Anomaly Detection on Gaia Data Using Diffusion Model

Kelvin H. M. Chan

Department of Physics

McGill University

Montréal, Québec, Canada

April 15, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

The Gaia space observatory has shown unprecedented ability to perform astrometric and photometric measurements of the luminous objects in the Milky Way. Benefiting from the amount and precision of measurements and the number of accurately measured features present in the Gaia dataset, looking for anomalies in the dataset may lead to the discovery of unexplored physics. Recent rapid development in deep learning allows the capture of complex patterns across multiple applications. In particular, generative modelling has proven to be a powerful method to detect high-order anomalies that the human eye cannot detect. In this thesis, we employ a state-of-the-art diffusion model, Diffusion Time Estimation, on Gaia's second data release to look for point anomalies. Then, we will explore the nature of the anomalies and look at why they are classified as anomalies.

Abrégé

Le télescope spatial Gaia a démontré une capacité sans précédent à effectuer des mesures astrométriques et photométriques des objets lumineux dans la Voie lactée. Grâce à la quantité et à la précision de leur mesure, ainsi que du nombre de caractéristiques précisément mesurées dans l'ensemble de données Gaia, la détection d'anomalies dans cette population d'objects pourrait conduire à la découverte de nouvelle physique. Le développement rapide de l'apprentissage profond au cours des dernières année permet la capture de correlations complexes à travers de multiples applications. En particulier, les modèles génératifs se sont avérées être une méthode puissante pour détecter des anomalies de haut niveau qui ne peuvent pas être détectées par l'œil humain. Dans ce mémoire, nous utilisons un modèle de diffusion à la pointe de la technologie, l'Estimation du Temps de Diffusion, sur la deuxième publication des données de Gaia pour rechercher des anomalies ponctuelles dans les données. Ensuite, nous explorerons la nature des anomalies et examinerons pourquoi elles sont classées comme telles.

Acknowledgements

First and most importantly, I would like to thank my supervisor, Katelin Schutz, for her guidance and support during my time at McGill. This project would not be possible without the knowledge and patience of her. I would also like to thank the members of our research group, and the collaborators in the computer science department for all the discussion and suggestions.

Contents

Abstract i

Abrégé ii

Acknowledgements iii

List of Figures vii

List of Tables viii

1 Introduction **1**

1.1 Thesis objective 1

1.2 Thesis organization 3

2 Gaia and dataset **4**

2.1 Gaia 4

2.1.1 A brief review of astrometry 5

2.1.2 Gaia 6

2.1.3 Objectives of Gaia 8

2.2 Dataset 11
2.2.1 Description of data 11
2.2.2 Data preprocessing 12

3 Anomaly detection **15**

3.1 Background 16
3.2 Review of anomaly detection on Gaia dataset 18
3.3 Diffusion Time Estimation 21
3.4 Experiments 25

4 Results and discussion **32**

4.1 Classifying anomalies 32
4.2 Analysis 35
4.2.1 Unbound stars 37
4.2.2 Extinction 40

5 Conclusion **43**

List of Figures

3.1 H-R diagram for 50000 randomly sampled objects from Gaia DR2. 26

3.2 Above: H-R diagram with artificial anomalies planted, with standardization as data preprocessing. Below: The same as above, but with quantile transform as data preprocessing. 29

3.3 Distribution of scaled luminosity using standardization versus quantile transform as feature scaling. 30

4.1 Distribution of anomaly score of the input data determined by the 2 runs of the DTE model. 33

4.2 Distribution of anomaly score of the input data determined by KNN. 34

4.3 Pairwise plot for the input parameters. 36

4.4 3-dimensional quiver plot for the unbound stars. 39

4.5 The colour excess of anomalous data compared to normal data. Above: the colour excess found by Apsis in Gaia DR2. Below: the colour excess found using `dustmap`. 41

List of Tables

4.1 Gaia DR2 source ID, total velocity to escape velocity ratio, and anomaly scores found by the 3 algorithms for the unbound stars. 38

List of Acronyms

DDPM	denoising diffusion probabilistic models.
Dec	declination.
DGM	deep generative model.
DL	deep learning.
DTE	diffusion time estimation.
GAN	generative adversarial network.
GC	Galactic Centre.
KNN	K-nearest neighbours.
ML	machine learning.
MW	Milky Way.
RA	right ascension.

Chapter 1

Introduction

1.1 Thesis objective

Astrometry is a branch of science that measures the positions and kinematics of celestial objects (1). It is essential to understanding the Milky Way (MW) since we can know the structure and evolution of the MW from the measurement of the 6D phase space of the luminous objects.

Gaia is a space observatory designed to measure the full 6D astrometry (3D spatial and 3D velocity) of the objects in the MW (2; 3). The volume of data generated by this observatory is unprecedented due to its ability to measure down to a magnitude of 20. Besides, compared to its predecessors, the improvement in accuracy of Gaia is substantial: it is expected to measure more than 1 billion objects with a precision of 10 μas, mapping 1% of stars in the

MW. With its ability, the measurements of Gaia can improve our current understanding of the MW. In particular, since Gaia can measure a significant amount of objects in the MW, this sampling provides an opportunity to find any object that deviates from the majority, known as point anomaly detection.

The detection of point anomaly is usually performed by machine learning (ML) algorithms such as K-nearest neighbours and clustering (4). However, with advances in deep learning (DL) in recent years, we can now make use of DL algorithms to detect complex anomalies. With more layers than traditional ML algorithms, DL algorithms possess a much larger number of parameters, from which they can learn expressive representations from complex data (5; 6). In particular, diffusion models (7) add noise to the input data, then learn to generate samples that are as close to the original data as possible. Anomalies can be found if the generated sample and original data have a large distance in the input space.

In this thesis, we employ a state-of-the-art diffusion model called diffusion time estimation (DTE) (8) to detect point anomalies in Gaia data. DTE adds noise to the data at each time step, but instead of learning to generate data and calculate the distance between the generated and input sample, we treat the anomalies as noise, so the model can simply learn the diffusion time step, which we can think of as learning the noisiness of each data. The authors show that DTE maintains high accuracy in identifying anomalies compared to other models while achieving high efficiency in terms of training time. From the result of the model, we analyse why the model considers the anomalies as anomalous.

1.2 Thesis organization

The thesis is organised as follows. In Chapter 2, we provide a review on Gaia and its dataset. In particular, we give an overview of astrometry, the improvement of Gaia and its objective, as well as describe the dataset and how we preprocess it. Chapter 3 reviews the methodology of the thesis. We review the current effort to detect anomalies in the Gaia dataset, motivate the application of the diffusion model to the Gaia dataset, and then test the algorithm by planting artificial anomalies. In Chapter 4, we present the results. We describe how we classify anomalies from the result of the model, and then analyse the nature of the anomalies. Finally, we conclude in Chapter 5.

Chapter 2

Gaia and dataset

2.1 Gaia

Gaia is a space observatory of the European Space Agency, surveying the MW since its launch in 2013 (3). It is designed to make unprecedentedly accurate astrometric measurements of point-like luminous objects down to a magnitude of 20, mostly stars in the MW. However, in addition to astrometry, high-quality multicolour photometry and spectroscopy measurements can also be obtained. The primary objective of the mission is to understand the evolution and structure of the MW.

2.1.1 A brief review of astrometry

Astrometry involves the accurate measurement of the positions and motions of celestial objects. The position of a large subset of celestial bodies can provide insight into the content and distribution of matter in the wider physical system. Meanwhile, analyzing the kinematics gives information about the gravitational field and the orbits of the bodies. These combine to provide a model-independent way to understand the structure and evolution of the wider system that these celestial bodies belong to.

Observing and recording the positions of astrophysical objects is accessible even to ancient people, thus astrometry naturally has a long history (1). Ancient Greek Hipparchus compiled the first star catalogue with 1000 stars at 1-degree accuracy in the second century BCE. Ever since then, astrometric accuracy roughly follows a logarithmic improvement (1) until the twentieth century, when ground-based astrometry hit its limit (1).

The Earth's turbulent atmosphere produces flickering effects, which make the measurements error-prone. Besides, under Earth's gravity, the telescope's weight will cause distortion, introducing additional noise. In addition, the ground-based telescope can only observe a part of the sky at a time. The proposal of space-based astrometric measurement in the 1960s solved all these problems by bringing the telescope above the atmosphere and beyond Earth's gravity.

Before Gaia, Hipparcos was the first and only space astrometry experiment, operating between 1989 and 1993. It produced two catalogues of stars that differ in accuracy. The

Hipparcos Catalogue (9) is higher in astrometric accuracy, it contains positions, proper motions and photometry in two bands for 120,000 stars brighter than magnitude 9 at an accuracy of $\mathcal{O}(1)$ milliarcsec (mas). The Tycho Catalogue (10; 11) is less precise and hence it allows more stars to be included in the catalogue. The Tycho Catalogue contains positions, proper motions and photometric data in two bands for 2.5 million stars down to magnitude 11.5 at an accuracy of $\mathcal{O}(10)$ mas.

2.1.2 Gaia

As the successor to the Hipparcos mission, Gaia shows substantial improvements in terms of both the number of measurements and their accuracy. Gaia is expected to measure 1 billion objects ($\sim$ 1% of the stars in the MW) down to magnitude 20 at a precision of $\mathcal{O}(10)$ μas throughout the MW (3). Two identical telescopes separated by 106.5° measure light within a wavelength of 330 - 1050 nm (G band; G stands for Gaia).

RA and Dec can be measured simply by recording the position of the point source in the sky. However, the measurement of parallax and proper motion is more complicated, since a single measurement is a combined effect of parallax and proper motion. Parallax is the change of apparent position of nearby stars compared to far away background stars that appear motionless, due to Earth's orbit around the Sun. We can measure the distance of nearby stars by simple trigonometry. On the other hand, proper motion is due to the motion of the stars, also compared against background stars. Since parallax motion repeats

annually, while proper motion is a continuous linear effect, years of repeated observation allow for resolving their degeneracy.

Besides matter distribution and kinematics, the photometry of Gaia also plays a significant role in advancing our understanding of the MW by providing measurements in different colour bands. In addition to measuring light flux in the G band, Gaia is also equipped with low-resolution spectrophotometers in the BP-band (blue photometer, 330 - 680 nm) and RP-band (red photometer, 640 - 1050 nm).

Lastly, Gaia is also equipped with a medium-resolution radial-velocity spectrometer (RVS) in the wavelength of 845 - 872 nm, which, as the name suggests, measures the radial velocity of the objects by their spectrum. Due to the Doppler effect, the radial motion of objects will lead to a blueshift or redshift of their spectrum. By comparing with known spectral lines, radial velocity can be obtained. This is the first time a space astrometry experiment is able to measure radial velocity.

From Gaia's astrometry, photometry, and spectrometry, the Gaia astrophysical parameters inference system (Apsis) (12) can then infer astrophysical parameters essential to characterize celestial objects, including effective temperature T_{eff}, extinction A_G, colour excess $E(G_{BP} - GRP)$, radius, luminosity and metallicity. Apsis uses supervised machine learning to infer the parameters, meaning that a machine learning model is trained on a set of data where the parameters are known, and then applied to infer the parameters of the actual observation data.

2.1.3 Objectives of Gaia

Despite the long history of astrometry, a large part of our Galaxy remains unknown to us. In the following, we summarize the most important scientific case of Gaia (2; 3), followed by the major discoveries of Gaia since its launch.

1. **Structure, dynamics and evolution of the Galaxy**: Gaia is specifically designed to study these aspects of our Galaxy. By studying the 6D phase space of the stars, meaning the 3-dimension position and 3-dimension velocity, we can study the dynamics and structure of the MW. The observation of the 3D position of the brightest light sources can give rise to the distribution of luminous objects in the Galaxy. Moreover, by tracking the motions of stars, we can model the gravitational potential and thus access the matter distribution of the Galaxy. This is important since most of the matter in the Galaxy is not luminous. The matter content in our Galaxy is made up of mostly dark matter and a small amount of luminous baryonic matter. Since dark matter cannot be directly probed by telescopes, studying luminous objects in the Galaxy is the only way we can understand the structure of the Galaxy.

   Historically, one of the evidence for the discovery of dark matter is the galaxy rotation curve (13), which is obtained by combining the position and kinematics of stars, plotting the velocity against the distance to the GC. By postulating the existence of dark matter, the discrepancy between the observation and theoretical rotation curve with just the luminous objects is explained. With the ability to

measure the 6D phase space of dimmer stars, Gaia data can be used to perform more detailed analyses for the luminous and dark matter distribution of the MW.

Also, with the kinematics and distribution of the luminous matter and dark matter, we can understand more about the formation and evolution of our Galaxy. The paradigm of galaxy formation is the hierarchical formation model (14), where the galaxies are formed through mergers of smaller galaxies. Under this formation process, different stages are likely to leave imprints in the structure of MW, which we can observe with the large number of measurements by Gaia. For example, stellar streams are thin ribbon-like arcs of stars orbiting the galactic centre, caused by the gravitational tidal disruption of dwarf galaxies and globular clusters by the host (15; 16). As a result, the discovery of stellar streams using Gaia data can provide information on the merging history of the MW. Moreover, the density fluctuation of stellar streams, resulting from the encounters between streams and dark matter subhalos, can also help probe the dark matter mass (17), due to the fact that if dark matter is less massive, there will be fewer subhalos. Understanding the nature of dark matter will also help in understanding the formation of galaxies since dark matter constitutes a large portion of galaxies.

2. **Star formation history of the MW**: Stellar luminosity, obtained by measured flux and parallax distance, and metallicity, estimated using RVS spectra by Apsis, can provide an accurate estimation of the age of stars. Older stars are more metallic due to the longer fusion time, while also becoming brighter and redder. Combining the stellar

age with the information on the structure and dynamics of the MW, Gaia makes it possible to deduce the star formation history of the Galaxy.

3. **Stellar physics and evolution**: One of the strengths of Gaia will be its parallax measurements. Along with Gaia's photometry, it is possible to derive a high-quality colour-magnitude diagram (H-R diagram). This can make us better understand stellar evolution (see Section 3.4 for more details of the H-R diagram) since the number of objects measured will cover most phases of stellar evolution. For example, Gaia will provide parallax for the faintest white dwarf for the first time.

Since the launch of Gaia, the data has enabled researchers to make groundbreaking discoveries. Here we review some of the most important ones.

Regarding the evolution of the MW, with Gaia DR2, (18) discover that the MW merged with another galaxy, referred to as Gaia-Enceladus, during its early stage. The stars that originated from Gaia-Enceladus cover the full sky, with motions very different from most stars in the MW. (19) shows that there are two phases in the MW history. The older phase started just 0.8 billion years after the Big Bang when the thick disk, which is a galactic structure filled with metal-poor stars, was formed. Two billion years later, the Gaia-Enceladus merger triggered the second phase of galactic formation when the thin disk with metal-rich stars was formed.

For the galactic structure, using Gaia DR2 data, (20) show that the warp of the MW is caused by a recent or ongoing encounter with a satellite galaxy. (21) argue that the

encounter with the Sagittarius dwarf galaxy triggered the star formation in the MW. (22) identify for the first time a snail-shell-like substructure in the phase space, further confirming the perturbation from the collision with a satellite galaxy.

For stellar physics, (23) report that the H-R diagram obtained from Gaia data shows for the first time there is a split of the white dwarf sequence into hydrogen white dwarfs and helium white dwarfs. (24) observes a gap near magnitude 10 in the main sequence on the H-R diagram with Gaia DR2 data.

These discoveries show that the high-quality Gaia data enables the astrophysics community to make a lot of progress in understanding the Galaxy in terms of its evolution, structure and composition.

2.2 Dataset

2.2.1 Description of data

The latest Gaia data release is the third data release (DR3), which contains the full astrometric solution for ∼1.5 billion sources at magnitude 3>G>21 (25). In this thesis, we use the data from Gaia Data Release 2 (DR2) (26) which contains the full astrometric solution for 1.3 billion sources. Although DR3 contains more objects, in particular, radial velocity data is more complete, a primitive study using DR2 data is beneficial to prove the effectiveness of the method, and also to focus on a few anomalies to investigate. The

astrometric parameters include the celestial coordinate right ascension (RA) α and declination (Dec) δ, parallax, and proper motion along the direction of the two celestial coordinates. It also contains radial velocities for 7.2 million sources, G-band photometry G for 1.7 billion sources, and BP-band G_{BP} and RP-band G_{RP} photometry for 1.3 billion sources. BP and RP spectra and RVS spectra are not released in DR2 but are released in DR3. Photometric time series and radial velocity time series are also not present until DR3.

2.2.2 Data preprocessing

The full dataset has 94 columns of features, in addition to the above 9 features, it also includes errors of the above measurements and correlation between pairs of the above measurements. 5 astrophysics parameters inferred using only photometry and parallax by Apsis are also present, including stellar effective temperature, stellar radius, stellar luminosity, line-of-sight extinction, and line-of-sight reddening.

To select only the most precise data, we filter the data following Section 2.1 in (23):

1. parallax_over_error>10
2. phot_g_mean_flux_over_error>50
3. phot_rp_mean_flux_over_error>20
4. phot_bp_mean_flux_over_error>20

5. phot_bp_rp_excess_factor<1.3+0.06*power(phot_bp_mean_mag-phot_rp_mean_mag,2)

6. phot_bp_rp_excess_factor>1.0+0.015*power(phot_bp_mean_mag-phot_rp_mean_mag,2)

7. visibility_periods_used>8

8. astrometric_chi2_al/(astrometric_n_good_obs_al-5)<1.44*greatest(1,exp(-0.4*(phot_g_mean_mag-19.5)))

To remove strong outliers, we keep data with at least 8 visibility periods. We also only keep data with 10% relative precision in parallax since parallax error would reduce the accuracy in magnitude estimation. We also filter data with 50% precision in G and 20% precision in G_{BP} and G_{RP} to remove variable stars, which would not be our interest in this thesis as anomalies since they are studied in (27). The fifth and sixth filters are applied to remove stars whose BP and RP fluxes are heavily impacted by nearby sources. The last filter is to remove observations that do not fit the single-star parallax model well, possibly due to two stars separated by a very small angle being mistaken for a single object (28).

After applying these filters, 29,952,901 objects remain in the dataset. Since parallax is more accurate for nearby foreground stars, the remaining dataset is biased towards the neighbourhood of the Sun. In fact, all the objects in the remaining dataset are within 1000 kpc of the Sun. However, this bias would not cause a problem for us, and in fact, can be an advantage since we can first focus on stars that are anomalous in a more local setting.

Of the 94 columns of data, many of them are errors of measurements, and correlations

between pairs of measurements, for both of which are not included as input features. Since the errors are not intrinsic features of the sources, but instead come from the instrument, feeding errors into the model would introduce external biases. We are applying a state-of-the-art DL algorithm to detect anomalies (see Sec. 3), which we would expect to model important relationships between features. Hence, if the correlation between pairs of features is important to classify anomalies, the algorithm will be able to learn such correlations directly from the measurements. As a result, we also exclude correlations as input features.

For the same reason, we also remove the 5 astrophysics parameters inferred using Apsis, since in DR2 they are inferred using just photometry and parallax measurements and do not provide additional information to the algorithm. As a result, only 9 features are fed into the anomaly detection algorithm, which are 3 positional features (RA, Dec, parallax), 3 velocity features (proper motion along RA, proper motion along Dec, radial velocity), and photometry of 3 bands (G, BP, RP). Since not every object contains data on all 9 features, we drop the objects with missing data and only keep those with 9 features as input data to the algorithm. Dropping missing data would not introduce bias for us since we are not focusing on the whole population of stars, but instead on point anomalies which individually deviate from other input data.

To further process the data, we apply quantile transform to scale the input data, where original data are transformed to a normal distribution. We argue in Sec. 3.4 the reason we apply quantile transformation instead of standardization as feature scaling.

Chapter 3

Anomaly detection

Anomaly detection, also known as novelty detection or outlier detection, is an active research area to detect data instances that deviate from the majority of the dataset. This means that the anomalies are the data instances that do not conform to the behaviour of the underlying distribution of the whole dataset. It has applications in a wide range of domains from medicine to finance to security. In physics, it has particular importance in the experimental particle physics community to differentiate signals from background (29; 30). Similar to particle accelerators, telescopes also output high-dimensional data from which astrophysicists need to separate signals from a background, which can be done by employing anomaly detection algorithms. Hence, anomaly detection is also important in the astrophysics community to look for signatures in the sky (31; 32; 33; 34; 35; 36; 37; 38; 39; 40).

3.1 Background

Below we review some basics of anomaly detection (4; 5; 6). Anomaly detection algorithms can be broadly categorized into 4 types: supervised anomaly detection, unsupervised anomaly detection, semi-supervised anomaly detection, and weakly-supervised anomaly detection. Supervised anomaly detection algorithms are trained with a dataset which has fully labelled instances for both normal and anomalous classes; unsupervised anomaly detection algorithms are trained without prior knowledge of the anomaly label; semi-supervised anomaly detection assumes the availability of labelled normal data; weakly-supervised anomaly detection assumes the availability of partially labelled anomalous data. We apply an unsupervised technique (see Sec. 3.3) for the Gaia data since we are keeping our target anomalies general, meaning that we do not aim for any specific anomaly. By using the unsupervised algorithm, we expect the model to return any anomalies no matter the reason for not confining to the majority, and then we investigate the reason individually.

Meanwhile, anomalies can also be categorized into 3 types: point anomaly, contextual anomaly, and group anomaly. Point anomalies are individual data instances that are considered anomalous compared to other data instances; contextual anomalies are individual data instances that are only anomalous if considered in a certain context; group anomalies, also known as collective anomalies, are a subset of data instances that are anomalous as a whole compared to other data instances, meanwhile, none of the members

of the group are anomalous. Among them, point anomalies are the most researched ones. In this manuscript, we only focus on point anomalies.

The most basic and common algorithms for point anomaly detection are K-nearest neighbours (KNN) and clustering (4). KNN algorithm finds k nearest neighbours for each sample, where k is an arbitrary integer. From the mean distances to the neighbours, we can classify the samples as anomalous if the distance is significantly higher than the distance of other samples. Meanwhile, the clustering algorithm groups data instances into clusters. Anomalies are detected by finding data instances that do not belong to any cluster, or are far away from the centroid of the nearest cluster. Many subsequent algorithms are either KNN-based or clustering-based.

However, with the advanced development of DL algorithms since the 2010s, it is now possible to exploit deep algorithms to perform the task. Classical machine learning models, such as KNN, clustering or regression, have a simple structure and resemble more closely with traditional statistics methods. DL, on the other hand, has a multi-layered architecture which involves orders of magnitude more parameters than classical machine learning algorithms. Due to this expressiveness, DL models are expected to learn much more complex structures in the data, and as a result, perform better on large-scale data. Hence, deep anomaly detection can be more effective and accurate in identifying anomalies than their classical counterpart for data in scale and dimension like Gaia's (5; 6).

One significant breakthrough recently in DL is generative artificial intelligence. The

underlying principle relies on the deep generative model (DGM), which estimates the likelihood of observations and creates new samples from the underlying distribution. Interestingly, since DGM learns the data distribution, it can also be applied to find anomalies that do not conform to the underlying distribution.

In particular, generative adversarial network (GAN) has been a major approach for anomaly detection since its early use by (41) (AnoGAN) in 2017. The intuition is that since the generative network is optimized to generate instances that are as close to the original data instances as possible, the normal data which form the majority of the whole dataset would have a small difference between the generated data and original data. Meanwhile, the anomalous data instances which have different distributions from normal data, will have a large difference to the generated data.

3.2 Review of anomaly detection on Gaia dataset

There are multiple efforts to identify anomalies within the massive Gaia data, in this section we review the motivation, method and results of other anomaly detection methods on Gaia. We show that among existing research on finding Gaia data anomalies, a gap exists in finding general point anomalies. Finding general points anomalies is important since we do not limit the nature of the anomalies. As a result, we can consider a wider possibility during the investigation. However, the downside is that it can be hard to retrieve the reason for the points to be anomalous.

(35; 36) use ANODE (ANOmaly detection with Density Estimation), an unsupervised machine learning algorithm to specifically look for stellar streams. ANODE (30) is based on normalising flows (42; 43) to estimate the localised probability density of the signal compared to the background. Normalising flows are a generative class of models that learns to transform a simple probability distribution, usually a Gaussian distribution, into a more complex one by a series of invertible and differentiable transformations. The authors transform the angular position into Euclidean coordinates, and also use proper motions along these two directions, colour, and magnitude, to identify overdense regions with respect to the background non-stream stars. They identified 102 stellar streams with high significance with DR2 data, where only 10 had been previously identified.

(39) also target stellar streams particularly. They use CWoLA (Classification Without Labels) (44) to identify stellar streams in a weakly-supervised setting. Originally designed for identifying particles in high-energy physics experiments, CWoLA is a weakly-supervised algorithm that is trained to distinguish sky regions M_1 with a higher proportion of signal and a lower proportion of background, versus sky regions M_2 with a lower proportion of signal and a higher proportion of background. Due to the fact that a classifier trained to distinguish M_1 and M_2 is the same as a classifier for distinguishing signal from background, the model can be directly applied to identify a localised anomaly from the background. As in (35; 36), they also consider Euclidean sky position, proper motions along the direction of the two sky positions, colour and magnitude. They identified ~ 2000 stars likely to belong

to the GD-1 stellar stream.

(40) make use of autoencoder to search for dark matter subhalo-associated stars. Autoencoders have a bottleneck structure and can be separated into two parts: encoder and decoder. The encoder network performs dimensionality reduction, mapping the original input into a lower dimensional latent space, and then the decoder network reconstructs data from the compressed representation. For an ideal model, the reconstructed and the actual inputs should be identical. When stars are near dark matter subhalos, gravitational perturbations will leave imprints on stellar positions and velocities. The encoder first maps the 6D phase space to the lower dimension latent space, and then the decoder decodes them back to the original 6 dimensions. Since background-like samples constitute the majority of the sample, the trained encoder-decoder network will be optimized for reconstructing background-like samples. Meanwhile, for signals which are not distributed like background, the model will find large reconstruction errors The authors use synthetic Gaia DR2 data of 3 MW-like galaxies from FIRE simulations (45), and show 80% true positive rate and 15% false positive rate.

Despite the effort to detect anomalies in Gaia data, all of the above literature assumes a particular nature of the anomalies to detect. This thesis aims to address this gap by looking for general point anomalies that may be anomalous due to any reason, or even due to unknown physics.

3.3 Diffusion Time Estimation

A more recent DGM is score-based diffusion models, which become popular since the proposal of denoising diffusion probabilistic models (DDPM) by (7) in 2020. A diffusion process is a stochastic process with a probability distribution that evolves with time following the diffusion equation, inspired by non-equilibrium statistical physics. The idea of the diffusion model is to systematically destroy the structure in a data distribution through an iterative forward diffusion process (46). Then the model learns a reverse diffusion process that restores structure in data, resulting in a model that is capable of generating samples matching the original data distribution even in high-dimensional spaces (7).

To mathematically describe DDPM, we denote the data as $\mathbf{x}_0$ and the corresponding distribution as $q\Big(\mathbf{x}_0\Big)$. The forward diffusion process is a Markov chain, meaning its distribution at time t only depends on its distribution at time t-1 but not at other times. It gradually adds Gaussian noise to the data, with subsequent latent variable $\mathbf{x}_t$ for $t = 1, ..., T$ defined by

$$\mathbf{x}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \tag{3.1}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{\epsilon}_t; \mathbf{0}, \mathbf{I})$ is Gaussian noise with mean 0 and variance 1, and $\beta_t \in (0, 1)$ is the variance schedule. This equation means that at each time $t = 1, ..., T$ we keep adding noise with mean 0 and a predetermined variance β_t, in order to destroy information in the data gradually. The drift $\sqrt{1 - \beta_t}$ is added so that the variance of each latent variable is kept at

1. Hence, writing Eq. 3.1 in terms of the distribution of latent variable $\mathbf{x}_1, ..., \mathbf{x}_T$ is

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{3.2}$$

meaning that the distribution of $\mathbf{x}_t$ at t conditioned on its lagged value at $t-1$, $\mathbf{x}_{t-1}$, is normal distributed with mean being its lagged value $\mathbf{x}_{t-1}$ multiplied by $\sqrt{1-\beta_t}$, and the variance being β_t. Over time when $T \to \infty$, $\mathbf{x}_T$ has an isotropic Gaussian distribution.

We can express $\mathbf{x}_t$ at any arbitrary time t in terms of input data $\mathbf{x}_0$ only, by reparametrization of β_t into $\tilde{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau = \prod_{\tau=1}^{t}(1-\beta_\tau)$:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_t \\
&= \cdots \\
&= \sqrt{\tilde{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\tilde{\alpha}_t}\boldsymbol{\epsilon}_t
\end{aligned} \tag{3.3}$$

where the third line is due to the fact that adding two normal distributions with variance $\sigma_1^2 = \alpha_t(1-\alpha_{t-1})$ and $\sigma_2^2 = 1-\alpha_t$ results in a normal distribution with variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 = \alpha_t(1-\alpha_{t-1}) + 1 - \alpha_t = 1 - \alpha_t\alpha_{t-1}$. Hence, we can express the distribution of x_t

conditioned on the input data x_0 as

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\tilde{\alpha}_t}\mathbf{x}_0, (1-\tilde{\alpha}_t)\mathbf{I}). \tag{3.4}$$

The goal of DDPM is to recreate a sample from the original distribution from Gaussian noise, by learning the approximation of the distribution of reverse process $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{w})$ governed by deep neural network $\mu(\mathbf{x}_t, \mathbf{w}, t)$ with parameters $\mathbf{w}$:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{w}) = \mathcal{N}(\mathbf{x}_{t-1}|\mu(\mathbf{x}_t, \mathbf{w}, t), \beta_t\mathbf{I}). \tag{3.5}$$

A common approach for training a diffusion model to detect anomalies is to treat the anomalies as the noise from forward diffusion, such that we can reverse diffuse the input samples and thus use the reconstruction distance to identify anomalies (47; 48; 49). However, (8) proposed a much simpler yet powerful approach called Diffusion Time Estimation (DTE). DTE use forward diffusion to create noisy samples as a way to simulate anomalous samples. Since noisy samples are expected to cover the entire feature space, they should cover potential anomalies as well. Then, a neural network is trained to predict the diffusion time corresponding to the noisy samples. Thus, an anomalous sample would look "noisy" to the model, and it would estimate a high diffusion time step, hence predicting a high anomaly score.

Since DTE does not model the reverse diffusion process like DDPM does, (8) shows

that DTE inference time is shorter than DDPM by 3 orders of magnitude. It is essential to apply an efficient algorithm like DTE to apply on data on the scale of Gaia, since it would be extremely computationally expensive to run slower algorithms such as KNN, on 30 million objects in the processed Gaia data across multiple features. Meanwhile, both models achieve high accuracy with similar AUC ROC (~ 0.8) evaluated with 57 anomaly detection benchmark datasets ADBench (Anomaly Detection Benchmark) (50). AUC ROC, the Area Under the Curve of the Receiver Operating Characteristics curve, is a commonly used metric to illustrate the accuracy of classification models. The ROC curve is a plot of false positive rates versus true positive rates at different thresholds. An ideal model with perfect prediction will have an AUC ROC of 1 while an AUC ROC of 0 shows the model is predicting everything wrong. Therefore, DTE is much more efficient and accurate compared to other anomaly detection models.

The application of an ML algorithm on a data type that the algorithm is not typically used for could lead to varying performance. Since the diffusion model is typically used for image data, it is important to prove the model is robust on general data like Gaia data, which may not have strong neighbouring pixel relationships like image data do, before we apply it to Gaia data. While (47; 48; 49) focus on the application of diffusion model on image data anomalies, (8) does not assume any data type or the nature of the anomaly during their analysis. Hence, given the data type, dimensionality and number of sources in the Gaia data, combined with the accuracy and efficiency of DTE, we expect DTE would be

a robust choice for identifying anomalies in Gaia data.

We add noise to each input sample until a random timestep between 0 and 300, according to the noise schedule $\beta = 0.0001, \cdots, 0.01$. A multilayer perceptron, with a hidden size of [512,512], is then trained to classify the timesteps of each noisy sample with a batch size of 512. We choose Adam as the optimizer with a learning rate of 0.003 and a weight decay of 0.0005. The model is then applied to the original input sample to classify the timesteps which is then normalized to a score between 0 and 6.

3.4 Experiments

To check if DTE is able to catch potential anomalies in the Gaia data, we have planted artificial anomalies, making use of the Hertzsprung-Russell diagram (H-R diagram). The H-R diagram illustrates the empirical relationship between stellar temperature and luminosity. It has the stellar temperature on the x-axis and log luminosity (or magnitude) on the y-axis. 3.1 shows the H-R diagram for 50000 randomly sampled objects from Gaia DR2, with luminosity expressed in absolute G-band magnitude and colour expressed in BP-band magnitude minus RP-band magnitude (BP-RP). A higher BP-RP value means the object is redder due to the fact that a higher magnitude corresponds to lower luminosity.

Plotting magnitude and colour of observations on H-R diagrams shows that the stars are not randomly distributed in the colour-luminosity space, but instead are confined to certain regions, where each region corresponds to an evolutionary stage of stars.

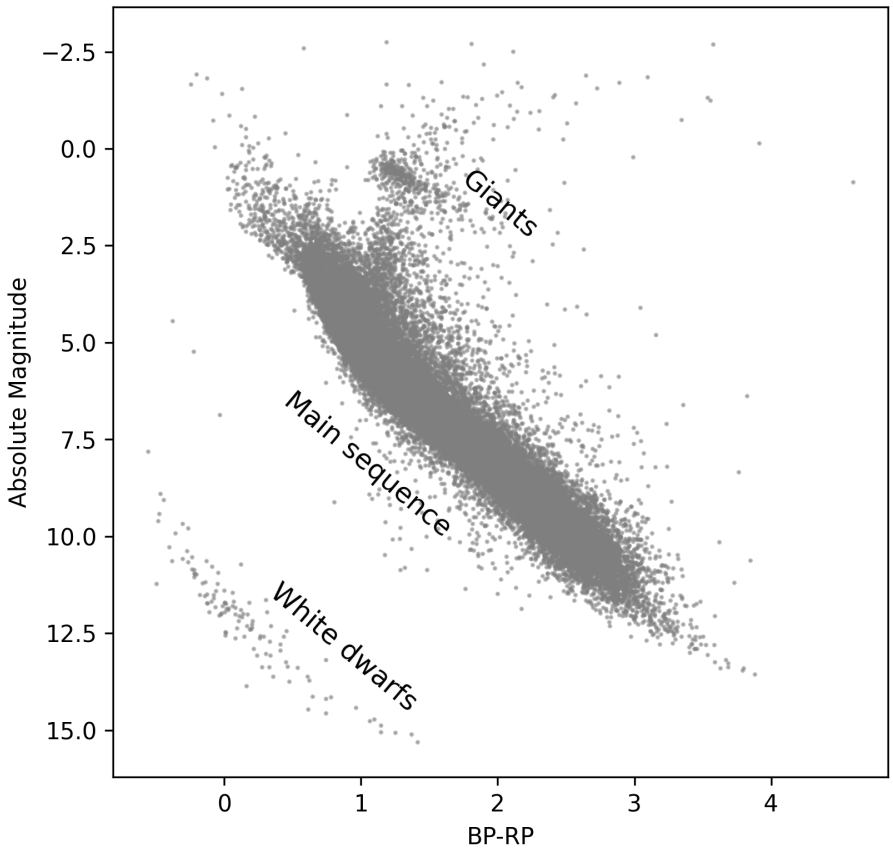

Figure 3.1: H-R diagram for 50000 randomly sampled objects from Gaia DR2.

Most of the stars are main-sequence stars, which occupy a diagonal band from the top left (high luminosity and blue) to the bottom right (low luminosity and red). Stars spend most of their lifetime being main sequence stars. During this time, they contract under self-gravity and heat themselves up, and start to fuse hydrogen into helium, generating energy to provide pressure against the self-gravity, thus would be in a hydrostatic equilibrium. The resulting equation of hydrostatic equilibrium combined with the equation of state and the equation of radiative transport gives temperature scaled as the square of radius. Meanwhile, Stefan-Boltzmann law, $L = \sigma 4\pi R^2 T^4$, states the luminosity T is proportional to the star's surface area $4\pi R^2$ and the fourth power of effective temperature T^4 which is determined by the blackbody spectrum, up to a constant σ called Stefan-Boltzmann constant. Hence, we obtain a scale relationship between luminosity and temperature, that luminosity scaled as the eighth power of temperature, which explains the diagonal band in the H-R diagram.

Therefore, from the laws of physics, we know where in the colour-luminosity space we would not expect main-sequence stars. This would be a robust check for the algorithm because if we plant anomalies that violate this result, we would expect the model to catch these anomalies. To plant anomalies, we randomly pick sources and change their brightness in order the bring them to a position where they are not expected to exist in the H-R diagram.

Since different features may have substantially different ranges and orders of magnitude, it is always beneficial to rescale the data to avoid placing importance on some particular

features simply because they have large values. The best practice in data preprocessing before feeding the data into a machine learning model is to standardize the data, which essentially means calculating the standard score, or to normalize the data, where each data is divided by the maximum of each feature, hence each data has a value between 0 and 1 after normalization.

However, we find in fig.3.2 that when we standardize the data, the model does not find the planted anomalies with low luminosity anomalous. Meanwhile, for planted anomalies with high luminosity, the model indeed finds them to be anomalous. This is because, as seen in the H-R diagram, important information regarding luminosity is extracted in log space, since the difference in luminosity for the brightest and dimmest stars can span multiple orders of magnitude. Besides, we show in fig. 3.3 that the standardized luminosity is highly skewed to the low luminosity end. Hence, the $\mathcal{O}(10^{-1})$ difference in the linear space between the planted anomalies and the normal main sequence stars does not seem anomalous to the model, but instead are considered normal since they are inside the strongly peaked range of values.

As a remedy, we apply quantile transform instead of standardization for feature scaling. Quantile transform is also known as inverse cdf (cumulative distribution function) transform, as the name suggests, is obtained by inverting the cdf. This means that the value after the transform is the corresponding y-axis value, the cumulative probability in the cdf. The resulting distribution is a uniform distribution in the range (0,1), which is then transformed

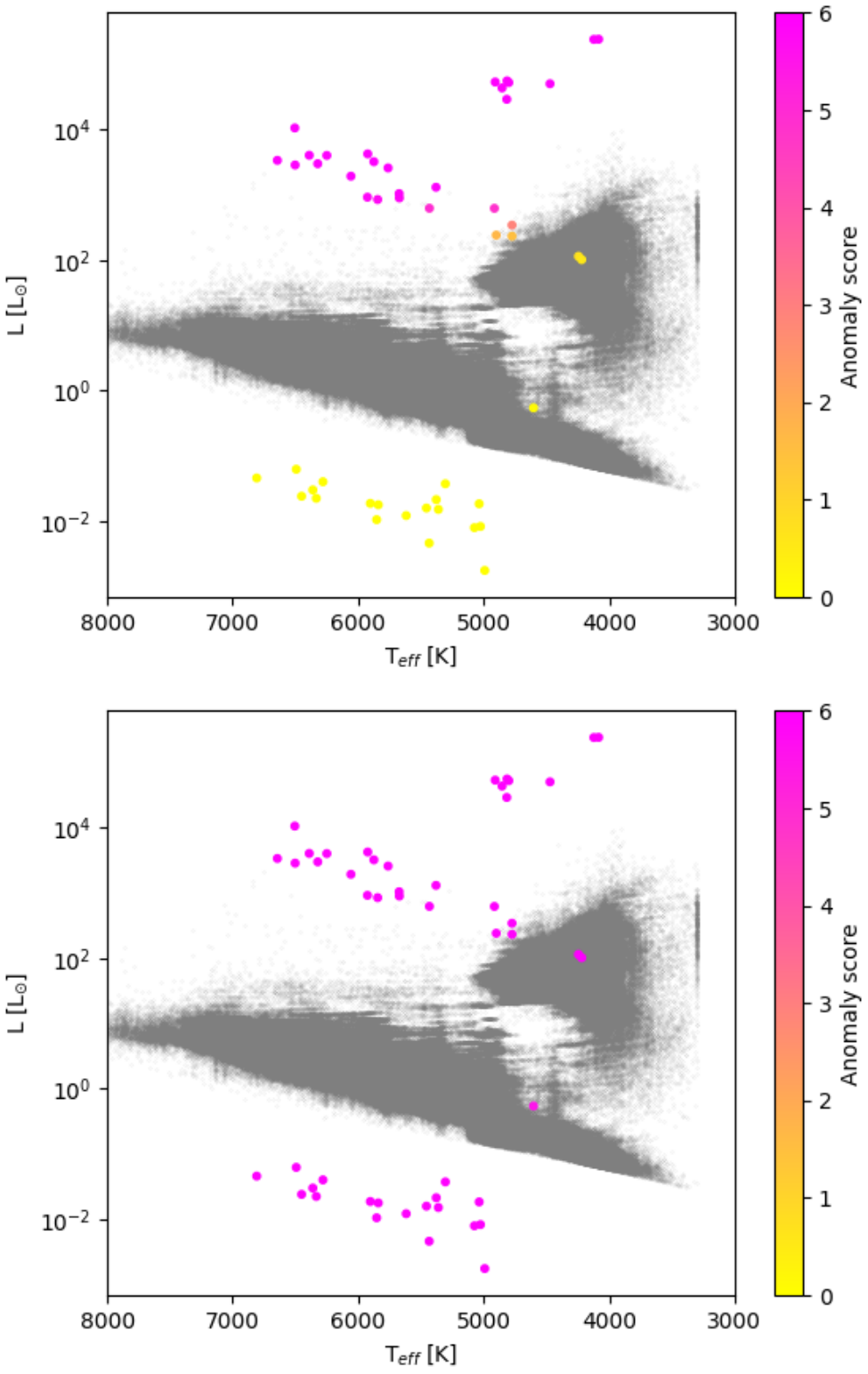

Figure 3.2: Above: H-R diagram with artificial anomalies planted, with standardization as data preprocessing. Below: The same as above, but with quantile transform as data preprocessing.

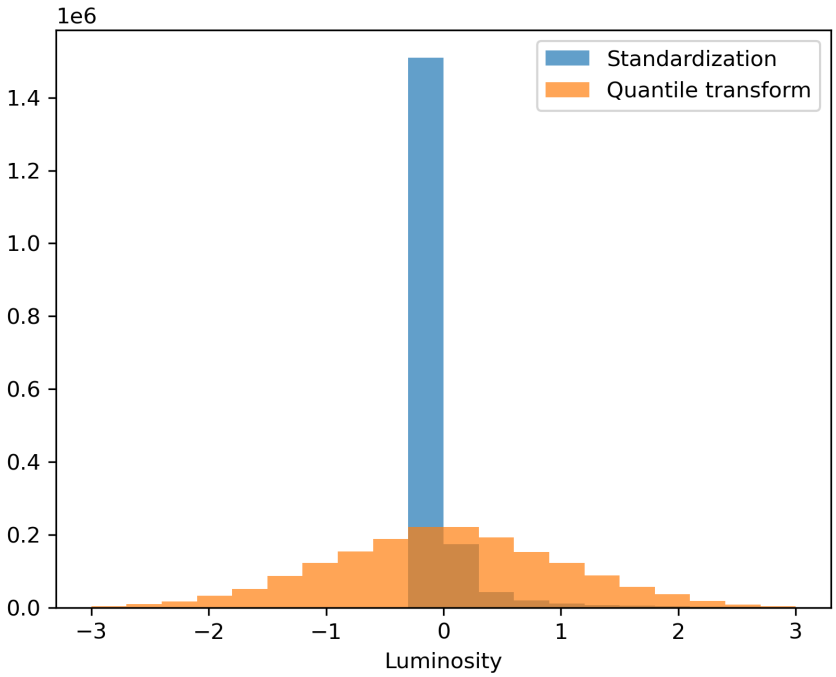

Figure 3.3: Distribution of scaled luminosity using standardization versus quantile transform as feature scaling.

to a normal distribution. Hence, the transformation spreads out dense regions while bringing data in the sparse region closer together. Thus, it places importance on the relative rank rather than the actual distance between data. This would increase the robustness of the model, especially for skewed data. Also, since the data is redistributed to a Gaussian, it can reduce the impact of extreme anomalies in the data. Therefore, it can avoid anomalies being caught simply because they are extreme outliers in one feature, which is not the most interesting target in our study, since we are applying DL algorithm which we expect to find more complex anomalies. As a result, we expect the model will focus on data that are anomalous in higher dimensional space. No additional standardisation is necessary since the data are already in a standardised form after the quantile transformation.

As seen in fig. 3.3, after applying quantile transform, the data smooth out and are no longer closely packed at low luminosity, which would help the model extract features that are interesting in the log space. Our experiment shows that by applying the quantile transform, both planted high-luminosity stars and low-luminosity stars are found to be anomalous.

Chapter 4

Results and discussion

4.1 Classifying anomalies

DTE model is not a categorical model that returns whether each data is an anomaly. Instead, it gives an anomaly score in the range of 0 to 6, with 0 being the least probability to be anomalous and 6 being very probably to be anomalous. As a result, we have to apply an arbitrary threshold for determining whether each input data is anomalous.

Fig. 4.1 shows the distribution of the anomaly score determined by DTE. We choose a cutoff at 5.3 so that about 0.1% of the data is anomalous. It is a conservative choice considering the proximity to the maximum anomaly score. We run DTE twice to avoid the model picking the anomalies randomly, or only because they are added noise first. There is about 80% of overlapping anomalies found by the two runs of DTE. The consistency

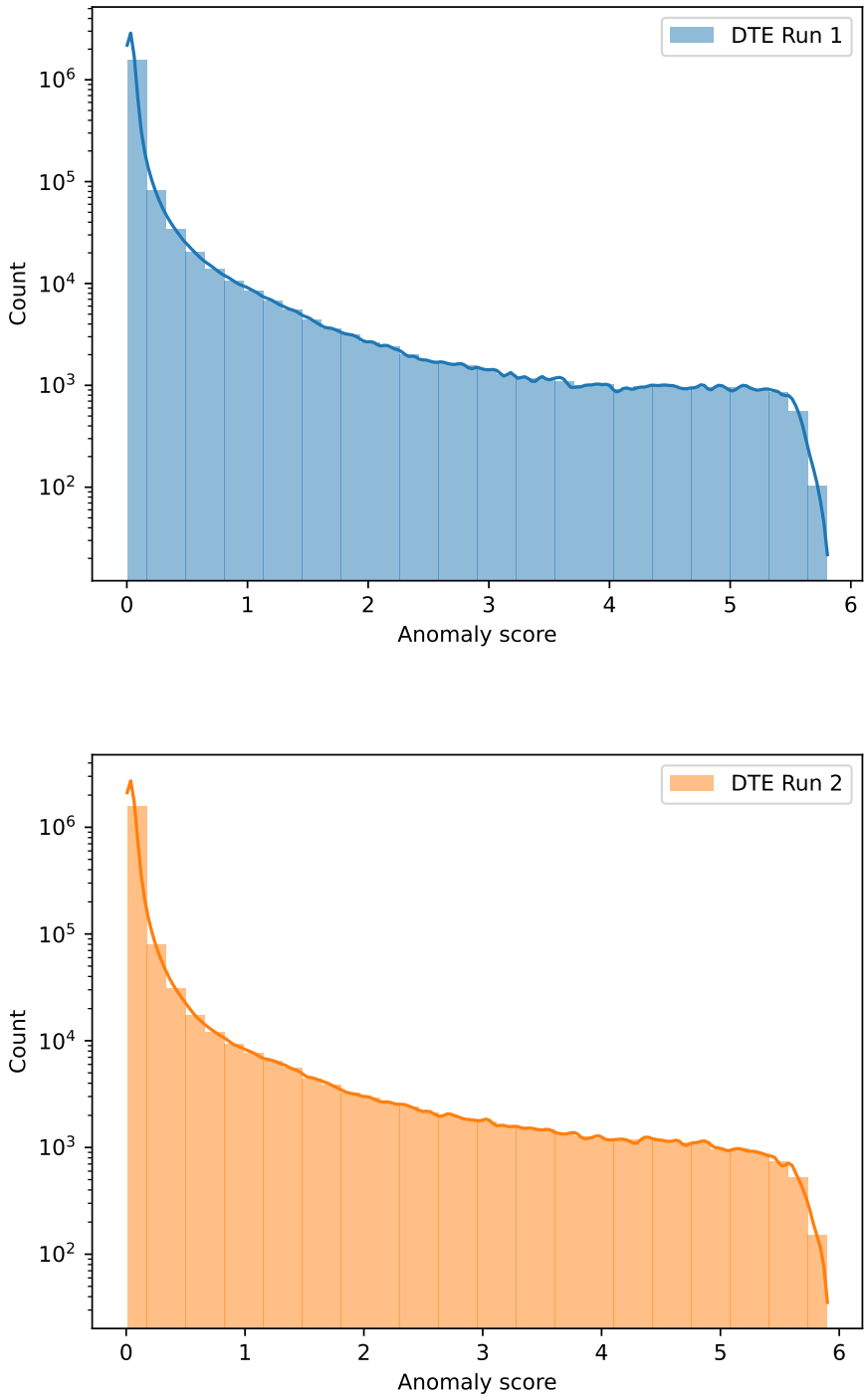

Figure 4.1: Distribution of anomaly score of the input data determined by the 2 runs of the DTE model.

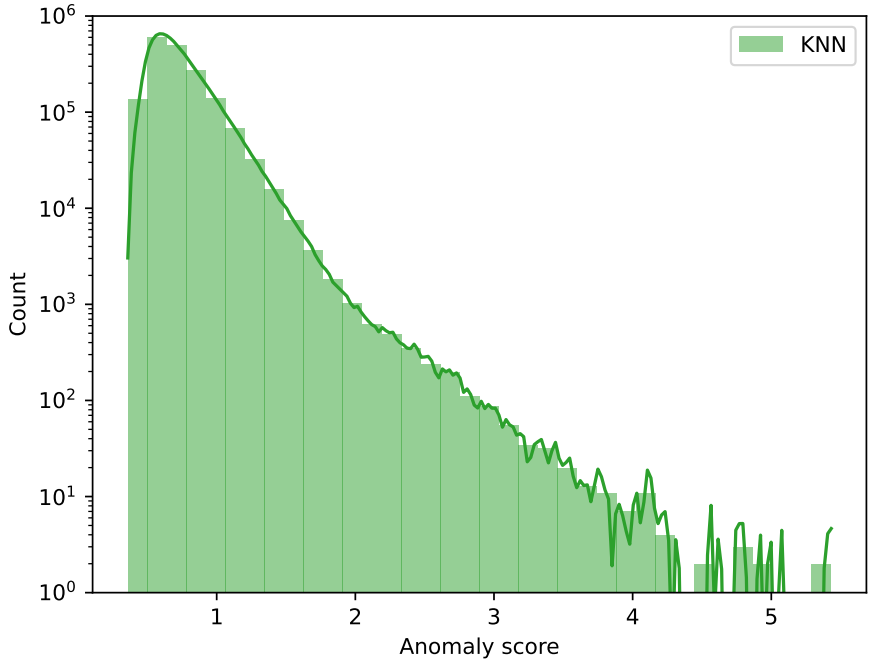

Figure 4.2: Distribution of anomaly score of the input data determined by KNN.

shows we can rule out the model finding random anomalies, or significantly influenced by initialization. For robustness, we only consider anomalies picked by both runs of DTE in the following analysis. Fig. 4.1 also shows the distribution of the anomaly score of the second run of DTE.

We also run the K-nearest neighbours (KNN) anomaly detection from `PyOD` (51) for comparison. KNN is a classical machine learning algorithm that classifies input data based on the class of its K nearest neighbours. To detect anomalies, an anomaly score is calculated by finding its average distance to its K nearest neighbours (52). We show in Fig 4.2 the distribution of anomaly score determined by KNN on the same Gaia dataset.

As in the case of DTE, we apply a cutoff arbitrarily at 2.2 for the anomaly score determined by KNN such that roughly 0.1% of the data is anomalies. Only 25% of the anomalies picked by DTE are also picked by KNN. This is expected because of the nature of the two algorithms. KNN is a classical algorithm while DTE is a deep algorithm, hence we would expect KNN to catch only low-level anomalies that are anomalous distance-wise. If there is a more abstract underlying structure in the input data, we would expect a deep model to capture the structure and as a result, pick different anomalies.

4.2 Analysis

Fig. 4.3 shows a pairwise scatter plot of the anomalies in each of the input features. We do not include the position in RA and Dec in the pairwise plot since they do not provide useful information. We show the G-band flux in log scale to resemble the y-axis of the H-R diagram. We also show only the difference between RP-band log flux and BP-band log flux instead of the flux of these two bands separately, since they just grow linearly with G-band flux and do not possess much information. This difference resembles colour in the H-R diagram since the higher the difference the redder the object is.

For normal input data, we show the contour in grey in the background for comparison. We also separate the anomalies into two groups: those only picked by DTE are shown in orange; while those picked jointly by DTE and KNN are shown in blue.

As discussed in 4.1, DL methods can pick anomalies that are anomalous in a higher-level

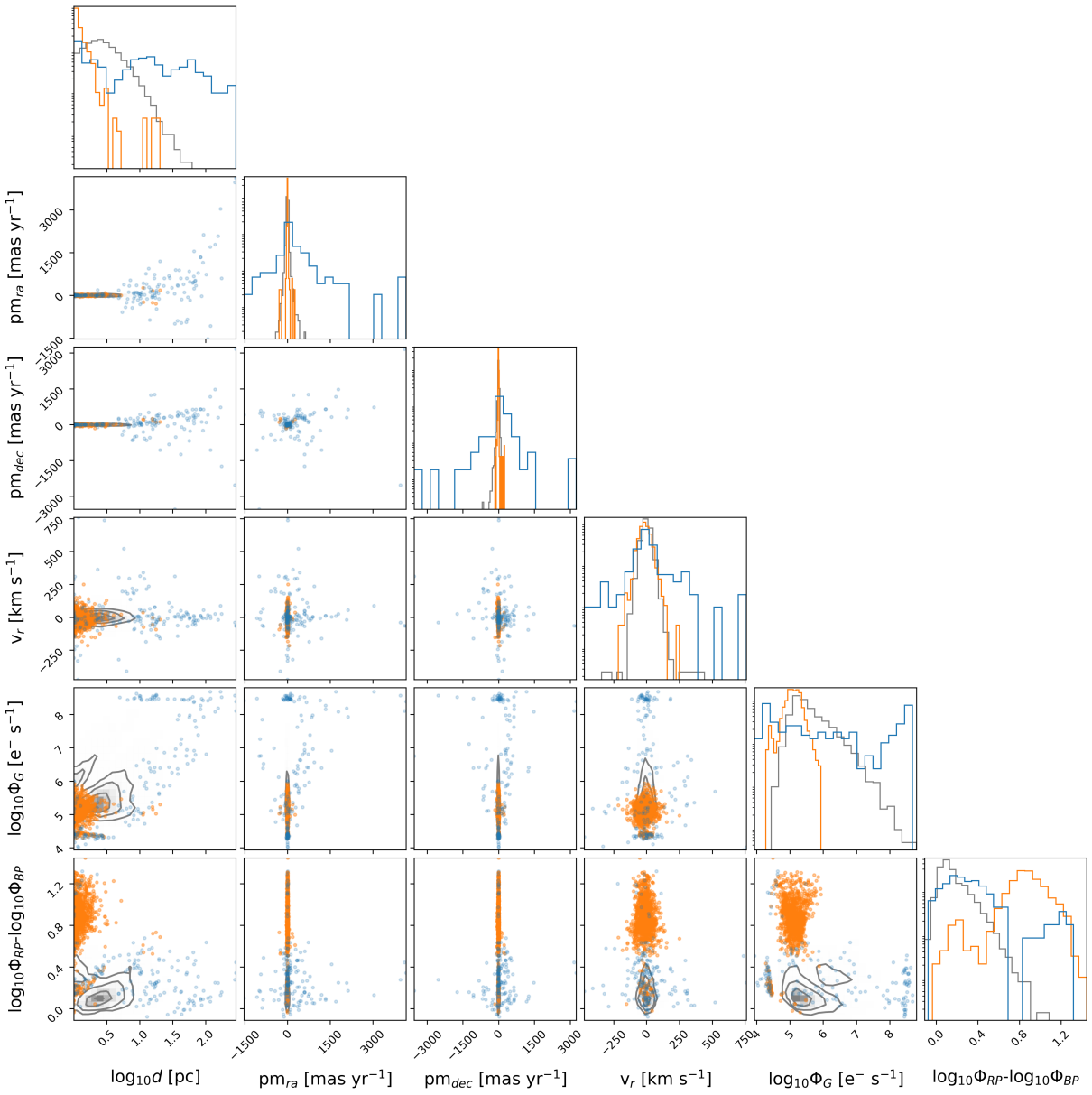

Figure 4.3: Pairwise plot for the input parameters.

latent space. This is reflected in the H-R diagram in Fig. 4.3 (the $\log_{10}\Phi_G$ vs $\log_{10}\Phi_{RP}$-$\log_{10}\Phi_{BP}$ plot). The anomalies caught solely by DTE mainly deviate from the normal data in the colour space, which is high-order since it is not part of the input features. Meanwhile, for the other features that are present in the input data, the anomalies picked only by DTE have a similar distribution to the normal data instances. On the other hand, the anomalies picked jointly by KNN and DTE are mostly outliers that deviate significantly from the normal data distribution in velocity, distance and flux.

4.2.1 Unbound stars

Since the KNN-DTE-overlapped anomalies are mostly outliers in terms of velocity, we consider the possibility that they can be unbound from the Galaxy. There are three classes of objects that sit at the high tail of the velocity distribution. Halo stars (53) are stars that do not follow orbits around the Galactic Centre (GC) within the disk, they can be unbound if they are the debris of tidally disrupted satellites. Runaway stars (RSs) (54) are O or B-type stars ejected from the disk, formed either by encounters between stars in dense systems, such as young star clusters, or supernova explosions in binary systems. Hypervelocity stars (HVSs) (55) are the result of the three-body interaction between a binary system and the supermassive black hole in the GC. Due to this close encounter, they can reach a velocity of $\sim$ 1000 km s^{-1}, high enough to escape from the gravitational field of the Galaxy.

Source ID	v_{tot}/v_{esc}	DTE score	DTE2 score	KNN score
5231593594752514304	1.634498	5.209758	5.322006	2.907877
2251311188142608000	1.683953	5.551567	5.689544	4.564000
5878409248569969792	1.575172	0.167402	1.709670	2.418901
5956359499060605824	1.131206	0.124300	1.264231	2.168586
4065480978657619968	1.161136	5.157624	5.713779	3.825398
4296894160078561280	1.625867	5.653924	5.733806	4.016321
5412495010218365568	1.248013	5.740820	5.730458	3.444774

Table 4.1: Gaia DR2 source ID, total velocity to escape velocity ratio, and anomaly scores found by the 3 algorithms for the unbound stars.

We convert the proper motion and radial velocity to total velocity v_{tot}. To do this, we first convert the proper motion to radial velocity to velocity galactocentric Euclidean coordinate using `galpy` (56), with the x-axis pointing towards the GC, the y-axis being the tangential direction, and the z-axis being perpendicular to the disk. We assume that the distance between the Sun and the GC is $d_{\odot} = 8.2$kpc, the Sun's height above the disk is $z_{\odot} = 25$pc, and the Sun's tangential velocity about the GC is $v_{g,\odot} = 248\text{km s}^{-1}$ (14). We also transform their position from RA, Dec and parallax to galactocentric Euclidean coordinate using `galpy`. From the galactocentric coordinate, we obtain the escape velocity v_{esc} using again `galpy` with MWPotential2014 as the model for the Milky Way gravitation potential.

We found 7 objects with total velocity v_{tot} larger than escape velocity v_{esp}. We report the Gaia DR2 source ID, total velocity to escape velocity ratio, and the anomaly scores in Table 4.1. There are 5 objects with very high anomaly scores and 2 with very low scores. Although only 3 objects pass the anomaly threshold stated in sec. 4.1, in the following we consider all 5 objects with high scores as anomalies. In fig. 4.4 we show the 3-dimension

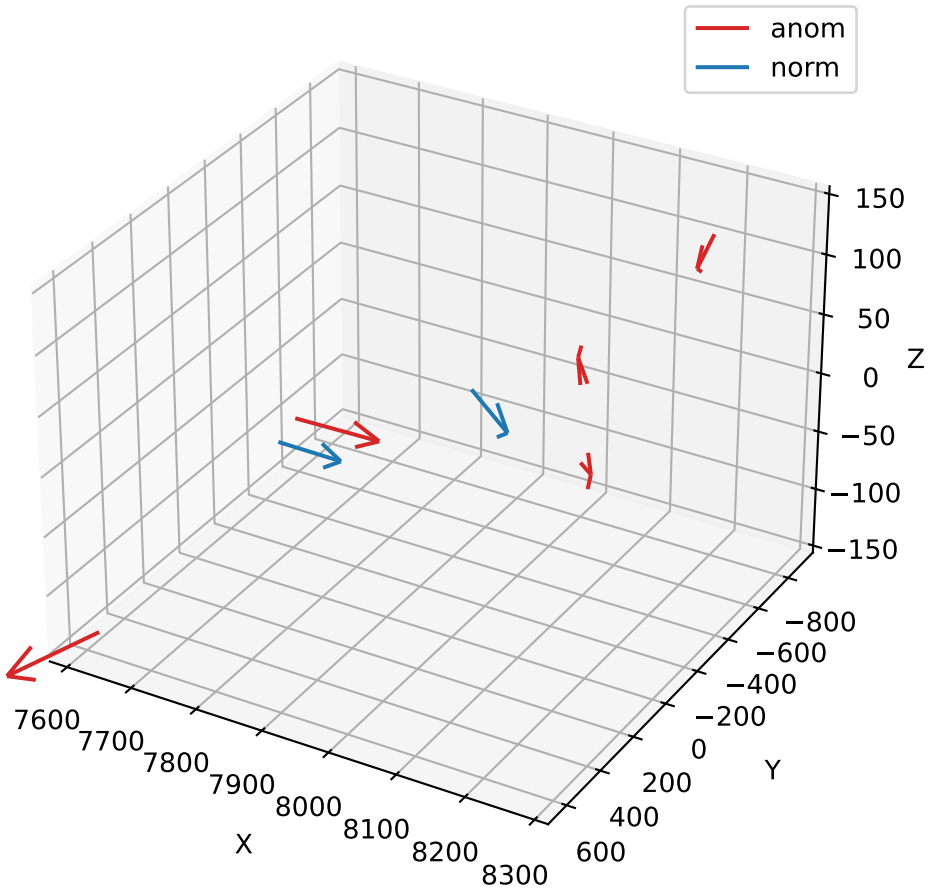

Figure 4.4: 3-dimensional quiver plot for the unbound stars.

quiver plot for the unbound stars. It shows that two of the unbound stars actually are going against the direction of rotation of the galactic disk.

4.2.2 Extinction

Another observation from the pairwise plot is that anomalies have another peak at a redder colour. Hence, we also consider the possibility that the anomalies have high red colour excess. Due to the presence of dust between the sources and the observatory, light will be absorbed and scattered by the dust, thus reducing light flux. The effect of this extinction on shorter wavelengths is higher than that of longer wavelengths, hence it will lead to reddening of light.

In fig. 4.5, we plot the histogram of colour excess of anomalies compared with normal data. We use the *e_bp_min_rp_val* parameter, the BP band magnitude and RP band magnitude difference, inferred by Apsis in Gaia DR2. We also use `dustmap` (57) to find the ebv, excess blue band minus visible band parameter, using the spatial coordinate determined by Gaia. With both ways of finding the reddening, we show that the anomalous data have high reddening compared to normal data.

Although further analysis is needed to confirm if this lead to any new discovery, the fact that the anomalies have a different distribution on reddening, which is not an input feature, confirms again the ability of our method to detect complex point anomalies in the enormous Gaia data. This would be significant when anomaly detection is performed on data releases

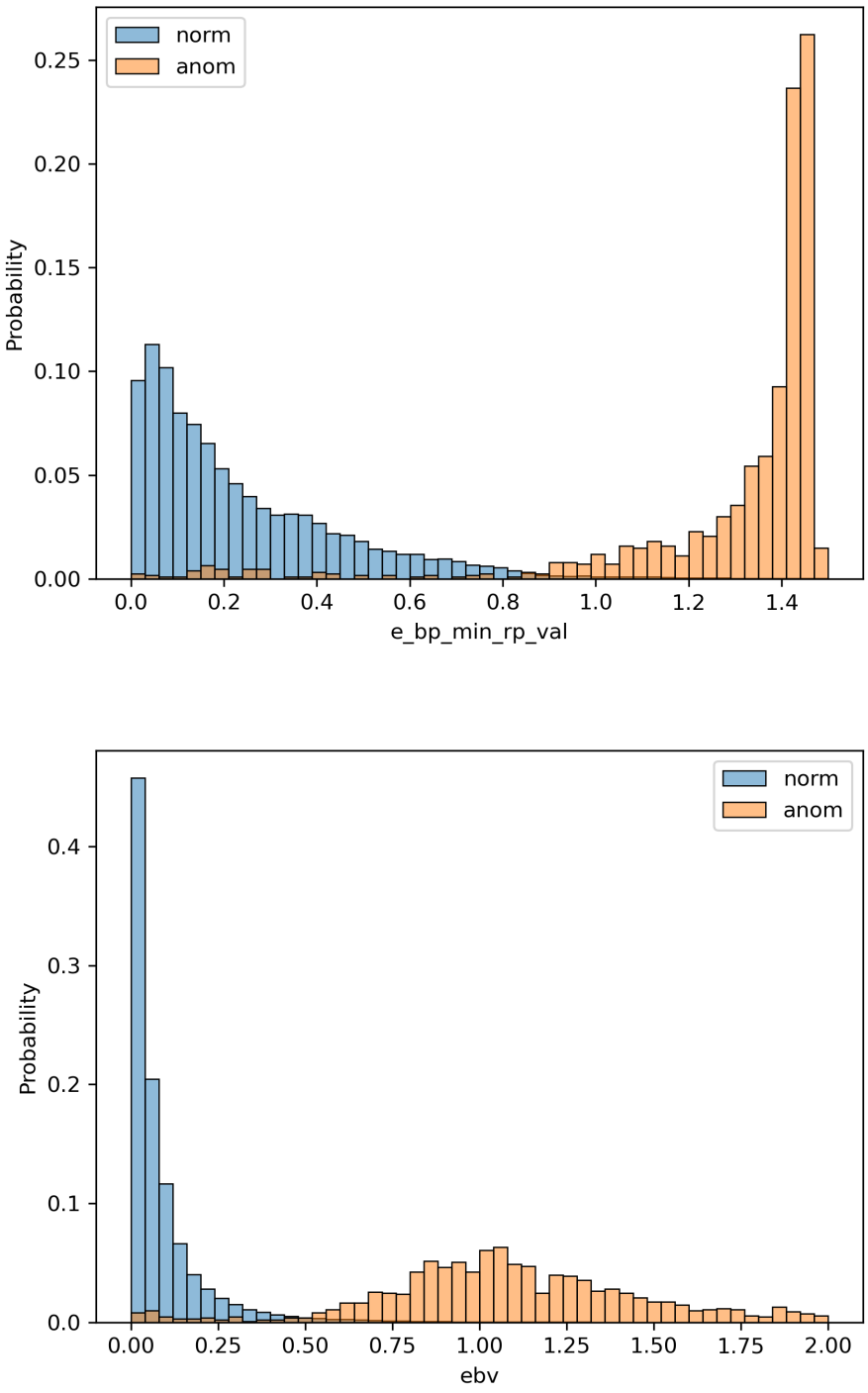

Figure 4.5: The colour excess of anomalous data compared to normal data. Above: the colour excess found by Apsis in Gaia DR2. Below: the colour excess found using `dustmap`.

after DR2 which contains more objects but also more features. The efficient yet powerful algorithm can be used to quickly identify anomalous data in newer data releases.

Chapter 5

Conclusion

In this thesis, we detect point anomalies in Gaia space telescope's data, in order to search for stars in the MW that deviate from the underlying distribution of the whole dataset. We apply a state-of-the-art deep learning algorithm, Diffusion Time Estimation, on Gaia's second data release to perform anomaly detection in an unsupervised manner. Our experiment with artificial anomalies on the H-R diagram shows that the efficient yet accurate algorithm is a robust method to detect point anomalies in the Gaia data, if we perform feature scaling by quantile transform instead of standardization. We then apply the model without the artificial anomalies to calculate the anomaly score, and apply an arbitrary threshold on the anomaly score to isolate 0.1 of anomalous data. We also apply a more traditional algorithm, KNN, to compare the results. Using a pairwise plot of the input features, we observe that the anomalies have different distributions in velocity and colour compared to the normal data.

Our analysis demonstrates that 7 objects in our filtered data have a velocity higher than the escape velocity, meaning they are unbound from the galactic potential. We also show that the normal and anomalous data have a different distribution in reddening, a quantity not present as an input, further confirming the power of the model in finding complex anomalies. Thus, further analysis of the nature of the anomalies, and application on newer data releases, may lead to more discovery in addition to those investigated in this thesis.

Bibliography

[1] M. Perryman, *The History of Astrometry*, *Eur. Phys. J. H* **37** (2012) 745, [1209.3563].

[2] M. A. C. Perryman, K. S. De Boer, G. Gilmore, E. Høg, M. G. Lattanzi, L. Lindegren et al., *GAIA: Composition, formation and evolution of the Galaxy*, *Astronomy & Astrophysics* **369** (Apr., 2001) 339–363.

[3] Gaia collaboration, T. Prusti et al., *The Gaia Mission*, *Astron. Astrophys.* **595** (2016) A1, [1609.04153].

[4] V. Chandola, A. Banerjee and V. Kumar, *Anomaly detection: A survey*, *ACM Comput. Surv.* **41** (2009) 15:1–15:58.

[5] R. Chalapathy and S. Chawla, *Deep learning for anomaly detection: A survey*, *ArXiv* **abs/1901.03407** (2019) .

[6] G. Pang, C. Shen, L. Cao and A. van den Hengel, *Deep learning for anomaly detection*, *ACM Computing Surveys (CSUR)* **54** (2020) 1 – 38.

[7] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, .

[8] V. Livernoche, V. Jain, Y. Hezaveh and S. Ravanbakhsh, *On diffusion modeling for anomaly detection*, 2023.

[9] M. A. C. Perryman et al., *The Hipparcos catalogue*, *Astron. Astrophys.* **323** (1997) L49–L52.

[10] E. Hoeg, G. Bässgen, U. Bastian, D. Egret, C. Fabricius, V. Großmann et al., *The TYCHO Catalogue*, **323** (July, 1997) L57–L60.

[11] E. Hog, C. Fabricius, V. V. Makarov, S. Urban, T. Corbin, G. Wycoff et al., *The Tycho-2 catalogue of the 2.5 million brightest stars*, *Astron. Astrophys.* **355** (2000) L27–L30.

[12] C. A. L. Bailer-Jones et al., *The Gaia astrophysical parameters inference system (Apsis). Pre-launch description*, *Astron. Astrophys.* **559** (2013) A74, [1309.2157].

[13] G. Bertone and D. Hooper, *History of dark matter*, *Rev. Mod. Phys.* **90** (2018) 045002, [1605.04909].

[14] J. Bland-Hawthorn and O. Gerhard, *The galaxy in context: Structural, kinematic, and integrated properties*, *Annual Review of Astronomy and Astrophysics* **54** (2016) 529–596.

[15] A. Helmi, *Streams, substructures, and the early history of the milky way*, *Annual Review of Astronomy and Astrophysics* (2020) .

[16] C. Mateu, *galstreams: A library of milky way stellar stream footprints and tracks*, 2022.

[17] N. Banik, G. Bertone, J. Bovy and N. Bozorgnia, *Probing the nature of dark matter particles with stellar streams*, *JCAP* **07** (2018) 061, [1804.04384].

[18] A. Helmi, C. Babusiaux, H. H. Koppelman, D. Massari, J. Veljanoski and A. G. A. Brown, *The merger that led to the formation of the milky way's inner stellar halo and thick disk*, *Nature* **563** (2018) 85 – 88.

[19] M. Xiang and H.-W. Rix, *A time-resolved picture of our milky way's early formation history*, *Nature* **603** (2022) 599 – 603.

[20] E. Poggio, R. Drimmel, R. Andrae, C. A. L. Bailer-Jones, M. Fouesneau, M. G. Lattanzi et al., *Evidence of a dynamically evolving galactic warp*, *Nature Astronomy* **4** (2019) 590 – 596.

[21] T. Ruiz-Lara, C. Gallart, E. J. Bernard and S. Cassisi, *The recurrent impact of the sagittarius dwarf on the milky way star formation history*, *arXiv: Astrophysics of Galaxies* (2020) .

[22] T. Antoja, A. Helmi, M. Romero-Gómez, D. Katz, C. Babusiaux, C. Babusiaux et al., *A dynamically young and perturbed milky way disk*, *Nature* **561** (2018) 360 – 362.

[23] Gaia collaboration, C. Babusiaux et al., *Gaia Data Release 2: Observational Hertzsprung-Russell diagrams*, *Astron. Astrophys.* **616** (2018) A10, [1804.09378].

[24] W.-C. J. , T. J. Henry, D. R. Gies and N. C. Hambly, *A gap in the lower main sequence revealed by gaia data release 2*, *The Astrophysical Journal Letters* **861** (2018) .

[25] Gaia collaboration, A. Vallenari et al., *Gaia Data Release 3 - Summary of the content and survey properties*, *Astron. Astrophys.* **674** (2023) A1, [2208.00211].

[26] Gaia collaboration, A. G. A. Brown et al., *Gaia Data Release 2: Summary of the contents and survey properties*, *Astron. Astrophys.* **616** (2018) A1, [1804.09365].

[27] Mowlavi, N., Lecoeur-Taïbi, I., Lebzelter, T., Rimoldini, L., Lorenz, D., Audard, M. et al., *Gaia Data Release 2 - The first Gaia catalogue of long-period variable candidates*, *A&A* **618** (2018) A58.

[28] Lindegren, L., Hernández, J., Bombrun, A., Klioner, S., Bastian, U., Ramos-Lerate, M. et al., *Gaia Data Release 2 - The astrometric solution*, *A&A* **616** (2018) A2.

[29] K. Fraser, S. Homiller, R. K. Mishra, B. Ostdiek and M. D. Schwartz, *Challenges for unsupervised anomaly detection in particle physics*, *JHEP* **03** (2022) 066, [2110.06948].

[30] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, *Phys. Rev. D* **101** (2020) 075042, [2001.04990].

[31] R. Zhang and Q. Zou, *Time series prediction and anomaly detection of light curve using lstm neural network*, *Journal of Physics: Conference Series* **1061** (2018) .

[32] E. E. O. Ishida, M. V. Kornilov, K. L. Malanchev, M. V. Pruzhinskaya, A. A. Volnova, V. S. Korolev et al., *Active anomaly detection for time-domain discoveries*, *ArXiv* **abs/1909.13260** (2019) .

[33] M. D'Addona, G. Riccio, S. Cavuoti, C. Tortora and M. Brescia, *Anomaly detection in astrophysics: A comparison between unsupervised deep and machine learning on kids data*, *Intelligent Astrophysics* (2020) .

[34] K. Storey-Fisher, M. Huertas-Company, N. Ramachandra, F. Lanusse, A. Leauthaud, Y. Luo et al., *Anomaly detection in astronomical images with generative adversarial networks*, *arXiv: Astrophysics of Galaxies* (2020) .

[35] D. Shih, M. R. Buckley, L. Necib and J. Tamanas, *via machinae: Searching for stellar streams using unsupervised machine learning*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 5992–6007, [2104.12789].

[36] D. Shih, M. R. Buckley and L. Necib, *Via Machinae 2.0: Full-Sky, Model-Agnostic Search for Stellar Streams in Gaia DR2*, 2303.01529.

[37] M. Mesarcik, A.-J. Boonstra, M. Iacobelli, E. Ranguelova, C. T. A. M. de Laat and R. van Nieuwpoort, *The road to discovery: machine learning-driven anomaly detection in radio astronomy spectrograms*, *ArXiv* **abs/2307.01054** (2023) .

[38] A. Nelles, Z. S. Meyers, J. A. A. Aguilar, P. Allison, D. Z. Besson, A. Bishop et al., *Anomaly detection in early data from the radio neutrino observatory greenland*, *Proceedings of 38th International Cosmic Ray Conference — PoS(ICRC2023)* (2023) .

[39] M. Pettee, S. Thanvantri, B. Nachman, D. Shih, M. R. Buckley and J. H. Collins, *Weakly-Supervised Anomaly Detection in the Milky Way*, 2305.03761.

[40] A. Bazarov, M. Benito, G. Hütsi, R. Kipper, J. Pata and S. Põder, *Sensitivity estimation for dark matter subhalos in synthetic Gaia DR2 using deep learning*, *Astron. Comput.* **41** (2022) 100667, [2203.08161].

[41] T. Schlegl, P. Seeboeck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, *Unsupervised anomaly detection with generative adversarial networks to guide marker discovery*, .

[42] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, *J. Mach. Learn. Res.* **22** (2019) 57:1–57:64.

[43] I. Kobyzev, S. Prince and M. A. Brubaker, *Normalizing flows: An introduction and*

review of current methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** (2020) 3964–3979.

[44] E. M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174, [1708.02949].

[45] R. E. Sanderson, A. Wetzel, S. Loebman, S. Sharma, P. F. Hopkins, S. Garrison-Kimmel et al., *Synthetic gaia surveys from the fire cosmological simulations of milky way-mass galaxies*, *The Astrophysical Journal Supplement Series* **246** (Jan., 2020) 6.

[46] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2256–2265, PMLR, 07–09 Jul, 2015.

[47] J. Wolleb, F. Bieder, R. Sandkühler and P. C. Cattin, *Diffusion models for medical anomaly detection*, in *International Conference on Medical image computing and computer-assisted intervention*, pp. 35–45, Springer, 2022.

[48] H. Zhang, Z. Wang, Z. Wu and Y.-G. Jiang, *Diffusionad: Denoising diffusion for anomaly detection*, *CoRR* **abs/2303.08730** (2023) .

[49] J. Wyatt, A. Leach, S. M. Schmon and C. G. Willcocks, *Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 650–656, June, 2022.

[50] S. Han, X. Hu, H. Huang, M. Jiang and Y. Zhao, *Adbench: Anomaly detection benchmark*, *ArXiv* **abs/2206.09426** (2022) .

[51] Y. Zhao, Z. Nasrullah and Z. Li, *Pyod: A python toolbox for scalable outlier detection*, *Journal of Machine Learning Research* **20** (2019) 1–7.

[52] S. Ramaswamy, R. Rastogi and K. Shim, *Efficient algorithms for mining outliers from large data sets*, in *ACM SIGMOD Conference*, 2000.

[53] M. C. Smith, N. W. Evans, V. A. Belokurov, P. C. Hewett, D. M. Bramich, G. Gilmore et al., *Kinematics of sdss subdwarfs: structure and substructure of the milky way halo*, *Monthly Notices of the Royal Astronomical Society* **399** (2009) 1223–1237.

[54] A. Blaauw, *On the origin of the O- and B-type stars with high velocities (the "run-away" stars), and some related problems*, **15** (May, 1961) 265.

[55] J. G. Hills, *Hyper-velocity and tidal stars from binaries disrupted by a massive galactic black hole*, *Nature* **331** (1988) 687–689.

[56] J. Bovy, *galpy: A python library for galactic dynamics*, *The Astrophysical Journal Supplement Series* **216** (2014) .

[57] G. Green, *dustmaps: A Python interface for maps of interstellar dust*, *The Journal of Open Source Software* **3** (Jun, 2018) 695.