Draft version June 26, 2023
Typeset using LaTeX **twocolumn** style in AASTeX631

# GROUP 2307 - Dormant Black Holes in Binary Systems

Gabriele Bertinelli, Martina Cacciola, Yelyzaveta Pervysheva, Wageesha W. W. Liyanage, and Poornima A. W. W. Mudiyanselage

## ABSTRACT

In Gaia DR3, there are many candidate stars in binary systems with a dark companion. At least two of these are confirmed to be binary systems with a dormant black hole (Gaia BH1 and Gaia BH2). The goal of the project is to aid in understanding how these systems form and which are the relevant features that ruled the evolution. The research work involves comparing data from SEVN simulations with observations: using machine learning techniques, our goal is to label the properties of the known sources and to understand the importance of each feature in the evolution. The classification results prove reliability and coherence, but fail in framing entirely the astrophysical interpretation, leading to a challenge towards a more efficient model or a different dataset.

## 1. INTRODUCTION

With the advent of gravitational wave astrophysics and advances in precise astrometry of numerous stellar sources, the search for compact objects is thriving. Rarely seen massive binaries containing a compact object play a vital role in the evolution towards compact object mergers. With Gaia Data Release 3 (DR3)[1], the first Gaia astrometric orbital solutions for binary sources have become available, revealing a large number of such binary candidates. In the present report, we are interested in the rare cases in which a black hole (BH) is evolving in a binary system with a Main Sequence star (MS). Since these systems are extremely difficult to observe, we deal with a small set of non-interacting BH+MS candidates, among which four are confirmed systems: BH1 (El-Badry et al., 2022), VFTS 243 (Shenar et al., 2022), HD130298 (Shenar, T. et al., 2022), BH2 (El-Badry et al., 2022). The comparison is made with respect to a larger dataset, obtained via SEVN (Stellar EVolution for N-body), a rapid binary population synthesis code (Iorio et al., 2022). It gets as input the initial conditions of stars or binaries (masses, spin, semi-major axis, eccentricity, etc.) and evolves them. Stellar evolution is calculated by interpolating pre-computed sets of PARSEC stellar tracks (Bressan et al., 2012). On the other hand, binary evolution is implemented by means of analytic and semi-analytic prescriptions. The implementation proves to be flexible, so that every model can be easily changed or updated.
We aim to understand what kind of processes these systems are likely to experience during their lifetime: using machine learning techniques, such as Deep Neural Network and XGBoost, we tried to infer and label the properties of the known sources and understand the importance that each feature has had in the evolution of the systems. We tried, moreover, to match some of the candidates to the simulated systems in order to retrieve the guessed full evolution history of those candidates.

[1] https://www.cosmos.esa.int/web/gaia/data-release-3

## 2. METHODS

### 2.1. *Selecting the systems from simulated data*

We work on three datasets, each corresponding to a different metallicity ($Z = 0.02$, 0.001, 0.0001) and value of the efficiency of the common envelope ($\alpha = 0.5$, 1, 5). The BH+MS star systems have been selected, with care to consider only those inside a binary (with a semi-major axis not null). To discriminate between interacting and non-interacting systems, we must take into account the Roche Lobe overflow radius. The radius of the sphere is described by the following approximated equation (Eggleton, 1983):

$$\frac{r_1}{a} = \frac{0.49\, q^{2/3}}{0.6\, q^{2/3} + \ln\left(1 + q^{1/3}\right)} \tag{1}$$

where $r_1$ is the radius of the sphere whose volume approximates the Roche Lobe of mass $M_1$, $a$ is the orbital separation of the system and $q = M_1/M_2$ is the ratio of the two masses. If the radius of the star or the BH in the system is greater than the Roche Lobe radius $r_1$, then the binary is classified as interacting. Otherwise, for $r \leq r_1$ it is non-interacting.
Using the classical Keplerian formula, a column corresponding to the period is then estimated. For visualization purposes only, we also calculate the time spent by the system in the BH+MS phase, the so-called Elapsed

`BWorldtime`, to assess the relative importance of that phase in the overall evolution. The initial time value is extracted from non-interacting systems. By selecting bound systems composed of a BH and a star that is no longer in MS phase, we compute the time at which each system exited the BH+MS phase.

### 2.2. *Selecting the evolution channel*

The SEVN simulation records each evolutionary phase the two objects underwent, so we can add a label to the selected ones to indicate the type of interaction that occurred when the system consisted of two stars (`BEvent`). We divide the systems according to the following: at least one Common Envelope (`CE`), stable Mass Transfer (`MT`) and Non-Interacting (`NI`). The data are then restricted to meaningful quantities, among them we have: the BH/star masses (`Mass_0` and `Mass_1`), the eccentricity (`Eccentricity`), the logarithm of period (`logP`), the elapsed time spent by the system in the BH+MS phase (`bwt_elapsed`). At this point, we produce a scatter plot with multiple subplots (Fig. 1), each showing the relationship between two quantities in the space of parameters. To consider the past evolution, the data is grouped on the basis of the type of interaction, in order to superimpose the scatter points of different colors to differentiate the history of the objects in each bin. The color scale of the bins is based on the number of systems inside. The points representing the confirmed systems are displayed with relative error bars and with varying sizes based on the elapsed time in the BH+MS phase. We aim to gain insights into the position in the phase space occupied by the four confirmed BH+MS systems. For a meaningful comparison, it is recommended to select simulated data that have solar-like metallicity ($z = 0.02$). This recommendation is based on the fact that three of the four real systems have metallicities similar to those of the Sun.

### 2.3. *Classification models*

In order to have the final dataset ready for the training of the classification models, we add more randomness to it performing several shuffling. In this way, we remove any patterns or biases that might exist in the original order of the data. Then, we rebalance and optimize the dataset, ensuring that each label category has the same number of samples: in such a way, we improve generalization, preventing any misleading evaluation metrics and avoiding biases toward the majority class.

#### 2.3.1. *Deep Neural Network*

We implement a Deep Neural Network (DNN) that consists of multiple layers of interconnected nodes. The first layer receives the input data, each node representing a specific feature. Each node computes the output for the next layer using activation functions, which introduce non-linearity into the DNN, enabling it to learn complex patterns. It is particularly effective for supervised learning tasks, such as classification, where input data has corresponding labels or target values. The connections between neurons are assigned weights that determine the strength or importance of the connection. The DNN learns to make accurate predictions by iteratively updating the weights to minimize the difference between its predictions and the true labels in the training data. This process is typically done using an optimization algorithm, such as gradient descent (in our case `Adam`, an extended version of stochastic gradient descent).

In our experimentation, we conducted several tests and used `GridSearchCV` to determine the optimal structure for the DNN. The chosen model architecture consists of four hidden layers, each comprising 20 neurons. Additionally, an input layer with the same number of neurons as features and an output layer with three neurons (equal to the number of labels to be investigated) were incorporated. The activation function employed in the hidden layers is the Rectified Linear Unit (`ReLU`). Since the model aims to classify new data, we selected the `categorical cross-entropy` loss function: it is suitable for multiclass classification tasks and facilitates the training process by penalizing incorrect class predictions. We set the number of epochs at 150 and the batch size at 750, in order to strike a balance between model performance and computational efficiency. The DNN model was trained using all the features present in the dataset. A scaling operation was performed before training for certain features, namely `Mass_BH`, `Mass_1`, `Eccentricity` and `logP`. By doing so, we prevent certain features from dominating the learning process solely based on their larger values. This allows the gradient descent algorithm to flow more smoothly and converge faster towards the minimum of the cost function. Therefore, the model better interprets the relative importance and contribution of each feature when making predictions or determining the importance of the features in the labeling process. After training the model, we proceeded to plot the performance metrics. Specifically, we show the accuracy and loss values over the course of training to assess the learning progress of the model (Fig. **??**). Furthermore, we generated a confusion matrix during the validation phase, which displays the number of correctly predicted labels for each class and the number of misclassifications (Fig. **??**). Furthermore, we created a "classification report", providing a

comprehensive summary of various classification performance metrics: precision, recall, F1 score and accuracy, for each class and the accuracy of the model. We observe that the three labels are recognized accurately in more than 95% of cases, and the overall accuracy of the model is slightly above 97%. These metrics indicate that the model performs exceptionally well in classifying the data and achieving high accuracy across all classes.

#### 2.3.2. *DNN for classification*

In order to effectively classify the dataset that contains candidate BH+MS systems, the model must be trained using the exact features present in that particular dataset. These features include `Mass_BH`, `Mass_1`, and `logP`. It is crucial to note that these features have been appropriately scaled, just as in the previous model, to ensure consistency and enhance the training process. The overall structure of the DNN used for this classification task remains largely unchanged. However, a notable addition has been made by incorporating a dropout layer after each hidden layer, excluding the input and output layers. Dropout layers are implemented to prevent overfitting and improve the generalization capability of the model. By introducing dropout, the DNN becomes more resilient and less prone to relying too heavily on specific neurons or features during training. Based on the findings of GridSearchCV, we decided to set the number of epochs to 150. The batch size value of 1000 was selected for computational efficiency purposes during model training, as it has been observed not to have a significant impact on model accuracy. These parameter values were determined to strike a balance between model performance and computational efficiency. We observed that the NI label is accurately recognized in over 97% of cases, while the other two labels are recognized in approximately 88% of cases (Fig.**??**). The overall accuracy of the model is around 91%. The lower accuracy can probably be attributed to the fact that the model is trained on only three features, which may not provide enough information to accurately capture complex patterns and relationships in the data.

#### 2.3.3. *XGBoost*

In addition to employing a DNN, we conducted experiments using various classification algorithms in their default configurations. These algorithms were evaluated and ranked according to their performance metrics, including Accuracy, Precision, Recall, and F1-Score. The results shown in Fig.**??** indicate that `RandomForestClassifier`, `BaggingClassifier`, `ExtraTreesClassifier`, `DecisionTreeClassifier`, and `XGBClassifier` demonstrated similar performances. In particular, all of these algorithms are based on the principles of decision trees and ensemble methods. After careful consideration, we have chosen to utilize XGBoost due to its ability to handle complex relationships between features and its efficiency in processing large datasets.

`XGBoost` (eXtreme Gradient Boosting) is an ensemble learning method that combines weak predictive models, usually decision trees, to create a strong predictive model. The algorithm works iteratively by training and adding decision trees to the ensemble. During each iteration, it identifies the areas where the previous trees have made errors and focuses on correcting those errors in the subsequent trees. This process, known as gradient boosting, helps improve the model's accuracy. `XGBoost` starts with an initial weak predictive model and then constructs new trees that minimize the loss function by learning from the gradients of the errors. The predictions from all the trees are combined to make a final prediction. Once the training is complete, `XGBoost` produces a final prediction by aggregating the predictions from all the trees.

After conducting a `GridSearchCV` analysis on multiple hyperparameters, our findings reveal that the default `XGBoost` model consistently outperforms other models when considering all available input features. The model performs extremely well, achieving accuracy rates of more than 97% for the three labels and overall accuracy of more than 98% (Fig.**??**).

#### 2.3.4. *XGB for classification*

Using the XGBoost model, we proceed with the classification of the dataset containing candidate BH+MS systems, using the following features `Mass_BH`, `Mass_1` and `logP`, as explained before. After performing a `GridSearchCV`, the optimal hyperparameters were determined to be: `learning_rate`=1, `n_estimators`=300, and `max_depth`=7. The overall accuracy of the model is approximately 92%. When assessing the accuracy of specific labels, it was observed that the accuracy of CE and MT labels is around 90%, while for the NI label, the accuracy reaches 98%. These findings indicate the model's proficiency in correctly classifying instances across different labels, with particularly high accuracy for the NI category.

### 2.4. *SHAP explainer*

SHAP (SHapley Additive exPlanations) values are a method used to explain the predictions of machine learning models, based on cooperative game theory (Lundberg and Lee, 2017). SHAP values aim to quantify the impact of each feature on the prediction outcome by considering all possible combinations of features. The contribution of each feature is determined by comparing

the predictions with and without the feature included. SHAP values offer a unified framework for measuring feature importance that is consistent and locally accurate. They also satisfy desirable properties such as consistency, meaning that if a feature is removed or added, the SHAP values change accordingly. In our study, we used three types of visualization to gain insight into the importance of each feature. Among them, we selected the violin type, a summary plot for each evolution channel: it helps in understanding the impact of each feature on the model's predictions. The vertical axis represents the features of the model. The features are presented in descending order of importance. The horizontal axis represents the range of SHAP values. It spans from negative to positive values, indicating the direction and magnitude of the feature's influence on the predictions. The color shows whether that variable is high (in red) or low (in blue) for that observation. Each violin-shaped distribution represents the density of the SHAP values for a specific feature. A wider section of the violin indicates a higher density of SHAP values, indicating a stronger influence of the feature on the predictions. The plot can also reveal the presence of different subgroups or distinct patterns in the data that affect the predictions differently (Figs. **??** and **??**).

### 2.5. *Candidates Classification*

After the training of the models, we want to classify the BH+MS candidates in the Gaia DR3 dataset. The dataset has been cleaned and processed, narrowing to information on the masses of the two objects and their periods, with upper and lower limits, and a measure of the reliability of the data called *plating*. The four confirmed BH+MS systems have also been added to the candidate dataset. To proceed with the analysis, we generated 10000 Gaussian samples for each feature of every system, which closely adhere to the statistical characteristics of the respective features in each system. Thus, we start with the actual classification of the generated dataset: first with the DNN model and after with the XGBoost model. During the classification process, each binary system is associated with a percentage value for each label, which provides insight into the success and level of certainty in the classification. By setting a threshold based on statistical significance, (1, 2, or 3 standard deviations), we can identify systems that have been classified successfully and with a reasonable level of confidence (Tables 2.5 and 2.5).

Subsequently, we compare the distribution of simulated data from SEVN, used for model training, with the candidate binary systems (Andrew et al., 2022). We narrow down the analysis to those categorized as 'gold' since their inferred parameters are more reliable compared to 'silver' and 'bronze'. The reason for this is that when observing a binary system of this kind, the orbital period is the only well-constrained measurement. Masses, on the other hand, are not as straightforward to measure, and numerical values should always be treated with caution. Furthermore, we have chosen, from the simulated data, the system with a metallicity of $z$=0.02. This decision was made because the systems observed by GAIA exhibit, overall, solar-like metallicity. By comparing the simulated data with the observed gold systems (Fig. **??**), we can assess how well the SEVN model captures the characteristics of real binary systems. If the simulated data closely aligns with the gold systems, it provides evidence that the model is accurately capturing the underlying physics governing these binary systems. However, discrepancies between the simulated and observed distributions could indicate potential limitations or missing aspects in the simulations, or poor measurement of the physical parameters of the observed systems.

| **Label** | $\sigma \geq 1$ | $\sigma \geq 2$ | $\sigma \geq 3$ |
|---|---|---|---|
| NI | 1301 (28.0%) | 762 (16.4%) | 418 (9.0%) |
| CE | 527 (11.34%) | 61 (1.31%) | 17 (0.37%) |
| MT | 679 (14.61%) | 2 (0.04%) | 2 (0.04%) |

**Table 1.** Percentage value for each label, divided by levels of statistical significance. Systems not well defined: 2139 (46.04%) (referring to $1\sigma$).

| **Label** | $\sigma \geq 1$ | $\sigma \geq 2$ | $\sigma \geq 3$ |
|---|---|---|---|
| NI | 1271 (27.36%) | 648 (13.95%) | 231 (4.97%) |
| CE | 393 (8.46%) | 45 (0.97%) | 6 (0.13%) |
| MT | 543 (11.69%) | 2 (0.04%) | 2 (0.04%) |

**Table 2.** Percentage value for each label, divided by levels of statistical significance. Systems not well defined: 2439 (52.5 %) (referring to $1\sigma$).

#### 2.5.1. *Finding similar systems to the real candidates*

We also aimed to identify systems in the SEVN simulation that closely resemble the selected real candidates. Our approach involved calculating the similarity between systems based on the BH mass, the MS mass, and the period (all quantities taken both from SEVN and the real candidates). The equation used is A1. Thus, we predict the history of the selected real candidates, whether they have undergone a `CE` phase, `MT`, or remain `NI`. We determine this by leveraging the history of the 1000 most similar simulated systems. By analyzing their respective histories, we assign a percentage possibility to each scenario for the real candidates. In order to predict the evolution history of our selected

candidate system, we employed a method that involved studying the history of the 10 most similar systems we identified. Our analysis focused solely on the key properties of the BH mass, the MS mass, the period, the semi-major axis, and eccentricity. To gather insights into the potential evolution of our candidate system, we delved into the simulated histories of these 10 comparable systems. By meticulously examining the changes in the aforementioned properties over time, we have assumed that the mean values for each property might be a possible likely evolutionary history of the real candidate system we selected.

## 3. RESULTS

In this section, we present the astrophysical interpretation of the results obtained, in order to check the consistency of the machine-learning approach (both for the DNN model and the XGBoost) with the established theoretical framework. We are extracting the main pieces of information from the SHAP explainer's outcomes (Figs. **??** and **??**).

**Non-Interacting** (`NI`):

- `logP` (Period): High values of the period contribute greatly. This is because systems that have never interacted, and therefore have not developed orbital energy dissipation mechanisms, are characterized by very long periods.

- `Mass_BH`: Intermediate mass values appear favoured. This preference can be attributed to the fact that stars, which subsequently underwent the process of evolving into BHs, and have not interacted with their companions, retain higher mass values. Consequently, these stars evolve into BHs with greater mass compared to systems involving interactions with companions.

- `z`: Based on observations of stellar tracks, it is evident that stars with higher metallicities (e.g., $z$=0.02) tend to exhibit a more pronounced expansion phase at the onset of helium-burning compared to stars with lower metallicities. At the same $M_{ZAMS}$, these stars demonstrate a propensity to attain larger radii more rapidly. Consequently, such stars are more susceptible to exceeding their Roche lobes, initiating a mass transfer process. Thus, the model should show a preference for low-medium values (e.g. $z$=0.001, $z$=0.0001) of metallicity regarding systems categorized as non-interacting, while high values for systems that are evolved via mass transfer or common envelope. In the context of this label, the influence of metallicity on label selection lacks definitive evidence. Nevertheless, in the DNN model, it is consistent with theoretical expectations that lower metallicity values are favoured. On the other hand, this feature interpretation from the XGBoost is inconsistent from an astrophysical point of view.

- `Eccentricity`: In the DNN model, medium to high values of eccentricity contribute slightly to label `NI`. The reason is related to the fact that they did not dissipate orbital energy and thus did not circularize the orbit. In the XGBoost, the interpretation is inconsistent.

- `Mass_1`: It seems not to contribute to labeling.

- `alpha`: It seems not to contribute to labeling.

**(stable) Mass Transfer** (`MT`):

- `logP` (Period): Low period values contribute significantly because, through mass transfer mechanisms, the systems dissipated orbital energy by shortening the period.

- `Mass_1` & `Mass_BH`: When a Roche Lobe (RL) overflow happens, it changes the mass ratio, the masses and the semi-major axis of the binary system. As a consequence, the RL shrinks or expands. If the RL shrinks faster than the donor's radius (or if the RL expands more slowly than the donor's radius) because of the adiabatic response of the star to mass loss, the mass transfer becomes unstable on a dynamical timescale, leading to a stellar merger or a CE configuration. Population-synthesis codes usually implement a simplified formalism in which the mass transfer stability is evaluated by comparing the mass ratio $q = M_d/M_a$ (where $M_d$ and $M_a$ are the mass of the donor, in our case most of the time the objects that became a BH, and accretor star, respectively), with some critical value $q_c$. If the mass ratio is larger than $q_c$, the mass transfer is considered unstable on a dynamical time scale. The critical mass ratio is usually assumed to be large ($> 2$) for stars with radiative envelopes (e.g. MS stars, stars in the Hertzsprung-gap phase, and pure-He stars). If the RL is smaller than the core radius of the donor star, the mass transfer is always considered unstable, ignoring the chosen stability criterion. If both stars have a radius $R \geq RL$, we assume that the evolution leads either to a CE (when at least one of the two stars has a clear core-envelope separation) or to a stellar merger (Iorio et al., 2023). Following this introduction, we can now delve into an explanation

for why medium to high mass values are generally preferred. During their star-star evolutionary phase, it is likely that the mass ratio of the two objects was below $q_c$. Consequently, the information about $q$ is expected to have influenced the mass values in the BH-MS phase, playing a significant role in determining the classification of the system as a `MT` system. It should be noted that it is not crystal clear whether the model has really identified the importance of astrophysically correct values in assigning this label.

- `z`: As stated previously, high metallicity values contribute to an amplified likelihood of a star undergoing mass transfer episodes.
- `Eccentricity`: Slight contribution is given by intermediate values of eccentricity.
- `alpha`: It seems not to contribute to labeling. Moreover, in the XGBoost, the feature interpretation is inconsistent.

(at least one) **Common Envelope** (`CE`):

- `Mass_BH` & `Mass_1`: Based on the previous information, it is possible to think the mass ratio during the star-star phase was above the critical value $q_c$, indicating the occurrence of at least one `CE` phase in the system's evolution. However, it should be noted that it is not crystal clear whether the model has really identified the importance of astrophysically correct values in assigning this label. In order to enhance the model's understanding and recognition of the threshold ($q_c$) described above, one potential avenue for improvement in future iterations is the inclusion of additional training data. This could involve incorporating information regarding the initial mass ratio of the binary system, or the mass ratio in each phase experienced by the donor star. In our training experiments, we found that including the mass ratio of the BH-MS system as a feature led to a notable decrease of at least 10 percentage points in the model's accuracy. Therefore, we decided to exclude this feature to prioritize higher accuracy in the categorization process.
- `logP` (Period) and `Eccentricity`: Low values of period and eccentricity contribute greatly. Through one or more common envelopes the system interacts and loses orbital energy resulting in period reduction and circularization of the orbit.
- `z`: This feature interpretation seems to be inconsistent from an astrophysical point of view.
- `alpha`: It seems not to contribute to labeling.

## 4. CONCLUSIONS

The comparison of the two models' classification results reveals a consistent agreement between them, indicating reliability and coherence. Moreover, a close correspondence is observed between the simulated data and the Gaia observations in the `logP-Mass_BH` phase space (Fig. **??**), suggesting a good congruence. However, when considering the MS masses (`Mass_1`), a significant discrepancy arises, particularly for values below 10 $M_\odot$. The discrepancy poses a challenge in providing a definitive and unequivocal explanation, as the current models do not anticipate MS stars in MS-BH systems to be so lightweight. Several potential explanations can be explored within a scientific context:

- Re-evaluating mass transfer mechanisms: It may be necessary to revisit the role of mass transfer in the evolutionary processes of MS-BH systems. This re-evaluation could involve investigating scenarios in which the mass loss experienced by stars during mass transfer events is more significant than currently accounted for in the models. The hypothesis, although being among the less favourable, involves assuming that MS-BHs form after at least one mass transfer event.
- Limitations of current stellar tracks: The simulations, specifically the SEVN model, employ stellar tracks with a minimum zero-age main sequence mass ($M_{ZAMS}$) of 2.2 $M_\odot$. Considering that Gaia primarily observes stars around this mass range, the absence of stellar tracks for lighter objects in the model might contribute to the observed gap in the simulations. This discrepancy could be addressed by incorporating stellar tracks that encompass a broader range of masses, thereby enabling the simulation of lighter MS objects. Notably, when examining more massive MS stars (represented by the pair of squares on the right in the last row of graphs), the observed data aligns with the simulation, indicating a match between observation and expectation.

In conclusion, there is still a non-negligible degree of inconsistency between the theoretical background and the predictions of the models. Our machine-learning approach does not seem to frame effectively the astrophysical aspect in the features' interpretation. Room for improvement can be found in implementing in a more efficient way the structure of the classification models, or in enlarging the dataset at our disposal in the training phase.

## 5. WORKLOAD SUBDIVISION

The group has worked most of the time in synergy, by keeping in contact and updating each other during the whole period of the project. Thus, each one has helped and given their contribution when there was a need in the challenges encountered by the colleagues. In particular, Bertinelli has put his effort into the preprocessing and classification sections, and also in giving the astrophysical interpretation of the results. Cacciola has worked on the part of preprocessing the dataset, the setting of the DNN and the writing of the report with its relative conclusion and take-home message. Liyanage has contributed to the building and testing of the first model proposed, the DNN. Pervysheva then proceeded with testing different classifiers and with the XGBoost building and testing. Mudiyanselage focused in particular on finding a metric to evaluate quantitatively the similarity with the observed ones.

## ACKNOWLEDGMENTS

We thank the DEMOBLACK group for allowing us to use their hardware and the generated data, without which this project could not have taken place. We would also like to express our thanks to Giuliano Iorio and Erika Korb for their valuable advices and support during the course of this project.

## References

Andrew, S. et al. (Sept. 2022). 'Binary parameters from astrometric and spectroscopic errors – candidate hierarchical triples and massive dark companions in iGaia/i DR3.' *Monthly Notices of the Royal Astronomical Society* 516.3, pp. 3661–3684. DOI: 10.1093/mnras/stac2532.

El-Badry, K. et al. (Nov. 2022). 'A Sun-like star orbiting a black hole.' *Monthly Notices of the Royal Astronomical Society* 518.1, pp. 1057–1085. DOI: 10.1093/mnras/stac3140.

Bressan, A. et al. (Nov. 2012). 'PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code.' *Monthly Notices of the Royal Astronomical Society* 427.1, pp. 127–145. DOI: 10.1111/j.1365-2966.2012.21948.x.

Eggleton, P. P. (1983). 'Approximations to the radii of Roche lobes.' *The Astrophysical journal*.

Iorio, G. et al. (2022). 'Compact object mergers: exploring uncertainties from stellar and binary evolution with SEVN.' DOI: 10.48550/ARXIV.2211.11774.

Iorio, G. et al. (June 2023). 'Compact object mergers: exploring uncertainties from stellar and binary evolution with scpsevn/scp.' *Monthly Notices of the Royal Astronomical Society*. DOI: 10.1093/mnras/stad1630.

Lundberg, S. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. arXiv: 1705.07874 [cs.AI].

Shenar, T. et al. (July 2022). 'An X-ray quiet black hole born with a negligible kick in a massive binary within the Large Magellanic Cloud.' DOI: 10.1038/s41550-022-01730-y.

Shenar, T. et al. (2022). 'The Tarantula Massive Binary Monitoring - VI. Characterisation of hidden companions in 51 single-lined O-type binaries: A flat mass-ratio distribution and black-hole binary candidates.' *A&A* 665, A148. DOI: 10.1051/0004-6361/202244245.

# APPENDIX

## A. EQUATIONS

$$d_{\mathrm{Sevn,Real}} = \sqrt{\left(\frac{M_{\mathrm{BH_{SEVN}}} - M_{\mathrm{BH_R}}}{M_{\mathrm{BH_{upper}}} - M_{\mathrm{BH_{lower}}}}\right)^2 + \left(\frac{M_{\mathrm{MS_{SEVN}}} - M_{\mathrm{MS_R}}}{M_{\mathrm{MS_{upper}}} - M_{\mathrm{MS_{lower}}}}\right)^2 + \left(\frac{P_{\mathrm{SEVN}} - P_{\mathrm{R}}}{P_{\mathrm{upper}} - P_{\mathrm{lower}}}\right)^2} \tag{A1}$$

## B. FIGURES

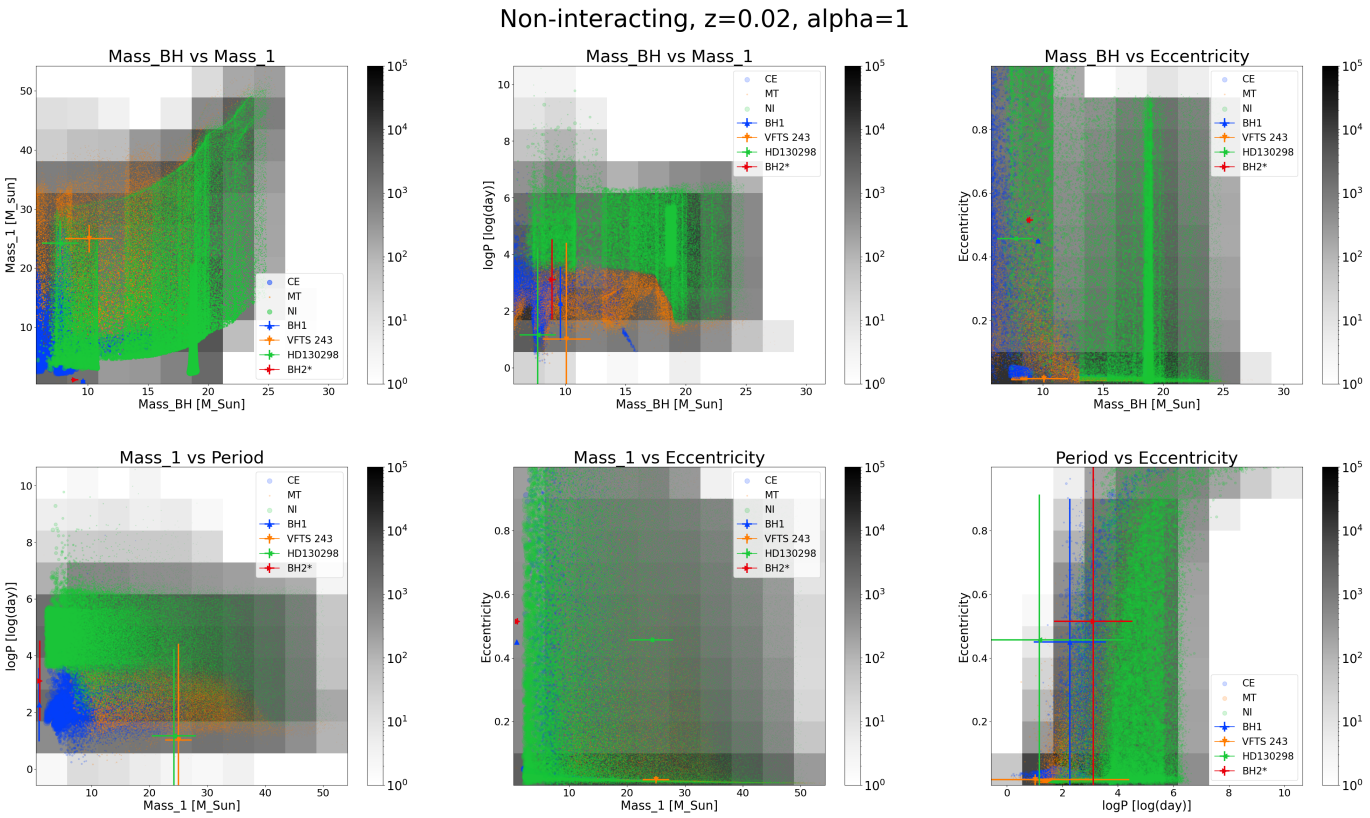


**Figure 1.** Visualization of the distribution of the selected systems in phase space. The background layer of the plot consists of a 2D histogram. Each bin in the histogram is colour-coded based on the number of systems that fall into that bin. The colour scale is displayed in a colorbar and follows a logarithmic scale. In the foreground layer, the simulated systems are plotted and categorized according to their labels: CE, MT, and NI. The points representing the systems are displayed with varying sizes based on the elapsed time in the BH+MS phase. The four confirmed systems are plotted as points with their relative error bars.

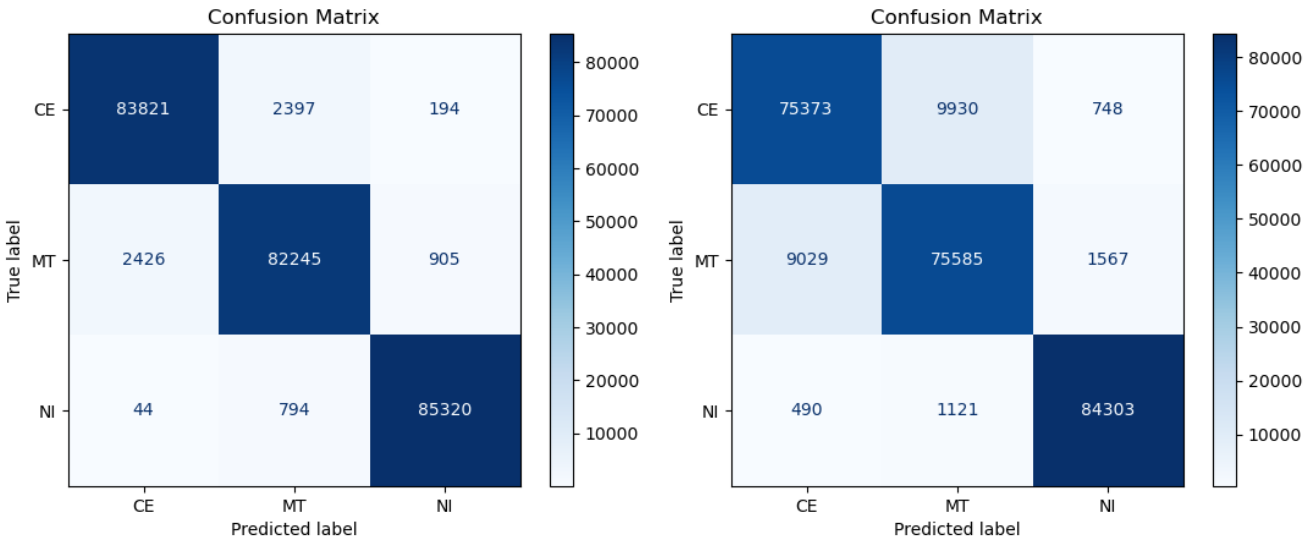


**Figure 2.** Left panel: Confusion matrix of the DNN model, trained using all the features present in the dataset, namely `Mass_BH`, `Mass_1`, `Eccentricity`, `logP`, `z`, and `alpha`. Right panel: Confusion matrix of the DNN model, trained with the features present in the dataset, i.e. `Mass_BH`, `Mass_1` and `logP`.

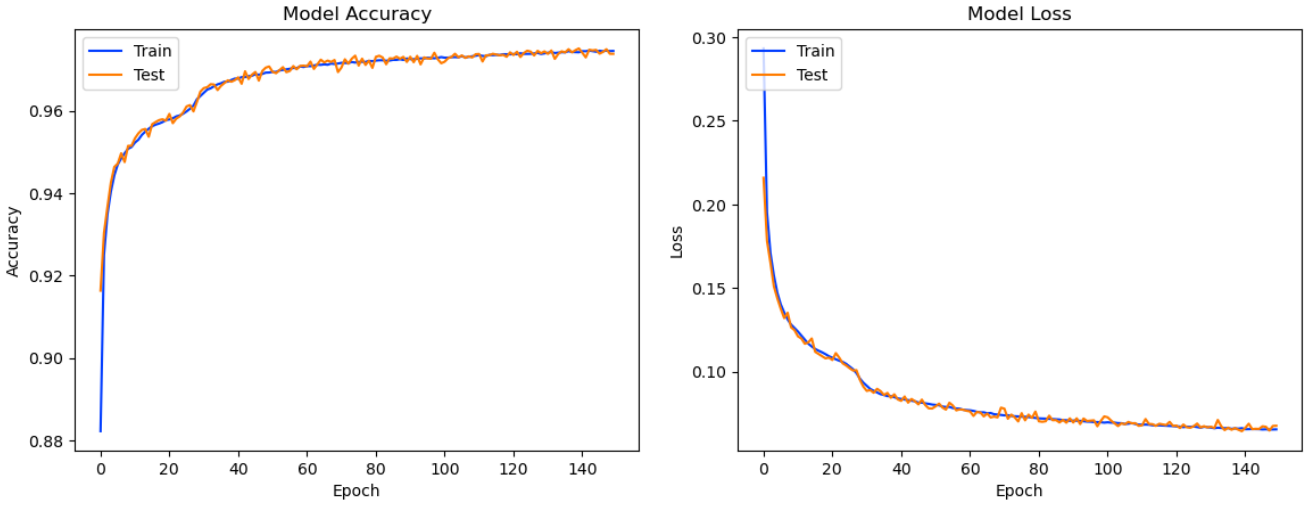


**Figure 3.** Left panel: Evolution of the accuracy of the DNN model for training and validation sets, using all the features. Right panel: Evolution of the loss of the DNN model for training and validation sets, using all the features.

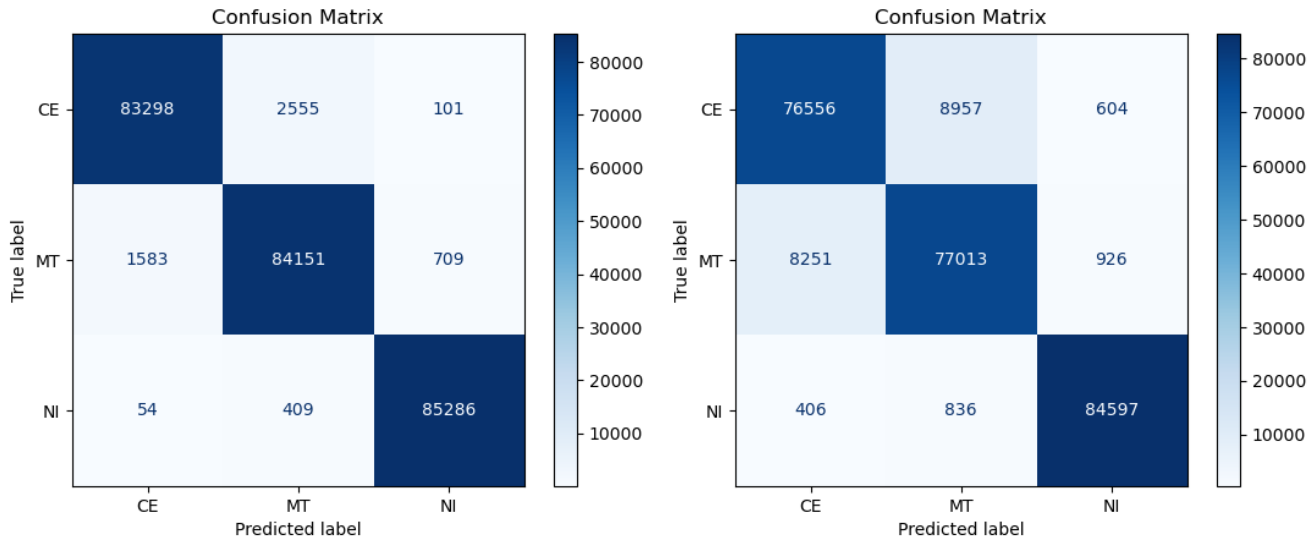


**Figure 4.** Left panel: Confusion matrix of the XGBoost model, trained using all the features present in the dataset, namely `Mass_BH`, `Mass_1`, `Eccentricity`, `logP`, `z`, and `alpha`. Right panel: Confusion matrix of the XGBoost model, trained with the features present in the dataset, i.e. `Mass_BH`, `Mass_1` and `logP`.

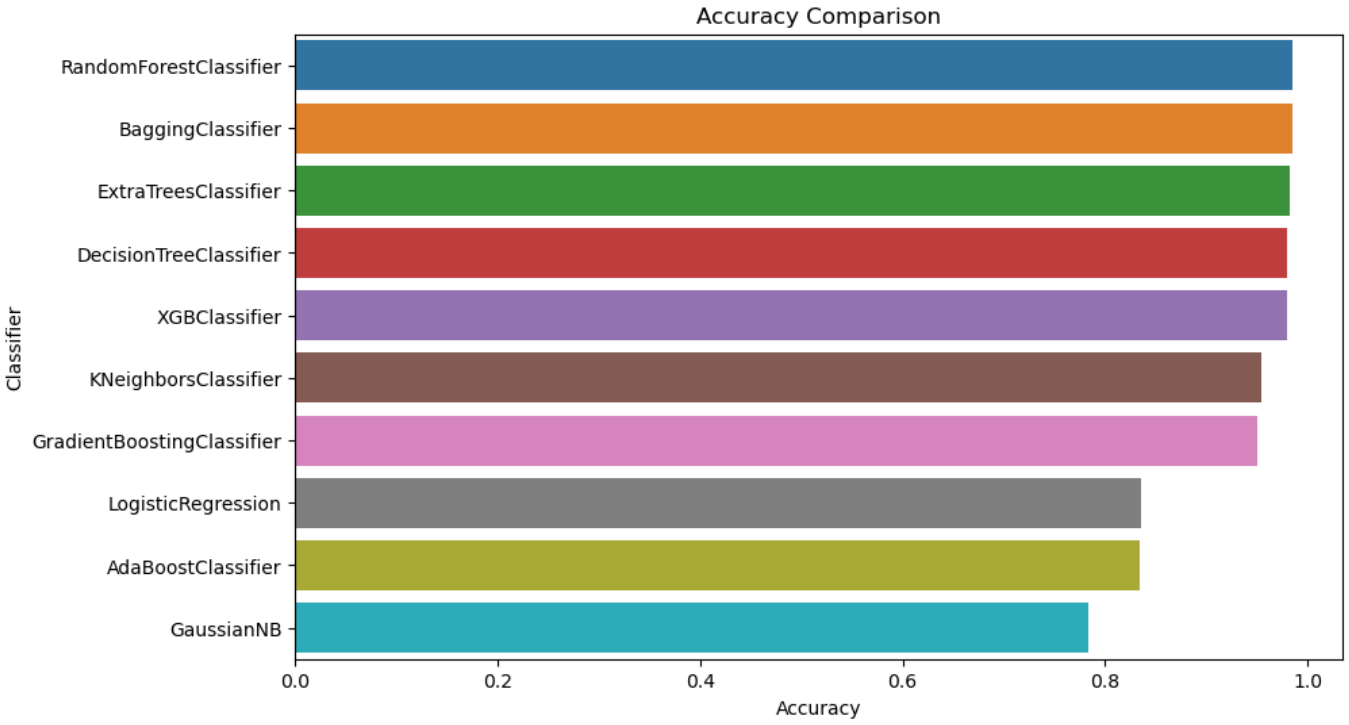


**Figure 5.** Evaluation of the accuracy performance of different classification algorithms in their default configurations.

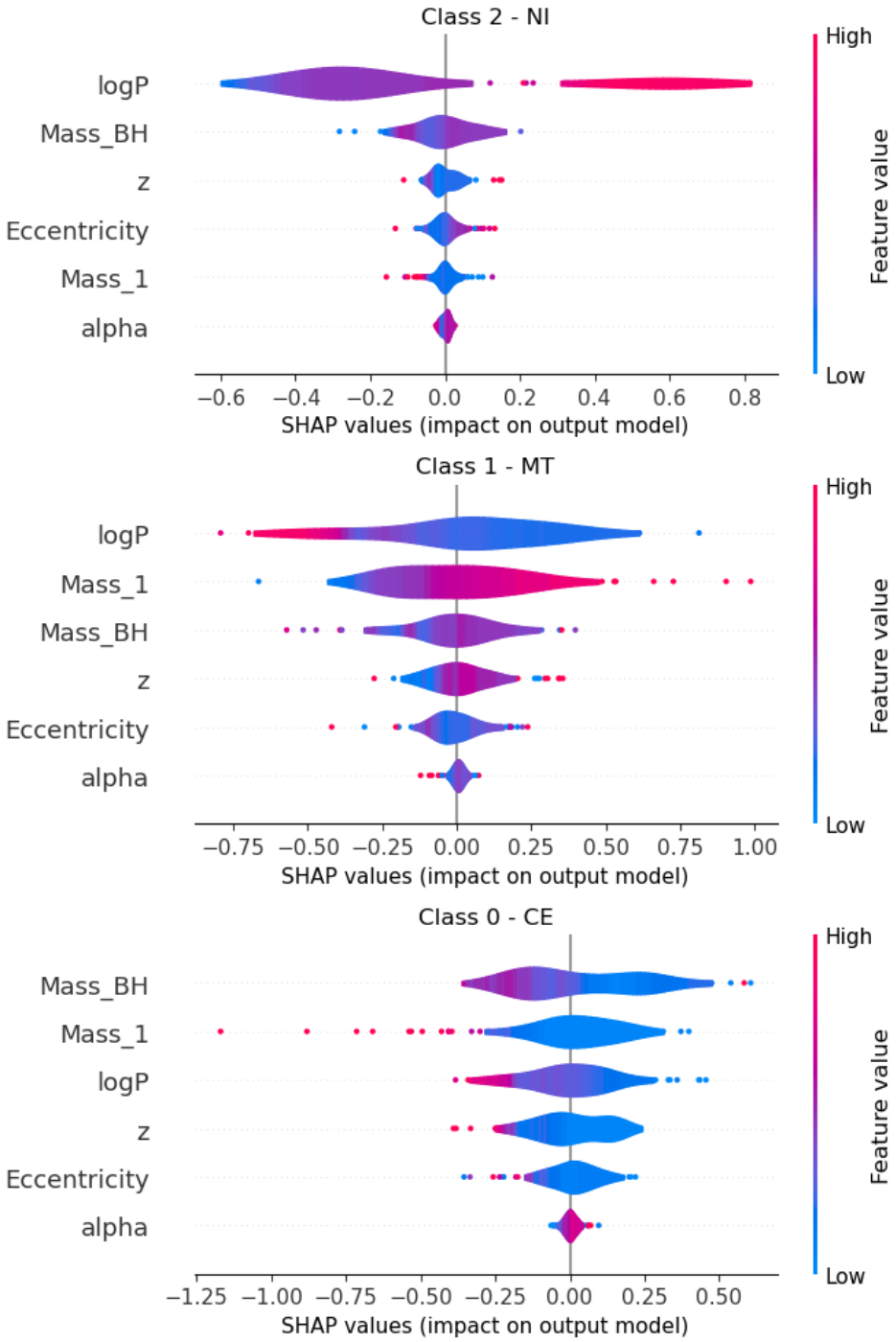


**Figure 6.** Plot of the SHAP values for each label, referring to the DNN model. Each violin-shaped distribution represents the density of the SHAP values for a specific feature. The width of the violin at a given point represents the density or frequency of SHAP values at that point.

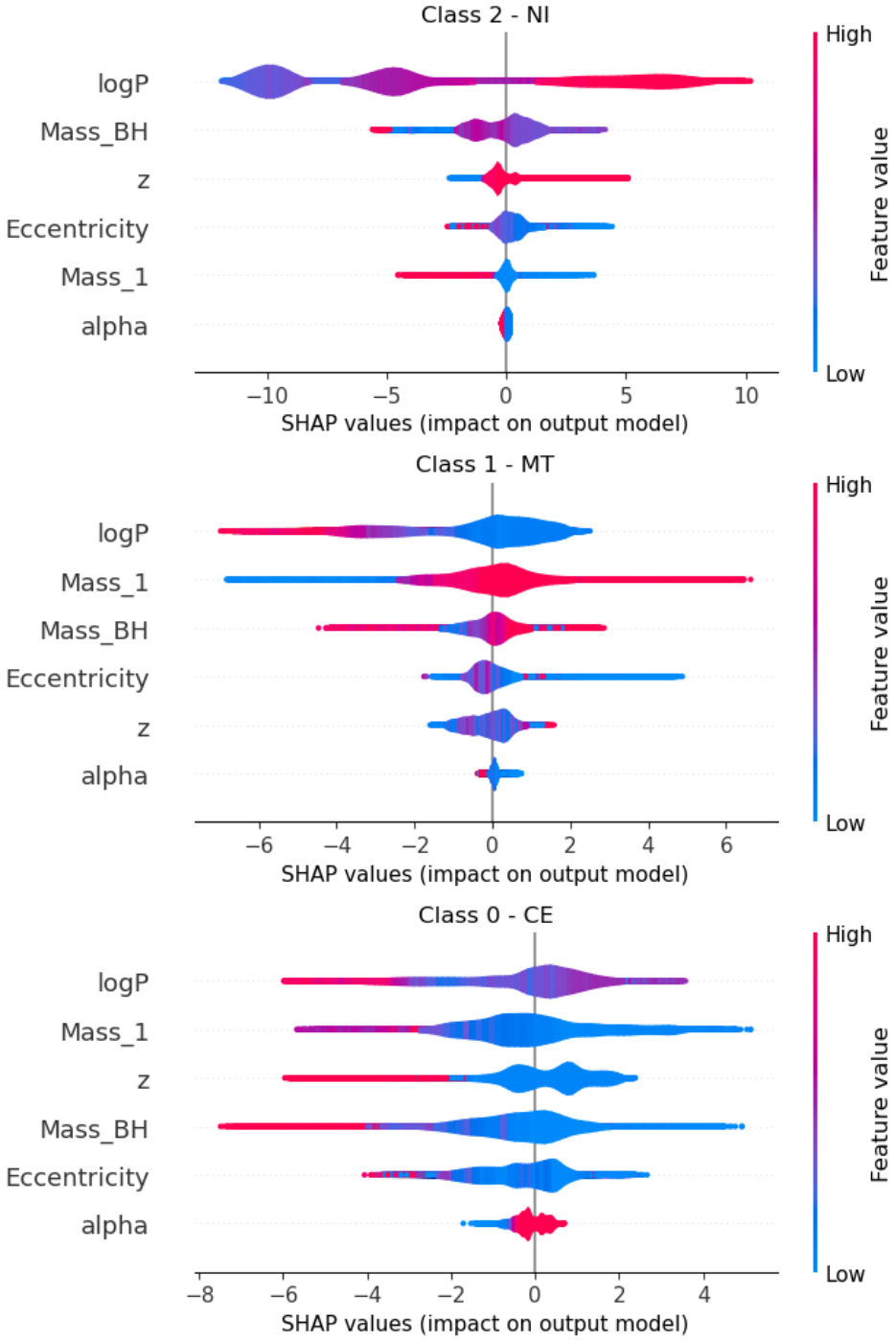


**Figure 7.** Plot of the SHAP values for each label, referring to the XGBoost model. Each violin-shaped distribution represents the density of the SHAP values for a specific feature. The width of the violin at a given point represents the density or frequency of SHAP values at that point.

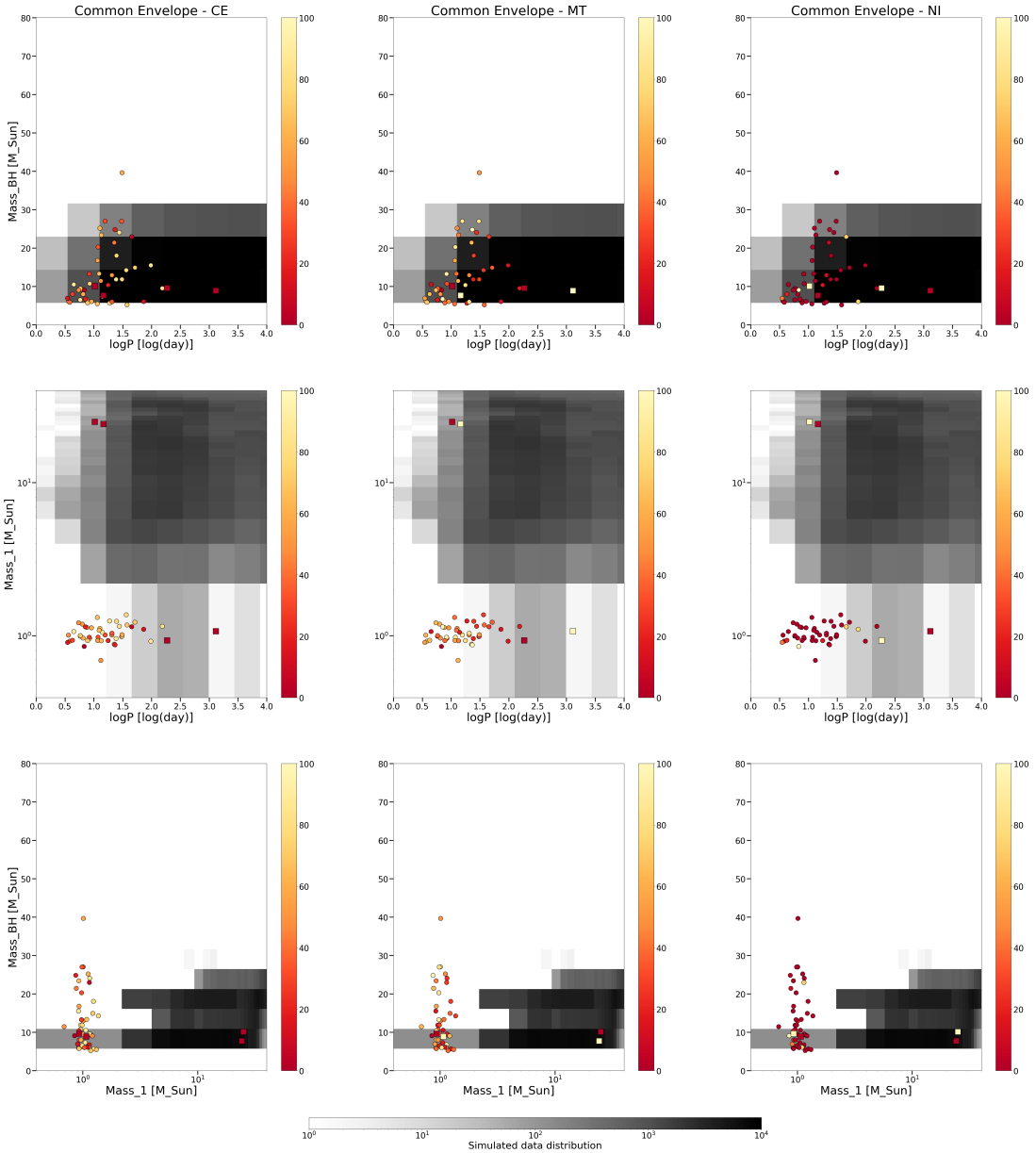


**Figure 8.** Comparison of the distribution of simulated data from SEVN, used for model training, with the candidate binary systems. The analysis is restricted to candidate systems categorized as "gold" and simulated data with $z$=0.02. The colour of each point refers to the percentage of that specific label given by the classification model (in this case the DNN).