Bulk classification and analysis of TESS γ Doradus stars using machine learning methods

JESSIE TAYLOR

MPhys Physics with Astrophysics

Master of Science by Research

University of York

Physics, Engineering, and Technology

August 2023

Abstract

Supervised machine learning was used to classify γ Doradus stars present in the TESS-SPOC data pipeline, to investigate the efficacy of fast bulk classification of pulsating stars in minimally-processed data. A fully-connected neural network was set up and trained as a binary classifier using built catalogues of previously confirmed pulsators of four types present in the γ Doradus instability strip, as well as known non-variable stars. During validation, the model obtained a 94.4% precision score. The trained network was then input with binned Lomb-Scargle periodograms of 173,398 stars within the ranges $T_{\text{eff}} = 6500$ - 7500 K and Tmag = 9.0 - 12.0. The total time for the network to classify all candidates was 11.1 minutes, with a pre-processing time of ~5 ms per lightcurve. The probability distributions and HR diagram positions of the output classifications were analysed and a small set of the results visually verified. It was found that a classifier confidence threshold of 77.4% was most suitable and yielded 7,749 potential γ Doradus candidates, representing 4.47% of the analysed set. Of 100 of these visually checked, only seven were misclassified EB stars, and three likely rotational variables. Eight of the classifications showed evidence of p-mode pulsations suggestive of γ Doradus and δ Scuti hybrids. This investigation shows a way in which only minimal treatment of TESS lightcurve data is necessary for high quality classifications of pulsators, allowing for quick identification in large datasets. This is important, as a large and diverse pool of candidates is necessary for thorough investigation and testing of stellar evolution models.

The code used for this research, as well as a CSV file containing the catalogue are available at `github.com/jessie-taylor/gDorClassifier`

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Acknowledgements

A huge thanks to my supervisor Emily Brunsden, you have guided me and helped me believe I can do things I never previously thought possible.

Phil Lightfoot, thank you for being so generous as to add me to your calendar, even when sleep wasn't granted that opportunity.

Jacan Chaplais, your patience, knowledge of python documentation that can rival Google, and guidance in coding and neural networks has been invaluable. I owe much of my accomplishments to the momentary distress in your eyes as I described the paths I was pursuing; without those, I wouldn't have achieved even half of what I have now.

Philip Jennings, Claire Rutter, Shannon, and Wendy, you are beautiful friends and you have all helped me so much.

This research made use of Lightkurve, a Python package for Kepler and TESS data analysis (Lightkurve Collaboration, 2018).

Contents

1 Introduction

Stars play a crucial role in enhancing our understanding of the Universe. Throughout history people have tracked their movements, and tried to connect celestial patterns with earthly occurrences. As techniques have advanced in astronomy, a much greater accuracy has been gained in tracking these movements and fluctuations, but for different reasons. Not only are we now able to track the positions of these stars, but our ability to take precise measurements of properties such as variations in brightness and exact colour temperature have led to a plethora of new information, which can be used to reveal a lot more about the Universe as a whole. Far from the primitive naked-eye observations our ancestors were limited to, modern tools, such as precise photometry and spectroscopy, have allowed us to make great leaps in our knowledge of the cosmos. We are now able to use advanced tools and methods to peek *within* stars and uncover vital information such as which elements make up stars. This has therefore led to an understanding of how stars and their fusing of simple hydrogen and helium eventually leads to the formation of planets and dust clouds, and the evolution of our universe on a larger scale.

1.1 Variable Stars

The first known recorded account of variability in stars was three millennia ago, by ancient Egyptian scribes. Contained within the Cairo Calendar is a recorded period of 2.85 days, which evidence suggests is the periodicity of the eclipsing binary star Algol [1]. It wouldn't be until 1638, when Johannes Holwarda noticed and recorded the variability of the star Omicron Ceti, that variable stars would be recognised in modern astronomy[2]. Since then, the field and our understanding has expanded greatly, with different techniques employed in order to probe these stars for further information. We are able to use photometry to accurately track luminosity changes in even the faintest of stars, and space-based observatories have allowed for constant observations, unrestricted by the day-night cycle of the Earth, or pseudovariability originating from terrestrial atmospheric perturbations.

Most stars only have small changes in observed brightness over periods shorter than their evolutionary timescale, usually originating from magnetic effects causing spots on their surfaces, coupled with the rotation of the star. However, there are processes which can occur within some stars that cause intrinsic oscillations, resulting in observed variability at the surface from expansion and contraction of the photosphere. In the case of certain types of these pulsators, oscillations can even originate from areas as deep as those directly surrounding the core. These offer a unique opportunity of insight into the layers and processes that exist below a star's surface, as the variability presented at the surface is therefore dependent on their internal structure, size, and composition. This allows us to probe deep within the stars, solely through observations of surface variations [3].

1.2 Oscillation Mechanisms of Stars

There are three main mechanisms which can cause a star to begin to oscillate. The simplest is in binary star systems, where the tidal forces due to the orbits of the stars in the system induce oscillations in one-another. The ways in which these are presented depend upon the nature of the orbit - a circular orbit forces oscillations in the star with a period equal to the orbital period of the sister star. The oscillations observed in a more eccentrically orbiting system are more complicated, but are still driven by the same mechanism, and hence maintain strong relation to the orbital period [4]. Although the tidally driven pulsations in binary systems are a result of external forces, rather than any intrinsic processes, the pulsations still propagate through the entire body, and so are still valuable in the study of variable stars.

For pulsating solar-like stars and some red giants, the main driving mechanism is through stochastically driven pulsations in a process called convective blocking. These occur due to a sharp interface gradient between a radiative zone and a convective envelope. This occurs in stars with a deep convection zone that extends to the surface of the stellar body. The luminosity transport from the inner radiative zone to the threshold of the more opaque convective zone leads to an area with modulation in the opacity, causing to small oscillations at the inner edge of the convective envelope [5]. These cannot be driven throughout the entire convective envelope due to its large size. However, evidence of these pulsations can be present at the surface due to global resonance that travels through the rest of the envelope [6]. Due to the fact that the global oscillations presented at the surface of the star are due to the resonance, these oscillations are lower in frequency than oscillations created through other mechanisms.

In some types of stars there exist H and He ionisation zones, which have a very high opacity to the radiation being emitted from the core. The gas in these regions absorbs that energy and swells, breaking hydrostatic equilibrium. Once expanded, the region becomes less opaque due to ionisation of the element in the zone and thus energy is able to radiate through and escape. As the energy escapes, and it cools, inward pressure exerted from the surrounding layers overcomes the outward pressure of the expansion, and so the expanded region falls back to its smaller volume. During the recompression, the material in the layer recombines and so restores the original opacity, allowing for the cycle to begin once again [7]. During this process the envelope is acting as a heat engine, and produces waves of higher frequency than those observed in stochastically driven stars. This mechanism of driving stellar pulsations is referred to as the κ-mechanism[3].

During most of their life-cycle, stars can typically be modelled to be in hydrostatic and thermodynamic equilibrium. That is, that there are two main forces in stars - the internal pressure gradient pushing outward and the gravitational force acting radially inward [8]. In hydrostatic equilibrium, both of these forces are equal, leading the star to be stable in radius, and not collapse under the force of its own gravity, nor does it expand until the point it can no longer be considered one celestial body. The forces which provide the outward pressure gradient come from the thermal gas pressure of the star, with the energy provided to the gas originating from within the core. This is where nuclear burning occurs and the star fuses elements that make up its composition in nucleosynthesis, emitting large amounts of energy outwards to other areas of the star [9].

However, stellar interiors are much more complex than just a fusion core and a single envelope of gas. As the energy travels from the core, it must pass through a series of layers, which are either radiative or convective in their ability to dissipate heat. Factors such as the elemental composition of a star can have an effect on the structure within, which can be dependent on the age; the generation it belongs to; and cosmic location in which it is formed. Low-mass main sequence stars have large convective envelopes as their most exterior layer, where the heat energy generated within is transferred to the surface for radiative emission. Stars also contain multiple elements of different masses, especially as their cores fuse fuel elements together to create more massive elements through nuclear burning. This can lead to envelopes of different materials settling at different distances from the core, which may have material drawn from them into other layers through the motion of convection [10]. These envelopes often have different behaviours in terms of how energy is transferred through them, due to factors such as differing densities and opacities to heat energy[11].

In intrinsically variable stars, while hydrostatic equilibrium holds true over longer timescales, in shorter periods there are processes which occur within the different regions within stars which can lead to pulsations. This leads to certain types of stars expressing unique oscillations that are characteristic of that type of star. This can be visualised on a Hertzsprung-Russell (HR) diagram, as in Fig. 1.1, where the luminosity of stars is plotted against the effective temperature, showing distributions of different types of stars in a way which helps us understand their evolution. As shown in the diagram, it can be seen that stars which exhibit different types of pulsations can be grouped together spatially on the plot, according to their effective temperatures and luminosities. The HR diagram is also useful in that it can allow us to extrapolate other information about stars, such as mass.

Different pulsators exhibit variability originating from different areas and processes in the star, but in all of them, this comes from a short-timescale breaking of the hydrostatic equilibrium of the body, where the internal pressure and gravitational force no longer are equal. The driving mechanisms for this can vary depending on the type of star, but in all, the energy required for driving pulsations is delivered through mechanisms which shed thermal energy from central regions to the surface of the star for emission. This energy is lost during pul-

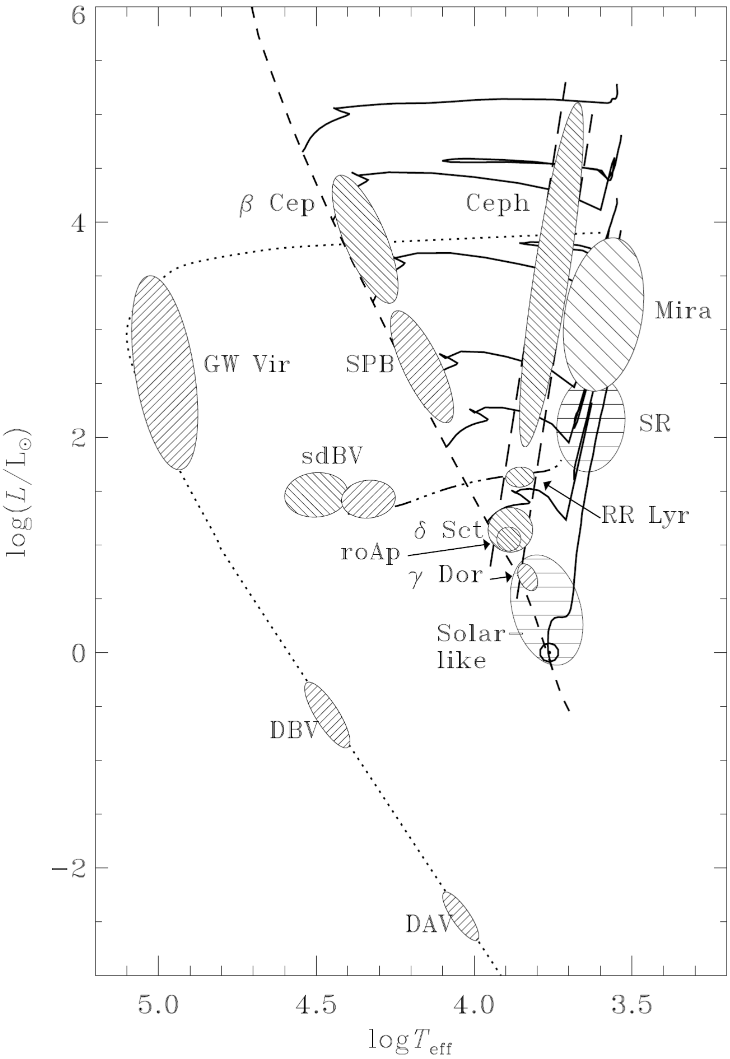

Figure 1.1: Hertzsprung-Russell diagram showing where groups of different types of pulsators lie on the plot. The dashed line from around 4.7 - 3.7 log $T_{\rm eff}$ represents the zero-age Main Sequence, and the two parallel long-dashed lines indicate the boundaries of the Cepheid instability strip. Image from [3].

sation cycles rather than through surface emission, as the star also damps the expansion with a *restoring* force [3]. This restoration is done through one of the two types of restoring force present in pulsators: pressure or gravity.

Depending upon the type of restoring force that occurs, the pulsations are referred to as p-mode or g-mode. In stars which exhibit p-mode pulsations the restoring force is pressure, and the motions of the waves are acoustic in nature. These are typical in higher pulsation frequency stars, such as δ Scuti and β Cephei stars. The p-mode pulsations are confined to exterior regions of stellar bodies due to changes in the density of stellar material as they try to travel towards the interior, because of the corresponding change in sound speed. This means that p-mode pulsations are also more sensitive to changes in the most outward regions of the stars [3]. Their sibling, g-mode pulsations, or gravity-mode pulsations, are when the restoring force is buoyancy. This leads to g-mode oscillations being more trapped in more central regions of the star, beneath the convective envelope, and hence much more sensitive to conditions deep within the star, around the core [12]. An exception here is in stars where the g-mode pulsations are trapped by a deep convective envelope surrounding the core, such as in Red Giants [13]. Some g-mode pulsations are found in stars exhibiting lower frequency oscillations, such as Slowly Pulsating B stars, and γ Doradus. Fig. 1.2 shows these two types of modes of pulsations and how they are constrained to certain parts of stars. As g-modes travel, and are sensitive to conditions around the core, they can offer a unique insight into the conditions at the very centre of the star. Due to the size of the convective zone in the Sun and other solar-like stars, these modes are currently unable to be detected on the surface due to the evanescence of the region. This makes their amplitudes much less than those detected from granulation [14]. However, in some stars, the outer convective layer can be harmonically excited by the g-mode activity and present with detectable pulsations at the surface [3].

There exists another variation in the way a star can oscillate; with radial or nonradial modes. In radial modes of oscillation, the entirety of the star surface swells and contracts with spherical symmetry. For the fundamental mode of radial oscillation, this includes the entirety of the stellar interior also moving, with the core as a node and the surface as an antinode. For higher modes of oscillation, nodes exist radially outward from the centre in which the matter is not displaced, as in a longitudinal standing wave. In nonradial pulsations surface nodes are present, around which the surface expands and collapses, in opposite directions on each side of the node, as on a transverse standing wave. The number of nodes present in both radial and nonradial modes is dependent on the type of driving mechanism present, as well as internal structure [3].

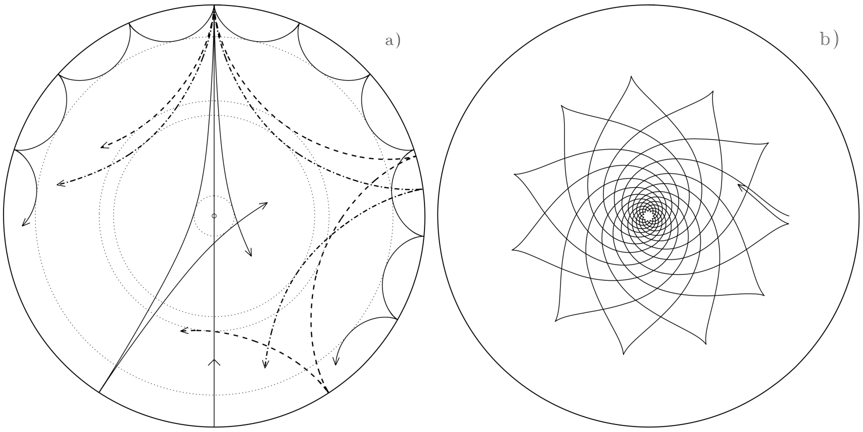

Figure 1.2: Ray path diagram of p-mode pulsations (a) and g-mode pulsations (b), modelled at higher frequencies to demonstrate how they are confined to different regions in the stellar interior. In (a) the path of the p-mode waves is curved away from the interior by changing sound speed as it penetrates deeper. In (b), the g-mode waves are seen to be trapped by the first convective layer from the core, and so reflect back through the central region near the core. Despite the trapping of g-modes within the inner layers, their motion can still excite modes of oscillation in the outer layers, producing oscillatory behaviour at the surface. Image from [3].

1.3 Asteroseismology

Asteroseismology is a relatively young field of astronomy in which stellar pulsations are analysed to probe internal structure. As pulsations originate from the interiors of stars, and then propagate through the rest of the body through reflections, studying these pulsations through the frequencies and modes presented on the surfaces of stars allows us to infer the internal structures of stars. Asteroseismology first began emerging as a field around the early 1990s, stemming from observations of the Sun during the 1970s using observations of changes in limb darkening at the edge of the solar disk [15]. This discovery opened up new pathways into investigating the nature of stars, and allowed a view into internal properties which could be used to confirm theories such as stellar evolution and internal rotation [16][17]. Using techniques from these initial helioseismological studies, it wasn't long until these techniques began to be used on more distant solar-like oscillators [18][19], as well as other types of oscillators [20][21][22].

Because of the importance of g-modes in enabling us to "see" within the stars, work is being done on finding methods to detect these in the Sun [23], but results so far have been inconclusive [24]. Being able to infer the internal structure through accurate mode-identification of as many types of stars as possible allows us to test models of stellar evolution, and learn more about the conditions in which stars formed, and how.

1.4 Measuring Variability in Stars

Whether a star is oscillating in radial or non-radial modes, these cyclical expansions present changes at the surface, which can be measured through observation at a distance. Early examples of scientific measurements of variability were taken using ground-based telescopes, observing over an extended period of time and taking photometric measurements of the brightness, which could then be put together to manually create a lightcurve [25]. As the stars vary in size and luminosity during their pulsation cycle, these can be measured from our vantage point on Earth. Spectroscopy is also a very useful tool and can be used to perform mode identification from the Doppler shifting of the light emitted from the surface in motion [26].

Ground-based observing of stars presents many challenges in classical astronomy, and even more so in asteroseismology. From the ground, atmospheric effects limit the quality of data obtainable from observations. Limiting factors, such as the day-night cycle of our planet, preclude the detection and recording of oscillations with periods longer than the available night-time or visibility of the star due to the Earth's rotation. This introduces aliases in the data. On top of this, ground-based observing is also weather-dependent.

Space-based Observations

To solve the problems of ground-based observing, we are now able to utilise space telescopes to observe and collect data for us. One of the earlier space-based observation programmes was *Hipparcos*, launched in 1989, with the primary mission of precision astrometry [27]. Its continuous scanning of the sky meant that the data also could be used for a secondary purpose - to detect variability in distant stars, which was exploited and allowed for the discovery of unprecedented numbers of new pulsators in a short time [28][29]. Another mission of great importance in asteroseismology was *Kepler*. *Kepler* was launched in 2009, this time with the express purpose of obtaining time-series lightcurves from continuous observations of around 150,000 stars in a fixed field now known as the *Kepler* Field. With an array of 42 CCDs viewing the field, it performed a photometric survey with the intention of finding exoplanets by recording transits observed across the stellar disks. The nature of its mission and greater precision than previous instruments made *Kepler* of great importance in asteroseismology and again allowed for great leaps in pulsation detection and asteroseismic analysis [30][31][32][19].

The TESS mission

Following on from the work of *Kepler*; launched in 2018, the *TESS* spacecraft was designed as its replacement, allowing for a much more comprehensive view of the sky. Rather than observing one set area of the sky as *Kepler* did, *TESS* uses an array of four cameras to take photometric data in strips of the sky before being rotated to the next region [33]. The pattern of observations, as well as periods of observation for the initial mission are shown in Fig 1.3. *TESS* observations yielded both short-cadence data of two minutes for specific target stars, and long-cadence 30 minute full-frame exposures for the rest of the field of the cameras [34]. This makes the cadence of the full-frame data have a Nyquist frequency of 24 d^{-1}, suitable for viewing higher frequency p-mode pulsations in stars studied in asteroseismology. The extended *TESS* mission now yields 10 min cadence full-frame lightcurve data, and 20 s short-cadence data for selected targets [35].

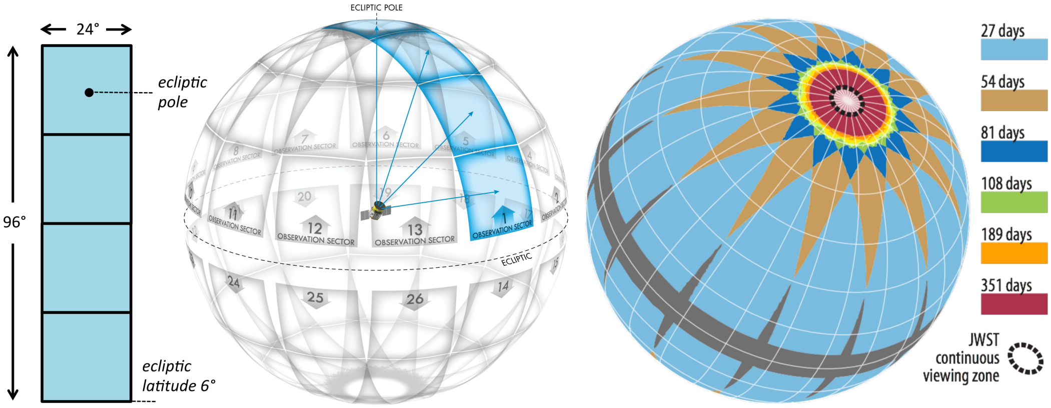

Figure 1.3: Left: The field the four TESS cameras are simultaneously able to observe. Centre: Projection of the instrument view onto the hemispheres of observation, showing each observation sector. Right: The observing strategy, showing the period of time each observation sector will be viewed for. Image from [33].

1.5 Classifying stars

While the main focus of asteroseismology is to study the interiors of stars, it is important that the types of pulsators useful in these studies can be easily identified in the large amounts of data that are now available. Having large catalogues of these pulsators is imperative to allow for a greater sample of stars to study and test our models on, spanning the entirety of the instability strip. This allows for a greater understanding of properties of stars that exhibit these oscillations, and the diversity of conditions under which certain modes to arise. This, in turn, enables the confirmation and improvement of oscillation theory models.

As different types of pulsations and frequencies are unique to each type of oscillator, identification of these types of stars by performing a Discrete Fourier Transform (DFT) on the time-series data is able to be carried out [36]. In Fig

1.4 it can be seen that the different modes of pulsation occupy unique frequency ranges. As these ranges are dictated by the structure and composition of the star, and are coherent across each pulsator type, different patterns of oscillation present in a periodogram (constructed as the square modulus of a DFT) can be used to identify the type of star. The g-mode pulsations present in KIC 11145123 (Fig 1.4) are that of the type of star γ Doradus. This is a type with prominent deep g-mode pulsations, making it a good candidate for asteroseismic analysis. The presence of both g-mode and p-mode pulsations present signify that this is a hybrid of two types of pulsator.

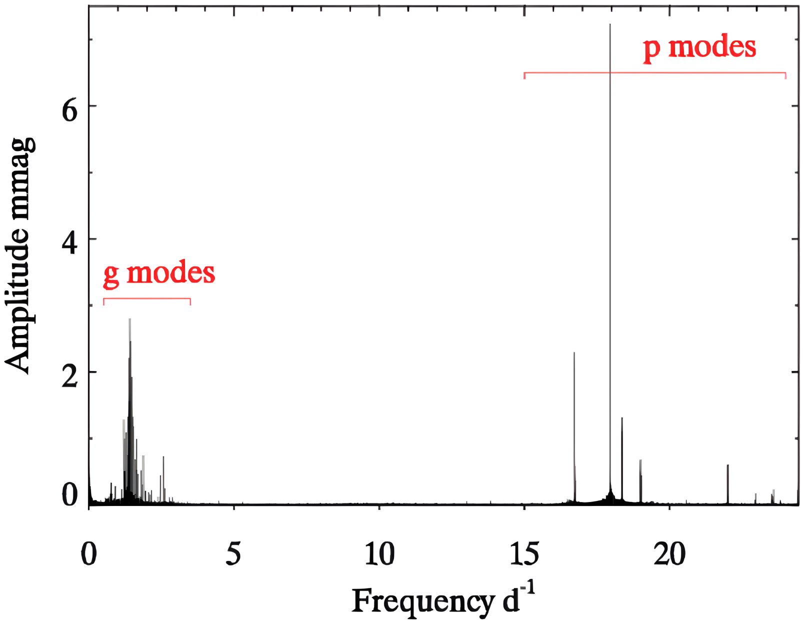

Figure 1.4: DFT amplitude spectrum of Kepler lightcurve of star KIC 11145123, a main sequence A class star. Both g- and p-mode frequencies can be seen present in different regions of the spectrum, as this is a hybrid of two types of pulsator. Image from [37].

1.6 γ Doradus

γ Doradus are a type of pulsator first defined in 1999 which exhibit high-order non-radial g-mode pulsations in the frequency range 0.3 to 3 d. They are spectral type late-A to early-F stars located on the cooler edge of the instability strip, and are of mass between around 1.3 to 2 $M_\odot$ [38][39]. They have convective cores surrounded by radiative zones, which are enveloped by a shallow convective region, leading to convective blocking being a driving mechanism of pulsations [38][40]. Because of the deep near-core pulsations intrinsic to this type of star, they are of great interest in asteroseismology [41], given that the period spacings present are sensitive to conditions at the centre of the star [42]. Rotation periods of this class of star are also similar to their pulsation periods, which may affect their oscillations [3]. Recent research has been able to use *TESS* data to find the core rotation rates, which are an important component in understanding stellar evolution models [43] [44].

1.7 γ Doradus - δ Scuti Hybrids

Other than solar-like oscillators, γ Doradus shares part of its region on the HR diagram with another type of pulsator, δ Scuti [45], as shown in Fig 1.5. δ Scuti are p-mode pulsators which usually can easily be distinguished from the lower frequency pulsators of the γ Doradus class. However, in some γ Doradus stars, p-mode pulsations are able to be excited. KIC 11145123 (Fig 1.4) is an example of a γ Doradus-δ Scuti hybrid that exhibits both g-mode and p-mode oscillations of these two types of pulsators. As both γ Doradus and δ Scuti share the same region on the HR diagram, consideration must be made of hybrids of the two when identifying new stars, and how their existence may affect classification.

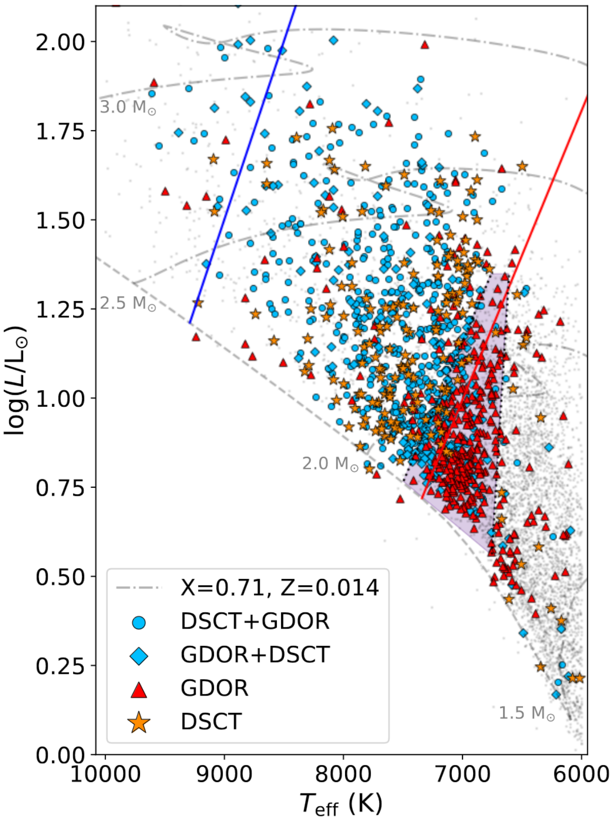

Figure 1.5: A HR diagram showing the distributions of pulsators within the instability strips of γ Doradus (shaded area bounded by dashed lines) and δ Scuti (blue and red lines). The solid line represents the zero-age main sequence. This plot was obtained from [46], in which these stars were identified. The grey dots represent stars not identified as either type of oscillator (nor a hybrid of the two) in the study.

1.8 Machine Learning

Machine learning is a term used in computing for any method of solving a problem or performing a task where the solution is not explicitly coded. Unlike traditional programming, machine learning involves creating algorithms in a way that allow the machine to learn how to perform the task on its own. This has many applications, and is especially common now, as data collection methods have become more advanced and automated, allowing data sets to reach sizes where human analysis would be too time consuming to be feasible. One of these applications which has great use for astronomical data is classification of observed objects in data gathered from large sky surveys [47]. There are two main approaches in machine learning for classifying data: supervised and unsupervised; and which is most suited to the problem depends on the nature of the task and the data being used [48]. For unsupervised learning, algorithms find features in the data which can be used to organise the data into different clusters (discrete groups) with no human guidance on input data [49]. [50] Supervised learning, in contrast to unsupervised learning, reduces the number of possible sets the data can be categorised as by requiring labelling of the input data, in effect enforcing which features the algorithms look at to ensure the categorisation is as desired [51]. As pre-labelling the input data defines and constrains the output classifications of the algorithm, supervised learning is therefore the best applied method when the desired output classifications are already known.

An increasingly common way of solving supervised classification problems in the past decade has been to employ machine learning utilising Artificial Neural Networks (ANNs), consisting of layers of neurons. These are nodes in the network which perform mathematical operations on the incoming information, which then feed the outputs to each neuron in the subsequent layer. The layout of an ANN is shown in Fig 1.6. The input layer consists of neurons the data features being fed into the network are initially sent to, i.e. the pixels in an image for an image classification problem. Each neuron in the input layer then sends its outputs to the next layer, by connections which each have an associated numerical weight, and it repeats this process for each layer until arriving at the output layer. In an ANN set up for a classification problem, there are an equal number of output neurons to classifications that are required. The way in which the network learns is by then checking the outputs of those classifications against what the desired output should be for the training set of data and adjusting the weights using an optimiser function at the end of each epoch (once all data has passed through the algorithm). Note that this is only true in supervised learning; in unsupervised learning the network is looking for patterns in the data itself. Machine learning algorithms are also analysed by measuring loss during the learning procedure, which is the measure of accuracy of classification of each individual piece of data during an epoch; essentially an error. Each layer between the input and output layer is referred to as a hidden layer. Due to the large number of nodes available in an ANN for analysing data, it is one of the preferred methods for classification of complex data [52]. Large networks of neurons processing input information allow for greater ability to identify and classify from non-linear relationships in data, as well as pick out complex features without explicit instruction. This is an advantage over other algorithms such as linear and logistic regression, which assume a linear relationship dependent on one variable [53][54]. As the output of an ANN is a probability of input data belonging to a classification, it also has an advantage over other types of machine learning approaches in classification problems, such as support vector machines and logistic regression which only provide the predicted classification as a “yes” or “no” [55][56]. A major disadvantage of ANNs however is their opacity, where the decisions being made in classification cannot be seen. This is in contrast to a decision tree algorithm, where splits toward each decision branch of neurons are made for each individual feature, and are transparent in their decision making. However, this transparency comes at the cost of being limited in classifying complex features, which may require a large number of splits which is also time consuming to set up [57]. It is these factors which have made the utilisation of ANNs for processing complex astronomical data increasingly popular in contemporary research.

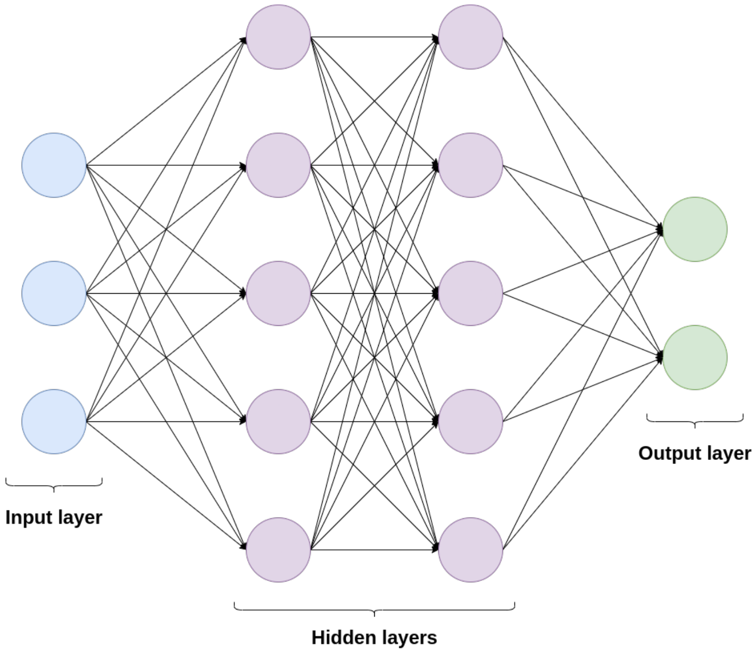

Figure 1.6: A schematic of the structure of a shallow neural network containing two hidden layers, with neurons represented by circles. Between each layer neurons pass their outputs and associated weights (represented by the connecting arrows) into the next layer through a connection to each of the neurons in the next layer. In a classification problem, the input layer neurons are fed the input data to be classified, and the output layer neurons represent the different classes. The numbers of neurons in each layer are unique to each ANN.

The operation each neuron performs is given by,

$$h_i = \sigma \sum_{j=1}^{N} (V_{ij} x_j) \tag{1.1}$$

where h_i is the output of the neuron i with inputs x_j, σ the activation function, N the number of inputs, and V_{ij} the weights [58]. In a hidden layer, an extra term is added within the sum in Eq 1.1 as a bias, so the neuron may have a constant value that is trainable [59]. The activation function, σ, is a non-linear function and may be selected according to the problem, a common example of one is the Sigmoid function [60].

There is much room for flexibility in the architecture of a neural network. The number of hidden layers needs to be selected when setting up an ANN, which is referred to as the depth of the network. More complex problems, such as image classification, may require a deep neural network in order to be able to obtain meaningful results; whereas simpler problems may only require a shallow network with as few as one hidden layer. The number of neurons in the input layer must equal the number of features from the input data, and the number of output neurons is equal to the number of classes. However, the numbers of neurons in the hidden layers may be arbitrary. The dimensions of the hidden layers are usually selected according to how well the network performs on the data, and adjusted to find the best configuration. Likewise, there are many other parameters, called hyperparameters, a network has which control how the model learns. The best values for these must be found through trial-and-error, to find the most optimised version of the network.

A problem in machine learning on small sets of data is the greater risk of overfitting, where the algorithm learns how to classify the specific data fed into it, rather than by features it should ideally be identifying [61]. This can be checked for by utilising a separate verification dataset, which the model has not seen previously, to measure accuracy of classification.

In the age of continuous full-sky surveys, in order to be able to fully use the vast amounts of data being collected, we must employ automatic systems, such as machine learning. Previous research has been done on utilising neural networks to analyse star data in bulk from both the *Kepler* and *TESS* missions, but these commonly use heavy pre-processing which limits their throughput [41][62][63] [64].

1.9 Aims of this research

In order for asteroseismic analysis and research to be done on these types of pulsators, it is important that catalogues of pre-classified stars are available. Previous work has mostly been done exclusively on *Kepler* data, in order to identify types of pulsators crucial in understanding the mechanics of stars. With the launch of the TESS mission and its accompanying *Tess Input Catalog* (TIC) now listing $\sim 1.5 \times 10^9$ possible point source targets in its current revision, *TIC-8* [65], it is more valuable than ever to have a way of quickly and reliably processing data in order to be able to identify objects of interest. As TESS is optimised for observations of brighter stars (unlike *Kepler*), this also opens avenues for ground-based follow up investigation of catalogues created from its data.

The goal of this research was to find a method of identifying candidate stars that have a strong indication of γ Doradus pulsations being present. Due to the size of the datasets involved, a method which required minimal processing and human review was sought. The end result of this was a catalogue of γ Doradus candidate stars which can be further verified and studied in future asteroseismological research.

2 Data Collection and Processing

In order to be able to use TESS data for machine learning, it must first be processed in a way which allows the network to be able to gather meaningful information from it. TESS data is pre-processed before publication, however, depending on the type of research being carried out, it may require further processing.

2.1 TESS-SPOC Pipeline

The data from TESS, once received from the satellite through the *Deep Space Network*, is sent to the TESS *Science Operations Center*, where the raw data is then processed in the *Science Processing Operations Center* (SPOC) at the NASA *Ames Research Center*. Here, the full-frame images collected by TESS are calibrated, and used to produce calibrated pixels, lightcurves, and centroids for all target stars. The full-frame images, recorded with a cadence of 1800s (30 minutes), were used for the initial TESS mission long-cadence data provided through the SPOC pipeline. The TESS-SPOC pipeline data was used in this research due to the availability of Pre-search Data Conditioning Simple Aperture Photometry (PDCSAP) flux data, in which systematic errors have had corrections applied [66]. The files are finally made available through the *Mikulski Archive for Space Telescopes* (MAST) [66], and are accessible through the *Lightkurve* Application Programming Interface (API) [67].

2.2 Candidate Selection

Training Data

For the training data, a selection of different star types were used for training the model. Rather than attempting to include every type of variable star, a training set was built which consisted of stars which may either be misclassified as γ Doradus, or could also be common in the dataset which the final model will be trying to classify γ Doradus stars from. These included four types of stars in the γ Doradus instability strip in which variability may be observed, as well as non-variable stars. The other types of objects which may exhibit variability and hence were included in the training set were Eclipsing Binaries (EBs), δ Scuti, and RR Lyrae. Both EB and RR Lyrae were chosen due to the possibility of the observed variability frequencies being in the ranges expected from γ Doradus, and therefore having a high chance of being misclassified as γ Doradus unless the model was explicitly trained using them. δ Scuti stars were included because of the possibility of γ Doradus - δ Scuti hybrids being misclassified as γ Doradus, despite typically having much stronger p-mode pulsations. While stars previously classified as hybrids of γ Doradus and δ Scuti were not included explicitly in the training data, many were noted to be present in some of the δ Scuti examples. Contamination of the γ Doradus training data with hybrid stars was deemed undesirable as it could lead to confusion between the classes of γ Doradus and δ Scuti, and result in misclassifications. For this reason, targets in the γ Doradus set which contained significant p-mode pulsations were removed from the training data. The target stars for the different classifications were obtained from a review of catalogues created in previous research (as listed below). Each catalogue was evaluated for its suitability for training an accurate model. It was important that each training set was not only as large and accurate as possible, but also contained a diverse range of each type, to avoid training the model only to detect "perfect" examples, rather than all those that may be present in the TESS data. Visual inspection of Lomb-Scargle periodograms of stars in these catalogues was then performed to ensure misclassified stars were not included, in order to reduce the chance of the classifier model learning from the wrong types of stars for each category. Examples of each of the types of stars present in the final training set are shown in Fig 2.1.

The γ Doradus and δ Scuti obtained were from those classified in Skarka et al. (2022) [46]. These data were chosen due to the study's robustness, having been classified using the lightcurve data and then further verified with spectroscopic analysis. This is of particular importance in the γ Doradus set, where rotationally variable stars may appear similar to intrinsically variable γ Doradus stars. Rotationally variable stars are stars whose observed brightness fluctuations arise from spots on the star's surfaces, with the brightness variation arising from the rotation of the star. From this, 387 γ Doradus star TIC IDs were obtained, and 162 δ Scuti IDs (differences between these numbers and the numbers in the paper's data result from there being some *Kepler* IDs not having a corresponding TIC ID in the *Simbad* database). An extra set of δ Scuti IDs was then obtained from Murphy (2019) [68], who took care to remove other types of pulsating stars that are also present in the other classifications that will be used in the training of this model. A final visual review of periodograms of the δ Scuti targets revealed many which contain strong g-mode pulsations, as well as some with no visible distinguishable pulsation features; these were manually removed from the training set. While more γ Doradus stars could have been obtained using Debosscher et al. (2011) [63], it is a set known to contain errors, so was not used in this work [41]. Examples of three γ Doradus light curves in the final training set and their associated Lomb-Scargle periodograms are shown in Fig 2.2.

The non-variable set was also obtained from stars Skarka et al. (2022) did not identify as any type of pulsators. This catalogue was chosen for this set as these stars are all within the same T_{eff} ranges as the other types of stars used to train the model (A-F spectral types). From this, a set of 1,587 non-variable TIC IDs were obtained. After download, the periodograms of these were then visually checked to remove any in which significant frequencies were present, with peaks above around half the amplitude of background noise. Those which had visible pulsation activity of multiple frequencies in one region were also removed to ensure no intrinsic pulsators were included in the set.

EB stars were obtained through the *Kepler Eclipsing Binary Catalog* (KEBC). This is a catalogue built up through several revisions [69][70][71] so can be relied upon for reliable classification. This catalogue contains many stars which TIC IDs were unavailable for through the Simbad database [72], however, 2920 TIC IDs were able to be obtained for the EB training set. EBs were deemed of particular importance to have in the training set, as their pulsation frequencies can often overlap with γ Doradus, but they generally have a more distinctive shape in the rest of their Fourier transforms. Both detached and contact EB systems were included in this set to reduce the chance of the trained model misclassifying any type of EB as a γ Doradus pulsator.

The RR Lyrae training data contained both type-ab and type-c, these were included due to the likelihood of them being in the same effective temperature T_{eff} ranges as γ Doradus [73]. Type-ab RR-Lyrae examples were obtained from Drake et al. (2013) [74], as well as from Nemec et al. (2013) [75], which included type-c RR Lyrae pulsators, although fewer due to the lack of them being detected in the *Kepler* field. In total, 1378 RR Lyrae TIC IDs were added to the set.

Unclassified Data

As there is such a large amount of TESS data available, in the interest of timeliness the candidates for analysis were selected according to two criteria in order to create a reasonably-sized set of data to analyse: T_{eff}, and TESS magnitude (Tmag). The data that was obtained for the training set was analysed for each Tmag recorded in the TIC, and the mean of the full set was calculated to be $\overline{\text{Tmag}} = 10.3$, with only few outliers lying outside of the $9.0 \leq \text{Tmag} \leq 12.0$ range, so this range was chosen for the criteria. The Tmag of the stars matching the majority of the training set is important, as it can greatly effect how the lightcurve of the star appears. A star with a brighter magnitude will have a greater signal-to-noise ratio, and too bright may overexpose in the CCD. The T_{eff} ranges chosen were 6500-7500 K, as that is the region of the instability strip in which γ Doradus stars lie [40]. The Python package *Astroquery* [76] was used in order to obtain TIC IDs for targets within these ranges.

2.3 Obtaining Data from the TESS-SPOC Pipeline

Once the TIC IDs for target stars were collected for each set, the *Lightkurve* API was used to download lightcurves from the MAST archives [67]. All lightcurves used in this research were long-cadence data, as the corresponding Nyquist frequency of $24\,\mathrm{d}^{-1}$ allows for enough frequency space to be seen to distinguish other pulsators from γ Doradus.

For the training sets, the numbers of each type of star that were able to be obtained through the TESS-SPOC pipeline are represented in Table 2.1. Due to very low availability of some types of these stars through this pipeline, the search was not limited to one certain sector, however, the distribution of stars that were obtained was 89% in sector 14, as shown in Fig 2.3. The uneven distribution of target stars across the sectors is largely due to sectors 14 and 15 crossing into the *Kepler* Field [77], where the majority of the target stars are located.

The distributions across the sectors of the data obtained for the unclassified stars is shown in Fig 2.4.

Table 2.1: Counts of how many of each classification were present in the final training set. Note that the differences in numbers in the final count vary from those in Fig 2.3 due to removal of stars with pulsations present in the non-variable classification.

Classification	**No. of lightcurves**	**Source of catalogue**
Non-variable	659	[46]
γ Doradus	387	[46]
δ Scuti	708	[46][68]
EB	501	[71]
RR Lyrae	46	[74][75]

2.4 Preparing the Data

Once the lightcurves were obtained through the *Lightkurve* API, a Lomb-Scargle periodogram was created for each of the data [78][79]. Feeding a periodogram into the network, rather than the full lightcurve, allows for less data points to be used while still retaining information of the periodicity of the original lightcurve. The Lomb-Scargle periodogram is used frequently in astronomy, as it remains reliable when performed on data that is unevenly sampled. It also has the advantage of being able to be quickly performed, taking an order of magnitude less than other methods [74].

Both the power and corresponding frequency values from the periodogram were then then stored in a Python dictionary, along with the TIC ID, lightcurve metadata, and target classifications (for the training data). These were then dumped into a JSON format to allow for fast access. The types of stars were separated into five different classes during the beginning of training so problems between classifications of specific types of stars could be identified, and the final set represented as a binary problem with γ Doradus with class identifier 1, and all others in class 0. The training data was then split to retain 10% of each classification for verification during training.

Multiple separate sets of mock training data were also prepared, so the architecture of the network could first be tested on simpler data. These were designed so it could be made certain the model was able to identify periodograms with amplitude peaks in specific areas without the intricacies and variation of the training data. These consisted of sets of periodogram-like structured data, containing noise as background, and one amplitude peak at a specific frequency, individual to each target classification they were assigned. As the unbalanced set of the real training data could also cause problems during training, this was also tested by creating sets mirroring the imbalance in data in the real data. Further sets were also created with peaks closer to the amplitude of the noise, simulating low SNR data, to ensure the model could distinguish these before the training data was used.

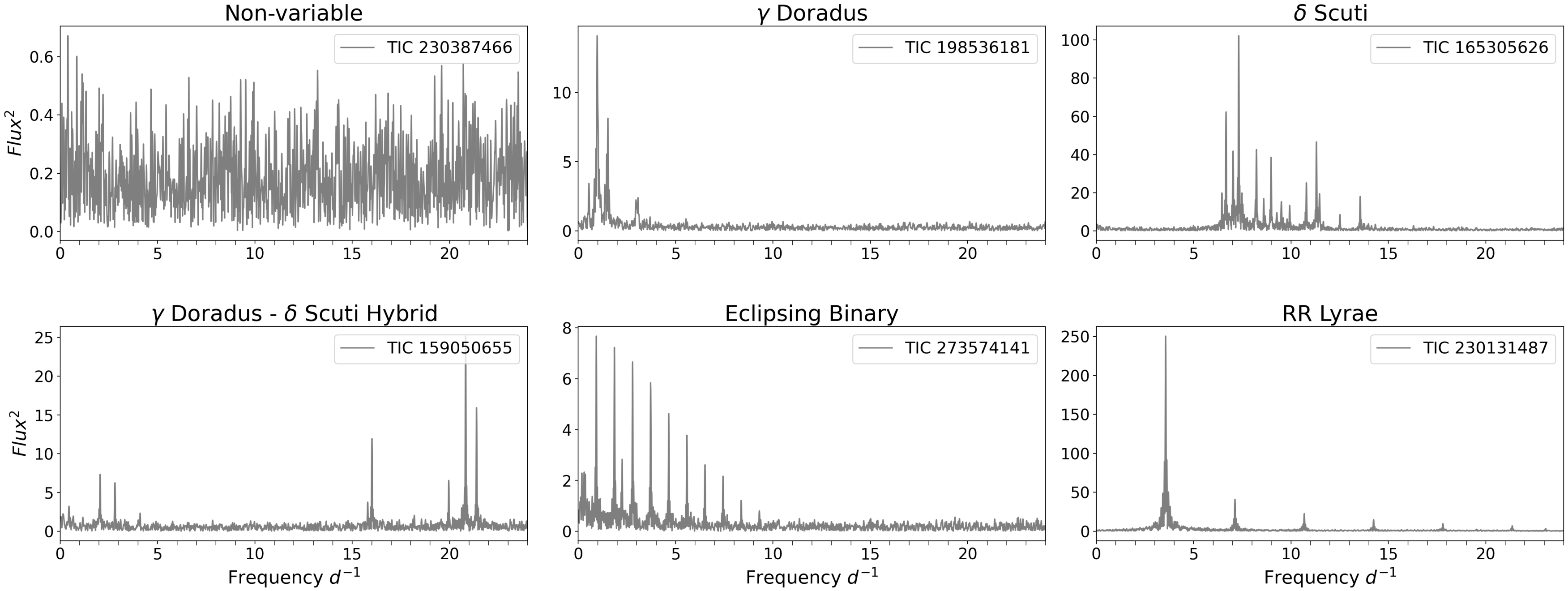

Figure 2.1: Lomb-Scargle periodograms of the types of stars included in the training set. Hybrids are also shown here as they were left present in the δ Scuti set. The variations in the y-axis scales are due to variations in the SNR of the lightcurve data, as well as strength of pulsations.

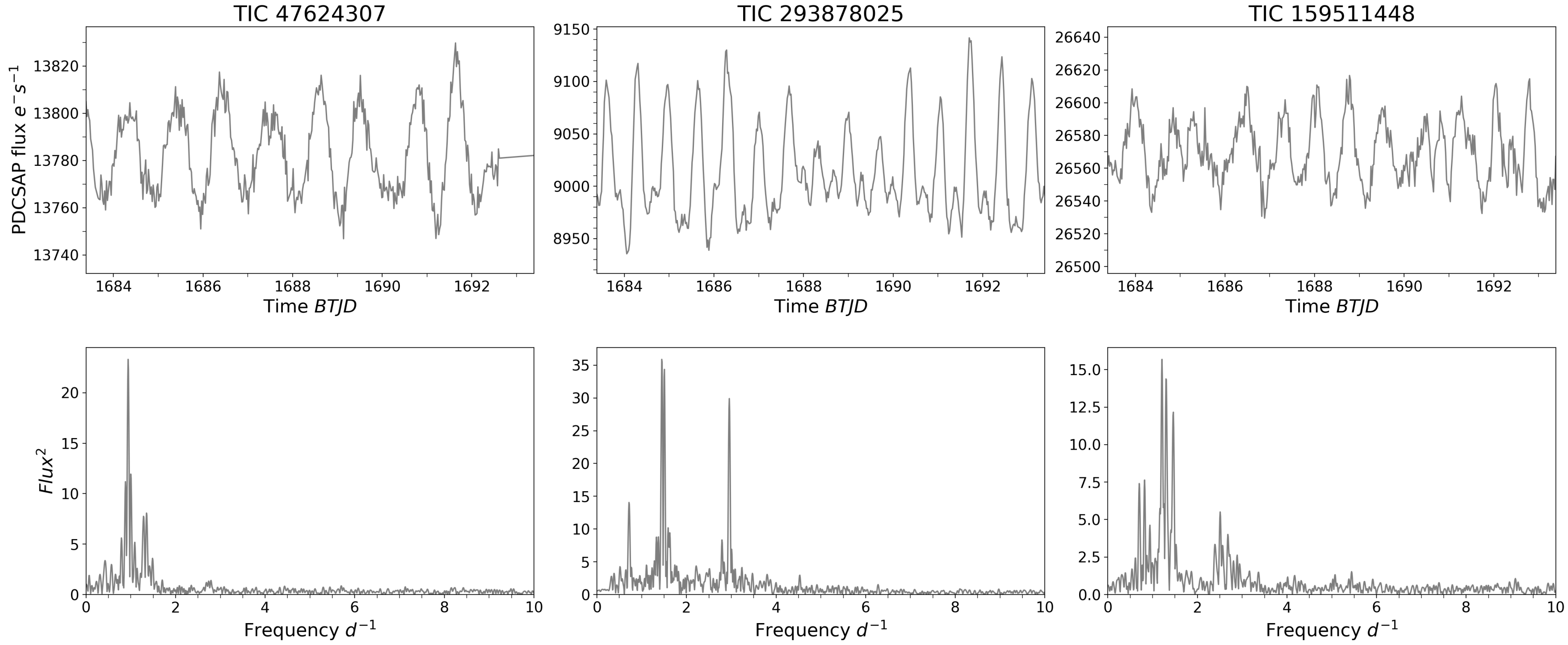

Figure 2.2: Examples of γ Doradus stars that were included in the set. The top figures are of the TESS-SPOC light curves, using Pre-search Data Conditioned Simple Aperture Photometry (PDCSAP) flux, which is corrected for instrumental variation. The light curve x-axes are in Barycentric TESS Julian Date (BJD - 2457000.0). The corresponding Lomb-Scargle periodograms are displayed beneath each light curve. Note that only 10 days are displayed on the light curve plot, to show more detail of the structure of variability.

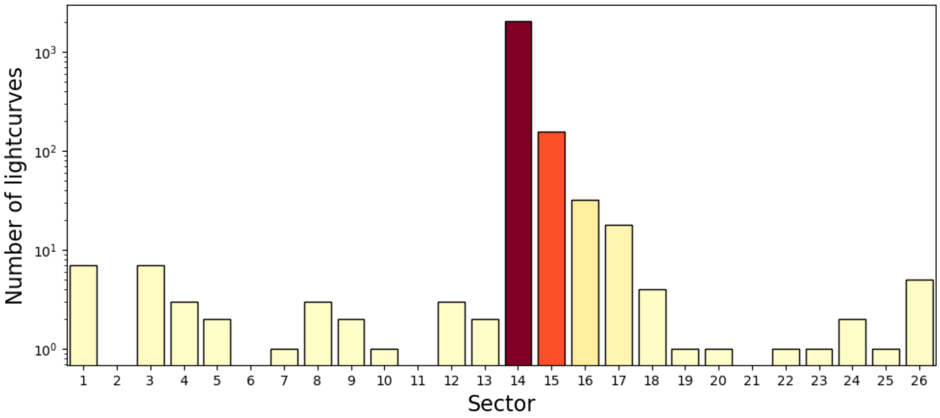

Figure 2.3: Distribution of sectors in which the downloaded lightcurves originate from in the SPOC routine. The y-axis is represented logarithmically. The number of lightcurves gathered from sector 14 was 2045, ten times more than that of sector 15 with 159 lightcurves yielded.

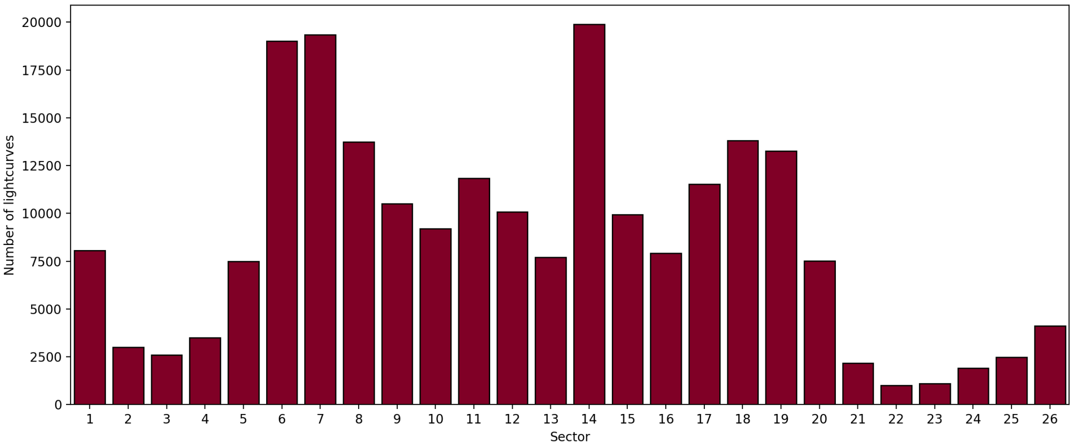

Figure 2.4: Distribution of sectors in which the unclassified TESS-SPOC data were obtained for classification. The upper bound of sector 26 is due to 1800 s cadence data being unavailable for the TESS mission extension.

3 Setting Up & Training the Model

The model was created using PyTorch [80] in Python 3.10. In order for the network to be able to effectively and efficiently learn to classify data, an architecture that was appropriate for the work being done needed to be chosen.

3.1 Initialising the Network

The model was constructed using fully-connected layers, with the Parametric Rectified Linear Unit (PreLU) activation function, as it is well suited to classification problems [81]. As there were a low number of training examples for the network to learn from, dropout was included. This forces the network to ignore random nodes during training and reduce the chance of overfitting the data, where the model learns the noise of the training set and thus is only able to properly classify the original training data [82]. A Softmax activation function was implemented on the final output layer of the network, as the desired output of the network is probabilities of the classes [83]. Due to the potential problems of training a model with such an unbalanced training set, Sigmoid Focal Loss was used for the loss function due to its ability of weighting each class according to the balance when calculating loss [84].

Rather than using a smoothing function on the frequency spectra, which could remove discrete pulsation frequencies present in some stars, a binning method was used when running the data through the model. Having this treatment at this point, rather than in the pre-treating of the data allows for flexibility in changing this value, and thus finding the most effective value. A higher number of bins, such as that close to the number of original data points allows for greater detail of the data, but negatively affects performance. A lower number of bins effectively smooths noise, thus reducing the amount of attention the network can pay to it. It also reduces the time taken to run each example through the network, as the network can be smaller in dimension. There is a minimum value at which the data are useful, as when too low the peaks are also smoothed over and averaged to a much larger area. The most effective value for the binning was investigated during the optimisation of the network.

3.2 Optimising The Network

In order to find the best configuration for the network during training, it was run multiple times and hyperparameters adjusted between each run, to determine which values produced the most accurate results on the training and verification data. The performance of the models were then analysed by looking at the loss and performance over time. The metric which was predominantly used for the performance was precision, as it is much more important that the data classified as γ Doradus contains fewer incorrect classifications, than it missing some γ Doradus pulsators. However, recall was also measured to ensure the number missed was within a reasonable range. Initially the sets of mock data were used to verify the architecture was set up correctly for this type of data, and then the real data introduced. Through feeding it through successively more difficult tasks, each of the hyperparameters could be focussed on and adjusted one at a time, so the differences each individual hyperparameter made on the training and classification could be properly understood. This approach also meant that low SNR, imbalance, and ability to discern peaks and peak ranges could be treated as separate problems and each optimised individually.

4 Results

Through the random search method, the hyperparameters which gave the most suitable optimisation were found. It was found that reducing the number of bins for the periodogram data had a great effect on the precision of the trained model, with more bins being necessary in order to obtain suitable precision. This is most likely due to the loss of finer detail in the input data at lower bin numbers. The most precise model was obtained with a resolution of 2639 bins for each DFT (reduced from ∼3014). This was the number of input neurons which were used for the final version of the network. Each subsequent hidden layer was found to be most effective when it had 854 nodes, and a total of 14 of these layers were found to give the best results when training. This configuration gave a training time per epoch of 27s on a single GPU configuration.

For the hyperparameters in the Sigmoid Focal Loss function, it was found that a γ value of 3.7977 was most the effective to account for the imbalance in the target classifications of the data it was trained and verified on. The α value was not included in the random search, to reduce the number of hyperparameters varied and thus reduce time required for the random search, but preliminary tests found the most effective value was 0.50, so that was used in the final model.

4.1 Training the model

Once the best hyperparameters were obtained using a random search, the model was trained for 450 epochs and the results from each epoch recorded and analysed. The values of precision and recall obtained from each epoch were then used, along with loss trends, to select the best epoch to use as the model for classifying the TESS-SPOC data. As can be seen in Fig 4.1, there is an inverse relationship between the precision and recall for each training epoch. A higher number of true positives (true γ Doradus classifications) identified came at the cost of also having the potential γ Doradus set polluted with more false-positive misclassifications. As this model is designed to be used on very large datasets, a lower precision can potentially mean misclassifications in amounts far too high to be able to manually verify, and so a higher precision at the expense of recall was chosen for the model.

The final model chosen was from the 10th training epoch and had precision, recall, and F1 scores of 94.4%, 44.7%, and 0.607, respectively, during validation testing during training.

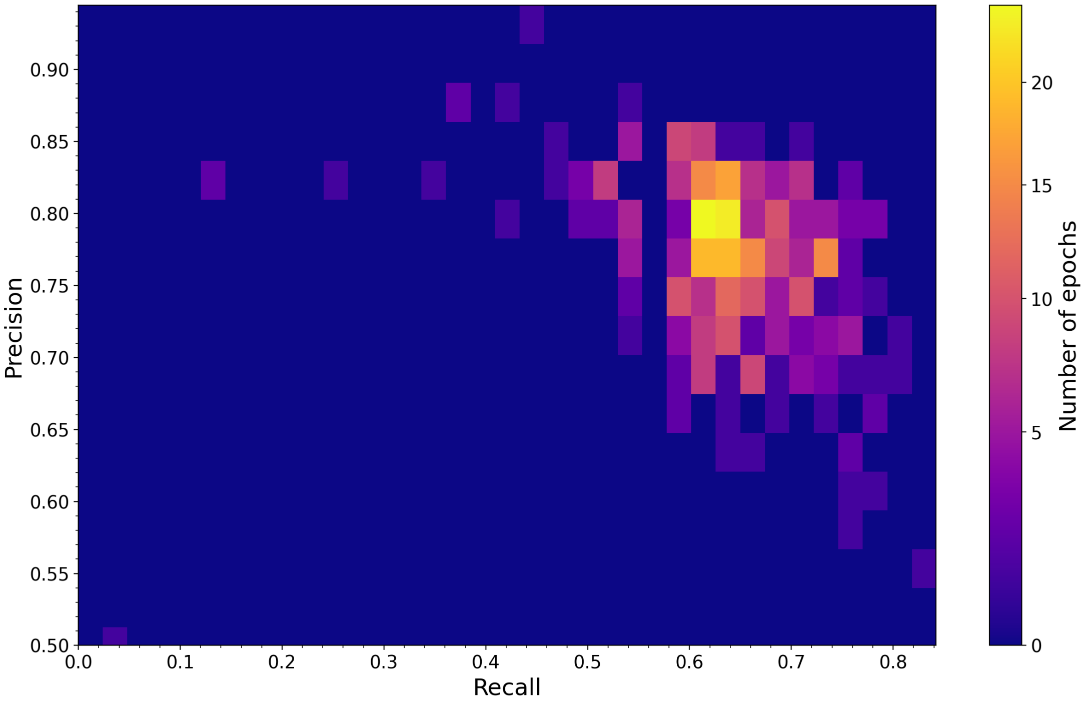

Figure 4.1: Heatmap showing distribution of training epochs and their resulting precision and recall values from validation tests of each model iteration during training. Values from all 450 training epochs were used in this plot.

A confusion matrix showing a breakdown of the class labels for each star in the validation data is shown in Fig 4.2. The low recall score is made obvious here, showing the false negatives (γ Doradus stars predicted as non-γ Doradus) as highly populated. The high precision of the trained model is also visible in the confusion matrix, with all but one other type of star having no false positive predictions. The only false positive prediction was from an EB star, which is unsurprising as it is common for variations in EB systems to have long periods and therefore frequencies which occupy the same region as γ Doradus stars, as can be seen in Fig 2.1. The false positive EB star compared to a true γ Doradus star can be seen in Fig 4.3. This star, TIC 122717066, is a spectroscopic binary star [85], and the more compact “comb”-like structure of its periodogram is not visible at the resolution offered by the data used in this research, which most likely contributed to the network being unable to identify it correctly.

The probability distribution in Fig 4.4 shows the classifier’s confidence of predicted γ Doradus predictions from the training validation data, and the data misclassified in the confusion matrix in Fig 4.2 can be seen highlighted in red. It can be seen that there is a clear trend forming in the distribution of correct classifications, with the peak lying at lower confidence levels, which is expected with a model with a lower recall. Notably, the highest confidence obtained for a star was for the incorrectly classified EB star, which may be due to the lack of visible structure of the periodogram.

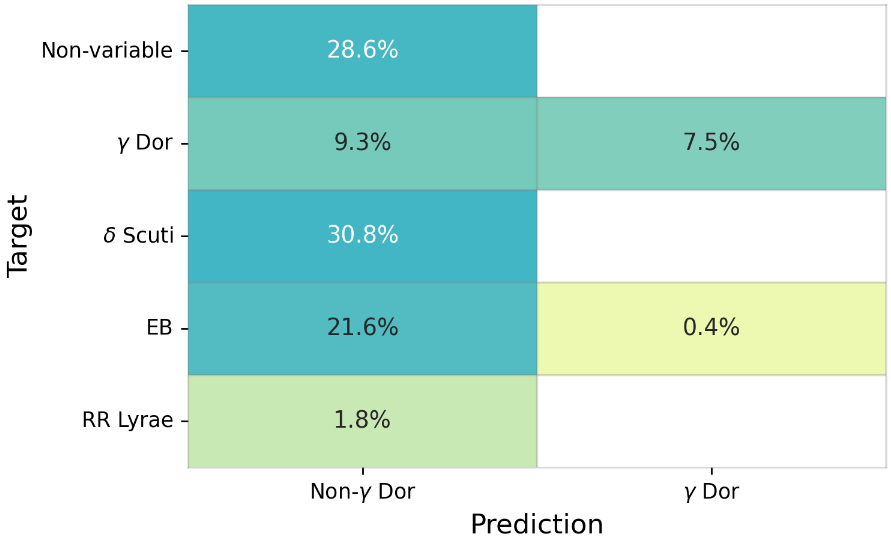

Figure 4.2: Heatmapped confusion matrix with all stars in the validation set shown with their original labels. Unmarked white boxes indicate no classifications made for that cell, and the percentages in each cell show how much the data with the corresponding training labels (targets) and classifier predictions represent of the total verification data. Cell colour is proportional to the value.

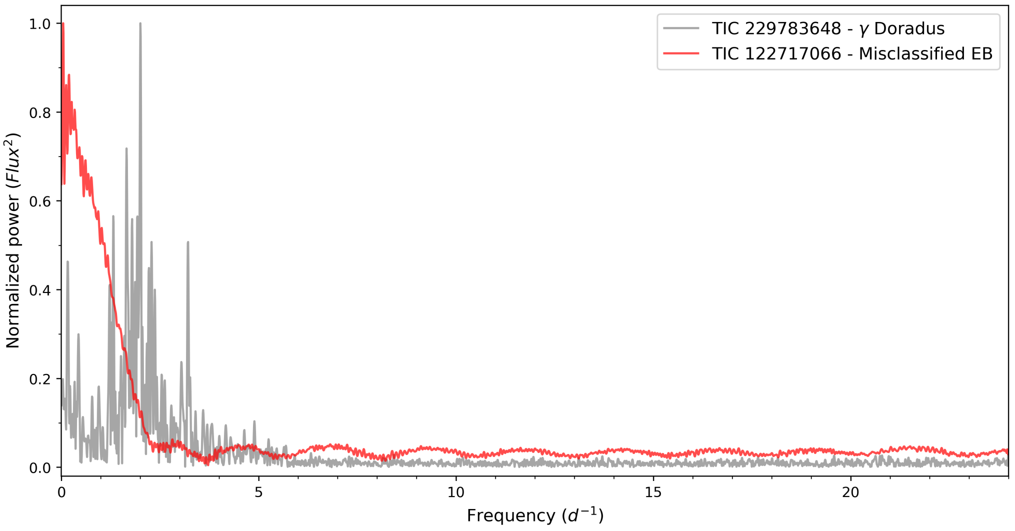

Figure 4.3: A comparison of a misclassified EB star compared to a correctly classified γ Doradus star. Powers of the frequency spectra have been normalised, as the EB has two orders of magnitude greater power on the periodogram data. It can be seen that the most prominent frequencies lie in similar regions, which is most likely the reason for the misclassification. This is not a typical EB periodogram, which an example of can be see in Fig 2.1.

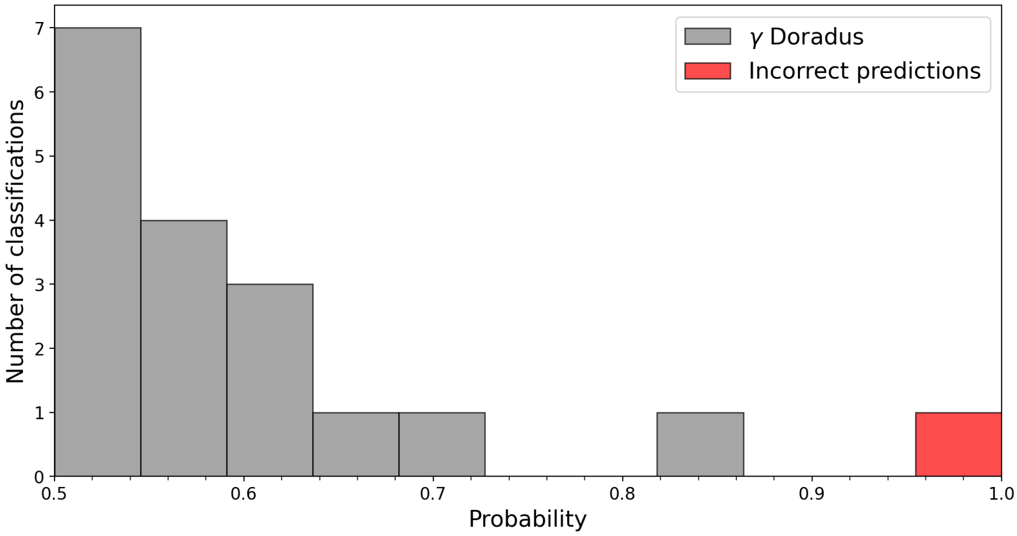

Figure 4.4: Probability distribution of stars in validation set classified as γ Doradus. The incorrect classifications are highlighted in red.

4.2 Classifying TESS Data

In total, 173,398 stars were run through the classifier model, taking a total amount of time of 11.1 minutes (3.843 milliseconds per star), not including pre-processing time to generate the Lomb-Scargle periodograms (~5 ms per lightcurve). 30,993 of the TESS-SPOC stars were classified as potential γ Doradus candidates, representing 17.87% of the total set. The classifier's confidence scores for each of the γ Doradus predictions can be viewed as a probability distribution in Fig 4.5. From this, it can also be seen that half of the classifications were obtained with a confidence level of 64.0% or above, and only 25% of them held a confidence level of 56.6% or below. The top 25% of classifications were able to be identified with a confidence of 77.4% or above, representing 7,749 total potential candidates, or 4.47% of the total set.

Fig 4.6 shows the stars classified with a 50% confidence threshold plotted on a HR diagram in comparison to the distribution of stars the classifier identified as non-γ Doradus. The region in the centre of the instability strip can clearly be seen as containing the largest amount of γ Doradus classifications, verifying that the model used was able to correctly identify pulsators of this type. However, there still remains a large amount of stars at lower temperatures which were classified as γ Doradus on the main sequence. This problem is rectified by increasing the confidence threshold for a classification to only include positive classifications in the 75th percentile, as seen in Fig 4.7. Here, the vast majority of classifications are confined to well within the boundaries of the instability strip, and this can be seen more clearly in Fig 4.8, where the distribution of positive classifications peaks only within the temperature ranges 6900 to 7200 K.

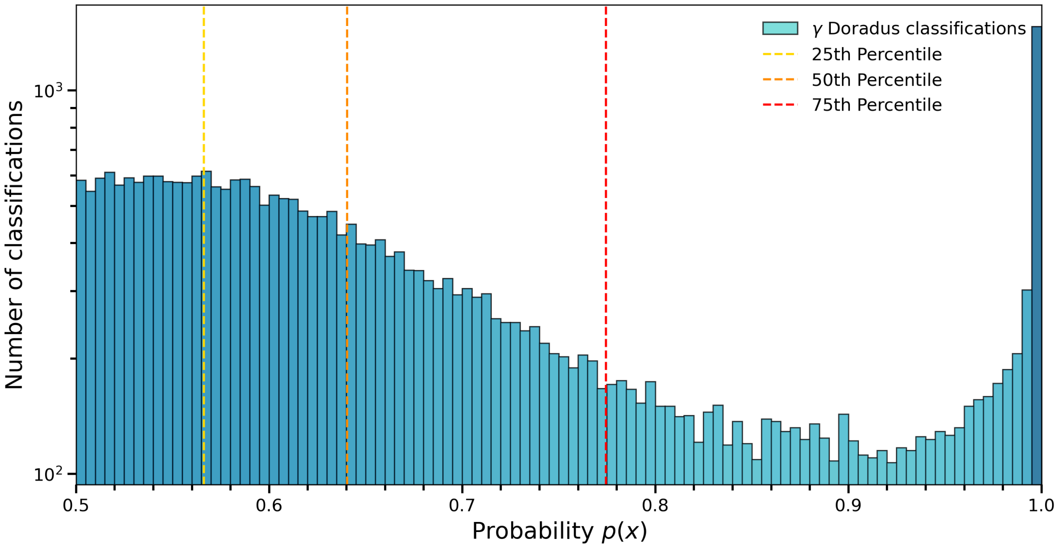

Figure 4.5: Histogram of probability density for stars classified as γ Doradus from the TESS-SPOC data. The percentile cut-offs are marked by the coloured dashed lines, with 50% of the γ Doradus classification lying above the orange line. Due to the large number of high confidence classifications near 1.0, the y-axis has been displayed logarithmically. Despite the large spread of classifications at lower probabilities, the majority of classifications lie at very high confidences, as evidenced by the location of the 50th percentile marker.

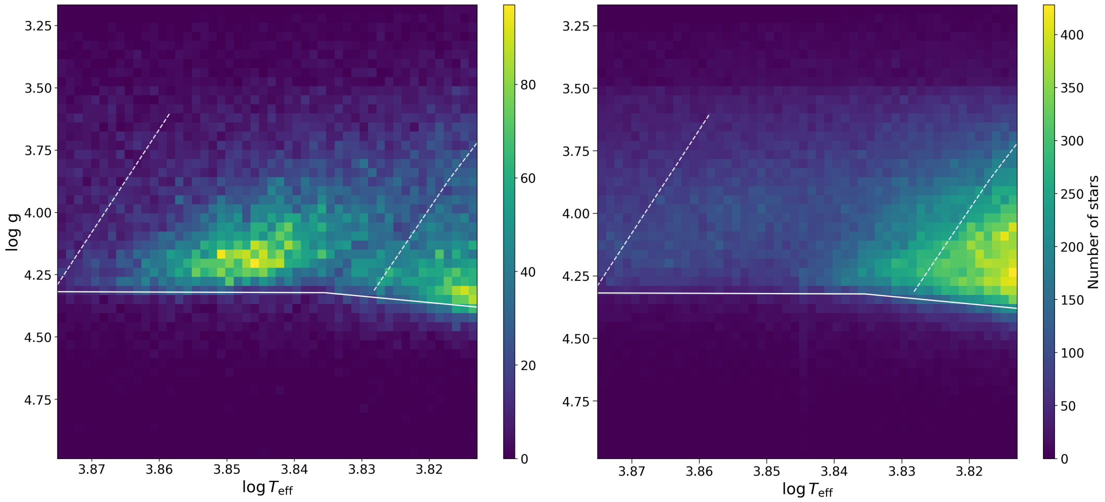

Figure 4.6: Two HR diagrams, represented as heatmaps. The left panel shows stars classified as γ Doradus by the classifier, and the right panel contains stars which were not identified as γ Doradus. The solid line represents the zero-age main sequence, and the dashed line the boundaries of the γ Doradus instability strip, obtained from [86]. The majority of the stars classified lie within the instability strip, with the exception of at lower temperatures, where the dataset contained the most stars. These classifications are obtained with a 50% confidence threshold.

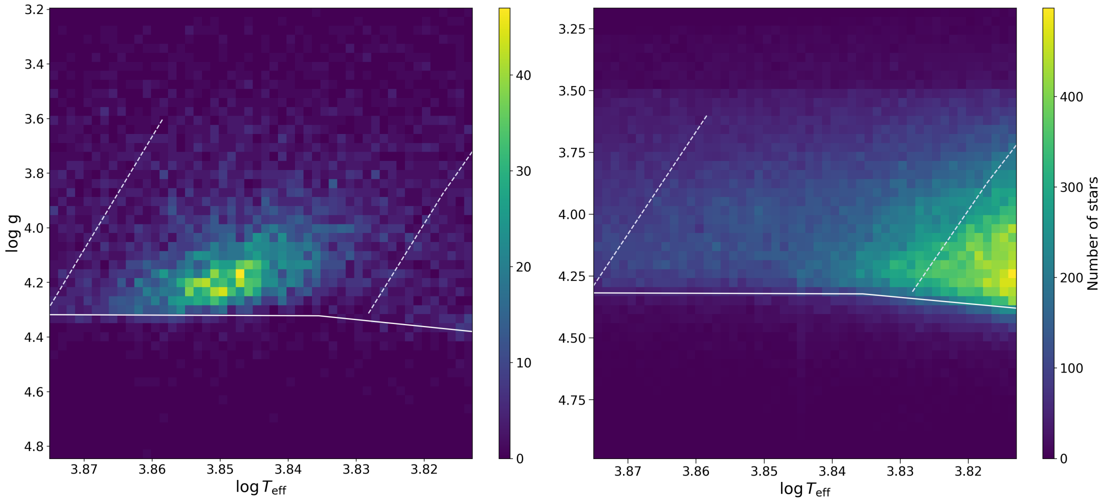

Figure 4.7: As in Fig 4.6, this plot represents the TESS-SPOC data in a HR diagram, with the γ Doradus classifications on the left, and the rest of the data run through the classifier in the right panel. This is with the confidence level set at a higher threshold of 77.4%, representing the 25% of the data the classifier gave the greatest probability for, as obtained from Fig 4.5. The majority of data eliminated by this higher confidence threshold lie outside of the instability strip.

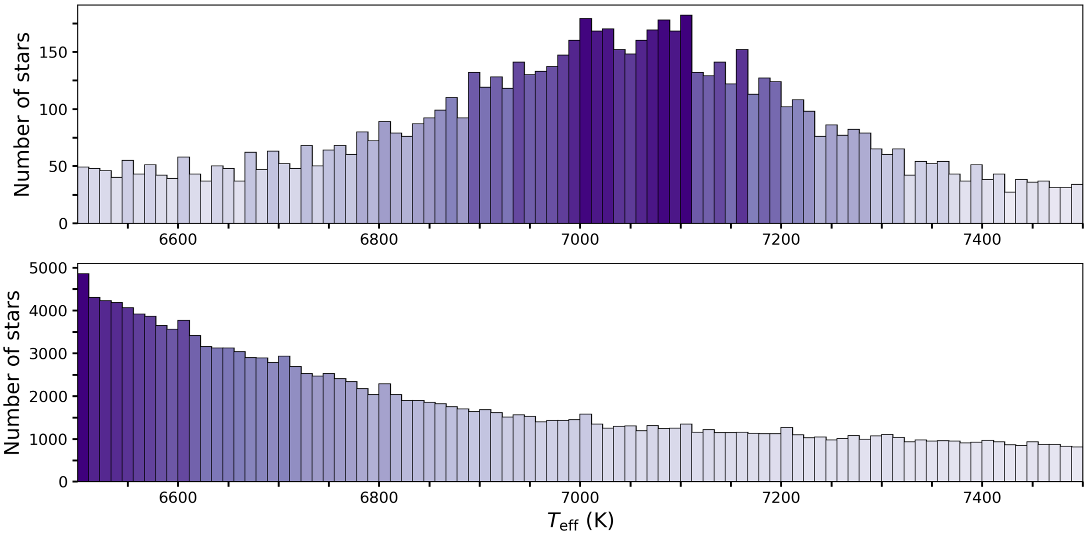

Figure 4.8: Histogram of T_{eff} distributions of stars classified as γ Doradus (top frame), and those not identified as γ Doradus pulsators (bottom frame) for the 75th percentile of classification confidences. The colour gradient of the plot maps to bar sizes. The highest density of γ Doradus classifications are within the ranges of 6900 and 7200 K. Introducing a higher confidence threshold has greatly reduced the amount of lower temperature classifications that are present outside the instability strip limit in Fig 4.6.

5 Discussion

From the distributions of γ Doradus classifications on the HR plots in Fig 4.6 and Fig 4.7, the classifier network can provide results that would be expected of a distribution of γ Doradus variables. This is further confirmed through the T_{eff} distribution of classifications in Fig 4.8. Due to the density of classifications in the lower temperatures outside of the expected region in the 50% threshold HR diagram, it is obvious that a higher classification threshold is able to provide more reliable results and so that is what was used for the rest of this analysis. The shape of the distribution of the positive classifications in the HR diagram for higher confidence classifications agrees very well with the theoretical instability strip, which drifts to lower temperatures as the stars age; ∼6850-7360 K at zero-age, and ∼6560-7000 K for terminal-age main sequence stars [87]. This can be seen as the distribution tends towards lower temperatures as the log(g) value decreases.

From the results with the higher confidence threshold applied, γ Doradus classifications made up 4.47% of the total stars within the T_{eff} and Tmag ranges. This is in agreement with, although slightly lower than, the 6.01% proportion found in *Kepler* data using feature-based machine learning in [41], and 7.48% from non-neural network classification of DFTs of *Kepler* data in [40].

The plot of the recall vs precision during training, Fig 4.1, shows that in the model there was little trade-off between precision and recall, as most values were high on recall - this may have been influenced by the fact that a limited validation set was used. There is a slight trend in lower precision at higher recall, which will be due to the model “overclassifying” non-γ Doradus stars as γ Doradus. This was avoided in the final model by ensuring that the precision of the version of the model used was more important than the recall. For building a catalogue, it was deemed that the reliability of given γ Doradus classifications was more important than their overall number.

Using a small and limited training set of data can have detrimental effects on a network’s ability to train. Problems that may arise because of this include a propensity for the network to very quickly learn the full set of training data and begin overfitting. It can also make trends in accuracy metrics more difficult to identify as each example represents a much larger part of the set of data than if a large set was used. This leads to much less stable metrics over training epochs, as each individual reclassification can shift the values significantly. Because of this, wider trends had to be used in determining what was happening during training, and so the point at which overfitting may be occurring was harder to detect. However, building a better picture of how the network was training by utilising more than one metric (training and validation accuracy, precision, recall, and loss) provided more reliable insight. Imbalances in training data sets also leads to issues with learning. In this research, as the problem was approached as a binary problem, the minority class was γ Doradus, meaning that the type of star it was most important for the classifier to learn was much more difficult for it to be able to properly understand the nuance and diversity of these types of stars [88]. The implementation of the sigmoid focal loss function allowed this problem to be managed, but it is obvious that this is not a replacement for a good quality and numerous dataset.

The result of finding higher numbers of input and hidden neurons is unsurprising, as lower numbers of input neurons require less detail to be present in the input data. This however does come with the cost of longer processing times, although as the number of hidden network layers was able to be kept relatively low, the final computation time of 3.843 ms per star classification is well within reason and makes this approach more than suitable for classification of pulsators from large sets of data.

The issue of confusion between EB stars and γ Doradus is an issue with the model, and has been in previous classification research using machine learning [41][64]. As machine learning is an “opaque” method, i.e. it is impossible to know the criteria the network is using for classification, it is hard to identify why the EB star was misclassified with such a high confidence in the verification data. It is most likely that the presence of variability in the same frequency ranges as γ Doradus stars is the greatest factor in this. From a visual inspection of 100 of the classifications, seven of these were found to likely be EB stars, and all had their highest peaks (if not all peaks) within the γ Doradus pulsation frequency range, supporting this hypothesis. Three other non-γ Doradus pulsators were identified, with frequency power spectra of potentially rotationally variable stars, however, these types of stars can be difficult to distinguish even through visual inspection, as discussed in [46]. It is a possibility that training for rotational variables by adding a catalogue of these into the training data may have improved the ability for the network to discern between the two. The confidence levels of the suspected rotational variables were all below 80%, while the average for the misclassified EBs was 92.53%. As many of these were very likely EB stars from visual inspection, it is believed that these misclassifications are a result of low numbers of training examples of both sets. Using a larger training set may have given the network the opportunity to learn more of the nuances between these types of stars. Within the set of 100, there were also eight γ Doradus-δ Scuti hybrid candidates identified, all with the δ Scuti p-mode pulsations with lower amplitudes than the γ Doradus pulsations. Accounting for the findings of the visual search, it is reasonable to assume that this catalogue contains ~6,974 γ Doradus pulsators, of which, 9% are hybrids with δ Scuti.

The greatest limitation in this research was the amount of training data. As most of the catalogues the training data was obtained from were from investigations on data from other missions, the process of obtaining the data included finding TIC identifiers for the stars from *Simbad*, and then searching MAST for data associated with that TIC. For some catalogues, the data needed to be obtained from a cone search as no identifiers were listed in the catalogues. These extra steps included meant that for some catalogues, the number of stars that lightcurve data was able to be obtained for was limited. Adjusting the cone search parameters may have yielded more results, reducing the number of lightcurves which were unable to be obtained, and using this method for all catalogues may have yielded larger training sets. During visual inspection of the periodograms when verifying the results it was noted that the classifications made with high confidences all had high signal-to-noise ratios, meaning that a lot of γ Doradus stars must be missing from the set. This is also obvious when comparing Figs 4.4 and 4.5, in which the distribution of true γ Doradus during validation matches that of the lower confidence predictions obtained from the TESS-SPOC data, which have been excluded from the final set of classifications here.

Further research into this method could include investigating how well the model performs for stars of other variable types. As there is currently only a limited amount of known γ Doradus to train the network on, the ability for detection of other pulsators that are more well documented may be higher.

6 Conclusions

A fast machine learning algorithm was implemented to classify γ Doradus pulsators from TESS data. Lomb-Scargle periodograms constructed from TESS-SPOC lightcurves for 173,398 unique stars with effective temperatures in the range 6500-7500 K and TESS magnitudes between 9.0-12.0 were obtained. These were then run through a classifier which had been trained on five separate classifications of stars: non-variable, γ Doradus, δ Scuti, EB, and RR Lyrae. The classifier was set up to treat this as a binary problem, so all star types other than γ Doradus were put under one class label during training. The distributions of these classifications on a HR diagram were then analysed and the 75th percentile of classification probability was used for the final classifications. From this, 7,749 high quality classifications were obtained. A portion of these were then visually checked to ensure robustness and it was found that of the 100 classifications visually checked, 90% of the set contained likely γ Doradus stars. The majority of misclassifications were from EB stars with variability frequencies in the same range as pulsations of γ Doradus g-modes.

This research has shown that the method of using a neural network classifier on minimally-treated lightcurve TESS-SPOC data is a reasonable approach that can yield reliable results. The short computation times of $\sim$5 ms to pre-process and $\sim$4 ms for classifying each lightcurve make this method of classification well suited for use in bulk classification in pipelines of large-scale sky surveys.

It is the intention that this research makes possible for asteroseismological investigation into a wider range of γ Doradus stars than previously available. This catalogue can also be used for higher quality model training in future research, allowing for much less imbalanced and larger training sets, as the small number of classifications currently available has been a limitation in training networks. Utilisation of networks such as this, which allow bulk classification in large datasets, can greatly aid in the advance of our knowledge of stellar evolution and structure by providing researchers a diverse range of examples of even the most rare types of pulsators.

References

[1] L. Jetsu, S. Porceddu, and J. Lyytinen et al. Did the Ancient Egyptians Record the Period of the Eclipsing Binary Algol—The Raging One? *The Astrophysical Journal*, 773(1):1, 2013.

[2] T. Jayasinghe, K.Z. Stanek, and C.S. Kochanek et al. The ASAS-SN Catalogue of Variable Stars - II. Uniform Classification of 412 000 Known Variables. *Monthly Notices of the Royal Astronomical Society*, 486(2):1907–1943, 2019.

[3] C. Aerts, J. Christensen-Dalsgaard, and D.W. Kurtz. *Asteroseismology.* Springer Science & Business Media, 2010.

[4] P. Kumar, C.O. Ao, and E.J. Quataert. Tidal Excitation of Modes in Binary Systems with Applications to Binary Pulsars. *The Astrophysical Journal*, 449:294–309, 1995.

[5] Y. Li. Bisystem Oscillation Theory of Stars - Part Two - Excitation Mechanisms. *Astronomy and Astrophyics*, 257:145–152, 1992.

[6] J.A. Guzik, A.B Kaye, and P.A Bradley et al. Driving the Gravity-Mode Pulsations in γ Doradus Variables. *The Astrophysical Journal*, 542(1):L57–L60, October 2000.

[7] W.A. Dziembowski. The New Opacities and B-Star Pulsations. *Symposium - International Astronomical Union*, 162:55–66, 1994.

[8] A.C. Phillips. *The Physics of Stars.* Manchester Physics Series. Chichester: Wiley, 1994.

[9] S. Basu and W.J. Chaplin. *Asteroseismic Data Analysis: Foundations and Techniques.* Princeton: Princeton University Press, 2017.

[10] T. Padmanabhan. *Theoretical Astrophysics: Volume 2, Stars and Stellar Systems*, volume 2. Cambridge: Cambridge University Press, 2000.

[11] T.M. Brown and R.L. Gilliland. Asteroseismology. *Annual Review of Astronomy and Astrophysics*, 32(1):37–82, 1994.

[12] M. S. Cunha and T. S. Metcalfe. Asteroseismic Signatures of Small Convective Cores. *The Astrophysical Journal*, 666(1):413, 2007.

[13] H. Saio, P.R. Wood, and M. Takayama et al. Oscillatory Convective Modes in Red Giants: A Possible Explanation of the Long Secondary Periods. *Monthly Notices of the Royal Astronomical Society*, 452(4):3863–3868, 2015.

[14] R.A. García and J. Ballot. Asteroseismology of solar-type stars. *Living Reviews in Solar Physics*, 16(1):4, 2019.

[15] H. A. Hill, R. T. Stebbins, and T. M. Brown. *Recent Solar Oblateness Observations: Data, Interpretation, and Significance for Experimental Relativity*, pages 622–628. Boston, MA: Springer US, 1976.

[16] R. Scuflaire, M. Gabriel, A. Noels, and A. Boury. Oscillatory Periods in the Sun and Theoretical Models With or Without Mixing. *Astronomy and Astrophysics*, 45(1):15–18, 1975.

[17] J. Christensen-Dalsgaard. Helioseismology. *Rev. Mod. Phys.*, 74:1073–1129, 2002.

[18] W.J. Chaplin, Y. Elsworth, and G.R. Davies et al. Super-Nyquist Asteroseismology of Solar-like Oscillators with *Kepler* and K2 – Expanding the Asteroseismic Cohort at the Base of the Red Giant Branch. *Monthly Notices of the Royal Astronomical Society*, 445(1):946–954, 2014.

[19] W.J. Chaplin, H. Kjeldsen, and J. Christensen-Dalsgaard et al. Ensemble Asteroseismology of Solar-Type Stars with the NASA *Kepler* Mission. *Science*, 332(6026):213–216, 2011.

[20] H. Saio and M. Takeuti. The Evolutionary Stage of the RRs star SX Phe. In *Current Problems in Stellar Pulsation Instabilities*, pages 317–327. 1980.

[21] H. Ando. A New Method for Determining the Internal Rotational Angular Velocity of the Stars. *Astrophysics and Space Science*, 73(1):159–174, 1980.

[22] U. Lee. Stability of the Delta Scuti Stars Against Nonradial Oscillations with Low Degrees ℓ. *Publications Astronomical Society of Japan*, 37(2):279–291, 1985.

[23] R.A. García, S. Turck-Chièze, and S.J. Jiménez-Reyes et al. Tracking Solar Gravity Modes: The Dynamics of the Solar Core. *Science*, 316(5831):1591–1593, 2007.

[24] H. Schunker, J. Schou, and P. Gaulme et al. Fragile Detection of Solar g -Modes by Fossat et al. *Solar Physics*, 293(6):95, 2018.

[25] A. Muir and W. Wehlau. The Light Variation of Delta Scuti. *The Astrophysical Journal*, 205:155–161, 1976.

[26] Y.H. Yang, Y.H. Chen, and M.Y. Tang. Asteroseismology of the DAV Star KUV 08368+4026. *Monthly Notices of the Royal Astronomical Society*, 522(4):6094–6101, 2023.

[27] M.A.C. Perryman, L. Lindegren, and J. Kovalevsky et al. The *Hipparcos* Catalogue. *Astronomy and Astrophysics*, 323:L49–L52, 1997.

[28] C. Waelkens, C. Aerts, and E. Kestens et al. Study of an Unbiased Sample of B Stars Observed with *Hipparcos*: The Discovery of a Large Amount of New Slowly Pulsating B Stars. *Astronomy and Astrophysics*, 330:215–221, 1998.

[29] C. Aerts, L. Eyer, and E. Kestens. The Discovery of New Gamma Doradus Stars from the *Hipparcos* Mission. *Astronomy and Astrophysics*, 337:790–796, 1998.

[30] H. Kjeldsen, J. Christensen-Dalsgaard, and R. Handberg et al. The *Kepler* Asteroseismic Investigation: Scientific goals and First Results. *Astronomische Nachrichten*, 331:966, 2010.

[31] D.G. Koch, W.J. Borucki, and G. Basri et al. *Kepler* Mission Design, Realized Photometric Performance, and Early Science. *The Astrophysical Journal*, 713(2):L79, 2010.

[32] M.N. Fanelli, J.M. Jenkins, and S.T. Bryson et al. *Kepler* Data Processing Handbook. *NASA Ames Research Center, Moffett Field*, 2011.

[33] G.R. Ricker, R.K. Vanderspek, and D.W. Latham et al. The Transiting Exoplanet Survey Satellite Mission. *American Astronomical Society Meeting Abstracts*, 224:113–02, 2014.

[34] M.M. Fausnaugh, D.A. Caldwell, and J.M. Jenkins et al. TESS data release notes: Sector 1, DR1. Technical report, 2018.

[35] D. Huber, T.R. White, and T.S Metcalfe et al. A 20 Second Cadence View of Solar-type Stars and Their Planets with TESS: Asteroseismology of Solar Analogs and a Recharacterization of π MEN c. *The Astronomical Journal*, 163(2):79, 2022.

[36] L.A. Balona, J.A. Guzik, and K. Uytterhoeven et al. The *Kepler* View of γ Doradus Stars. *Monthly Notices of the Royal Astronomical Society*, 415(4):3531–3538, 2011.

[37] D.W. Kurtz, H. Saio, and M. Takata et al. Asteroseismic Measurement of Surface-to-Core Rotation in a Main-Sequence A Star, KIC 11145123. *Monthly Notices of the Royal Astronomical Society*, 444:102–116, 2014.

[38] R. Ouazzani, S.J.A.J. Salmon, and V. Antoci et al. A New Asteroseismic Diagnostic for Internal Rotation in γ Doradus Stars. *Monthly Notices of the Royal Astronomical Society*, 465(2):2294–2309, 2016.

[39] A.B. Kaye, G. Handler, and K. Krisciunas et al. γ Doradus Stars: Defining a New Class of Pulsating Variables. *Publications of the Astronomical Society of the Pacific*, 111(761):840–844, 1999.

[40] P.A. Bradley, J.A. Guzik, and L.F. Miles et al. Results of a Search for γ Dor and δ Sct Stars With the *Kepler* Spacecraft. *The Astronomical Journal*, 149(2):68, 2015.

[41] N.H. Barbara, T.R. Bedding, and B.D. Fulcher et al. Classifying *Kepler* Light Curves for 12 000 A and F Stars Using Supervised Feature-based Machine Learning. *Monthly Notices of the Royal Astronomical Society*, 514(2):2793–2804, 2022.

[42] G. Li, T. Van Reeth, and T.R. Bedding et al. Gravity-mode Period Spacings and Near-Core Rotation Rates of 611 γ Doradus Stars with *Kepler*. *Monthly Notices of the Royal Astronomical Society*, 491(3):3586–3605, 2020.

[43] C. Aerts and S. Mathis. Mode Coupling Coefficients Between the Convective Core and Radiative Envelope of γ Doradus and Slowly Pulsating B Stars. *Astronomy and Astrophysics*, 677:A68, 2023.

[44] T. Van Reeth, P. De Cat, and J. Van Beeck et al. The Near-Core Rotation of HD 112429. A γ Doradus Star with TESS Photometry and Legacy Spectroscopy. *Astronomy and Astrophysics*, 662:A58, 2022.

[45] G. Handler. The Domain of Doradus Variables in the Hertzsprung-Russell Diagram. *Monthly Notices of the Royal Astronomical Society*, 309(2):L19–L23, 1999.

[46] M. Skarka, J. Žák, and M. Fedurco et al. Periodic Variable A-F Spectral Type Stars in the Northern TESS Continuous Viewing Zone. I. Identification and Classification. *Astronomy and Astrophysics*, 666:A142, 2022.

[47] F.Z. Zeraatgari, F. Hafezianzadeh, and Y Zhang et al. Machine Learning-based Photometric Classification of Galaxies, Quasars, Emission-line Galaxies, and Stars. *Monthly Notices of the Royal Astronomical Society*, 527(3):4677–4689, 2023.

[48] I.H. Sarker. Machine Learning: Algorithms, Real-world Applications and Research Directions. *SN computer science*, 2(3):160, 2021.

[49] J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc, Sebastopol, California, 2016.

[50] M. Mohssen, K. Muhammad, and B. Eihab. *Machine Learning: Algorithms and Applications.* Boca Raton FL: CRC Press, 2017.

[51] B.C. Love. Comparing Supervised and Unsupervised Category Learning. *Psychonomic Bulletin Review*, 9(4):829–835, 2002.

[52] S. Dreiseitl and L. Ohno-Machado. Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics*, 35(5):352–359, 2002.

[53] D. Maulud and A.M. Abdulazeez. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147, 2020.

[54] S. Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18, 2014.

[55] E. Mayoraz and E. Alpaydin. Support Vector Machines for Multi-class Classification. In José Mira and Juan V. Sánchez-Andrés, editors, *Engineering Applications of Bio-Inspired Artificial Neural Networks*, pages 833–842. Heidelberg: Springer Berlin Heidelberg, 1999.

[56] A. Das. *Encyclopedia of Quality of Life and Well-Being Research, Logistic Regression*, pages 3680–3682. Dordrecht: Springer Netherlands, 2014.

[57] C. Jacobsen, U. Zscherpel, and P. Perner. A Comparison Between Neural Networks and Decision Trees. In Petra Perner and Maria Petrou, editors, *Machine Learning and Data Mining in Pattern Recognition*, pages 144–158. Heidelberg: Springer Berlin Heidelberg, 1999.

[58] S.C. Wang. *Artificial Neural Network*, pages 81–100. Boston, MA: Springer US, 2003.

[59] O.I. Abiodun, A. Jantan, and A.E. Omolara et al. State-of-the-Art in Artificial Neural Network Applications: A Survey. *Heliyon*, 4(11), 2018.

[60] S. Narayan. The Generalized Sigmoid Activation Function: Competitive Supervised Learning. *Information Sciences*, 99(1):69–82, 1997.

[61] S. Marsland. *Machine Learning: An Algorithmic Perspective.* New York: Chapman and Hall/CRC, 2011.

[62] M. Sai Jakka. Assessing Exoplanet Habitability through Data-driven Approaches: A Comprehensive Literature Review. *arXiv e-prints*, 2305.11204, 2023.

[63] J. Debosscher, J. Blomme, and C. Aerts et al. Global Stellar Variability Study in the Field-of-View of the *Kepler* Satellite. *Astronomy and Astrophysics*, 529:A89, 2011.

[64] J. Audenaert, J. S. Kuszlewicz, R. Handberg et al, and The T'DA collaboration. TESS Data for Asteroseismology (T'DA) Stellar Variability Classification Pipeline: Setup and Application to the *Kepler* Q9 Data. *The Astronomical Journal*, 162(5):209, 2021.

[65] K.G. Stassun, R.J. Oelkers, and M. Paegert et al. The Revised TESS Input Catalog and Candidate Target List. *The Astronomical Journal*, 158(4):138, 2019.

[66] J.M. Jenkins, J.D. Twicken, and S. McCauliff et al. The TESS Science Processing Operations Center. In Gianluca Chiozzi and Juan C. Guzman, editors, *Software and Cyberinfrastructure for Astronomy IV*, volume 9913 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 99133E, 2016.

[67] Lightkurve Collaboration, J.V.D.M. Cardoso, and C. Hedges et al. Lightkurve: *Kepler* and TESS Time Series Analysis in Python. Astrophysics Source Code Library, 2018.

[68] S.J. Murphy, D. Hey, and T. Van Reeth et al. Gaia-derived Luminosities of *Kepler* A/F Stars and the Pulsator Fraction Across the δ Scuti Instability Strip. *Monthly Notices of the Royal Astronomical Society*, 485(2):2380–2400, 2019.

[69] A. Prša, N. Batalha, and R.W. Slawson et al. Kepler Eclipsing Binary Stars. I. Catalog and Principal Characterization of 1879 Eclipsing Binaries in the First Data Release. *The Astronomical journal*, 141(3):83, 2011.

[70] R.W. Slawson, A. Prša, and W.F. Welsh et al. *Kepler* Eclipsing Binary Stars. II. 2165 Eclipsing Binaries in the Second Data Release. *The Astronomical Journal*, 142(5):160, 2011.

[71] B. Kirk, K. Conroy, and A. Prša et al. *Kepler* Eclipsing Binary Stars. VII. The Catalog of Eclipsing Binaries Found in the Entire *Kepler* Data Set. *The Astronomical Journal*, 151(3):68, 2016.

[72] M. Wenger, F. Ochsenbein, and D. Egret et al. The SIMBAD Astronomical Database. The CDS Reference Database for Astronomical Objects. *Astronomy and Astrophysics*, 143:9–22, 2000.

[73] V.A. Marsakov, M.L. Gozha, and V.V. Koval'. Masses of RR Lyrae Stars with Different Chemical Abundances in the Galactic Field. *Astronomy Reports*, 63(3):203–211, 2019.

[74] A.J. Drake, M. Catelan, and S.G. Djorgovski et al. Probing the Outer Galactic Halo with RR Lyrae from the Catalina Surveys. *The Astrophysical Journal*, 763(1):32, 2013.

[75] J.M. Nemec, J.G. Cohen, and V. Ripepi et al. Metal Abundances, Radial Velocities, and Other Physical Characteristics for the RR Lyrae Stars in the *Kepler* Field*. *The Astrophysical Journal*, 773(2):181, 2013.

[76] A. Ginsburg, B.M. Sipőcz, and C.E. Brasseur et al. Astroquery: An Astronomical Web-querying Package in Python. *The Astronomical Journal*, 157:98, 2019.

[77] D. Jontof-Hutter, P.A. Dalba, and J.H. Livingston. TESS Observations of *Kepler* Systems with Transit Timing Variations. *The Astronomical Journal*, 164(2):42, 2022.

[78] N.R. Lomb. Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophysics and Space Science*, 39(2):447–462, 1976.

[79] J.D. Scargle. Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data. *The Astrophysical Journal*, 263:835–853, 1982.

[80] A. Paszke, S. Gross, and F. Massa et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[81] K. He, X. Zhang, and S. Ren et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, 2015.

[82] N. Srivastava, G. Hinton, and A. Krizhevsky et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[83] J. Bridle. Training Stochastic Model Recognition Algorithms as Networks can lead to Maximum Mutual Information Estimation of Parameters. In D. Touretzky, editor, *Advances in Neural Information Pocessing Systems*, volume 2. Morgan-Kaufmann, 1989.

[84] T.Y. Lin, P. Goyal, and R. Girshick et al. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[85] Gaia Collaboration. *Gaia* DR3 Part 3. Non-Single Stars, 2022.

[86] A.N. Cox (eds.). *Allen's Astrophysical Quantities.* Los Alamos, NM: Springer, 4 edition, 2002.

[87] P.B. Warner, A.B. Kaye, and J.A. Guzik. A Theoretical γ Doradus Instability Strip. *The Astrophysical Journal*, 593(2):1049–1055, 2003.

[88] B Krawczyk. Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.