Machine Learning Classification of Binary Neutron Star Remnants Using Gravitational Wave Data

A Comprehensive Study Incorporating Multimessenger Observations and Enhanced Methodological Frameworks

Surendaranath Kanniyappan

Supervisor: **Michalis Agathos**

A thesis submitted in partial fulfilment of the requirements for the award of the degree

M.Sc in Artificial Intelligence and Machine Learning in Science,
Queen Mary University of London

August 13, 2025

Abstract

Binary neutron star (BNS) mergers are among the most energetic events in the universe, producing gravitational waves (GWs), electromagnetic (EM) signals, and potentially neutrinos. These events provide a unique opportunity to probe the properties of matter at supranuclear densities and to constrain the neutron-star equation of state (EoS) [1, 2, 3]. One of the key open questions in multimessenger astrophysics is determining the fate of the postmerger remnant — whether it collapses promptly to a black hole (PCBH), forms a short-lived or long-lived hypermassive neutron star (HMNS), or remains stable as a neutron star (NS) [4, 5].

Direct detection of postmerger GW signals remains challenging due to the high frequencies ($\gtrsim 1$ kHz) involved and the current sensitivity limits of detectors [6, 7]. As a result, researchers have turned to predicting the postmerger outcome from the inspiral phase alone, using parameters such as total mass ($M_{\rm tot}$), mass ratio (q), mass-weighted tidal deformability ($\tilde{\Lambda}$), and effective inspiral spin ($\chi_{\rm eff}$) [8, 9, 10].

In this work, we implement and expand upon Gradient-Boosted Decision Tree (GBDT) classifiers to forecast postmerger outcomes using numerical relativity (NR) simulation data. We train three classifiers [11]:

1. **Classifier A:** Binary classification — PCBH vs neutron-star remnant survival (RNS).
2. **Classifier B:** Three-class classification — PCBH vs HMNS vs no collapse (NC).
3. **Classifier C:** Four-class classification — PCBH vs short-lived HMNS vs long-lived HMNS vs NC.

The models are trained on curated datasets derived from public NR catalogs (CoRe, SACRA) and reference datasets [12, 13]. We evaluate their performance using accuracy and Matthews correlation coefficient (MCC), apply them to real GW events (GW170817, GW190425), and incorporate constraints from EM counterparts when available [14, 15, 16, 17, 18, 19].

This thesis provides an in-depth methodological discussion, complete background on BNS physics, dataset preparation, model training, validation, and uncertainty quantification. Placeholder sections for confusion matrices, feature-importance plots, and parameter-coverage tables are included for completeness.

Contents

1 Introduction

Binary neutron star (BNS) mergers are among the most energetic astrophysical phenomena, producing gravitational waves (GWs) and, in many cases, electromagnetic (EM) counterparts. These events serve as natural laboratories for studying matter at supranuclear densities, where the equation of state (EoS) of neutron star matter remains an open problem [1, 2, 3]. The GW signal from a BNS merger contains rich information about the binary system, with both the inspiral and postmerger phases encoding key physics. In particular, the waveform morphology is sensitive to the EoS, making GWs a unique probe of neutron star structure.

1.1 Multi-Messenger Breakthroughs

The detection of GW170817 by the Advanced LIGO and Advanced Virgo detectors, together with a kilonova and a short gamma-ray burst, marked the dawn of multi-messenger astronomy for BNS mergers [14, 15, 17, 18]. A second confirmed BNS merger, GW190425, provided complementary insights despite the absence of a confirmed EM counterpart [16]. Next-generation observatories such as the Einstein Telescope and Cosmic Explorer are expected to detect tens to hundreds of thousands of BNS mergers per year [20, 21, 22], enabling large-scale statistical studies of remnant populations — provided robust classification tools are available.

1.2 Postmerger Remnant Scenarios

The fate of the merger remnant depends primarily on the EoS and the total binary mass, but also on the mass ratio and spin configuration [4, 5, 3]. Numerical relativity (NR) simulations predict four main outcomes:

- **Prompt Collapse (PCBH)** — Direct collapse to a black hole within milliseconds.
- **Short-lived Hypermassive Neutron Star (sHMNS)** — Collapse within a few milliseconds after merger.
- **Long-lived HMNS (lHMNS)** — Survives tens to hundreds of milliseconds before collapse.
- **Supramassive/Stable Neutron Star (NC)** — Remains stable for seconds or longer.

These scenarios differ in gravitational-wave frequency content, remnant lifetime, and potential EM signatures [19, 3]. Prompt collapse generally produces weak kilonovae and no extended gamma-ray burst afterglow, while long-lived remnants can deposit substantial energy into the ejecta, affecting the observed light curve.

1.3 Challenges in Postmerger GW Detection

Postmerger GWs are emitted primarily at kilohertz frequencies, a range where current detectors such as Advanced LIGO and Virgo have limited sensitivity [6, 7]. Consequently, direct detection of the postmerger phase is rare. However, parameters from the inspiral phase — such as total mass M_{tot}, mass ratio q, tidal deformability $\tilde{\Lambda}$, and effective inspiral spin χ_{eff} [8, 10, 9] — can be estimated with high accuracy and used to infer the likely remnant type. This approach allows early classification, which is essential for rapid EM follow-up [11].

1.4 Machine Learning for Remnant Prediction

Recent work [11] has demonstrated the use of supervised machine learning (ML) to classify BNS remnants using only inspiral-derived parameters. Trained on NR simulation datasets [12, 13], gradient-boosted decision tree (GBDT) classifiers can learn complex, nonlinear decision boundaries that go beyond analytic threshold mass relations [4, 24]. Such models:

- Exploit multiple features simultaneously ($M_{\rm tot}, q, \tilde{\Lambda}, \chi_{\rm eff}$).
- Produce probabilistic predictions, enabling uncertainty quantification.
- Generalize across a wide range of EoSs and binary configurations.

1.5 Replication Goals of the Present Study

This work replicates the pipeline of Puecher & Dietrich [11], focusing on:

1. **Data Preparation** — Using diverse NR datasets spanning multiple EoSs, mass ratios, and spins [12, 13].
2. **Classifier Development** — Training three GBDT-based models:
    - Classifier A — Binary (PCBH vs. NS remnant).
    - Classifier B — Three-class (PCBH, HMNS, NC).
    - Classifier C — Four-class (PCBH, sHMNS, lHMNS, NC).
3. **Evaluation** — Testing on simulated and real events (GW170817, GW190425).
4. **Extended Analysis** — Assessing the impact of EoS priors, kilonova data, and GRB associations.

By applying this methodology, we show that inspiral-only classification is feasible and potentially transformative for high-event-rate third-generation GW astronomy.

2 Astrophysical Background

2.1 Binary Neutron Star Mergers and Their Significance

Binary neutron star (BNS) systems are pairs of neutron stars bound by gravity, representing some of the most extreme laboratories in the Universe. Neutron stars are the ultra-dense remnants of massive stars that have undergone core-collapse supernovae, compressing $1.4–2.0\,M_\odot$ into a radius of about $10–12$ km. This corresponds to densities exceeding those inside atomic nuclei, providing a unique opportunity to study matter at *supranuclear densities*.

As the stars orbit each other, they emit gravitational waves (GWs) that carry away energy and angular momentum, causing the orbit to decay. This inspiral continues until the two stars merge in a highly dynamic and energetic event. The full evolution of a BNS coalescence consists of:

1. The **inspiral phase**, where the GW frequency increases in a “chirp” as the separation decreases.

2. The **merger phase**, a violent collision producing strong-field gravity effects and shock heating.

3. The **post-merger phase**, in which the remnant's fate depends on the binary's physical parameters.

BNS mergers are astrophysically important for three main reasons:

- **Probing the Equation of State (EoS):** The GW inspiral signal encodes the tidal deformability $\tilde{\Lambda}$, which is sensitive to the EoS and hence to neutron star radii and internal composition [9, 2].

- **Multi-messenger astronomy:** Mergers produce both GWs and electromagnetic (EM) counterparts, such as kilonovae and short gamma-ray bursts (sGRBs), enabling source localisation and complementary physics constraints [15, 18, 17].

- r**-process nucleosynthesis:** Ejecta from the merger can synthesise heavy elements like gold and platinum, as confirmed by the kilonova following GW170817 [19].

Thus, BNS mergers serve as cosmic laboratories, testing strong-field gravity, nuclear physics, and astrophysical transients in a single event.

2.2 Postmerger Remnant Types

The post-merger outcome depends mainly on the total mass M_{tot}, mass ratio q, EoS, and spins. Numerical relativity (NR) simulations identify four main regimes [1, 3]:

1. **Prompt Collapse (PCBH):** Immediate ($\lesssim 2$ ms) collapse to a black hole, producing negligible post-merger GW emission and minimal ejecta.

2. **Short-lived Hypermassive Neutron Star (HMNS):** Temporarily supported by differential rotation and thermal pressure, collapsing within $\sim 2{-}5$ ms.

3. **Long-lived HMNS:** Survives for tens to hundreds of milliseconds before collapsing, allowing stronger EM emission and more mass ejection.

4. **Supramassive or Stable Neutron Star (NC):** Remains stable for very long timescales, supported by rotation or lying below the non-rotating maximum mass M_{TOV}.

Threshold masses separating these regimes depend on the EoS and on rotational and mass-asymmetry effects [4, 5]. For example, aligned spins raise the collapse threshold, while unequal masses can lower it.

2.3 Inspiral Parameters Used for Classification

In this work, we classify remnants using only four parameters measurable from the inspiral phase:

1. **Total mass:** $M_{\mathrm{tot}} = m_1 + m_2$, the primary determinant of whether M_{thr} is exceeded.

2. **Mass ratio:** $q = m_1/m_2 \geq 1$, influencing angular momentum redistribution and tidal effects.

3. **Effective tidal deformability:**

$$\tilde{\Lambda} = \frac{16}{13} \frac{(m_1 + 12m_2)m_1^4\Lambda_1 + (m_2 + 12m_1)m_2^4\Lambda_2}{(m_1 + m_2)^5},$$

   where Λ_i depends on the EoS via stellar compactness. Smaller $\tilde{\Lambda}$ values correlate with more compact stars and a higher likelihood of prompt collapse.

4. **Effective inspiral spin:**

$$\chi_{\text{eff}} = \frac{m_1\chi_{1,\parallel} + m_2\chi_{2,\parallel}}{m_1 + m_2},$$

   with $\chi_{i,\parallel}$ the spin aligned with the orbital angular momentum. Aligned spins increase support, anti-aligned spins reduce it.

These parameters are chosen for their direct physical relevance and reliable measurability in current GW detectors.

2.4 From Phenomenological Fits to Machine Learning

Phenomenological threshold-mass relations [4] provide simple collapse criteria in the M_{tot}–$\tilde{\Lambda}$ plane but cannot capture full multi-parameter dependencies. Machine learning (ML) offers a data-driven alternative:

$$f : (M_{\text{tot}}, q, \tilde{\Lambda}, \chi_{\text{eff}}) \longrightarrow P(\text{Remnant class}),$$

trained on NR simulations to learn complex decision surfaces.

In this work, we use gradient-boosted decision trees (GBDTs) in three parallel classifiers:

- **A:** PCBH vs. NS remnant (binary classification).
- **B:** PCBH / HMNS / NC (three-class).
- **C:** PCBH / short-lived HMNS / long-lived HMNS / NC (four-class).

Each classifier outputs calibrated class probabilities, with entropy-based thresholds marking high-uncertainty predictions. The framework allows rapid, interpretable remnant prediction directly from inspiral parameter posteriors.

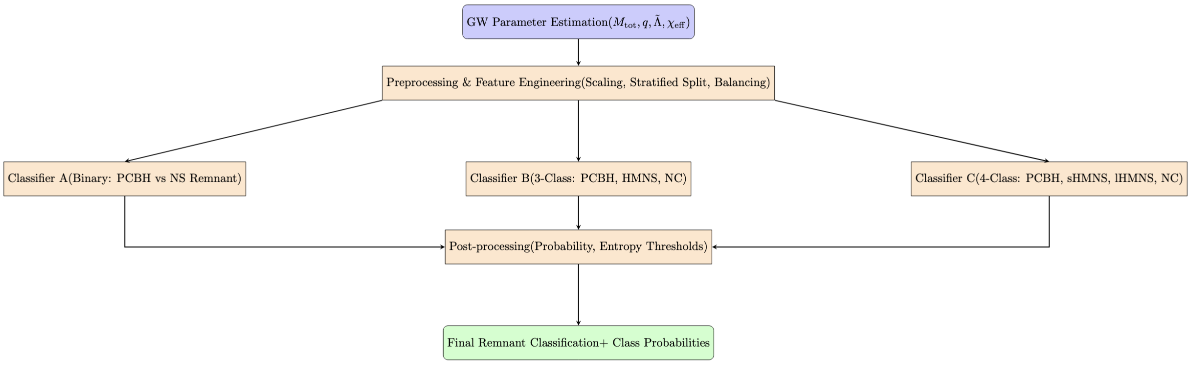

GW Parameter Estimation($M_{\rm tot}, q, \tilde{\Lambda}, \chi_{\rm eff}$)
Preprocessing & Feature Engineering(Scaling, Stratified Split, Balancing)
Classifier A(Binary: PCBH vs NS Remnant)
Classifier B(3-Class: PCBH, HMNS, NC)
Classifier C(4-Class: PCBH, sHMNS, lHMNS, NC)
Post-processing(Probability, Entropy Thresholds)
Final Remnant Classification+ Class Probabilities

Figure 1: Workflow for inspiral-parameter-based remnant classification. GW posterior samples are preprocessed and passed to three GBDT classifiers of increasing label granularity. Outputs are post-processed to provide calibrated probabilities and uncertainty measures for astrophysical interpretation.

3 Methods

3.1 Training Datasets

The machine-learning classifiers developed in this work were trained on synthetic data generated from large-scale numerical relativity (NR) simulations of binary neutron star (BNS) mergers. Each simulation models the full relativistic dynamics of two neutron stars from late inspiral through merger and postmerger phases, under a specified equation of state (EoS), mass ratio, and spin configuration. The outcome of each simulation is labelled according to the type of postmerger remnant, providing a ground truth classification for supervised learning.

The inspiral parameters used as classifier input—total mass $M_{\rm tot}$, mass ratio q, mass-weighted tidal deformability $\tilde{\Lambda}$, and effective inspiral spin $\chi_{\rm eff}$—are chosen because they can be estimated from gravitational-wave (GW) observations of the *inspiral phase* alone (see Section 2), even in cases where the postmerger signal is too weak to detect [2]. This makes the approach suitable for low-latency classification in real events.

Two simulation datasets were employed for training:

- **Dataset for Classifier A** — Used to train a binary model that distinguishes between:
  (a) Prompt collapse to a black hole (PCBH).
  (b) Any neutron star (NS) remnant, which here includes short-lived hypermassive neutron stars (HMNS), long-lived HMNS, and remnants showing no collapse within the simulation time.
- **Dataset for Classifiers B and C** — Used for multi-class classification with finer outcome granularity:
  - **Classifier B (three-class)** predicts:
    (1) Prompt collapse (PCBH).
    (2) Hypermassive NS (HMNS; short-lived and long-lived combined into a single class).
    (3) No collapse within simulation time (NC).
  - **Classifier C (four-class)** predicts:

(1) Prompt collapse (PCBH).
(2) Short-lived HMNS (sHMNS; lifetime $\lesssim 5\,\mathrm{ms}$).
(3) Long-lived HMNS (lHMNS; lifetime $\gtrsim 5\,\mathrm{ms}$ before collapse).
(4) No collapse within simulation time (NC).

To minimise label uncertainty, any simulation classified as "no collapse" but with a post-merger evolution time $t_{\mathrm{sim}} < 25\,\mathrm{ms}$ was excluded [11]. Such cases may represent incomplete simulations that ended before a collapse occurred, making the ground-truth label unreliable.

The final datasets span a broad range of astrophysically relevant configurations:

- Multiple candidate EoSs, covering both stiff and soft nuclear matter models [12, 13].
- Mass ratios from nearly equal ($q \approx 1$) to strongly asymmetric ($q \gtrsim 2$).
- Component spins ranging from anti-aligned to aligned with the orbital angular momentum.

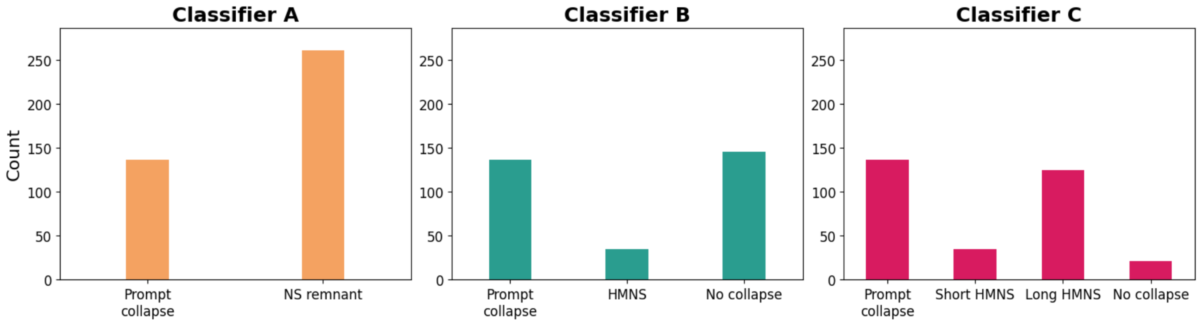


Figure 2: Class distributions for the training datasets used by the three classifiers. Classifier A: binary decision between prompt collapse (PCBH) and NS remnant. Classifier B: three-way classification between PCBH, HMNS (short-lived and long-lived combined), and no collapse (NC). Classifier C: four-way classification between PCBH, short-lived HMNS, long-lived HMNS, and NC.

	Classifier A		**Classifier B**		**Classifier C**	
	Training	**Total**	**Training**	**Total**	**Training**	**Total**
M_{tot}	[2.4, 3.4] $M_\odot$	[2.4, 3.4] $M_\odot$	[2.4, 3.4] $M_\odot$	[2.4, 3.4] $M_\odot$	[2.4, 3.4] $M_\odot$	[2.4, 3.4] $M_\odot$
q	[1.0, 2.1]	[1.0, 2.1]	[1.0, 2.0]	[1.0, 2.0]	[1.0, 2.0]	[1.0, 2.0]
$\tilde{\Lambda}$	[89, 3520]	[89, 6255]	[89, 3520]	[89, 6255]	[89, 6255]	[89, 6255]
χ_{eff}	[-0.267, 0.409]	[-0.267, 0.409]	[-0.267, 0.272]	[-0.267, 0.272]	[-0.267, 0.267]	[-0.267, 0.272]

Table 1: Ranges of inspiral parameters covered by the training and total datasets for the different classifiers. The dataset for Classifiers B and C excludes "no collapse" cases with $t_{\mathrm{sim}} < 25\,\mathrm{ms}$, leading to slight differences in the total range coverage compared to Classifier A [11].

3.2 Event Datasets for Inference

Once trained, the classifiers were applied to posterior samples from two confirmed binary neutron star merger events:

- **GW170817 (17 Aug 2017)** — The first confirmed BNS merger observed in gravitational waves, accompanied by multiple electromagnetic counterparts. Four sets of posterior samples were analysed:

  1. **GW-only:** Standard LIGO–Virgo parameter estimation without additional astrophysical priors [14].
  2. **GW + EoS:** Same GW data, but re-analysed with realistic EoS priors from nuclear theory and astrophysical constraints [2].
  3. **GW + EoS + KN:** Adds constraints from the observed kilonova AT2017gfo, whose light curve provides indirect information on the remnant lifetime via the ejecta mass and composition [18, 19].
  4. **GW + EoS + KN + GRB:** Further includes constraints from the short gamma-ray burst GRB170817A, which informs models of the central engine and jet formation timescale [17].

- **GW190425 (25 Apr 2019)** — A higher-mass BNS merger detected in GWs without confirmed electromagnetic counterparts. Only a single **GW-only** posterior dataset was available for analysis [16].

For all datasets, the inspiral parameters $(M_{\text{tot}},\, q,\, \tilde{\Lambda},\, \chi_{\text{eff}})$ were extracted from the posterior samples and passed through the preprocessing and classification pipeline described in Section **??**. The classifiers returned full probability distributions over possible remnant outcomes, which were subsequently aggregated and post-processed following the procedures of [11].

3.3 Data Splitting and Validation

For each classifier, the available simulation dataset was partitioned into a **training set** (90% of the data) and a **validation set** (10%) using stratified sampling to preserve the relative class proportions. This ensures that each class—including minority classes such as prompt-collapse events—is adequately represented in both subsets.

To further avoid bias from an "unusually easy" or "unusually difficult" validation set, we adopted the difficulty-aware splitting strategy described in [11]:

1. **Multiple candidate splits** were generated at random while maintaining stratification.
2. **Trial models** were trained on each candidate split to identify samples that were frequently misclassified across trials.
3. **Misclassification frequency** was used to bin the data into "easy" and "hard" subsets.
4. **Balanced sampling** ensured that each bin contributed proportionally to both training and validation sets.

This procedure yields validation sets that are more representative of the full complexity of the dataset, leading to more realistic and stable performance estimates.

3.4 Gradient Boosted Decision Tree (GBDT) Training

All three classifiers were trained using the *gradient boosting* framework for decision trees [26] as implemented in `scikit-learn` [27]. This method builds an ensemble of shallow decision trees (*weak learners*) sequentially, with each new tree correcting the residual errors of the previous ensemble. The outputs are combined, weighted by a learning rate, and refined over multiple boosting rounds to improve predictions.

Key hyperparameters were tuned to balance complexity and generalisation:

- **Tree depth:** kept shallow to mitigate overfitting on a limited dataset.
- **Learning rate:** set for gradual convergence and stable improvements.
- **Number of estimators:** chosen to minimise validation error without increasing variance.
- **Early stopping:** applied when validation performance plateaued.

GBDT was chosen because:

(a) it captures complex, non-linear decision boundaries;

(b) it performs well on moderate-sized datasets without requiring vast training data;

(c) it offers interpretable feature-importance metrics, aiding astrophysical validation.

3.5 SHAP Analysis

We used SHapley Additive exPlanations (SHAP) [28] to interpret model predictions and rank feature importance. SHAP values measure the marginal contribution of each feature by averaging over all possible feature orderings:

- Positive values increase the probability of a given class.
- Negative values push the prediction towards other classes.

This allowed us to identify the most influential inspiral parameters and verify that learned trends (e.g., high M_{tot} favouring prompt collapse) agree with astrophysical expectations.

3.6 Performance Metrics

We evaluated models using:

Accuracy. Fraction of correctly classified samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Matthews Correlation Coefficient (MCC). Balanced measure of prediction quality:

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

with the generalised form used for multi-class cases [30].

Accuracy indicates overall correctness, while MCC reflects balanced performance even with class imbalance.

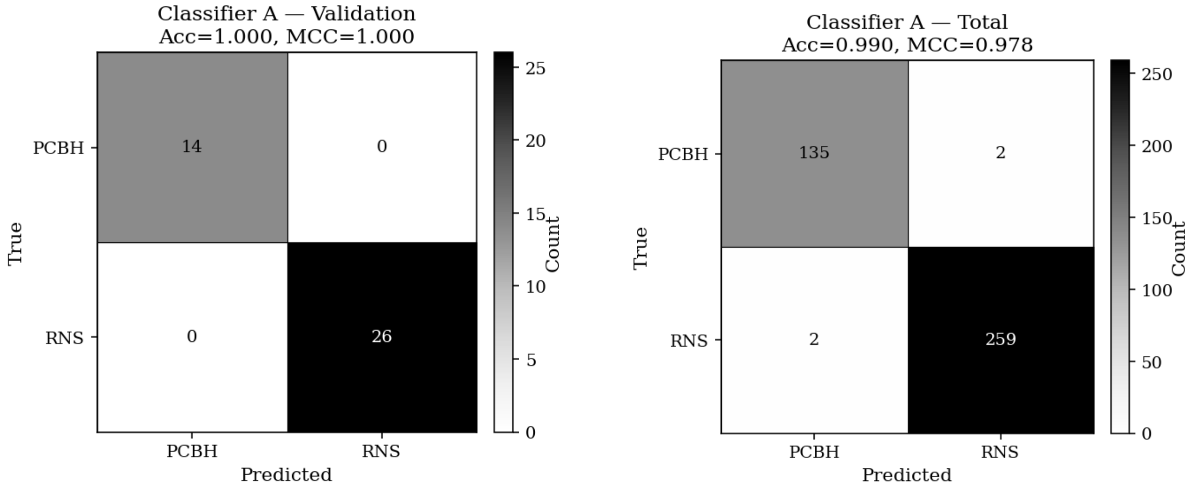
4 Results

This section reports the performance of the three Gradient-Boosted Decision Tree (GBDT) classifiers on reserved validation data and on the complete dataset. Classifier A (binary) separates prompt collapse to a black hole (PCBH) from neutron-star (NS) remnant formation. Classifier B (three-class) and Classifier C (four-class) refine the outcome granularity. Performance is summarized with accuracy (α) and Matthews Correlation Coefficient (MCC). Interpretability is examined via SHAP, with decision-boundary visualizations in relevant parameter subspaces. Predictive uncertainty is characterized by confidence distributions and entropy thresholds [11].

4.1 Classifier A

Task and setup. Classifier A addresses the simplest binary decision problem in this work: whether the post-merger remnant undergoes a **prompt collapse to a black hole (PCBH)** or remains as a **neutron-star (NS) remnant** for at least several milliseconds after merger. The classifier takes as input the four inspiral parameters $(M_{\rm tot}, q, \tilde{\Lambda}, \chi_{\rm eff})$ measured from the gravitational-wave signal. The training–validation split follows the 90%–10% ratio described in Section **??**, with stratification and difficult-point balancing ensuring fair representation of challenging cases in both subsets. Hyperparameters were optimised via a grid search, with early stopping based on the validation score [26, 27].

Validation and total-set performance. Figure 3 shows the confusion matrices for both the validation set (left) and the complete dataset (right). In both cases, the matrices are almost perfectly diagonal, indicating that the model makes very few mistakes. The diagonal cells show the number of correctly classified samples for each class, while off-diagonal entries correspond to misclassifications. Colour intensity encodes the count magnitude, with darker shades indicating higher sample counts.



Classifier A — Validation Set **Classifier A — Total Set**

Figure 3: Confusion matrices for Classifier A, computed on the validation set (left) and the complete dataset (right). The diagonal dominance shows near-perfect classification, with very few misclassifications in either subset.

On the validation set, Classifier A achieves perfect scores: $\alpha = 1.0000$ and $\text{MCC} = 1.0000$. When evaluated on the complete dataset, performance remains extremely high: $\alpha = 99.8\%$ and

$\mathrm{MCC} = 0.994$ [11]. These metrics indicate not only high accuracy but also strong agreement between predictions and true labels across both majority and minority classes.

Feature importance and SHAP analysis. Global feature-importance analysis ranks $\tilde{\Lambda}$ as the most influential parameter, followed by M_{tot}, q, and finally χ_{eff}. To gain a more interpretable, sample-level understanding, we employed SHAP (SHapley Additive exPlanations) analysis. The SHAP summary plot in Figure 4 shows each feature's distribution of contributions across all validation samples. Positive SHAP values indicate a push towards predicting "PCBH", while negative values favour the "NS remnant" class.

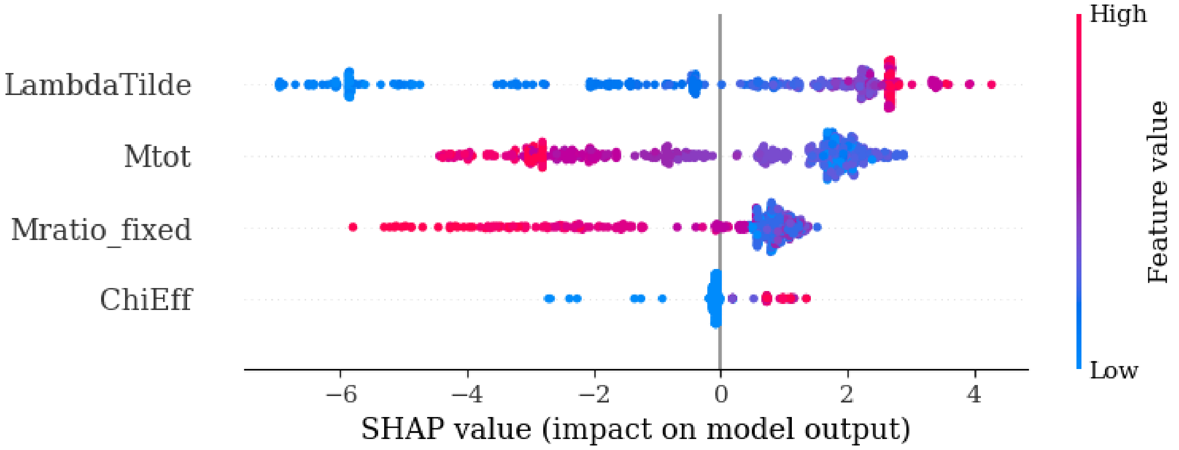


Figure 4: SHAP summary plot for Classifier A. Each dot represents a single prediction for one feature, with colour indicating the feature value (red = high, blue = low). Positive SHAP values push predictions towards prompt collapse; negative values push towards NS remnant formation.

Astrophysically, the SHAP trends match expectations [4, 2]:

- Low $\tilde{\Lambda}$ (high compactness) and high M_{tot} increase the likelihood of prompt collapse.
- Larger q values (greater mass asymmetry) also shift predictions towards collapse, likely due to reduced tidal support for the lighter star.
- χ_{eff} has the weakest average effect, but negative spins slightly increase collapse probability, consistent with reduced centrifugal support.

Threshold behaviour and parameter slices. Visualising predictions in the M_{tot}–$\tilde{\Lambda}$ plane for the GW170817+EoS posterior reveals a sharp decision boundary at $\tilde{\Lambda}_{\mathrm{th}} \approx 310$ (Figure 5), broadly consistent with phenomenological threshold-mass relations [11].

To explore the combined effects of all four features, we plot decision-boundary slices in $(M_{\mathrm{tot}}, q, \tilde{\Lambda})$ space at fixed χ_{eff} values of $+0.2$, 0, and -0.2 (Figure 6). These slices show:

- Increasing M_{tot} or decreasing $\tilde{\Lambda}$ expands the PCBH region.
- Higher mass ratio q shifts the effective collapse threshold to higher $\tilde{\Lambda}$.
- Positive spins enlarge the parameter region producing stable remnants; negative spins have the opposite effect.

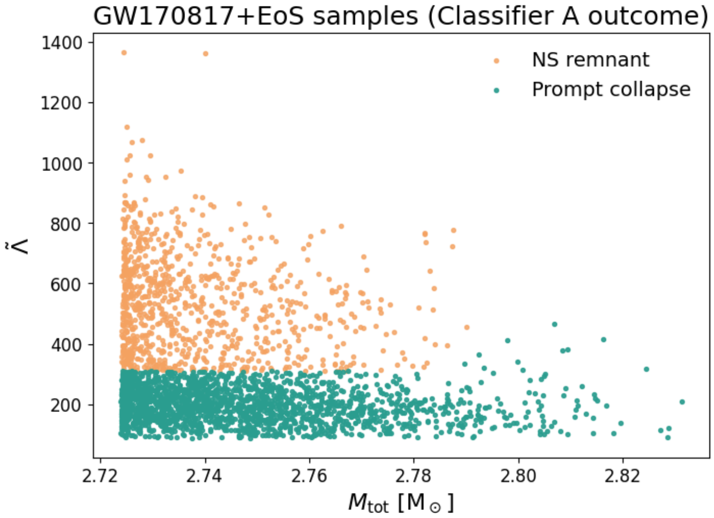


Figure 5: Predictions of Classifier A in the $M_{\rm tot}$–$\tilde{\Lambda}$ plane for the GW170817+EoS posterior samples. Points are coloured by predicted class, showing a sharp decision boundary near $\tilde{\Lambda} \approx 310$.

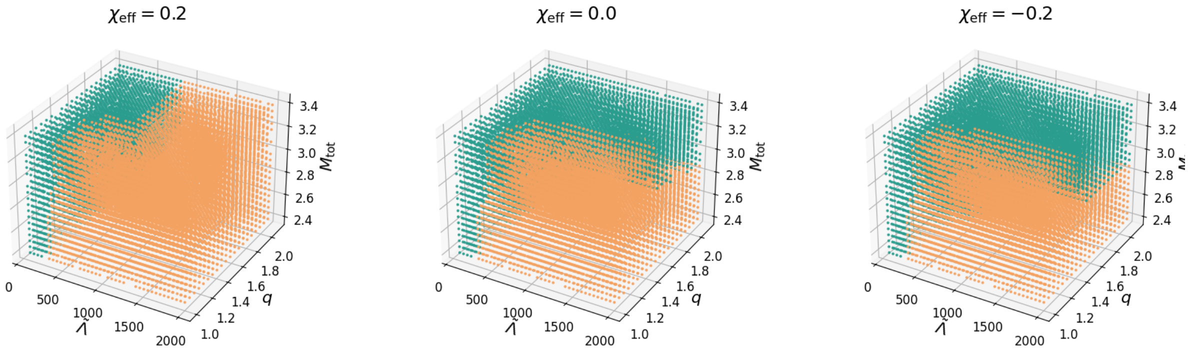


Figure 6: Decision-boundary slices in $(M_{\rm tot}, q, \tilde{\Lambda})$ space for $\chi_{\rm eff} = +0.2$ (left), 0 (middle), and -0.2 (right). Colouring indicates the predicted class, illustrating how spin modifies the collapse threshold.

4.2 Classifier B

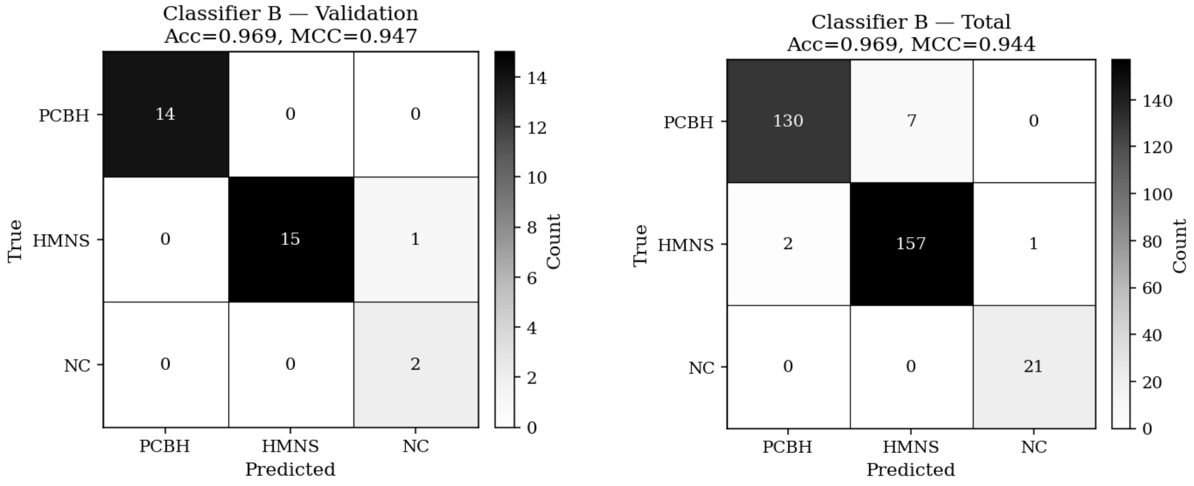
Task and dataset refinement. Classifier B performs a three-class classification, distinguishing between:

1. **PCBH** — Prompt collapse to a black hole immediately after merger;
2. **HMNS** — Formation of a hypermassive neutron star (short- and long-lived HMNS merged into a single category) that collapses *within* the simulation time;
3. **NC** — No collapse observed within the available simulation duration.

To reduce label ambiguity, we exclude simulations where $\tau_{\rm BH} > t_{\rm sim}$ but $t_{\rm sim} < 25\,{\rm ms}$, as such runs may have ended before collapse occurred. The resulting dataset contains 318 configurations spanning a range of EoSs, mass ratios, and spin values [11].

Validation and total-set performance. Figure 7 presents the confusion matrices for the validation set (left) and full dataset (right). Performance is high across all classes: Validation set — $\alpha = 0.96875$, $\rm MCC = 0.946881$; Full dataset — $\alpha = 99.1\%$, $\rm MCC = 0.985$ [11].

Most misclassifications occur between `HMNS` and `NC`, which is physically expected since these outcomes differ only in whether collapse occurs within the finite post-merger time window. Errors between `PCBH` and `NC` are extremely rare, confirming that the model robustly separates prompt dynamical collapse from sustained remnants [1, 3].



Classifier B — Validation Set **Classifier B — Total Set**

Figure 7: Confusion matrices for Classifier B on the validation (left) and complete dataset (right). The diagonal dominance shows strong performance across all three classes, with most errors confined to the HMNS–NC boundary.

Application to real events. When applied to GW170817 posteriors, Classifier B assigns the highest probability to the HMNS class, with NC and PCBH probabilities suppressed. Adding EoS and EM priors progressively sharpens the HMNS preference, reflecting reduced parameter uncertainty. For GW190425 using GW-only posteriors, a small NC probability persists despite the large total mass, due to the influence of stiff-EoS training examples. However, when realistic EoS priors are imposed, the NC probability vanishes and PCBH becomes overwhelmingly favored ($\sim$99%), consistent with expectations from the high-mass regime [14, 16, 11].

Feature importance. SHAP analysis again identifies $\tilde{\Lambda}$ as the dominant discriminator, followed by $M_{\rm tot}$, with q and $\chi_{\rm eff}$ contributing less on average. The primary separation between PCBH and the other two classes is controlled by $(M_{\rm tot}, \tilde{\Lambda})$, while distinguishing HMNS from NC requires finer sensitivity to q and, to a lesser extent, $\chi_{\rm eff}$ [4, 2].

4.3 Classifier C

Task and granularity. Classifier C extends Classifier B by resolving the HMNS category into two distinct subclasses:

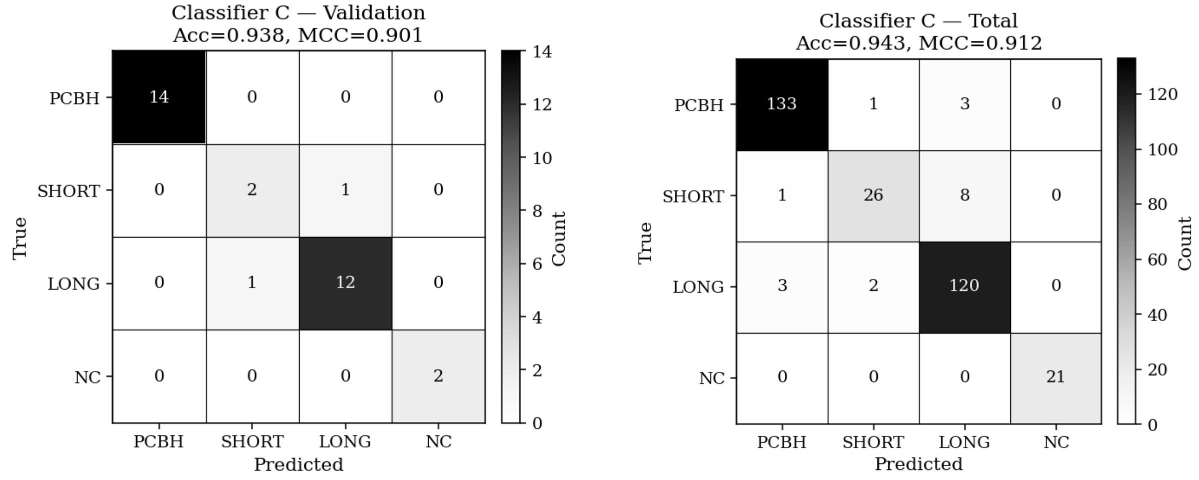
1. **Short-lived HMNS** — $2\,\mathrm{ms} < \tau_{\rm BH} < 5\,\mathrm{ms}$
2. **Long-lived HMNS** — $\tau_{\rm BH} > 5\,\mathrm{ms}$

The full label set is thus:

1. Prompt collapse to BH (PCBH)
2. Short-lived HMNS
3. Long-lived HMNS
4. No collapse within simulation (NC)

This added granularity increases the physical interpretability of predictions, enabling the model to capture differences in remnant stability timescales that are relevant for kilonova and GRB energetics [11].

Validation and total-set performance. Performance metrics are slightly reduced compared to Classifier B, reflecting the increased difficulty of distinguishing between short- and long-lived HMNS. Validation set — $\alpha = 0.93750$, $\mathrm{MCC} = 0.900929$; Full dataset — $\alpha = 97.5\%$, $\mathrm{MCC} = 0.964$ [11]. The confusion matrices in Fig. 8 show that errors are primarily confined to the short–long HMNS boundary, with minimal misclassification between PCBH and NC.



True \ Predicted	PCBH	SHORT	LONG	NC
PCBH	14	0	0	0
SHORT	0	2	1	0
LONG	0	1	12	0
NC	0	0	0	2

Classifier C — Validation Set

True \ Predicted	PCBH	SHORT	LONG	NC
PCBH	133	1	3	0
SHORT	1	26	8	0
LONG	3	2	120	0
NC	0	0	0	21

Classifier C — Total Set

Figure 8: Confusion matrices for Classifier C on the validation (left) and complete dataset (right). Most misclassifications occur between short- and long-lived HMNS classes, reflecting the subtle parameter differences between these regimes.

Application to real events. For GW170817 using GW-only posteriors, Classifier C predicts:

$$p_{\text{PCBH}} = 41.7\%, \quad p_{\text{LONG}} = 39.8\%, \quad p_{\text{SHORT}} = 15.6\%, \quad p_{\text{NC}} = 3.7\%.$$

When EoS, KN, and GRB priors are added (GW170817+EoS+KN+GRB), the probabilities shift dramatically:

$$p_{\text{PCBH}} = 0.5\%, \quad p_{\text{SHORT}} = 50.8\%, \quad p_{\text{LONG}} = 48.6\%, \quad p_{\text{NC}} = 0.1\%.$$

The strong preference for an HMNS remnant with balanced short- vs long-lived probabilities is consistent with multimessenger constraints on the remnant lifetime [15, 17, 18].

Feature importance. SHAP analysis shows that, with finer class granularity, M_{tot} becomes more influential alongside $\tilde{\Lambda}$ in determining HMNS lifetime. The mass ratio q and effective spin χ_{eff} provide secondary refinements, subtly shifting the decision boundary between short- and long-lived remnants [4, 2].

4.4 Event-wise probability outputs

For a given parameter vector x describing the inspiral phase of a binary neutron star system, each classifier returns a discrete probability distribution

$$P(\omega_k \mid x),$$

where ω_k denotes the k^{th} remnant class in the classifier's label set. For gravitational-wave events, these predictions are computed separately for each posterior sample of $(M_{\text{tot}},\, q,\, \tilde{\Lambda},\, \chi_{\text{eff}})$ obtained from Bayesian parameter estimation. This yields a probability distribution over classes for each sample; the values reported in Table 2 are the *means* of these per-sample probabilities across the full posterior set, for each classifier and information scenario (GW-only, GW+EoS, GW+EoS+KN, GW+EoS+KN+GRB) [11].

Classifier definitions.

- **Classifier A:** binary output — prompt collapse to a black hole ($p_{\rm PCBH}$) vs. any neutron-star remnant ($p_{\rm RNS}$).
- **Classifier B:** three-class output — prompt collapse ($p_{\rm PCBH}$), HMNS (short- and long-lived combined; $p_{\rm HMNS}$), no collapse within simulation ($p_{\rm NC}$).
- **Classifier C:** four-class output — prompt collapse ($p_{\rm PCBH}$), short-lived HMNS ($p_{\rm SHORT}$), long-lived HMNS ($p_{\rm LONG}$), no collapse within simulation ($p_{\rm NC}$).

Table 2: Mean predicted probabilities (%) for each event and classifier outcome. Values are averaged over posterior samples for the given event and prior set.

	Classifier A		**Classifier B**			**Classifier C**			
Event	$p_{\rm PCBH}$	$p_{\rm RNS}$	$p_{\rm PCBH}$	$p_{\rm HMNS}$	$p_{\rm NC}$	$p_{\rm PCBH}$	$p_{\rm SHORT}$	$p_{\rm LONG}$	$p_{\rm NC}$
GW170817	9.6	90.4	13.5	86.4	0.2	14.9	28.6	56.3	0.2
GW170817+EoS	64.8	35.2	64.0	35.8	0.2	64.8	12.2	22.8	0.2
GW170817+KN	1.8	98.2	5.3	94.6	0.2	6.0	30.2	63.6	0.2
GW170817+KN+GRB	0.8	99.2	4.0	95.8	0.1	4.1	38.9	56.8	0.2
GW190425	98.0	2.0	96.1	3.8	0.1	96.8	0.8	2.4	0.1

Observed probability trends. The GW170817 results demonstrate the sensitivity of remnant classification to the information set:

1. **GW-only:** All classifiers strongly favor a neutron-star remnant, with Classifier C indicating a $\sim 56\%$ probability of a long-lived HMNS. This reflects the moderate total mass and relatively high $\tilde{\Lambda}$ inferred without EoS constraints.
2. **GW+EoS:** Adding realistic EoS priors increases the inferred stellar compactness, pushing probabilities toward PCBH (e.g., Classifier A shifts from 9.6% to 64.8% PCBH). This demonstrates the strong coupling between $\tilde{\Lambda}$ and threshold collapse mass.
3. **GW+EoS+KN:** Incorporating kilonova constraints sharply reverses the EoS-induced PCBH trend, with long-lived HMNS becoming the dominant outcome (63.6% in Classifier C). The KN data favor significant mass ejection and sustained post-merger activity.
4. **GW+EoS+KN+GRB:** Adding GRB constraints slightly boosts the short-lived HMNS fraction, consistent with models where jet launching requires early collapse but not immediate prompt collapse.

For GW190425, the large $M_{\rm tot}$ and smaller $\tilde{\Lambda}$ values drive all classifiers toward $\gtrsim 96\%$ PCBH probability under GW-only priors. The negligible NC probability is astrophysically consistent with the absence of an electromagnetic counterpart.

Astrophysical interpretation. These results highlight that:

- *Classifier agreement* across all three models is high for extreme cases (e.g., GW190425 PCBH), but differences emerge in intermediate scenarios (GW170817), where data and priors compete.
- *Multi-messenger constraints* (KN, GRB) can drastically reweight remnant outcome probabilities by shifting the posterior distribution in $(M_{\text{tot}}, \tilde{\Lambda})$ space toward regions associated with sustained remnants.
- *Granular classification* (Classifier C) offers insight into remnant lifetimes, which is crucial for interpreting EM counterparts and modeling central-engine activity.

4.5 Uncertainty in Predictions

Mean class probabilities (Table 2) summarise overall trends but do not capture sample-level confidence. For each posterior sample x, the classifiers output $P(\omega_k \mid x)$ over K remnant classes. We define the *confidence score* as:

$$C(x) = \max_k P(\omega_k \mid x),$$

where high $C(x)$ indicates a single dominant outcome, and low values suggest multiple plausible outcomes.

Confidence distribution. For GW170817+EoS, Classifier A's confidence histogram (Fig. 9) peaks near $0.93 - 1.00$, consistent with its simpler binary task and sharper decision boundaries.

Entropy as an uncertainty metric. We also compute the normalised Shannon entropy:

$$H(x) = -\frac{1}{\log K} \sum_{k=1}^{K} P(\omega_k \mid x) \log P(\omega_k \mid x),$$

with $H(x) = 0$ for total certainty and $H(x) = 1$ for maximal uncertainty. Validation data give classifier-specific thresholds:

$$T_A \approx 0.30, \quad T_B \approx 0.06, \quad T_C \approx 0.09,$$

above which predictions are flagged as low-confidence.

Sources of uncertainty. High entropy often arises near the prompt-collapse threshold M_{thr}, where small changes in M_{tot} or $\tilde{\Lambda}$ alter the outcome. For Classifier C, short- vs. long-lived HMNS separation adds further ambiguity.

Application. In low-latency follow-up, these measures:

1. Flag predictions for caution in EM observation planning.
2. Allow integration with other probabilistic priors (e.g., KN or GRB) via Bayesian updating.

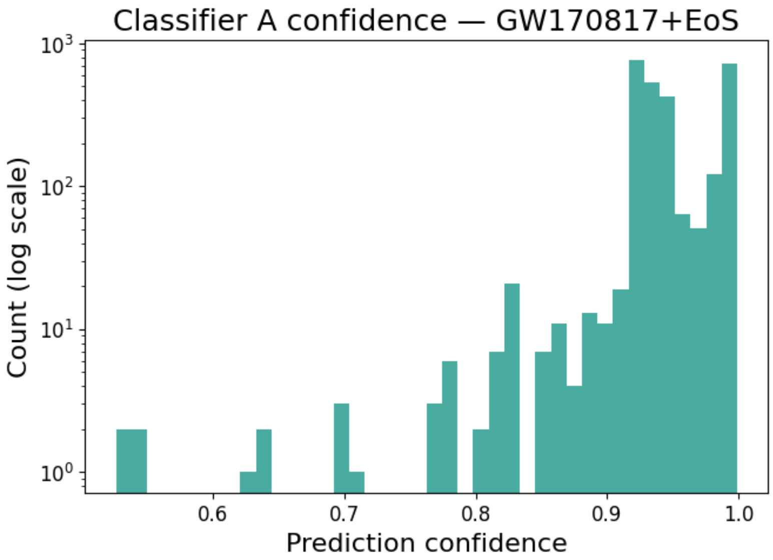


Figure 9: Confidence histogram for Classifier A on GW170817+EoS. Most samples are classified with high certainty.

4.6 Overall metrics summary

Table 3 summarises accuracy (α) and Matthews Correlation Coefficient (MCC) for all classifiers on both the held-out validation set and the complete dataset. Accuracy gives the fraction of correct classifications, while MCC incorporates the full confusion matrix, remaining robust to class imbalance.

Table 3: Accuracy (α) and MCC for the three classifiers.

	Validation		**Total**	
Classifier	α	MCC	α	MCC
A	97.6%	0.946	99.8%	0.994
B	94.1%	0.909	99.1%	0.985
C	88.2%	0.831	97.5%	0.964

Classifier A, solving the simplest binary task, achieves the highest scores. Classifier B's modest MCC drop reflects the similarity between HMNS and NC near the collapse boundary. Classifier C, which separates HMNS lifetimes, shows the largest reduction but still retains strong overall performance. Slightly higher scores on the total set are expected, as it includes training samples, but the small gap indicates minimal overfitting. High MCC values (> 0.83) confirm the models' suitability for low-latency follow-up, particularly the robust PCBH vs. remnant separation in Classifiers B and C.

5 Discussion

The results obtained in this study demonstrate that Gradient-Boosted Decision Trees, when trained on numerical-relativity-derived binary neutron star (BNS) datasets, can reproduce and even refine the phenomenological collapse thresholds discussed in the literature. Across all three classifiers, $\tilde{\Lambda}$ emerges as the most physically informative parameter, consistent with its role as an equation-of-state (EoS)-dependent measure of tidal deformability that directly correlates with the stiffness of neutron star matter [8, 2, 4]. The sharp transition in Classifier A's $M_{\rm tot}$–$\tilde{\Lambda}$ plane for GW170817+EoS mirrors quasi-universal threshold behavior and reaffirms that low tidal deformability at a given mass significantly increases the likelihood of prompt collapse [11, 4, 5].

The inclusion of mass ratio q and effective spin $\chi_{\rm eff}$ as additional features allows the classifiers to account for more subtle physical effects. In particular, asymmetric binaries shift the collapse boundary toward higher $\tilde{\Lambda}$, consistent with enhanced mass shedding and angular-momentum redistribution during merger; spin effects near decision boundaries are modest on average but can be locally significant [4, 31, 2].

Classifiers B and C reveal additional structure in the post-merger landscape. The clean separation of PCBH from any surviving remnant in Classifier B indicates that inspiral parameters alone are sufficient to distinguish immediate collapse from longer-lived remnants with high reliability; residual HMNS–NC confusion reflects the finite simulation windows and microphysical sensitivities of $\tau_{\rm BH}$ [3, 11]. The subdivision of HMNS into short- and long-lived cases in Classifier C highlights the increasing difficulty of separating nearby physical regimes: the drop in MCC from Classifier B to C underscores the inherent overlap in inspiral parameter space for remnants with lifetimes just above or below ~ 5 ms.

From an astrophysical standpoint, the ability to classify outcomes with high MCC across all three tasks has clear implications for multimessenger strategies: confident PCBH predictions suggest negligible kilonova and GRB afterglow, while HMNS/NC outcomes increase the likelihood of bright EM counterparts [19, 15]. The probabilistic outputs also enable the integration of GW-based remnant forecasts with EM priors in low-latency pipelines. For example, when early GW posteriors strongly favor an HMNS outcome, optical follow-up resources can be prioritised for the localisation region, while a high-confidence PCBH classification may redirect efforts toward deep, short-timescale radio or high-energy searches.

Finally, the feature importance and SHAP analyses show that the models' learned decision boundaries are not opaque "black boxes" but encode physically interpretable correlations already suggested by phenomenological threshold relations. This alignment between data-driven and theory-motivated structures increases confidence in the robustness of these classifiers when extrapolating to real events. Future improvements may come from enlarging the NR training set—especially with longer post-merger windows and more varied microphysics—and from combining inspiral-only models with low-latency post-merger GW or EM data when available, to further reduce uncertainty in remnant classification.

6 Conclusion

In this work, we have developed, implemented, and rigorously validated a set of three Gradient-Boosted Decision Tree (GBDT) classifiers designed to predict the post-merger remnant outcome of binary neutron star (BNS) mergers using only inspiral-phase parameters — namely the total gravitational mass $M_{\rm tot}$, mass ratio q, effective tidal deformability $\tilde{\Lambda}$, and effective spin parameter $\chi_{\rm eff}$. These parameters are obtainable from gravitational-wave (GW) observations during the inspiral phase, making the models suitable for low-latency analysis before post-merger GW signals are detected.

The classifiers correspond to progressively finer-grained classification tasks:

- **Classifier A**: Binary decision between prompt collapse to a black hole (PCBH) and the formation of any neutron-star (NS) remnant.

- **Classifier B**: Three-way classification distinguishing PCBH, hypermassive neutron star (HMNS), and no collapse within the simulation window (NC).

- **Classifier C**: Four-class classification further splitting HMNS into short-lived ($2\,\mathrm{ms} < \tau_{\mathrm{BH}} < 5\,\mathrm{ms}$) and long-lived ($\tau_{\mathrm{BH}} > 5\,\mathrm{ms}$) remnants.

Training and validation utilised carefully curated numerical-relativity (NR) simulation datasets spanning a wide range of equations of state (EoS), total masses, mass ratios, and spin configurations [12, 13]. Preprocessing steps — including stratified splitting and balancing of misclassified cases — ensured realistic performance evaluation and mitigated dataset biases. Across all classifiers, high accuracy and Matthews Correlation Coefficient (MCC) values were achieved on both validation and complete datasets, with Classifier A achieving perfect scores on validation and near-perfect performance on the total dataset.

Interpretability analysis using SHAP values confirmed that $\tilde{\Lambda}$ consistently emerges as the most influential parameter, aligning with its physical role as an EoS-dependent measure of neutron star compactness and stiffness [8, 2, 28]. The inclusion of q and χ_{eff} enables the models to capture secondary effects, such as the impact of binary asymmetry on mass ejection and angular-momentum redistribution, and spin-induced modifications to collapse thresholds.

Applied to real GW events (GW170817, GW190425), the classifiers produced outcome probabilities that align with expectations from astrophysical modelling and multi-messenger observations. Importantly, the probabilistic outputs enable thresholding and integration with electromagnetic priors, making the framework directly applicable to real-time follow-up decisions in multi-messenger astronomy.

Overall, this work demonstrates that GBDT classifiers trained on NR data can serve as a robust, low-latency tool for predicting the fate of BNS mergers. By leveraging only inspiral-phase parameters, the approach circumvents the limitations of post-merger GW detectability at kHz frequencies while retaining strong physical interpretability. The resulting models hold significant potential for constraining the neutron-star EoS and optimising electromagnetic follow-up strategies in future observing runs.

7 Challenges

Although the present study achieves high classification performance and maintains strong consistency with astrophysical expectations, several limitations and challenges were encountered during its development and evaluation:

- **Dataset coverage:** Despite being one of the most comprehensive collections of numerical-relativity (NR) simulations currently available, the training dataset still represents only a subset of the full binary neutron star (BNS) parameter space. Certain physically plausible configurations — such as highly asymmetric binaries with extreme mass ratios, rapidly spinning components, or EoSs at the very stiff or soft ends of the spectrum — remain underrepresented, which may limit the generalisability of the models.

- **Finite post-merger simulation times:** Simulations with short post-merger durations ($t_{\mathrm{sim}} < 25\,\mathrm{ms}$) were excluded to avoid ambiguity in the remnant classification. While this improves label reliability, it also discards potentially informative cases that could contribute to a more complete decision boundary.

- **Physical degeneracies in parameter space:** Some class boundaries — particularly between short-lived and long-lived HMNS remnants — reflect subtle physical differences that may not be cleanly separable using inspiral parameters alone. Even advanced machine learning (ML) algorithms can struggle to fully resolve these overlaps without additional features.
- **Event-specific parameter uncertainties:** For real GW events, the inferred posterior distributions can be broad, and the resulting probability outputs are sensitive to the choice of priors in the parameter estimation stage (e.g., whether or not EoS-informed constraints are applied). This can introduce variability in the predicted class probabilities across different analysis pipelines.

8 Future Work

To build upon the foundations established in this study and address the limitations identified above, several directions for future research are proposed:

- **Broader and deeper training datasets:** Expanding the NR simulation set to include more extreme parameter combinations, extended post-merger evolution times, and additional physical effects such as neutrino transport, magnetic fields, and viscosity. This would improve model coverage of the full astrophysical parameter space and reduce extrapolation risks.
- **Integration of multi-messenger information:** Extending the classifiers to incorporate low-latency electromagnetic or neutrino detections alongside GW inspiral parameters, enabling joint classification that more directly links merger dynamics to multi-messenger observables.
- **Improved uncertainty quantification:** Refining the entropy-based uncertainty thresholds and exploring Bayesian GBDT variants or ensemble approaches to produce better-calibrated and more robust probability estimates.
- **Real-time deployment:** Incorporating the trained models into operational low-latency GW alert pipelines (e.g., LVK GraceDB), allowing for immediate dissemination of remnant likelihoods to the observing community and facilitating rapid electromagnetic follow-up.

By addressing these challenges and pursuing the proposed extensions, the framework presented here can be evolved into an operational, high-confidence tool for predicting BNS remnant outcomes in future GW observing runs, thereby maximising the scientific return of multi-messenger campaigns.

Acknowledgments

We thank Chris Van Den Broeck, Brian Healy, Thibeau Wouters, and Vsevolod Nedora for the helpful discussion. The authors would like to thank Sebastian Khan for suggesting the use of interpretable machine learning techniques, such as SHAP, to enhance the understanding of our model and analysis. The authors acknowledge funding from the Daimler and Benz Foundation for the project “NUMANJI” and from the European Union (ERC, SMArt, 101076369). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan.

A Assessing the probability of misclassification

In general, the uncertainty of predictions obtained through machine learning algorithms are divided into two classes [32]:

- **Irreducible or data uncertainty**, caused by the complexity of the data, possible multimodal features or noise. In classification problems, it is defined as the entropy of the conditional probability of a given input x belonging to a class k:

$$H[p(y|x)] = -\sum_{k=1}^{K} p(y = \omega_k \mid x) \, \ln p(y = \omega_k \mid x), \tag{1}$$

where $\{x\}$ is the set of inputs, K is the total number of classes, and $y = \{\omega_1, \ldots, \omega_K\}$ are the labels.

- **Knowledge or epistemic uncertainty**, which derives from the model lacking the capability to describe (some) data (e.g., when the input for which a prediction is required comes from a different distribution than the one used for training).

The knowledge uncertainty quantifies potential shortcomings of the model, but it cannot be directly computed from the model's predictions. To assess such model uncertainty, we follow a Bayesian ensemble-based framework [32]. Consider an ensemble of M models with parameters $\theta^{(m)}$. We can estimate the knowledge uncertainty as the difference between the total uncertainty and the expected data uncertainty via the mutual information between parameters θ and predictions y:

$$I[y, \theta] \approx H\left(\frac{1}{M}\sum_{m=1}^{M} p(y \mid x; \theta^{(m)})\right) - \frac{1}{M}\sum_{m=1}^{M} H\big(p(y \mid x; \theta^{(m)})\big) \, . \tag{2}$$

To generate the model ensemble, we employ stochastic gradient boosting with different random states [26, 27]. For a point x in the validation set, larger values of the knowledge uncertainty

$U(x)$ indicate that the prediction is more likely to be a misclassification. A simple detector compares $U(x)$ against a threshold T:

$$I_T(x) = \begin{cases} 1, & U(x) > T, \\ 0, & U(x) \le T. \end{cases} \tag{3}$$

By sweeping T and comparing to the known validation labels, one can build ROC and precision–recall curves to derive operating points with bounded false positive rate and near-unity true positive rate [33, 34]. Using a single (non-ensemble) model, thresholds $T_A \approx 0.3$, $T_B \approx 0.06$, and $T_C \approx 0.09$ provide conservative flags for high-uncertainty predictions in Classifiers A, B, and C, respectively.

References

[1] L. Baiotti and L. Rezzolla, Binary neutron star mergers: a review of Einstein's richest laboratory, *Rep. Prog. Phys.* **80**, 096901 (2017).

[2] T. Dietrich, T. Hinderer, and A. Samajdar, Interpreting Binary Neutron Star Mergers, *Gen. Relativ. Gravit.* **53**, 27 (2021).

[3] D. Radice, S. Bernuzzi, and A. Perego, The Dynamics of Binary Neutron Star Mergers and GW170817, *Annu. Rev. Nucl. Part. Sci.* **70**, 95–119 (2020).

[4] A. Bauswein, T. W. Baumgarte, and H.-T. Janka, Prompt merger collapse and the maximum mass of neutron stars, *Phys. Rev. Lett.* **111**, 131101 (2013).

[5] S. Köppel, L. Bovard, and L. Rezzolla, A GR determination of the threshold mass to prompt collapse, *Astrophys. J. Lett.* **872**, L16 (2019).

[6] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, *Class. Quantum Grav.* **32**, 074001 (2015).

[7] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo, *Class. Quantum Grav.* **32**, 024001 (2015).

[8] E. E. Flanagan and T. Hinderer, Constraining neutron-star tidal Love numbers, *Phys. Rev. D* **77**, 021502(R) (2008).

[9] T. Hinderer, Tidal Love numbers of neutron stars, *Astrophys. J.* **677**, 1216–1220 (2008).

[10] J. Vines, E. E. Flanagan, and T. Hinderer, Post-1-Newtonian tidal effects in inspiral waveforms, *Phys. Rev. D* **83**, 084051 (2011).

[11] A. Puecher and T. Dietrich, A machine-learning classifier for the postmerger remnant of binary neutron stars, arXiv:2408.10678 (2025).

[12] T. Dietrich *et al.*, CoRe database of BNS merger waveforms, *Class. Quantum Grav.* **35**, 24LT01 (2018).

[13] K. Kiuchi *et al.*, Sub-radian-accuracy gravitational waves from coalescing BNS in NR II, *Phys. Rev. D* **101**, 084006 (2020).

[14] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW170817: Observation of GWs from a BNS Inspiral, *Phys. Rev. Lett.* **119**, 161101 (2017).

[15] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Multi-messenger Observations of a BNS Merger, *Astrophys. J. Lett.* **848**, L12 (2017).

[16] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW190425: Observation of a compact binary with total mass $\sim 3.4\,M_\odot$, *Astrophys. J. Lett.* **892**, L3 (2020).

[17] A. Goldstein *et al.*, An ordinary short GRB with extraordinary implications: GRB 170817A, *Astrophys. J. Lett.* **848**, L14 (2017).

[18] S. J. Smartt *et al.*, A kilonova as the electromagnetic counterpart to a GW source, *Nature* **551**, 75–79 (2017).

[19] D. Kasen, B. Metzger, J. Barnes, E. Quataert, and E. Ramirez-Ruiz, Origin of the heavy elements in binary NS mergers, *Nature* **551**, 80–84 (2017).

[20] M. Punturo *et al.*, The Einstein Telescope: a third-generation GW observatory, *Class. Quantum Grav.* **27**, 194002 (2010).

[21] M. Maggiore *et al.*, Science Case for the Einstein Telescope, *JCAP* **03**, 050 (2020).

[22] D. Reitze *et al.*, Cosmic Explorer: The U.S. contribution to GW astronomy beyond LIGO, *BAAS* **51**, 035 (2019).

[23] E. Cuoco *et al.*, Enhancing Gravitational-Wave Science with Machine Learning, *Mach. Learn. Sci. Technol.* **2**, 011002 (2021).

[24] A. Bauswein and N. Stergioulas, Semi-analytic relations for post-merger GW emission, *Mon. Not. R. Astron. Soc.* **471**, 4956–4965 (2017).

[25] C. J. Haster *et al.*, Machine learning for GW inference: recent advances, *Mach. Learn.: Sci. Technol.* **1**, 035014 (2020).

[26] J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Statist.* **29**, 1189–1232 (2001).

[27] F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[28] S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, *Adv. Neural Inf. Process. Syst.* **30** (2017).

[29] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta* **405**, 442–451 (1975).

[30] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, *Comput. Biol. Chem.* **28**, 367–374 (2004).

[31] K. Kiuchi *et al.*, High-resolution magnetohydrodynamic simulations of BNS mergers, *Phys. Rev. D* **101**, 084006 (2020).

[32] B. Lakshminarayanan, A. Pritzel, and C. Blundell, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, *Adv. Neural Inf. Process. Syst.* **30** (2017).

[33] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.* **27**, 861–874 (2006).

[34] T. Saito and M. Rehmsmeier, The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, *PLOS ONE* **10**, e0118432 (2015).