# Galaxy Morphology Classifier Proposal

David Thornton

October 2025

**Abstract**

This proposal includes a brief overview of galaxy morphology classifications and machine learning concepts. The proposed project is a galaxy classifications system which has a heavy focus on model trustability, through explainability techniques and the implementation of uncertainty quantification. The project will also include an in-depth literature review on relevant concepts and methods of implementing these concepts in the current literature. The project will be developed using the crisp-DM methodology for machine learning.

# 1 Introduction

Galaxies can be classified in many different ways with different techniques. One of the most common ways is to classify them by their morphology, their physical components and structure. Classifying galaxies this way can provide insight into their formation and life cycle, as well as the general composition of the universe. However, there are a massive number of galaxies in our sky, many of which have already been captured as images. The time and resources it would take for these images to be classified by people would be truly immense. Another, better alternative to this is to train a machine learning model to carry out this task. However, machine learning models aren't as trustworthy as people and for scientific applications such as research of galaxy systems this is an issue. This project aims to solve these issues using modern techniques.

To understand what this project involves, there are two main concepts that need to be understood, galaxy morphology classification and machine learning.

## 1.1 Galaxy Morphology

There are many ways to classify galaxy types, including luminosity (brightness), mass, spin, structure, and distance from us to name a few (Roberts and Haynes 1994). However, not all of these can be discovered from images alone, so, for the purposes of this project, we will focus on the morphology/structure of the galaxy types. Galaxy morphologies are often categorized based on Hubble Sequencing (Conselice 2006) into five main categories, each with its own subcategories, as follows.

### 1.1.1 Spiral

Spiral galaxies can be identified by their flat, disk-like structure, made up of arms that spiral out from a central bulge, an area of densely packed stars in the center of the galaxy. These galaxies are typically younger, with young blue stars along their disks. (Conselice 2006)

Figure 1: Spiral Image (Kazmierczak n.d.)

This galaxy type can be further broken down into subcategories:

- Bulge type: Spiral galaxies can either have a bar-shaped bulge or a circular bulge
- Bulge size: How big the bulge is by comparison to the arms
- Number of arms: These galaxies can have one to 4 arms or sometimes even more
- Tightness of arms: How tightly wound the arms are.

(Baillard et al. 2011)

### 1.1.2 Lenticular

Lenticular galaxies are like spiral galaxies, featuring a central bulge and flat disk shape. However, they do not feature strong/noticeable arms, star formation, or large amounts of dust or gas. These are believed to form from spiral galaxies due to effects such as a lack of gas. Due to the lack of arms featured in this galaxy, their only subcategories (which we are concerned with) are the shape and size of the bulge. (Conselice 2006)

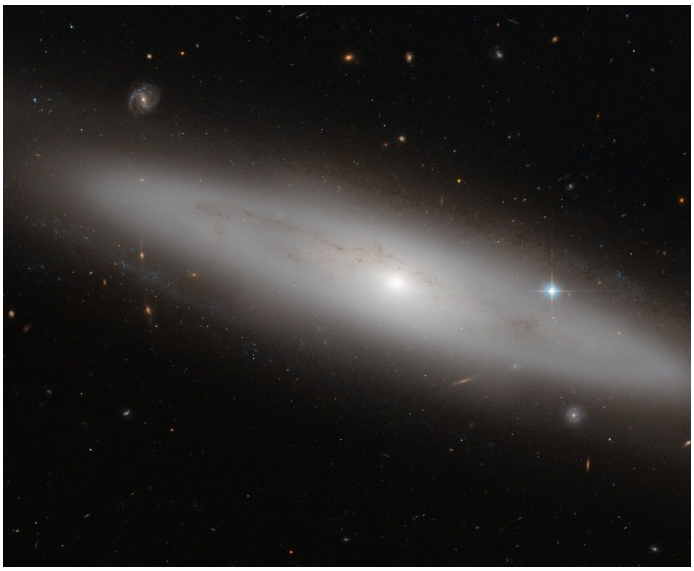

Figure 2: Lenticular Image (Kazmierczak n.d.)

#### 1.1.3 Elliptical

Elliptical galaxies can be identified by their lack of a disk structure, often being described as a "smooth" shape. These galaxies are believed to have formed from older galaxies that have collided with each other, resulting in their loss of a defined structure. Most of these galaxies are older, being formed from earlier galaxies, and hence have mostly dimmer red/yellow stars. (Conselice 2006)

Figure 3: Elliptical Image (Kazmierczak n.d.)

These galaxies can be subcategorised by their shape, including:

- Round: Appearing spherical
- In between or ovular: Having an oval shape
- Cigar shaped: Having a longer, narrower shape (like a cigar)

(Baillard et al. 2011)

#### 1.1.4 Irregular

Irregular galaxies are generally smaller and younger than other galaxy types mentioned previously, in their initial stages of formation. These galaxies don't have any defining shape or structure, and as a result aren't often subcategorised (Conselice 2006).

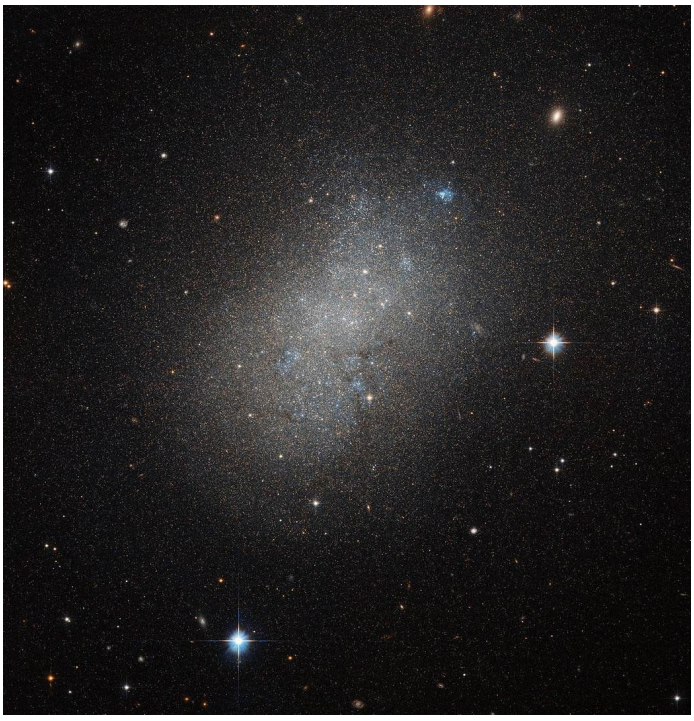

Figure 4: Irregular Image (Kazmierczak n.d.)

#### 1.1.5 Interacting Galaxies

These are galaxies that are affected by other interstellar objects, most commonly other galaxies, merging/interacting with each other. This is by far the least common galaxy type, since galaxies don't spend a lot of time in this state (Conselice 2006).

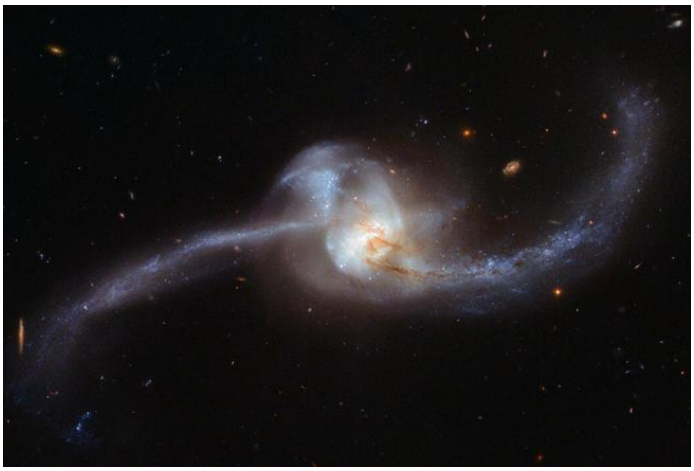

Figure 5: Merging Image (Kazmierczak n.d.)

A common sixth categorisation is often included for "on-edge galaxies". This is because when a disk-shaped galaxy (spiral or lenticular) is viewed from the edge of the disk, it's very difficult or even impossible to identify whether the galaxy features arms or not.

## 1.2 Machine Learning

Machine learning is a branch of data science that deals with systems that can improve automatically through experience, or training. Machine learning models are trained using data, split into features (what we tell the model) and labels (what we want the model to predict). The model makes its own decisions on how to make predictions, changing its parameters throughout training iterations to become more accurate at predicting the correct label. The model can then be used in practical applications to make predictions on unlabelled data. (Jordan and Mitchell 2015)

### 1.2.1 Computer Vision

Computer vision is a field of machine learning that allows machines to interpret and understand content in images or videos. This is done by training a model to extract and process features or structural components from visual data such as edges, shapes, textures, and patterns. Convolutional neural networks are often used in this process. These features are then used in other layers or models to perform linear regression, clustering or classification tasks, which is the purpose of our model. (Szeliski 2022)

### 1.2.2 Uncertainty Quantification

Uncertainty quantification is a machine learning concept which involves training a model to not only make predictions on data, but to also describe how sure of the predictions it is. There are different methods of implementing this, many of which involve having the model return percentage representations of how likely a classification is to be true, rather than simply returning the classification itself. (Abdar et al. 2021)

### 1.2.3 Machine Learning Explainability

Machine learning explainability is the process of explaining and understanding how a model makes predictions. In some models, this process is more straightforward. In decision tree models for example, we can plainly see how the model makes decisions at each level of the tree. These are called white box models. Black box models, however, don't produce direct explanations of how they make decisions, which can make them less trustworthy. To eliminate this issue, we can implement external methods to help understand how the model makes its decisions, such as by analysing how important the features that it is given are to making a prediction. (Linardatos, Papastefanopoulos, and Kotsiantis 2020)

### 1.2.4 Hierarchical Model

Hierarchical models often contain sub models which perform different tasks. Hierarchical models start with a core model, which will classify higher level, more general labels. Based on this initial classification, feature data will then be passed into different sub models, which will carry out classifications on subcategories associated with a given category. (Serrano-Pérez, Díaz Hernández, and Sucar 2024)

# 2 Justification and Motivation

In most cases, galaxies can be identified by people with relatively high accuracy. However, there are extremely large datasets of galaxy images, some having hundreds of millions of images, many of which are unlabelled. Categorising these images will aid in researching galaxy morphologies and the composition of the universe. However, there are too many images to be reliably labelled by people, and the number of these images is only increasing. The implementation of machine learning in galaxy classification can aid in resolving this issue. By training machine learning models to classify galaxy images the time it takes to process these massive datasets can be greatly reduced. Additionally, machine learning models can often find associations people can't, and understand some datasets better than people can, especially in multidimensional data. This may also lead to machine learning models being more accurate in some cases.

# 3 Aim, Objectives and Research Questions

## 3.1 Main Project Aim

The aim of the project is to develop a hierarchical galaxy morphology classification model that classifies galaxies into the basic Hubble types, then refining classifications into subcategories within the Hubble sequence. The model will implement level specific explainability and uncertainty quantification for better interpretability and enhanced trustability.

## 3.2 Research Questions

- Can the addition of uncertainty quantification in galaxy morphology classification models increase model accuracy and trustability?
- How effective is implementing explainability in galaxy classification systems at informing better decision making within model design?
- How effective are hierarchical models at classifying data with a hierarchical classification structure, such as galaxy types, when compared to flat classification models?

By adding uncertainty quantification to a classification model, only trusting predictions with higher certainty and having predictions with lower certainty be reassessed, can a model become (artificially) more accurate? Given this is true, this artificial increase in accuracy will make the model more trustworthy, and hence more useful in practical applications.

Regardless of whether the application of uncertainty quantification increases accuracy and trustability as described, the model will still need to be rigorously tested to ensure it is as accurate as possible. One of the most important components of enhancing model accuracy is providing the model with relevant features. However, finding good features is a significant challenge for any model, especially in computer vision systems.

Can testing a model with explainability measures such as SHAP provide the necessary insight needed to ensure a model is given relevant features? If this is the case, there should be a noticeable increase in model accuracy when features with high explainability scores are used. Additionally, further insight into how the model makes its predictions will greatly increase model trustability.

Flat classification models are often used to classify data, even in cases when the data labels follow a hierarchical structure, such as in galaxy classifications following Hubble Sequencing. This is done even though models can be structured with the same hierarchical structures as the data they are classifying. How do hierarchical models compare against flat models when classifying data with a hierarchical structure?

### 3.3 Objectives

- Conduct research on relevant concepts in machine learning and galaxy morphology.
- Research current literature on galaxy classification systems.
- Analyse possible datasets.
- Train baseline model on single dataset.
- Refine feature extraction process.
- Test different models.
- Compare refined models to baseline and other published models.

## 4 Proposed Methodologies

Throughout project development, the CRISP-DM framework for machine learning will be followed, which can be broken down into the following steps.

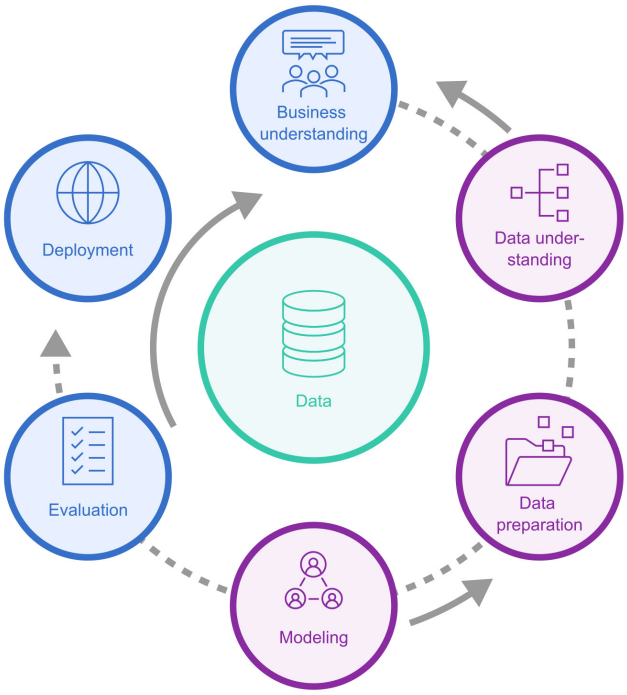


Figure 6: Crisp-DM life cycle (Institute n.d.)

- **Business Understanding:** Researching current literature on relevant concepts, such as galaxy morphologies, machine learning and how machine learning can be used in galaxy classification.
- **Data Understanding:** Using business understanding to select a dataset most suited to the project aim and taking steps to further understand it.
- **Data Preparation:** Feature engineering the dataset using our understanding of the project aim and the dataset.
- **Modelling:** Design and train different models with features created in the data preparation phase.
- **Evaluation:** Comparing the performance of different models and techniques to maximise the model's overall performance.
- **Deployment:** Using the model to predict galaxy classifications and (possibly) deploy an application for others to use the model.

As this is a research based project with some elements of experimentation, the process will likely not be as linear as described, often returning to research for a better understanding of concepts, experimenting with different datasets and data preparation techniques and building several different models over the course of the project. Despite this, most of the research for this project will be carried out as part of a literature review in the earlier stages of development, covering core concepts and techniques in the current literature. Soon after, a basic model will be developed, which will act as a baseline to compare later iterations to.

A literature review will be carried out to advance the author's understanding of the current techniques involved in galaxy classification and general image classification systems. The author will also research current machine learning explainability and uncertainty quantification techniques and evaluate which would be suitable for testing in the project. The author will research available datasets that could be used.

The author will assess the quality and usability of the available datasets and evaluate which would be most suitable for the project. The author will further analyse the chosen datasets to gain a better understanding of its content and how it should specifically be used.

Data will be prepared for modelling after sufficient understanding of the data has been achieved and at the start of every new model iteration. The data will be processed using techniques discovered during the data understanding phase and in previous model iterations.

Once data has been prepared it will be used in model training. The first models will be designed based on findings from the literature review and data understanding phases. Subsequent models will also be designed with understandings gained through evaluation of previous models.

After each model iteration has been designed and trained, the model's performance will be evaluated. Specifically, the model will be assessed on its performance compared to the base model and other studies. How the features and model design affected the performance will also be assessed.

Once a model meeting the project requirements has been developed it will be used to classify unlabelled datasets to evaluate how its classifications align with current understandings of galaxy morphologies. A website may also be published to give the public access to the model if there is sufficient time to do so before the project deadline.

Several frameworks and technologies will be utilized throughout the development of this project. Python will be used as the main programming language, along with python libraries such as NumPy, Pandas, Matplotlib, Scikit-learn, PyTorch, TensorFlow and others. Jupyter Notebook will be used as the main IDE. It has been specifically designed for data science projects and is easily integrated with cloud-based architectures such as Google Colab. For these reasons it is the most suitable option for a project such as this. The project will be completed with both personal hardware and cloud computing resources.

# 5 Contributions and Benefits

The issue with using machine learning for galaxy classification is the same issue with most machine learning tasks, a lack of trustability. Even the most well trained and accurate models can often be untrustworthy. This is partially because they only give predictions on what they think the classification is, without any insight as to how sure they are of the prediction. The addition of uncertainty quantification will increase the trustability of the model by giving an understanding of how sure the model is of its predictions. This will increase the reliability of the model. However, due to the additional computational requirements and complexity, implementing uncertainty quantification isn't commonly done in galaxy classification models. The addition of uncertainty quantification to galaxy classification will increase the trustability of the model.

Another factor limiting model trustability is a lack of understanding of how the model is making its predictions. This is particularly an issue in black box models such as convolutional neural networks, which are commonly used in image classification, including galaxy image classification. The implementation of machine learning explainability techniques can increase model trustability by providing insight to how these predictions are made. Additionally, explainability techniques can help to identify relevant features and their importance on decision making within models, resulting in better feature selections and model performance. The addition of explainability measures in our model will increase model trustability and understanding of features.

The implementation of uncertainty quantification and explainability techniques in the model will increase the trustability of the model's predictions and understanding of model behaviours.

# 6 Scope and Limitations

The main limitations of this project will be the time constraints and technological requirements. To alleviate these constraints, financial decisions may come into effect.

Creating a classification model with many labels will need a lot of iterations and experimentation to achieve satisfactory results, however, given the time constraints of 30 weeks to complete the project, it may be difficult to achieve the desired results. Additionally, model training with large datasets and image processing, which will need to be completed, takes a substantial amount of time, further adding to time constraints. Large amounts of data processing also means a considerable amount of compute will be needed for the project. This creates substantial hardware requirements for the project. The implementation of cloud computing solutions could reduce both the computational constraints, with more substantial processing power, and as a result the time constraints since this processing power will also aid in speeding up computational tasks. However, many of these cloud computing solutions require paid subscriptions, replacing time and technical constraints with greater financial burden. Throughout the course of the project, the impact of each of these limitations will need to be assessed to evaluate what actions to take.

# 7 Ethics

There are no considerable ethical concerns associated with this project, as all datasets under consideration are publicly available. Additionally, no personal data will need to be collected or used in the completion of the project. The main ethical concern for the project is to ensure authenticity of work and accrediting the work of others used throughout.

# 8 Feasibility

Galaxy classification using machine learning has been heavily researched for several decades and has been proven to be a reliable method of classification. There are also many studies on uncertainty quantification and methods for machine learning explainability, in both general applications and computer vision specifically.

There are many large, well structured, public datasets of galaxy images that can be used for the purposes of this project. Open source nonprofit projects like Galaxy Zoo have organised large datasets to be labelled by the public, such as the galaxy zoo 2 dataset, and have also taken extra steps to account for unreliable and biased classifications. Some of these projects were developed specifically with machine learning experimentation in mind, such as the Galaxy10 dataset. Other projects, such as EFIGI, have had professionals in the field of astronomy classify galaxy images, providing datasets with highly reliable labelling.

The author of this project is familiar with the technologies and concepts involved and has a basic understanding of techniques and methodologies commonly used in similar projects. However, further research into common techniques will need to be conducted to complete the project to a suitable standard. For this reason, the author will conduct further research on common techniques and available resources in the form of a literature review and will converse with professionals in the field for further insight and guidance.

# 9 Expected Results

The expected result of this project is that a galaxy classification system will have been developed which is not only accurate in its predictions but is highly trustable.

The model is expected to be accurate on testing data and to provide insight on unlabelled data which correlates to the current scientific understanding of galaxy morphologies and the composition of the universe.

The model is expected to provide better insight to large, unlabelled datasets with more reliability and trustability.

The model is expected to provide valuable insights into how machine learning models identify galaxy morphologies and what impacts a model's confidence in its predictions.

The project is expected to accurately assess what the most efficient techniques are for galaxy image classification systems. It is expected to do this by training several iterations, using different data processing techniques and model configurations to find the one that is the most accurate.

The project is expected to provide valuable insight into the effectiveness of machine learning explainability and uncertainty quantification techniques. It is expected to do this by comparing different techniques and assessing what value each brings.

# 10 Conclusion

The aim of this project is to develop a hierarchical galaxy classification system using explainability and uncertainty quantification techniques. The model will be developed with both accuracy and trustability in mind. The project will be developed using the crisp-DM methodology, involving stages for business and data understanding, data preparation, modelling, evaluation and deployment. An in-depth literature review will be conducted to compare current techniques and improve the author's understanding of machine learning concepts before and throughout modelling. The model will be developed using Python, Jupyter notebook and other libraries. Limitations in development include the short project time frame and heavy technical requirements.

# 11 Project Plan

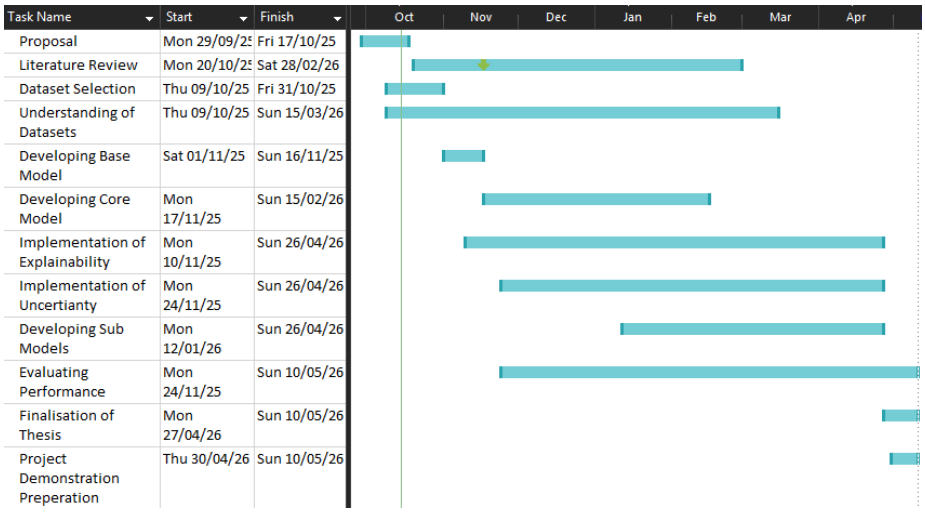


Figure 7: Project Gantt Chart

The first step of the project is to write the proposal document. This will also require background research and decisions on methodologies and technologies used throughout the project. Project aims, objectives and research questions will also be defined during this step. The proposal will be written from the start of the project until the proposal deadline.

The literature review will take place throughout the course of the project and will be updated as more research is done or whenever further investigation is needed. However, the first official version of the literature review will need to be completed by the 16th of November to be submitted for the assignment related to it.

Before training the model, datasets will need to be evaluated to choose the most suitable for the project. Once dataset(s) have been chosen, further analysis will need to be conducted to ensure meaningful features and accurate labels are used. Understanding of the data will be developed throughout the development of the model.

A base model will be trained early in the project, to be used as a base for further models to be compared to and to gain a better understanding of the data and techniques involved.

After the base model, a core model will be developed. This model will classify higher level galaxy classifications. Once the base model has reached a sufficient standard, sub models will be developed to identify subcategories of the galaxy types classified by the main model. The core model may be developed further alongside the sub models if necessary.

Explainability measures will be implemented throughout the project devel-

opment, starting in the base model to gain a better understanding of the data. Uncertainty quantification will be implemented in the core and sub models in the beginning of their development but likely won't be featured in the base model.

Model performance will be evaluated throughout development at the end of each model iteration. A more structured evaluation will take place at the end of the model's development to provide performance metrics for the thesis document.

The project thesis will be written/contributed to throughout the project development. A two-week period will be given to finalising the project thesis before the project deadline. The project demonstration will be made and rehearsed alongside the finalisation of the project thesis.

# References


Abdar, Moloud et al. (2021). "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information fusion* 76, pp. 243–297.

Baillard, Anthony et al. (2011). "The EFIGI catalogue of 4458 nearby galaxies with detailed morphology". In: *Astronomy & Astrophysics* 532, A74.

Conselice, Christopher J (2006). "The fundamental properties of galaxies and a new galaxy classification system". In: *Monthly Notices of the Royal Astronomical Society* 373.4, pp. 1389–1408.

Institute, Fraunhofer (n.d.). *Implementation strategy for a data-mining project with CRISP-DM in surface technology*. URL: `https://www.ist.fraunhofer.de/en/expertise/simulation-digital-services/data-acquisition-model-based-process-optimization/crisp-dm-surface-technology.html`.

Jordan, Michael I and Tom M Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245, pp. 255–260.

Kazmierczak, Jeanette (n.d.). *Galaxy Types*. URL: `https://science.nasa.gov/universe/galaxies/types/`. (accessed: 22.10.2024).

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1, p. 18.

Roberts, Morton S and Martha P Haynes (1994). "Physical parameters along the Hubble sequence". In: *Annual Review of Astronomy and Astrophysics, Volume 32, 1994, pp. 115-152.* 32, pp. 115–152.

Serrano-Pérez, Jonathan, Raquel Díaz Hernández, and L Enrique Sucar (2024). "Bayesian and convolutional networks for hierarchical morphological classification of galaxies". In: *Experimental Astronomy* 58.2, p. 5.

Szeliski, Richard (2022). *Computer vision: algorithms and applications*. Springer Nature.