Machine learning methods for the search for L&T brown dwarfs in the data of modern sky surveys.[★]

Aleksandra Avdeeva[a,b,*,1]

[a]*HSE University, 20 Myasnitskaya St., Moscow, 101000, Russia*
[b]*Institute of Astronomy RAS, 48 Pyatnitskaya St., Moscow, 119017, Russia*

ARTICLE INFO

ABSTRACT

According to various estimates, brown dwarfs (BD) should account for up to 25 percent of all objects in the Galaxy. However, few of them are discovered and well-studied, both individually and as a population. Homogeneous and complete samples of brown dwarfs are needed for these kinds of studies. Due to their weakness, spectral studies of brown dwarfs are rather laborious. For this reason, creating a significant reliable sample of brown dwarfs, confirmed by spectroscopic observations, seems unattainable at the moment. Numerous attempts have been made to search for and create a set of brown dwarfs using their colours as a decision rule applied to a vast amount of survey data. In this work, we use machine learning methods such as Random Forest Classifier, XGBoost, SVM Classifier and TabNet on PanStarrs DR1, 2MASS and WISE data to distinguish L and T brown dwarfs from objects of other spectral and luminosity classes. The explanation of the models is discussed. We also compare our models with classical decision rules, proving their efficiency and relevance.

1. Introduction

Brown dwarfs are substellar objects that were theoretically predicted (Kumar, 1963; Hayashi and Nakano, 1963) and then discovered 30 years later (Rebolo et al., 1995; Nakajima et al., 1995). Since then, the search (Luhman, 2013; Burningham et al., 2013; Carnero Rosell et al., 2019) and systematic study of known brown dwarfs (Kirkpatrick et al., 1999; Skrzypek et al., 2016; Kirkpatrick et al., 2021) has not stopped. Their mass is insufficient to start and maintain stable hydrogen fusion, which causes them to cool over time. The peak of the radiation intensity falls into the infrared range, so objects are rather faint in the visible spectrum. In the spectral classification, brown dwarfs occupy spectral types L, T and Y.

According to studies (Mužić et al., 2017), the number of brown dwarfs in the Galaxy ranges from 25 to 100 billion objects (with the total number of objects ranging from 100 to 500 billion). Homogeneous and complete samples of brown dwarfs are needed for various kinds of studies: kinematic studies (Smith et al., 2014), studies of binary stars with brown dwarfs (Lodieu et al., 2014), and studies of the parameters of the Galaxy. Brown dwarfs occupy the boundary between stars and planets, and studying their properties helps to refine our understanding of this boundary. Complete and uniform catalogues enable to identify and characterize brown dwarfs with greater accuracy, allowing for a better determination of the lower mass limit for stellar formation and the upper mass limit for planet formation. Moreover, brown dwarfs share similarities with giant exoplanets, making them valuable analogs for studying exoplanetary atmospheres.

By studying the atmospheres of brown dwarfs, similar to exoplanets, we can gain insights into the processes and conditions that govern exoplanet atmospheres, including the presence of clouds, atmospheric composition, and thermal profiles.

Probably the most topical issue regarding brown dwarfs is the L/T transition (Artigau et al., 2009; Gillon et al., 2013; Khandrika et al., 2013). The L/T transition in brown dwarfs is a fascinating phenomenon that is characterized by a sharp change in the near-infrared colours and brightness of brown dwarfs. It is believed to be driven by several possible mechanisms. Cloud models link the sharp transition to the sinking of dust clouds below the photosphere (Marley and Robinson, 2015; Charnay et al., 2018). Instability in the carbon chemistry of brown dwarf atmospheres has been proposed as another mechanism contributing to the L/T transition (Tremblin et al., 2019). Adiabatic convection triggered by this instability can lead to variability across the L/T spectral sequence. Cloud dispersal has emerged as a potential mechanism for the L/T transition (Tan and Showman, 2019). It has been suggested that clouds with larger particle sizes dissipate more easily than clouds with smaller particle sizes. The shift from L to T spectral types may be accompanied by a change from small to large particles, leading to the fragmentation of clouds and a transition to cloud-free atmospheres (Burningham et al., 2017; Saumon and Marley, 2008). A detailed overview of the problem can be found in Vos et al. (2019).

To uncover the nature of the phenomenon mentioned above, as well as to refine the statistical characteristics of brown dwarfs, complete and uniform catalogues of brown dwarfs are required. While spectroscopy is essential for confirming the nature of a brown dwarf and studying its detailed properties, conducting spectroscopic observations for a large number of objects across the entire sky is time-consuming and resource-intensive. Photometric surveys, on

the other hand, can cover a much larger area of the sky and efficiently capture data on numerous celestial objects simultaneously.

By employing colour selection techniques in photometric surveys, one can identify objects that exhibit colours indicative of brown dwarf characteristics. The advantage of using photometric surveys is that they allow for systematic and wide-scale screening of potential brown dwarf candidates, helping to identify promising targets for subsequent spectroscopic observations.

As an illustration, Skrzypek et al. (2016) accomplished this feat by effectively employing data from three surveys: SDSS, UKIDSS, and WISE. They employed a specific colour selection criterion, namely $(Y - J)_{Vega} > 0.8$ and $J < 17.5$, as a decision rule. Through this approach, they were able to discern approximately 1300 brown dwarfs within an area of 3000 square degrees, which corresponds to approximately 7.5% of the celestial sphere. An additional noteworthy implementation in the quest for brown dwarfs is exemplified by Carnero Rosell et al. (2019). Their study also incorporated a decision rule utilizing data from the DES, VHS, and WISE surveys. The following criteria were applied: $(i-z) > 1.2$, $(z-Y) > 0.15$, $(Y_{AB}-J_{Vega}) > 1.6$, and $z < 22$. The imposition of a magnitude limit on the z band was necessary to ensure the completeness of the dataset, thereby excluding any missing values. Within an area spanning 2400 square degrees about 5.8% of the celestial sphere, approximately 12 thousand brown dwarfs were successfully identified through their approach. Carnero Rosell et al. (2019) also presents a comprehensive review of other colour selection works.

When it comes to using colour selection to search for brown dwarfs, incorporating machine learning methods can provide significant advantages. Machine learning techniques can enhance the effectiveness and efficiency of the colour selection process by leveraging large datasets and complex algorithms to identify patterns and make more accurate predictions.

Machine learning methods can help uncover subtle relationships and correlations in multi-dimensional colour space, allowing for the identification of distinct colour signatures associated with brown dwarfs. This can be particularly valuable when dealing with complex and overlapping colour distributions between different objects.

ML methods are increasingly being used for classifying astronomical objects due to the vast amount of data collected in the past decades. For example, Maravelias et al. (2022) combined Support Vector Machine (SVM), Random Forest (RF) and Multilayer Perceptron to classify massive stars in nearby galaxies. The accuracy of the test dataset was 83%. Applying it to other galaxies (not included in the dataset), namely IC 1613, WLM and Sextans A, achieved an accuracy result at the level of 70%, which the authors attribute to different metallicity and extinction effects. The missing data were filled in with simple averages and the *Iterative Imputer* method of the Scikit-learn library (Pedregosa et al., 2011). The Iterative Imputer method calculates the missing values

based on the present values for the features in the same manner as regression models do. This method, being at the same time more robust, showed a better performance in the work.

The interpretable machine learning techniques (Localized General Matrix LVQ and RF) were used in Mohammadi et al. (2022) to detect extragalactic Ultra-compact dwarfs and Globular Clusters. Authors analysed the importance of features and compared them with features that carry physical information of the objects.

This work aims to develop an additional tool for the search for brown dwarfs in large photometric surveys with machine learning methods. That is, based on the set of magnitudes and colours of the object, the model must determine whether the given object is a brown dwarf or not. We also compare our results with some classical decision rules: Burningham et al. (2013) and Carnero Rosell et al. (2019). The summary of these rules is shown in Tab. 2. The tool will be used in future work for the search for previously undiscovered brown dwarfs.

This paper is organised as follows. Sec. 2 describes the dataset and preprocessing of the data, including feature engineering, augmentation and the approach to handling the missing values. In Sec. 3 we apply the machine learning methods to the dataset and explore the importance of features. Finally, in Sec. 4 we compare the performance of machine learning models and the classical decision rules and discuss the robustness of the models.

2. Building the dataset and preprocessing

The dataset is based on L and T brown dwarfs from the Best et al. (2018) catalogue. The catalogue contains information on 1601 L and T type brown dwarfs and 8287 M type red dwarfs, the spectral class closest in physical characteristics to brown dwarfs. Magnitudes in 12 photometric bands and their errors are provided: g, r, i, z, y of Pan-STARRS 1 (Chambers et al., 2016), J, H, Ks of 2MASS space mission (Cutri et al., 2003) and $W1, W2, W3, W4$ of WISE space mission (Cutri et al., 2021).

The catalogue also contains astrometrical information: position, parallax and proper motion. In addition, references to the literature are presented, from which data on proper motion and parallax was taken.

For our Machine Learning models we consider brown dwarfs to be a positive class. To create a representative distribution of negative class objects, we examined the distribution of 100 thousand stars from Gaia DR3 (Collaboration et al., 2016, 2022) by absolute magnitude M_G (Fig. 2a). We have selected 1791 objects from A0 to K9 spectral class in the proportions observed in Fig. 2a from the Simbad[1] database of astronomical objects. Objects that are presented in Simbad are usually well-studied and have solid spectral classifications. Gaia data seem to be short regarding M-type dwarfs, especially, after M3, so we adopt their distribution

---

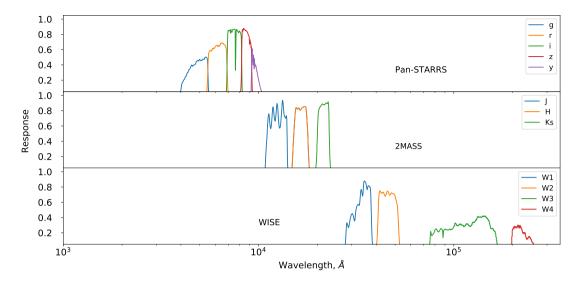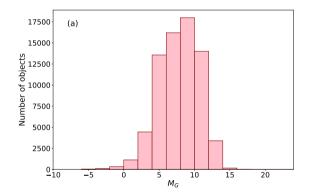[1] http://simbad.cds.unistra.fr/simbad/

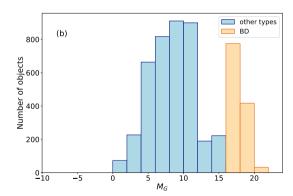**Figure 1:** Response curves of photometric systems.

from Best et al. (2018). The distribution of the dataset we have obtained is shown in Fig. 2b.

It should be noted, the obtained distribution is most likely incomplete, both in M-dwarfs and in earlier spectral classes. The deficiency in M-dwarfs could be due to Best et al. (2018) being not complete for M-type dwarfs outside of 10-pc. The objects of A0 to K9 might tend to be brighter than the actual population and in fact have an underdensity in the range between 12^m and 15^m in i_{PS1} (Fig 3a), which we, however, assume to be not relevant in case of using only colour indices as features.

The objects selected from Simbad were cross-matched with data from the Pan-STARRS DR1, 2MASS, and ALL-WISE catalogues. We chose the matching radius to be $1''$, which is a reasonable value for most surveys for objects with low proper motions, including those used in this paper.

The resulting dataset contains 5669 objects, 1601 of which are of the positive target class. The dataset is available online[2]. As the peak intensity of brown dwarfs falls in the infrared part of the spectrum, their magnitudes in the optical photometric bands (bands g,r,i) are most likely almost beyond the sensitivity limit of the telescope and therefore are missing from the data. The g and r bands of the Pan-STARRS data are missing for almost all objects, so we don't use these magnitudes as features. The i band values are missing for about one-third of the objects of the positive class and a small number of objects of the negative class. Magnitude values in this band are important to us, also for comparison with the classical rules, so we keep them. We also remove $W3$ and $W4$ as they are of poor quality, 90th percentile of the error of magnitude measurement for both magnitudes is about half of a magnitude, which is very noisy for the classification problem we want to solve. As a result, we have 7 magnitudes for each object: i_{PS1}, z_{PS1}, y_{PS1}, J, H, Ks, $W1$, $W2$.

Figure 2: Absolute colour distribution for the Gaia sample (a) and the dataset used in this work (b).

2.1. Train, validation and test

The data is divided into training (60%), validation (20%) and test (20%) sets in a stratified fashion using using sci-kit learn *train_test_split* method. The features are scaled using *StandardScaler* and model hyperparameters are selected on

---

[2]https://github.com/iamaleksandra/ML-Brown-Dwarfs/

the validation set using *optuna* (Akiba et al., 2019). The final performance of the model is calculated on the test set.

2.2. Augmentation

As the number of objects of negative class is almost 2.5 times higher than the number of positive class objects, we use augmentation with Gaussian-distributed noise as an oversampling method to make the dataset more balanced. The data was divided into training, validation and test sub-samples prior to augmentation in order to not let models see the augmented data in the test sample, while being trained on the original prototype of these augmented objects.

We augment the data of positive and negative classes separately, using all objects of positive class and only objects in the range between 12^m and 15^m in i_{PS1} in negative class objects. For each feature error, we calculate the mean and the standard deviation. We then generate normally distributed noise values with the same parameters of the distribution. The noise is added to all values of the corresponding features, which do not have any missed values. Thus we have 8364 objects of which 4155 are positive and 4209 are negative.

2.3. Feature engineering

In a classification of astronomical objects, as in various types of astronomical problems, the colours of objects are even more important than the magnitudes. Colours are characteristic of the energy distribution in the spectrum and are almost independent of distance. To take this into account when classifying, we have added several features - colour indices: $(i - z)_{PS1}$, $(i - y)_{PS1}$, $(z - y)_{PS1}$, $z_{PS1} - J$, $y_{PS1} - J$, $J - H$, $H - Ks$, $Ks - W1$, $W1 - W2$. They are also often used as the colour selection to distinguish brown dwarfs from other objects. We only use colours of the most spectroscopically adjacent magnitudes (see Fig. 1). The two exceptions are $z_{PS1} - J$ since it is commonly used as a colour selection and proved to be useful, and $(i - y)_{PS1}$ since it turned out to be extremely useful for classification (see Sec. 4 for details).

After this procedure, the table contains 17 features for each of the 8364 objects, listed in Tab. 1. Fig. 3 shows how objects of the target class look compared to objects of all other classes in a two-dimensional slice of feature space. We use all of the magnitudes and colours simultaneously (even though the latter ones are the linear combination of the first ones by design) since we process the missing values independently, which sometimes violates these relations (for the details see Sec. 2.4).

As one can see (Fig. 3a), the upper limit of magnitude i_{PS1} differs a lot for our positive and negative objects. This is due to the procedure of building the dataset, i.e. Simbad has four times more objects with $i_{SDSS} > 20$ and a spectral type $\geq L0$ than objects with a spectral type $< L0$. Catalogues, however, contain large amount of faint objects, the majority of which are not brown dwarfs. We therefore should avoid models based primarily on PS1 magnitudes. Although our dataset can be called balanced regarding other magnitudes (from 2MASS and WISE), the magnitude is not only a

function of the luminosity of an object, but also a function of the distance to it, so it is not advisable to rely on these magnitudes as well.

Thus, three cases for each approach are investigated: all magnitudes and colours are used as features (we call this case "all features"), no Pan-STARRS magnitudes are used ("w/o PS magnitudes") and no magnitudes used at all ("only colours").
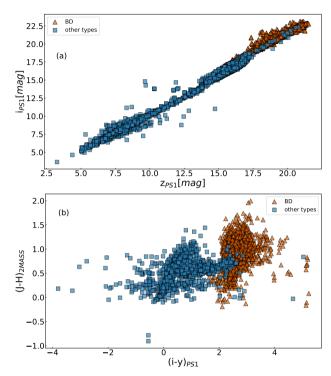

Figure 3: Augmented data: objects of different classes on the magnitude-magnitude (a) and colour-magnitude (b) diagrams. Brown triangles are the objects of positive class (L&T brown dwarfs), blue squares are the objects of negative class. Here blue points are plotted on top of the brown triangles.

2.4. Processing of missing values

As it was mentioned earlier, the dataset has a significant number of missing values. At the part of the spectrum with longer wavelength, this is most likely due to the sensitivity limit of the telescope: brown dwarfs are rather faint objects, and their emission maximum falls into the infrared part of the spectrum. Missing values at the shorter wavelength part of the spectrum seem to occur due to poor-quality measurements or artefacts.

To deal with the missing values, Maravelias et al. (2022) used imputing by means and the Iterative Imputer of the Sckit-learn library, showing that Iterative Imputer provides results more robust and effective in terms of classification.

We test the method by additionally throwing out the magnitude values for 5 percent of objects, which do not have missing values of the particular feature. Then we impute these values using the Iterative Imputer and compare the results to the original feature values for the object.

Tab.1 shows the results of the imputation with the following parameters of the Iterative Imputer:

```
estimator=ExtraTreesRegressor
(n_estimators=150
max_features=14
max_depth=15
min_samples_split=12
initial_strategy='median'
max_iter=20)
```

Tab. 1 contains the information about the fraction of missing values of a particular feature in the dataset and the number of objects that were withheld for the testing. We also compare the 90th percentile error of the feature measurement (the error of magnitude is usually given in the catalogue and the error of the colour index is calculated as the square root of the sum of the squared errors of the magnitudes involved) with the 90th percentile discrepancy in the actual value of the feature and the value predicted by the imputer.
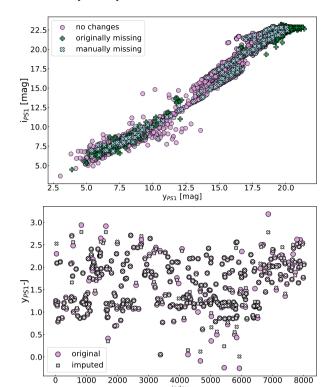
In most cases, the 90th percentile for the discrepancy between the imputed values and the original ones is compatible with the 90th percentile for the error of the feature. Even though the colour values calculated in the Sec. 2.3 are directly related to the magnitude values, it was decided to apply Iterative Imputer to them independently. This allows one to achieve better results and avoid large errors in the calculation of colour values, as can be seen from Tab. 1.

In Figure 4, we can see an example of imputing. The top panel displays a magnitude-magnitude diagram containing both the original and imputed data. It's worth noting that a significant portion of the missing data points are located in the fainter range of the i magnitude. The bottom panel compares the original data to the imputed values for the same objects. Although there are some discrepancies up to 0.5^m in the $y_{PS1} - J$ color, most of the values are predicted accurately. Specifically, for 90% of the stars, $y_{PS1} - J$ value have an error less than 0.063^m.

3. Application of machine learning

Four approaches are tested during the work: Random Forest (RF), Support Vector Machines (SVM), XGBoost and TabNet. As it was said in Sec. 2, we investigate three cases for each approach: "all features", "w/o PS magnitudes" and "only colours". We calculate the score and feature importance for each model, using *SHAP* (Lundberg and Lee, 2017) for RF, XGBoost, SVM and TabNet. Although TabNet has built-in capabilities of calculating importance of features,based on the attention mechanism's dynamic selection of input features, we use *SHAP* as well so as we could compare the results properly.

The *SHAP* method works by evaluating the model's prediction for each instance while permuting the values of a specific feature. This involves shuffling the values of the chosen feature while keeping the rest of the features unchanged. The difference between the model's prediction with the original feature values and the prediction with the

Figure 4: An example of imputed values for $y_{PS1} - J$ and y_{PS1} features. (Top) Original data (pink circles), initially missing data (green plus) and manually missing values (blue crosses) on the colour-magnitude diagram. (Bottom) Comparison of the original data (pink circles) and the values, imputed via the Iterative Imputer method (white crosses). An index is used to avoid contamination from the imputed values of other features.

permuted feature values is used to calculate the Shapley value.

We compare all models to the classical decision rules Carnero Rosell et al. (2019) and Burningham et al. (2013). The Matthews correlation coefficient (MCC) is chosen as the primary metric since it takes into account both False Positive and False Negative predictions. It can be calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FP)(TP + FN)}}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. We also provide the precision and recall scores for they are more intuitive. Precision is a proportion of relevant instances among retrieved instances, while recall is the proportion of relevant instances that have been retrieved. They are defined as:

$$PRECISION = \frac{TP}{TP + FP} \qquad RECALL = \frac{TP}{TP + FN}$$

The classical decision rules and the result of their application to the test dataset after augmentation and imputation

Table 1
Properties of the dataset and the results of the imputing test for every feature for the train part of the dataset. The table includes the fraction of missing values in the dataset (after augmentation) and the number of objects which were withheld for the test (5% of the objects for which the value of the feature was presented). We compare the 90th percentile for the discrepancy between the imputed values and the original ones with the 90th percentile for the error in the measurement of the feature value. The error of the colour index is calculated as the square root of the sum of the squared errors of the magnitudes involved.

Feature	Fraction of missing values	Number of objects withheld	90th percentile of the error	90th percentile of discrepancy
i_{PS1}	17%	208	0.050	0.070
z_{PS1}	5.5%	237	0.050	0.091
y_{PS1}	2.2%	245	0.060	0.109
J	8.8%	228	0.100	0.101
H	8.8%	228	0.120	0.107
Ks	8.9%	228	0.110	0.075
W1	2.1%	245	0.141	0.085
W2	2.0%	242	0.080	0.096
$(i-z)_{PS1}$	18.4%	204	0.067	0.050
$(i-y)_{PS1}$	18.0%	205	0.072	0.038
$(z-y)_{PS1}$	6.7%	234	0.073	0.077
z_{PS1}-J	12.3%	219	0.122	0.048
y_{PS1}-J	10.9%	223	0.130	0.063
J-H	8.8%	228	0.164	0.154
H-Ks	9.1%	225	0.170	0.145
Ks-W1	9.9%	226	0.183	0.143
W1-W2	2.2%	245	0.168	0.176

are summarized in Tab. 2. The MCC score was calculated on the test part of the dataset. It should be mentioned, that even though the filters in different surveys have similar names (i.e. y_{PS1} and Y_{DES}), they are not identical to each other. Therefore, it is not entirely correct to apply the rules made for one survey to the magnitudes of the other survey. However, we have estimated that for our dataset the magnitudes of the same name differ within 0.2 mag which does not increase the score in any case. Also, it is worth mentioning that Burningham et al. (2013) was originally devoted to T-type dwarfs exclusively, yet it shows great performance on L dwarfs, so we use it as a decision rule for both L and T dwarfs.

Although the performance of the decision rules is reasonably high, the actual number of false positive and false negative classifications grows with the number of objects, and this becomes important when we have millions of objects, as in most modern sky surveys (PanSTARRS - 1.9 billion objects, 2MASS - 470 million objects, ALLWISE - 560 million objects), so it makes sense to make an effort to increase the performance and these values (Tab. 2) are the baselines we want to outperform.

The work made use of the following Python software packages: Sckit-learn 1.0.2 version, optuna 2.10.1 version, Tabnet from Pytorch 4.0 version.
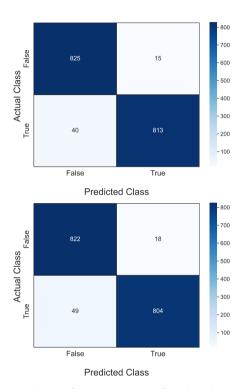
3.1. Random Forest

The decision tree concept is quite similar to the classical colour selections that are traditionally used in astronomy to classify objects. While automated decision trees can be much more efficient than classic decision rules, they tend to overfit, i.e. they learn too much about the data they are trained on and can fail when applied to data they have not seen before. The solution to this problem can be a random forest approach (RF) - an ensemble of decision trees. In this case, the decision about the class an object belongs to, is made based on which class the greater number of trees voted for.

Using *optuna*, we selected the maximum tree depth, the minimum number of samples required to split an internal node, the criterion and the maximum number of features in a node. Tuned parameters and corresponding scores are presented in Tab. 3. Note that the scores are not guaranteed and depend on the dataset the model is tested on, e.g. the

Table 2
Decision rules from the literature.

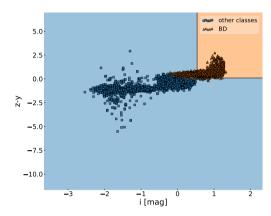| Author	Rule	MCC
Carnero Rosell et al. (2019)	$(i - z) > 1.2$, $(z - Y) > 0.15$, $(Y_{AB} - J_{Vega}) > 1.6$, $z < 22$	0.935
Burningham et al. (2013)	$(z - J)_{Vega} > 2.5$, $J < 17.5$	0.921

Figure 5: The confusion matrices for the decision rules Carnero Rosell et al. (2019) on the top and Burningham et al. (2013) on the bottom applied to the test part of the dataset.

imputing of missing values. For a more objective evaluation of the model we use bootstrap technique (see confidence intervals, Sec. 4 and Fig. 9).

In Fig. 6, an example of a two-dimensional cut shows the separating boundary between classes defined by the RF model. It should be noted that this is a slice in multidimensional parameter space, so it does not represent the number of mismatches. In fact, the performance is very good, meaning that the decision rules underlying that model might be quite complex and cannot be represented on a 2D diagram.

Fig. 10 shows the importance of features for the RF model in all three cases. Seeing Pan-STARRS magnitudes, the model primarily relies on i magnitudes. Without optical magnitudes, $i - y$ becomes the most important feature. This becomes even more pronounced for the RF model that relies on colours only.

Figure 6: Slice of the separating boundary in the feature space according to the RF model.

Table 3
Random forest hyperparameters for three sets of features. Number of trees is 500 for all of the models. Number in max features is the fraction of all available features.

Hyperparameter	All features	No PS magnitudes	Only colours
Max depth	11	13	12
Min samples split	20	9	8
Max features	sqrt	sqrt	0.7
Criterion	entropy	entropy	gini
MCC test score	0.983	0.986	0.975
MCC train score	0.987	0.990	0.990
Precision	0.992	0.992	0.988
Recall	0.992	0.994	0.987

3.2. XGBoost

Boosting is a very popular machine-learning technique. It is a type of ensemble learning that uses the output of the previous model as input to the next one. Instead of training models individually, boosting trains models sequentially, with each new model trained to correct the errors of previous ones. At each iteration, correctly predicted results are given less weight, and incorrectly predicted results are given more weight. It then uses the weighted average to get the final result.

We are also interested in boosting from the point of view that often a certain principle of handling missing values is built into the algorithm. Two popular boosting models are CatBoost (Prokhorenkova et al., 2017) and XGBoost (Chen

Table 4
XGBoost hyperparameters for three sets of features.

Hyperparameter	All features	No PS magnitudes	Only colours
Max depth	15	15	10
Learning rate	0.340	0.126	0.033
Subsample	0.05	0.04	0.11
Gamma	0.323	0.548	0.996
Reg alpha	0.82	0.48	0.03
MCC test score	0.978	0.972	0.969
MCC train score	0.980	0.978	0.974
Precision	0.992	0.989	0.985
Recall	0.985	0.982	0.985

Figure 7: Slice of the separating boundary in the space of features according to the SVM model.

Table 5
SVM classifier hyperparameters for three sets of features.

Hyperparameter	All features	No PS magnitudes	Only colours
Kernel	rbf	linear	rbf
C	1.150	0.729	0.792
Gamma	0.298	0.066	0.453
MCC test score	0.981	0.984	0.958
MCC train score	0.982	0.980	0.968
Precision	0.989	0.986	0.972
Recall	0.992	0.998	0.986

and Guestrin, 2016). CatBoost only has built-in filling in the missing values with some specific values, but XGBoost uses the following strategy: each node is assigned a default solution, and this works well in many cases. Therefore, we choose XGBoost for the task and also compare the performance of the model on the missing values filled in by the default method of XGBoost and filling with the Iterative Imputer.

On a test dataset with the default missing value algorithm, XGBoost gives MCC = 0.96. When training and testing on data in which we filled the missing values using the Iterative Imputer, the performance reaches MCC = 0.986. Thus, we conclude that Iterative Imputer in this case is not only a more robust method but also has a better effect on model performance.

The number of estimators is fixed at 500. The following hyperparameters were optimized with optuna: maximum tree depth, learning rate (the step size reduction used during the update to prevent overfitting), subsample ratio of the training instances, reg_alpha (regularization term) and gamma - the minimum loss reduction required to create a further partition on a leaf node of the tree. Optimized values are presented in Tab. 4.

The feature importances is presented in Fig. 10. Trained on all features XGBoost relies mainly on the i magnitude of the Pan-STARRS survey, as does RF. If Pan-STARRS magnitudes are excluded from the feature list, $i - z$, $z - J$ and $i - y$ colours became the dominant features, however, the J magnitude of the 2MASS survey remains important.

3.3. Support Vector Machine

The support vector machine (Cortes and Vapnik, 1995) is another widely used and well-developed method. The principle of the support vector machine is to find a line, surface or hypersurface that would separate classes in the feature space. The fitting process maximizes the distance from each point to the decision boundary (reference vector).

We tuned the regularization parameter C, the kernel type and the kernel coefficient ('gamma', for 'rbf' kernel). The decision function was set to one-versus-one ('ovo') since it is
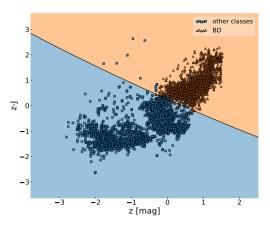
a binary classification and weights of classes were automatically adjusted inversely proportional to class frequencies in the input data. For tuned parameters in every case see Tab. 5. It is seen from the table that models are very similar for all of the cases and so are the most important features (see Fig. 10). As SVM mainly rely on colour indices, the importance distributions do not change significantly when some or all of the magnitudes get excluded.

Fig. 7 shows a separating boundry constructed by the SVM model using a 2D cut as an example. It is worth noting that this is only a slice in which the remaining feature values are taken in some neighbourhood of the mean, so a large number of points that fell into the wrong area does not mean an actual misclassification.
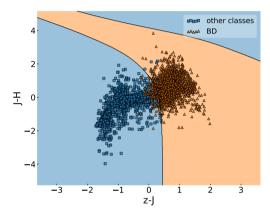
3.4. TabNet

TabNet (Arik and Pfister, 2019) is a deep learning neural network that utilises attention to select important features at each step of the decision-making process so that only the most important features are used. In this case, the choice of features depends on the object, and, for example, it can be different for each row of the training data set. In the end, you can see which features the model has focused on the most.

TabNet consists of several steps, each step is a block of components, with the number of steps being a hyperparameter. Each step gets its vote in the final classification, which

Table 6
TabNet hyperparameters for the different sets of features.

Hyperparameter	All features	No PS magnitudes	Only colours
N_d	16	12	12
N_a	16	28	12
Number of steps	3	3	2
Gamma	1.2	1.6	1.6
N_shared	2	2	1
MCC test score	0.983	0.986	0.975
MCC train score	0.987	0.990	0.990
Precision	0.992	0.992	0.988
Recall	0.992	0.994	0.987

Figure 8: Slice of the separating boundary in the space of features according to the TabNet model.

mimics the ensemble classification. Other hyperparameters are the width of the decision prediction layer (N_d), the width of the attention embedding for each mask (N_a), number of shared Gated Linear Units at each step (N_shared) and the coefficient for feature reusage in the masks (Gamma). We fitted the hyperparameters of TabNet on optuna using MCC score as the metric. The fitted parameters are presented in Tab. 6.

The model is trained using gradient descent optimization with Adam as the optimizer. The validation part of a dataset is used in training to prevent overfitting, so the result of the model training is the configuration, that provides the best scores, both on training and validation data. A separating boundary is shown in Fig. 8, as it can be seen it is more complex then the boundaries of other models.

4. Results and discussion

In this section, we summarise the results of the application of machine learning methods to the brown dwarf search problem.

We trained four models on the data: Random Forest Classifier, SVM Classifier, XGBoost Classifier and TabNet Classifier.

Confidence intervals for the MCC scores of the obtained models are presented in Fig. 9. Confidence intervals are calculated via the bootstrap method with 100 samples half the length of the test data set. The coloured box is the interval from the 25th percentile to the 75th percentile and the median value is represented by a black line. Error bars show the minimum and the maximum value with the outliers marked as diamonds. The baseline values obtained using decision rules from the literature are represented in Fig. 9 as dashed lines.

All models on the full set of features and the features without PS1 magnitudes provide almost the same results, but some performed slightly better. If we use only magnitude features, the performance decreases. However, such models still outperform the baselines.

Although the performance of the models is nearly the same, they differ in terms of robustness. The importance of features of the models is presented in Fig. 10. Random Forest and XGBoost rely primarily on the i magnitude in their decisions, if any. At the same time, SVM and Tabnet on the full set of features seem more robust, as they rely mainly on colour indices. Tabnet also tends to pay a lot of attention to the magnitudes, although it performs the best in the case of using only colours. Table 7 shows true positive, true negative, false positive and false negative values for all of the models obtained on the test part of the dataset.

According to Carnero Rosell et al. (2019) and Skrzypek et al. (2015), $(i - z)_{PS1}$ and $y_{PS1} - J$ colour indices are expected to be the most important feature since they have the largest variation across the M/L transition. While the $(i - z)_{PS1}$ colour index is important to SVM classifiers and in some cases for RF and TabNet, most of the models do not consider them essential.

It is expected that $z_{PS1} - J$ is the most important feature as Burningham et al. (2013), which we used as a baseline, is nearly entirely based on the above colour. However, it only plays a secondary role in most of the models. Unlike the previous works, we revealed the importance of the $(i-y)_{PS1}$ colour index. It is the most important feature in most cases and the colour selection $(i - y)_{PS1} > 1.88$ alone gives an MCC score of 0.968 on the testing data. Other colour indices could be important in the case of multi-class classification, for example, L and T-type dwarfs differ drastically in $W1 - W2$, although it has almost no relevance for the problem investigated in the present article.

5. Conclusion

In this paper, we compiled a dataset of L and T type brown dwarfs (marked as positive class) and objects of other spectral types (marked as negative class) from literature. It is important to acknowledge potential limitations and biases that could impact the dataset's representativeness and generalizability. While we have compiled the data set in such a way as to reproduce the observed distribution in absolute magnitudes, the distribution in apparent magnitudes may be not representative, shifted towards brighter objects. This

Table 7
True positive (TP), true negative (TN), false positive (FP) and false negative (FN) values as well as Precison and Recall scores for four models: Random Forest (RF), XGBoost, Support Vector Machine (SVM), and TabNet. Each model was trained on three sets of features, labeled as "All features", "w/o PS magnitudes", and "only colours".

Model	TP	TN	FP	FN	Precision	Recall
Random Forest						
All features	846	7	846	7	0.992	0.992
W/o PS magnitudes	848	7	833	5	0.992	0.994
Only colours	842	10	830	11	0.988	0.987
XGBoost						
All features	840	6	834	13	0.992	0.985
W/o PS magnitudes	838	9	831	15	0.989	0.982
Only colours	840	13	827	13	0.985	0.985
SVM						
All features	846	9	831	7	0.989	0.992
W/o PS magnitudes	851	12	828	2	0.986	0.998
Only colours	841	24	816	12	0.972	0.986
TabNet						
All features	850	9	831	3	0.992	0.992
W/o PS magnitudes	851	12	828	2	0.992	0.994
Only colours	846	13	827	7	0.988	0.987

most likely does not affect the performance of the model in the case of "only colours" mode, but may have an impact on models trained on magnitudes too. The incompleteness in M type dwarfs that stems from limitations in the Best et al. (2018) catalogue introduces a bias in the dataset towards the included objects, potentially impacting the generalizability of the results.

Brown dwarfs are faint astronomical sources with peak intensities that fall into the infra-red part of the spectrum, so it is harder to carry out the measurements in optical and far-red filters, such as i and z. We imputed missing values with Iterative Imputer and explored the result. We also imputed the colour indices regardless of corresponding magnitudes in order to reduce the error of imputing. For most of the magnitude features, the imputing error is compatible with the measurement error. For colour index features, imputing error is usually much lower than the corresponding measurement error. Thus, the imputing part of preprocessing is considered to be successful.

Four models, namely, Random Forest Classifier, SVM Classifier, XGBoost Classifier and TabNet Classifier were trained to distinguish brown dwarfs among all objects according to their photometric data. The classification results for all models are consistently high, all models outperform the baselines. However, tree-like models (RF and XGBoost) tend to exploit the faintness of the objects, which is less preferable in terms of robustness of the model. On the contrary, SVM primarily relies on colour indices, which reduce the possibility of misclassification of other types of faint sources.

Examining the features of the models, we have found that the $(i − y)_{PS1}$ colour index is the most important feature in most cases. This colour index can be used subsequently and independently in brown dwarf classification problems. We also confirm that the $z_{PS1} − J$, $(i − z)_{PS1}$ and $(z − y)_{PS1}$ colour indices are important and powerful features in brown dwarf classification.

Along with the overall success of the work, there are several limitations to the applicability of the results, that are yet to be solved. First, potential biases of the dataset were emphasised, including a bias towards brighter objects. These biases can affect the generalizability and reliability of the findings. Second, it is known, that highly red-shifted quasars might also be a source of contamination when distinguishing brown dwarfs from other types of objects (Carnero Rosell et al., 2019). This issue can be resolved using colour cuts in most cases, but can still remain for some cases (Reed et al., 2017). At this point, we haven't considered quasars. Third, while the imputing magnitudes and colour indices independently allow us to predict colour indices much more accurately, this leads to the colour index not being always a linear combination of magnitudes involved. Multi-class classification will also be in scope of future work.
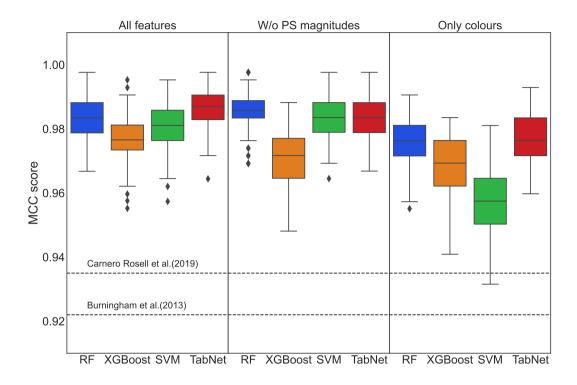
Acknowledgements

Figure 9: Confidence intervals for the scores of models

References

Kumar, S.S., The structure of stars of very low mass., The Astrophysical Journal 137 (1963) 1121.

Hayashi, C. and Nakano, T., Evolution of stars of small masses in the pre-main-sequence stages, Progress of Theoretical Physics 30 (1963) 460–474.

Rebolo, R., Zapatero Osorio, M.R., and Martín, E.L., Discovery of a brown dwarf in the pleiades star cluster, Nature 377 (1995) 129–131.

Nakajima, T., Oppenheimer, B.R., Kulkarni, S.R., et al., Discovery of a cool brown dwarf, Nature 378 (1995) 463–465.

Luhman, K.L., Discovery of a binary brown dwarf at 2 pc from the sun, The Astrophysical Journal Letter 767 (2013) L1.

Burningham, B., Cardoso, C.V., Smith, L., et al., 76 t dwarfs from the ukidss las: benchmarks, kinematics and an updated space density, Monthly Notices of the Royal Astronomical Society 433 (2013) 457–497.

Carnero Rosell, A., Santiago, B., dal Ponte, M., et al., Brown dwarf census with the dark energy survey year 3 data and the thin disc scale height of early l types, Monthly Notices of the Royal Astronomical Society 489 (2019) 5301–5325.

Kirkpatrick, J.D., Reid, I.N., Liebert, J., et al., Dwarfs Cooler than "M": The Definition of Spectral Type "L" Using Discoveries from the 2 Micron All-Sky Survey (2MASS), The Astrophysical Journal 519 (1999) 802–833.

Skrzypek, N., Warren, S.J., and Faherty, J.K., Vizier online data catalog: Photometric brown-dwarf classification (skrzypek+, 2016), VizieR Online Data Catalog (2016) J/A+A/589/A49.

Kirkpatrick, J.D., Gelino, C.R., Faherty, J.K., et al., The field substellar mass function based on the full-sky 20 pc census of 525 l, t, and y dwarfs, Astrophysical Journal Supplement Series 253 (2021) 7.

Mužić, K., Schödel, R., Scholz, A., et al., The low-mass content of the massive young star cluster rcw 38, Monthly Notices of the Royal Astronomical Society 471 (2017) 3699–3712.

Smith, L., Lucas, P.W., Bunce, R., et al., High proper motion objects from the UKIDSS Galactic plane survey, Monthly Notices of the Royal Astronomical Society 443 (2014) 2327–2341.

Lodieu, N., Pérez-Garrido, A., Béjar, V.J.S., et al., Binary frequency of planet-host stars at wide separations. A new brown dwarf companion to a planet-host star, Astronomy and Astrophysics 569 (2014) A120.

Artigau, É., Bouchard, S., Doyon, R., et al., Photometric variability of the t2.5 brown dwarf simp j013656.5+093347: Evidence for evolving weather patterns, The Astrophysical Journal 701 (2009) 1534 – 1539.

Gillon, M., Triaud, A.H.M.J., Jehin, E., et al., Fast-evolving weather for the coolest of our two new substellar neighbours, Astronomy & Astrophysics 555 (2013) L5.

Khandrika, H.G., Burgasser, A.J., Melis, C., et al., A search for photometric variability in l- and t-type brown dwarf atmospheres, The Astronomical Journal 145 (2013).

Marley, M.S. and Robinson, T.D., On the cool side: Modeling the atmospheres of brown dwarfs and giant planets, Annual Review of Astronomy and Astrophysics 53 (2015) 279–323.

Charnay, B., Bézard, B., Baudino, J.L., et al., A self-consistent cloud model for brown dwarfs and young giant exoplanets: Comparison with photometric and spectroscopic observations, The Astrophysical Journal 854 (2018) 172.

Tremblin, P., Padioleau, T., Phillips, M.W., et al., Thermo-compositional diabatic convection in the atmospheres of brown dwarfs and in earth's atmosphere and oceans, The Astrophysical Journal 876 (2019) 144.

Tan, X. and Showman, A.P., Atmospheric variability driven by radiative cloud feedback in brown dwarfs and directly imaged extrasolar giant planets, The Astrophysical Journal 874 (2019) 111.

Burningham, B., Marley, M.S., Line, M.R., et al., Retrieval of atmospheric properties of cloudy l dwarfs, Monthly Notices of the Royal Astronomical Society 470 (2017) 1177–1197.

Saumon, D. and Marley, M.S., The evolution of l and t dwarfs in color-magnitude diagrams, The Astrophysical Journal 689 (2008) 1327–1344.
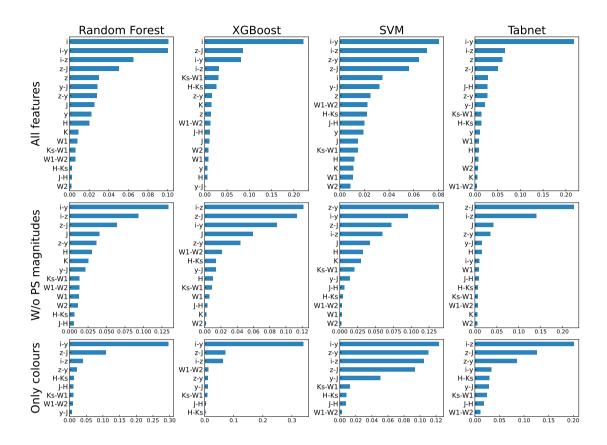
Figure 10: Importance of features for all models. For RF, XGBoost and SVM models we calculate the importance of each feature using *SHAP* and for TabNet we use built-in capabilities of the model.

Vos, J.M., Allers, K., Apai, D., et al., Astro2020 white paper: The l/t transition (2019).

Maravelias, G., Bonanos, A.Z., Tramper, F., et al., A machine-learning photometric classifier for massive stars in nearby galaxies i. the method, arXiv e-prints (2022) arXiv:2203.08125.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

Mohammadi, M., Mutatiina, J., Saifollahi, T., et al., Detection of extragalactic ultra-compact dwarfs and globular clusters using explainable AI techniques, Astronomy and Computing 39 (2022) 100555.

Best, W.M.J., Magnier, E.A., Liu, M.C., et al., Photometry and proper motions of m, l, and t dwarfs from the pan-starrs1 3π survey, Astrophysical Journal Supplement Series 234 (2018) 1.

Chambers, K.C., Magnier, E.A., Metcalfe, N., et al., The pan-starrs1 surveys, arXiv e-prints (2016) arXiv:1612.05560.

Cutri, R.M., Skrutskie, M.F., van Dyk, S., et al., Vizier online data catalog: 2mass all-sky catalog of point sources (cutri+ 2003), VizieR Online Data Catalog (2003) II/246.

Cutri, R.M., Wright, E.L., Conrow, T., et al., Vizier online data catalog: Allwise data release (cutri+ 2013), VizieR Online Data Catalog (2021) II/328.

Collaboration, G., Prusti, T., de Bruijne, J.H.J., et al., The gaia mission, Astronomy and Astrophysics 595 (2016) A1.

Collaboration, G., Vallenari, A., Brown, A.G.A., et al., Gaia data release 3: Summary of the content and survey properties, arXiv e-prints (2022)

arXiv:2208.00211.

Akiba, T., Sano, S., Yanase, T., et al., Optuna: A next-generation hyperparameter optimization framework (2019).

Lundberg, S.M. and Lee, S.I., A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., et al. (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Prokhorenkova, L., Gusev, G., Vorobev, A., et al., CatBoost: unbiased boosting with categorical features, arXiv e-prints (2017) arXiv:1706.09516.

Chen, T. and Guestrin, C., Xgboost: A scalable tree boosting system (2016).

Cortes, C. and Vapnik, V., Support-vector networks, Machine learning 20 (1995) 273–297.

Arik, S.O. and Pfister, T., Tabnet: Attentive interpretable tabular learning, arXiv e-prints (2019) arXiv:1908.07442.

Skrzypek, N., Warren, S.J., Faherty, J.K., et al., Photometric brown-dwarf classification, Astronomy & Astrophysics 574 (2015) A78.

Reed, S.L., McMahon, R.G., Martini, P., et al., Eight new luminous z ≥ 6 quasars discovered via sed model fitting of vista, wise and dark energy survey year 1 observations, Monthly Notices of the Royal Astronomical Society 468 (2017) 4702–4718.