LIÈGE université Library

https://lib.uliege.be https://matheo.uliege.be

# Mémoire, COLLÉGIALITÉ

**Auteur :** Bozet, Quentin
**Promoteur(s) :** Sluse, Dominique; Delchambre, Ludovic
**Faculté :** Faculté des Sciences
**Diplôme :** Master en sciences spatiales, à finalité approfondie
**Année académique :** 2023-2024
**URI/URL :** http://hdl.handle.net/2268.2/19851

*Avertissement à l'attention des usagers :*

*Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.*

*Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.*

LIÈGE université
Sciences

# Searching for strongly lensed galaxies in large sky surveys using machine learning techniques

Graduation thesis submitted in partial fulfilment of the requirements for the degree of Master in Space Sciences, Research focus, Faculty of Sciences

**Quentin Bozet**

Director
Dominique Sluse
Service
Multi-wavelength Extragalactic and Galactic Astrophysics
Co-director
Ludovic Delchambre
Service
Group of AstroPhysics and High-Energies

Academic year
2023 - 2024

**Abstract**

In this work, we are going to investigate the possibility of using machine learning algorithms to detect strong gravitational lenses. This work is a follow up of a previous master thesis on the subject [18]. In [18], a significant amount of preprocessing was performed on the dataset, which was taken from [30]. The first steps to build a machine learning model to detect lenses were also taken. Our first task was thus to reorganize and make more readable the code of [18].

Two classifiers were used to build the models: random forests and extremely randomized trees [11]. In order to compare and discriminate between models, we drew ROC curves and compared them thanks to a method developed in [28]. We have shown that machine learning models provide accurate and reliable lens detection. Models using principal components offered the best results. Their advantage also lies in their interpretability.

Finally, we compared the performance of the machine learning model developed during the course of this work to a deep learning model built in [30].

# Contents

**1 Introduction** **1**
1.1 Context and general considerations . . . . . 1
1.2 Historical introduction and overview of detection methods . . . . . 3

**2 Description of the lensing phenomenon** **6**
2.1 Lensing of point-like masses . . . . . 6
2.2 Description of strong lensing . . . . . 9

**3 Origin and treatment of the data** **15**
3.1 Source of the data . . . . . 15
3.2 Preprocessing of the data . . . . . 18

**4 Theoretical framework of the machine learning model** **22**
4.1 General considerations . . . . . 22
4.2 Description of the algorithms . . . . . 23
4.3 Theory of ROC curves . . . . . 25

**5 Description and results of the machine learning model** **32**
5.1 Description of the model . . . . . 32
5.2 ROC curves of different classifiers . . . . . 36
5.3 Analysis of the bootstrapped ROC curves . . . . . 42

**6 Comparison with the deep learning model** **52**
6.1 Presentation of the convolutional neural network model . . . . . 52
6.2 ROC curves of EfficientNet . . . . . 53
6.3 Discussion of the results . . . . . 54

**7 Conclusion** **58**
7.1 Main results and discussion . . . . . 58
7.2 Future perspectives . . . . . 59

**Appendices** **65**

**A Morphological parameters of galaxies** **66**

# Chapter 1

# Introduction

## 1.1 Context and general considerations

Gravitational lensing is the ability of an object to distort the light coming from a distant source. Gravitational lensing can be strong or weak. Strong lensing produces distinct images while weak lensing does not. In the case of strong lensing, the source is then observed in double or more in the sky. Such situation is illustrated in figure 1.1. Strong lensing is used for example to study the distribution of dark matter in galaxies, determine the structure of galaxy clusters and measure the Hubble constant. Weak lensing on the other hand can help measure the correlation function of density fluctuations. More applications of gravitational lensing can be found in [39]. Strong lensing is the subject of this master thesis. The source and the lens can be any type of astronomical objects. Here, we will focus on the case where both the source and the lens are galaxies. As lensing is a rare phenomenon, specific procedures must be implemented in order to detect lenses in the sky.

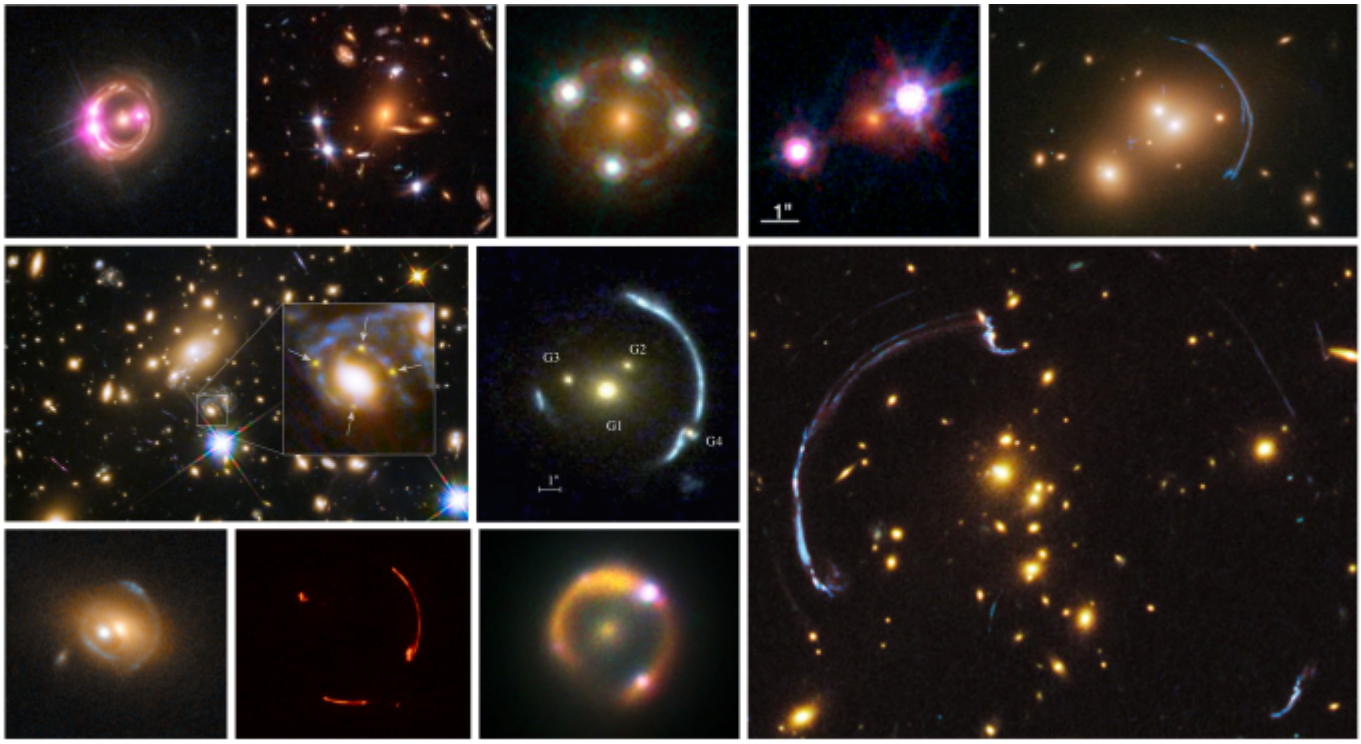


Figure 1.1: Sample of lensed sources. These include quasars, galaxies and supernovae. This image is taken from [20]

Much time has passed since the discovery of the first lensed quasar, Q0957+561, in 1979 by Walsh, Carswell, and Weymann [37]. Indeed, the era of fortuitous discoveries of gravitational lenses is long gone. Nowadays systematic procedures to detect them in increasingly large datasets are transforming the field of lens detection. Among these procedures, we can find artifical neural networks (referred to as neural networks hereafter). In 2021, Savary et al. applied a convolutional neural network ("CNN" [19]) to the detection of strong gravitational lenses. Their results and methodology are reported in [30].

Neural networks are now ubiquitous in the field of lens detection as they are able to produce accurate results. They however suffer from a lack of interpretability and long training times. The aim of this master thesis is to investigate the possibility to use simpler, more interpretable machine learning algorithms in order to detect new strong lenses. A first master thesis [18] was already written on the subject. The present work is therefore a follow up to this first master thesis.

The remaining of the present document is structured as follows. We will first give a historical introduction to the detection of gravitational lenses as well as briefly review

the methods to detect them. Then, in chapter 2, we shall introduce the physics behind the phenomenon of strong gravitational lensing. Afterwards, in chapters 3, 4 and 5, we will present the data and the model developed during the course of our investigation and present its results. In chapter 6, we will introduce the model created in [30] and compare its results with the results of our model. Finally, in chapter 7, we will conclude this work.

## 1.2 Historical introduction and overview of detection methods

This section is a brief review of the history of detection of strong lenses. Most of the information given in this section was based on [20] and partly on [38]. Technically, the first observed gravitational lens is the Sun. Indeed, the prediction made by Albert Einstein in 1916 of the deflection of light rays by the Sun and its observation in 1919 by Arthur Eddington was one of the most striking experimental confirmation of general relativity. The Sun is however not a strong lens. Chwolson in 1924 was the first to raise the possibility of multiple images of a lensed source. Einstein himself, in 1936, made further calculations on the subject but concluded that the effect would not be observable as he considered only a star-star system. Zwicky then realized in 1937 that more massive objects such as galaxy clusters would be more appropriate to detect the strong lensing effect.

As already mentioned, the first detected strongly lensed source was a quasar in 1979. Gravitational lenses were first randomly found during the investigation of other phenomena. The very first lenses were exclusively quasars. Most notably, the Einstein cross, or Q2237+0305, was discovered in 1985 by Huchra et al. [13] . One of the brightest lensed quasars, APM08279+5255, was also found serendipitously. Among these fortuitous discoveries was the lensed quasar with the lowest source redshift, RXJ1131-1231, by Sluse et al. in 2003 [33]. Lensed galaxies were also found in the early days of lens detection such as the galaxies deflected by the galaxy cluster Abell 370. Although automated search algorithms become more prominent, many strong lenses are still found serendipitously.

Gradually, more structured methods began to emerge in order to find new strong gravitational lenses. Among them is the magnification selection of images. Indeed, for strong lenses the magnification theorem holds (see [31] for details): at least one of the duplicated images is as bright as the original source. To discover new lenses it would therefore be reasonable to look among the brightest objects in the sky or to concentrate on objects that have unphysical brightness as these can be the consequences of magnification.

A non-exhaustive list of lenses discovered thanks to this method can be found in [20].

Another way to look for new lenses is to find clearly spatially separated objects, that can be visually distinguished, as these could potentially be the results of the lensing effect. This is called visual inspection. Interestingly, citizen science played an important role in that kind of endeavor. Citizen science is the term that designates the projects where amateur scientists analyze entire catalogues of images looking for lenses. The first project of this kind, Galaxy Zoo [21], was started not to detect lenses but to study the morphology of galaxies. More followed such as The Zooniverse [32] and Planet Hunters [9]. Note that these projects are not limited to astronomy. The first citizen science project that was designed to detect strong lenses is Space Warps [24]. Visual inspection is well-suited for repetitive tasks where pattern recognition is important and where computers can not yet achieve acceptable performances. However, for magnified images where the separation is of the order of the point spread function of the instrument or lower, spatial separation is not so clear and a detailed pixel by pixel analysis of the catalogue must be carried out. A machine learning or deep learning algorithm is appropriate to perform this task

Machine learning has been applied to find gravitational lenses only in recent times. The first conclusive efforts to apply machine learning and deep learning to the systematic search of lenses date back only to late 2017. The advent of large sky surveys and the increase of available computational power made the use of machine learning possible. The most successful deep learning algorithm used to search for new lenses is the supervised convolutional neural network. These were imagined by Yann LeCun [19] and found applications in various fields such as image recognition, colorization of black and white images, medicine and of course physics. They overall lead to excellent results and a detailed account of their performance to detect strong gravitational lenses can be found in [20].

One other possible approach is to use ensemble machine learning algorithms such as decision trees, random forests and extremely randomized trees. The main advantages of these methods are their higher interpretability and lower computing time compared to CNNs. For now, these ensemble methods were mainly applied to the search of lensed quasars, most notably by Khramtsov et al. [15] and Delchambre et al. [8]. It is this approach that we will follow in the present work.

# Chapter 2

# Description of the lensing phenomenon

## 2.1 Lensing of point-like masses

We will now describe the physics of gravitational lensing. The account of the lensing phenomenon given in this chapter follows closely the one given in [39]. Moreover, all the points not discussed in details here can be found in [39].

We are first going to consider a system where both the lens and the source are point-like masses. We work in a coordinate system centered on Earth. Let $\alpha$ be the true angle between the source and the Earth and $\beta$ be the angle between the image of the source and Earth. These two angles are different because the light rays are bent by the lens. The situation is depicted in figure 2.1.

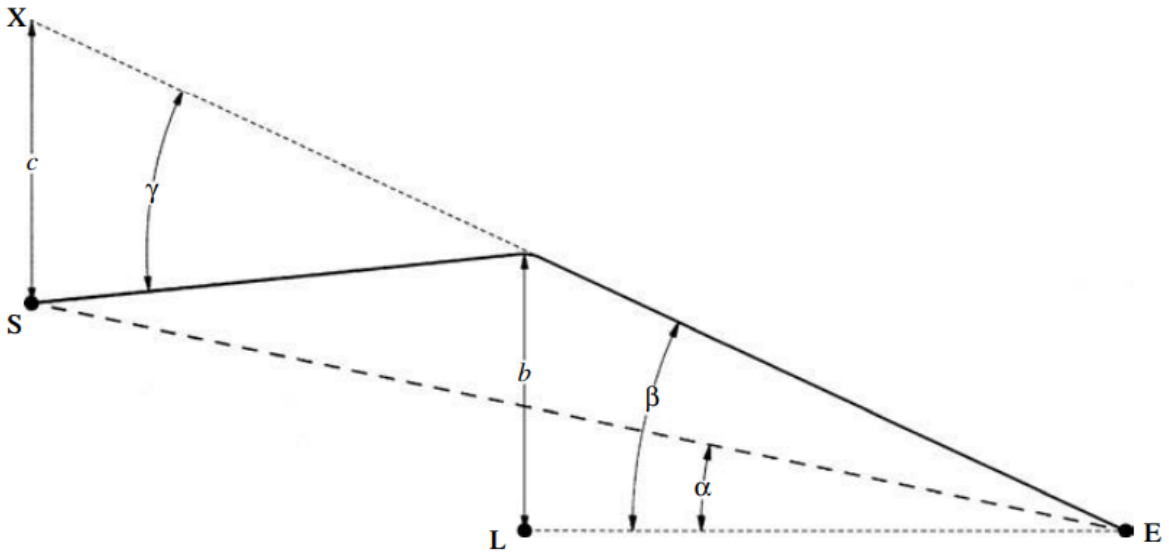


Figure 2.1: Illustration of the different variables used to calculate the lensing effect. This image is taken from [39]

.

We assume that the angle $\beta$ is very small and that the considered objects are not too far apart. So

$$\tan(\beta) \approx \beta \tag{2.1}$$

This implies that

$$b = \beta d_A(\mathrm{EL}) \tag{2.2}$$

where $d_A(\mathrm{EL})$ is the angular diameter distance between the Earth and the lens. This distance accounts for the expansion of the Universe and is given by

$$d_A(\mathrm{EL}) = a(t_\mathrm{L}) r_\mathrm{E}(\mathrm{L}) \tag{2.3}$$

In this equation $a(t_\mathrm{L})$ is the scale factor of the Universe at time $t_\mathrm{L}$ where the light ray emitted by S arrives at L and $r_\mathrm{E}(\mathrm{L})$ is the radial coordinate of L in the Robertson–Walker coordinate system centered on E. From figure 2.1 we can directly deduce that

$$\gamma d_A(\mathrm{LS}) = c = (\beta - \alpha)\, d_A(\mathrm{ES}) \tag{2.4}$$

The angle $\gamma$ can be calculated thanks to general relativity. It is given by

$$\gamma = \frac{4\mathrm{MG}}{b} \tag{2.5}$$

Combining equations 2.2, 2.4 and 2.5, we find

$$(\beta - \alpha)\,\beta = \gamma\,\frac{d_A(\mathrm{LS})}{d_A(\mathrm{ES})}\,\frac{b}{d_A(\mathrm{EL})} = \frac{4\mathrm{MG}d_A(\mathrm{LS})}{d_A(\mathrm{ES})d_A(\mathrm{EL})} \equiv \beta_{\mathrm{E}}^2 \tag{2.6}$$

This last equation is called the lens equation. It admits two solutions, corresponding to the angles of two images in the sky, denoted $\beta_\pm$

$$\beta_\pm = \frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} + \beta_{\mathrm{E}}^2} \tag{2.7}$$

As it is impossible to extract the value of $\alpha$ from the data, we are only able to get an upper bound on the mass of the galaxy via $\beta_{\mathrm{E}}^2$

$$|\beta_+ - \beta_-|^2 \geq 4\beta_{\mathrm{E}}^2 \tag{2.8}$$

If we can measure both $\beta_+$ and $\beta_-$, we can obtain the value of the mass of the galaxy. Indeed, if we multiply the two roots, we obtain

$$\beta_+\beta_- = -\beta_{\mathrm{E}}^2 \tag{2.9}$$

We have thus shown one possible application of gravitational lensing: estimating the masses of galaxies. Note that in the special case where $\alpha = 0$ we observe a continuous ring, called an Einstein ring rather than two separate images. An example of such ring, the Cosmic Horseshoe is given in figure 2.2.

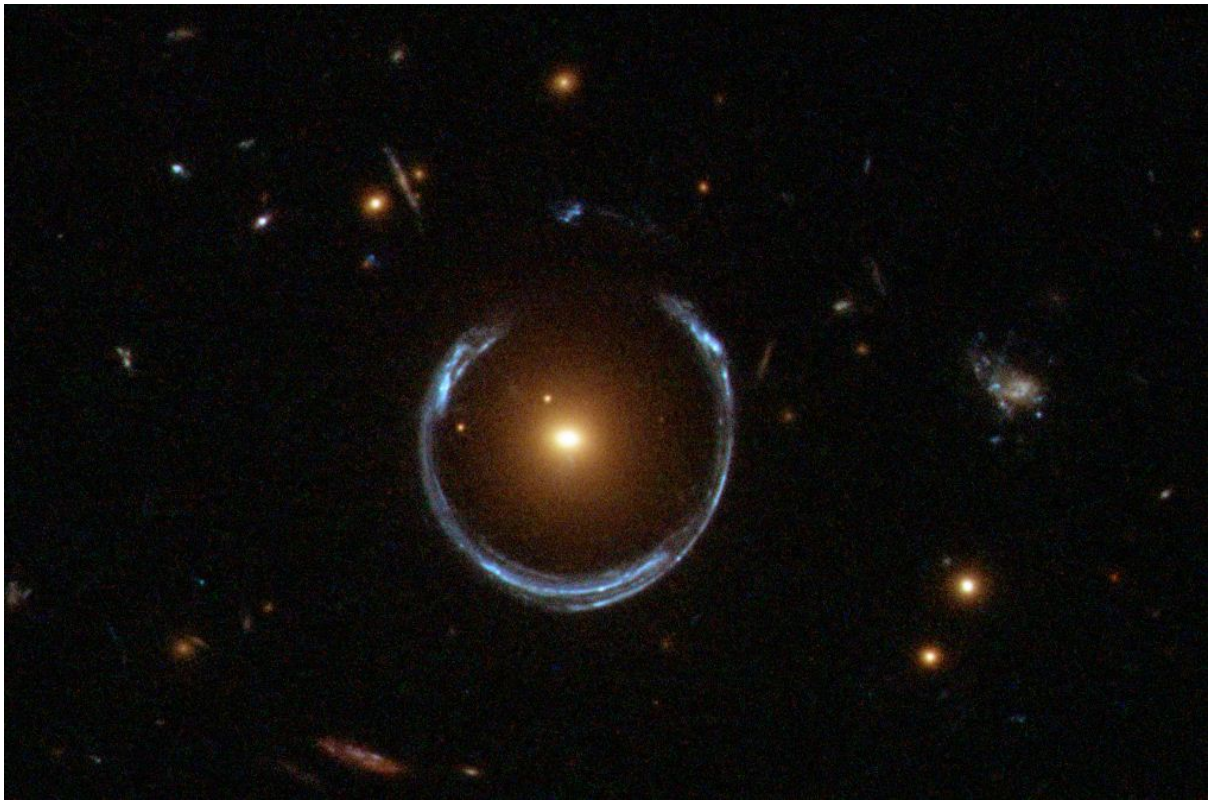

Figure 2.2: The Cosmic Horsehoe is a gravitationally lensed system of two galaxies. This figure is taken from [20]

## 2.2 Description of strong lensing

We will now investigate the magnification effect characteristic of strong lensing. As we will only work with strong lenses, we will refer to strong lenses simply as lenses from now on. In the first place, we can calculate the apparent luminosity of the different images of the source. We will refer to positions in the receiving area relative to some fixed point Y that lies on a line that goes from the source to the lens and continues past Earth. The distance from the observation point, E, to this line is denoted h. The situation is illustrated in figure 2.3.

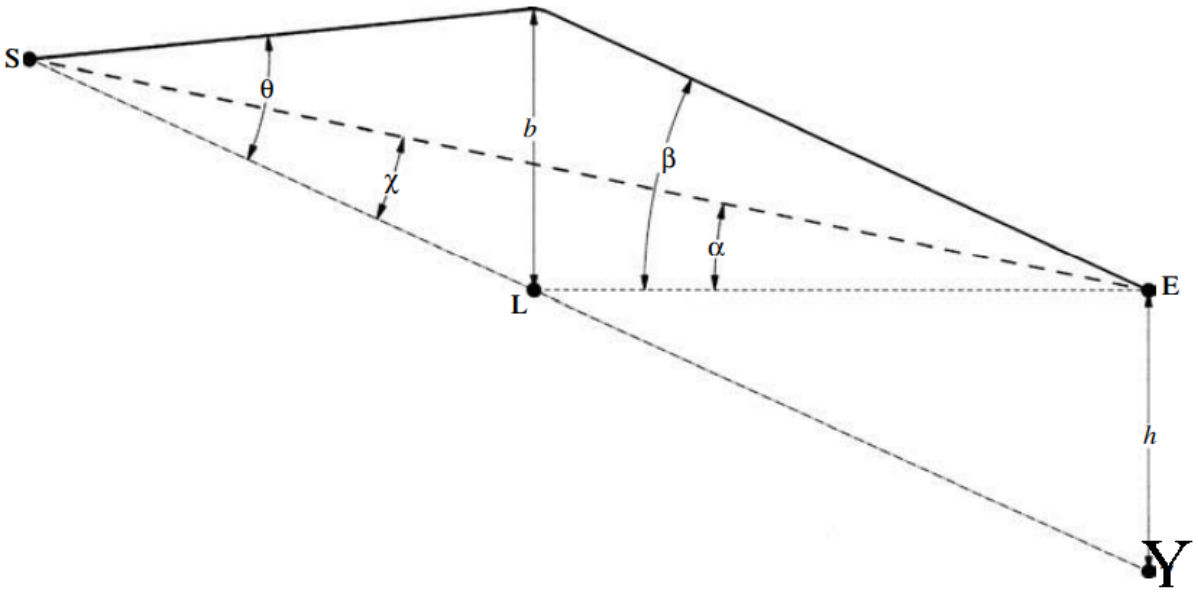


Figure 2.3: Illustration of the different variables used to calculate the magnification effect. This image is taken and corrected from [39]

The apparent luminosity perceived between polar angles $\theta$ and $\theta + \mathrm{d}\theta$, azimuthal angles $\phi$ and $\phi + \mathrm{d}\phi$ and area delimited with height $\mathrm{d}h$ and width $h\mathrm{d}\phi$ is

$$l = \left| \frac{\mathrm{L}\ \theta\ \mathrm{d}\theta\ \mathrm{d}\phi/4\pi}{h\ \mathrm{d}h\ \mathrm{d}\phi\ (1+z_S)^2} \right| = \frac{\mathrm{L}}{4\pi\ (1+z_S)^2} \left| \frac{\theta\ \mathrm{d}\theta}{h\ \mathrm{d}h} \right| \tag{2.10}$$

where L is the luminosity of the source and $z_S$ is the redshift of the source. We deduce the following equality from figure 2.3

$$h = d_A(\mathrm{SE})\chi = d_A(\mathrm{SE})d_A(\mathrm{EL})\alpha/d_A(\mathrm{SL}) \tag{2.11}$$

Moreover, we obtain also by inspecting the same figure

$$\theta = \frac{b}{d_A(\mathrm{SL})} = \beta \frac{d_A(\mathrm{EL})}{d_A(\mathrm{SL})} \tag{2.12}$$

Taking these considerations into account, we can write equation 2.10 as

$$l = l_0 \left| \frac{\beta \mathrm{d}\beta}{\alpha \mathrm{d}\alpha} \right| \tag{2.13}$$

where $l_0$ is the apparent luminosity that would be observed if there was no lensing effect

$$l_0 = \frac{\mathrm{L}}{4\pi(1+z_S)^2 d_A^2(\mathrm{SE})} \tag{2.14}$$

We will now look into the particular case of a point-like lens. The lens equation 2.6 then takes the form

$$\alpha = \beta - \frac{\beta_E^2}{\beta} \tag{2.15}$$

From this last equation, we get

$$\frac{\alpha}{\beta}\frac{\mathrm{d}\alpha}{\mathrm{d}\beta} = 1 - \frac{\beta_E^4}{\beta^4} \tag{2.16}$$

The perceived luminosity is hence given by

$$l = \frac{l_0}{\left|1 - \frac{\beta_E^4}{\beta^4}\right|} \tag{2.17}$$

If the the distance from the lens to the line joining the source and the observer, noted $d$, is small, then

$$\alpha \ll \beta_E \tag{2.18}$$

Thus, the two roots of the lens equation become

$$\beta_\pm = \pm\beta_E + \frac{\alpha}{2} \tag{2.19}$$

We can see by putting this result in equation 2.17 that the images are magnified by a factor $|\beta_E/2\alpha|$. This is the expression of the magnification effect due to strong lensing. The point-like mass approximation for the lens can be used to describe microlensing phenomena. Microlensing is widely used to detect exoplanets and it is the only way to detect wandering exoplanets.

Note that if $\alpha \gg \beta_E$ that is, if $d$ is large, then

$$\beta_+ = \alpha \tag{2.20}$$

and

$$-\beta_- = \frac{\beta_E^2}{\alpha} \tag{2.21}$$

Since $\beta_+ \gg \beta_E$ and $-\beta_- \ll \beta_E$, the image corresponding to $\beta_-$ is not observable while the image corresponding to $\beta_+$ is visible and its brightness is not affected by gravitational lensing. Strong lensing can only occur if

$$\alpha \leq \beta_E \tag{2.22}$$

This means that the maximum proper distance to the lens, noted $d_{\max}$, is given by

$$d_{\max} = \beta_E d_A(\mathrm{EL}) \tag{2.23}$$

$$= \sqrt{\frac{4\mathrm{MG}d_A(\mathrm{LS})d_A(\mathrm{EL})}{d_A(\mathrm{ES})}} \tag{2.24}$$

$$= \sqrt{\frac{4\mathrm{MG}r_{\mathrm{L}}(\mathrm{S})r_{\mathrm{E}}(\mathrm{L})a(t_{\mathrm{L}})}{r_{\mathrm{E}}(\mathrm{S})}} \tag{2.25}$$

$$= \sqrt{\frac{4\mathrm{MG}r_{\mathrm{L}}(\mathrm{S})r_{\mathrm{E}}(\mathrm{L})a(t_{\mathrm{E}})}{r_{\mathrm{E}}(\mathrm{S})}}\frac{1}{\sqrt{1+z_{\mathrm{L}}}} \tag{2.26}$$

where $t_{\mathrm{E}}$ is the time the light ray arrives on Earth and $z_{\mathrm{L}}$ is the redshift of the lens. We have the following identity

$$1 + z_{\mathrm{L}} = \frac{a(t_{\mathrm{E}})}{a(t_{\mathrm{L}})} \tag{2.27}$$

In this last equation, $t_{\mathrm{L}}$ is the time needed for the light ray to reach the lens from the source. From the fact that the proper distance between the source and the lens is equal in both the coordinate system centered on the lens and centered on the Earth we can deduce the equality

$$\int_0^{r_{\mathrm{L}}(\mathrm{S})} \frac{\mathrm{d}r}{\sqrt{1-\mathrm{K}r^2}} = \int_{r_{\mathrm{E}}(\mathrm{L})}^{r_{\mathrm{E}}(\mathrm{S})} \frac{\mathrm{d}r}{\sqrt{1-\mathrm{K}r^2}} \tag{2.28}$$

where K is the curvature of spacetime, $r_{\mathrm{L}}(\mathrm{S})$ the coordinate of the source in the coordinate system centered on the lens, $r_{\mathrm{E}}(\mathrm{L})$ the coordinate of the lens in the coordinate system centered on the Earth and $r_{\mathrm{E}}(\mathrm{S})$ the coordinate of the source in the coordinate system centered on the Earth. We can then obtain

$$r_{\mathrm{L}}(\mathrm{S}) = r_{\mathrm{E}}(\mathrm{S})\sqrt{1-\mathrm{K}r_{\mathrm{E}}^2(\mathrm{L})} - r_{\mathrm{E}}(\mathrm{L})\sqrt{1-\mathrm{K}r_{\mathrm{E}}^2(\mathrm{S})} \tag{2.29}$$

Combining all the results we got so far, the total number of strongly lensed object is given by the integral

$$N_{\mathrm{S}} = \int_0^{r_{\mathrm{E}}(\mathrm{S})} \frac{\mathrm{d}r_{\mathrm{E}}(\mathrm{L})a(t_{\mathrm{L}})}{\sqrt{1-\mathrm{K}r_{\mathrm{E}}^2(\mathrm{L})}} \int_0^{+\infty} \pi d_{\max}^2 n(t_{\mathrm{L}}, M)\mathrm{d}M \tag{2.30}$$

where $n(t, M)$ is the number density of lenses of mass $M$ at time $t$.

For sources located at small redshifts, the integral 2.30 can be solved exactly and

does not depend sensitively on the cosmological model. However for sources at large redshifts, it can be shown by studying the behaviour of 2.30 that the probability of strong lensing depends sensitively on the cosmological model. Thus the probability of strong lensing can be used to discriminate between cosmological models. To date, $O(10^3)$ galaxy-galaxy lenses were discovered [10], which gives an indication on the low occurence of this phenomenon. Another application of strong lensing is to measure the Hubble constant, $H_0$. More specifically, lensed quasars and supernovae can be used to measure $H_0$. See [4] for more details. However, the most important application of strong lensing is the estimation of the mass distributions of galaxies. Note that we have only discussed the point-like approximation for the lenses. This was done for simplicity. Normally, the galaxies must be treated as an extended distribution of mass. The full discussion about extended lenses can be found in [39].

# Chapter 3

# Origin and treatment of the data

## 3.1 Source of the data

Before presenting the machine learning model, we are going to present the dataset used to train the model. As this master thesis is the follow up of [18], we used the same dataset. The dataset was first exploited to train a CNN by Savary et al. [30].

As explained in [18] the dataset is built from images that are taken from the Canada France Imaging Survey (CFIS in short). It was a legacy survey employing the Canada-France-Hawaii Telescope (CFHT), located on top of Mauna Kea, Hawaii. This telescope is multi-band and works in the optical domain, near infrared and ultraviolet and is part of the Ultraviolet Near Infrared Optical Northern Survey (UNIONS). UNIONS uses multiple telescopes. The characteristics of the telescope are the following

- It is a 3.6m telescope
- Its resolution is 0.187 arcsecond/pixel
- It is equipped with MegaCam, a wide-field optical imaging device constructed from 40 2048x4612 pixels CCD cells

This list is inspired by the features listed in [18]. The aim of the survey was to cover 8000 deg$^2$ of northern sky in the $u$ band and 4800 deg$^2$ in the $r$ band. The images used to train both the CNN of Savary et al. [30] and our model are from the second Data Release of CFIS. The total footprint of CFIS and the footprint of the second Data Release of CFIS can be observed in figure 3.1. Similarly to what is done in [30], we will only use $r$ band images. The images taken from CFIS consist only in luminous red galaxies (or LRGs in short), as their lensing cross section is the largest.

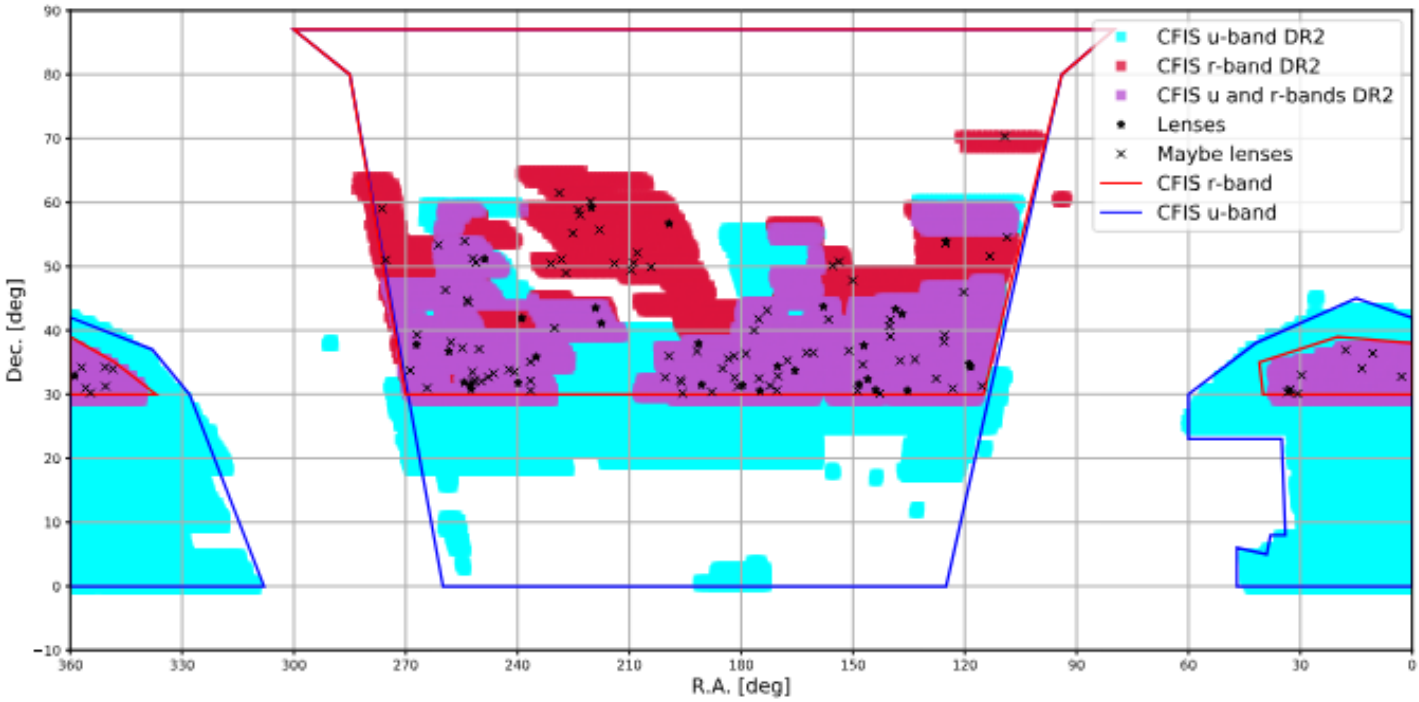


Figure 3.1: The total footprint of the CFIS in the different photometric bands is shown in plain traits while the specific footprint of the second data release is the colored area in the figure. This figure is taken from [30]

As we want to train a model that can detect gravitational lenses, we need a dataset that is divided in two parts: images that are not gravitational lenses (that will be our negative class in our classification model) and images that are gravitational lenses (that will be our positive class in our classification model).

The non lensed images are randomly selected from luminous red galaxies taken from the CFIS. The lensed images are built thanks to a simulation that use a galaxy from the Hubble Space Telescope (HST) as a background source (more precisely from HST/ACS F814W images). A LRG from the CFIS is randomly selected to be the deflector of a background galaxy from the HST images. As a result, we get a lensed source from the HST and a deflector LRG on in the same image. As already mentioned, this dataset was first constructed by Savary et al. and more details on its construction can be found in [30]. Examples of lensed and non-lensed galaxies are shown in figure 3.2 and 3.3.

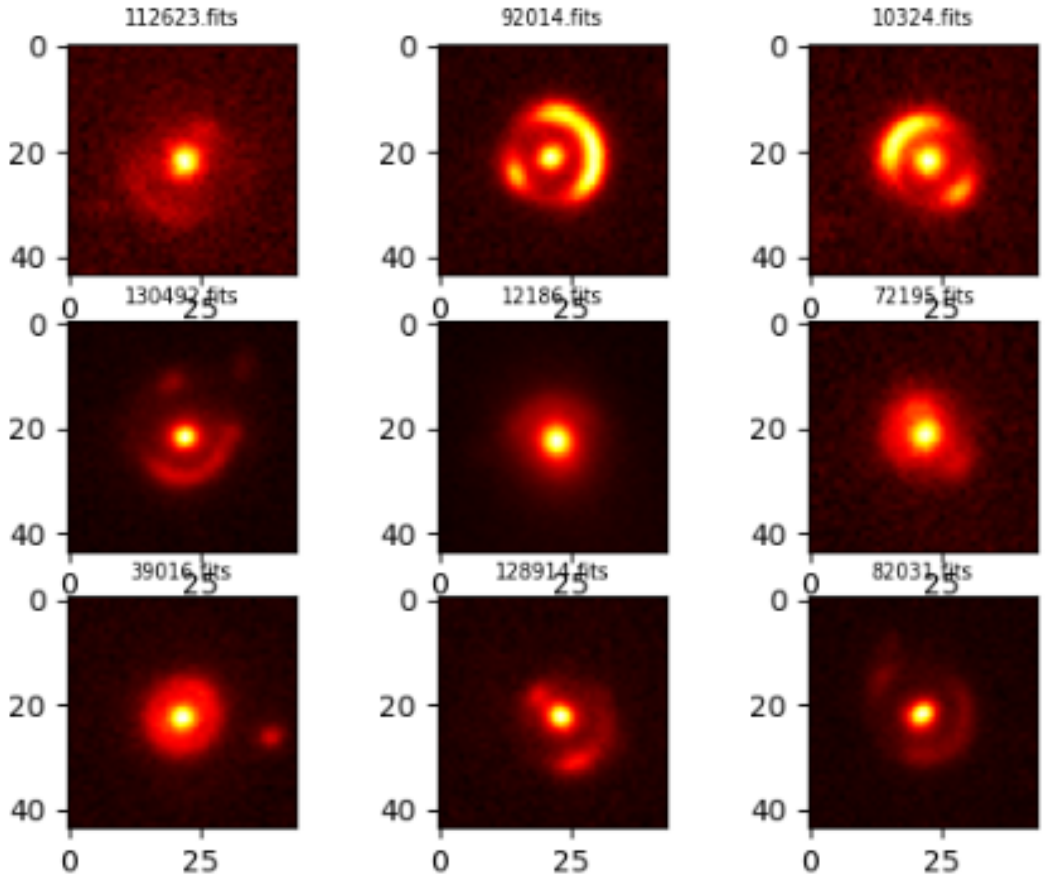


Figure 3.2: Examples of simulated lenses where the source comes from an image taken by the HST and the deflector is a LRG from the CFIS dataset. This figure is taken from [18]

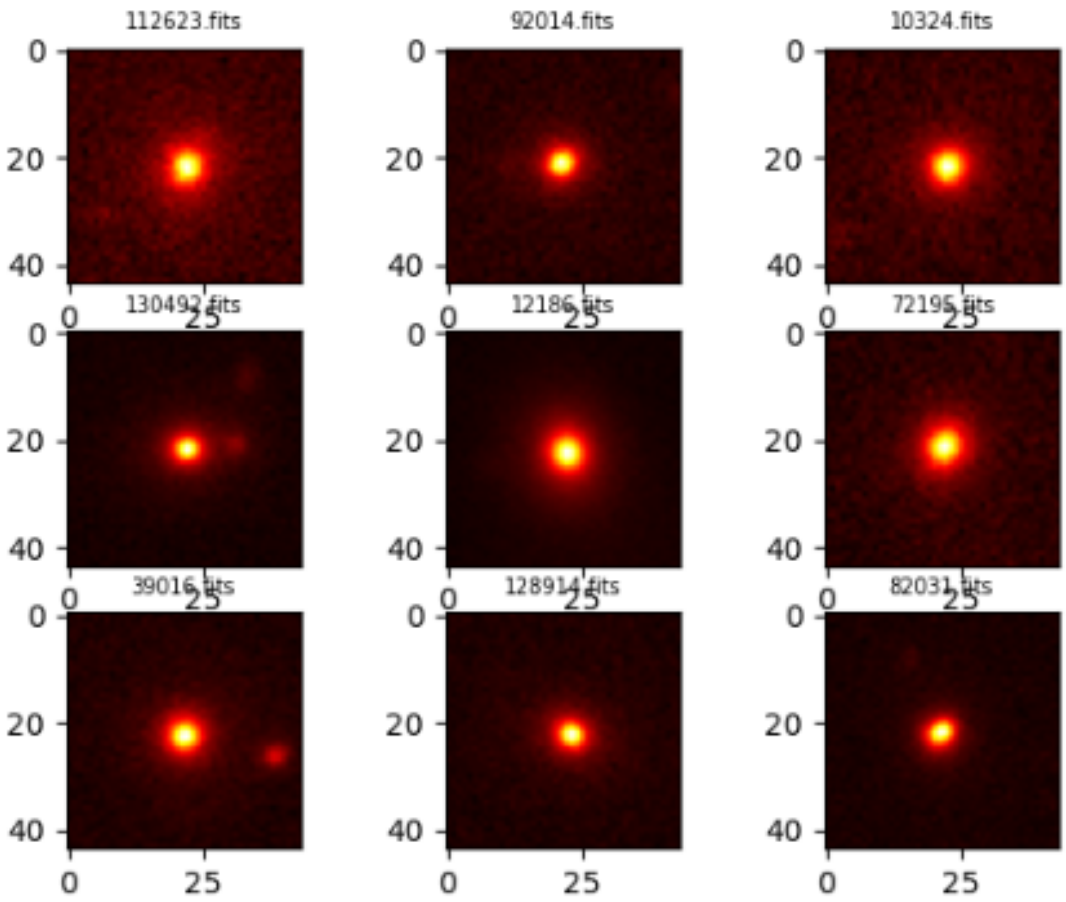


Figure 3.3: Examples of LRGs from the CFIS dataset. This figure is taken from [18]

## 3.2 Preprocessing of the data

Once we have our dataset, a significant amount of preprocessing must still be carried out. This was one of the main task of the master thesis done previously by Clément Laisney [18]. Therefore, we only summarize the main steps of data preprocesssing and refer the reader to [18] for further details. The account given here follows of course closely the one given in [18].

First, a sanity check was performed on the dataset. More specifically, the correctness of the root mean square error (RMS) files provided with the non lensed data was checked. Since the photons coming from distant galaxies are collected thanks to a photodiode, the intensity expressed in analog to digital units, noted $I_{\mathrm{ADU}}$, can be expressed as

$$I_{\mathrm{ADU}} = I_{\mathrm{e\_}} \times g \tag{3.1}$$

where $I_{e\_}$ is the intensity of the electron flux and $g$ is the gain of the photodiode. Since the electrons collected follow a Poisson distribution with variance $\sigma_e^2$, we have

$$I_{e\_} = \sigma_e^2 \tag{3.2}$$

Thus, we obtain

$$\sigma_e^2 = \frac{I_{ADU}}{g} \tag{3.3}$$

The total RMS of the image must therefore be the RMS associated to the object itself plus the RMS of the sky background

$$\mathrm{RMS}^2 = \sigma_{sky}^2 + \sigma_e^2 = \frac{I_{ADU}}{g} + \sigma_{sky}^2 \tag{3.4}$$

Hence, we must have a linear relationship between the $\mathrm{RMS}^2$ of the pixel and the intensity in ADU. This is the case in our dataset and an illustration of this relationship for several non lensed LRGs is provided in figure 3.4. As RMS files were not provided with the lensed data (i.e. the original image with the lensed source), a RMS was computed in [18]. These RMS files were needed in order to develop a procedure to remove the contaminant sources and the sky background.

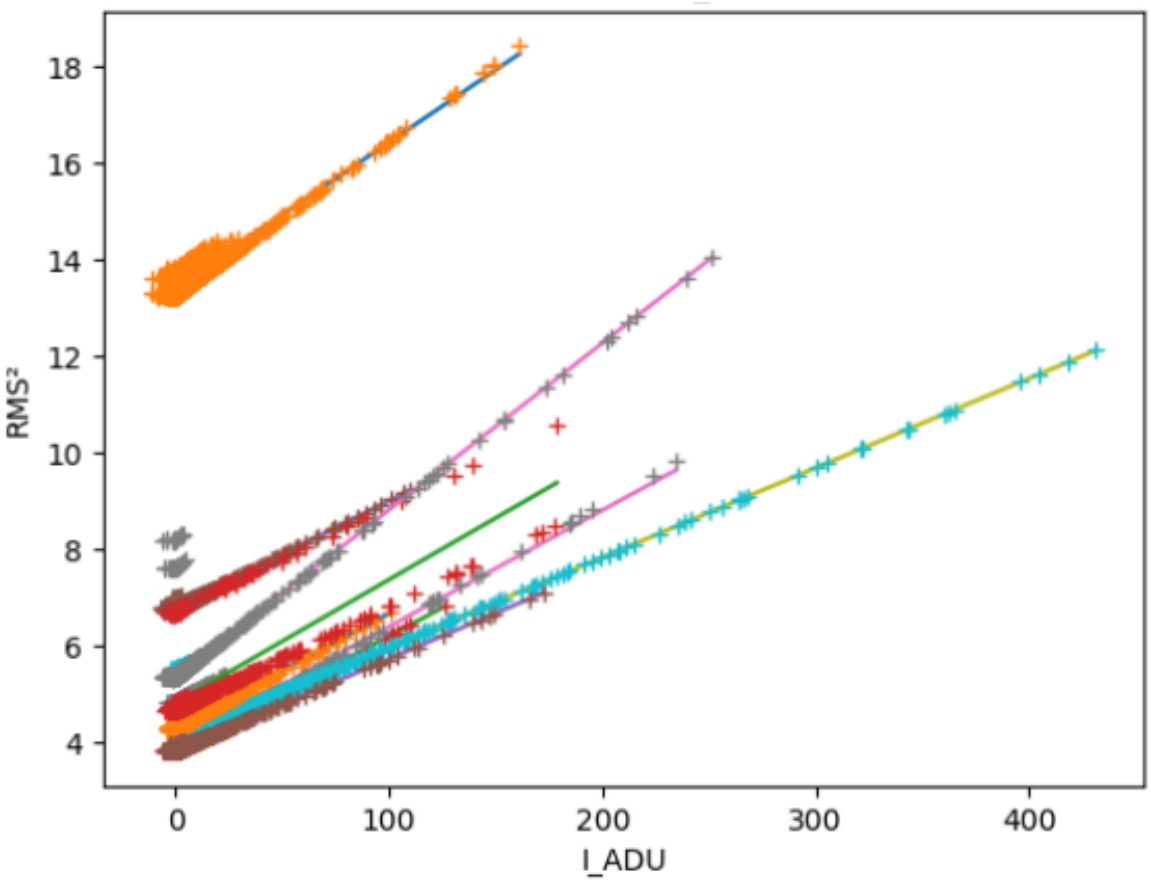


Figure 3.4: RMS² as a function of the intensity in ADU for nine different non lensed LRG images. This figure is taken from [18]

Finally, possible contaminant sources and the sky background had still to be removed, both from lensed and non lensed data. A procedure using Python functions from the photutils package [5] was designed in [18] and applied to the whole dataset.

# Chapter 4

# Theoretical framework of the machine learning model

## 4.1 General considerations

Determining whether a given galaxy is a lens or not is a binary classification problem. We would indeed like to assign to a particular instance a label that can take two possible values (lensed or not). We will do so thanks to a set of measurements $\mathbf{X}$, which are also called features. In order to decide which label to assign to an instance, we will reduce this vector $\mathbf{X}$ to a single scalar score with the help of a function $S(\mathbf{X})$. The function $S$ is called a classifier. In this work, we will use two different classifiers, random forests and extremely randomized trees, to be explained in the next section.

First, we will describe how the dataset is organized. The dataset is split in two parts: the train set and the test set. The model $S(\mathbf{X})$ is trained on the train set, as the name suggests. In the training phase, the free parameters of the model are adjusted by minimizing a loss function. The accuracy and reliability of the model is then measured on the test set. The test set must consist in data on which the model was not trained.

The score is generally comprised between 0 and 1 and is interpreted as the probability of a given instance to belong to the positive class. It is however only an interpretation and it is important to keep in mind that it is not a true probability. This score will be denoted as $o$ in what follows. We will decide that a particular instance belongs to the positive class if its score is greater than a threshold value $t$. It goes without saying that the classification is never perfect and in order to analyze the performance of a given model we introduce the following quantities

- the marginal probability that the instance $s$ belongs to the positive class; $p(\mathbf{p})$
- the marginal probability that the instance $s$ belongs to the negative class; $p(\mathbf{n})$
- the probability that the instance $s$ will be classified as positive while it is negative in reality; $p(o > t|\mathbf{n})$. It is the false positive rate, written $fp$.
- the probability that the instance $s$ will be classified as negative while it is positive in reality; $p(o \leq t|\mathbf{p})$. It is the false negative rate, written $fn$.
- the true positive rate, noted $tp$, is the probability of correct classification for a positive object $s$; $p(o > t|\mathbf{p})$
- the true negative rate, noted $tn$, is the probability of correct classification for a negative object $s$; $p(o \leq t|\mathbf{n})$

The vocabulary and notations of this section are borrowed from [17].

## 4.2 Description of the algorithms

As already mentioned, we will use two types of classifiers in this work: random forests and extremely randomized trees (or ERT in short). We are now going to explain them very briefly.

Both algorithms are built upon a simpler algorithm: decision trees. The construction of a decision tree involves two stages

- first, we split the feature spaces into $M$ distinct and non-overlapping regions. The optimal split can be found by minimizing a loss function.
- if an instance $s$ fits in region $R_i$, then the most common class of this region is assigned to $s$.
- we repeat this procedure for the subregions that we just obtained

The optimal boundaries of the regions can be determined thanks to a least square error minimization procedure. A very simple hypothetical classification tree to search for lenses is given in figure 4.1. Decision trees present a number of advantages, most notably they are readily interpreted. However, in most applications they clearly lack accuracy and the resulting variance of a decision tree is generally high. We therefore have to use other ensemble learning methods such as random forests and ERT.

We will begin by discussing random forests. They are based on bootstrapped aggregation, or bagging in short. We begin by building a large number of decision trees on bootstrapped data samples (i.e. samples drawn with replacement). In a random forest, only a subset of the entire possible features is used in order to create a new branch in the tree. This is done to decorrelate the resulting trees. This deccorelation will of course be partial. When all the trees are built, the score assigned to an object $s$ will be the mean of the scores predicted by the different trees. The explanation of decision trees and random forests given here is based upon the one given in [14] and more details can be found there.

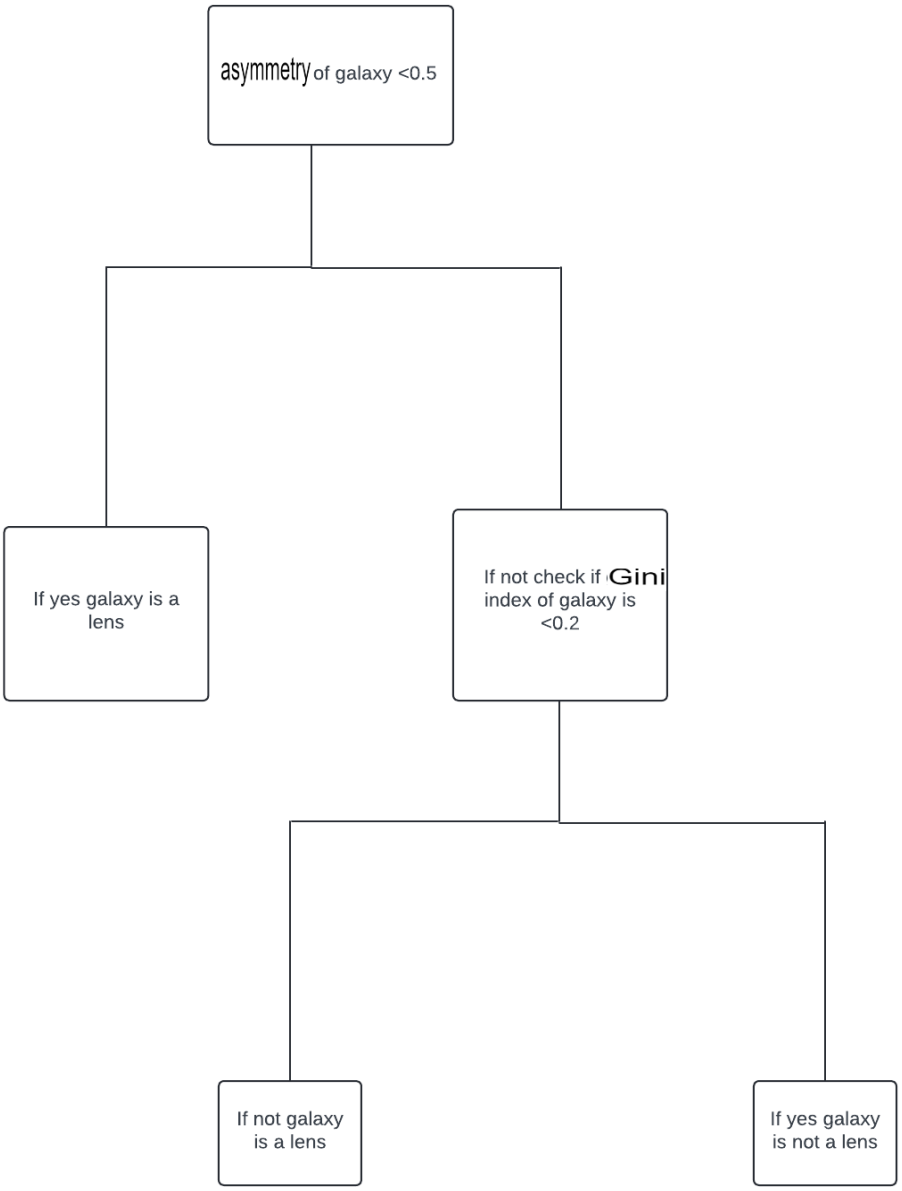


Figure 4.1: Toy decision tree to determine whether a galaxy is lensed or not. Two features (asymmetry and Gini index) are used to build this model.

Extremely randomized trees were first introduced by Geurts et al. in [11]. The rationale behind ERT is the same as the one behind random forests: reduce the variance of classical decision trees. In order to achieve this goal, randomization is implemented while constructing the tree. More specifically, cut points for new leaves are determined randomly at first for a definite number $K$ of attributes. The conserved split out of the $K$ drawn will be the one to maximize the score of its attribute. A large number of these randomized trees is built and the score of a particular instance will be again the mean of the scores given by the different trees. For random forests the random element of the algorithm is brought thanks to the bootstrap, while for ERT the trees are constructed on the non-bootstrapped dataset. Compared to random forests, extremely randomized trees offer, in general, similar performance but take less time to train [11]. This reduction of training time is given by a constant factor [11].

We can also use principal components instead of the attributes. Principal components are linear combinations of the original parameters. The coefficients of the linear combinations are chosen such that the principal components are linearly uncorrelated. The resulting vectors in parameter space are orthogonal and point in the direction of maximal variance not spanned by the other principal components. They are labeled such that the first principal component explains the most variance, the second, the second most variance and so on. The first principal components is noted $Z1$, the second $Z2$ and so forth. This is a major improvement over original parameters as these can potentially be correlated or introduce noise. More details on principal components can be found in [14].

## 4.3 Theory of ROC curves

We will now present the theory behind the concept of ROC curves. The acronym ROC stands for receiver operating characteristic. The origin of ROC curves dates back to World War II, in 1941 more precisely [17]. They were used to help detect enemy objects. Afterwards, they were soon used in the civilian world, most notably in medical science [17]. In modern times, they are increasingly used in machine learning and deep learning in order to assess the performance of different models. It is in this context that we will introduce the theory of ROC curves.

First, let us define a ROC curve. According to [17] a ROC curve is " a graph showing true positive rate on the vertical axis and false positive rate on the horizontal axis, as the classification threshold $t$ varies". We thus have one ROC curve per model $S(\mathbf{X})$. We can deduce general features from this definition.

The worst classifier that we can think of would be a completely random classifier. It would allocate an individual to negative or positive populations with the same probability, such that $p(s|\mathbf{n}) = p(s|\mathbf{p}) = p(s)$. In such classifier we will therefore have $tp = fp$. The resulting ROC curve is thus a diagonal line joining the points (0,0) and (1,1). This line is called the chance diagonal.

Now suppose that we have a perfect classifier. For this classifier there will therefore be at least one value of $t$ where we have $tp = 1$ and $fp = 0$. By definition of the true positive rate $tp = p(o > t|\mathbf{p})$, for smaller values of $t$ we will still have $tp = 1$ while the false positive rate will vary between 0 and 1. Conversely, for larger values of $t$, the false positive rate $fp$ will be 0 (by definition of $fp$, $fp = p(o > t|\mathbf{n})$) whereas the true positive rate will vary from 1 to 0. The ROC curve of this perfect classifier will thus be a straight line that goes from (0,0) to (0,1) succeeded by a line ranging from (0,1) to (1,1). ROC curves of real classifiers will be continuous curves that stand between the perfect ROC curve and the chance diagonal. Such examples are given in figure 4.2.

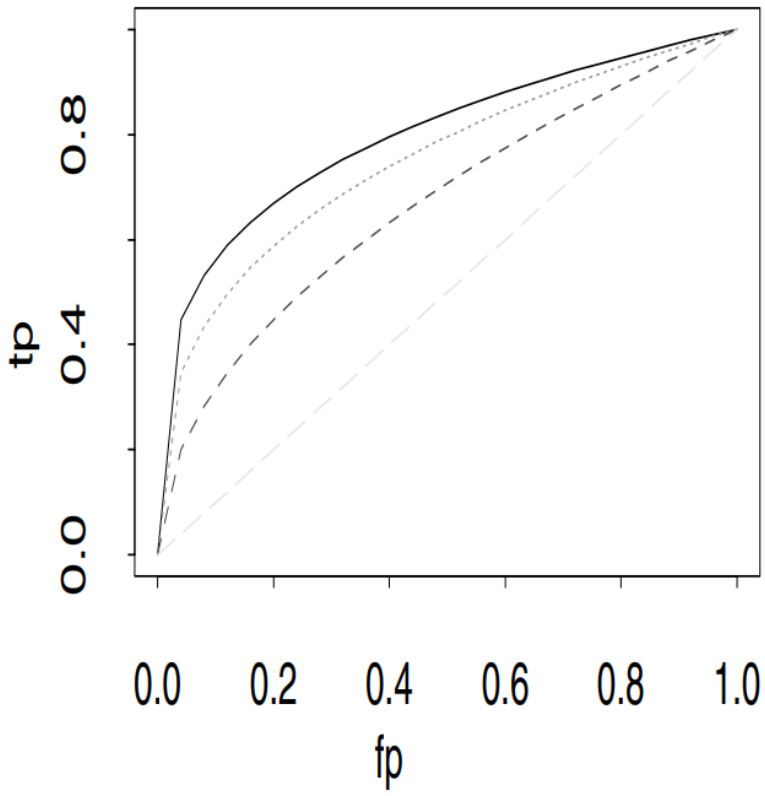


Figure 4.2: Examples of ROC curves. The chance diagonal is also shown in the graph. Figure taken from [17]

In order to assess the performance of different classifiers, one must be able to compare

different ROC curves. One way to do it is to compare a set of summary indices. We will introduce only of them: the area under the ROC curve, or AUC in short. It is defined as

$$\text{AUC} = \int_0^1 y(x)\mathrm{d}x \tag{4.1}$$

where $y(x)$ is the true positive rate as a function of the false positive rate $x$. We can say that it is the function associated to the ROC curve. From the considerations of the previous paragraphs, we can deduce that the AUC for a perfect classifier is 1 whereas it is 0.5 for a random classifier. A simple interpretation of the AUC is that it is the average true positive rate for the whole domain of false positive rates, that spans from 0 to 1. This follows directly from the definition of the average value of a function. We thus expect that better classifiers will have higher AUCs compared to lesser classifiers, as classifiers with a higher AUC give on average a higher true positive rate. Most information in the first part of this section was found in [17] where more summary indices and properties of ROC curves can be found.

We will now describe how ROC curves can be used in the context of machine learning. The approach given thereafter was first pioneered by Fawcett and Provost in [28] and we will follow closely the account given in this defining paper. ROC curves can indeed be used to compare different machine learning models.

A first simple approach to compare machine learning models would be to compare their classification accuracy, defined as the ratio between the number of correctly predicted instances and the total number of predictions, and choose the model with the highest classification accuracy. However, in order for this method to be relevant, we implicitly assume that the target environment will be constant (the conditions in which we obtain the data stay the same over time) and relatively balanced (there are approximately the same numbers of positive and negative instances). This is clearly not the case for the problem studied in this master thesis i.e the search for gravitational lenses. In fact, as lensed galaxies are very rare, the class distribution is very skewed towards the negative class. We call this phenomenon class imbalacing. So let's imagine that the lenses appear in a 999:1 ratio, any model that classifies all the galaxies as non-lensed will have a 99.9 % classification accuracy. It is therefore pointless to compare the classification of different models in this extreme scenario. Classification accuracy does not incorporate the cost of misclassification either.

A robust way with to compare different models would be to compare their ROC

curves as it enables us to see how the different models perform for chosen false positive rates. ROC curves allow us to analyze the performance of the models in a way that is independent of the possible skewness of the class distribution. As mentioned earlier, one way to perform the comparison of the ROC curves is to compare the different AUCs. Due to the average nature of the AUC, this is sufficient only if the ROC curves do not cross each others. Indeed, if the ROC curves cross this means that one model is better than the other on a limited range of false positive rates. As a consequence it is not clear which one to chose. Moreover, there is no mention of the misclassification cost in the AUC.

This is where the ROC convex hull method comes into place. We first notice that the expected cost associated to the point ($fp$, $tp$) in ROC space is given by

$$p(\mathbf{p}) \times (1 - tp) \times c(\mathbf{N}, \mathbf{p}) + p(\mathbf{n}) \times fp \times c(\mathbf{Y}, \mathbf{n}) \tag{4.2}$$

where $c(\mathbf{N}, \mathbf{p})$ is defined as the cost of false positive and $c(\mathbf{Y}, \mathbf{n})$ the cost of false negatives. Thus two models with ($fp_1$, $tp_1$) and ($fp_2$, $tp_2$) will perform equally well if

$$\frac{tp_2 - tp_1}{fp_2 - fp_1} = \frac{c(\mathbf{Y}, \mathbf{n})p(\mathbf{n})}{c(\mathbf{N}, \mathbf{p})p(\mathbf{p})} \tag{4.3}$$

The term on the right side of equation 4.3 defines the slope of so called iso-performance lines. As we want models with maximal performance, optimal points in the ROC space will be located in the upper-left corner of the ROC space. More formally, the best models will be located on the boundary of the convex hull of the different ROC curves in consideration. Hence, models that lie under this boundary can be discarded (for example models B and D in figure 4.3). The situation is depicted in figure 4.3.

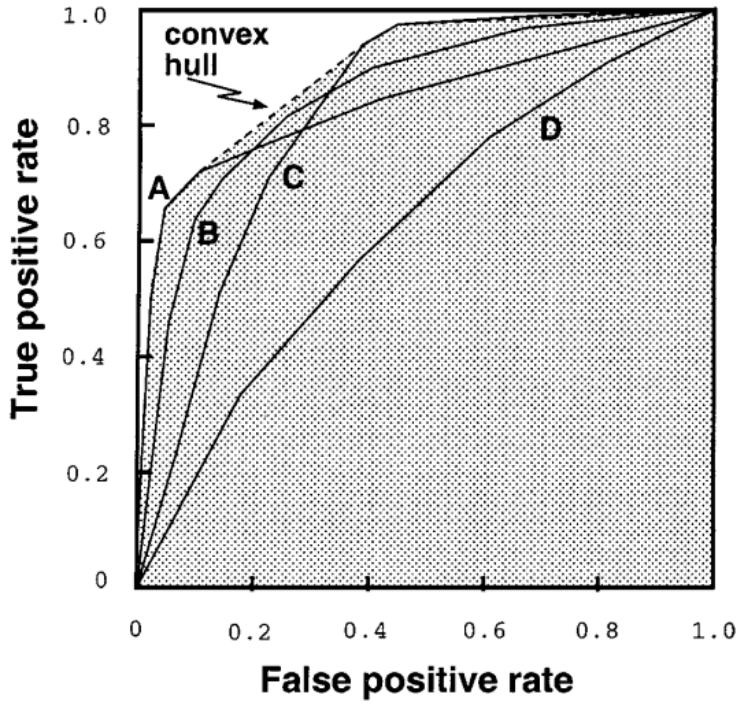


Figure 4.3: Hypothetical ROC curves with their convex hull. Figure taken from [28].

Of course, the optimal model will depend on the specific class distribution and on the objective (this objective could be high purity or high completeness). As an example, let us consider two fictitious scenarios, taken from [28]. In both scenarios, there are five times more negative instances of the negative class than the positive class. In scenario 1, false positive and false negative errors are equally expensive (i.e. $c(\mathbf{N}, \mathbf{p}) = c(\mathbf{Y}, \mathbf{n})$). In scenario 2, a false negative is 25 times more costly than a false positive. From equation 4.3, we deduce the slopes of the iso-performance lines in both scenarios. It is 5 in the first scenario and 1/5 in the second. Lines $\alpha$ and $\beta$ drawn in figure 4.4 are iso-performance lines with respectively slope 5 and 1/5 that intersect with the convex hull of ROC curves A and C. We therefore conclude that model A is optimal in the first scenario while model C is optimal in the second scenario.

In reality, one does not know precisely the costs or true class distribution. A sensitivity analysis over the different ranges of costs and distributions must therefore be carried out. It is the task that we will accomplish for different machine learning models in chapters 5 and 6.

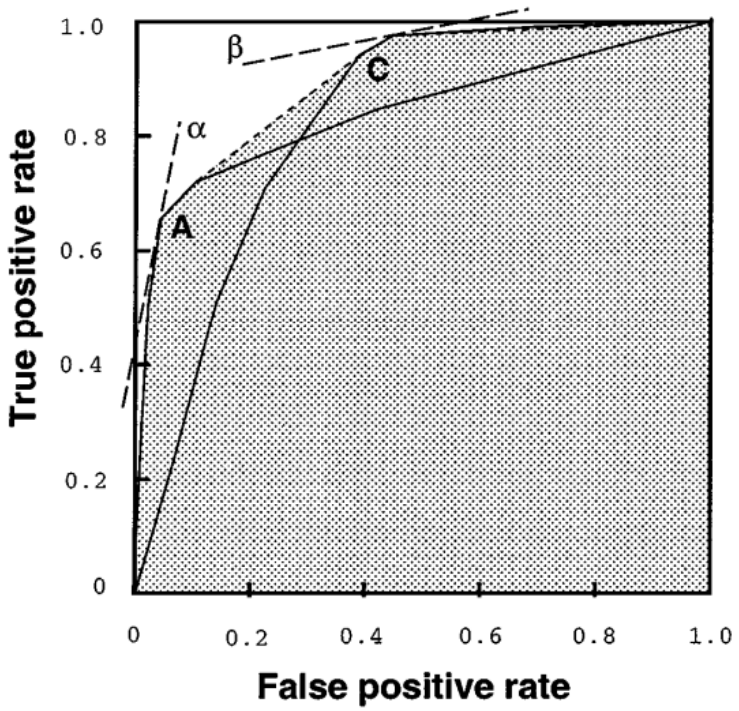


Figure 4.4: ROC curves with iso-perfomance lines. Figure taken from [28].

# Chapter 5

# Description and results of the machine learning model

## 5.1 Description of the model

Now that the general framework is set up, we can describe in details how the final machine learning model was built. As was seen previously, the goal is to construct the best (in the sense of section 4.3) possible machine learning model that takes a set of attributes of the picture, noted $\mathbf{X}$, as input and returns a the probability that a particular galaxy is lensed or not. This is done thanks to a classifier $S$. The image is then classified as a lens if this probability is greater than a threshold $t$.

We will start by contextualizing the present work. As mentioned earlier, this master thesis is a follow up of [18]. In [18], a significant amount of preprocessing was carried out. The first steps of building the machine learning model were also performed. Our first task was therefore to clean up, comment and reorganize the code of [18]. The second and more important task was to develop a more rigorous framework to evaluate the performance of machine learning models in the context of lens search. With this in mind, we searched the literature and found the method described in section 4.3 along with other tools to be explained later in this chapter.

In order to build efficient models, one needs to know which attributes to include. This is called feature selection. A review of the topic of feature selection can be found in [36]. A possible way to determine the most relevant attributes is to sort them by increasing information gain. The information gain of a feature $X_\mathrm{i}$ with respect to the classes $C$ is defined as

$$IG(X_\mathrm{i}) = H(X_\mathrm{i}) - H(X_\mathrm{i}|C) \tag{5.1}$$

where $H(X_\mathrm{i})$ is the entropy of $X_\mathrm{i}$ and $H(X_\mathrm{i}|C)$ is the entropy of $X_\mathrm{i}$ when we know that it belongs to class $C$. These are defined as

$$H(X_\mathrm{i}) = -\sum_j p(x_\mathrm{j}) \log_2(p(x_\mathrm{j})) \tag{5.2}$$

$$H(X_\mathrm{i}|C) = -\sum_{k=0,1} p(c_\mathrm{k}) \sum_j p(x_\mathrm{j}|c_\mathrm{k}) \log_2(p(x_\mathrm{j}|c_\mathrm{k})) \tag{5.3}$$

where $x_\mathrm{j}$ denote the possible values taken by the feature $X_\mathrm{i}$ and $c_0$ and $c_1$ are the two possible classes. This metric has the advantage to be computationally simple and to possess a simple interpretation in terms of information theory. It is indeed the information that we obtain by adding the feature $X_\mathrm{i}$ to the model.

The attributes were calculated thanks to statmorph [29]. statmorph is a Python package that enables us to compute parametric and non-parametric features of galaxies. A complete list of the possible non-parametric features computable thanks to statmorph along with their descriptions is given in appendix A. There are 23 parameters that we can calculate with statmorph. Here, we will describe only the four most relevant parameters. These are, listed by decreasing information gain,

- The most significant parameter is the the Gini-$M_{20}$ bulge index [34] [29]. It uses the normalized second order moment of the brightest 20% of the pixels $M_{20}$ [22]. It is defined as

  $$M_{20} = \log(\frac{\sum_i M_i}{M_\mathrm{tot}}) \tag{5.4}$$

  where $M_i$ is the second order moment of pixel $i$, $f_i$ is the fraction of the flux carried by pixel $i$ (subject to the constraint that $\sum_i^n f_i = 0.2$) and $M_\mathrm{tot}$ is expressed as

  $$M_\mathrm{tot} = \sum_i^n f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{5.5}$$

  $(x_c, y_c)$ is the position of the center of the galaxy defined such that $M_\mathrm{tot}$ is minimized. For reasons given in [34] the Gini-$M_{20}$ bulge index is the position of a galaxy along a line in the Gini-$M_{20}$ plane that is perpendicular to the line

  $$F(G, M_{20}) = -0.693 M_{20} + 4.95 G - 3.96 \tag{5.6}$$

  with the origin in $(0.565, -1.679)$. This index increases for bulge dominated systems, hence the name.

- The Gini index [22], already widely used in economics [12], quantifies the inequality

of the light distribution of the galaxy. It is defined by the equation

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j| \tag{5.7}$$

where $n$ is the total number of pixels, $\bar{X}$ the mean value of the intensities of the pixels and $X_i$ the intensity of pixel $i$. It lies in the interval $[0, 1]$. 0 corresponds to a completely egalitarian distribution while 1 is an extremely unequal distribution.

- The concentration [7] quantifies how spread out the luminosity profile of a galaxy is. It given by

$$C = 5 \log(\frac{r_{80}}{r_{20}}) \tag{5.8}$$

where $r_{80}$ is the radius of the region where 80 % of the flux lies and where $r_{20}$ is the radius of the region where 20 % of the flux lies. The central pixel is defined such that the asymmetry is minimized (see definition below). It is strongly correlated to the $M_{20}$ due to their similar definitions.

- As its name suggests, asymmetry enables us to quantify the asymmetry of the luminosity profile of the galaxy [1]. Its defining equation is

$$A = \frac{\sum_{i,j} |I(i,j) - I_{180}(i,j)|}{\sum_{i,j} I(i,j)} - B_{180} \tag{5.9}$$

where $I(i, j)$ is the intensity of the pixel located at coordinate $(i, j)$, $I_{180}(i, j)$ is the intensity of the pixel located at coordinate $(i, j)$ on the image rotated by 180° around the central pixel and $B_{180}$ is the average asymmetry of the background. $B_{180}$ is the magnitude of the difference between the background pixels and the background pixels of the image rotated by 180°.

A pairplot of these features is given in figure 5.1. This allows us to visualize the relevance of these parameters as the positive and negative classes are neatly separated in the pairplot, especially for the first three. Indeed, we can intuitively expect sharp inequalities and asymmetries in the light distribution of a lens.

We also tried to compute the Sersic parameters of the galaxies. It however proved impossible to obtain reliable estimations of Sersic parameters. Indeed, both statmorph [29] and galfit [25] [26] (another software specifically designed to calculate Sersic parameters) would flag the results as unreliable for more than half of the dataset. This is likely due to the background of the provided images. We therefore abandoned the idea to use Sersic parameters for the machine learning models.

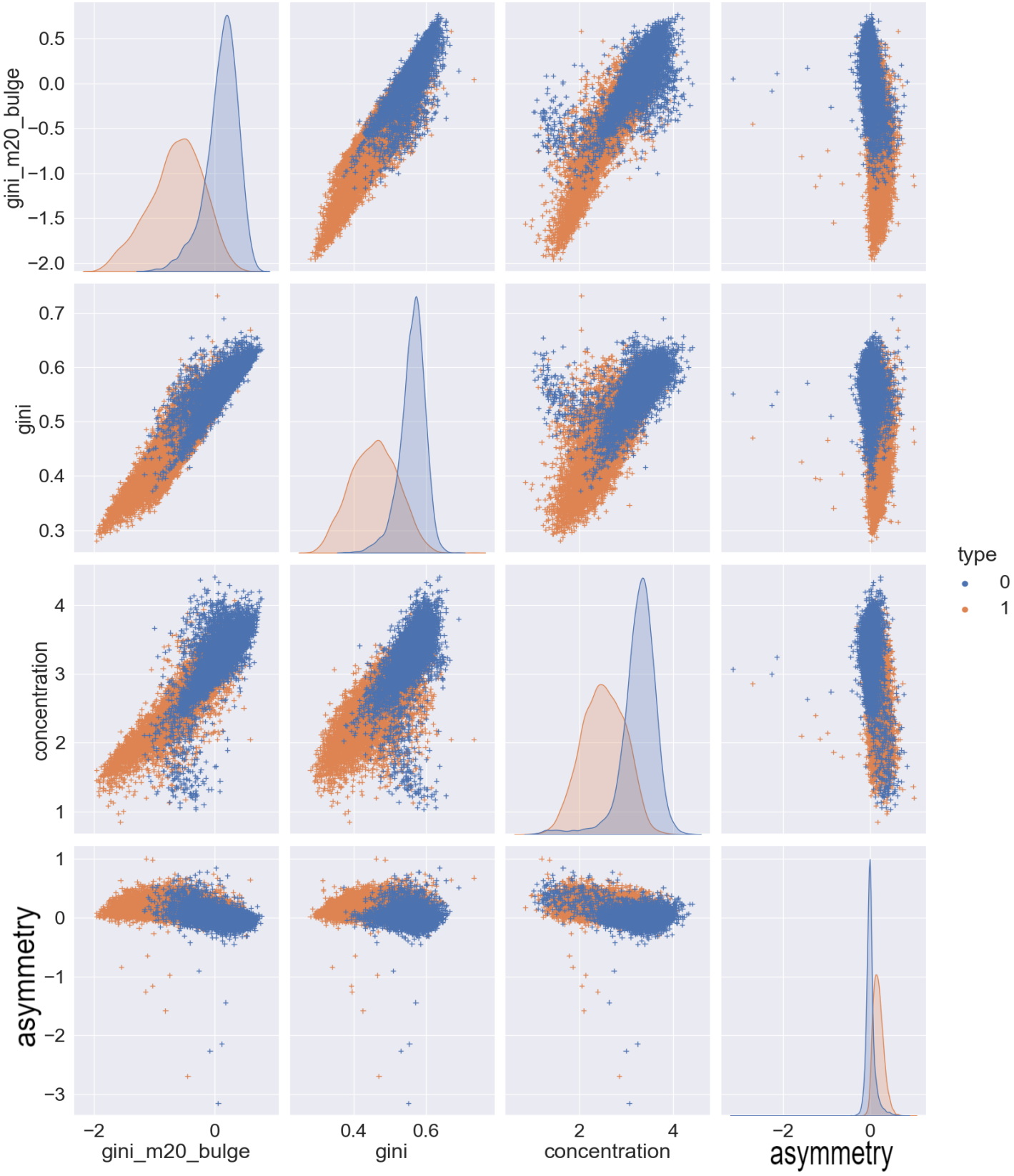


Figure 5.1: Pairplot of the four most relevant attributes. Orange points correspond to lenses. Blue points are non-lensed galaxies. The four most relevant attributes are defined in the main text.

## 5.2 ROC curves of different classifiers

We can now describe the course of action used to build the best machine learning models. First, we computed the ROC curves of random forests using 10 to 23 parameters with the most information gain. The same procedure was carried out for extremely randomized trees. The results are displayed in figures 5.2 and 5.3 (only a few of the models are shown in the figures for readability). We can immediately conclude from the figures that the random forests or ERT containing 10 parameters are not optimal. Indeed, the ROC curves for the models that include 10 parameters lies below the others. We know from section 4.3 that such ROC curve are not optimal. That conclusion holds for both random forests and extremely randomized trees. The eleventh most important parameter in terms of information gain is the signal to noise ratio. It is defined in appendix A. The ROC curves are subject to random fluctuations due to the fact that they are subject to significant random noise and only a rough conclusion can be drawn at this stage. A more analysis of theses curves is postponed to section 5.3.

We also created models using principal components as inputs. Similarly to what was done for bare parameters, we computed the principal components and sorted them by decreasing order of information gain. We turned to principal components mainly because they are linearly uncorrelated, which is not the case for the original parameters. The results are reported in table 5.1. This table illustrates the fact that the principal components that explain the most variance are not necessarily the most relevant. In figures 5.4 and 5.5, we drew the ROC curves of random forests and extremely randomized trees that used the tenth to fifteenth most relevant principal components in terms of information gain, as over 90 % of variance is explained with these 15 principal components. We can already note a significant improvement compared to models using the attributes. Indeed, we can achieve a higher true positive rate for a given false positive rate. As the curves are subject to random fluctuations, no best models can be deduced from these figures.

| Principal component | PVE | cumulative PVE |
|---|---|---|
| Z1 | 29.69 % | 29.69 % |
| Z12 | 1.12 % | 30.8 % |
| Z8 | 3.39 % | 34.2 % |
| Z10 | 1.88 % | 36.08 % |
| Z13 | 0.99 % | 37.07 % |
| Z9 | 2.97 % | 40.04 % |
| Z19 | 0.1 % | 40.1 % |
| Z11 | 1.4 % | 41.5 % |
| Z5 | 4.8 % | 46.3 % |
| Z3 | 9.7 % | 56.008 % |
| Z6 | 4.2 % | 60.2 % |
| Z2 | 25.8 % | 86.02 % |
| Z18 | 0.1 % | 86.1 % |
| Z16 | 0.3 % | 56.4 % |
| Z7 | 4.02 % | 90.4 % |
| Z14 | 0.7 % | 91.1 % |
| Z20 | 0.04 % | 91.1 % |
| Z22 | 0.01 % | 91.1 % |
| Z15 | 0.4 % | 91.6 % |
| Z17 | 0.2 % | 91.8 % |
| Z21 | 0.03 % | 91.8 % |
| Z24 | $9.10^{-31}$ % | 91.8 % |
| Z23 | $10^{-30}$ % | 91.8 % |
| Z4 | 8.2 % | 100 % |

Table 5.1: Table containing the principal components, ranked by decreasing order of information gain, along with their percentage of variance explained (PVE) and the cumulative PVE of all the previous and current rows of the array.

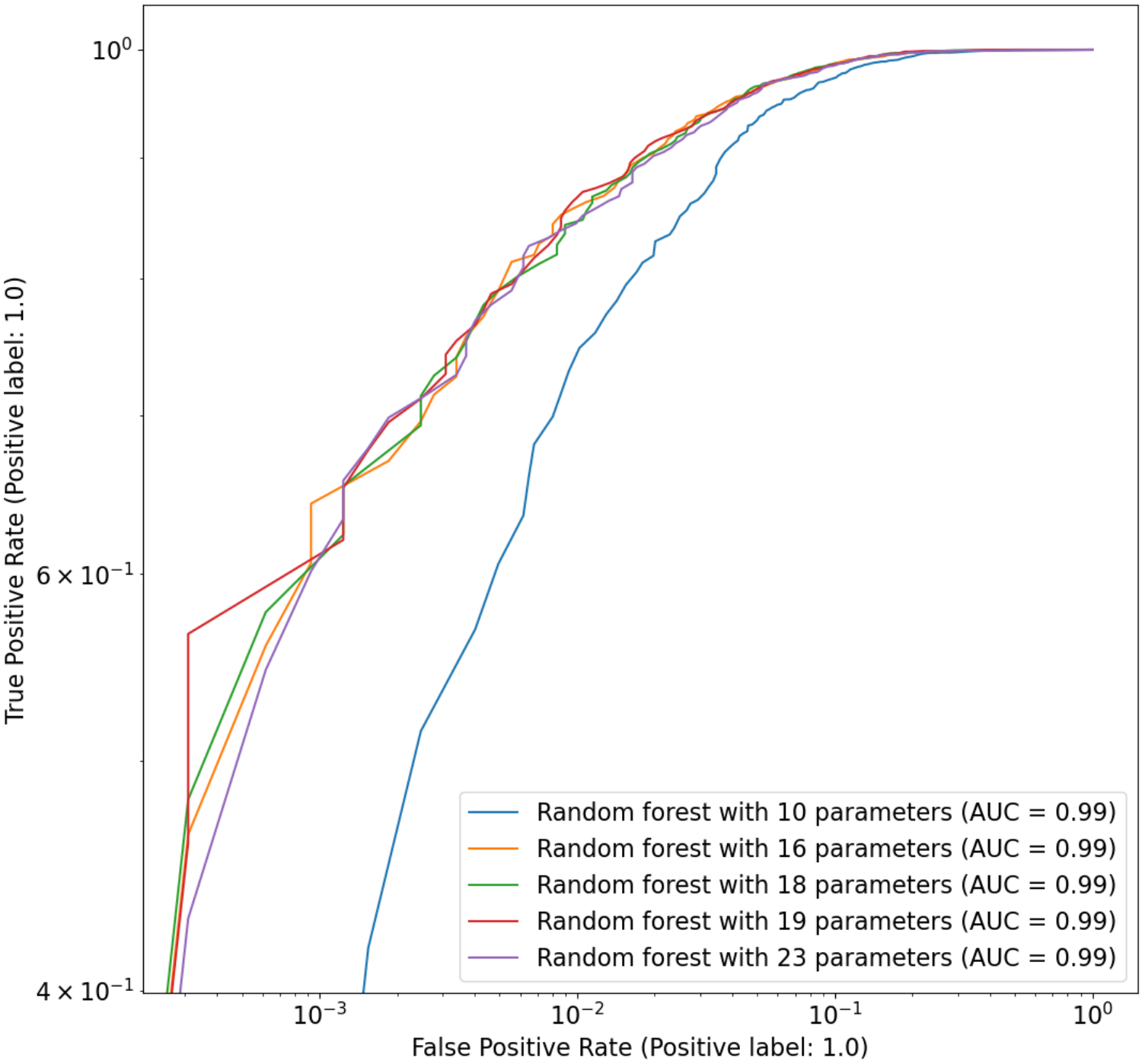


Figure 5.2: Five empirical ROC curves for random forests with different sets of attributes. Both x and y axes are in logarithmic scales. The AUCs of the models are also displayed in the legend.

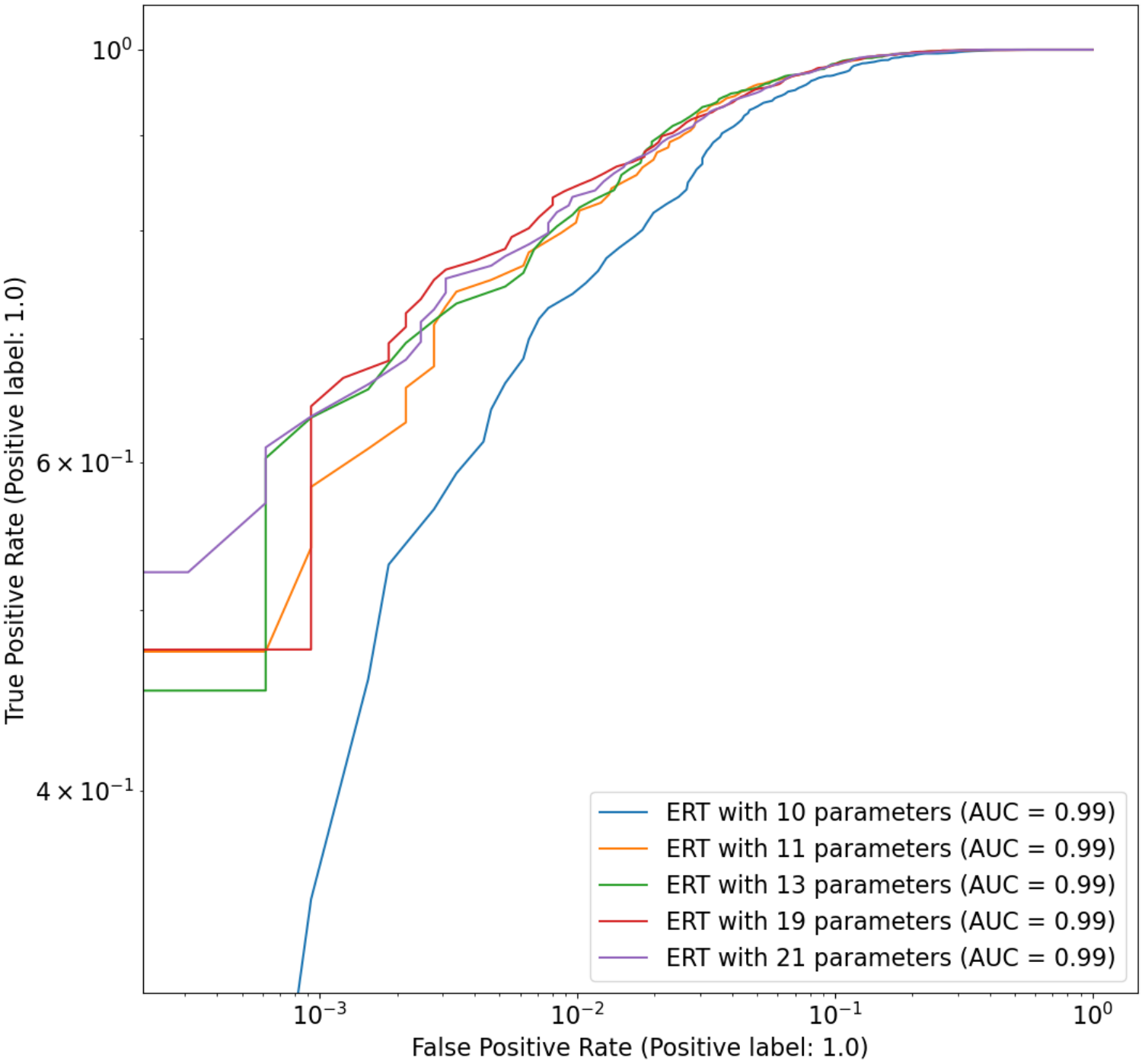


Figure 5.3: Five empirical ROC curves for extremely randomized trees with different sets of attributes. Both x and y axes are in logarithmic scales. The AUCs of the models are also displayed in the legend.

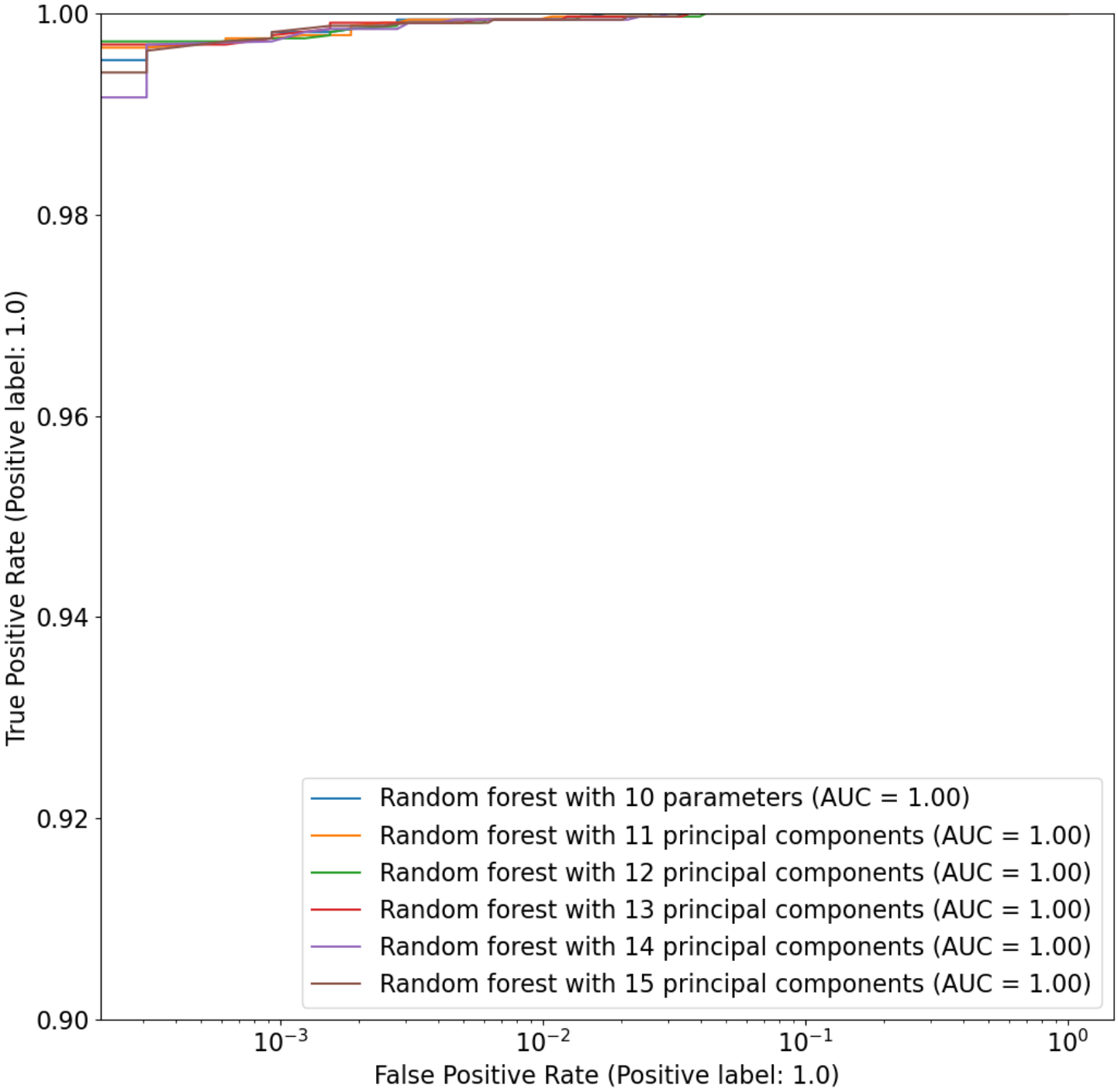


Figure 5.4: Six empirical ROC curves for random forests with different sets of principal components. The x axis is in logarithmic scale. The AUCs of the models are also displayed in the legend.

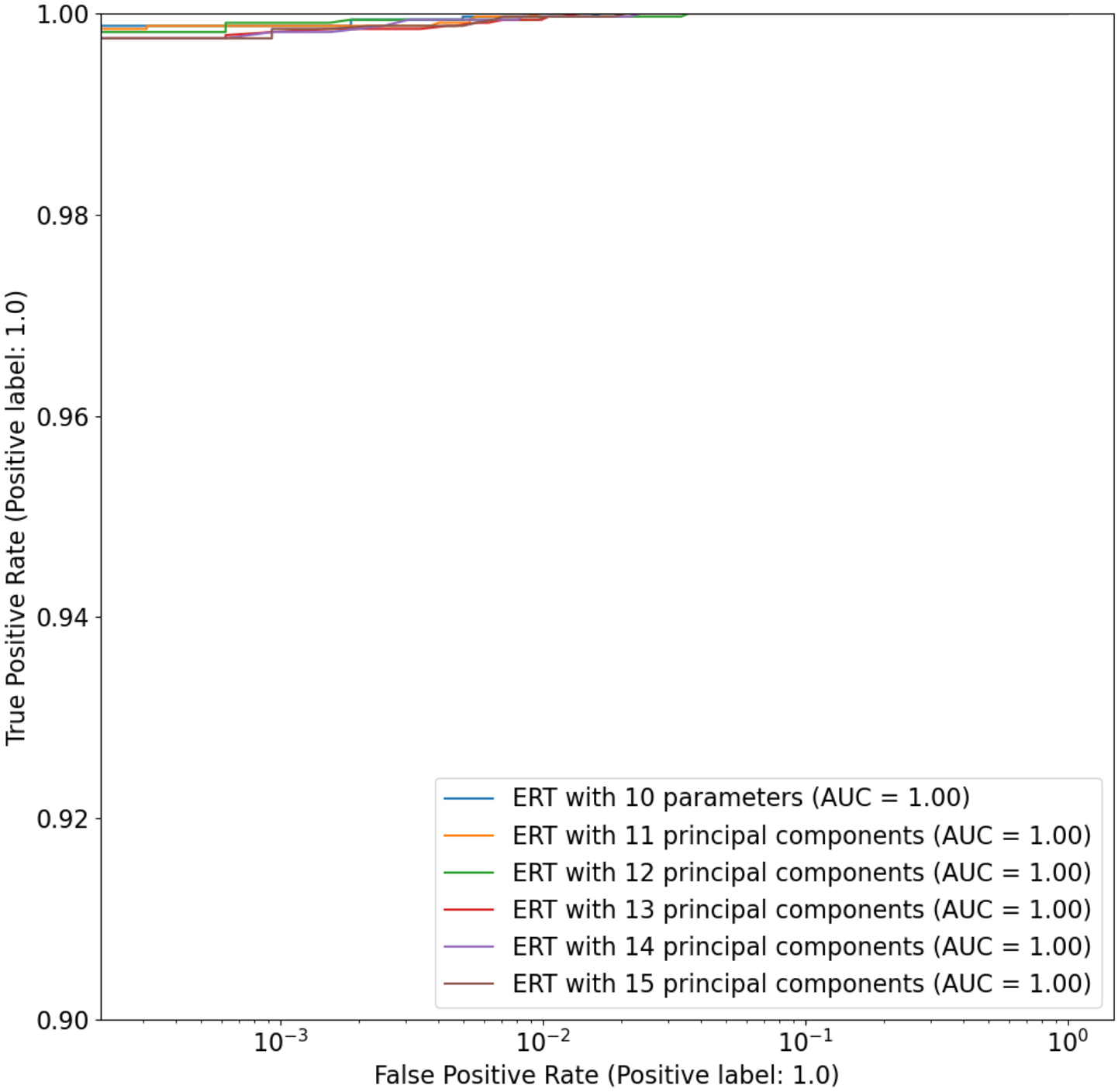


Figure 5.5: Six empirical ROC curves for extremely randomized trees with different sets of principal components. The x axis is in logarithmic scale. The AUCs of the models are also displayed in the legend.

## 5.3 Analysis of the bootstrapped ROC curves

We have to keep in mind that since the observations contain noise, the ROC curve shown in the previous section are themselves random variables with unknown variance. Indeed, both the false positive rate and false negative rate are probabilities. Since we are working on a limited sample, we can only rely on empirical estimators to compute the probabilities. In order to obtain a good approximation of the true ROC curves, we bootstrapped different empirical ROC curves. Bootstrapping is a technique to obtain reliable estimates of random quantities if we have at our disposal only a limited sample. For the objects at hand (i.e. ROC curves), it consists in calculating the ROC curves for data samples drawn from the original sample with replacement. The ROC curves are then averaged to obtain an estimator of the ROC curve with a variance that is greatly reduced.

As bootstrapping is a computationally intensive procedure, we could only bootstrap a limited number of the empirical ROC curves presented in the last section. For the models using attributes, we noticed that both for random forests and extremely randomized trees with 19 to 23 parameters laid on the convex hulls of empirical ROC curves for false positive rates up to $10^{-3}$. This can be seen on figures 5.2 and 5.3. We therefore chose to bootstrap these ROC curves. The results are reported in figures 5.6 and 5.7.

We can apply the analysis technique developed in section 4.3 to figures 5.6 and 5.7. We can see that in both figures, the convex hull consists only in one curve. The model associated to this curve is therefore the best model. We observe that the best random forest uses 23 parameters while the best ERT uses 19 parameters. In figure 5.11, we drew the ROC curves of these two models in the same graph. We can see that the two ROC curves intersect. However, using the method developed in section 4.3, we can say that this crossing occurs for regions of low misclassification cost, which is not of interest here.

We repeated the same procedure for classifiers that use principal components as input. We chose to bootstrap the curves of models that include from 10 to 15 principal components, as they contain most of the variance, as shown in table 5.1. The results for ERT and random forests are displayed in figures 5.9 and 5.8. We see that again for each classifier the convex hull is just one curve. Hence, the best ERT uses 12 principal components whereas the best random forest uses 15 principal components. In figure 5.10, we plotted the ROC curves of the two best models on the same graph. We immediately see that the ERT with 12 principal components is better than the random forest.

We can finish this chapter by comparing figures 5.10 and 5.11. Overall, models using

principal components are far better than the models using bare attributes. Moreover, they use fewer principal components, compared to the number of bare attributes needed to describe the data. This is due to the fact that the principal components are uncorrelated and hence account for the redundancy between attributes. The principal components carry therefore more information. Tables 5.2, 5.3, 5.4 and 5.5 summarize our findings. We can conclude that the best machine learning model to search for strong lenses found in this work is an ERT model using 12 principal components.

| ERT with 12 principal components | | |
|---|---|---|
| False positive rate | True positive rate | 95 % confidence interval |
| $10^{-3}$ | 0.985 | [0.9702,0.993] |
| $10^{-2}$ | 0.997 | [0.995,0.998] |
| $10^{-1}$ | 1 | [0.999,1] |

Table 5.2: Table containing the true positive rates for given false negative rates and the bounds of the 95% confidence intervals on these true positive rates for an ERT with 12 principal components.

| Random forest using 15 principal components | | |
|---|---|---|
| False positive rate | True positive rate | 95 % confidence interval |
| $10^{-3}$ | 0.926 | [0.842,0.972] |
| $10^{-2}$ | 0.983 | [0.961,0.994] |
| $10^{-1}$ | 1 | [0.999,1] |

Table 5.3: Table containing the true positive rates for given false negative rates and the bounds of the 95% confidence intervals on these true positive rates for a random forest with 15 principal components.

| ERT using 19 attributes | | |
|---|---|---|
| False positive rate | True positive rate | 95 % confidence interval |
| $10^{-3}$ | 0.562 | [0.467,0.631] |
| $10^{-2}$ | 0.823 | [0.8005,0.843] |
| $10^{-1}$ | 0.984 | [0.981,0.986] |

Table 5.4: Table containing the true positive rates for given false negative rates and the bounds of the 95% confidence intervals on these true positive rates for an ERT with 19 attributes.

| Random forest using 23 attributes | | |
|---|---|---|
| False positive rate | True positive rate | 95 % confidence interval |
| $10^{-3}$ | 0.548 | [0.422,0.636] |
| $10^{-2}$ | 0.819 | [0.798,0.834] |
| $10^{-1}$ | 0.985 | [0.982,0.986] |

Table 5.5: Table containing the true positive rates for given false negative rates and the bounds of the 95% confidence intervals on these true positive rates for a random forest with 23 attributes.

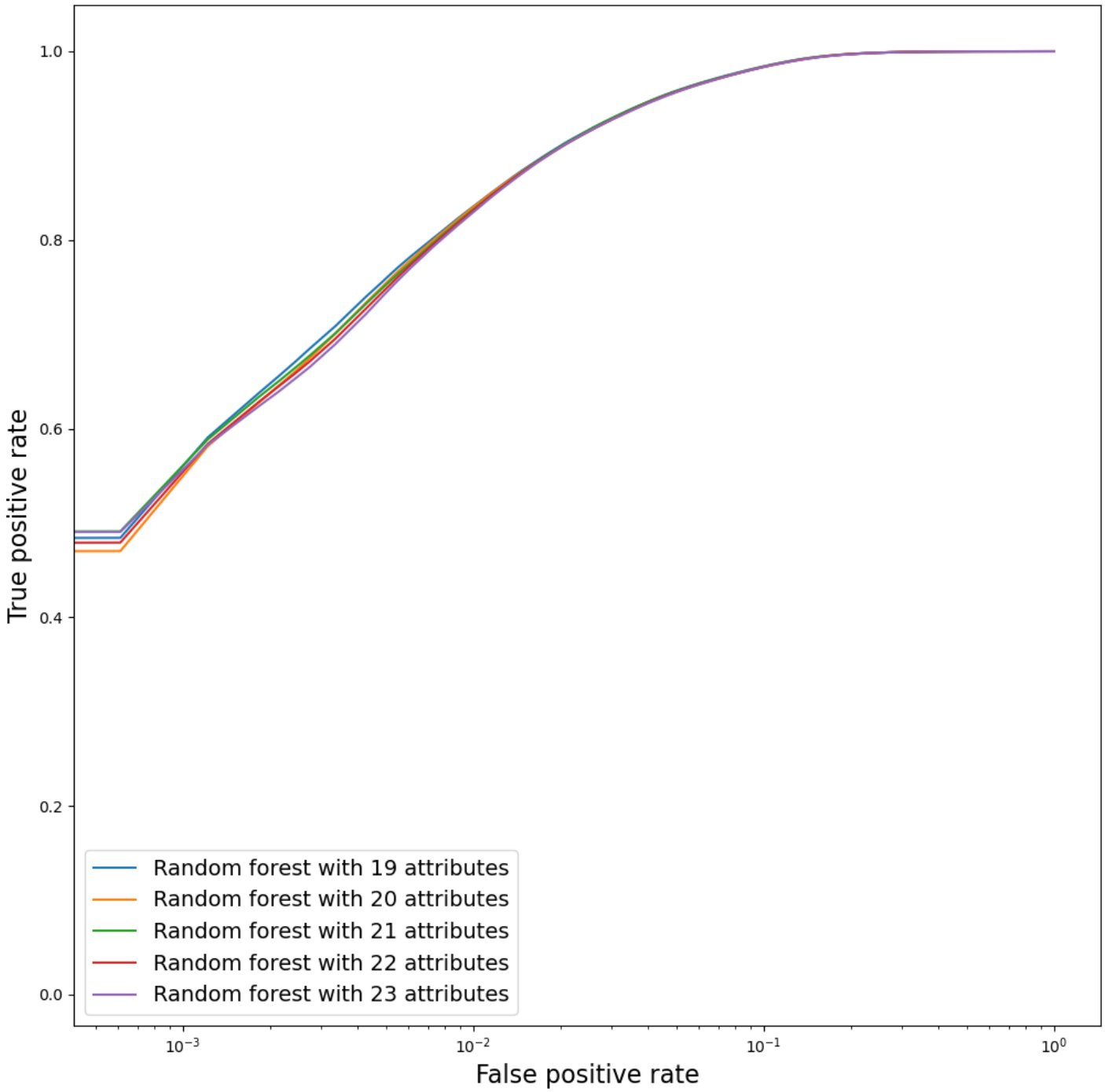


Figure 5.6: Bootstrapped ROC curves of five different random forests built using the attributes as input. X axis is in logarithmic scale.

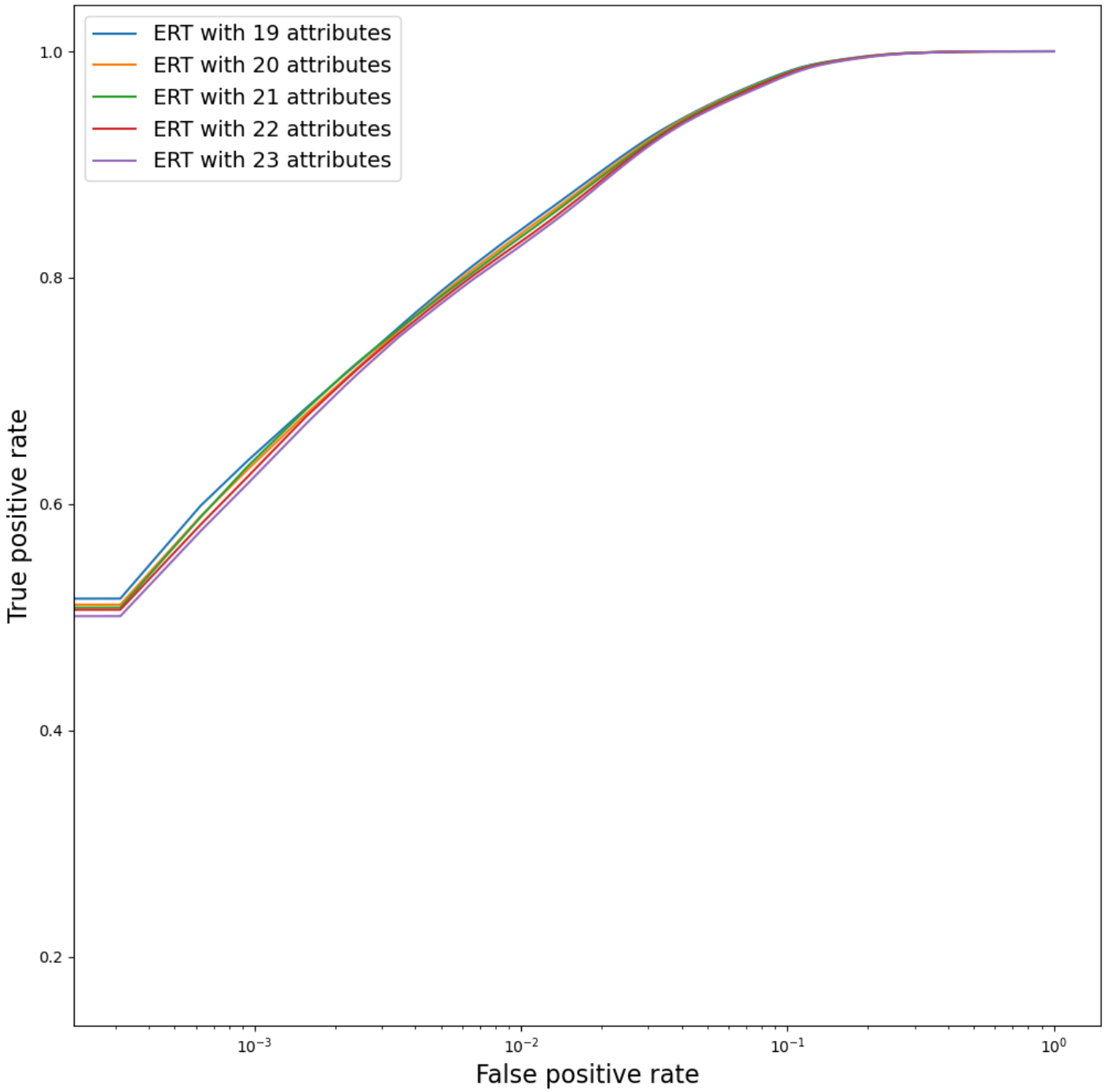


Figure 5.7: Bootstrapped ROC curves of five different extremely randomized trees built using the attributes as input. X axis is in logarithmic scale.

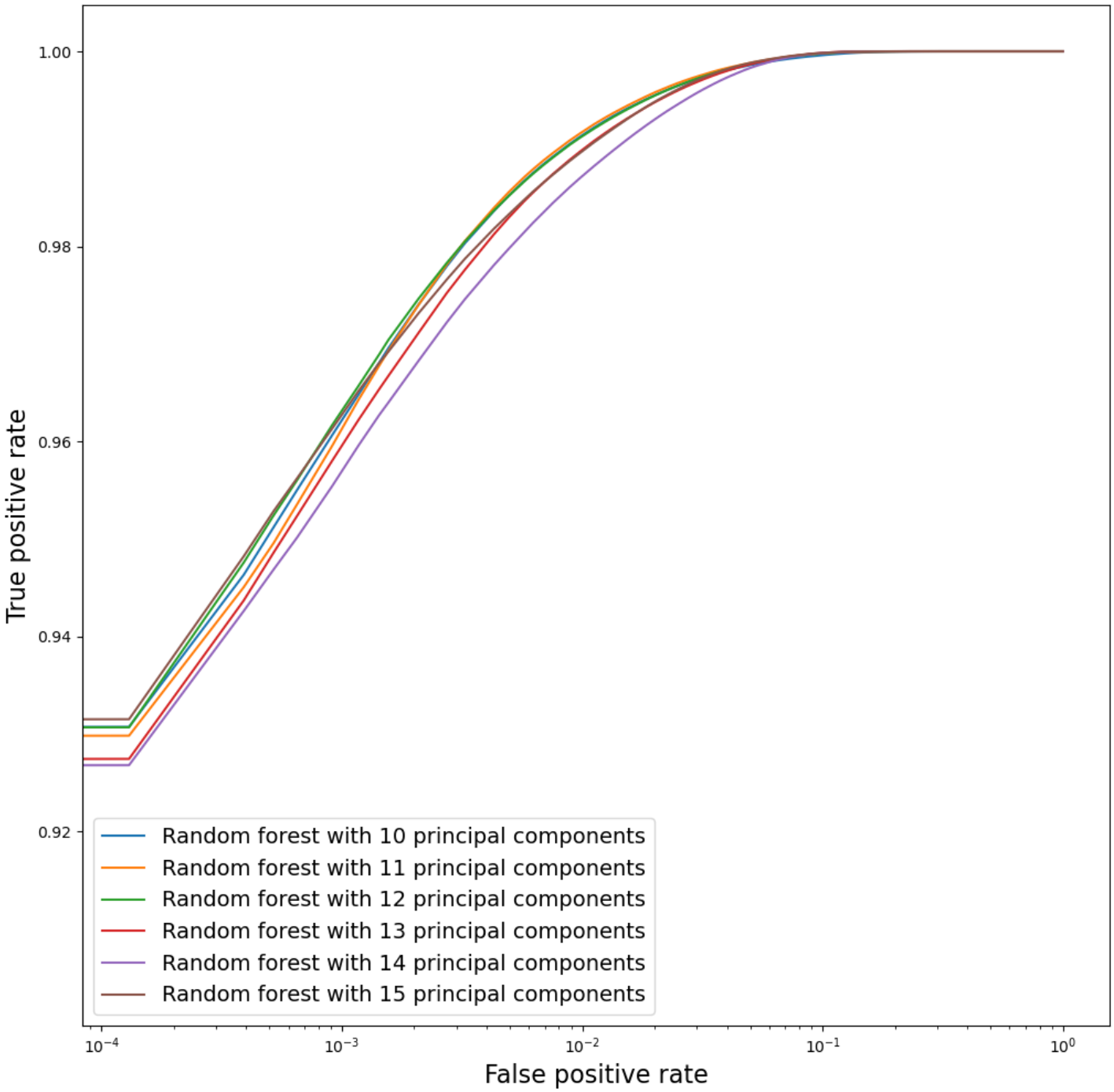


Figure 5.8: Bootstrapped ROC curves of six different random forests built using the principal components as input. X axis is in logarithmic scale.

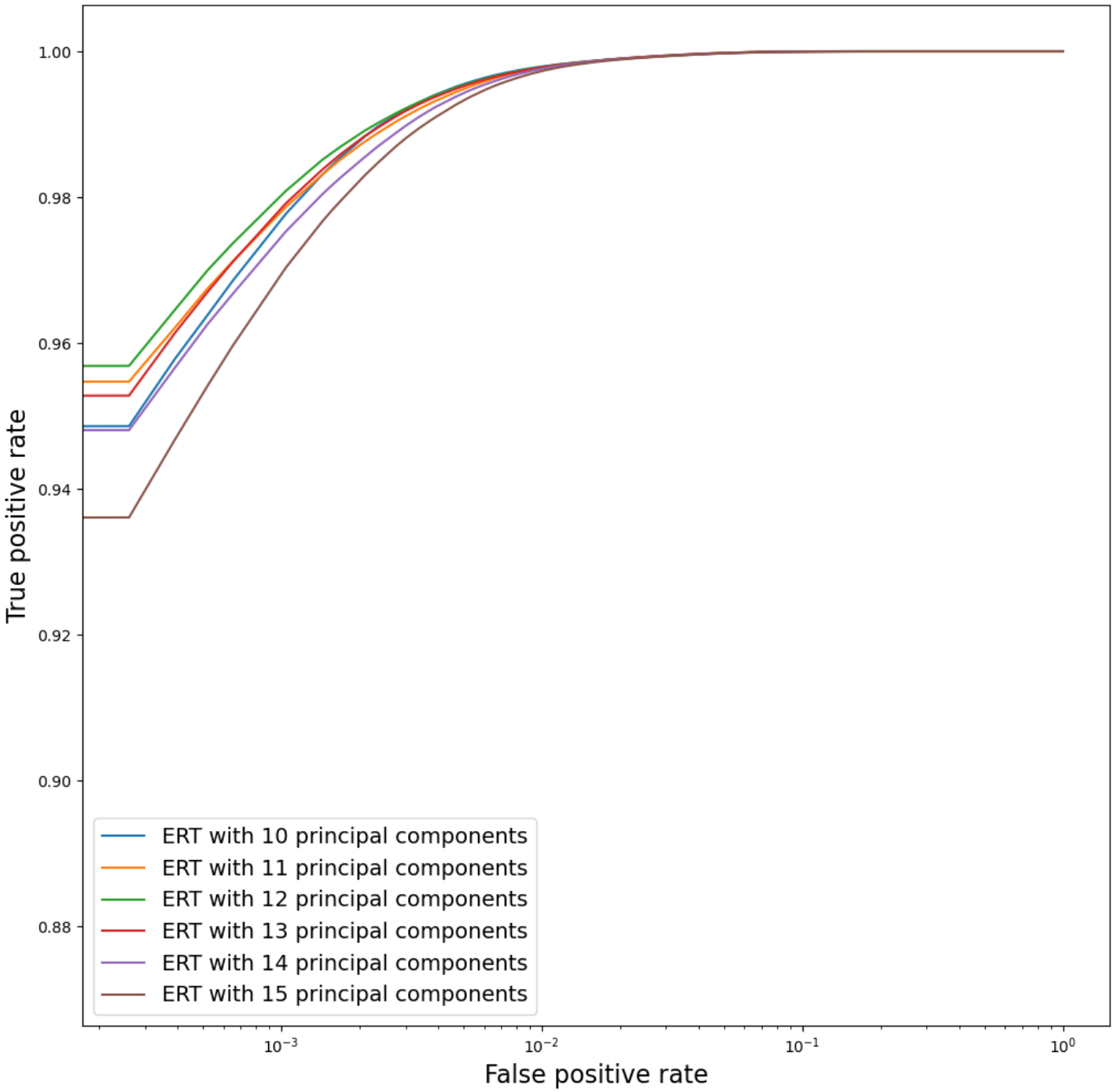


Figure 5.9: Bootstrapped ROC curves of six different extremely randomized trees built using the principal components as input. X axis is in logarithmic scale.

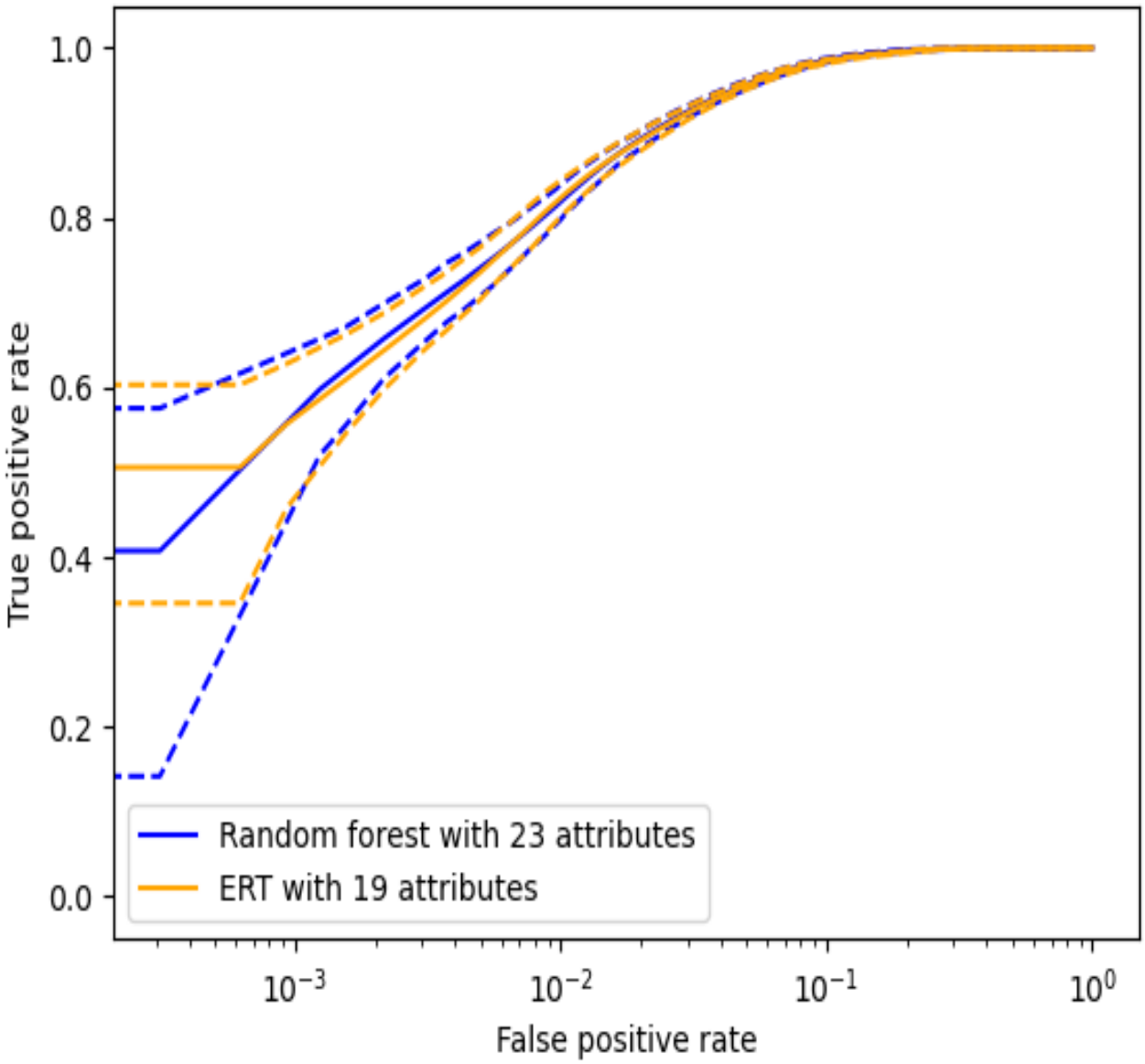


Figure 5.10: Bootstrapped ROC curves of the two best models built with the attributes. X axis is in logarithmic scale. The dashed lines correspond to the bounds of the 95 % confidence interval.

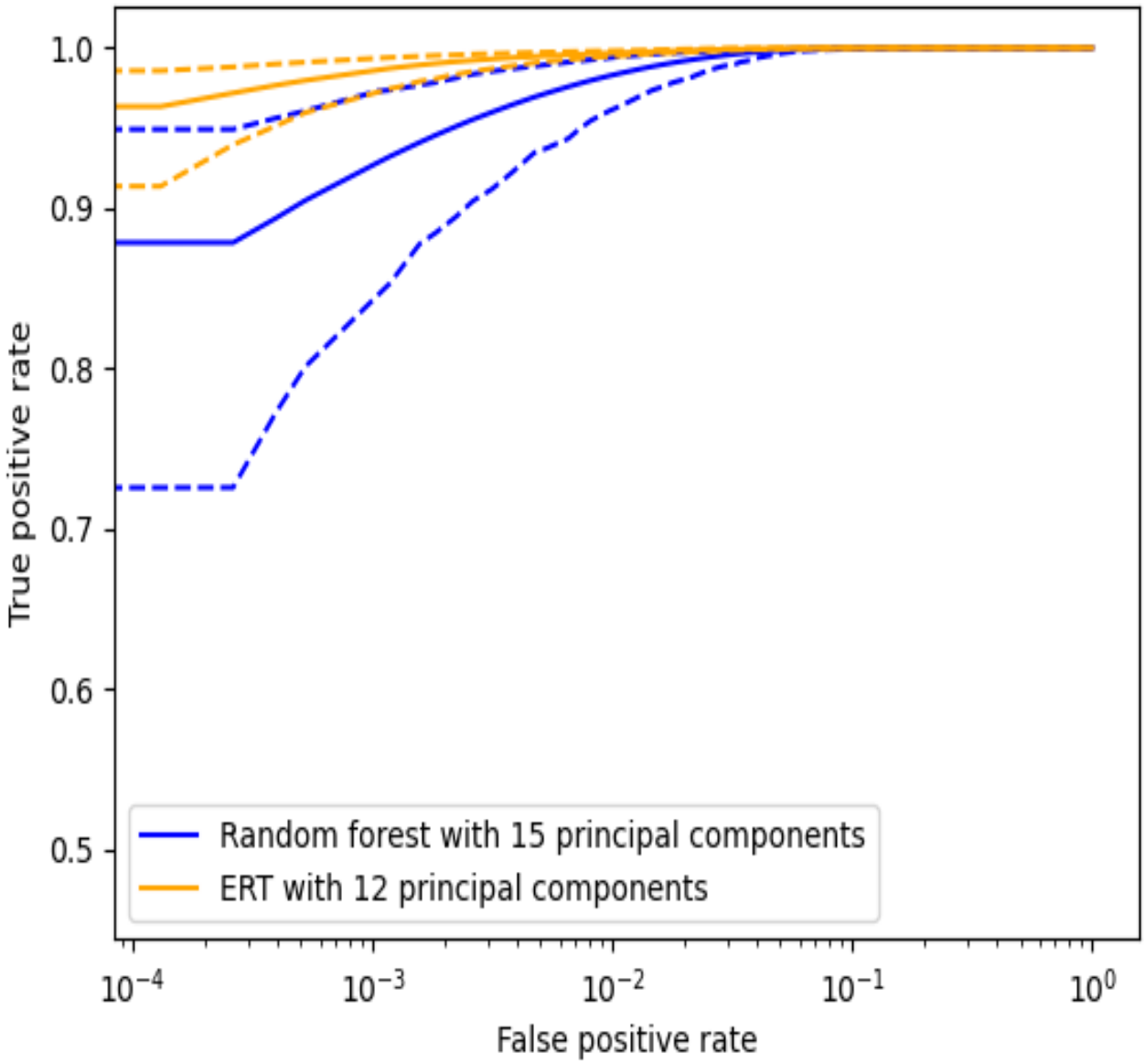


Figure 5.11: Bootstrapped ROC curves of the two best models built with the principal components. X axis is in logarithmic scale. The dashed lines correspond to the bounds of the 95 % confidence interval.

# Chapter 6

# Comparison with the deep learning model

## 6.1 Presentation of the convolutional neural network model

In this chapter, we are going to compare the machine learning model developed in the previous chapter with the deep learning model of [30]. We will begin by describing in more details the neural network created in [30], then we will proceed to the comparison.

We are first going to detail the functioning of the deep learning model developed by Savary et al. in [30]. As mentioned numerous times during the course of this work, the authors used a convolutional neural network as their classifier. CNNs are inspired by biology and they consist in feeding the data to successive mathematical functions called layers. The bulk of these layers in CNNs consist in convolution products. The parameters of the CNN are computed by minimizing a loss function. This minimization is performed thanks to optimization algorithms such as Adagrad or gradient descent. CNNs are especially efficient to process images. More on CNNs can be found in [23].

The authors of [30] made use of a particular architecture of CNNs called EfficientNet [35]. More specifically, the B0 version of EfficientNet. Savary et al. adapted the implementation of B0 from Keras Python library [6] to treat images of $44 \times 44$ pixels which is the size of the images of the training set. Additionally, different neural networks were trained on different subsets of the training sets. As mentioned earlier, the dataset is the same as the one we used to train our machine learning model. The predictions of these networks were then averaged to obtain the final score. This technique is called

ensemble-averaging. As a reminder, a similar technique is used to construct random forests and extremely randomized trees. In these cases, the average of different models is done by performing the average of the outputs of the different trees.

As the predictions of CNNs are not invariant under rotation, predictions were made for the original images as well as rotated and flipped versions of the original images. Three rotations were performed, by 90°, 180° and 270° and flips of the original and rotated images. The final prediction is an average of these predictions.

## 6.2 ROC curves of EfficientNet

In order to compare the models developed in chapter 5 to EfficientNet, we will again compare their respective ROC curves. In order to draw the ROC curve of EfficientNet, we needed its test set as well as the scores that were given to the galaxies of the test set. The test set consisted in 1060 galaxies, 50 % of which were images of galaxies taken from the Pan-STARRS1 survey that did not exhibit any lens feature and 50 % of which were simulated lenses taken from the same dataset that we trained and tested our machine learning model with.

A file containing the results of EfficientNet on the test was provided by the authors of [30]. Following the same procedure as the one described in the previous chapter, we first drew the empirical ROC curve of the deep learning model. The result is displayed in figure 6.1. We then bootstrapped this ROC curve to obtain a reliable approximation of the true ROC curve. This is shown in figure 6.2.

As we can see in both figure 6.1 and figure 6.2, the model developed in [30] appears to be a perfect classifier that clearly outperforms all the models developed in chapter 5. This result can be explained by several factors. First, we did not have at our disposal the results on the whole test set. We only had the results for 450 galaxies out of the 1060 reported in [30]. Secondly, even if we had the complete results, 1060 galaxies, with 50 % of lenses, is a too small sample to produced reliable ROC curves. Indeed, most statistics are high quality approximations only if the sample size is sufficiently large, which was not the case here. So as to perform a more robust analysis of their results, the authors of [30] could have included more galaxies in their test set. Drawing the ROC curve in [30] would have made this problem more blatant.

We can give a more quantitative assessment of these considerations. We can indeed

compute the probability of obtaining a "perfect" ROC curve for the machine learning model developed in chapter 5, namely an ERT with 12 principal components and with a test set of 450 images. To do so, we trained the model in similar conditions than the one used to train EfficientNet in [30], then we bootstrapped the ROC curves obtained with a test set of 450 images. We could then calculate the probability of obtaining a perfect ROC curve. This probability is 99.2 %. Hence, this number clearly gives an indication of the issue encountered in the comparison with EfficientNet.

## 6.3 Discussion of the results

Despite the problem of statistical significance detected in the methodology of [30], it is expected that EfficientNet is still better than simple machine learning models such as the ones created in chapter 5.

The main reasons would be their higher dimensional parameter space and their ability to perform feature selection automatically. Despite this, we have shown in chapter 5 that machine learning algorithms, especially if principal components are used as inputs, could still be highly accurate. Moreover, their interpretability allow them to be used in combination of deep learning with the purpose of discovering the possible biases of the former.

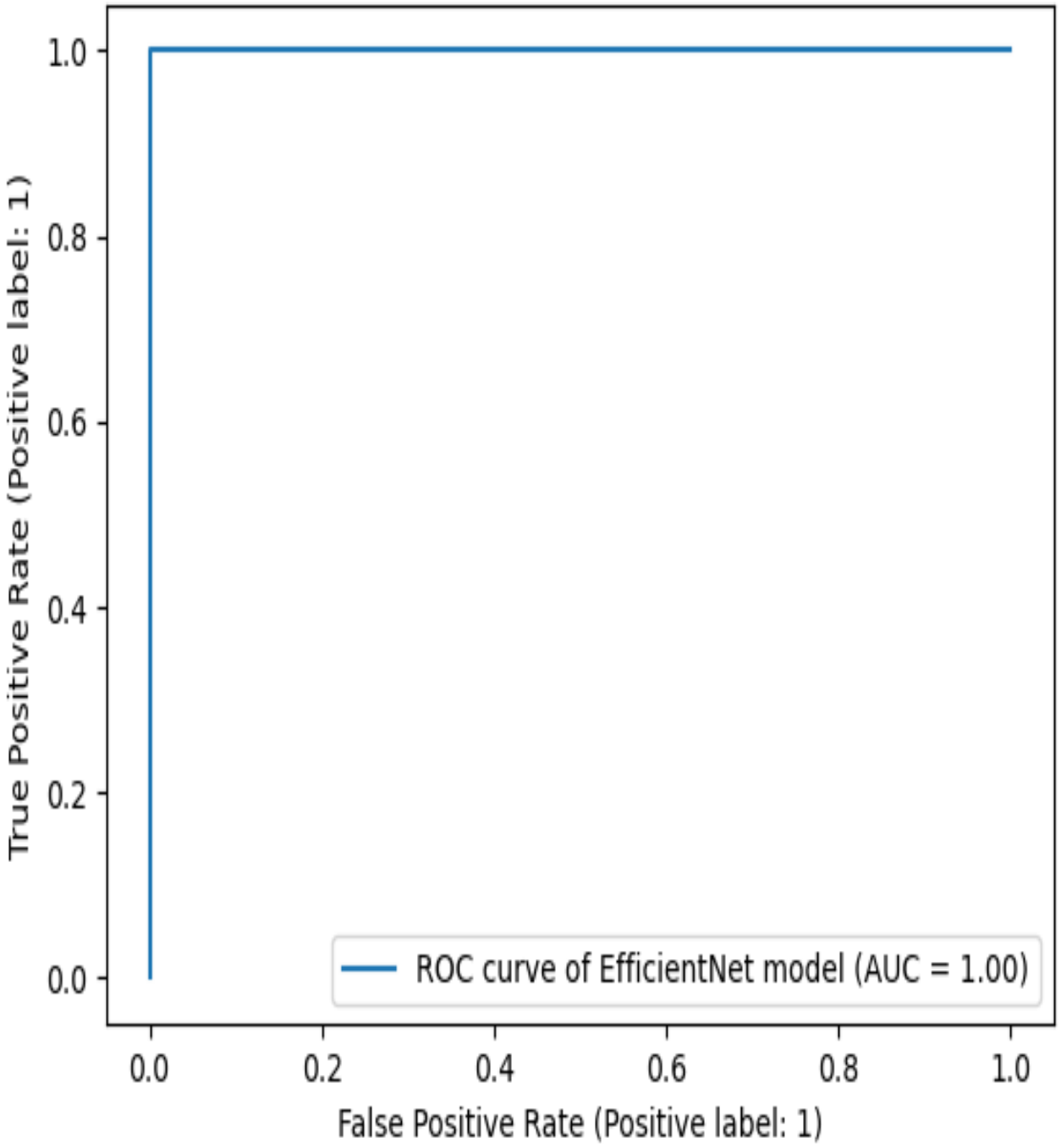


Figure 6.1: Empirical ROC curve of Efficient NET

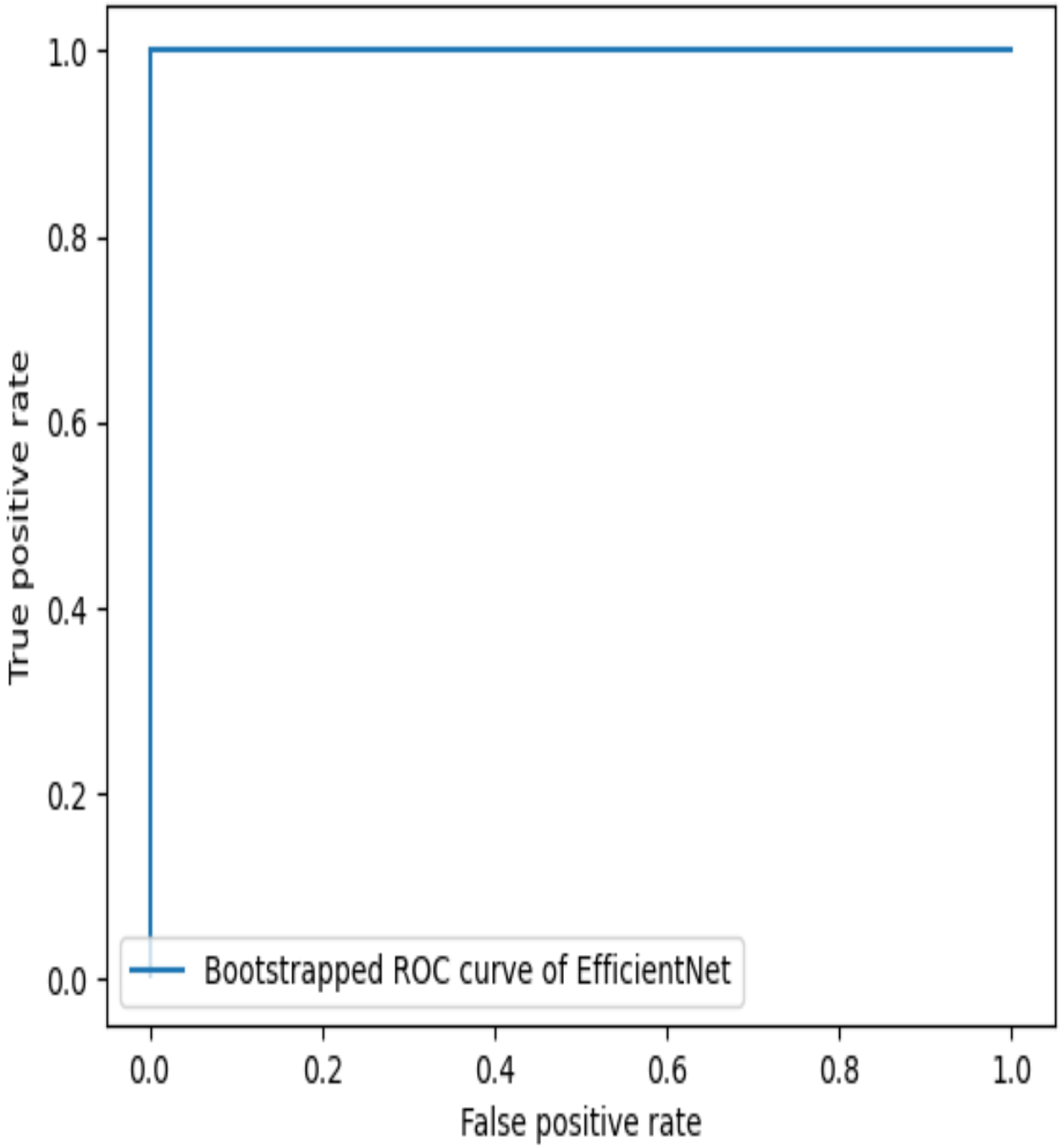


Figure 6.2: Bootstrapped ROC curve of Efficient NET

# Chapter 7

# Conclusion

## 7.1 Main results and discussion

We are now going to conclude this work. We first started by reviewing the physical principles governing gravitational lenses. Then, we presented the origin of the data used in this work. Afterwards, we introduced the classification problem that we were facing, along with the algorithms used as classifiers. We also presented the tools needed to compare different models. This led to the introduction of ROC curves. A substantial amount of time was devoted to find a way to analyse ROC curves properly. A robust way to compare ROC curves was developed in [28] and it is this method that we used in our work.

As was mentioned several times during the course of this master thesis, this work is a follow up of a previous master thesis [18]. In [18], data preprocessing was completed and preliminary steps were taken in the construction of a reliable machine learning model. The dataset used in [18] and in our work was originally created to train the deep learning model of [30]. It consisted in red luminous galaxies taken from the CFIS survey as negative instances and simulated lenses combining images of the CFIS and the Hubble Space Telescope as positive instances.

The main goal of the present work was thus to extend and build more rigorous foundations of machine learning models destined to look for gravitational lenses. We also wished to compare the results of a machine learning model with the ones of a deep learning model, more specifically, the deep learning model of [30]. We began by refactoring and reorganizing the code of [18] and put it on a github.

Once the theoretical foundations and context were set up, we proceeded to build several different machine learning models. We first had to select features. Again, an important

amount of time was invested in trying to find the most robust way to select features. Finally, we determined that comparing the information gain of the different features was the most robust way to select features, because it is computationally efficient, model independent and the method is readily interpretable. Feature selection was therefore performed using the information gain of the different features. We also could not compute in a reliable way Sersic parameters for the dataset, despite using different softwares to do so, most likely because of the background of the images. We hence decided to discard the Sersic parameters and to build a model without using them.

Afterwards, we built the ROC curves of the different machine learning models and compared them. The best models were built using principal components as input. These can be considered to be accurate models. The model that was chosen in the end was an ERT model using 12 principal components. It achieved a 98.5 % true positive rate for a $10^{-3}$ false positive rate. All other machine learning models considered here were inferior in terms of true positive rates for a given false negative rates.

Lastly, we compared the performance of the deep learning model developed in [30] and our machine learning models. The deep learning model appeared to be perfect. However, the results lacked statistical significance.

Indeed, we have shown in chapter 6 that a perfect ROC curve for our machine learning model could be obtained with a 99.2 % probability if the test set only consisted in 450 galaxies. Therefore, the results and analysis presented in chapter 6 are not entirely robust. We however expect the deep learning model to be better than the simple machine learning models built earlier, at the expense of interpretability.

## 7.2 Future perspectives

We will finally share some future perspectives. The natural next step would be to search for lenses using machine learning models in a real dataset. This would give more hindsight on the limitations of such models.

It would also be instructive to perform a deeper and more rigorous comparison of deep learning and machine learning models. In particular, it is necessary to run EfficientNet on more samples. It would be interesting as well to run them both on real datasets and examine the strengths and weaknesses of each type of models with respect to the different galaxy types. We expect both types of models to perform less well on real datasets and

the gap between the two approaches to be reduced. Indeed, significant differences arise between a dataset of simulated lenses and a dataset of real lenses, most notably in terms of noise and structure of the images.

# Bibliography

[1] Bershady, M. A., Jangren, A., and Conselice, C. J. (2000). Structural and photometric classification of galaxies. i. calibration based on a nearby galaxy sample. *The Astronomical Journal*, 119(6):2645.

[2] Bertin, E. (2017). *SExtractor Documentation*.

[3] Binney, J. and Tremaine, S. (2011). *Galactic dynamics*, volume 13. Princeton university press.

[4] Birrer, S., Millon, M., Sluse, D., Shajib, A., Courbin, F., Koopmans, L., Suyu, S., and Treu, T. (2022). Time-delay cosmography: Measuring the hubble constant and other cosmological parameters with strong gravitational lensing. *arXiv preprint arXiv:2210.10833*.

[5] Bradley, L., Sipőcz, B., Robitaille, T., Tollerud, E., Vinícius, Z., Deil, C., Barbary, K., Wilson, T. J., Busko, I., Donath, A., Günther, H. M., Cara, M., Lim, P. L., Meßlinger, S., Burnett, Z., Conseil, S., Droettboom, M., Bostroem, A., Bray, E. M., Bratholm, L. A., Jamieson, W., Ginsburg, A., Barentsen, G., Craig, M., Pascual, S., Rathi, S., Perrin, M., Morris, B. M., and Perren, G. (2024). *astropy/photutils: 1.12.0*. Zenodo.

[6] Chollet, F. et al. (2015). Keras. `https://keras.io`.

[7] Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1.

[8] Delchambre, L., Krone-Martins, A., Wertz, O., Ducourant, C., Galluccio, L., Klüter, J., Mignard, F., Teixeira, R., Djorgovski, S., Stern, D., et al. (2019). Gaia gral: Gaia dr2 gravitational lens systems-iii. a systematic blind search for new lensed systems. *Astronomy & Astrophysics*, 622:A165.

[9] Eisner, N. L., Barragán, O., Lintott, C., Aigrain, S., Nicholson, B., Boyajian, T. S., Howell, S., Johnston, C., Lakeland, B., Miller, G., et al. (2021). Planet hunters TESS II:

findings from the first two years of TESS. *Monthly Notices of the Royal Astronomical Society*, 501(4):4669–4690.

[10] Ferrami, G. and Wyithe, S. (2024). A model for galaxy-galaxy strong lensing statistics in surveys. *arXiv preprint arXiv:2404.03143*.

[11] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

[12] Gini, C. (1921). Measurement of inequality of incomes. *The economic journal*, 31(121):124–125.

[13] Huchra, J., Gorenstein, M., Kent, S., Shapiro, I., Smith, G., Horine, E., and Perley, R. (1985). 2237+ 0305: A new and unusual gravitational lens. *Astronomical Journal (ISSN 0004-6256), vol. 90, May 1985, p. 691-696. Research supported by the Smithsonian Institution.*, 90:691–696.

[14] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer Nature.

[15] Khramtsov, V., Sergeyev, A., Spiniello, C., Tortora, C., Napolitano, N. R., Agnello, A., Getman, F., de Jong, J. T., Kuijken, K., Radovich, M., et al. (2019). Kids-squad-ii. machine learning selection of bright extragalactic objects to search for new gravitationally lensed quasars. *Astronomy & Astrophysics*, 632:A56.

[16] Kieser, R., Reynisson, P., and Mulligan, T. J. (2005). Definition of signal-to-noise ratio and its critical role in split-beam measurements. *ICES Journal of Marine Science*, 62(1):123–130.

[17] Krzanowski, W. J. and Hand, D. J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC.

[18] Laisney, C. (2023). *Innovative techniques to find strongly lensed systems*. Université de Liège, Liège, Belgique.

[19] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[20] Lemon, C., Courbin, F., More, A., Schechter, P., Cañameras, R., Delchambre, L., Leung, C., Shu, Y., Spiniello, C., Hezaveh, Y., et al. (2024). Searching for strong gravitational lenses. *Space Science Reviews*, 220(2):23.

[21] Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

[22] Lotz, J. M., Primack, J., and Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163.

[23] Louppe, G. (2024). *Deep Learning* [Lecture notes]. `https://github.com/glouppe/info8010-deep-learning?tab=readme-ov-file`. Université de Liège, Liège, Belgique.

[24] Marshall, P. J., Verma, A., More, A., Davis, C. P., More, S., Kapadia, A., Parrish, M., Snyder, C., Wilcox, J., Baeten, E., et al. (2016). Space warps–i. crowdsourcing the discovery of gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 455(2):1171–1190.

[25] Peng, C. Y., Ho, L. C., Impey, C. D., and Rix, H.-W. (2002). Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124(1):266.

[26] Peng, C. Y., Ho, L. C., Impey, C. D., and Rix, H.-W. (2010). Detailed decomposition of galaxy images. ii. beyond axisymmetric models. *The Astronomical Journal*, 139(6):2097.

[27] Petrosian, V. (1976). Surface brightness and evolution of galaxies. *Astrophysical Journal, vol. 209, Oct. 1, 1976, pt. 2, p. L1-L5.*, 209:L1–L5.

[28] Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine learning*, 42:203–231.

[29] Rodriguez-Gomez, V., Snyder, G. F., Lotz, J. M., Nelson, D., Pillepich, A., Springel, V., Genel, S., Weinberger, R., Tacchella, S., Pakmor, R., et al. (2019). The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to pan-STARRS observations. *Monthly Notices of the Royal Astronomical Society*, 483(3):4140–4159.

[30] Savary, E., Rojas, K., Maus, M., Clément, B., Courbin, F., Gavazzi, R., Chan, J., Lemon, C., Vernardos, G., Cañameras, R., et al. (2022). Strong lensing in unions: Toward a pipeline from discovery to modeling. *Astronomy & Astrophysics*, 666:A1.

[31] Schneider, P. (1984). The amplification caused by gravitational bending of light. *Astronomy and Astrophysics (ISSN 0004-6361), vol. 140, no. 1, Nov. 1984, p. 119-124.*, 140:119–124.

[32] Simpson, R., Page, K. R., and De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054.

[33] Sluse, D., Surdej, J., Claeskens, J.-F., Hutsemekers, D., Jean, C., Courbin, F., Nakos, T., Billeres, M., and Khmil, S. (2003). A quadruply imaged quasar with an optical Einstein ring candidate: 1rxs j113155. 4–123155. *Astronomy & Astrophysics*, 406(2):L43–L46.

[34] Snyder, G. F., Torrey, P., Lotz, J. M., Genel, S., McBride, C. K., Vogelsberger, M., Pillepich, A., Nelson, D., Sales, L. V., Sijacki, D., et al. (2015). Galaxy morphology and star formation in the Illustris simulation at z= 0. *Monthly Notices of the Royal Astronomical Society*, 454(2):1886–1908.

[35] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

[36] Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37.

[37] Walsh, D., Carswell, R. F., and Weymann, R. J. (1979). 0957+ 561 A, B: twin quasistellar objects or gravitational lens? *Nature*, 279(5712):381–384.

[38] Weinberg, S. (1972). *Gravitation and cosmology: principles and applications of the general theory of relativity*. OUP Oxford.

[39] Weinberg, S. (2008). *Cosmology*. OUP Oxford.

# Appendices

# Appendix A

# Morphological parameters of galaxies

We are now going to list all the parameters that can be computed thanks to statmorph [29]. These are

- The Petrosian radius [27] [22], $r_p$, is defined such that the ratio between the surface brightness of the image at $r_p$ and the mean surface brightness enclosed within a surface of radius $r_p$ is equal to 0.2

$$0.2 = \frac{\mu(r_p)}{\bar{\mu}(r < r_p)} \tag{A.1}$$

- The effective radius [3], $R_e$, is defined as the radius of the surface that contains 50 % of the light of the galaxy.

- The smoothness [22], $S$, defined by

$$S = \sum_{i,j} \frac{|I(i,j) - I_S(i,j)|}{|I(i,j)|} - B_S \tag{A.2}$$

  where $I_S(i,j)$ is the intensity of pixel $(i,j)$ smoothed by a box of width $0.25r_p$ and $B_S$ is the averaged smoothness of the background pixels.

- The signal to noise ratio [16] is defined to be the ration between the power of the useful signal and the power of the noisy part of the signal

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}} \tag{A.3}$$

- The coordinates of the geometrical center of the image (average of the positions of the pixels weighted by their intensity).

- The coordinates of the center of the image that minimizes the asymmetry.

- The ellipticity [3]

$$\epsilon = 1 - \frac{b}{a} \tag{A.4}$$

  where $a$ is the semi-major axis and $b$ the semi-minor axis of the galaxy. $a$ and $b$ can be calculated with one of the two centers mentioned above in this list.

- The elongation [2] defined as $\frac{a}{b}$.

- The orientation [2] is angle between $a$ and the $x$ axis of the image.

- The Gini-$M_{20}$ merger index [34] [29] is the position of a galaxy along a line in the Gini-$M_{20}$ plane that is perpendicular to the line

$$S(G, M_{20}) = 0.139M_{20} + 0.99G - 0.327 \tag{A.5}$$

  with the origin in (0.565, -1.679).