318 (cit. on p. 215).

Savić, Đ. V., Jankov, I., et al. (Aug. 2023). 'The LSST AGN Data Challenge: Selection Methods'. In: ApJ 953.2, 138, p. 138. DOI: 10.3847/1538-4357/ace31a (cit. on pp. 36, 87, 88, 154).

Saz Parkinson, P. M., Xu, H., et al. (Mar. 2016). 'Classification and Ranking of Fermi LAT Gamma-ray Sources from the 3FGL Catalog using Machine Learning Techniques'. In: ApJ 820.1, 8, p. 8. DOI: 10.3847/0004-637X/820/1/8 (cit. on p. 25).

Schapire, R. E. (June 1990). 'The strength of weak learnability'. In: *Machine Learning* 5.2, pp. 197–227. ISSN: 1573-0565. DOI: 10.1007/BF00116037 (cit. on p. 25).

Schapire, R. E., Freund, Y., et al. (1998). 'Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods'. In: *The Annals of Statistics* 26.5, pp. 1651–1686. ISSN: 00905364. URL: http://www.jstor.org/stable/120016 (visited on 05/12/2023) (cit. on p. 26).

Schawinski, K., Thomas, D., et al. (Dec. 2007). 'Observational evidence for AGN feedback in early-type galaxies'. In: MNRAS 382.4, pp. 1415–1431. DOI: 10.1111/j.1365-2966.2007.12487.x (cit. on p. 8).

Schechter, P. (Jan. 1976). 'An analytic expression for the luminosity function for galaxies.' In: ApJ 203, pp. 297–306. DOI: 10.1086/154079 (cit. on pp. 6, 214).

Schmidt, M. (Mar. 1963). '3C 273 : A Star-Like Object with Large Red-Shift'. In: Nature 197.4872, p. 1040. DOI: 10.1038/1971040a0 (cit. on p. 6).

Schmidt, M. (Feb. 1968). 'Space Distribution and Luminosity Functions of Quasi-Stellar Radio Sources'. In: ApJ 151, p. 393. DOI: 10.1086/149446 (cit. on pp. 6, 212).

Schneider, D. P., Richards, G. T., et al. (June 2010). 'The Sloan Digital Sky Survey Quasar Catalog. V. Seventh Data Release'. In: AJ 139.6, 2360, p. 2360. DOI: 10.1088/0004-6256/139/6/2360 (cit. on p. 87).

Schneider, P. C., Freund, S., et al. (May 2022). 'The eROSITA Final Equatorial-Depth Survey (eFEDS). The Stellar Counterparts of eROSITA sources identified by machine learning and Bayesian algorithms'. In: A&A 661, A6, A6. DOI: 10.1051/0004-6361/202141133 (cit. on p. 23).

Schuecker, P. (Jan. 1993). 'Automated Galaxy Redshift Measurements from Very Low Dispersion Objective Prism Schmidt Plates'. In: ApJS 84, p. 39. DOI: 10.1086/191744 (cit. on p. 16).

Scott, D. and Rees, M. J. (Dec. 1990). 'The 21-cm line at high redshift: a diagnostic for the origin of large scale structure'. In: MNRAS 247, p. 510 (cit. on p. 122).

Scoville, N., Aussel, H., et al. (Sept. 2007). 'The Cosmic Evolution Survey (COSMOS): Overview'. In: ApJS 172.1, pp. 1–8. DOI: 10.1086/516585 (cit. on p. 15).

Secrest, N. J., Dudik, R. P., et al. (Nov. 2015). 'Identification of 1.4 Million Active Galactic Nuclei in the Mid-Infrared using WISE Data'. In: ApJS 221.1, 12, p. 12. DOI: 10.1088/0067-0049/221/1/12 (cit. on p. 113).

Selina, R. J., Murphy, E. J., et al. (Dec. 2018). 'The ngVLA Reference Design'. In: *Science with a Next Generation Very Large Array*. Ed. by E. Murphy. Vol. 517. Astronomical Society of the Pacific Conference Series, p. 15. DOI: 10.48550/arXiv.1810.08197 (cit. on p. 5).

Selina, R., Murphy, E., and Beasley, A. (Jan. 2023). 'The ngVLA: A Technical Overview'. In: *American Astronomical Society Meeting Abstracts*. Vol. 55. American Astronomical Society Meeting Abstracts, 357.02, p. 357.02 (cit. on p. 5).

Sen, S., Agarwal, S., et al. (Feb. 2022). 'Astronomical big data processing using machine learning: A comprehensive review'. In: *Experimental Astronomy* 53.1, pp. 1–43. DOI: 10.1007/s10686-021-09827-4 (cit. on p. 24).

Seyfert, C. K. (Jan. 1943). 'Nuclear Emission in Spiral Nebulae.' In: ApJ 97, p. 28. DOI: 10.1086/144488 (cit. on p. 6).

Shakura, N. I. and Sunyaev, R. A. (Jan. 1973). 'Black holes in binary systems. Observational appearance.' In: A&A 24, pp. 337–355 (cit. on p. 2).

Shapley, L. S. (1953). 'A Value for n-Person Games'. In: *Contributions to the Theory of Games (AM-28), Volume II*. Vol. 1. Princeton University Press, pp. 307–318. DOI: 10.1515/9781400881970-018 (cit. on p. 28).

Shen, X., Hopkins, P. F., et al. (Jan. 2020). 'The bolometric quasar luminosity function at z = 0-7'. In: MNRAS 495.3, pp. 3252–3275. DOI: 10.1093/mnras/staa1381 (cit. on p. 1).

Shimwell, T. W., Hardcastle, M. J., et al. (Mar. 2022). 'The LOFAR Two-metre Sky Survey. V. Second data release'. In: A&A 659, A1, A1. DOI: 10.1051/0004-6361/202142484 (cit. on p. 145).

Shimwell, T. W., Röttgering, H. J. A., et al. (Feb. 2017). 'The LOFAR Two-metre Sky Survey. I. Survey description and preliminary data release'. In: A&A 598, A104, A104. DOI: 10.1051/0004-6361/201629313 (cit. on pp. 5, 35).

Shimwell, T. W., Tasse, C., et al. (Feb. 2019). 'The LOFAR Two-metre Sky Survey. II. First data release'. In: A&A 622, A1, A1. DOI: 10.1051/0004-6361/201833559 (cit. on pp. 21, 34, 35, 134, 154).

Shwartz-Ziv, R. and Armon, A. (2022). 'Tabular data: Deep learning is not all you need'. In: *Information Fusion* 81, pp. 84–90. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.11.011 (cit. on p. 63).

Shy, S., Tak, H., et al. (July 2022). 'Incorporating Measurement Error in Astronomical Object Classification'. In: AJ 164.1, 6, p. 6. DOI: 10.3847/1538-3881/ac6e64 (cit. on pp. 39, 153).

Silva, L., Schurer, A., et al. (Jan. 2011). 'Modelling the spectral energy distribution of galaxies: introducing the artificial neural network'. In: MNRAS 410.3, pp. 2043–2056. DOI: 10.1111/j.1365-2966.2010.17580.x (cit. on p. 16).

Silva Filho, T., Song, H., et al. (Sept. 2023). 'Classifier calibration: a survey on how to assess and improve predicted class probabilities'. In: *Machine Learning* 112.9, pp. 3211–3260. ISSN: 1573-0565. DOI: 10.1007/s10994-023-06336-7 (cit. on p. 61).

Simpson, C., Rawlings, S., et al. (Apr. 2012). 'Radio imaging of the Subaru/XMM-Newton Deep Field- III. Evolution of the radio luminosity function beyond z= 1'. In: MNRAS 421.4, pp. 3060–3083. DOI: 10.1111/j.1365-2966.2012.20529.x (cit. on pp. 122, 135).

Singh, V., Beelen, A., et al. (Sept. 2014). 'Multiwavelength characterization of faint ultra steep spectrum radio sources: A search for high-redshift radio galaxies'. In: A&A 569, A52, A52. DOI: 10.1051/0004-6361/201423644 (cit. on p. 5).

Sipple, J. and Lidz, A. (Jan. 2024). 'The Star Formation Efficiency during Reionization as Inferred from the Hubble Frontier Fields'. In: ApJ 961.1, 50, p. 50. DOI: 10.3847/1538-4357/ad06a7 (cit. on p. 1).

Skrutskie, M. F., Cutri, R. M., et al. (Feb. 2006). 'The Two Micron All Sky Survey (2MASS)'. In: AJ 131.2, pp. 1163–1183. DOI: 10.1086/498708 (cit. on p. 18).

Šlaus, B., Smolčić, V., et al. (June 2020). 'The XXL Survey. XLI. Radio AGN luminosity functions based on the GMRT 610 MHz continuum observations'. In: A&A 638, A46, A46. DOI: 10.1051/0004-6361/201937258 (cit. on pp. 122, 129, 135).

Šlaus, B., Smolčić, V., et al. (Apr. 2024). 'The XXL survey. LII. The evolution of radio AGN LF determined via parametric methods from GMRT, ATCA, VLA, and Cambridge interferometer observations'. In: A&A 684, A19, A19. DOI: 10.1051/0004-6361/202346947 (cit. on pp. 132–134, 137).

Slob, M. M., Callingham, J. R., et al. (Dec. 2022). 'Extragalactic peaked-spectrum radio sources at low frequencies are young radio galaxies'. In: A&A 668, A186, A186. DOI: 10.1051/0004-6361/202244651 (cit. on p. 137).

Smith, M. J. and Geach, J. E. (May 2023). 'Astronomia ex machina: a history, primer and outlook on neural networks in astronomy'. In: *Royal Society Open Science* 10.5, 221454, p. 221454. DOI: 10.1098/rsos.221454 (cit. on p. 152).

Smolčić, V., Novak, M., et al. (June 2017). 'The VLA-COSMOS 3 GHz Large Project: Cosmic evolution of radio AGN and implications for radio-mode feedback since z 5'. In: A&A 602, A6, A6. DOI: 10.1051/0004-6361/201730685 (cit. on p. 137).

Smolčić, V., Schinnerer, E., et al. (Jan. 2009a). 'The Dust-Unbiased Cosmic Star-Formation History from the 20 CM VLA-COSMOS Survey'. In: ApJ 690.1, pp. 610–618. DOI: 10.1088/0004-637X/690/1/610 (cit. on p. 215).

Smolčić, V., Zamorani, G., et al. (May 2009b). 'Cosmic Evolution of Radio Selected Active Galactic Nuclei in the Cosmos Field'. In: ApJ 696.1, pp. 24–39. DOI: 10.1088/0004-637X/696/1/24 (cit. on p. 14).

Snyder, J. (1987). *Map Projections–a Working Manual*. Professional paper. U.S. Government Printing Office. ISBN: 9780318235622. DOI: 10.3133/pp1395 (cit. on p. 15).

— (1997). *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press. ISBN: 9780226767475 (cit. on p. 15).

Sola, J. and Sevilla, J. (1997). 'Importance of input data normalization for the application of neural networks to complex industrial problems'. In: *IEEE Transactions on Nuclear Science* 44.3, pp. 1464–1468. DOI: 10.1109/23.589532 (cit. on p. 48).

Sollich, P. and Krogh, A. (1995). 'Learning with ensembles: How overfitting can be useful'. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M. Mozer, and M. Hasselmo. Vol. 8. MIT Press, pp. 190–196. DOI: 10.5555/2998828.2998855 (cit. on p. 25).

Sommer, M. W., Basu, K., et al. (May 2011). 'Redshift evolution of the 1.4 GHz volume averaged radio luminosity function in clusters of galaxies'. In: A&A 529, A124, A124. DOI: 10.1051/0004-6361/201016150 (cit. on p. 211).

Sørenson, T. (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content*. Biologiske skrifter. I kommission hos E. Munksgaard (cit. on p. 55).

Sravan, N., Graham, M. J., et al. (July 2023). 'Machine-directed gravitational-wave counterpart discovery'. In: *arXiv e-prints*, arXiv:2307.09213, arXiv:2307.09213. DOI: 10.48550/arXiv.2307.09213 (cit. on p. 24).

Stark, D. P. (Sept. 2016). 'Galaxies in the First Billion Years After the Big Bang'. In: ARA&A 54, pp. 761–803. DOI: 10.1146/annurev-astro-081915-023417 (cit. on p. 214).

Steidel, C. C., Adelberger, K. L., et al. (July 1999). 'Lyman-Break Galaxies at z>∼4 and the Evolution of the Ultraviolet Luminosity Density at High Redshift'. In: ApJ 519.1, pp. 1–17. DOI: 10.1086/307363 (cit. on p. 6).

Steidel, C. C., Giavalisco, M., et al. (May 1996a). 'Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at Redshifts Z > 3'. In: ApJ 462, p. L17. DOI: 10.1086/310029 (cit. on p. 17).

Steidel, C. C., Giavalisco, M., et al. (Aug. 1996b). 'Spectroscopy of Lyman Break Galaxies in the Hubble Deep Field'. In: AJ 112, p. 352. DOI: 10.1086/118019 (cit. on p. 17).

Steidel, C. C. and Hamilton, D. (Sept. 1992). 'Deep Imaging of redshift QSO Fields Below the Lyman Limit. I. The Field of Q0000-263 and galaxies at Z= 3.4'. In: AJ 104, p. 941. DOI: 10.1086/116287 (cit. on p. 17).

Stern, D. (July 2015). 'The X-Ray to Mid-infrared Relation of AGNs at High Luminosity'. In: ApJ 807.2, 129, p. 129. DOI: 10.1088/0004-637X/807/2/129 (cit. on p. 7).

Stern, D., Assef, R. J., et al. (July 2012). 'Mid-infrared Selection of Active Galactic Nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected Active Galactic Nuclei in COSMOS'. In: ApJ 753.1, 30, p. 30. DOI: 10.1088/0004-637X/753/1/30 (cit. on pp. xxvi, 11, 84).

Stern, D., Eisenhardt, P., et al. (Sept. 2005). 'Mid-Infrared Selection of Active Galaxies'. In: ApJ 631.1, pp. 163–168. DOI: 10.1086/432523 (cit. on p. 10).

Stone, M. (1974). 'Cross-Validatory Choice and Assessment of Statistical Predictions'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 111–133. DOI: 10.1111/j.2517-6161.1974.tb00994.x (cit. on p. 64).

Storey-Fisher, K., Hogg, D. W., et al. (Mar. 2024). 'Quaia, the Gaia-unWISE Quasar Catalog: An All-sky Spectroscopic Quasar Sample'. In: ApJ 964.1, 69, p. 69. DOI: 10.3847/1538-4357/ad1328 (cit. on pp. 12, 22, 119).

Storey-Fisher, K., Huertas-Company, M., et al. (Dec. 2021). 'Anomaly detection in Hyper Suprime-Cam galaxy images with generative adversarial networks'. In: MNRAS 508.2, pp. 2946–2963. DOI: 10.1093/mnras/stab2589 (cit. on p. 24).

Stoughton, C., Lupton, R. H., et al. (Jan. 2002). 'Sloan Digital Sky Survey: Early Data Release'. In: AJ 123.1, pp. 485–548. DOI: 10.1086/324741 (cit. on p. 16).

Suganuma, M., Yoshii, Y., et al. (Mar. 2006). 'Reverberation Measurements of the Inner Radius of the Dust Torus in Nearby Seyfert 1 Galaxies'. In: ApJ 639.1, pp. 46–63. DOI: 10.1086/499326 (cit. on p. 12).

Surana, S., Wadadekar, Y., et al. (Apr. 2020). 'Predicting star formation properties of galaxies using deep learning'. In: MNRAS 493.4, pp. 4808–4815. DOI: 10.1093/mnras/staa537 (cit. on p. 21).

Sutherland, W. and Saunders, W. (Dec. 1992). 'On the likelihood ratio for source identification.' In: MNRAS 259, pp. 413–420. DOI: 10.1093/mnras/259.3.413 (cit. on p. 22).

Sweijen, F., van Weeren, R. J., et al. (Jan. 2022). 'Deep sub-arcsecond wide-field imaging of the Lockman Hole field at 144 MHz'. In: *Nature Astronomy* 6, pp. 350–356. DOI: 10.1038/s41550-021-01573-z (cit. on p. 22).

Tacchella, S., Conroy, C., et al. (Feb. 2022). 'Fast, Slow, Early, Late: Quenching Massive Galaxies at z ∼ 0.8'. In: ApJ 926.2, 134, p. 134. DOI: 10.3847/1538-4357/ac449b (cit. on p. 19).

Tanaka, M., Coupon, J., et al. (Jan. 2018). 'Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1'. In: PASJ 70, S9, S9. DOI: 10.1093/pasj/psx077 (cit. on pp. 15, 16).

Taylor, M. B. (Dec. 2005). 'TOPCAT & STIL: Starlink Table/VOTable Processing Software'. In: *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell, M. Britton, and R. Ebert. Vol. 347. Astronomical Society of the Pacific Conference Series, p. 29. ISBN: 978-1-58381-281-5 (cit. on p. 169).

Thomas, N., Davé, R., et al. (May 2021). 'The radio galaxy population in the SIMBA simulations'. In: MNRAS 503.3, pp. 3492–3509. DOI: 10.1093/mnras/stab654 (cit. on p. 6).

Thorne, J. E., Robotham, A., et al. (Mar. 2022a). *AGN Unification Diagram*. DOI: 10.5281/zenodo.6381013 (cit. on p. 3).

Thorne, J. E., Robotham, A. S. G., et al. (Feb. 2022b). 'Deep Extragalactic VIsible Legacy Survey (DEVILS): identification of AGN through SED fitting and the evolution of the bolometric AGN luminosity function'. In: MNRAS 509.4, pp. 4940–4961. DOI: 10.1093/mnras/stab3208 (cit. on pp. 84, 90).

Toba, Y., Oyabu, S., et al. (June 2014). 'Luminosity and Redshift Dependence of the Covering Factor of Active Galactic Nuclei viewed with WISE and Sloan Digital Sky Survey'. In: ApJ 788.1, 45, p. 45. DOI: 10.1088/0004-637X/788/1/45 (cit. on pp. 7, 11).

Tong, G., Li, F., and Allen, A. S. (2019). 'Missing Data'. In: *Principles and Practice of Clinical Trials*. Ed. by S. Piantadosi and C. L. Meinert. Cham: Springer International Publishing, pp. 1–21. ISBN: 978-3-319-52677-5. DOI: 10.1007/978-3-319-52677-5_117-1 (cit. on p. 41).

Tonry, J. and Davis, M. (Oct. 1979). 'A survey of galaxy redshifts. I. Data reduction techniques.' In: AJ 84, pp. 1511–1525. DOI: 10.1086/112569 (cit. on p. 16).

Toth, M. J., Goran, M. I., et al. (1993). 'Examination of data normalization procedures for expressing peak VO2 data'. In: *Journal of applied physiology* 75.5, pp. 2288–2292. DOI: 10.1152/jappl.1993.75.5.2288 (cit. on p. 48).

Trenti, M. and Stiavelli, M. (Apr. 2008). 'Cosmic Variance and Its Effect on the Luminosity Function Determination in Deep High-z Surveys'. In: ApJ 676.2, pp. 767–780. DOI: 10.1086/528674 (cit. on p. 133).

Tripodi, R., Feruglio, C., et al. (Sept. 2022). 'Black hole and host galaxy growth in an isolated z ∼ 6 QSO observed with ALMA'. In: A&A 665, A107, A107. DOI: 10.1051/0004-6361/202243920 (cit. on p. 1).

Troyer, J., Starkey, D., et al. (Mar. 2016). 'Correlated X-ray/ultraviolet/optical variability in NGC 6814'. In: MNRAS 456.4, pp. 4040–4050. DOI: 10.1093/mnras/stv2862 (cit. on p. 12).

Tulio Ribeiro, M., Singh, S., and Guestrin, C. (Feb. 2016). '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In: *arXiv e-prints*, arXiv:1602.04938, arXiv:1602.04938. DOI: 10.48550/arXiv.1602.04938 (cit. on p. 28).

U, V. (July 2022). 'The Role of AGN in Luminous Infrared Galaxies from the Multiwavelength Perspective'. In: *Universe* 8.8, 392, p. 392. DOI: 10.3390/universe8080392 (cit. on p. 2).

Ulmer-Moll, S., Santos, N. C., et al. (Oct. 2019). 'Beyond the exoplanet mass-radius relation'. In: A&A 630, A135, A135. DOI: 10.1051/0004-6361/201936049 (cit. on p. 28).

Urry, C. (June 2004). 'AGN Unification: An Update'. In: *AGN Physics with the Sloan Digital Sky Survey*. Ed. by G. T. Richards and P. B. Hall. Vol. 311. Astronomical Society of the Pacific Conference Series, p. 49. DOI: 10.48550/arXiv.astro-ph/0312545 (cit. on pp. 2, 7).

Urry, C. M. and Padovani, P. (Sept. 1995). 'Unified Schemes for Radio-Loud Active Galactic Nuclei'. In: PASP 107, p. 803. DOI: 10.1086/133630 (cit. on pp. 2, 7).

Uttley, P., Edelson, R., et al. (Feb. 2003). 'Correlated Long-Term Optical and X-Ray Variations in NGC 5548'. In: ApJ 584.2, pp. L53–L56. DOI: 10.1086/373887 (cit. on p. 12).

Uzgil, B. D., Oesch, P. A., et al. (May 2021). 'The ALMA Spectroscopic Survey in the HUDF: A Search for [C II] Emitters at 6 ≤ z ≤ 8'. In: ApJ 912.1, 67, p. 67. DOI: 10.3847/1538-4357/abe86b (cit. on p. 17).

Van Calster, B., McLernon, D. J., et al. (Dec. 2019). 'Calibration: the Achilles heel of predictive analytics'. In: *BMC Medicine* 17.1, p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1466-7 (cit. on p. 61).

van den Busch, J. L., Hildebrandt, H., et al. (Oct. 2020). 'Testing KiDS cross-correlation redshifts with simulations'. In: A&A 642, A200, A200. DOI: 10.1051/0004-6361/202038835 (cit. on p. 15).

van der Velden, E. (Feb. 2020). 'CMasher: Scientific colormaps for making accessible, informative and 'cmashing' plots'. In: *The Journal of Open Source Software* 5.46, 2004, p. 2004. DOI: 10.21105/joss.02004 (cit. on p. 170).

van der Vlugt, D., Algera, H. S. B., et al. (Jan. 2021). 'An Ultradeep Multiband VLA Survey of the Faint Radio Sky (COSMOS-XS): Source Catalog and Number Counts'. In: ApJ 907.1, 5, p. 5. DOI: 10.3847/1538-4357/abcaa3 (cit. on p. 134).

van der Vlugt, D., Hodge, J. A., et al. (Dec. 2022). 'An Ultra-deep Multiband Very Large Array (VLA) Survey of the Faint Radio Sky (COSMOS-XS): New Constraints on the Cosmic Star Formation History'. In: ApJ 941.1, 10, p. 10. DOI: 10.3847/1538-4357/ac99db (cit. on pp. 122, 129, 131–134, 215).

van Haarlem, M. P., Wise, M. W., et al. (July 2013). 'LOFAR: The LOw-Frequency ARray'. In: A&A 556, A2, A2. DOI: 10.1051/0004-6361/201220873 (cit. on pp. 34, 167).

van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd. USA: Butterworth-Heinemann. ISBN: 0408709294 (cit. on p. 55).

Vanden Berk, D. E., Wilhite, B. C., et al. (Feb. 2004). 'The Ensemble Photometric Variability of ~25,000 Quasars in the Sloan Digital Sky Survey'. In: ApJ 601.2, pp. 692–714. DOI: 10.1086/380563 (cit. on p. 12).

Vanschoren, J. (2019). 'Meta-Learning'. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by F. Hutter, L. Kotthoff, and J. Vanschoren. Cham: Springer International Publishing, pp. 35–61. ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5_2 (cit. on p. 26).

Vanzella, E., Cristiani, S., et al. (Aug. 2004). 'Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS'. In: A&A 423, pp. 761–776. DOI: 10.1051/0004-6361:20040176 (cit. on p. 25).

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer. ISBN: 9780387945590 (cit. on p. 23).

Vardoulaki, E., Jiménez Andrade, E. F., et al. (Apr. 2021). 'FR-type radio sources at 3 GHz VLA-COSMOS: Relation to physical properties and large-scale environment'. In: A&A 648, A102, A102. DOI: 10.1051/0004-6361/202039488 (cit. on p. 24).

Veilleux, S. and Osterbrock, D. E. (Feb. 1987). 'Spectral Classification of Emission-Line Galaxies'. In: ApJS 63, p. 295. DOI: 10.1086/191166 (cit. on p. 7).

Véron-Cetty, M. .-. and Véron, P. (Jan. 1984). 'A catalogue of Quasars and Active Nuclei.' In: *European Southern Observatory Scientific Report* 1, p. 1 (cit. on p. 14).

— (1985). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 4 (cit. on p. 14).

— (1987). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 5 (cit. on p. 14).

— (Jan. 1989). 'A Catalogue of Quasars and Active Nuclei (4th Edition).' In: *European Southern Observatory Scientific Report* 7, p. 1 (cit. on p. 14).

— (1991). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 10 (cit. on p. 14).

— (1993). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 13 (cit. on p. 14).

— (1996). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 17 (cit. on p. 14).

— (1998). 'A Catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 18 (cit. on p. 14).

— (2000). 'A catalogue of quasars and active nuclei'. In: *European Southern Observatory Scientific Report* 19 (cit. on p. 14).

— (July 2001). 'A catalogue of quasars and active nuclei: 10th edition'. In: A&A 374, pp. 92–94. DOI: 10.1051/0004-6361:20010718 (cit. on p. 14).

— (Dec. 2003). 'A catalogue of quasars and active nuclei: 11th edition'. In: A&A 412, pp. 399–403. DOI: 10.1051/0004-6361:20034225 (cit. on p. 14).

— (Aug. 2006). 'A catalogue of quasars and active nuclei: 12th edition'. In: A&A 455.2, pp. 773–777. DOI: 10.1051/0004-6361:20065177 (cit. on p. 14).

— (July 2010). 'A catalogue of quasars and active nuclei: 13th edition'. In: A&A 518, A10, A10. DOI: 10.1051/0004-6361/201014188 (cit. on p. 14).

Villaescusa-Navarro, F., Anglés-Alcázar, D., et al. (July 2021). 'The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations'. In: ApJ 915.1, 71, p. 71. DOI: 10.3847/1538-4357/abf7ba (cit. on p. 27).

Virtanen, P., Gommers, R., et al. (2020). 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on pp. 20, 170).

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Cham: Springer International Publishing. ISBN: 978-3-031-06649-8. DOI: 10.1007/978-3-031-06649-8 (cit. on pp. 38, 153).

Vuttipittayamongkol, P., Elyan, E., and Petrovski, A. (2021). 'On the class overlap problem in imbalanced data classification'. In: *Knowledge-Based Systems* 212, p. 106631. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2020.106631 (cit. on p. 68).

Wagstaff, K. L., Huff, E., and Rebbapragada, U. (July 2022). 'Machine-Assisted Discovery Through Identification and Explanation of Anomalies in Astronomical Surveys'. In: *Astronomical Society of the Pacific Conference Series*. Ed. by J. E. Ruiz, F. Pierfedereci, and P. Teuben. Vol. 532. Astronomical Society of the Pacific Conference Series, p. 183 (cit. on p. 24).

Walcher, J., Groves, B., et al. (Jan. 2011). 'Fitting the integrated spectral energy distributions of galaxies'. In: Ap&SS 331, pp. 1–52. DOI: 10.1007/s10509-010-0458-z (cit. on p. 84).

Warren, S. J., Hewett, P. C., et al. (Jan. 1987). 'First observation of a quasar with a redshift of 4'. In: Nature 325.6100, pp. 131–133. DOI: 10.1038/325131a0 (cit. on p. 17).

Wasleske, E. J. and Baldassare, V. F. (Aug. 2023). 'X-Ray Emission of Ultraviolet Variable Active Galactic Nucleus Candidates'. In: AJ 166.2, 64, p. 64. DOI: 10.3847/1538-3881/ace16b (cit. on p. 7).

Wenger, M., Ochsenbein, F., et al. (Apr. 2000). 'The SIMBAD astronomical database. The CDS reference database for astronomical objects'. In: A&AS 143, pp. 9–22. DOI: 10.1051/aas:2000332 (cit. on p. 169).

Wenzl, L., Schindler, J.-T., et al. (Aug. 2021). 'Random Forests as a Viable Method to Select and Discover High-redshift Quasars'. In: AJ 162.2, 72, p. 72. DOI: 10.3847/1538-3881/ac0254 (cit. on p. 24).

Werner, M. W., Roellig, T. L., et al. (Sept. 2004). 'The Spitzer Space Telescope Mission'. In: ApJS 154.1, pp. 1–9. DOI: 10.1086/422992 (cit. on p. 8).

Whittam, I. H., Prescott, M., et al. (Jan. 2024). 'MIGHTEE: Multi-wavelength counterparts in the COSMOS field'. In: MNRAS 527.2, pp. 3231–3245. DOI: 10.1093/mnras/stad3307 (cit. on p. 22).

Wilber, A. G., Dabbech, A., et al. (July 2023). 'Scalable precision wide-field imaging in radio interferometry - II. AIRI validated on ASKAP data'. In: MNRAS 522.4, pp. 5576–5587. DOI: 10.1093/mnras/stad1353 (cit. on p. 24).

Wilks, D. S. (1990). 'On the Combination of Forecast Probabilities for Consecutive Precipitation Periods'. In: *Weather and Forecasting* 5.4, pp. 640–650. DOI: 10.1175/1520-0434(1990)005<0640:OTCOFP>2.0.CO;2 (cit. on p. 61).

Williams, W. L., Calistro Rivera, G., et al. (Apr. 2018). 'LOFAR-Boötes: properties of high- and low-excitation radio galaxies at 0.5 < z < 2.0'. In: MNRAS 475.3, pp. 3429–3452. DOI: 10.1093/mnras/sty026 (cit. on pp. 5, 137).

Williams, W. L. and Röttgering, H. J. A. (June 2015). 'Radio-AGN feedback: when the little ones were monsters'. In: MNRAS 450.2, pp. 1538–1545. DOI: 10.1093/mnras/stv692 (cit. on p. 13).

Willott, C. J., Rawlings, S., et al. (Apr. 2001). 'The radio luminosity function from the low-frequency 3CRR, 6CE and 7CRS complete samples'. In: MNRAS 322.3, pp. 536–552. DOI: 10.1046/j.1365-8711.2001.04101.x (cit. on p. 137).

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. ISBN: 9780080890364 (cit. on p. 21).

Wolpert, D. H. (1992). 'Stacked generalization'. In: *Neural Networks* 5.2, pp. 241–259. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80023-1 (cit. on p. 26).

Wolstencroft, R. D., Savage, A., et al. (Nov. 1986). 'The identification of IRAS point sources- I. A 304 deg2 field centred on the South Galactic Pole.' In: MNRAS 223, pp. 279–302. DOI: 10.1093/mnras/223.2.279 (cit. on p. 22).

Wright, E. L., Eisenhardt, P. R. M., et al. (Dec. 2010). 'The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance'. In: AJ 140.6, pp. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868 (cit. on pp. 8, 18, 21).

Wu, C., Wong, O. I., et al. (Jan. 2019). 'Radio Galaxy Zoo: CLARAN - a deep learning classifier for radio morphologies'. In: MNRAS 482.1, pp. 1211–1230. DOI: 10.1093/mnras/sty2646 (cit. on p. 90).

Xue, Y. Q., Luo, B., et al. (July 2011). 'The Chandra Deep Field-South Survey: 4 Ms Source Catalogs'. In: ApJS 195.1, 10, p. 10. DOI: 10.1088/0067-0049/195/1/10 (cit. on p. 22).

Yan, L., Donoso, E., et al. (Mar. 2013). 'Characterizing the Mid-infrared Extragalactic Sky with WISE and SDSS'. In: AJ 145.3, 55, p. 55. DOI: 10.1088/0004-6256/145/3/55 (cit. on p. 113).

Yan, W., Brandt, W. N., et al. (July 2023). 'The Most Obscured AGNs in the XMM-SERVS Fields'. In: ApJ 951.1, 27, p. 27. DOI: 10.3847/1538-4357/accea6 (cit. on p. 4).

Yang, H., Zhou, L., et al. (Feb. 2023). 'Data mining techniques on astronomical spectra data - II. Classification analysis'. In: MNRAS 518.4, pp. 5904–5928. DOI: 10.1093/mnras/stac3292 (cit. on p. 152).

Yang, J. (Sept. 2021). 'Fast TreeSHAP: Accelerating SHAP Value Computation for Trees'. In: *arXiv e-prints*, arXiv:2109.09847, arXiv:2109.09847. DOI: 10.48550/arXiv.2109.09847 (cit. on p. 99).

Yang, Q. and Shen, Y. (Jan. 2023). 'A Southern Photometric Quasar Catalog from the Dark Energy Survey Data Release 2'. In: ApJS 264.1, 9, p. 9. DOI: 10.3847/1538-4365/ac9ea8 (cit. on p. 119).

Ye, H., Sweijen, F., et al. (Sept. 2023). '1-arcsecond imaging strategy for the LoTSS survey using the International LOFAR Telescope'. In: *arXiv e-prints*, arXiv:2309.16560, arXiv:2309.16560. DOI: 10.48550/arXiv.2309.16560 (cit. on p. 22).

Yeo, I.-K. and Johnson, R. A. (Dec. 2000). 'A new family of power transformations to improve normality or symmetry'. In: *Biometrika* 87.4, pp. 954–959. ISSN: 0006-3444. DOI: 10.1093/biomet/87.4.954 (cit. on p. 48).

Yerushalmy, J. (1947). 'Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques'. In: *Public Health Reports (1896-1970)* 62.40, pp. 1432–1449. ISSN: 00946214. DOI: 10.2307/4586294 (cit. on p. 56).

Yong, S. Y. and Ong, C. S. (Sept. 2023). 'Uncertainty quantification of the virial black hole mass with conformal prediction'. In: MNRAS 524.2, pp. 3116–3129. DOI: 10.1093/mnras/stad2080 (cit. on p. 153).

York, D. G., Adelman, J., et al. (Sept. 2000). 'The Sloan Digital Sky Survey: Technical Summary'. In: AJ 120.3, pp. 1579–1587. DOI: 10.1086/301513 (cit. on pp. 15, 36, 37).

Yuan, F., Lidman, C., et al. (Sept. 2015). 'OzDES multifibre spectroscopy for the Dark Energy Survey: first-year operation and results'. In: MNRAS 452.3, pp. 3047–3063. DOI: 10.1093/mnras/stv1507 (cit. on p. 92).

Yuan, Z., Jarvis, M. J., and Wang, J. (May 2020). 'A Flexible Method for Estimating Luminosity Functions via Kernel Density Estimation'. In: ApJS 248.1, 1, p. 1. DOI: 10.3847/1538-4365/ab855b (cit. on p. 213).

Yuan, Z. and Wang, J. (June 2013). 'A graphical analysis of the systematic error of classical binned methods in constructing luminosity functions'. In: Ap&SS 345.2, pp. 305–313. DOI: 10.1007/s10509-013-1402-9 (cit. on p. 134).

Yuan, Z., Wang, J., et al. (Sept. 2017). 'A Mixture Evolution Scenario of the AGN Radio Luminosity Function. II. Do Low- and High-power Radio-loud AGNs Evolve Differently?' In: ApJ 846.1, 78, p. 78. DOI: 10.3847/1538-4357/aa8463 (cit. on p. 144).

Yuan, Z., Zhang, X., et al. (May 2022). 'A Flexible Method for Estimating Luminosity Functions via Kernel Density Estimation. II. Generalization and Python Implementation'. In: ApJS 260.1, 10, p. 10. DOI: 10.3847/1538-4365/ac596a (cit. on pp. 131, 132, 213).

Yule, G. U. (1912). 'On the Methods of Measuring Association Between Two Attributes'. In: *Journal of the Royal Statistical Society* 75.6, pp. 579–652. ISSN: 09528385. DOI: 10.2307/2340126 (cit. on p. 55).

Zajaček, M., Busch, G., et al. (Oct. 2019). 'Radio spectral index distribution of SDSS-FIRST sources across optical diagnostic diagrams'. In: A&A 630, A83, A83. DOI: 10.1051/0004-6361/201833388 (cit. on p. 13).

Zammit, M. A. and Adami, K. Z. (Nov. 2023). 'Machine Learning Applications in Jupiter-host Star Classification using Stellar Spectra'. In: MNRAS. DOI: 10.1093/mnras/stad3668 (cit. on p. 26).

Zeraatgari, F. Z., Hafezianzadeh, F., et al. (Jan. 2024). 'Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars'. In: MNRAS 527.3, pp. 4677–4689. DOI: 10.1093/mnras/stad3436 (cit. on pp. 113, 152).

Zhang, Y. and Zhao, Y. (May 2015). 'Astronomy in the Big Data Era'. In: *Data Science Journal* 14, p. 11. DOI: 10.5334/dsj-2015-011 (cit. on p. 19).

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly. ISBN: 9781491953242 (cit. on p. 45).

Zitlau, R., Hoyle, B., et al. (Aug. 2016). 'Stacking for machine learning redshifts applied to SDSS galaxies'. In: MNRAS 460.3, pp. 3152–3162. DOI: 10.1093/mnras/stw1454 (cit. on p. 26).

Zou, H., Gao, J., et al. (May 2019). 'Photometric Redshifts and Stellar Masses for Galaxies from the DESI Legacy Imaging Surveys'. In: ApJS 242.1, 8, p. 8. DOI: 10.3847/1538-4365/ab1847 (cit. on p. 120).

Appendices

This page intentionally left blank.

A

Luminosity function formulation

As presented in Sect. 1.1, fluxes (F) can be used to obtain an estimate of the luminosities (L) emitted by an astrophysical source. The intial formula to obtain luminosities (L) from fluxes is:

$$L = 4\pi r^2 F\,, \tag{A.1}$$

with r being the distance from the observer to the emitting source. This expression does not take any change in the path of the light into account and assumes no interactions of it with intervening materials.

The first modification that can be applied is related to the distance at which the source is from the observer. Thus, a correction for its redshift must be applied (e.g. Sandage 1962). Additionally, a term with the correction when transforming between observed and rest-frame fluxes, that is between the waveband of the detector and that of the observed source can be included (K-correction; Oke and Sandage 1968). For the specific case of radio bands, it is possible to assume their continuum spectra can be modelled as a power law in the form $S_\nu \propto \nu^\alpha$ (Sommer et al. 2011), with α, the spectral index, typically taking values around $\alpha \sim -0.7$. Then, and taking into account the mentioned corrections, Eq. A.1 can be written, for rest-frame luminosities, as:

$$L_{\nu,\mathrm{rest}} = 4\pi D_L^2 S_{\nu,\mathrm{obs}} \frac{\mathcal{K}}{(1+z)}\,, \tag{A.2}$$

in which D_L represents the luminosity distance to the source (Hogg 1999), $S_{\nu,\mathrm{obs}}$ corresponds to the observed-frame flux at frequency ν, and $\mathcal{K}$ stands for the aforementioned K-correction. For the radio emission, the K-correction can take the form $(1+z)^{-\alpha}$ (Condon et al. 2002; Radcliffe et al. 2018; Cochrane et al. 2023), with α being the same spectral index, and then, Eq. A.2 leads to:

$$L_{\nu,\mathrm{rest}} = \frac{4\pi D_L^2}{(1+z)^{1+\alpha}} S_{\nu,\mathrm{obs}} \,. \tag{A.3}$$

Considering that the radio emission has been assumed to have the shape of a power law, it is possible to use this fact to transform luminosities in different frequencies without complex procedures (Delhaize et al. 2017). If a luminosity is measured at a frequency ν_a, the way to convert it into the luminosity at frequency ν_b, using the spectral index, α, is through:

$$L_{\nu_b} = \frac{4\pi D_L^2}{(1+z)^{1+\alpha}} \left(\frac{\nu_b}{\nu_a}\right)^{\alpha} S_{\nu_a} \,. \tag{A.4}$$

As described in Sect. 1.1, LFs can provide a measure of the evolution of the density of sources in different luminosity and redshift bins. It is possible to define the LF using a differential approach. In this way, the LF, ϕ, the number of sources per unit luminosity, L, and unit volume, V, can be defined as:

$$\phi\,(L,z) = \frac{\mathrm{d}^2\mathbb{d}}{\mathrm{d}V\mathrm{d}L}\,(L,z)\ , \tag{A.5}$$

in which $\mathbb{d}$ corresponds to the number of objects with luminosity between L and $L + \mathrm{d}L$ and in a volume of range V and $V + \mathrm{d}V$ (which is a function of the redshift z). Equation A.5 can also be written using the logarithm of the luminosity as:

$$\phi\left(\log_{10} L, z\right) = \frac{\mathrm{d}^2\mathbb{d}}{\mathrm{d}V\mathrm{d}\left(\log_{10} L\right)} \,. \tag{A.6}$$

Using a sample of measured luminosities, it is possible to estimate the LF using different methods that take into account the general properties of the observations from which the measurements were obtained (e.g. the survey depth and the area it covers). One way in which LFs can be quantified is by the use of the $\langle V/V_{\mathrm{max}}\rangle$ method (Kafka 1967; Rowan-Robinson 1968; Schmidt 1968). An application of this method is the $1/V_{\mathrm{max}}$ technique (e.g. Felten 1976; Avni and Bahcall 1980; Maccacaro et al. 1991; Ellis et al. 1996). The $1/V_{\mathrm{max}}$ technique has been devised to measure the space number density ($\mathrm{d}\mathbb{d}\mathrm{d}V$) via the small interval approximation:

$$\frac{\mathrm{d}\mathbb{d}}{\mathrm{d}V} \approx \sum_{i=1}^{\mathbb{d}} \frac{1}{V_a(i)} \,, \tag{A.7}$$

where $V_a(i)$ corresponds to the volume in which the object i, with luminosity $L(i)$ might have been still detected in the redshift bin Δz (Page and Carrera 2000). With this approximation, the

luminosity function can be estimated as:

$$\phi_{1/V_a}(L, z) = \frac{1}{\Delta L} \sum_{i=1}^{\mathbb{d}} \frac{1}{V_a(i)} \,. \tag{A.8}$$

Another technique for the estimation of the LF values is called binned approximation. First described by Page and Carrera (2000), it refines the region in which each step of the LF is calculated as assuming a dependence on the redshift of the sources (as originally presented in the definition of LF, Eq. A.5), concept which the $1/V_{\rm max}$ technique omits.

For the binned LF, it is assumed that the plane $L - z$ is divided into bins limited by fixed redshift values and luminosity values that depend on the conditions of the studied survey (i.e. the minimum luminosity at which a source can be detected). Therefore, the LF can be approximated as:

$$\phi \approx \phi_{\rm binned} = \frac{\mathbb{d}}{\int_{z_{\rm min}}^{z_{\rm max}} \int_{L_{\rm min}(L)}^{L_{\rm max}} \frac{dV}{dz} dz dL} \,, \tag{A.9}$$

where $\mathbb{d}$ corresponds to the number of sources in each bin. In this case, Eq. A.9 needs to be calculated for the central point of each relevant bin of the plane $L - z$.

Finally, further methods to estimate the values of the LF can consider more complex mathematical expressions. That is the case of its calculation using a kernel density estimation (KDE) approach (Abramson 1982; Davies et al. 2018b). As presented by Yuan et al. (2020, 2022), its implementation incorporates the addition of several kernels (one per source in the sample). Then, its complexity arises at the determination of the bandwidths (and additional parameters) of the kernels. The final expression of the KDE-based LF is then:

$$\phi \approx \phi_{\rm KDE} = \frac{\mathbb{d}\,(z_2 - z_1)\,\hat{f}\,(x, y|h_1, h_2)}{(z - z_1)\,(z_2 - z)\,\Omega \frac{\mathrm{d}V}{\mathrm{d}z}} \,, \tag{A.10}$$

where z_1 and z_2 are the redshift limits of the studied sample, $\hat{f}(x, y|h_1, h_2)$ corresponds to the density function of the pair (x, y) given the KDE bandwidths h_1, h_2, which is the equivalent of (z, L) in the KDE space, and $\mathbb{d}$ is the size of such sample.

As expected, all the previous methods allow to count the sources that are located in a specific area of the sky. However, this count might be biased by the conditions of the measurements in the used catalogues. A selection function, $\mathcal{P}$, is used to correct the sources

count. $\mathcal{P}$ summarises the corrections that the distribution of sources must suffer in order to be as close as possible to our best guess of their real distribution. In this way, $\mathbb{d}$ becomes $\mathbb{d}_{\rm eff}$, the effective number of sources in the sample.

All the previous methods are non-parametric, meaning that they do not assume any analytic form or shape of the studied LF. Contrary to that, several parametric models have been developed to determine an expression for the LF. The use of analytical expression for the formulation of LFs can be used to compare their shapes and evolution among several techniques and datasets.

For the analysis of source densities, a number of functional forms have been proposed for the parametrisation of LFs. One of the most used expressions is that of Schechter (1976), which was derived for galaxies in optical wavelengths. The Schechter function consists of a power law with an exponential decline at the bright end. It is usually formulated as follows:

$$\phi\left(L\right) = \phi^{*}\left(\frac{L}{L^{*}}\right)^{\alpha_{\rm LF}} \exp\left(-\frac{L}{L^{*}}\right), \tag{A.11}$$

where $\alpha_{\rm LF}$ corresponds to the exponent of the power law, L^{*} indicates the luminosity at which the decline of the function starts, and ϕ^{*} is a normalising factor. The value for all three factors can be obtained from the fit of the function to the measured data points. The Schechter LF is typically used in optical or UV studies of galaxies (for a review of its use in high-redshift UV sources see, for instance, Stark 2016).

For the study of radio sources, different formulations have been proposed. Typically, a double power law (one for fainter sources and the other for brighter sources) is used (e.g. Dunlop and Peacock 1990; Brown et al. 2001; Mauch and Sadler 2007) following the expression

$$\phi\left(L\right) = \frac{\phi^{*}}{\left(\frac{L^{*}}{L}\right)^{\alpha_{\rm LF}} + \left(\frac{L^{*}}{L}\right)^{\beta_{\rm LF}}}. \tag{A.12}$$

with ϕ^{*}, L^{*}, $\alpha_{\rm LF}$, and $\beta_{\rm LF}$ parameters to be obtained from the fit to the measured data. On the other hand, Condon (1984, 1989) presented a RLF that has a hyperbolic form (as compiled by Condon et al. 2002)

$$\phi\left(L\right) = 28.83 - \frac{3}{2}\log_{10}\left(L_{1.4\,\rm GHz}\right) + Y - \left[B^{2} + \frac{\left(\log_{10}\left(L_{1.4\,\rm GHz}\right) - X\right)^{2}}{W^{2}}\right]^{1/2}, \tag{A.13}$$

in which Y, B, X, and Y are parameters fitted to the data and the selected cosmology. This

parametrisation has been used, for example, by Machalski and Godlowski (2000), Condon et al. (2002), and Gupta et al. (2017, 2020).

Alternatively, and from the study of FIR sources, Saunders et al. (1990) developed a LF that has the general shape (following the proposed form by Sandage et al. 1979)

$$\phi\left(L\right) = \phi^{*}\left(\frac{L^{*}}{L}\right)^{(1-\alpha_{\rm LF})} \exp\left[-\frac{1}{2\sigma^{2}}\log^{2}\left(1+\frac{L}{L^{*}}\right)\right], \tag{A.14}$$

in which a power law dominates at $L < L^{*}$ and a Gaussian at $L > L^{*}$. The new parameter, $\sigma_{\rm LF}$ controls the bright end of the distribution. Such formulation has been used, for example, by Smolčić et al. (2009a) and van der Vlugt et al. (2022). While this formulation offers a useful framework, it may not fully capture the complexities of RL AGN as discussed by (Mauch and Sadler 2007). This difference highlights the potential advantages of using Eq. A.12 for RLFs.

This page intentionally left blank.

B

Sample of predicted radio-detectable AGN

The columns shown in the prediction results for sources in both HETDEX and S82 fields (Chapter 3) are described in Table B.1. Its first column shows the name of the variables while the second column presents their description. Tables B.2, B.3, B.4, and B.5 display the prediction information for the first 20 sources in their respective subsets (HETDEX test subset, S82 labelled sources, HETDEX unlabelled sources, and S82 unlabelled sources, respectively) sorted by decreasing predicted redshift. Full datasets and models from prediction pipeline can be obtained from Carvajal et al. (2023b) at https://zenodo.org/doi/10.5281/zenodo.10220008. Cutouts for the newly predicted radio-AGN described in Tables B.4 and B.5 (unlabelled HETDEX and S82 subsets, respectively) are included as well in Fig. B.1 for HETDEX and Fig. B.2 for S82. Images have been obtained, in the case of the HETDEX field, from the LoTSS-DR2 survey and, for the sources in the S82 field, from the VLAS82 survey.

Table B.1: Table columns descriptions.

ID	Internal identification number
RA_ICRS	Right Ascension (in degrees) of source in CW
DE_ICRS	Declination (in degrees) of source in CW
Name	Name of source as it appears in CW catalogue
band_num	Number of non-radio bands with valid measurement per source (cf. Sect. 2.5)
class	1 if source is confirmed AGN by MQC. 0 if spectroscopically confirmed as SFG in SDSS-DR16. Sources with no value do not have spectroscopic classification in this catalogue
Sint_LOFAR (or Fint_VLAS82)[a]	Imputed integrated flux (in mJy) of source from LOFAR or VLAS82
Sint_LOFAR_non_imp (or Fint_VLAS82_non_imp)[a]	Non imputed integrated flux (in mJy) of source from LOFAR or VLAS82
W1mproPM	Imputed *W1* magnitude of source
W2mproPM	Imputed *W2* magnitude of source
gmag	Imputed *g* magnitude of source
rmag	Imputed *r* magnitude of source
imag	Imputed *i* magnitude of source
zmag	Imputed *z* magnitude of source
ymag	Imputed *y* magnitude of source
W3mag	Imputed *W3* magnitude of source
W4mag	Imputed *W4* magnitude of source
Jmag	Imputed *J* magnitude of source
Hmag	Imputed *H* magnitude of source
Kmag	Imputed *Ks* magnitude of source
Score_AGN	Score from meta AGN-SFG classifier to be AGN
Prob_AGN	Probability from calibrated meta AGN-SFG classifier AGN
LOFAR_detect	1 if source has been detected on LoTSS-DR1 or in analogue surveys for different fields (see Sects. 2.1 and 2.2). 0 otherwise
Score_radio_AGN	Score from meta radio detection model to be detected in radio
Prob_radio_AGN	Probability from calibrated radio detection model to be detected in radio
radio_AGN	class × LOFAR_detect. 1 if source is AGN and has been detected in radio. 0 otherwise
Score_rAGN	Score_AGN × Score_radio. Score of source to be AGN detected in radio
Prob_rAGN	Prob_AGN × Prob_radio. Probability of source to be AGN detected in radio
Z	Spectroscopic redshift as listed by the MQC (if available)
pred_Z	Redshift value predicted by our model

[a] Sources from HETDEX field have columns with LOFAR suffix. Sources from VLAS82 have VLAS82 suffix.

Table B.2: Predicted and original properties for 20 sources in testing subset with probability of being AGN higher than 75 % and probability of being detected in the radio higher than 75 %. Sources sorted by decreasing predicted redshift.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	class	Score_AGN	Prob_AGN	LOFAR_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	z	pred_z
1 902 745	166.180954	50.515976	12	1	0.500103	0.978645	1	0.907702	0.758989	0.453945	0.742781	0.166	1.5759
4 006 444	219.706543	45.888813	9	1	0.500119	0.987831	0	0.904532	0.755586	0.452374	0.746392	1.927	1.5116
10 137 651	192.613922	50.668171	8	0	0.500047	0.853210	1	0.917604	0.770030	0.458845	0.656998	0.317	1.3737
3 755 378	162.411804	55.768215	9	1	0.500105	0.979757	0	0.962190	0.832508	0.481196	0.815656	1.878	1.3653
8 236 546	198.748169	51.977886	9	1	0.500105	0.980115	0	0.936183	0.792880	0.468190	0.777114	0.599	1.1378
11 978 048	220.690002	46.463566	9	1	0.500109	0.982264	1	0.923282	0.776683	0.461741	0.762908	0.246	1.0272
13 056 274	227.962921	52.427773	12	1	0.500120	0.988429	1	0.908502	0.759856	0.454360	0.751064	1.135	0.9497
7 493 576	200.010376	48.593803	9	1	0.500117	0.986877	0	0.908724	0.760098	0.454468	0.750123	1.290	0.8404
1 512 922	179.778275	53.112061	12	1	0.500121	0.988717	0	0.930354	0.785357	0.465290	0.776496	0.482	0.6856
4 452 097	215.366074	47.503265	12	1	0.500120	0.988429	1	0.923160	0.776538	0.461691	0.767553	0.372	0.6683
6 024 210	188.937332	48.454311	9	1	0.500055	0.886451	1	0.943275	0.802576	0.471689	0.711444	1.023	0.6543
1 457 919	175.730774	52.230663	9	1	0.500115	0.985849	1	0.913986	0.765919	0.457098	0.755080	1.239	0.6217
11 526 075	194.468185	56.491222	9	1	0.500111	0.983666	0	0.955346	0.820837	0.477779	0.807429	0.500	0.5822
11 572 993	225.653214	46.226662	9	1	0.500088	0.963304	0	0.914355	0.766333	0.457258	0.738212	0.433	0.5215
12 312 511	222.086884	49.540306	9	1	0.500039	0.816701	1	0.952610	0.816468	0.476343	0.666810	0.426	0.4347
5 078 026	207.513031	48.828613	9	1	0.500118	0.987521	1	0.899876	0.750692	0.450045	0.741324	0.310	0.4133
6 822 082	184.805313	51.346802	12	1	0.500069	0.930237	1	0.933342	0.789168	0.466736	0.734113	0.171	0.2993
8 103 101	197.249725	49.852257	12	1	0.500046	0.851377	1	0.928487	0.783022	0.464286	0.666646	0.169	0.1889
10 677 338	190.747543	52.878620	12	1	0.500101	0.976491	1	0.943539	0.802950	0.471865	0.784073	0.170	0.1850
11 364 079	184.971283	55.751850	12	1	0.500111	0.983489	1	0.971107	0.849799	0.485661	0.835768	0.107	0.1599

Table B.3: Properties for the 20 sources in S82 with highest predicted redshift on labelled sources with probability of being AGN higher than 75 % and probability of being detected in the radio higher than 75 %. Sources sorted by decreasing predicted redshift.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	class	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	z	pred_z
3 412 405	341.279388	1.240154	8	1	0.500075	0.942365	0	0.899861	0.750676	0.449998	0.707411	2.773	2.3840
2 152 145	16.782793	0.268437	9	1	0.500041	0.824738	0	0.925780	0.779694	0.462928	0.643044	2.474	2.2683
3 550 919	341.773010	1.390980	8	1	0.500082	0.955373	0	0.968683	0.844813	0.484422	0.807112	1.322	2.2577
1 321 215	342.140900	−0.369507	8	1	0.500106	0.980467	0	0.927200	0.781431	0.463698	0.766168	2.182	1.6667
1 761 104	26.202965	−0.031640	7	1	0.500030	0.758513	0	0.918961	0.771597	0.459508	0.585267	0.662	1.5522
262 099	342.219910	−1.170675	8	1	0.500096	0.971927	0	0.910198	0.761710	0.455186	0.740327	0.689	1.4672
373 370	344.050629	−1.085508	8	1	0.500058	0.897328	1	0.951657	0.814980	0.475884	0.731305	0.651	1.3394
1 992 304	346.538452	0.145603	12	1	0.500120	0.988346	0	0.939947	0.797945	0.470087	0.788646	1.344	1.2874
1 802 982	10.864357	0.000554	9	1	0.500091	0.966860	1	0.911189	0.762802	0.455677	0.737523	1.110	1.2290
1 265 539	24.285915	−0.410777	9	1	0.500119	0.987655	0	0.945336	0.805523	0.472780	0.795578	1.220	1.1823
2 940 564	336.693878	0.869805	9	1	0.500112	0.984409	1	0.938661	0.796195	0.469436	0.783781	2.247	0.9585
1 434 167	345.759491	−0.283792	7	1	0.500068	0.927101	0	0.907829	0.759125	0.453976	0.703786	0.639	0.8271
2 978 289	13.052857	0.899000	9	1	0.500082	0.954432	1	0.967873	0.843199	0.484016	0.804776	0.689	0.8231
2 129 339	27.427147	0.250460	12	1	0.500121	0.988512	0	0.915034	0.767100	0.457627	0.758287	0.552	0.7927
2 267 890	25.954859	0.357989	9	1	0.500042	0.828909	0	0.954102	0.818832	0.477091	0.678737	1.323	0.7924
185 533	31.697111	−1.232174	12	1	0.500086	0.959946	0	0.947128	0.808136	0.473645	0.775767	0.254	0.6576
197 420	22.591999	−1.222085	7	1	0.500057	0.893232	0	0.903913	0.754928	0.452008	0.674325	0.527	0.6325
2 742 138	17.874889	0.718236	9	1	0.500067	0.923836	1	0.932483	0.788063	0.466304	0.728041	0.501	0.6055
2 693 416	30.275667	0.681665	9	1	0.500115	0.986050	1	0.921805	0.774927	0.461009	0.764118	0.517	0.5838
1 306 677	23.234962	−0.380373	9	1	0.500072	0.935919	1	0.909428	0.760866	0.454779	0.712109	0.095	0.5826

Table B.4: Predicted and original properties for the 20 sources from unlabelled sources in the HETDEX field with probability of being AGN higher than 75 % and probability of being detected in the radio higher than 75 %. Sources sorted by decreasing predicted redshift.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	pred_z
10 961 010	184.426529	53.141224	5	0.500046	0.849058	0	0.900656	0.751504	0.450369	0.638070	4.2693
778 887	168.165359	46.682114	6	0.500035	0.793882	0	0.948099	0.809574	0.474083	0.642706	4.0131
6 865 314	184.201157	52.477249	6	0.500058	0.898994	0	0.944490	0.804305	0.472300	0.723065	3.6694
6 709 724	184.289200	49.760670	5	0.500042	0.831476	0	0.940897	0.799252	0.470488	0.664559	3.3578
13 576 233	216.509552	51.169346	9	0.500031	0.767079	0	0.927260	0.781505	0.463659	0.599476	2.2669
9 717 970	196.101685	54.323223	9	0.500084	0.957497	1	0.931452	0.786747	0.465804	0.753308	2.1750
7 130 068	201.169128	46.119282	8	0.500032	0.776064	0	0.946665	0.807456	0.473363	0.626638	2.1231
6 916 244	182.151321	51.299366	9	0.500100	0.975984	0	0.937245	0.794292	0.468716	0.775216	2.0945
6 839 868	183.741501	51.902031	7	0.500031	0.763152	0	0.956488	0.822706	0.478273	0.627850	1.8873
1 371 388	173.322937	50.434879	9	0.500099	0.975291	0	0.948189	0.809709	0.474189	0.789702	1.8784
4 125 564	206.366638	46.239353	9	0.500067	0.924855	1	0.921111	0.774109	0.460618	0.715939	1.8395
7 795 661	196.923538	47.662121	8	0.500076	0.944120	1	0.976973	0.863019	0.488561	0.814793	1.8079
15 125 585	206.967789	56.156597	9	0.500106	0.980537	1	0.963244	0.834415	0.481724	0.818174	1.7941
4 189 572	206.137772	47.411507	9	0.500095	0.971627	0	0.903082	0.754049	0.451627	0.732655	1.7869
2 874 018	174.158203	54.144489	4	0.500034	0.784184	1	0.927848	0.782230	0.463955	0.613412	1.7449
5 027 408	208.582047	48.194042	8	0.500094	0.970293	0	0.951168	0.814224	0.475674	0.790035	1.7066
6 280 199	184.405975	46.924526	7	0.500054	0.884605	0	0.925043	0.778801	0.462572	0.688931	1.7064
8 815 801	201.293304	52.171860	4	0.500046	0.849989	1	0.908481	0.759834	0.454282	0.645850	1.7015
4 426 189	214.601044	47.052723	6	0.500029	0.752459	0	0.933165	0.788939	0.466609	0.593644	1.6943
10 796 259	187.873978	55.745792	7	0.500109	0.982826	0	0.970626	0.848790	0.485419	0.834212	1.6763

Table B.5: Predicted and original properties for the 20 sources from unlabelled sources in the S82 field with probability of being AGN higher than 75 % and probability of being detected in the radio higher than 75 %. Sources sorted by decreasing predicted redshift.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	pred_z
2 927 274	342.541626	0.859697	7	0.500075	0.942167	0	0.907758	0.759049	0.453947	0.715151	2.9072
1 214 845	17.745253	−0.448830	7	0.500031	0.766428	0	0.930906	0.786054	0.465482	0.602453	2.8768
1 283 463	342.137756	−0.397631	6	0.500029	0.752459	0	0.980341	0.871561	0.490199	0.655814	2.7214
2 870 731	342.317749	0.816674	7	0.500043	0.836019	0	0.917919	0.770393	0.458999	0.644063	2.4490
3 565 273	29.208164	1.416150	8	0.500077	0.945073	0	0.981683	0.875213	0.490917	0.827140	2.3102
3 524 618	12.516848	1.354527	6	0.500059	0.899981	0	0.902679	0.753624	0.451393	0.678248	2.0308
77 878	17.529116	−1.335152	9	0.500112	0.984184	0	0.901816	0.752717	0.451009	0.740812	1.9694
999 027	12.822252	−0.610582	9	0.500104	0.979320	0	0.929686	0.784517	0.464940	0.768293	1.9058
2 347 779	336.604706	0.419295	9	0.500038	0.807251	1	0.914791	0.766825	0.457430	0.619020	1.8332
1 125 054	339.766754	−0.515258	4	0.500035	0.791488	1	0.925146	0.778925	0.462605	0.616510	1.7549
7953	26.582243	−1.450007	8	0.500086	0.960778	0	0.921340	0.774379	0.460749	0.744006	1.7320
460 908	331.872528	−1.019004	8	0.500104	0.979320	0	0.931471	0.786770	0.465833	0.770500	1.7062
213 737	13.658834	−1.208855	7	0.500060	0.905106	0	0.929352	0.784099	0.464732	0.709692	1.7039
2 937 396	349.847290	0.867445	8	0.500045	0.846709	1	0.952573	0.816409	0.476330	0.691261	1.6961
133 501	334.452972	−1.278622	12	0.500110	0.983371	0	0.915613	0.767756	0.457908	0.754989	1.6638
3 296 696	335.001831	1.144865	4	0.500040	0.821557	1	0.923509	0.776955	0.461792	0.638313	1.6422
1 487 824	25.028767	−0.242225	4	0.500037	0.805545	0	0.912676	0.764452	0.456372	0.615801	1.6227
2 720 760	341.770447	0.702141	8	0.500088	0.963046	0	0.903075	0.754041	0.451617	0.726176	1.5965
1 986 718	336.789001	0.141474	8	0.500105	0.979901	0	0.926490	0.780561	0.463342	0.764873	1.5874
2 916 309	20.715971	0.851349	4	0.500032	0.774159	0	0.951093	0.814108	0.475577	0.630249	1.5730

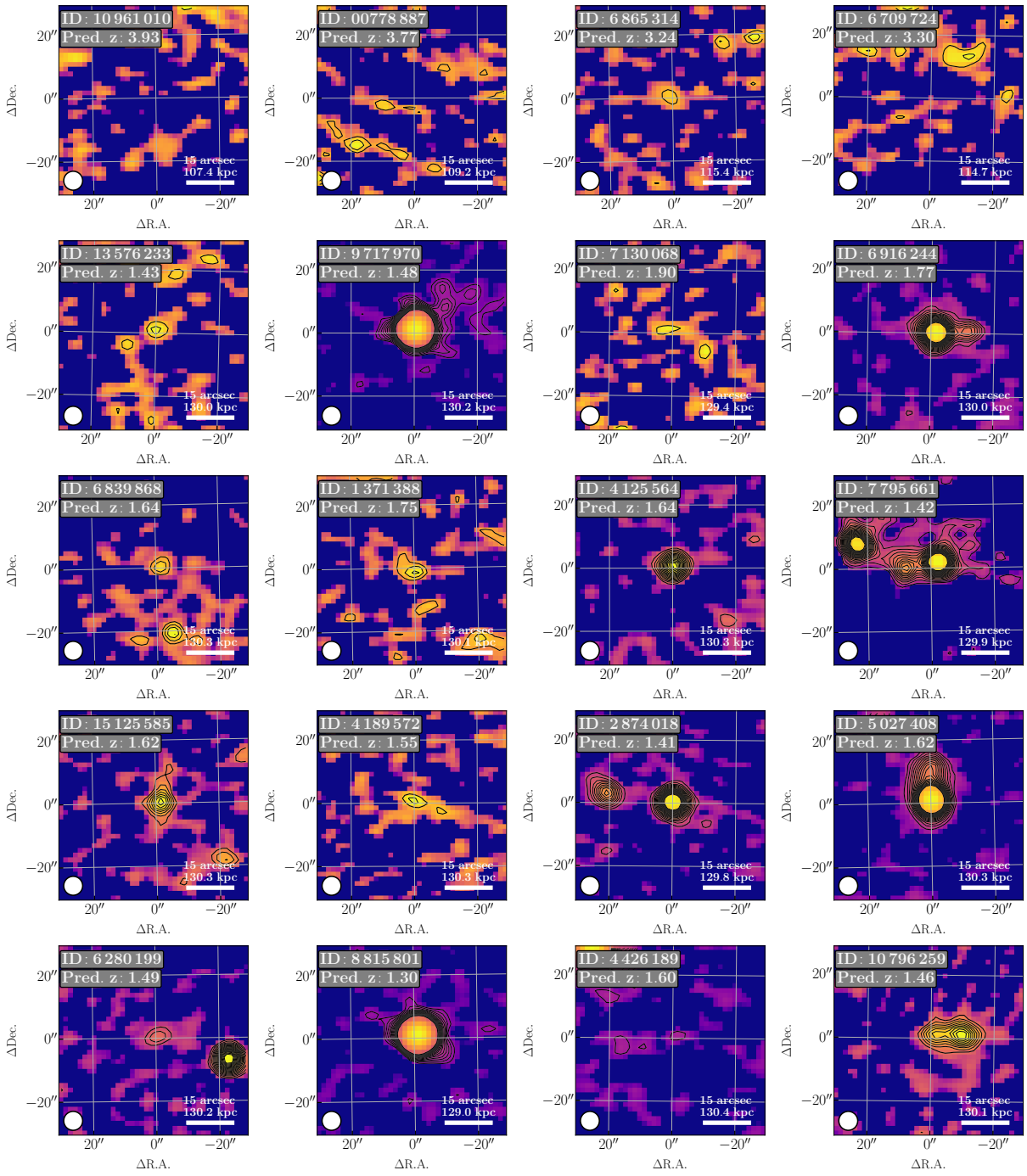


Figure B.1: Postage stamps of CW-detected sources predicted (with high probability) to be radio-detected AGN from the unlabelled sources in the HETDEX field. Sources, labelled by their identification number (in the top right corner of each cutout) are described in Table B.4. Bottom left corner of each image shows the size of the primary beam of the LoTSS-DR2 survey ($\sim 28\,\mathrm{arcsec}^2$), while the bar in the bottom right corner presents a scale of $15\,''$ in physical units using the predicted redshift. As in Fig. 5.21 and for displaying purposes, all emission below $1\,\sigma$ of the distribution of each image has been set to zero (0). Pixels with brighter colours represent areas with more radio emission in the observed area and black contours show emission at $2\,\sigma$ to $20\,\sigma$ levels (using the nominal noise level of the survey).

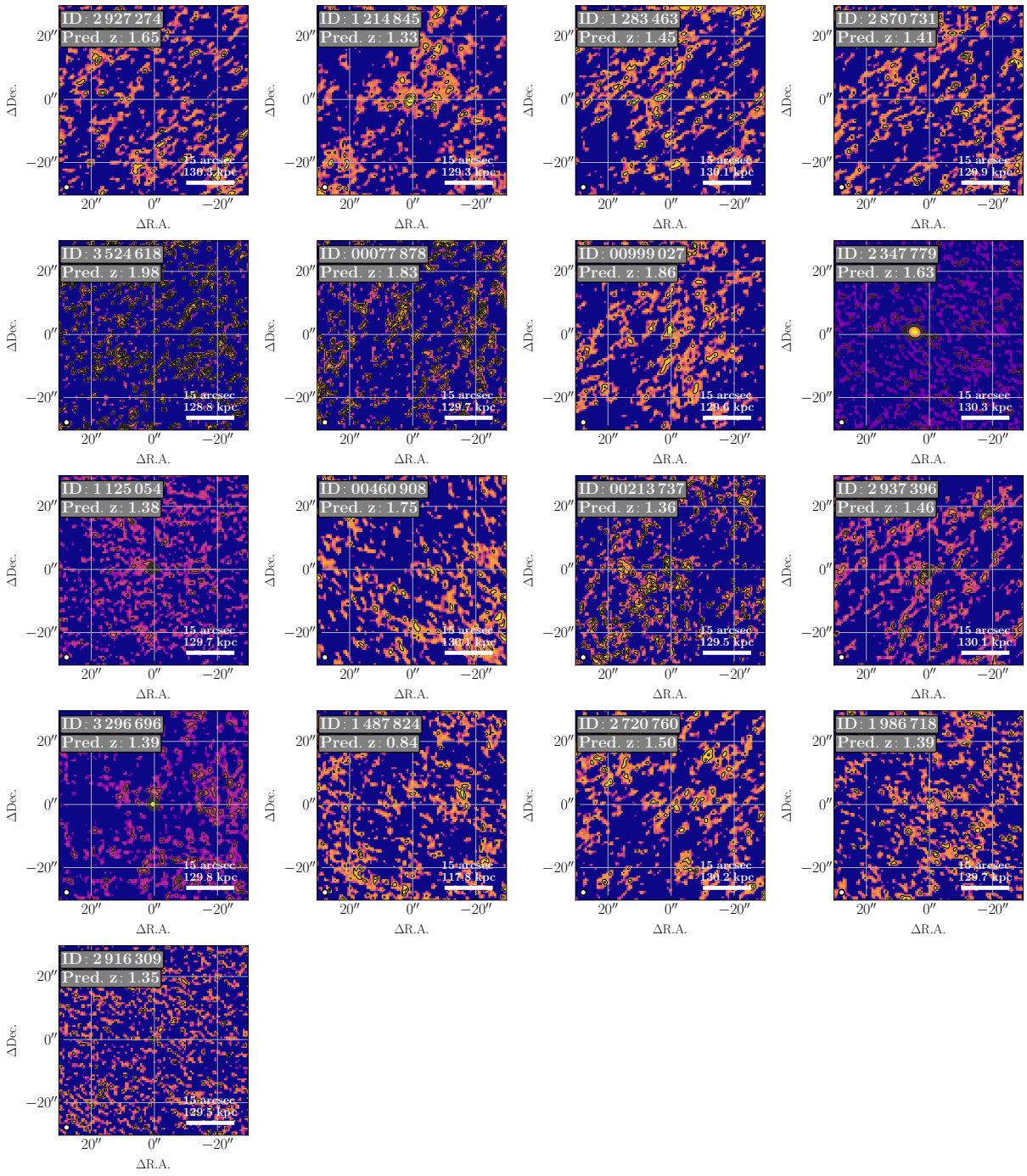


Figure B.2: Postage stamps of CW-detected sources predicted (with high probability) to be radio-detected AGN from the unlabelled sources in the S82 field. Sources, labelled by their identification number (in the top right corner of each cutout) are described in Table B.5. Bottom left corner of each image shows the size of the primary beam of the VLAS82 survey ($\sim 2.5\,\mathrm{arcsec}^2$), while the bar in the bottom right corner presents a scale of $15''$ in physical units using the predicted redshift. As in Fig B.1 and for displaying purposes, all emission below $1\,\sigma$ of the distribution of each image has been set to zero (0). Pixels with brighter colours represent areas with more radio emission in the observed area and black contours show emission at $2\,\sigma$ to $20\,\sigma$ levels (using the nominal noise level of the survey).

C

Extended prediction pipeline

For some of the calculations in Chapter 5, a new instance of the prediction pipeline was trained and implemented. Its steps were kept similar to its original definition (Sect. 1.4) with some relevant differences. The first change is in the overall structure of the pipeline. The prediction for radio-detected AGN works as presented in Fig. 1.9, but a new branch has been added for the treatment of radio-detected SFGs (i.e. extragalactic sources without indications of AGN emission). This new branch replicates the steps of the original process and its stages are presented, graphically, in the flowchart of Fig. C.1.

The difference in the processing of the candidates start when sources are predicted to be SFGs (i.e. not to be AGN). Instead of being discarded, they are subjected to a series of models that replicate the process for predicted AGN. Thus, a new step that predicts their likelihood of being radio detectable is applied. The predicted radio-detectable SFGs have, then, their photometric redshifts predicted. Finally, the results from both branches, AGN and SFGs, are compiled into one single catalogue of source candidates.

The internal processing of each model remains the same as described in Chapter 5 and Fig. 2.6. The only difference can be found in the data collection. In particular, the cross-match of the CW-detected sources with the radio detections (for the training stages, from LoTSS-DR1). Instead of using a search radius of 1.″1 (as it is maintained for all the remaining ancillary catalogues), sources are cross-matched with a radius of 6″ with the radio catalogue. This increase of more than five times in distance (and close to 30 times in area) can be explained by the need of obtaining a larger fraction of sources with a radio counterpart (as the results of the cross-match itself are to be assessed in Sect. 5.3). A modified version of Table 2.3, with the new search radius for radio counterparts, is shown in Table C.1.

As expected, the number of non-radio counterparts remains the same. In contrast, the number of radio cross-matches has grown by more than a 100 % from the use of a 1.″1 search radius. This change can, in turn, alter the metrics for the radio detection models and those of the redshift predictions as well given the modification of the distribution of values of the feature

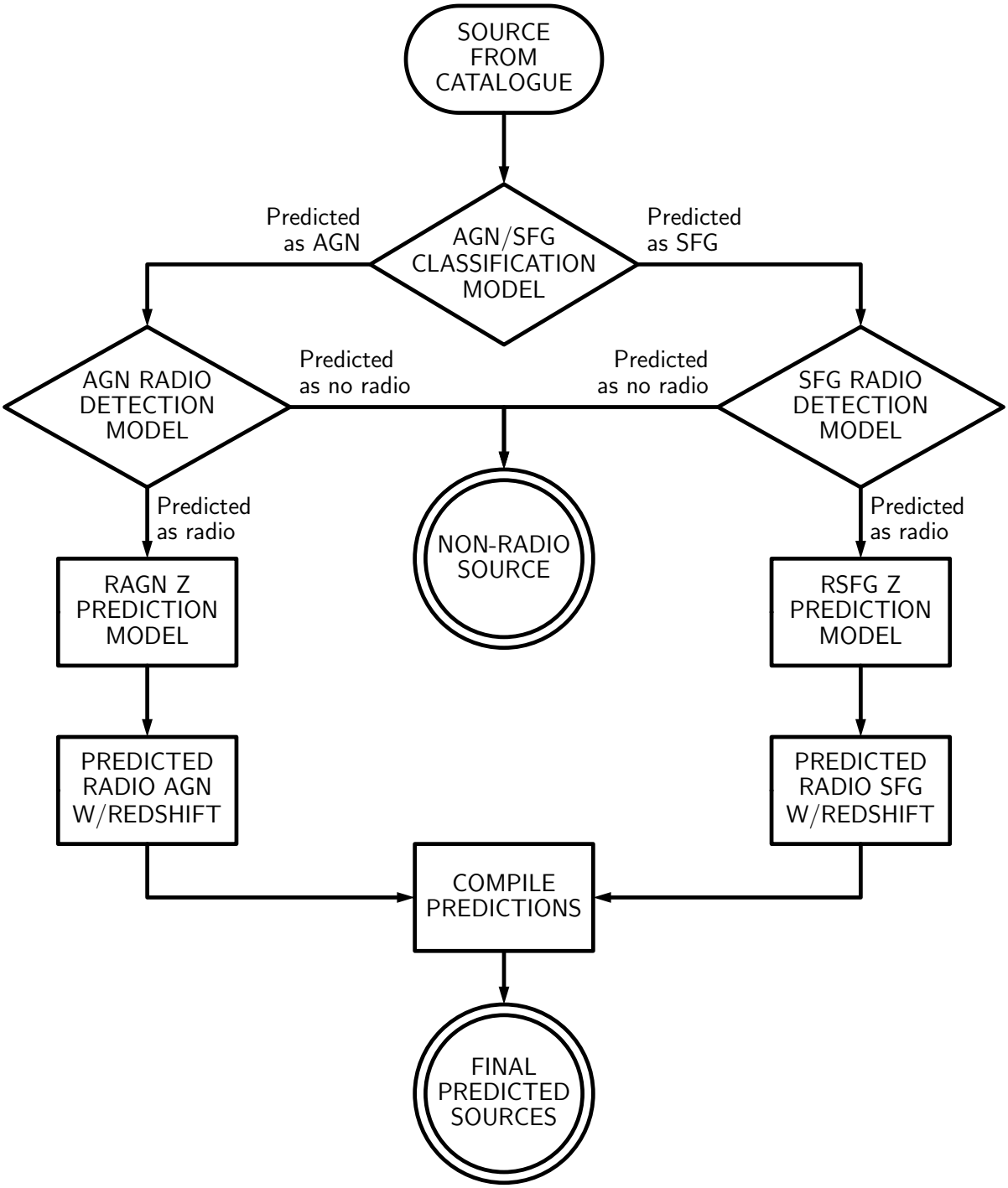


Figure C.1: Flowchart representing the proposed extended prediction pipeline used to predict the presence of radio-detected AGN and SFGs and their redshift values from IR-detected sources. Diamonds represent classification models and rectangles, regression model and intermediate data-collection steps. Double circles represent end states for the data in the pipeline.

Table C.1: Composition of initial catalogue and number of cross matches with additional surveys and catalogues for training of updated pipeline

Step	Survey	HETDEX	Stripe82
Base catalogue	CatWISE2020	15 136 878	3 590 306
Photometry cross-match	AllWISE	5 955 123	1 424 576
	Pan-STARRS	4 837 580	1 346 915
	2MASS	566 273	214 445
	LoTSS (6″)	382 431	...
	VLAS82 (6″)	...	17 706
Source identification	MQC (AGN)	50 538	17 743
	SDSS (SFGs)	68 196	4085

`radio_detect`. For our purposes, the differences in the metrics for this new set of models do not imply a degradation of the results they provide given that they have been corrected for these differences when needed.

C.1 Training and model selection

Results and metrics of the AGN-SFG classification model have not changed with the extension of the prediction pipeline and the inclusion of a larger number of radio counterparts. This prediction stage does not use any form of radio information for its training and predictions. We refer, then, the reader to Chapter 4 and Sect. 3.7.1 for the analysis of its results for the prediction of AGN.

The model for the classification of radio-detectable AGN is modified by the change in the distribution of values in its target feature, `radio_detect`. Nevertheless, the features selected for training remain the same as with the original model (i.e. `band_num`, `W4mag`, `g_r`, `g_i`, `r_i`, `r_z`, `i_z`, `z_y`, `z_W1`, `y_J`, `y_W1`, `J_H`, `H_K`, `K_W3`, `K_W4`, `W1_W2`, and `W2_W3`). For the selection of the meta and base models, `XGBoost` was adopted, since it obtained the best rank as the meta learner, while `GBC`, `RF`, `CatBoost`, and `ET` have been used as base learners. The justification for such selection is reported in Table C.2.

The training of the model for the prediction of radio detectability in predicted SFGs led to the use of 18 features (`W4mag`, `Kmag`, `g_r`, `g_W2`, `r_i`, `r_y`, `i_z`, `i_y`, `z_y`, `z_W2`, `y_J`, `y_W2`, `J_H`, `H_K`, `H_W3`, `W1_W2`, `W1_W3`, and `W3_W4`) together with its target, `radio_detect`. Additionally, and as seen in Table C.3, `RF` was selected as meta learner and `CatBoost`, `XGBoost`, `ET`, and

Table C.2: Performance rating for modified base models for the radio detection classification of AGN

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
GBC	36.98 ± 2.20	36.39 ± 2.26	66.47 ± 3.39	27.10 ± 1.99	1.25
RF	36.74 ± 1.99	35.90 ± 1.90	65.48 ± 1.91	26.98 ± 1.78	2.25
XGBoost	36.41 ± 1.82	34.47 ± 2.16	61.93 ± 3.18	27.18 ± 1.53	3.25
CatBoost	36.18 ± 1.64	35.21 ± 1.68	64.60 ± 2.46	26.55 ± 1.48	3.50
ET	35.08 ± 1.09	34.32 ± 1.48	64.22 ± 2.48	25.52 ± 0.83	4.75
No-skill	15.29 ± 0.76	0.00 ± 0.91	15.29 ± 0.76	15.29 ± 0.76	6.00

[a] Values and uncertainties as in Table 3.1.

GBC, as base models.

Table C.3: Performance rating for modified base models for the radio detection classification of SFGs

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
RF	39.39 ± 2.49	41.69 ± 2.36	76.09 ± 2.42	28.19 ± 2.12	1.25
CatBoost	39.24 ± 2.36	41.53 ± 2.39	75.90 ± 2.69	28.06 ± 1.97	2.25
GBC	37.69 ± 2.32	40.83 ± 2.40	77.39 ± 3.01	26.49 ± 1.89	3.50
XGBoost	38.64 ± 1.98	40.38 ± 2.23	73.55 ± 2.87	27.76 ± 1.57	4.00
ET	38.35 ± 2.57	40.69 ± 2.50	75.28 ± 2.78	27.31 ± 2.17	4.00
No-skill	14.22 ± 1.13	0.00 ± 1.32	14.22 ± 1.14	14.22 ± 1.12	6.00

[a] Values and uncertainties as in Table 3.1.

The model for the prediction of photometric redshifts of radio-detectable AGN is also modified but, in this case, by the change in number of sources in the training set, which is increased (by the larger search radius for radio-detected sources). Thus, the features selected for training are 17 (i.e. band_num, W4mag, g_r, g_W1, r_i, r_z, i_z, i_y, z_y, y_J, y_W1, J_H, H_K, K_W3, K_W4, W1_W2, and W1_W3). Despite three algorithms having the same mean rank, the algorithm selected to be meta learner is RF given that it presents the best value of σ_{NMAD}, which is the metric to be optimised during training. Its metrics (together with those from the base models) are reported in Table C.4.

The model for the prediction of photometric redshifts of radio-detectable SFGs is also modified and selects 14 features for its training: W4mag, g_r, r_i, r_z, i_z, i_y, z_y, y_J, y_W2, J_H, H_K, K_W3, W1_W2, and W1_W3. The selected algorithm to be meta learner is ET (again, because it presents the best value of σ_{NMAD}) and its metrics (together with those from the base models and an no-skill prediction) are presented in Table C.5.

For the sake of completeness, Table C.6 presents the optimised hyperparameters for all

Table C.4: Performance rating for modified base models for redshift value prediction on predicted radio detectable AGN

Model	$\sigma_{\rm MAD}$ ($\times 100$)	$\sigma_{\rm NMAD}$ ($\times 100$)	σ_z ($\times 100$)	$\sigma_z^{\rm N}$ ($\times 100$)	η ($\times 100$)	Rank
RF	17.54 ± 1.24	7.70 ± 0.34	41.55 ± 4.38	19.63 ± 2.57	18.30 ± 1.60	2.0
ET	18.79 ± 1.22	8.32 ± 0.39	40.81 ± 3.25	18.29 ± 1.86	19.56 ± 1.78	2.0
CatBoost	21.07 ± 1.15	9.97 ± 0.25	39.65 ± 2.52	18.06 ± 1.57	21.14 ± 2.17	2.2
XGBoost	22.92 ± 1.20	10.49 ± 0.54	42.16 ± 3.85	19.36 ± 2.16	23.53 ± 1.62	3.8
GBR	28.33 ± 1.48	13.24 ± 0.73	44.77 ± 3.59	20.21 ± 1.93	29.90 ± 1.82	5.0
No-skill	97.06 ± 4.52	39.95 ± 1.89	86.78 ± 1.92	48.17 ± 1.10	72.49 ± 1.96	6.0

[a] Algorithms sorted by increasing $\sigma_{\rm MAD}$ values.
[b] Uncertainties as in Table 3.1.

Table C.5: Performance rating for modified base models for redshift value prediction on predicted radio detectable SFGs

Model	$\sigma_{\rm MAD}$ ($\times 100$)	$\sigma_{\rm NMAD}$ ($\times 100$)	σ_z ($\times 100$)	$\sigma_z^{\rm N}$ ($\times 100$)	η ($\times 100$)	Rank
ET	3.85 ± 0.25	2.81 ± 0.18	9.89 ± 0.81	7.03 ± 1.06	2.91 ± 0.77	2.0
RF	3.89 ± 0.12	2.86 ± 0.12	9.78 ± 0.88	6.98 ± 1.10	3.08 ± 0.88	2.0
CatBoost	4.01 ± 0.20	2.96 ± 0.12	9.75 ± 0.72	6.99 ± 1.01	2.78 ± 0.56	2.0
XGBoost	4.31 ± 0.27	3.16 ± 0.17	9.98 ± 0.90	7.05 ± 1.05	3.13 ± 0.73	4.2
GBR	4.81 ± 0.16	3.54 ± 0.08	9.96 ± 0.76	7.05 ± 0.99	3.45 ± 1.05	4.6
No-skill	33.51 ± 1.86	21.70 ± 1.31	26.74 ± 0.67	20.94 ± 0.42	49.08 ± 1.93	6.0

[a] Algorithms sorted by increasing $\sigma_{\rm MAD}$ values.
[b] Uncertainties as in Table 3.1.

the five stacked models of the new prediction pipeline.

Table C.6: Hyperparameters values for meta-learners in modified pipeline after tuning.

AGN-SFG model (CatBoost)			
Parameter	Value	Parameter	Value
learning_rate	0.0075	random_strength	0.1
depth	6	l2_leaf_reg	10
Radio detection model for AGN (GradientBoosting)			
Parameter	Value	Parameter	Value
n_estimators	187	min_samples_leaf	2
learning_rate	0.0560	max_depth	9
subsample	0.3387	max_features	0.5248
min_samples_split	5		
Radio detection model for SFGs (RF)			
Parameter	Value	Parameter	Value
n_estimators	17	max_depth	6
min_impurity_decrease	0.0000	max_features	0.4280
bootstrap	False	criterion	gini
class_weight	balanced_subsample	min_samples_split	10
min_samples_leaf	3		
Redshift prediction model for rAGN (RF)			
Parameter	Value	Parameter	Value
n_estimators	187	max_depth	9
min_impurity_decrease	0.0000	max_features	0.6346
bootstrap	False	criterion	mae
min_samples_split	3	min_samples_leaf	5
Redshift prediction model for rGal (ET)			
Parameter	Value	Parameter	Value
n_estimators	100	criterion	mse
max_depth	None	min_impurity_decrease	0.0000
max_features	auto	bootstrap	False
min_samples_split	2	min_samples_leaf	1

[a] This table shows the parameters which were subject to tuning.
[b] Remaining hyperparameters used their default values as defined by their developers.

C.2 Application of stacked models

As stated in Sect. C.1, the model for the classification between AGN and SFGs has not suffered any change from the modifications of the new prediction pipeline. However, if the focus of the analyses is changed towards the prediction of SFGs, new metrics can be obtained from

the same model. These values are presented in Table C.7.

Table C.7: Resulting metrics of SFG-AGN classification model for the test subset and the labelled sources in S82 using two different threshold values. Opposite to Table 3.5, the results of this table are focused on the prediction of SFGs and are only shown for the PR-based thresholds.

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	PR	96.43 ± 0.33	91.85 ± 0.70	97.15 ± 0.31	95.84 ± 0.52
S82-label	PR	76.36 ± 1.37	70.67 ± 1.71	74.95 ± 1.90	77.60 ± 1.83

[a] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 3.5)

After the hyperparameter optimisation and probability calibration, new metrics were obtained for both radio detection models in the test subset and in the labelled sources in the S82 field. These values are presented, jointly, in Tables C.11 and C.11 for sources in the test subset and in the S82 labelled sources, respectively.

Table C.8: Resulting metrics of radio detection prediction model for the AGN and SFG branches of our modified pipeline. Models have been applied to the test subset and the labelled sources in S82 using the PR threshold values.

Branch	Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
AGN	HETDEX-test	PR	46.46 ± 3.25	34.86 ± 3.90	43.37 ± 3.18	49.46 ± 3.95
	S82-label	PR	23.46 ± 1.84	20.92 ± 2.51	13.55 ± 1.09	59.43 ± 4.59
	HETDEX-pipe	PR	44.82 ± 2.97	33.85 ± 3.67	42.34 ± 3.27	47.19 ± 3.28
	S82-pipe	PR	22.22 ± 1.99	20.25 ± 2.53	12.81 ± 1.24	56.63 ± 3.73
SFG	HETDEX-test	PR	45.88 ± 2.54	36.67 ± 3.08	46.23 ± 3.21	45.66 ± 2.62
	S82-label	PR	9.86 ± 5.64	9.90 ± 8.21	5.28 ± 3.00	35.00 ± 20.81
	HETDEX-pipe	PR	46.22 ± 3.11	36.73 ± 3.89	47.17 ± 3.94	45.51 ± 2.71
	S82-pipe	PR	23.17 ± 3.50	20.96 ± 4.68	13.72 ± 2.00	54.12 ± 9.71

[a] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 3.5)

If, then, both classification steps, the classification between AGN and SFGs, and the radio detection prediction for both kinds of sources, are joined into one single step, their metrics can be obtained. Those metrics are presented in Table C.9 for the sources in the test subset and in the labelled sources in the S82 field.

Metrics from the application of the redshift prediction models to sources in our test subset and the labelled sources in the S82 field are presented in Table C.10.

As in the main text, we also provide a measure of the base status of the data as to assess the improvement given by our modified pipeline in the prediction of sources. In particular,

Table C.9: Resulting metrics of joint classification prediction models for the selection of radio-AGN and radio-SFG. Models have been applied to the test subset and the labelled sources in S82 using the PR threshold values.

Branch	Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
rAGN	HETDEX-test	PR	46.22 ± 2.85	36.73 ± 3.13	47.17 ± 3.22	45.51 ± 2.96
	S82-label	PR	21.35 ± 3.32	18.93 ± 4.18	12.71 ± 2.07	28.04 ± 6.99
rSFG	HETDEX-test	PR	42.24 ± 2.19	36.94 ± 2.32	42.28 ± 2.16	42.26 ± 2.77
	S82-label	PR	5.57 ± 1.96	8.01 ± 2.93	2.84 ± 1.02	28.04 ± 8.30

[a] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 3.5)

Table C.10: Redshift prediction metrics for the test subset from HETDEX and S82 labelled sources as discussed in Sect. 3.7.4

Branch	Subset	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^N (×100)	η (×100)
AGN	HETDEX-test	14.78 ± 2.09	6.65 ± 0.70	38.63 ± 4.54	17.97 ± 2.90	16.58 ± 2.72
	S82-label	18.17 ± 2.73	8.81 ± 0.72	51.01 ± 3.41	22.98 ± 3.36	21.99 ± 1.88
	HETDEX-pipe	14.56 ± 2.34	6.65 ± 0.65	36.85 ± 5.90	23.04 ± 6.00	17.00 ± 3.08
	S82-pipe	17.92 ± 1.56	8.57 ± 0.72	43.59 ± 3.41	25.34 ± 3.36	21.77 ± 1.88
SFG	HETDEX-test	3.90 ± 0.34	2.93 ± 0.25	10.38 ± 1.06	7.01 ± 1.03	2.79 ± 0.97
	S82-label	6.77 ± 2.81	4.41 ± 1.61	14.27 ± 9.19	9.75 ± 7.97	5.54 ± 6.80
	HETDEX-pipe	3.55 ± 0.34	2.85 ± 0.30	14.35 ± 7.55	7.57 ± 1.61	2.79 ± 1.07
	S82-pipe	6.59 ± 1.32	4.85 ± 0.75	35.67 ± 16.66	11.91 ± 2.56	8.77 ± 2.73

[a] Values and uncertainties as in Table 3.5.

Table presents the no-skill metrics for the three classification steps, as defined in Sect. 3.1.1.

Table C.11: Results of no-skill selection of sources in different stages of pipeline to the labelled sources in the HETDEX test subset

Branch	Prediction (×100)	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
AGN	AGN	42.57	0.00	42.57	42.57
	Radio-label	16.46	0.00	16.46	16.46
	Radio-pipe	15.51	0.00	15.51	15.51
	Radio AGN	14.64	0.00	14.64	14.64
SFG	SFG	57.43	0.00	57.43	57.43
	Radio-label	14.69	0.00	14.69	14.69
	Radio-pipe	15.40	0.00	15.40	15.40
	Radio SFG	14.23	0.00	14.23	14.23

Table C.12: Results of no-skill selection of sources in different stages of pipeline to the labelled sources in the S82 labelled sources

Branch	Prediction (×100)	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
AGN	AGN	81.29	0.00	81.29	81.29
	Radio-label	4.92	0.00	4.92	4.92
	Radio-pipe	4.32	0.00	4.32	4.32
	Radio AGN	4.29	0.00	4.29	4.29
SFG	SFG	18.71	0.00	18.71	18.71
	Radio-label	1.74	0.00	1.74	1.74
	Radio-pipe	4.32	0.00	4.32	4.32
	Radio SFG	1.54	0.00	1.54	1.54

From Table C.9, it is possible to see that there is a relevant difference in the scores between the application of the classification steps in the test subset and the labelled sources from the S82 field. Their difference is even stronger for the classification in the SFG branch of our pipeline. Part of these differences can be explained by analysing the values from Tables C.11 and C.12, and the descriptions from Sect. 2.5.

While the number of AGN and SFGs in the HETDEX is somewhat balanced with a higher fraction of SFGs, the situation in the S82 is completely different. Most of the labelled sources in the S82 field are AGN, which contradicts most recent results on the detection of extragalactic sources (cf. Chapter 1). By virtue of the observational depth that the S82 field has been subject, most of its studies have been focused on high-redshift sources and, consequently, on bright

AGN. These investigations have increased the number of AGN while that of SFG has remained constant.

The differences in the number of detections can be the cause for the low scores in the SFG branch. While our model predicts typical number of SFGs in the S82 field (as given by the metrics in the test subset), there are not enough confirmed sources to match our estimations, decreasing the scores obtained by our models.