White dwarf Random Forest classification through Gaia spectral coefficients

Enrique Miguel García-Zamora¹, Santiago Torres¹,²,⋆⋆, and Alberto Rebassa-Mansergas¹,²

¹ Departament de Física, Universitat Politècnica de Catalunya, c/Estève Terrades 5, 08860 Castelldefels, Spain
² Institute for Space Studies of Catalonia, c/Gran Capità 2–4, Edif. Nexus 104, 08034 Barcelona, Spain

ABSTRACT

Context. The third data release of Gaia has provided approximately 220 million low resolution spectra. Among these, about 100 000 correspond to white dwarfs. The magnitude of this quantity of data precludes the possibility of performing spectral analysis and type determination by human inspection. In order to tackle this issue, we explore the possibility of utilising a machine learning approach, based on a Random Forest algorithm.

Aims. Our goal is to analyse the viability of the Random Forest algorithm for the spectral classification of the white dwarf population within 100 pc from the Sun, based on the Hermite coefficients of Gaia spectra.

Methods. We utilised the assigned spectral type from the Montreal White Dwarf Database for training and testing our Random Forest algorithm. Once validated, our algorithm model was applied to the rest of the unclassified white dwarfs within 100 pc. First, we started by classifying the two major spectral type groups of white dwarfs: hydrogen-rich (DA) and hydrogen-deficient (non-DA). Next, we explored the possibility of classifying the various spectral subtypes, including the secondary spectral types in some cases.

Results. Our Random Forest classification presented a very high recall (>80%) for DA and DB white dwarfs, and a very high precision (>90%) for DB, DQ, and DZ white dwarfs. As a result we have assigned a spectral type to 9 446 previously unclassified white dwarfs: 4 739 DAs, 76 DBs (60 of them DBAs), 4 437 DCs, 132 DZs, and 62 DQs (nine of them DQpec).

Conclusions. Despite the low resolution of Gaia spectra, the Random Forest algorithm applied to the Gaia spectral coefficients proves to be a highly valuable tool for spectral classification.

Key words. (Stars:) white dwarfs, atmospheres, catalogs

1. Introduction

White dwarfs are the remnants of stars with initial masses ≤8–10 M☉ (e.g. Althaus et al. 2010). They are basically composed of a degenerate core of typically half a solar mass that is surrounded by a thin partially degenerate atmospheric layer. Since nuclear reactions have practically ceased, the energy source in the deep interior of white dwarfs is primarily derived from gravothermal energy released by the ions and eventually provided by core crystallisation, phase separation, and other processes such as sedimentation of minor species (see Isern et al. 2022 for a recent review). The heat generated in the core of the white dwarf is radiated through the atmospheric envelope. Thus, this thin layer plays a capital role in the cooling of the white dwarf. In the canonical model, the outermost layer of a white dwarf is primarily composed of helium with a mass around 10⁻³ M☉, accounting for less than 2% of the total white dwarf mass. However, in the majority of cases (approximately 80%), there is an additional thinner layer of hydrogen with a mass between 10⁻¹¹ to 10⁻¹³ M☉, which overlays the helium layer.

From an observational point of view, spectroscopic analysis of white dwarf atmospheres enables the identification of atomic and molecular lines and bands. This fact has allowed a spectral classification of white dwarfs according to the presence of certain lines (Sion et al. 1983). Basically, white dwarfs are divided into those that present Balmer lines (i.e. hydrogen-rich white dwarfs, or DAs), and those that do not (generically called non-DAs). Among this last group, we may also find white dwarf spectra that exhibit absorption helium lines, He i or He ii, called DB and DQ, respectively; carbon features, either atomic or molecular, named DQ; metallic lines such as Ca ii or Fe ii, named DZ; or very weak lines or no features at all, thus showing a continuous spectrum, named DC. This general spectral classification relates to what is referred to as the primary spectral type (see Table 2 from Sion et al. 1983). However, it is common to identify lines from different elements in white dwarf spectra. For instance, we may find a DA with weaker helium lines or metallic lines additionally present, in which case these objects would be labelled as DAB or DAZ, respectively. The presence of a magnetic field or variability in the white dwarf spectrum, respectively, would add a secondary H or V to the primary spectral class.

Spectral classification of white dwarfs is of paramount importance for the determination of their stellar parameters such as temperature, surface gravity, mass, or luminosity. Moreover, our understanding of the physical evolution of the white dwarf population depends on the proper identification of their atmospheres. For instance, processes such as convective mixing or convective dilution in spectral evolution (e.g. Blouin et al. 2019 Cunningham et al. 2020), the presence of carbon in hydrogen-deficient atmospheres as a possible explanation of the Gaia colour-magnitude bifurcation (Camisassa et al. 2023 Blouin et al. 2023), the high ratio of DQ white dwarfs in the called Q branch (Tremblay et al. 2019 Cheng et al. 2019), or the origin of accreted material in white dwarfs (e.g. Zuckerman et al. 2007).
Farihi et al. (2010) are a few examples where a detailed identification of white dwarf spectra is required for a proper understanding of these issues. However, spectroscopic follow-up of white dwarfs is a time-consuming task. A volume complete spectroscopic sample is achieved up to 40 pc from the Sun (Tremblay et al. 2020; McCleery et al. 2020; O’Brien et al. 2023), where observations mostly performed from the William Herschel Telescope and the Gran Telescopio Canarias in the Northern Hemisphere and the Very Large Telescope and Southern Astrophysical Research Telescope in the Southern Hemisphere provide resolving powers \( R \approx 2000 \sim 3900 \). However, this is not the case up to 100 pc, where the percentage of spectrally labelled white dwarfs is roughly 20\% (e.g. Kilkic et al. 2020).

Nevertheless, the third Gaia mission Data Release (Gaia Collaboration et al. 2023) has provided astrometric data for nearly two billion objects and mean low resolution (30 \( \lesssim R \lesssim 100 \); Carrasco et al. 2021) BP and RP spectra of approximately 220 million sources (De Angeli et al. 2023). Of these, almost 100 000 correspond to candidates for white dwarf objects (Genetile Fusillo et al. 2021).

This enormous quantity of data prevents spectral classification by human inspection. With the recent increasing growth of large astronomical databases, other approaches based on machine learning artificial intelligence algorithms are absolutely necessary. These techniques are widely used nowadays in astrophysics, and particularly in the field of white dwarfs. Since the pioneering work of Torres et al. (1998) on the use of self-organising maps for the identification of halo stars, up to the most recent ones using the Random Forest algorithm in Galactic component identification (Torres et al. 2019), or their spectral identification (Echeverry et al. 2022; Montegriffo et al. 2023), or through deep learning techniques (Kong et al. 2018; Vincent et al. 2023), all of these approaches have been proven to be reliable methods in the automatic analysis of large white dwarf databases. Additional statistical classification methods have been performed, in particular in the spectral classification of white dwarfs. For instance, in Jiménez-Esteban et al. (2023) and Torres et al. (2023), the Virtual Observatory Spectral energy distribution Analyser tool (Bayo et al. 2008) was used to conduct an automated spectral energy distribution (SED) fitting of the 100 pc and 500 pc Gaia white dwarf samples, respectively, to different atmospheric models. These works allowed the authors to classify the samples into DA and non-DA white dwarfs with an accuracy of over 90 per cent.

In this work we apply a Random Forest algorithm specifically developed to classify, for the first time, the whole 100-pc white dwarf Gaia sample into their spectral types. Focusing the analysis to objects identified within this distance limit is capital, since it represents a nearly complete volume-limited sample, which potentially allows to derive accurate percentages of the different spectral type classes among white dwarfs. This approach based on artificial intelligence techniques represents a clear advantage respect other approaches we performed in Jiménez-Esteban et al. (2023) and Torres et al. (2023), since it does not require the use of theoretical atmospheric models. The models are subject to substantial uncertainties for temperatures below 5 500 K, which implies an unreliable classification for such cool white dwarfs. Instead, the Random Forest algorithm presented here relies on previously spectral type labelled white dwarfs covering all possible values of effective temperatures. Thus, we aim to obtain as much spectroscopic information as possible from the Gaia spectral coefficients. This includes not only a classification of the white dwarfs into their primary spectral types, but also attempting to classify them into different subcategories.

In Section 2 we explain the methodology applied. In Section 3 the validation tests performed on a subset of white dwarfs with spectral data assigned and their results are detailed. In Section 4 we apply the algorithm to the classification of white dwarfs in a 100-pc radius around the Sun. In Section 5 we identify the most relevant spectral coefficients used in the classification process. The performance of our Random Forest algorithm is compared in Section 6 to other recent classification methods. Finally, in Section 7 we present our conclusions.

2. The method: Random Forest classification of Gaia spectral coefficients

The Random Forest (Breiman 2001) is a widely used machine learning algorithm. From a set of labelled data, which is used to train the algorithm, an ensemble of decision trees, which is called a Random Forest, is created. Once this ensemble has been obtained, it can be used to classify new data in the given categories.

This algorithm has been widely used for the classification of stellar objects (see, for instance, Li et al. 2019; Plewa 2018 or Dubath et al. 2011) and, in particular, to the study of the white dwarf population: some examples already show the feasibility of using Random Forest for the identification of different Galactic white dwarf populations using Gaia data as input parameters (Torres et al. 2019), or distinguishing between spectra of isolated white dwarfs, main sequence objects and white dwarf-main sequence binaries (Echeverry et al. 2022). Moreover, a Random Forest algorithm has also been used for the selection of white dwarfs in the Gaia sample (Gaia Collaboration et al. 2021). Besides, a first attempt to classify white dwarfs into DAs and non-DAs using the spectral coefficients was performed by MonTEGRIFfo et al. (2023).

Here, following the line of the previous works, we aimed to apply the Random Forest algorithm for the spectral classification of the Gaia white dwarf population in a 100-pc radius around the Sun. In particular, our effort is employed in classifying the different sub-populations of the non-DA sample identified by Jiménez-Esteban et al. (2023) as well as to extend the classification to cool white dwarfs (\( \lesssim 5 \ 500 \) K) that the previous work did not consider due to the lack of accurate atmospheric models.

The Gaia spectra have low resolution (\( \Delta \lambda / \Delta \lambda \approx 100 \) and cover the 3 300-10 500 Å wavelength range (3 300-6 800 Å by the Blue Photometer (BP) and 6 400-10 500 Å by the Red Photometer (RP); Carrasco et al. 2021). One particularity of the Gaia spectra is that they are not provided as a typical series of flux values for certain wavelengths but rather as a set of 55 coefficients for each of the BP and RP spectrographs (i.e. 110 coefficients in total). These coefficients refer to the Hermite functions that act as the basis for the spectral representation (Carrasco et al. 2021). The spectra are internally calibrated in a pseudo-pixel scale, and they can also be transformed to an external calibration (i.e. flux versus wavelength representation) by using the specifically designed Python package GaiaXP\footnote{https://gaia-dpci.github.io/GaiaXPy-website/}.

As input data for the Random Forest algorithm, we use the 110 Hermite coefficients. It was demonstrated in Montegriffo et al. (2023) that the use of the coefficients provides better performance of the classification algorithm than when other input passbands were used. Moreover, the coefficient procedure can
be considered totally appropriate, as the different white dwarf spectral types are defined by their specific spectral features, and all this information is contained in the coefficients (see, for instance, Weiler et al. 2023 for a mathematical description applied to hydrogen lines). As a consequence, no external calibration was applied, since this process may introduce what is known as ‘wiggles’, or oscillatory behaviour. In our data, this effect would be more prominent at both ends of the externally calibrated spectra. These ‘wiggles’ are produced by the mathematical process used to obtain the spectra (De Angeli et al. 2023).

3. Training and validating the algorithm: The Montreal White Dwarf Database

The Montreal White Dwarf Database is a virtual database containing astrometric, photometric, and spectroscopic data including a spectral type classification for tens of thousands of white dwarfs (Dufour et al. 2017). At the beginning of 2023, a total of 41,570 white dwarfs are classified into their different spectral types, 2,905 of them within 100 pc from the Sun. In this work, the MWDD spectroscopic white dwarf classification within 100 pc is used as the input labelled sample for the training and validation tests of our Random Forest algorithm.

For the cross-validation, we adopt the stratified k-fold method. It consists in dividing the whole set into k folds, where k is a variable number (in this work we chose k = 10). Each fold has approximately the same number of objects, and the category ratio (the proportion of objects assigned to different spectral types) is kept as close as possible to the original category set ratio. For each fold, a Random Forest is trained with all nine remainder folds, and tested on it. The advantages over the random training-test split consists in avoiding the randomness of the subset divisions, the constancy of the subset proportions and the fact that the whole set is used for both training and testing.

As it is well established, the Random Forest performance tends to be optimal for balanced data sets (e.g. Breiman 2001). Consequently, the validation strategy followed consisted of keeping the classification samples as close as possible to a balanced sample. Thus, our first validation test consisted in classifying white dwarfs into DA and non-DA spectral types. The second one, focused on those labelled as non-DA, and classified them into DB, DC, DQ, and DZ types. Finally, the third one consisted of classifying the white dwarfs of a specific type into its different subtypes. For instance, DA white dwarfs are divided into DA, DAB, DAH, and DAZ, and similarly for the other spectral types.

In order to create the random forests and obtain the confusion matrices and classification metrics, the Python package scikit-learn (Pedregosa et al. 2011) was used.

3.1. First validation test: classifying the Gaia population with a MWDD type into DA and non-DA types

In the first place, we classify the whole sample of white dwarfs with labels in the MWDD within 100 pc into DA and non-DA types (2,905 objects; 1993 as DAs and 912 as non-DAs). Although the ratio of DA and non-DA, 68.6% and 31.4%, respectively, is not strictly balanced, the proportion of the two groups is large enough to ensure an optimal performance of the Random Forest algorithm and to avoid extreme imbalance effects.

The resulting confusion matrix of our Random Forest model is shown in the top panel of Figure 1. True labels (rows) correspond to the MWDD classification, while the predicted label (columns) is the one resulting from our Random Forest algorithm.

Fig. 1: Confusion matrices for our validation tests: DA vs non-DA (top panel), non-DA types (middle panel), and DB subtypes (bottom panel). As true label (rows) we adopted the MWDD classification, while the predicted label (columns) is the one resulting from our Random Forest algorithm.
Table 1: Hyperparameters and optimal values adopted in the first two validation tests.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Optimal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of estimators</td>
<td>10000, 10000</td>
</tr>
<tr>
<td>Maximum depth</td>
<td>50, 70</td>
</tr>
<tr>
<td>Minimum samples split</td>
<td>20, 2</td>
</tr>
<tr>
<td>Minimum samples leaf</td>
<td>1, 1</td>
</tr>
<tr>
<td>Maximum features</td>
<td>sqrt, sqrt</td>
</tr>
</tbody>
</table>

3.2. Second validation test: Classifying the Gaia non DA white dwarf population with a MWDD type into their subtypes

In our second validation test, we classify the non-DA white dwarfs (912 objects in total) into their spectral types (DB, DC, DQ, and DZ). The resulting confusion matrix is shown in the middle panel of Figure 1. The hyperparameters used and the metrics obtained can be found in Tables 1 and 2, respectively. The results obtained reveal a very good performance with an accuracy of 0.81. In particular, the algorithm presents a very good recall for DB white dwarfs (82%), an excellent recall for DCs (98%), and low recalls for both DQs and DZs (<50%). Two main reasons can be identified to account for these facts. First, the low recall may be caused by the low resolution spectra inherent to Gaia. That is, not very prominent spectral lines might be unnoticed in the Gaia low resolution spectra, which would result in the algorithm treating them as featureless, continuous spectra characteristic of DC white dwarfs. Second, DQ and DZ classes represent 12.8% and 13.7%, respectively, of the non-DA population used for training. Thus, imbalanced effects, worsening the performance of the algorithm, are likely to start to manifest.

However, it must be noted that, in spite of the low recall for DQ and DZ white dwarfs, their precision (as well as for DB white dwarfs) is excellent, that is 92% for the three types. False positives are almost absent. This implies that, while the algorithm does not find all DZ and DQ white dwarfs, the probability of a white dwarf belonging to the type it has been classified into is very high. This makes our algorithm highly useful for efficiently identifying white dwarfs of these spectral types within an unclassified population.

Finally, as a verification exercise, we have attempted to classify the entire sample into their primary subtypes; DA, DB, DC, DQ, and DZ. An excellent recall is achieved for DAs (97%), DBs show a very good recall (80%); and DQ and DZ recalls are certainly improvable (25 and 34%, respectively). Once more, DBs, DQs, and DZs show an excellent precision (89% for DBs and 91% for DQs and DZs). However, the scoring values are lower than the values obtained in the first two validation tests. The conclusion we extract from this result is that better results are obtained when the workflow includes a first DA/non-DA classification and a second, specific non-DA classification.

3.3. Third validation test: classifying the Gaia white dwarf population with a MWDD type into their secondary types

So far, we have demonstrated that the Random Forest algorithm based on the coefficients of Gaia spectra is a feasible tool for classifying white dwarfs into their primary spectral types. Now, we explore the possibility to classify the secondary types. Several factors prevent us from being optimistic about such performance. First, Gaia low resolution spectra appear to have limited capability to discern detailed features, such as distorted Balmer lines in DAZ white dwarfs or weak lines in other atmospheres. Second, subtype classes represent in many cases clearly imbalanced samples with respect to the predominant subtype, thus worsening the performance of the algorithm.

Nevertheless, even if we expect a very low recall in the classification of the majority of spectral subtypes, if a high precision is achieved, it would imply the identification of valuable objects. Consequently, the four considered spectral types (DA, DB, DQ, and DQpec) were divided into different subtypes and analysed separately. In the following sections we provide the technical details and a summary is given in Section 3.3.5.

3.3.1. DA subtype classification

The DA type was divided into pure DA, DAB, DAH, and DAZ. The classification test reveals the disappointing, but not unexpected, result that only one DAH white dwarf is correctly classified as such (1% recall). This shows that, in almost all cases, the fine magnetic splitting of the spectral lines produced by the intense magnetic field is too small to be discerned in the low resolution Gaia spectra. Additionally, one DA is misclassified as a DAH, resulting in a precision of 50% for DAHs. The rest of the subtypes are not recognised in the classification and are all misclassified as DAs. Subtle lines and the spectral resolution, as well as their low number compared to the initial sample, would explain these results.

3.3.2. DB subtype classification

In this test, the DB type was divided into pure DB, DAB, DAH, and DAZ. The classification test reveals the disappointing, but not unexpected, result that only one DAH white dwarf is correctly classified as such (1% recall). This shows that, in almost all cases, the fine magnetic splitting of the spectral lines produced by the intense magnetic field is too small to be discerned in the low resolution Gaia spectra. Additionally, one DA is misclassified as a DAH, resulting in a precision of 50% for DAHs.

The rest of the subtypes are not recognised in the classification and are all misclassified as DAs. Subtle lines and the spectral resolution, as well as their low number compared to the initial sample, would explain these results.

3.3.3. DQ subtype classification

The results of the Random Forest applied to the DQ subtype reveals that the algorithm is only able to distinguish DQpec white dwarfs, and even then only two out of seventeen (11% recall).
Table 2: Classification metrics for the first validation test in which we classify white dwarfs of the MWDD into DA and non-DA classes, and the second validation test in which non-DAs are classified into DB, DC, DQ, and DZ.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Validation test</th>
<th>DA</th>
<th>non-DA</th>
<th>DB</th>
<th>DC</th>
<th>DQ</th>
<th>DZ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test 1</td>
<td>0.95</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Recall</td>
<td>Test 2</td>
<td>-</td>
<td>-</td>
<td>0.82</td>
<td>0.98</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Test 1</td>
<td>0.91</td>
<td>0.87</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Precision</td>
<td>Test 2</td>
<td>-</td>
<td>-</td>
<td>0.92</td>
<td>0.78</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>F1 score</td>
<td>Test 1</td>
<td>0.93</td>
<td>0.83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Test 2</td>
<td>-</td>
<td>-</td>
<td>0.87</td>
<td>0.87</td>
<td>0.46</td>
<td>0.61</td>
</tr>
</tbody>
</table>

The other subtypes (DQA, DQZ, and DQZA), which comprise at most two objects each, are indistinguishable from DQs.

It is also worth noting, however, that the precision is also perfect for the DQpec subtype (100%), with no false positives from other subtypes. This implies that the few DQpec stars the algorithm may find have a very high probability of belonging to this subtype.

3.3.4. DZ subtype classification

Regarding DZs, the Random Forest algorithm is not able to distinguish DZH subtype from DZ white dwarfs. On the other hand, a single DZA (8% recall) is properly classified. However, three DZs are mislabelled as DZA, negatively impacting the precision for this subtype (25%).

3.3.5. Spectral subtype classification summary

From the analysis of the results of the Random Forest algorithm applied to the different spectral subtypes, we can conclude that the algorithm is mostly unable to classify secondary spectral subtypes, whether due to the imbalance of the dataset or the inherent low resolution of the Gaia spectra that prevents their spectral lines from being recognised by the algorithm.

A possible exception is the subtype DQpec, which, although shows a low recall, also presents a perfect precision. Its situation among DQ subtypes is similar to the situation of DQs among non-DA white dwarfs: low recall, but very high precision that might allow us to find candidates with a very high probability of actually belonging to the group is has been classified into.

Furthermore, although the DB/DBA classification may seem possible due to the good recovery of DBA white dwarfs, we must take this result cautiously. While the recall is reasonably good for DBAs (74%), it is just 50% for DBs. Additionally, the precision is only slightly superior to 50% for both subtypes. Consequently, we cannot assume that a white dwarf identified as a DB or DBA has a high probability of really being one.

4. Classifying the Gaia 100-pc white dwarf population

Once our Random Forest algorithm has been tested and validated, it can be applied to the unclassified Gaia 100-pc white dwarf population. A subgroup of these white dwarfs, namely those with $\text{BR} > \text{RP} < 0.86$ (equivalent to white dwarfs hotter than $\gtrsim 5500 \text{K}$), has already been classified into DAs and non-DAs by Jiménez-Esteban et al. (2023). To that end, synthetic photometry of all white dwarfs was generated using their spectra and the J-PAS (Benítez et al. 2014) filter system (Marín-Franch et al. 2012). These spectra were fitted using a collection of DA and DB atmospheric models, and a probability for each them belonging to the DA type was computed from the $\chi^2$ arising from the best fits. In this exercise, Jiménez-Esteban et al. (2023) adopted two approaches: model fits using all Gaia spectral coefficients and model fits using the truncated coefficients. The former case, defined as the VOSA-GJP estimator, provided better results, with an overall accuracy of 91%. This value is slightly higher than the accuracy we have obtained here using our Random Forest algorithm (90%). Although both classification performances can be considered practically equivalent, we hereafter adopt the DA and non-DA VOSA-GJP classification of Jiménez-Esteban et al. (2023) for all white dwarfs with $\text{BR} > \text{RP} < 0.86$. Thus, in this section we first apply our Random Forest model to those white dwarfs classified into DA and non-DA by Jiménez-Esteban et al. (2023) with the aim to obtain their spectral subtypes. Then, we expand the classification to those unclassified objects with colour $\text{BR} > \text{RP} > 0.86$, i.e. the cooler white dwarfs that we failed to identify in Jiménez-Esteban et al. (2023) due to the lack of accurate atmospheric models.

4.1. White dwarfs identified by VOSA-GJP

In this section, we analyse the objects classified in Jiménez-Esteban et al. (2023) with the aim of obtaining their spectral subtypes. In that classification, as mentioned before, white dwarves were assigned a probability, $P_{\text{DA}}$, of being DAs. Those with $P_{\text{DA}} > 0.5$ were classified as DAs, while those with $P_{\text{DA}} < 0.5$ were classified as non-DAs. A total of 5 823 white dwarfs with $\text{BR} > \text{RP} > 0.86$ are considered in this section; 4 157 of them classified as DAs and 1 666 classified as non-DAs.

4.1.1. DA white dwarfs identified by VOSA-GJP

Despite the poor performance of the algorithm in classifying secondary spectral types found in subsection 3.3.1, we attempted to find possible DAH or DAZ candidates among the group of 4 157 white dwarfs classified as DA in Jiménez-Esteban et al. (2023).

The result reveals that only two DAH candidates are found. Nonetheless, despite the very low number of DAHs, we consider it as a success for our algorithm, specially since the effect of magnetic fields in spectral lines is fine magnetic splitting, which is not easily noticeable in low resolution spectra. In Figure 2 we show the location of these two DAH candidates in the Gaia Hertzsprung-Russell (HR) diagram.
4.1.2. Non-DA white dwarfs identified by VOSA-GJP

We analyse now the white dwarfs that have been classified as non-DAs in Jiménez-Esteban et al. (2023) via adopting the VOSA-GJP estimator. In order to train the set, we once more resort to the MWDD. From the whole set, we derive a subset that mimics the conditions of the objects that will be classified (i.e. non-DA white dwarfs with $BR - RP < 0.86$). This left us with only 509 objects in the training set for the Random Forest algorithm, contrasting with 912 non-DA white dwarfs classified in the whole training set.

The classifying algorithm is then applied to the remaining 1666 objects in the test subset. The classification yields the following results: 76 objects are identified as DBs, 1429 as DCs, 40 as DQs, and 121 as DZs. The corresponding HR diagram of these classified objects is shown in Figure 3.

The HR diagram not only serves to illustrate the composition of the classified population, but it also allows us to check for consistency with expected white dwarf characteristics. For instance, no DB white dwarfs should be found below a certain temperature ($\approx 10000$ K). In Figure 3 DBs appear restricted to the top left, hotter region of the white dwarf sequence, while some DQ white dwarfs appear in the Q branch. All these factors reinforce the idea that our classification is essentially correct, and no spectral types appear outside of their expected locations.

Furthermore, as seen in our third validation test (see Section 3.3), the Random Forest algorithm is able to identify (although with low recall but with high precision) secondary subtypes of DBs and DQs. As we do not expect to find more DBs in the cooler region that remains to be analysed, we apply our classification algorithm to the set of 76 DB white dwarfs identified so far. The Random Forest identified 16 pure DB and 60 DBA objects. No other secondary subtypes (DBAH, DBAZ, DBQA, DBZ, DBZA, and DBE) were identified. In Figure 4 we depict the HR diagram location for the identified pure DBs and DBAs. We can check that no pure DBs are found with colours redder than $BP - RP \geq -0.1$ (i.e. effective temperatures cooler than $\approx 12000$ K).

The sample of identified DQs is analysed into its secondary types in Section 4.3.
method, as there are enough labelled objects to train the algorithm in the HR region of interest.

The results of applying our Random Forest to the cold sample of Jiménez-Esteban et al. (2023) can be seen in the HR diagram presented in Figure 5. The vast majority of white dwarfs (3,041; 84%) are classified as non-DAs (blue dots), while DAs (magenta dots) comprise only a small fraction (582; 16%) and none of them has BP − RP > 1.25.

This classification is consistent with the expected behaviour at temperatures lower than \( \approx 5,000 \, \text{K} \), since at this range the hydrogen in the white dwarf atmosphere remains mostly in the ground state. Thus, Balmer spectral lines would become too weak (or they simply disappear) to be detected in Gaia low resolution spectra and, consequently, the object would be classified as a featureless DC.

4.2.2. DA secondary type classification

As we have seen in Section 4.1.1, our Random Forest algorithm was able to find two DAHs. Thus, we applied the algorithm to the classified cold DA white dwarfs. From the 582 objects, none was classified as a DAH or DAZ; all of them were classified as DAs. This result is not entirely unexpected as it was already known that the Gaia resolution was, in almost all cases, insufficient for this purpose.

4.2.3. Non-DAs subtype classification.

Once the identification between DAs and non-DAs has been completed, the 3,041 found non-DAs were classified into DC, DQ, and DZ categories. DBs were discarded, as none can be found at these low temperatures. As in subsection 4.2.1, the whole MWDD classified set was used as the training data. In particular, we used 248 DCs, 19 DQ, and 26 DZ with colours BR − RP > 0.86.

The results shown in Figure 6 reveal that, as expected, the most prominent group are DCs: 3,008 objects representing 98.9% of the sample. Despite the low Gaia resolution, and possibly low signal-to-noise ratio, which impair the algorithm’s ability to correctly identify any spectral feature at this low temperature regime, the Random Forest algorithm was able to identify 22 DQs and 11 DZs. Taking into account that only 19 DQs and 26 DZs white dwarfs with colours BR − RP > 0.86 form the training sample, these newly found objects represent a 115.6% and 42.3% increment, respectively.

4.3. DQ secondary type classification

In our analysis, we have found 62 DQ so far. Of them, 40 with colours bluer than BP − RP = 0.86 (see Section 4.1.2) and 22 redder than that value (previous section). As demonstrated in our third validation test (see Section 3.3), the Random Forest is capable of identifying certain secondary DQ spectral types, although at the expense of low recall.

We apply our Random Forest algorithm to the 62 DQ-identified white dwarfs, with the aim of classifying them into the secondary spectral types (i.e. DQ, DQA, DQZ, DQZA, and DQpec). The result reveals that objects are thus classified only into two groups: 53 DQ and 9 DQpec white dwarfs. Figure 7 shows the corresponding HR diagram. Those objects classified as DQpec are typically cold, with BP − RP > 0.6, indicative that Swan bands are more easily distorted at low temperatures (e.g. Blouin & Dufour 2019).

5. Feature importance

As an ensemble learning method, the Random Forest algorithm constructs multiple decision trees combining their predictions to achieve the more accurate and stable result. In this construction, some features (variables or parameters of the sample) play a
more remarkable role than others. Even more, one can remove some features without significantly altering the result. In our case, the features are the 110 Gaia spectral coefficients. We aim to analyse which of them have the highest importance for each classification. The method used to compute the feature importance was the mean decrease in impurity (MDI), which is based on the decrease of node impurity averaged over the whole Random Forest. This can be understood as follows. When a decision tree is generated, decision nodes are created. Node impurity is a measurement of the amount of classes in a certain decision node. They are said to be pure if they only comprise one class. Therefore, the most important features in our analysis are the ones that reduce the node impurity the most across the forest. These will, of course, be dependent on the set that is being classified. For instance, the coefficients that rule the Balmer lines are capital in a DA non-DA classification, but of no importance in a non-DA classification.

In Figure 8 we show the feature importance obtained by the MDI method as a function of the Gaia spectral coefficients for the DA versus non-DA (top panel) and the non-DA subtype classification (bottom panel). Regarding the DA versus non-DA classification, the most important coefficients are approximately the 15 first red coefficients and the 20 first blue coefficients. Moreover, if we consider a 0.8% threshold (marked as black line in Fig. 8), we can eliminate most of the low-significant spectral coefficients, representing the remaining 73.6% of the information.

This result implies that a greater importance is placed in the BP information; indeed, all Balmer lines except Hα, He i lines, Swan bands and most metallic lines fall in the BP wavelength range. The most important feature, however, corresponds to the RP. We identify it with the Balmer Hα line. Since DAs show H features, it is predictable that the algorithm considers this spectral line as the most important to distinguish between DAs and non-DAs.

With respect to the non-DA classification into its spectral subtypes, the feature importance distribution (bottom panel) reveals that blue coefficients are the most relevant. Applying the same 0.8% threshold that in the previous case, approximately the first 30 coefficients contain 52% of the information. As most of the type-characteristic spectral lines (for instance, most He i lines, the Swan bands, or Ca ii lines) appear in the wavelength range covered by the BP, rather than in the RP range, this result is both expected and consistent with our previous knowledge.

6. The Gaia 100-pc sample classification summary

In this section, we present a summary of our white dwarf spectral classification. From the 9 446 classified white dwarfs, 4 737 have been classified as DA, 2 as DAH, 76 as DB, 4 437 as DC, 62 as DQ, and 132 as DZ. The original, labelled MWDD sample used as training comprises 2 905 objects; 1 845 DA, 90 DAH, 97 DB, 573 DC, 117 DQ, and 125 DZ.

Consequently, the number of classified objects within 100 pc from the Sun has increased by 257% for DAs, 2.2% for DAHs, 78.4% for DBs, 774% for DCs, 53% for DQs, and 105.6% for DZs.

Figure 9 shows the Gaia HR diagrams for the white dwarfs classified in this work as DA and its secondary types (left panels), while in Figure 10 are represented the corresponding subtypes for objects classified as DB, DC, DQ, and DZ by our algorithm (left panels). For completeness, we also show the HR diagrams including those white dwarfs previously classified in the MWDD (right panels).

Additionally, all the objects studied here are collected in a list, where we provide their corresponding spectral classification among other Gaia parameters. A representative excerpt of this catalogue is presented in Table 3. The whole catalogue can be found in the electronic version of the paper. Moreover, for illustrative purposes, in Annex A we show some examples of Gaia spectra corresponding to white dwarfs of different spectral types classified by our algorithm. These spectra are compared to the Gaia spectra of white dwarfs labelled in the MWDD.

7. Comparison to other automatic classification methods.

To assess the quality of the performance of our Random Forest algorithm, in this section, we compare it with other similar automated classification methods described in the literature. In particular, we analyse the results obtained by Vincent et al. (2023) in their white dwarf spectral classification using neural networks. Although the methodology is not the same, and neither the classification sample, the training sample, nor the input data are identical, we can establish a certain comparative analysis of the results. For instance, our work is focused on the Gaia 100 pc white dwarf sample for mainly primary spectral types (DA, DB, DC, DQ, and DZ) classification, while the work by Vincent et al. (2023) consists in a more general approach for white dwarf candidate selection and spectroscopic classification. This includes...
### Table 3: *Gaia* 100 pc white dwarf sample catalogue classified by our Random Forest algorithm into spectral types.

<table>
<thead>
<tr>
<th>Gaia Source ID</th>
<th>RA (deg)</th>
<th>Dec (deg)</th>
<th>$M_G$ (mag)</th>
<th>BP − RP (mag)</th>
<th>$d$ (pc)</th>
<th>S$_{type}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6617996741403360128</td>
<td>328.75</td>
<td>-27.84</td>
<td>14.42</td>
<td>0.91</td>
<td>23.5</td>
<td>DA</td>
</tr>
<tr>
<td>3022956969731332096</td>
<td>88.79</td>
<td>-4.18</td>
<td>15.19</td>
<td>1.13</td>
<td>6.44</td>
<td>DA</td>
</tr>
<tr>
<td>5314177402013456256</td>
<td>141.20</td>
<td>-69.26</td>
<td>14.59</td>
<td>0.89</td>
<td>22.57</td>
<td>DA</td>
</tr>
<tr>
<td>86313137249758912</td>
<td>176.43</td>
<td>63.10</td>
<td>14.29</td>
<td>0.92</td>
<td>24.01</td>
<td>DA</td>
</tr>
<tr>
<td>4789516515431781156</td>
<td>65.99</td>
<td>-45.85</td>
<td>14.29</td>
<td>0.89</td>
<td>29.94</td>
<td>DA</td>
</tr>
<tr>
<td>3501922067493606272</td>
<td>109.10</td>
<td>-23.30</td>
<td>13.24</td>
<td>0.88</td>
<td>41.89</td>
<td>DA</td>
</tr>
<tr>
<td>514233682564176256</td>
<td>25.67</td>
<td>-17.24</td>
<td>14.36</td>
<td>0.90</td>
<td>40.05</td>
<td>DA</td>
</tr>
<tr>
<td>690619780698867712</td>
<td>306.88</td>
<td>-7.60</td>
<td>14.63</td>
<td>0.91</td>
<td>51.38</td>
<td>DA</td>
</tr>
<tr>
<td>508635403471767680</td>
<td>49.30</td>
<td>-29.19</td>
<td>14.74</td>
<td>0.96</td>
<td>31.50</td>
<td>DA</td>
</tr>
<tr>
<td>543278933518999168</td>
<td>144.40</td>
<td>-38.87</td>
<td>14.27</td>
<td>0.86</td>
<td>34.50</td>
<td>DA</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>594525220254643432</td>
<td>266.95</td>
<td>-52.12</td>
<td>12.72</td>
<td>-0.11</td>
<td>38.92</td>
<td>DAH</td>
</tr>
<tr>
<td>360648387747801088</td>
<td>204.92</td>
<td>-7.22</td>
<td>12.57</td>
<td>-0.01</td>
<td>56.33</td>
<td>DAH</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>62894373222829440</td>
<td>339.03</td>
<td>-1.68</td>
<td>12.74</td>
<td>0.15</td>
<td>39.02</td>
<td>DBA</td>
</tr>
<tr>
<td>337513569070213632</td>
<td>93.44</td>
<td>20.84</td>
<td>11.98</td>
<td>0.26</td>
<td>57.98</td>
<td>DBA</td>
</tr>
<tr>
<td>56537140992422782</td>
<td>187.25</td>
<td>-26.96</td>
<td>10.97</td>
<td>-0.19</td>
<td>83.64</td>
<td>DBA</td>
</tr>
<tr>
<td>5791721605860588</td>
<td>124.65</td>
<td>-18.53</td>
<td>11.53</td>
<td>-0.17</td>
<td>87.66</td>
<td>DBA</td>
</tr>
<tr>
<td>160849786404134016</td>
<td>216.90</td>
<td>53.81</td>
<td>11.41</td>
<td>0.91</td>
<td>74.98</td>
<td>DBA</td>
</tr>
<tr>
<td>59918262256640320</td>
<td>52.15</td>
<td>-8.26</td>
<td>11.49</td>
<td>-0.21</td>
<td>87.54</td>
<td>DBA</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>464791423963623616</td>
<td>87.62</td>
<td>-76.81</td>
<td>14.63</td>
<td>1.09</td>
<td>85.91</td>
<td>DC</td>
</tr>
<tr>
<td>498383964752291504</td>
<td>21.02</td>
<td>-42.68</td>
<td>13.36</td>
<td>0.72</td>
<td>18.45</td>
<td>DC</td>
</tr>
<tr>
<td>4108828945319007744</td>
<td>256.67</td>
<td>-26.73</td>
<td>14.73</td>
<td>0.72</td>
<td>13.05</td>
<td>DC</td>
</tr>
<tr>
<td>119629521109841562</td>
<td>237.68</td>
<td>16.06</td>
<td>14.41</td>
<td>0.91</td>
<td>91.01</td>
<td>DC</td>
</tr>
<tr>
<td>492382540566074624</td>
<td>3.01</td>
<td>-54.02</td>
<td>15.37</td>
<td>1.36</td>
<td>73.78</td>
<td>DC</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5717278911884258176</td>
<td>115.09</td>
<td>-17.42</td>
<td>13.16</td>
<td>0.34</td>
<td>9.15</td>
<td>DQ</td>
</tr>
<tr>
<td>5332606522595645952</td>
<td>176.46</td>
<td>-64.84</td>
<td>13.09</td>
<td>0.29</td>
<td>4.64</td>
<td>DQ</td>
</tr>
<tr>
<td>946030529073021440</td>
<td>107.56</td>
<td>37.67</td>
<td>13.58</td>
<td>0.49</td>
<td>24.37</td>
<td>DQ</td>
</tr>
<tr>
<td>6797171606323993728</td>
<td>305.11</td>
<td>-30.45</td>
<td>12.39</td>
<td>0.10</td>
<td>17.46</td>
<td>DQ</td>
</tr>
<tr>
<td>1932612039116771456</td>
<td>341.64</td>
<td>40.41</td>
<td>11.53</td>
<td>-0.21</td>
<td>48.79</td>
<td>DQ</td>
</tr>
<tr>
<td>51090628850671088</td>
<td>57.83</td>
<td>19.63</td>
<td>14.29</td>
<td>1.02</td>
<td>79.04</td>
<td>DQ</td>
</tr>
<tr>
<td>6812238909081247270</td>
<td>329.21</td>
<td>-24.88</td>
<td>14.41</td>
<td>0.87</td>
<td>67.19</td>
<td>DQ</td>
</tr>
<tr>
<td>2201461976641376512</td>
<td>335.06</td>
<td>60.36</td>
<td>14.25</td>
<td>1.01</td>
<td>94.86</td>
<td>DQ</td>
</tr>
<tr>
<td>303082043081929808</td>
<td>115.74</td>
<td>-12.80</td>
<td>14.45</td>
<td>1.47</td>
<td>76.83</td>
<td>DQ</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>527107256109138176</td>
<td>124.31</td>
<td>124.31</td>
<td>15.21</td>
<td>1.36</td>
<td>38.88</td>
<td>DQpec</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6054148143441683072</td>
<td>184.36</td>
<td>-63.50</td>
<td>13.17</td>
<td>0.30</td>
<td>37.52</td>
<td>DZ</td>
</tr>
<tr>
<td>3327484804002004000</td>
<td>96.91</td>
<td>10.04</td>
<td>12.88</td>
<td>0.22</td>
<td>52.91</td>
<td>DZ</td>
</tr>
<tr>
<td>659883720324218240</td>
<td>335.91</td>
<td>-34.64</td>
<td>12.98</td>
<td>0.15</td>
<td>57.02</td>
<td>DZ</td>
</tr>
<tr>
<td>264529556512242688</td>
<td>350.31</td>
<td>1.04</td>
<td>15.07</td>
<td>1.21</td>
<td>61.06</td>
<td>DZ</td>
</tr>
<tr>
<td>6429838671917056512</td>
<td>306.32</td>
<td>-63.40</td>
<td>14.25</td>
<td>1.01</td>
<td>61.70</td>
<td>DZ</td>
</tr>
<tr>
<td>2067245446933999396</td>
<td>359.98</td>
<td>39.44</td>
<td>14.92</td>
<td>1.01</td>
<td>67.25</td>
<td>DZ</td>
</tr>
<tr>
<td>6505113009316709760</td>
<td>399.00</td>
<td>-55.82</td>
<td>14.73</td>
<td>1.03</td>
<td>31.91</td>
<td>DZ</td>
</tr>
<tr>
<td>4823007896975577088</td>
<td>81.84</td>
<td>-34.46</td>
<td>14.91</td>
<td>1.17</td>
<td>65.17</td>
<td>DZ</td>
</tr>
<tr>
<td>421772952902509024</td>
<td>306.73</td>
<td>-6.43</td>
<td>12.56</td>
<td>0.13</td>
<td>51.72</td>
<td>DZ</td>
</tr>
</tbody>
</table>
primary spectral types, and also other subtypes, such as DO, hot DQ, DAH, PG 1159 objects and various types of subdwarfs, as well as white dwarfs plus main sequence binaries. Moreover, the input data used in [Vincent et al. (2023)] comes mainly from both the Gaia parameter database and Sloan Digital Sky Survey spectra, while in our study, we only focused on the Hermite coefficients from Gaia spectra.

Nevertheless, for comparative purposes, we have constructed a confusion matrix with the spectral classification of the objects that appear in both the present work and [Vincent et al. (2023)]. Both catalogues were cross-matched and 1 103 objects were found in both tables. Of them, six were classified in Vincent et al. (2023) as cataclysmic variables and two as hot subdwarf stars; these objects were disregarded in the construction of the confusion matrix. As such, the resulting confusion matrix, shown in Figure 11, contains 1 095 objects.

The obtained confusion matrix is nearly diagonal, which indicates a general good agreement (86% accuracy considering the Vincent et al. (2023) classification as true labels). The only remarkable exception are magnetic DAH white dwarfs. For the 22 white dwarfs classified as DAH in Vincent et al. (2023), 19 of them are classified by our algorithm as DAs and 3 as DCs. In Section 3.3.1 we showed that our algorithm is practically unable to distinguish DAs from DAHs. The reason is due to the lack of the necessary Gaia spectral resolution to resolve the fine magnetic splitting. Therefore, this result for DAHs is entirely understandable.

Similarly, but to a lesser extent, the 104 DZ candidates in Vincent et al. (2023) are broadly in agreement with our classification (61 objects, 59%), but our algorithm classifies the discrepancies as DC, DA, or DQs. Once again, the low resolution of Gaia spectra prevents a more accurate classification.

We can conclude that, despite the quality of the input data (which, obviously, the better the quality data, the better the performance), our Random Forest algorithm is a feasible tool with very low computer time consumption and model-independent tuning parameters, allowing a reliable and robust classification of white dwarf spectra.

8. Conclusions

By using Artificial Intelligence techniques based on a Random Forest algorithm, we have analysed the information contained in the coefficients of Gaia spectra. Even though these spectra are of low resolution, we have verified their usefulness in classifying the population of white dwarfs into their different spectral types. In particular, we have classified the full 100 pc Gaia white dwarf population into their primary spectral types (i.e. DA, DB, DC, DZ, and DQ), and we have also identified some secondary types (DAH, DBA, and DQpec).

A summary of the main findings is as follows:

1. The Random Forest algorithm is able to classify DA and DC white dwarfs with excellent recall (>97%), DBs with very good recall (>80%), and DQs and DZs with improvable recall (<50%).
Fig. 10: As Fig. 8 but for the different spectral subtypes of non-DAs.
Fig. 11: Confusion matrix of objects that appear both in our classification and in Vincent et al. (2023).

2. In spite of the low recall, DQ and DZ white dwarfs are classified with an excellent precision (>90%).

3. While the algorithm performance is certainly improvable for the correct identification of DQ and DZ white dwarfs, its high precision for these spectral types, as well as DB, allows us to use the classifying algorithm as a white dwarf finder for these specific types.

4. With the possible exception of the DBA and DQpec subtypes, spectral subtypes do not seem to be recognised by the algorithm. Low resolution inherent to Gaia mean spectra seems to be the limiting factor for classification, as non-prominent spectral lines are not expected to be detected in them.

5. Our algorithm has identified 76 DB (most of them, 60, DBA), 60 DQ (9 of them DQpec), 132 DZ, and 2 DAH candidates in a 100-pc radius around the Sun. For comparison, the MWDD classified sample used in validation tests and as training material contained 117 DQ and 125 DZ white dwarfs.

In conclusion, this initial classification of the entire white dwarf population within 100 pc opens the door to more precise studies of mass distribution and luminosity function, among others, based on the spectral classification of these objects. In parallel, we have initiated a spectroscopic follow-up of a large sample of candidate objects to confirm their classification.

Acknowledgements. We acknowledge support from MINECO under the PID2020-117252GB-I00 grant and by the AGAUR/Generalitat de Catalunya grant SGR-386/2021. EMGZ also acknowledges financial support from Banco de Santander, under a Becas Santander Investigación grant SGR-386/2021. EMGZ also acknowledges financial support from Banco de Santander, under a Becas Santander Investigación grant SGR-386/2021.

References

Breiman, L. 2001, Machine Learning, 45, 5
Isern, J., Torres, S., & Rebassa-Mansergas, A. 2022, Frontiers in Astronomy and Space Sciences, 9, 6
Li, X.-R., Lin, Y.-T., & Qiu, K.-B. 2019, Research in Astronomy and Astrophysics, 19, 111
Appendix A: White dwarf Gaia spectra

In this Annex, we show examples of Gaia white dwarf spectra classified by our Random Forest algorithm. In all of them, the expected location of the Balmer spectral lines are shown, as well as some characteristic spectral lines for every spectral type. These include He i spectral lines for DB white dwarfs, Swan bands for DQs and a selection of metallic lines for DZs, detailed in subsection A.6. For comparative purposes, we accompanied each classified spectrum obtained by our algorithm with one spectrum corresponding to a white dwarf classified with the same type from the MWDD.

All shown spectra have been obtained from Gaia internally calibrated spectra, using the GaiaXPy Python package to transform them into wavelength-flux externally calibrated spectra. Oscillatory behaviour in the spectra, as explained in Section 2 at the blue and red extremes, are characteristic of Gaia externally calibrated spectra, and are caused by the behaviour of the Hermite polynomials used to convey them near the extremes.

Appendix A.1: DA spectra

In this section, we show a DA spectrum from a registered white dwarf in the MWDD and a spectrum classified as such in this work in Figure A.1. Balmer lines (Hα, Hβ, Hγ, Hδ, and Hϵ) are marked in both spectra. These lines are very prominent, which explains the excellent recall for DAs (see confusion matrix in the top panel of Figure 1.)

Appendix A.2: DAH spectra

Gaia spectra for those objects classified as DAH are shown in Figure A.2. As hydrogen-dominated DAs, Balmer lines are prominent; Zeeman effect magnetic splitting is not noticeable in the objects labelled as such in the MWDD nor in those identified by our Random Forest algorithm. This serves to illustrate the role of Gaia spectral low resolution in explaining the low recall for DAH white dwarfs by our algorithm.

Appendix A.3: DB spectra

The Gaia spectra of a white dwarf classified as DB by the MWDD and by our algorithm are presented in Figure A.3. Neutral helium lines are present and noticeable in the spectra, and the He i lines at 4 471, 5 015, 5 875, and 6 678 Å have been marked.

Appendix A.4: DC spectra

In Figure A.4 we show the Gaia DC spectra of a MWDD and here classified white dwarf. It can clearly be visualised a characteristic, featureless spectra (except for oscillations caused by Hermite polynomials behaviour). No spectral lines could be found; wavelengths corresponding to H i Balmer lines have been marked nonetheless to stress the featurelessness of these spectra.

Appendix A.5: DQ spectra

The defining characteristic of DQ white dwarfs is the presence of carbon spectral lines. Most atomic carbon lines are outside of the Gaia BP and RP spectral coverage. However, Swan bands, which are vibrational bands characteristic of diatomic carbon (C₂), are in the visible spectra. These are marked in the shown spectra (Figure A.5). Even though Swan bands comprise a high number of vibrational transitions, for clarity they have been marked at 4 600 and 5 050 Å.

Appendix A.6: DZ spectra

Metallic lines are present in DZ spectra. In these example (Figure A.6), we have marked spectral lines of elements that frequently form planetary matter: calcium (Ca ii at 3 933, 3 968, 8 542, and 8 662 Å), magnesium (Mg ii at 4 481 Å), oxygen (O i at 7 772, 7 774, and 7 775 Å), iron (Fe i at 4 578, 5 167, 5 227, and 5 269 Å) and silicon (Si ii at 6 347 and 6 371 Å).

Article number, page 13 of 15
Fig. A.1: Examples of Gaia spectra. *Left panel:* of a white dwarf classified as DA by MWDD. *Right panel:* of a white dwarf classified as DA by our algorithm.

Fig. A.2: As Fig. A.1 but for a DAH classified in MWDD (left panel) and a DAH classified by our algorithm (right panel).

Fig. A.3: As Fig. A.1 but for a DB classified in MWDD (left panel) and a DB classified by our algorithm (right panel).
Fig. A.4: As Fig. A.1 but for a DC classified in MWDD (left panel) and a DC classified by our algorithm (right panel).

Fig. A.5: As Fig. A.1 but for a DQ classified in MWDD (left panel) and a DQ classified by our algorithm (right panel).

Fig. A.6: As Fig. A.1 but for a DZ classified in MWDD (left panel) and a DZ classified by our algorithm (right panel).