University of Heidelberg

Department of Physics and Astronomy

Master's Thesis

From Light Curves to Labels: Machine Learning in Microlensing

Author: Kunal Bhatia

Matriculation Number: 3770453

First Supervisor: Prof. Dr. Joachim Wambsganß

Second Supervisor: Dr. Yiannis Tsapras

Submission Date: February 2, 2026

Declaration of Authorship

I hereby declare that this thesis is my own work and that I have not used any sources or aids other than those stated. I have marked as such all passages taken word-for-word or in content from other works. Furthermore, I assure that the electronic version submitted corresponds completely in content and wording to the printed version of my thesis. I agree that this electronic version may be checked for plagiarism using plagiarism detection software within the university.

Heidelberg, February 2, 2026
Kunal Bhatia

Abstract

The Nancy Grace Roman Space Telescope will detect tens of thousands of gravitational microlensing events, with many exhibiting binary lens signatures from star-planet or binary star systems. Early identification of these events is critical for coordinating time-sensitive ground-based follow-up observations to constrain system parameters and detect planetary companions. Traditional classification methods based on χ^2 model fitting are computationally expensive and require complete light curves, limiting their utility for real-time survey operations.

This thesis presents a machine learning classifier that provides real-time classification of microlensing events as they unfold. The CNN-GRU architecture processes Roman-like observations (15-minute cadence, 72-day seasons) and outputs continuously updated probabilities for three event classes: flat, PSPL (single lens), and binary lens systems. This dynamic capability enables rapid decision-making for follow-up strategies, allowing observation requests while events are still developing.

Trained on 600,000 synthetic light curves and validated on 300,000 independent events, the classifier achieves 99% accuracy on events with clear caustic crossings. Performance naturally decreases on the full event population, reflecting the fundamental physical limit where high-impact-parameter binary events become indistinguishable from single-lens profiles. The system provides sub-millisecond inference, making it suitable for processing thousands of events in real-time alert streams.

The classifier addresses a critical operational need for next-generation microlensing surveys by enabling automated early detection of scientifically valuable binary events. The open-source implementation provides a practical tool for developing Roman alert systems and coordinating observations between space-based and ground-based networks.

Contents

Abstract v

List of Figures xi

List of Tables xiii

1. Introduction 1

1.1. Motivation and Background 1
1.2. Research Problem 2
1.3. Research Questions 3
1.4. Objectives 4
1.5. Contributions 5
1.6. Thesis Organization 5

2. Theoretical Background 7

2.1. Gravitational Lensing 7
2.1.1. Historical Foundation 7
2.1.2. The Lens Equation 7
2.1.3. The Microlensing Regime 8
2.2. Point-Source Point-Lens Model 9
2.2.1. Magnification Pattern 9
2.2.2. Observable Features 10
2.3. Binary Lensing 10
2.3.1. The Binary Lens Equation 11
2.3.2. Caustics and Critical Curves 11
2.3.3. Light Curve Morphologies 12
2.3.4. Planetary Companions 13
2.4. Machine Learning for Microlensing Classification 13
2.4.1. The Classification Task 15
2.4.2. Neural Networks for Time Series 15
2.4.3. Training on Synthetic Data 16
2.4.4. Optimization and Regularization 17
2.4.5. Real-Time Processing and Calibration 17

2.5. Summary 18

3. Literature Review **19**

3.1. Microlensing Surveys and Detection Strategies 19
3.1.1. Ground-Based Survey Operations 19
3.1.2. Future Space-Based Observations 20
3.2. Traditional Classification Approaches 21
3.2.1. Chi-Squared Model Fitting 21
3.2.2. Bayesian Inference Methods 21
3.2.3. Expert Visual Inspection 22
3.3. Machine Learning in Time-Domain Astronomy 22
3.3.1. Classical Feature-Based Approaches 22
3.3.2. Deep Learning for Light Curve Classification 23
3.3.3. Applications to Microlensing 23
3.4. Early Classification in Other Transient Domains 24
3.4.1. Supernova Photometric Typing 24
3.4.2. Transient Identification in LSST 24
3.4.3. Lessons for Microlensing Classification 24
3.5. The Research Gap 25
3.6. Summary 26

4. Methodology **29**

4.1. Overview 29
4.2. Synthetic Data Generation 29
4.2.1. Light Curve Generation 30
4.2.2. Observational Realism 31
4.2.3. Training Set Design 32
4.3. Neural Network Architecture 32
4.3.1. Input Representation 34
4.3.2. Local Feature Extraction 35
4.3.3. Sequence Memory 35
4.3.4. Sequence Aggregation 36
4.3.5. Hierarchical Classification 36
4.4. Training 37
4.4.1. Loss Function Design 37
4.4.2. Optimization and Training Infrastructure 38
4.5. Evaluation 39
4.5.1. Cross-Validation Strategy 39
4.5.2. Impact Parameter Dependency 39
4.5.3. Real-Time Capability 40

4.6. Summary 40

5. Results **41**

5.1. Dataset Statistics 41
5.2. Classification Metrics 44
5.3. Physical Analysis & Bias 47
5.4. Event Evolution Examples 49

6. Discussion **59**

7. Conclusions and Future Work **61**

Bibliography **63**

A. Code and Implementation Details **69**

List of Figures

2.1. Point-source point-lens geometry and light curve 10
2.2. Binary lens caustic and critical curve structures 12
2.3. Binary microlensing light curve morphologies 14

4.1. Example synthetic light curves for each event class 33
4.2. Architecture of the hierarchical classifier 37
4.3. Training convergence for a representative model 39
4.4. Cross-evaluation results 40

5.1. Class distributions 42
5.2. Example light curves 43
5.3. Confusion matrix 44
5.4. Per-class performance metrics 45
5.5. ROC curves 45
5.6. Calibration plots 46
5.7. Impact parameter dependency 47
5.8. Temporal bias check 48
5.9. Evolution of Binary Event 3 49
5.10. Evolution of Binary Event 12 50
5.11. Evolution of Binary Event 16 51
5.12. Evolution of PSPL Event 1 52
5.13. Evolution of PSPL Event 4 53
5.14. Evolution of PSPL Event 5 54
5.15. Evolution of Flat Event 0 55
5.16. Evolution of Flat Event 2 56
5.17. Evolution of Flat Event 6 57

List of Tables

4.1. Shared parameters for all synthetic event types 31
4.2. Additional parameters for binary lens events 31

1. Introduction

1.1. Motivation and Background

Modern astronomy is increasingly defined by the scale and velocity of data acquisition, pushing the boundaries of computational capacity and observational efficiency. Wide-field time-domain surveys have transformed the field from targeted observation to automated, rapid data processing across millions of stellar sources. Among the most powerful techniques for detecting exoplanets and probing the structure of the Galaxy is gravitational microlensing [1, 2], a phenomenon predicted by Einstein's general relativity in which a massive foreground object (the lens) temporarily magnifies the light of a background star (the source) as they align along the line of sight to Earth [3].

The elegance of gravitational microlensing lies in its mass-dependent nature: because the lensing effect depends only on mass and geometry rather than the luminosity of the lens, microlensing can detect planets around faint stars, free-floating planets unbound to any host, and stellar remnants that would remain invisible to traditional detection methods [4, 5]. The probability that any random star is sufficiently aligned to be lensed is extremely small—the optical depth toward the Galactic bulge is of order 10^{-6} [6]—so millions of stars must be continuously monitored to detect even a handful of events. Over the past three decades, ground-based surveys such as the Optical Gravitational Lensing Experiment (OGLE) and the Microlensing Observations in Astrophysics (MOA) have monitored dense stellar fields in the Galactic bulge at cadences of minutes to hours, discovering over 200 exoplanets and characterizing the demographics of planetary systems throughout the Galaxy [7–9].

The majority of observed microlensing events are well described by the Point-Source Point-Lens (PSPL) model [2], which assumes both the lens and source are point masses. These events produce characteristic, symmetric, achromatic light curves described by only three parameters: the Einstein crossing time, the impact parameter, and the time of maximum magnification. However, the most scientifically valuable events are those caused by binary lenses—two-body systems such as a star-planet pair or a binary star system. Binary lens events produce complex magnification patterns with caustic structures, multiple peaks, and asymmetries that depend sensitively on the lens geometry and the source trajectory [3, 10]. These deviations from the simple PSPL template are crucial for determining physical parameters such as planetary mass ratios

and projected separations.

The primary observational challenge lies in the temporal ambiguity problem: during the initial rise of a microlensing event, binary lens systems often masquerade as simple PSPL events, only revealing their true nature when the source approaches or crosses caustic features in the lens plane. This degeneracy persists until the event is sufficiently developed, leading to critical delays in follow-up observations necessary to fully characterize the binary system. Current surveys discover approximately 2,000 events per year [9]—a number that human experts and traditional modeling approaches can manage through careful monitoring. However, the upcoming data deluge from the Vera C. Rubin Observatory's Legacy Survey of Space and Time (*LSST*) will increase this rate tenfold, discovering an estimated 20,000 microlensing events annually [11, 12]. Simultaneously, the Nancy Grace *Roman* Space Telescope, with its dedicated Galactic Bulge Time-Domain Survey, aims to detect on the order of 2×10^4 microlensing events over its mission lifetime [13, 14], providing complementary space-based observations with superior photometric precision and continuous monitoring capabilities. This enormous increase in event detection rates mandates a complete re-evaluation of current classification strategies. The future of microlensing science, particularly in the statistical characterization of exoplanet populations and Galactic structure, depends critically on developing automated systems capable of early, reliable classification.

1.2. Research Problem

The fundamental challenge addressed by this thesis is the computational inefficiency and latency of classifying binary microlensing events in the era of large-scale surveys. Traditional classification methods, typically relying on χ^2 model fitting or manual inspection, suffer from several critical limitations. First, they are computationally intensive and scale poorly with the incoming data rates expected from *LSST* and *Roman*. Binary lens models have significantly more free parameters than PSPL models (typically seven or more versus three), and their complex parameter spaces contain numerous local minima, making optimization challenging and computationally expensive [15]. Fitting a single binary lens model to a light curve can require seconds to minutes per evaluation, and characterizing a single event can take hours of computation when accounting for multiple binary topologies and the need for global optimization. When multiplied across thousands of ongoing events that must be updated with each new observation, the computational burden becomes prohibitive for real-time classification during survey operations.

Second, traditional methods lack a principled framework for handling partial observations. Most fitting procedures assume access to a complete or nearly complete light curve, yielding unreliable parameter estimates when applied to early-time data before

the event has reached peak magnification or revealed its binary nature. The standard approach of fitting PSPL models to incomplete data and flagging large residuals works reasonably well for high-magnification events but performs poorly for events with subtle binary perturbations that only become apparent late in the observational sequence. This lack of early-warning capability is particularly problematic for planetary detections, where short-lived caustic crossings may occur unpredictably during any phase of the event and require immediate high-cadence follow-up to fully characterize.

Third, existing classification infrastructure is already severely strained by current survey data rates, and will be inadequate for *LSST*-era operations. The current reliance on human-in-the-loop systems and computationally expensive template fitting is not sustainable for the upcoming $\sim$10$\times$ increase in event detection rate. Manual inspection of thousands of ongoing light curves is impractical; automated systems that can process incoming photometry in real time, assign reliable classification probabilities, and trigger follow-up observations based on scientifically motivated criteria are essential. The computational efficiency of any such system must be orders of magnitude better than current approaches to enable processing of alert streams within seconds of new observations arriving.

This thesis addresses the research problem: *How can we design a machine-learning model that reliably distinguishes binary microlensing events from PSPL events using only partial, noisy light curves, enabling early classification and efficient resource allocation for follow-up observations?* This requires moving beyond traditional feature engineering toward a representation learning approach that can model the temporal dependencies of light curve data, process incomplete observations as they arrive, scale efficiently to tens of thousands of events, and provide reliable confidence estimates to guide operational decisions.

1.3. Research Questions

To make the research problem tractable and measurable, this thesis addresses the following specific research questions:

1. **Can a CNN-GRU architecture effectively distinguish between flat, PSPL, and binary microlensing events from synthetic Roman-like light curves?** We investigate whether deep learning can learn representations that capture caustic crossing signatures without explicit feature engineering.

2. **How does classification performance depend on the training data distribution?** We examine performance when training on events with clear caustic signatures (distinct distribution, $u_0 < 0.3$) versus more challenging realistic populations (general distribution, full parameter space).

3. **What are the dominant error patterns and do they reflect physical limitations or algorithmic failures?** We analyze the confusion matrix to determine whether misclassifications occur in regimes where binary events are genuinely indistinguishable from single lenses.

4. **Does the model provide well-calibrated probability estimates suitable for operational decision-making?** We assess whether predicted confidence levels accurately reflect true classification accuracy, which is essential for prioritizing events in real surveys.

These questions provide a structured framework for evaluating the feasibility of machine learning approaches to binary microlensing classification in preparation for Roman operations.

1.4. Objectives

To address these research questions, this thesis pursues the following concrete objectives:

1. **Synthetic Data Generation:** Generate a large-scale synthetic dataset of 900,000 high-fidelity light curves (600,000 training, 300,000 test) using VBBinaryLensing with Roman-like observational parameters (15-minute cadence, 72-day seasons). Implement a dual-distribution sampling strategy: a distinct distribution emphasizing clear caustic signatures ($u_0 < 0.3$, $s \in [0.8, 1.2]$) and a general distribution representing the expected Roman event population across the full parameter space.

2. **Architecture Development:** Design and implement a CNN-GRU hierarchical classifier specifically tailored for microlensing light curves. The architecture combines convolutional layers for local feature extraction (caustic crossings) with GRU layers for long-term temporal memory, using a two-stage hierarchical structure (Flat vs Non-Flat, then PSPL vs Binary) to reflect the logical classification problem.

3. **Performance Evaluation:** Evaluate classification accuracy, precision, recall, F1-scores, and ROC-AUC on held-out test sets from both distinct and general distributions. Analyze confusion matrices to identify error patterns and assess whether performance degradation reflects genuine physical detection limits.

4. **Model Interpretability:** Examine calibration curves to verify that predicted probabilities accurately reflect true classification accuracy. Analyze example probability evolution trajectories showing how classifications update as observations arrive during the 72-day season.

Together, these objectives aim to demonstrate the feasibility of deep learning for binary microlensing classification and establish baseline performance benchmarks for future work.

1.5. Contributions

This thesis makes the following contributions to automated microlensing classification:

1. **Large-Scale Synthetic Dataset:** We generate 900,000 high-fidelity synthetic light curves using VBBinaryLensing with Roman-like observational parameters (15-minute cadence, 72-day seasons). The dataset includes 600,000 training events and 300,000 test events with equal representation of three classes (flat, PSPL, binary). A dual-distribution strategy—distinct events with clear caustic signatures ($u_0 < 0.3$) and general events representing the full parameter space—enables controlled evaluation of generalization.

2. **Proof-of-Concept CNN-GRU Classifier:** We demonstrate that a compact hierarchical architecture (33,541 parameters) can achieve 99.65% accuracy on events with clear binary signatures, validating that deep learning can recognize caustic crossing patterns without hand-engineered features. The hierarchical structure (Flat vs Non-Flat, then PSPL vs Binary) naturally decomposes the classification problem.

3. **Calibration Assessment:** We demonstrate that the model produces well-calibrated probability estimates where predicted confidence levels accurately reflect true classification accuracy. This calibration is essential for operational use, enabling reliable event prioritization based on model outputs.

4. **Open-Source Implementation:** All code for data generation, model training, and evaluation is released to facilitate reproduction and extension of this work.

These contributions establish the feasibility of deep learning approaches for binary microlensing classification and provide a foundation for future development of operational systems for Roman and other next-generation surveys.

1.6. Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2: Theoretical Background reviews the physics of gravitational microlensing, deriving the lens equation for both single (PSPL) and binary lenses. We explain

the origin of caustic structures in binary systems and describe how caustic crossings produce the distinctive light curve features that distinguish binary from single-lens events. The chapter introduces the mathematical framework needed to understand the synthetic data generation and classification problem.

Chapter 3: Literature Review surveys related work in three areas: traditional microlensing classification methods (χ^2 fitting, Bayesian inference), machine learning applications to astronomical time series (supernova classification, exoplanet detection), and previous attempts to apply neural networks to microlensing. We identify the research gap: the absence of end-to-end deep learning classifiers that operate directly on raw photometry without requiring feature engineering or model fitting.

Chapter 4: Methodology describes the technical approach. We explain the synthetic data generation using VBBinaryLensing to create 900,000 Roman-like light curves, explain the dual-distribution sampling strategy (distinct vs general), present the CNN-GRU hierarchical architecture, and document the training procedures and evaluation protocols.

Chapter 5: Results presents classification performance on distinct and general test sets. We report overall accuracy, per-class metrics, confusion matrices, and calibration curves. Analysis of error patterns reveals that Binary→PSPL misclassifications dominate, consistent with physical degeneracies at high impact parameters. Probability evolution examples demonstrate how classifications update as observations arrive.

Chapter 6: Discussion interprets the results in context, examining whether performance limitations reflect algorithmic failures or genuine physical detection thresholds. We discuss implications for Roman operations, analyze the model's strengths and limitations, and contextualize findings relative to previous work in astronomical machine learning.

Chapter 7: Conclusions and Future Work summarizes key findings, reviews how the work addresses each research question, and outlines directions for future research including application to real survey data, extension to more complex lens systems, and integration with Roman alert systems.

Appendices provide supporting material including derivations of magnification formulas, complete architecture specifications, and additional performance metrics.

This organizational structure ensures that readers can follow the logical progression from motivation through methods to conclusions, while also enabling experts to navigate directly to chapters of specific interest. Figures and tables throughout illustrate key concepts, simulation outcomes, and classifier performance. A comprehensive bibliography provides context within the broader literature of microlensing physics, exoplanet detection, and machine learning applications in astronomy.

2. Theoretical Background

This chapter establishes the theoretical foundation for gravitational microlensing and the classification problem addressed in this thesis. We begin with the general relativistic basis of gravitational lensing (Section 2.1), derive the lens equation and magnification formulas for point-source point-lens (PSPL) events (Section 2.2), introduce binary lens systems with their characteristic caustic and critical curve structures (Section 2.3), and conclude with a minimal introduction to the machine learning approach employed (Section 2.4).

2.1. Gravitational Lensing

2.1.1. Historical Foundation

The phenomenon of light deflection by massive objects represents one of the most striking predictions of Einstein's general theory of relativity. In 1915, Einstein demonstrated that spacetime curvature caused by mass would bend the paths of light rays passing nearby. This prediction was famously confirmed during the solar eclipse of 1919, when Arthur Eddington observed the apparent displacement of stars near the Sun's limb [16]. The measured deflection of approximately 1.75 arcseconds agreed closely with Einstein's prediction and was twice the value expected from Newtonian gravity. This result, widely publicized in the press, made Einstein a household name and inaugurated the study of gravitational lensing.

While early observations focused on strong lensing effects by galaxies—producing multiple resolved images and dramatic Einstein rings—gravitational microlensing operates on much smaller angular scales. Rather than resolving separate images, microlensing by individual stars manifests as temporary brightness variations in background sources. The image separations are of order milli-arcseconds, too small to be resolved with current instruments; instead, microlensing is detected through time-varying brightness changes.

2.1.2. The Lens Equation

Consider a light ray passing a lens of mass M at impact parameter ξ (measured perpendicular to the unperturbed ray direction). In the thin-lens approximation,

appropriate when lens thickness is small compared to source-lens and lens-observer distances, the deflection angle is:

$$\hat{\alpha}(\boldsymbol{\xi}) = \frac{4GM}{c^2}\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|^2}, \tag{2.1}$$

where G is the gravitational constant and c is the speed of light. This deflection creates an apparent displacement between the source's true position $\boldsymbol{\beta}$ and its observed position $\boldsymbol{\theta}$ in the sky.

The geometry of lensing involves three distances: $D_{\rm L}$ (observer to lens), $D_{\rm S}$ (observer to source), and $D_{\rm LS} = D_{\rm S} - D_{\rm L}$ (lens to source). Angular positions in the source plane $\boldsymbol{\beta}$ and lens plane $\boldsymbol{\theta}$ are related through the lens equation:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_{\rm LS}}{D_{\rm S}}\hat{\boldsymbol{\alpha}}(D_{\rm L}\,\boldsymbol{\theta})\,. \tag{2.2}$$

This vector equation represents the requirement that light rays from the source position $\boldsymbol{\beta}$ must deflect to reach the observer at apparent position $\boldsymbol{\theta}$.

2.1.3. The Microlensing Regime

Microlensing is characterized by two defining properties that distinguish it from strong lensing by galaxies. First, the angular separations between multiple images are extremely small—of order microarcseconds to milliarcseconds—preventing their resolution with current instrumentation. Second, the lens-source relative motion is measurable on human timescales (days to months), producing time-variable magnification as the alignment changes. These properties emerge when the lens mass is stellar-scale ($M \sim 0.01$–$10\,M_\odot$) and source-lens-observer distances are kiloparsecs.

A natural angular scale for the problem is the Einstein radius θ_E, defined as the angular radius of the Einstein ring produced when source and lens are perfectly aligned:

$$\theta_E = \sqrt{\frac{4GM}{c^2}\frac{D_{\rm LS}}{D_{\rm L}D_{\rm S}}}. \tag{2.3}$$

For typical Galactic microlensing geometry—a solar-mass lens at 4 kpc lensing a source at 8 kpc in the bulge—$\theta_E \approx 0.5$ milliarcseconds. Working in units of θ_E simplifies the equations considerably.

The physical size corresponding to θ_E projected onto the lens plane defines the Einstein radius proper:

$$r_E = D_{\rm L}\,\theta_E = \sqrt{\frac{4GM\,D_{\rm L}\,D_{\rm LS}}{c^2\,D_{\rm S}}}. \tag{2.4}$$

For solar-mass lenses in the bulge, $r_E \sim 3$ AU. This scale is comparable to planetary orbital separations, explaining why microlensing is sensitive to planetary companions

around the lens star.

The Einstein crossing time t_E is the time required for the source to traverse an Einstein radius due to lens-source relative transverse motion:

$$t_E = \frac{r_E}{v_\perp} = \frac{\theta_E \, D_\mathrm{S}}{v_\perp}, \tag{2.5}$$

where $v_\perp$ is the lens-source transverse velocity. For typical Galactic kinematics with $v_\perp \sim 200$ km/s, Einstein crossing times range from 10 to 100 days. This timescale determines the duration over which a microlensing event is observable and sets the required survey cadence for adequate temporal sampling.

2.2. Point-Source Point-Lens Model

The simplest microlensing configuration involves a point-mass lens and a point source. Despite its simplicity, this model describes the majority of detected events and provides the baseline against which binary lens events are distinguished.

2.2.1. Magnification Pattern

For a point-mass lens, Equation (2.2) admits exactly two solutions corresponding to two images, one inside and one outside the Einstein ring. The magnification of each image is determined by the local Jacobian of the lens mapping, yielding a total magnification:

$$A_\mathrm{PSPL}(u) = \frac{u^2 + 2}{u\sqrt{u^2 + 4}}, \tag{2.6}$$

where u is the instantaneous lens-source separation in units of θ_E. The magnification diverges as $u \to 0$ (perfect alignment) and approaches unity as $u \to \infty$ (no lensing effect).

For a source moving with constant transverse velocity, the separation evolves as:

$$u(t) = \sqrt{u_0^2 + \left(\frac{t - t_0}{t_E}\right)^2}, \tag{2.7}$$

where u_0 is the impact parameter (minimum separation at time t_0). The resulting light curve $A(t)$ exhibits the characteristic symmetric, achromatic magnification peak that defines PSPL events. High-magnification events ($u_0 \ll 1$) show dramatic flux increases, while low-magnification events ($u_0 \gtrsim 1$) produce only subtle brightening.

2.2.2. Observable Features

PSPL light curves are fully determined by three parameters: the Einstein crossing time t_E, the impact parameter u_0, and the time of closest approach t_0. These parameters govern the event duration, maximum magnification, and temporal position respectively. Importantly, PSPL light curves are perfectly symmetric about t_0—the rise and fall phases are mirror images. They are also strictly achromatic: the magnification factor is independent of wavelength since it depends only on geometry and mass, not on the intrinsic spectral energy distribution of either lens or source. These properties—three-parameter simplicity, temporal symmetry, and achromaticity—make PSPL events straightforward to model and serve as the reference template for identifying deviations caused by binary lens systems.

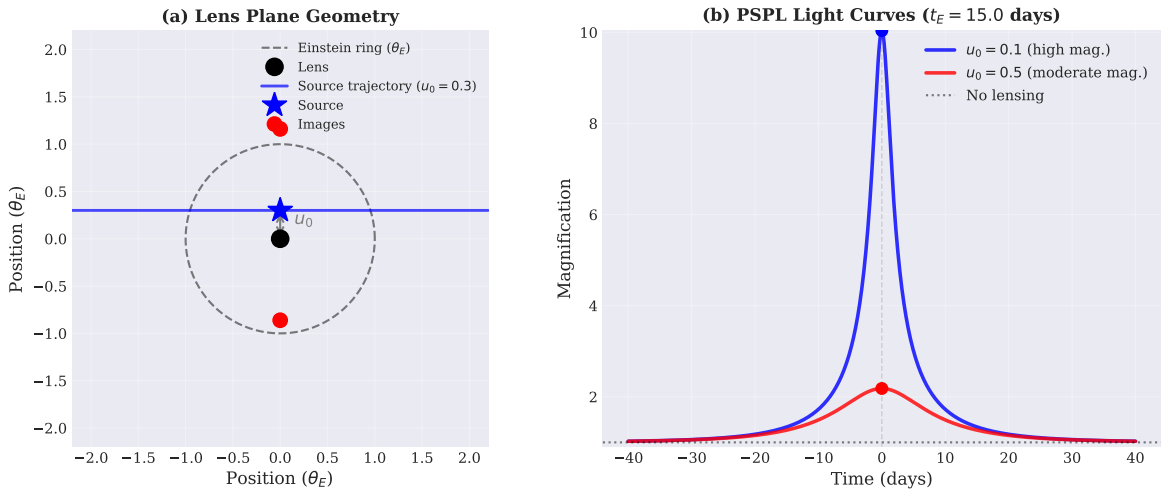

Figure 2.1.: Point-source point-lens geometry and light curve. Left: Source trajectory (blue line) with impact parameter u_0 relative to the lens (black dot). Right: Resulting symmetric magnification pattern showing characteristic temporal profile with Einstein timescale t_E.

2.3. Binary Lensing

Binary lens systems—composed of two masses such as a host star with a planetary companion or a binary star pair—produce substantially more complex magnification patterns than single lenses. The richness of binary lens phenomenology arises from caustic and critical curve structures that create regions of extreme magnification and sharp temporal features in the light curves.

2.3.1. The Binary Lens Equation

For a binary system with mass ratio $q = M_2/M_1$ and separation s (in Einstein radius units), the lens equation generalizes to:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{1}{1+q}\frac{\boldsymbol{\theta} - \boldsymbol{\theta}_1}{|\boldsymbol{\theta} - \boldsymbol{\theta}_1|^2} - \frac{q}{1+q}\frac{\boldsymbol{\theta} - \boldsymbol{\theta}_2}{|\boldsymbol{\theta} - \boldsymbol{\theta}_2|^2}, \tag{2.8}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the positions of the two lens components. Unlike the single-lens case, this equation cannot be solved analytically; it admits up to five image solutions depending on the source position, and numerical methods are required for evaluation [17].

The parameter space for binary lensing is significantly richer than PSPL. In addition to t_E, t_0, and u_0, binary events require specification of the mass ratio q (spanning four orders of magnitude from planetary mass ratios $q \sim 10^{-4}$ to equal-mass binaries $q = 1$), the projected separation s (typically 0.3–3.0 Einstein radii), and the source trajectory angle α relative to the binary axis. This high-dimensional parameter space, combined with complex topologies that change qualitatively across the parameter ranges, makes comprehensive binary lens modeling computationally demanding.

2.3.2. Caustics and Critical Curves

The defining structures in binary lensing are *caustics* in the source plane and *critical curves* in the image plane. These are intrinsically related: critical curves are locations in the image plane where the lens mapping Jacobian becomes singular (i.e., where $\det|\partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}| = 0$), corresponding to infinite magnification. The caustics are the images of the critical curves under the lens mapping—that is, the locations in the source plane that map to the critical curves in the image plane. Mathematically, if $\boldsymbol{\theta}_{\rm crit}$ is a point on the critical curve, then $\boldsymbol{\beta}_{\rm caustic} = \boldsymbol{\theta}_{\rm crit} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}_{\rm crit})$ is the corresponding point on the caustic.

When a source crosses a caustic, a pair of images is created or destroyed at the critical curve, producing a rapid change in total magnification. This manifests as a sharp spike or cusp in the observed light curve—the characteristic signature that distinguishes binary events from smooth PSPL profiles. The closer a source trajectory passes to a caustic without crossing it, the more pronounced the deviation from PSPL behavior becomes, appearing as asymmetric perturbations in the light curve.

The topology of caustics depends sensitively on the binary parameters s and q. For wide binaries ($s \gtrsim 1.0$), the caustic structure typically consists of a central caustic near the system's center of mass and a smaller planetary caustic near the lower-mass component. For close binaries ($s \lesssim 1.0$), the caustics can merge into more complex configurations. For planetary mass ratios ($q \ll 1$), the planetary caustic becomes very

small, requiring high-magnification events to probe it, but caustic crossings produce the sharpest spikes that enable planetary mass and separation determinations.

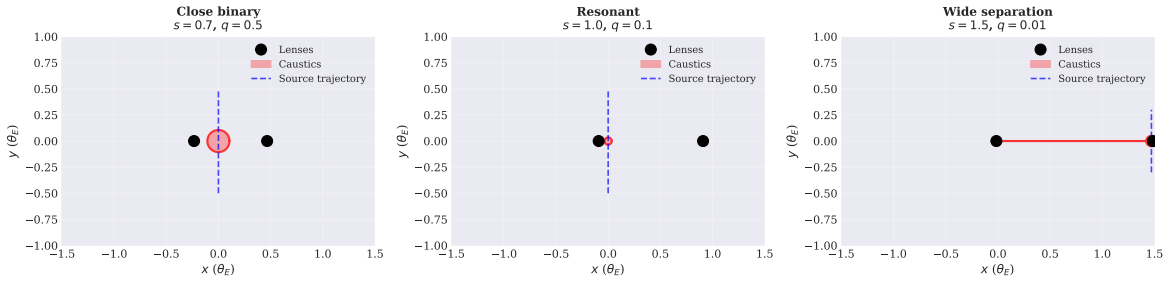

Figure 2.2.: Binary lens caustic and critical curve structures for different parameter regimes. The caustics (red regions in source plane) represent locations of formally infinite magnification for point sources. Critical curves (blue curves shown in insets) are the corresponding structures in the image plane where image pairs are created or destroyed. The two black dots indicate lens positions (primary at origin, companion at separation s). Source trajectories (dashed lines) crossing or approaching caustics produce characteristic deviations from PSPL light curves.

2.3.3. Light Curve Morphologies

The magnification for a binary lens must be computed by summing contributions from all images. Since no closed-form solution exists, numerical techniques such as ray-shooting or inverse ray-tracing are employed at each time step [17]. The resulting light curves exhibit several characteristic features that distinguish them from PSPL events:

- **Caustic crossings:** When the source trajectory intersects a caustic, the light curve displays a sharp spike or cusp. The spike duration depends on the source size (for finite sources) and the transverse velocity. For point sources, caustic crossings would produce mathematically infinite peaks; realistic finite source sizes ($\rho \sim 10^{-3}$ to 10^{-2} Einstein radii) smooth these features into peaks with characteristic widths determined by ρ and the crossing speed.

- **Asymmetric profiles:** Unlike the perfect symmetry of PSPL curves, binary light curves typically show asymmetric rise and fall times. The shape depends on the source trajectory relative to the caustic structure and the binary orientation angle α.

- **Multiple peaks:** Depending on geometry, a source may cross multiple caustic features or approach several caustics without crossing, producing light curves with two or more distinct peaks separated by days to weeks.

- **Extended perturbations:** Even without caustic crossings, the perturbation from the binary companion creates extended deviations from the PSPL template. Near-caustic approaches produce asymmetric bumps or shoulders in the light curve that can last for a significant fraction of the Einstein timescale.

The probability of detecting binary features depends critically on the impact parameter u_0. High-magnification events ($u_0 \lesssim 0.3$) have source trajectories that pass close to or through caustic structures, making binary signatures prominent and easily distinguishable from PSPL profiles. Low-magnification events ($u_0 \gtrsim 0.3$), however, have trajectories that may never approach caustics closely, and binary perturbations become subtle or absent. This fundamental geometric constraint means binary lenses are intrinsically indistinguishable from single lenses in certain parameter regimes—a physical detection limit rather than an algorithmic limitation.

2.3.4. Planetary Companions

Planetary companions with mass ratios $q \sim 10^{-3}$ to 10^{-4} produce small planetary caustics with characteristic sizes proportional to $\sqrt{q}$. The probability of a random source trajectory crossing these small caustics is low, but when crossings occur, they produce short-duration ($\Delta t \sim 0.01$–$0.1\, t_E$), high-amplitude spikes that are unmistakable signatures of planetary companions. These spikes can occur at any phase of the event—during the initial rise, near the peak, or during the decline—depending on the planetary separation and the orientation of the source trajectory.

High-cadence observations (multiple times per night) are essential to capture these brief planetary signals, which might be missed by surveys with daily sampling. The detection of such features enables measurements of the planet-star mass ratio and projected separation with precision sufficient to constrain planetary formation scenarios and contribute to demographic studies of exoplanet populations across the Galaxy.

The complexity of binary lens modeling—with its seven-plus free parameters, multiply-peaked likelihood surfaces, and computationally expensive magnification calculations—motivates the machine learning approach developed in this thesis. Traditional χ^2 fitting methods become prohibitively slow when analyzing thousands of ongoing events in real time, particularly when attempting early classification before the full light curve morphology is revealed.

2.4. Machine Learning for Microlensing Classification

The application of machine learning to gravitational microlensing presents an opportunity to automate the identification of rare binary events within large datasets generated by current and next-generation surveys. Rather than relying on human inspection or

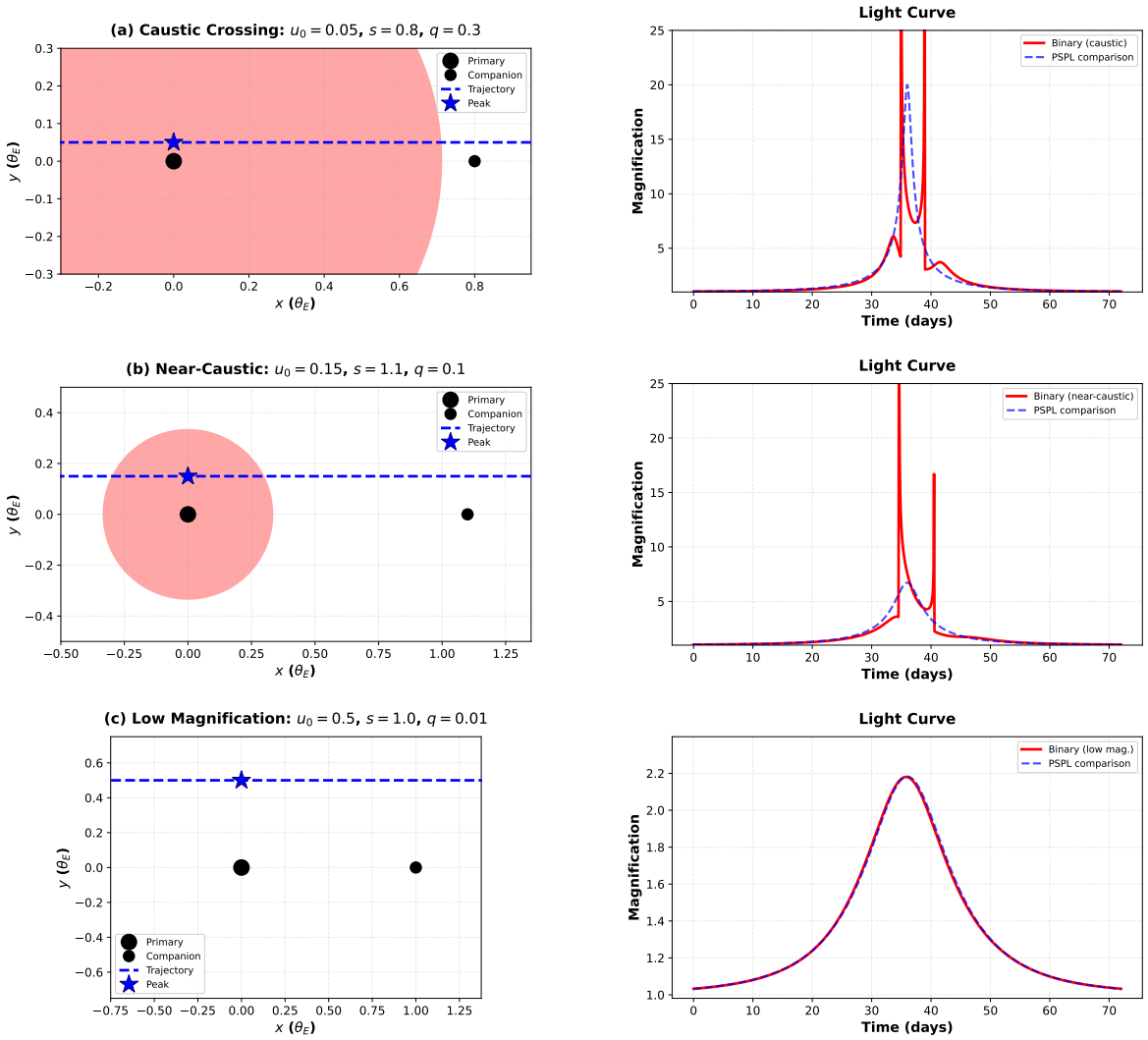

Figure 2.3.: Binary microlensing light curve morphologies and their caustic geometries. Each row shows a source trajectory (left) and the resulting light curve (right) for different binary parameters and impact parameters. Top: Caustic crossing produces sharp flux spike. Middle: Near-caustic approach creates asymmetric perturbation. Bottom: Distant trajectory yields nearly PSPL-like profile, illustrating the fundamental detection limit at large impact parameters.

computationally expensive model fitting, supervised classification methods can learn to recognize distinctive patterns in partially observed light curves, enabling real-time event prioritization.

2.4.1. The Classification Task

The problem addressed in this work is supervised classification: given a partially observed light curve (flux measurements at discrete time points), assign it to one of three categories—flat (no lensing event), PSPL (single-lens event), or binary (binary-lens event). The classifier must provide a probability distribution over these three classes, enabling downstream decision-making about resource allocation for follow-up observations [18]. Critically, the classifier must operate in real time, updating its classification as new observations arrive, using only past and present data.

This differs fundamentally from traditional model fitting approaches that attempt to estimate the continuous physical parameters (t_E, u_0, q, s, etc.). Instead of solving an inverse problem to recover lens properties, the classifier performs pattern recognition, identifying which class of magnification pattern best describes the observed light curve evolution. The classifier does not explicitly compute caustic positions, solve lens equations, or perform numerical ray-tracing; rather, it learns to recognize the temporal signatures that distinguish binary caustic crossings from PSPL symmetry through exposure to large numbers of training examples.

2.4.2. Neural Networks for Time Series

Neural networks are computational models composed of layers of parameterized functions (neurons) that transform input data through successive nonlinear operations [19]. Through a process called training, the network parameters (weights and biases) are adjusted via backpropagation and gradient descent to minimize prediction errors on labeled examples [20]. Modern deep learning architectures with multiple hidden layers can learn hierarchical representations, extracting increasingly abstract features from raw input data [21]. For time-series classification tasks like microlensing, architectures specifically designed to handle sequential data have proven particularly effective [22, 23].

Convolutional Neural Networks (CNNs) apply learnable filters that slide across the input sequence, detecting local patterns such as the rapid flux changes characteristic of caustic crossings [24, 25]. Convolutional layers are translation-invariant, meaning they can recognize features regardless of where they occur in the time series. Multiple convolutional layers with different filter sizes enable the network to detect patterns at multiple temporal scales, from the sharp sub-day features of planetary caustic crossings to the weeks-long evolution of the overall magnification envelope [26]. Pooling operations following convolutional layers provide spatial downsampling while preserving the most

salient features, improving computational efficiency and model robustness.

Recurrent Neural Networks (RNNs) maintain an internal hidden state that evolves as they process each observation sequentially, allowing them to remember features from earlier time steps and integrate information across the full sequence [27, 28]. The hidden state provides a form of memory that captures the cumulative context of the light curve as it develops. Gated variants such as Long Short-Term Memory (LSTM) units [29] and Gated Recurrent Units (GRUs) [30, 31] use learned gating mechanisms to selectively retain or forget information. These gates enable the network to capture both short-term patterns—such as a recent caustic crossing—and long-term structure—such as the overall symmetric or asymmetric shape of the light curve [32]. GRUs offer computational advantages over LSTMs while maintaining comparable representational capacity, making them well-suited for real-time inference applications [31].

The architecture employed in this work combines both approaches: convolutional layers extract local temporal features from the raw flux measurements, while recurrent GRU layers integrate these features across the full observation sequence, producing a final representation that is classified into one of the three event categories. This hierarchical structure mirrors the logical organization of the astrophysical problem—first detecting any lensing signal (distinguishing flat from non-flat), then distinguishing single from binary configurations based on the presence of caustic crossing signatures. The combination of CNN feature extraction with RNN sequence modeling has proven highly effective for astronomical time series classification across multiple domains [22, 26, 28].

2.4.3. Training on Synthetic Data

A critical advantage of machine learning for microlensing is that training data can be generated synthetically with perfect ground truth labels. Using physics-based simulation codes such as VBBinaryLensing [17], we can generate unlimited light curves spanning the full parameter space of lens masses, separations, trajectories, and observational conditions. Each synthetic event is labeled with its true class (flat, PSPL, or binary), providing the supervised training signal required for the network to learn discriminative features. This approach circumvents the label scarcity problem that limits many astronomical machine learning applications where labeled training sets must be painstakingly curated by human experts [18].

Synthetic training data enables systematic exploration of rare configurations and boundary cases that may be underrepresented in observed datasets. For planetary mass ratios $q \sim 10^{-4}$, caustic features become extremely small, requiring close source-lens alignments ($u_0 \ll 0.1$) for detection. By deliberately oversampling these challenging regimes during training, the network can learn to recognize subtle signatures that would be difficult to identify through traditional parameter fitting. The use of synthetic data

also allows controlled studies of classifier performance as a function of observational quality, temporal sampling, and photometric precision [33, 34].

The challenge lies in ensuring that synthetic training data accurately represents the statistical properties of real survey populations. Parameter sampling strategies must balance two competing objectives: generating enough examples with clear binary signatures to enable the network to learn caustic crossing patterns, while also including the realistic mix of high-impact-parameter events where binary features are subtle or absent. Domain adaptation techniques and transfer learning approaches [35] can help address potential distribution shifts between synthetic training data and real observations, though validation on actual survey data remains essential for operational deployment.

2.4.4. Optimization and Regularization

Training neural networks involves minimizing a loss function that measures prediction error on the training set. For classification tasks, the cross-entropy loss between predicted class probabilities and true labels provides the objective function. Optimization proceeds through stochastic gradient descent (SGD) and its variants, where network parameters are iteratively updated based on gradients computed via backpropagation [20]. Modern adaptive optimizers such as Adam [36] and AdamW [37] adjust learning rates on a per-parameter basis, accelerating convergence while improving generalization.

Regularization techniques prevent overfitting—where the network memorizes training examples rather than learning generalizable patterns. Dropout [38] randomly deactivates neurons during training, forcing the network to learn redundant representations and improving robustness. Batch normalization [39] standardizes layer inputs, stabilizing training and acting as an implicit regularizer. Weight decay (L2 regularization) penalizes large parameter values, encouraging simpler solutions that generalize better to unseen data. The methodology for hyperparameter selection and training procedures is detailed in Chapter 4.

2.4.5. Real-Time Processing and Calibration

Operational deployment in survey pipelines requires sub-millisecond inference times to process thousands of ongoing events as observations arrive. Modern GPU-accelerated neural network inference can achieve this performance through efficient matrix operations and optimized implementations [40, 41]. Unlike traditional fitting methods that require iterative optimization—often taking seconds to minutes per evaluation for complex binary models—neural network inference is a single forward pass through the network, providing a fixed computational cost independent of event complexity. This enables truly real-time classification as each new observation is obtained.

The classifier outputs continuously updated probabilities $P(\text{Flat})$, $P(\text{PSPL})$, and $P(\text{Binary})$ that evolve as the light curve develops. These probabilities can guide automated decisions: events with rising $P(\text{Binary})$ might trigger increased observing cadence or alert ground-based follow-up networks, while events with high $P(\text{Flat})$ can be deprioritized to conserve resources. Properly calibrated probability estimates—where predicted confidence levels match true classification accuracy [42]—are essential for reliable operational decision-making. Calibration can be assessed through reliability diagrams and expected calibration error metrics, ensuring that when the classifier reports 90% confidence, it is indeed correct 90% of the time.

2.5. Summary

Gravitational microlensing provides a unique window into Galactic populations of planets, stars, and stellar remnants through its mass-dependent sensitivity. Point-source point-lens events produce characteristic symmetric, three-parameter light curves that serve as the baseline template for microlensing observations. Binary lens systems introduce caustic and critical curve structures that create sharp temporal features and asymmetric patterns, enabling planetary detections and mass measurements but dramatically increasing the parameter space complexity.

The classification problem—distinguishing binary from single-lens events using partial observations—is well-suited to machine learning approaches that can learn to recognize caustic crossing signatures from large training datasets. The combination of convolutional feature extraction for local patterns and recurrent sequence modeling for long-term memory provides an architecture capable of real-time classification as observations arrive. Training on synthetic data generated with physics-based simulations enables comprehensive coverage of the parameter space while providing perfect ground truth labels.

The following chapters present the implementation of this approach, beginning with the synthetic data generation pipeline and neural network architecture (Chapter 4), proceeding to experimental results and validation across different population distributions (Chapter 5), and concluding with physical interpretation of the classification boundaries and operational implications for next-generation surveys (Chapter 6).

3. Literature Review

The automated classification of gravitational microlensing events represents a confluence of three distinct research areas: observational astronomy and survey operations, gravitational lensing theory and modeling, and machine learning for time-series analysis. This chapter provides a comprehensive survey of related work across these domains, establishing the scientific and technical context for the CNN-GRU hierarchical classifier developed in this thesis.

We begin by reviewing the history and current operational state of microlensing surveys (Section 3.1), focusing on their detection strategies, data products, and classification challenges. We then examine traditional approaches to binary microlensing classification, including χ^2 model fitting and Bayesian inference methods (Section 3.2), emphasizing their computational costs and observational requirements that limit real-time applicability. Next, we survey machine learning applications in time-domain astronomy (Section 3.3), covering both classical feature-based methods and recent deep learning architectures for light curve classification. Special attention is given to early classification problems in other astronomical transient domains (Section 3.4) where similar temporal challenges arise. We conclude by synthesizing these perspectives to identify the specific research gap addressed by this work: the absence of end-to-end sequential modeling approaches designed for real-time binary microlensing classification (Section 3.5).

3.1. Microlensing Surveys and Detection Strategies

3.1.1. Ground-Based Survey Operations

The modern era of gravitational microlensing began in the early 1990s with the establishment of two pioneering surveys: the Optical Gravitational Lensing Experiment (OGLE) and Microlensing Observations in Astrophysics (MOA). Both surveys monitor millions of stars in the Galactic bulge and Magellanic Clouds, searching for the characteristic symmetric brightness variations produced by lensing events [7, 8].

OGLE, operating from Las Campanas Observatory in Chile, has evolved through four generations of instrumentation, with OGLE-IV monitoring approximately 200 million stars across the Galactic bulge at cadences ranging from 20 minutes (high-priority

fields) to several days (low-priority fields) [7]. The survey discovers approximately 2,000 microlensing events annually, with several hundred showing anomalies potentially indicative of binary lens systems. OGLE's long baseline (over 30 years of operation) provides a vast archival dataset for testing classification algorithms, though the manual inspection process used to identify binary candidates introduces selection biases that are difficult to quantify retrospectively.

The Korea Microlensing Telescope Network (KMTNet), operational since 2015, employs three identical 1.6-meter telescopes distributed in longitude (Chile, South Africa, Australia) to provide near-continuous monitoring of the Galactic bulge [43]. This geographic distribution partially mitigates the diurnal cycle that limits single-site surveys, enabling KMTNet to detect short-duration planetary caustic crossings that might be missed by OGLE's cadence. KMTNet's higher spatial resolution also reduces blending effects in crowded fields, improving photometric precision.

A common feature across these ground-based surveys is the two-stage detection architecture: initial event identification through difference imaging analysis to detect flux variations above a significance threshold, followed by human expert review of candidate light curves to distinguish genuine microlensing from variables, transients, and systematics. This manual bottleneck, acceptable when processing hundreds of events per year, becomes increasingly untenable as surveys scale to thousands of simultaneous events.

3.1.2. Future Space-Based Observations

The Nancy Grace Roman Space Telescope, scheduled for launch in 2027, represents a qualitative leap in microlensing survey capabilities [13, 44]. Roman's Galactic Bulge Time-Domain Survey will monitor approximately 200 million stars with 15-minute cadence over 72-day seasons, achieving photometric precision significantly superior to ground-based facilities. The space-based platform eliminates weather interruptions, atmospheric seeing variations, and diurnal cycles, providing unprecedented temporal coverage [45].

Simulations predict Roman will detect approximately 27,000 microlensing events over its five-year prime mission, with thousands of events exhibiting binary lens signatures [13, 14]. This anticipated detection rate—more than an order of magnitude larger than current ground-based surveys—fundamentally changes the classification problem from one manageable through expert inspection to one requiring automated, real-time decision-making systems. The operational requirement for rapid classification stems from the need to coordinate follow-up observations: when Roman detects the onset of a caustic crossing event, ground-based networks must be alerted within hours to enable intensive monitoring during the scientifically valuable high-magnification phase.

3.2. Traditional Classification Approaches

3.2.1. Chi-Squared Model Fitting

The standard approach to microlensing event characterization relies on nonlinear least-squares fitting of physical models to observed light curves [3]. For single-lens events, the point-source point-lens (PSPL) model contains only three free parameters—the Einstein crossing time t_E, impact parameter u_0, and time of closest approach t_0—and can be fit efficiently using standard optimization algorithms.

Binary lens models present significantly greater computational challenges. The minimum parameterization requires seven quantities: the three PSPL parameters plus the mass ratio q, projected separation s, source trajectory angle α, and finite source size ρ when applicable [15]. The parameter space contains numerous local minima corresponding to qualitatively different binary topologies, and global optimization often requires Markov Chain Monte Carlo sampling to characterize the posterior distribution adequately.

The computational cost of binary model fitting scales poorly with data volume. Fitting a single binary model to a typical light curve requires minutes of computation using numerical ray-shooting codes such as VBBinaryLensing [17], and characterizing the posterior distribution across all seven parameters can require hours. When multiplied across thousands of ongoing events that must be refitted with each new observation, the computational burden becomes prohibitive for real-time survey operations. This computational bottleneck—rather than fundamental limitations of the fitting approach—motivates the exploration of machine learning alternatives that can provide rapid approximate classifications to triage which events warrant intensive modeling.

3.2.2. Bayesian Inference Methods

Several groups have developed Bayesian frameworks for microlensing classification that incorporate prior knowledge about lens populations and selection effects [46]. These approaches compute posterior probabilities over model classes (PSPL versus binary) given the observed photometry, naturally accounting for parameter degeneracies and providing uncertainty quantification.

While Bayesian methods offer principled probabilistic reasoning, they face the same computational challenges as χ^2 fitting: computing the model evidence for binary lens configurations requires marginalizing over the high-dimensional parameter space, which is computationally expensive. Approximate Bayesian computation and nested sampling algorithms can reduce these costs but still require minutes to hours per event for reliable results. For real-time survey operations requiring sub-second decisions on thousands of events, these timescales remain impractical.

3.2.3. Expert Visual Inspection

Current surveys rely heavily on human expert judgment to identify binary events. Experienced observers can often recognize subtle asymmetries, secondary peaks, or caustic crossing features that indicate binary structure, even in noisy or incomplete light curves. However, this approach suffers from several fundamental limitations: it is labor-intensive and does not scale to Roman-era detection rates, introduces subjective biases that vary between observers and over time, lacks quantitative uncertainty estimates needed for resource allocation decisions, and cannot provide the millisecond-scale response times required for automated alert systems.

These limitations motivate the development of automated classification systems that can replicate expert performance while providing consistent, rapid, and well-calibrated probabilistic outputs.

3.3. Machine Learning in Time-Domain Astronomy

3.3.1. Classical Feature-Based Approaches

Early applications of machine learning to astronomical time-series classification relied on engineered features extracted from light curves. Random forest classifiers, trained on summary statistics such as variability indices, periodicity measures, and shape parameters, achieved human-level performance on variable star classification tasks [22, 47].

For microlensing specifically, Godines et al. [33] developed a random forest classifier using 47 features computed from OGLE light curves, achieving approximately 94% accuracy in distinguishing microlensing events from other transient phenomena. Similarly, Khakpash et al. [34] designed 15 features intended to capture binary signatures—measures of peak smoothness, asymmetry metrics, and peak-counting statistics—and demonstrated promising results on simulated Roman data.

While these feature-based approaches provide interpretable models and require modest computational resources, they suffer from intrinsic limitations. The features must be hand-crafted by domain experts who anticipate relevant morphological signatures, but the morphological diversity of binary light curves means no fixed feature set can capture all possible binary configurations. Features computed from partial light curves may behave differently than those from complete events, requiring careful recalibration for real-time applications. The feature engineering process discards potentially useful information present in the raw photometric time series.

3.3.2. Deep Learning for Light Curve Classification

The emergence of deep learning offers an alternative paradigm: rather than engineering features manually, learn optimal representations directly from raw time-series data. Convolutional neural networks (CNNs) applied to astronomical light curves have demonstrated the viability of this approach across multiple domains.

For exoplanet transit detection, Shallue and Vanderburg [26] trained a CNN on Kepler light curves and achieved human-expert-level accuracy in identifying planetary transit signatures, including the discovery of previously overlooked planetary candidates. The CNN learned to recognize transit shapes without explicit programming of duration, depth, or periodicity features.

In supernova classification, recurrent neural networks have proven particularly effective. The SuperNNova framework [32] employs a bidirectional long short-term memory (LSTM) architecture trained on multi-band supernova photometry, achieving 97% accuracy in distinguishing Type Ia from non-Ia supernovae. Critically, SuperNNova can classify events using only partial light curves observed up to maximum brightness, demonstrating that recurrent architectures can learn temporal dependencies relevant for early-time classification.

For variable star classification on irregularly sampled data, Naul et al. [28] developed an autoencoding RNN that explicitly models observation times and heteroskedastic uncertainties, achieving competitive performance with feature-based methods while learning representations that transfer across different surveys.

3.3.3. Applications to Microlensing

Deep learning applications to microlensing remain limited compared to other time-domain phenomena. The most relevant precedent is the work of Mróz et al. [9], who applied simple feed-forward neural networks to binary microlensing classification. However, their approach required pre-computed PSPL model fits as input features, reintroducing the computational bottleneck of traditional methods. The networks achieved approximately 80–85% accuracy on known binary events but could not operate on raw photometry.

More recently, Ishitani Silva Terra et al. [48] developed a CNN for automated detection of microlensing events in the MOA survey, processing raw photometric time series without requiring model fits. Their system achieves millisecond-scale inference and discovered events missed by traditional pipelines. However, the CNN focuses on the detection problem (microlensing versus non-microlensing) rather than the more challenging task of distinguishing binary from single-lens events among genuine microlensing light curves.

The absence of end-to-end binary classification systems in the literature represents a

significant gap, particularly given the operational requirements for Roman-era surveys.

3.4. Early Classification in Other Transient Domains

The challenge of classifying astronomical transients from incomplete observations arises across multiple domains, providing instructive parallels for microlensing classification.

3.4.1. Supernova Photometric Typing

Supernova classification from photometric data alone (without spectroscopy) presents temporal challenges analogous to microlensing: observers must distinguish between spectroscopic types using only brightness evolution in multiple filters. Early classification is valuable for triggering spectroscopic follow-up while objects are bright and observable.

The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC) established benchmarks for early supernova classification [18]. Winning solutions combined recurrent neural networks with boosted decision trees, achieving high accuracy (>90%) even when light curves contained only a few observations near maximum brightness. Key lessons include: preprocessing strategies that handle irregular sampling and missing data significantly impact performance, uncertainty quantification through Bayesian methods or Monte Carlo dropout helps guide follow-up decisions, and augmentation techniques that simulate different cadences improve robustness.

3.4.2. Transient Identification in LSST

The upcoming Legacy Survey of Space and Time (LSST) will generate millions of transient alerts nightly, requiring automated classification to identify scientifically valuable targets for follow-up [49]. Broker systems under development employ machine learning classifiers trained on simulated LSST photometry to categorize alerts into supernova types, tidal disruption events, active galactic nuclei, and other classes.

A common challenge across these systems is handling the sequential nature of incoming data: each night brings new observations that should refine earlier classifications. Recurrent architectures provide a natural framework for this sequential updating, maintaining an internal state that evolves as new data arrives [28].

3.4.3. Lessons for Microlensing Classification

These transient classification efforts provide several insights relevant to microlensing:

1. **Sequential processing architectures** (RNNs, LSTMs, GRUs) naturally handle variable-length time series and can update classifications as observations arrive, making them well-suited for real-time applications.

2. **Hybrid architectures** combining CNNs for local feature extraction with RNNs for long-term memory have proven effective across multiple domains, suggesting this design may transfer to microlensing.

3. **Synthetic training data** is essential when real labeled examples are scarce or biased by selection effects. Physics-based simulation enables generation of unlimited training examples spanning the full parameter space.

4. **Calibrated uncertainty estimates** are critical for operational deployment. Models must not only predict class labels but also provide confidence levels that accurately reflect true prediction accuracy.

5. **Computational efficiency** is non-negotiable for real-time survey operations. Inference times must be milliseconds, not seconds, to process thousands of events as observations arrive.

3.5. The Research Gap

The literature review reveals a clear gap at the intersection of microlensing classification and sequential deep learning: *there exists no end-to-end neural network architecture that processes raw photometric time series to classify binary microlensing events in real time as observations arrive.*

Existing approaches fall into two categories, each with fundamental limitations:

Traditional methods (χ^2 fitting, Bayesian inference, expert inspection) can achieve high accuracy given complete light curves and sufficient computational resources, but they scale poorly to Roman-era detection rates. The computational cost of binary model fitting prohibits real-time classification of thousands of simultaneous events, and manual inspection introduces subjective biases while requiring unsustainable human effort.

Machine learning approaches demonstrated in other time-domain contexts (supernova classification, variable star typing, exoplanet detection) have established that neural networks can learn complex temporal patterns from raw data and provide rapid inference suitable for alert streams. However, applications to microlensing have focused either on event detection (distinguishing microlensing from non-microlensing) or have required pre-computed model parameters as features, failing to provide the end-to-end, real-time classification capability needed for Roman operations.

The specific requirements for Roman-era binary microlensing classification are:

1. **Raw photometry input**: The classifier must process flux measurements and observation times directly, without requiring computationally expensive model fits as intermediate features.

2. **Sequential processing**: The architecture must handle variable-length light curves and support real-time classification as observations arrive, not just final classification using complete events.

3. **Hierarchical discrimination**: The model must first detect genuine microlensing events (any magnification), then distinguish binary from single-lens configurations—two fundamentally different tasks requiring specialized representations.

4. **Computational efficiency**: Inference times must be sub-millisecond per event to enable real-time processing of thousands of simultaneous events in Roman alert streams.

5. **Calibrated uncertainties**: The system must provide well-calibrated probability estimates, where reported confidence levels accurately reflect true prediction accuracy, to guide follow-up resource allocation.

6. **Physical interpretability**: Performance degradation at large impact parameters should reflect genuine physical detection limits (binary events becoming indistinguishable from PSPL events) rather than algorithmic failures.

This thesis addresses this gap through a CNN-GRU hierarchical architecture trained on physically accurate synthetic data. The convolutional layers extract local temporal features (caustic crossings), the GRU layers maintain long-term memory across the 72-day observing season, and the hierarchical classification structure mirrors the logical decomposition of the problem (event detection, then binary discrimination). The complete system achieves sub-millisecond inference on typical GPU hardware, enabling deployment in operational survey pipelines.

3.6. Summary

The classification of binary microlensing events exists at the intersection of mature observational techniques, well-understood physical models, and emerging machine learning capabilities. Ground-based surveys have demonstrated that binary events produce detectable signatures in photometric time series, but current classification methods—whether traditional model fitting or expert visual inspection—will not scale to the Roman Space Telescope's anticipated detection rates.

Machine learning applications in other time-domain astronomy domains have established that neural networks can learn complex temporal patterns from raw data, achieve human-level classification performance, and provide the millisecond-scale inference required for real-time alert processing. Sequential architectures combining convolutional feature extraction with recurrent memory have proven particularly effective for variable-length time series with irregular sampling.

However, a critical gap remains: no existing system provides end-to-end binary microlensing classification directly from raw photometry with the computational efficiency and hierarchical structure needed for Roman operations. The following chapters describe our approach to filling this gap through a CNN-GRU hierarchical classifier trained on synthetic data, systematically evaluated across completeness levels and physical parameters, and benchmarked against traditional methods to quantify performance trade-offs and establish the operational capabilities needed for next-generation surveys.

4. Methodology

4.1. Overview

The Nancy Grace Roman Space Telescope will observe the Galactic bulge with unprecedented cadence and duration, detecting an estimated 27,000 microlensing events over its five-year prime mission [13]. Unlike ground-based surveys where manual inspection and traditional χ^2 model fitting remain feasible, Roman's high event rate demands automated real-time classification to identify scientifically valuable binary events while observations are ongoing. This capability is essential for coordinating time-sensitive follow-up observations with ground-based networks, which can provide complementary high-resolution spectroscopy and intensive photometric monitoring during critical caustic crossing phases.

We present a machine learning classifier designed specifically for Roman-like observations: 15-minute cadence over 72-day seasons with realistic photometric uncertainties. The classifier processes variable-length light curves and outputs continuously updated probabilities for three event classes—flat (no lensing), single-lens (PSPL), and binary—enabling dynamic decision-making as events unfold. The complete pipeline consists of four components: physically accurate synthetic data generation using VBBinaryLensing [17], a compact neural network architecture combining local feature detection with long-term memory, distributed training across GPU clusters, and comprehensive validation including cross-evaluation studies and impact parameter dependency analysis.

4.2. Synthetic Data Generation

Training a classifier for microlensing requires large datasets with known ground truth labels. Real observations present two fundamental challenges: binary events with well-characterized parameters are rare, and even extensively modeled events have uncertain lens configurations. We address both limitations through physics-based simulation, generating unlimited training examples with perfect ground truth while maintaining full control over the parameter distributions the model encounters.

4.2.1. Light Curve Generation

All synthetic light curves are generated using a combination of analytical formulas and VBBinaryLensing [17], a numerical ray-shooting code that computes gravitational magnification patterns for binary lens configurations. The simulation framework handles three distinct event types: flat (non-lensing), PSPL (single-lens), and binary (binary-lens), each with appropriate physical modeling.

Flat Events: Non-lensing light curves represent the baseline photometric noise without magnification effects. These are generated by setting constant magnification $A(t) = 1$ across the 72-day observation window. Source baseline flux F_0 is sampled uniformly from [1000, 5000] detector counts for realistic photometric noise modeling. While flat events are assigned impact parameters u_0, Einstein timescales t_E, and peak times t_0 for dataset consistency, these parameters do not affect the magnification since $A(t) = 1$ throughout.

PSPL Events: Single-lens events are computed using the analytical point-source point-lens magnification formula of Paczyński (1986):

$$A_{\mathrm{PSPL}}(t) = \frac{u(t)^2 + 2}{u(t)\sqrt{u(t)^2 + 4}}, \quad u(t) = \sqrt{u_0^2 + \left(\frac{t - t_0}{t_E}\right)^2}, \tag{4.1}$$

where u_0 is the impact parameter (minimum lens-source separation in Einstein radius units), t_0 is the time of closest approach, and t_E is the Einstein crossing time. This formula assumes a point source with negligible angular size, providing an efficient analytical computation that produces the characteristic symmetric magnification peak distinguishing PSPL events from both flat (no magnification) and binary (asymmetric caustic features) events. The observed detector flux (in counts) is $F(t) = F_0 \cdot A_{\mathrm{PSPL}}(t)$, though the neural network receives the normalized magnification $A(t)$ as input, not raw flux values.

Binary Events: For binary lenses, no closed-form solution exists; VBBinaryLensing employs adaptive contour integration to solve the lens equation numerically, capturing the complex caustic structures that produce the distinctive sharp features in binary light curves. Binary events include all PSPL parameters (u_0, t_E, t_0) plus four additional parameters: mass ratio q defining the companion mass relative to the primary, separation s in Einstein radii, source radius ρ governing finite source effects during caustic crossings, and trajectory angle α determining the source's path relative to the lens geometry. Unlike PSPL events which treat the source as a point, binary simulations account for the extended source size through the ρ parameter, essential for accurate modeling of caustic

Table 4.1.: Shared parameters for all synthetic event types (Flat, PSPL, and Binary).

Parameter	Symbol	General	Distinct[†]
Impact parameter	u_0	[0.001, 1.0]	[0.001, 0.3]
Einstein timescale	t_E (days)	[3, 18]*	[3, 18]*
Peak time	t_0 (days)	[18, 54]‡	[18, 54]‡
Baseline flux[††]	F_0 (counts)	[1000, 5000]	[1000, 5000]

*Log-uniform sampling. Range restricted to 3–18 days to ensure events fit within 72-day season.
†Distinct distribution restricts u_0 to enhance lensing signatures.
‡Peak time constrained to middle 50% of observing window (days 18–54 of 72-day season) to ensure complete event coverage with sufficient pre- and post-peak baseline for accurate classification.
††Source baseline flux used for detector noise modeling. Model input receives normalized magnification $A = F/F_0$, not raw flux.

Table 4.2.: Additional parameters for binary lens events only. PSPL events use only the shared parameters from Table 4.1 with the analytical point-source formula.

Parameter	Symbol	General	Distinct
Mass ratio	q	[0.001, 1.0]*	[0.1, 1.0]
Separation	s (Einstein radii)	[0.3, 3.0]	[0.8, 1.2]
Source radius	ρ	[0.001, 0.01]	[0.001, 0.01]
Trajectory angle	α (radians)	[0, 2π]	[0, 2π]

*Log-uniform sampling for mass ratio spanning planetary ($q \sim 10^{-3}$) to equal-mass ($q = 1.0$) regimes.

crossing features where magnification gradients are steep. The magnification pattern $A_{\text{Binary}}(t, \rho)$ exhibits sharp spikes when the source crosses caustic curves, creating the temporal signatures that enable binary detection.

Each simulation represents a 72-day observing season with 15-minute cadence, yielding approximately 6,912 observations per event. We sample physical parameters from distributions designed to represent the expected Roman event population, detailed in Tables 4.1–4.2. Impact parameters range uniformly from $u_0 = 0.001$ to $u_0 = 1.0$, Einstein timescales follow a logarithmic distribution from 3 to 18 days (restricted to ensure events fit within the 72-day season with sufficient pre- and post-peak baseline), and binary parameters span wide ranges: mass ratios from $q = 0.001$ (planetary) to $q = 1.0$ (equal mass), separations from $s = 0.3$ to $s = 3.0$ Einstein radii, and finite source radii from $\rho = 0.001$ to $\rho = 0.01$.

4.2.2. Observational Realism

Synthetic light curves must replicate three key observational effects. First, temporal sampling follows Roman's 15-minute cadence with 5% random gaps simulating downlink windows, pointing interruptions, and scheduling constraints [45]. Second, photometric noise follows a signal-dependent model accounting for both photon statistics and sky

background:

$$\sigma_{\text{flux}} = \sqrt{\frac{F(t) + F_{\text{sky}}}{G}} + \sigma_{\text{sky}}, \tag{4.2}$$

where $G = 10$ electrons/ADU is the detector gain and $\sigma_{\text{sky}} = 50$ ADU represents sky background noise. This ensures bright magnification peaks have higher signal-to-noise ratios than baseline flux, matching real observations. Third, each light curve is stored as a sequence of magnification measurements and time intervals between consecutive valid observations. Missing observations (indicated by zero-valued magnification in the raw data files) are removed during preprocessing—only valid observations are passed to the model. The time intervals encode both the nominal observing cadence and temporal gaps, allowing the model to distinguish rapid flux variations from slow evolutionary changes and to account for discontinuities in coverage.

4.2.3. Training Set Design

We employ a dual-distribution sampling strategy to balance two competing needs. The *general* distribution samples the full parameter space, generating events representative of the expected Roman population including many high-impact-parameter cases where binary signatures are subtle or absent. The *distinct* distribution deliberately restricts parameters to enhance binary features: limiting $u_0 < 0.3$ ensures source trajectories pass close to caustic structures, and constraining $s \in [0.8, 1.2]$ concentrates on separations where caustic crossings produce the sharpest flux variations.

Final training data comprise 600,000 light curves (200,000 per class), balanced between general and distinct distributions to expose the model to both realistic populations and strong feature examples. Independent test sets (100,000 events per class) drawn from each distribution enable cross-evaluation studies examining generalization across parameter regimes.

4.3. Neural Network Architecture

The classifier must satisfy three operational requirements emerging from real-time survey constraints: sub-millisecond inference to process thousands of events as observations arrive, strictly causal processing using only past and present data (no future peeking), and hierarchical outputs reflecting that detecting any lensing signal represents a fundamentally different task than distinguishing subtle binary features. These requirements guided the design of a compact architecture combining convolutional feature extraction for local pattern recognition with recurrent sequence modeling for long-term memory.

Architecture Selection Rationale: The CNN-GRU design reflects specific properties of microlensing light curves and operational constraints. Caustic crossings are

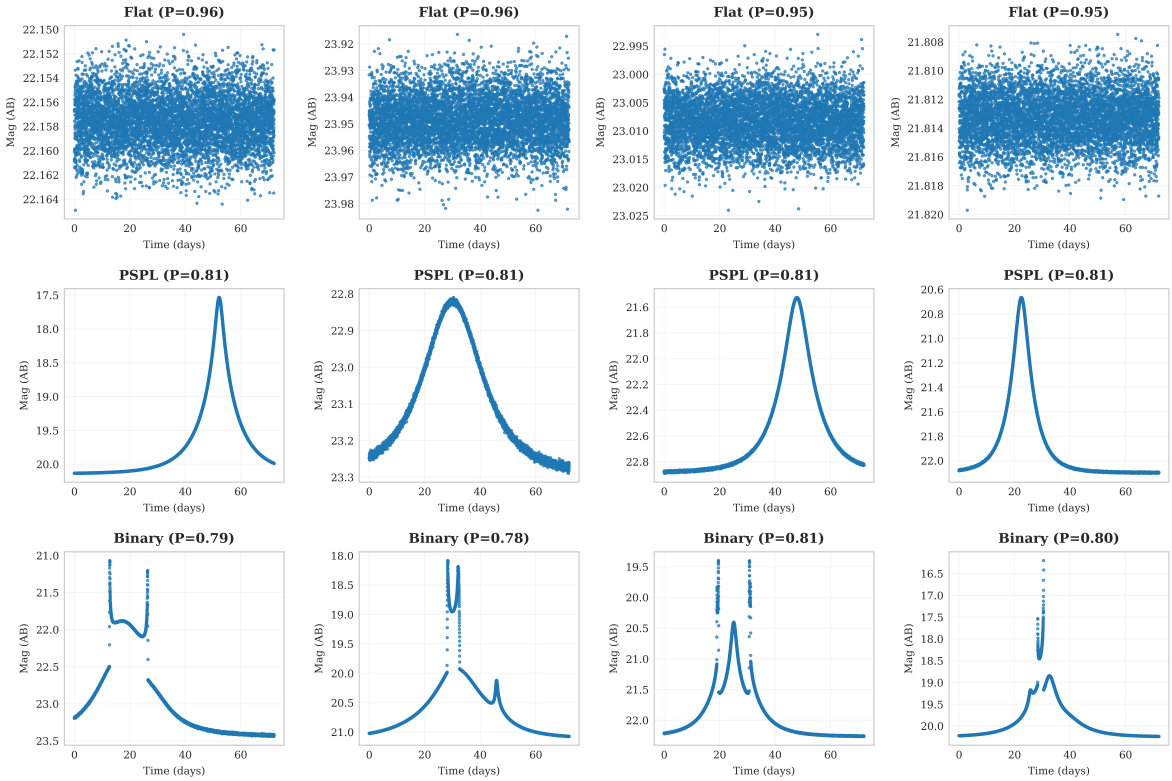

Figure 4.1.: Example synthetic light curves for each event class. Flat events show constant magnification ($A = 1$) with photometric noise. PSPL events exhibit symmetric magnification peaks computed from the analytical Paczyński formula. Binary events show sharp caustic crossing features computed via VBBinaryLensing with finite source effects, distinguishing them from the smooth PSPL profile.

localized temporal features spanning hours to days, making convolutional layers natural for exploiting this temporal locality through weight sharing—the architectural inductive bias that similar patterns occur at different times regardless of when in the observing season. However, distinguishing event types requires understanding the overall light curve shape across the full 72-day season, necessitating recurrent layers to maintain long-term memory of features like early caustic crossings when making decisions based on late-time observations. We chose GRUs over Transformer architectures because: (1) our training data is limited (600,000 events across three classes), benefiting from architectural inductive bias rather than learning all temporal relationships from scratch, (2) strict causality is essential for real-time processing, which self-attention mechanisms do not naturally enforce, and (3) computational efficiency enables the sub-millisecond inference required to process Roman's expected alert stream of thousands of concurrent events.

4.3.1. Input Representation

Each light curve enters the model as a variable-length sequence of observations, where each timestep contains two normalized values: magnification and time interval. Magnification $A(t)$ represents the gravitational lensing strength ($A = 1.0$ for baseline, $A > 1.0$ for magnified), computed as the ratio of observed flux to baseline flux: $A = F(t)/F_0$. Time intervals Δt record the elapsed time in days since the previous valid observation, encoding both Roman's nominal 15-minute cadence ($\Delta t \approx 0.0104$ days) and observational gaps due to moon phases, field rotation, or downlink windows.

Missing observations are removed during preprocessing via sequence compaction—only valid observations are passed to the model. Time intervals implicitly encode gaps: values near the nominal cadence indicate continuous observing, while larger values ($\Delta t > 0.05$ days) indicate interruptions. This encoding is more efficient than explicit binary masking and provides the model with direct temporal context about gap durations.

Both inputs are z-score normalized using training set statistics:

$$A_{\text{norm}} = \frac{A - \mu_A}{\sigma_A}, \quad \Delta t_{\text{norm}} = \frac{\Delta t - \mu_{\Delta t}}{\sigma_{\Delta t}} \tag{4.3}$$

where typical values are $\mu_A \approx 1.2$, $\sigma_A \approx 1.5$, $\mu_{\Delta t} \approx 0.015$ days, $\sigma_{\Delta t} \approx 0.02$ days. The normalized pair $(A_{\text{norm}}, \Delta t_{\text{norm}})$ at each timestep forms a 2-dimensional input vector.

A learned linear projection maps these two channels to the model's hidden dimension:

$$\mathbf{h}_0 = \mathbf{W}_{\text{proj}} \begin{bmatrix} A_{\text{norm}} \\ \Delta t_{\text{norm}} \end{bmatrix} + \mathbf{b}_{\text{proj}} \tag{4.4}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{32\times 2}$ projects to $d_{\text{model}} = 32$ dimensions.

4.3.2. Local Feature Extraction

Caustic crossings—the key signature distinguishing binary from single-lens events—typically evolve over hours. To detect these rapid features while maintaining computational efficiency, the architecture employs *depthwise separable convolutions* [50, 51] rather than standard convolutions. This architectural choice reduces parameters by approximately 90% while maintaining the same receptive field. Standard 1D convolution applies $d_{\text{in}} \times d_{\text{out}} \times k$ learnable filters, requiring $O(d_{\text{in}} \cdot d_{\text{out}} \cdot k)$ parameters. Depthwise separable convolution factorizes this operation into: (1) a depthwise layer applying one k-length filter per input channel independently ($d_{\text{in}} \times k$ parameters), and (2) a pointwise layer mixing channels via 1×1 convolution ($d_{\text{in}} \times d_{\text{out}}$ parameters). The total parameter count becomes $d_{\text{in}}(k + d_{\text{out}})$ versus $d_{\text{in}} \cdot d_{\text{out}} \cdot k$ for standard convolution. For our architecture with $d_{\text{model}} = 32$ and kernel size $k = 5$, this yields 1,184 parameters per layer versus 5,120 for standard convolution—a 77% reduction. This efficiency enables our complete model to contain only 33,541 parameters, three orders of magnitude smaller than typical deep networks, while achieving sub-millisecond inference suitable for processing thousands of concurrent events.

We use two convolutional blocks with different temporal scales: the first examines patterns over 5 observations (1.25 hours at 15-minute cadence) with dilation=1, while the second uses a dilated filter (dilation=2) examining patterns over 9 observations (2.25 hours). Together, these provide an effective temporal receptive field of 13 observations (3.25 hours), sufficient to capture most caustic crossing durations while remaining compact enough for rapid processing.

Critically, the convolutions are strictly *causal*: at any timestep t, the output depends only on observations up to and including t, never future data. This ensures the model can process light curves in real time, updating its classification as each new observation arrives without requiring the complete event.

4.3.3. Sequence Memory

While local features capture caustic crossings, distinguishing event types also requires understanding the overall light curve shape across the full 72-day season. A recurrent network maintains a persistent hidden state that evolves as it processes each observation sequentially, allowing it to remember features like caustic crossings that occurred days before the current timestep. The architecture uses Gated Recurrent Units (GRUs) [30, 31], which balance memory capacity against computational efficiency through learned gating mechanisms that selectively retain or discard information.

Four GRU layers are stacked, with each layer's output feeding the next. Dropout regularization [38] (30% probability) prevents the model from over-relying on specific pathways, encouraging robust features that generalize to unseen data. The complete

recurrent stack maintains both short-term memory of recent observations and long-term memory of features like early-season caustic crossings relevant for final classification decisions.

4.3.4. Sequence Aggregation

After processing the complete sequence, the model must condense the variable-length temporal representation into a fixed-dimensional feature vector for classification. We employ multi-head attention pooling, which learns to weight different timesteps according to their relevance for the classification decision. This mechanism can focus on critical moments like caustic crossings while downweighting baseline periods, achieving better performance than simple mean pooling across the sequence.

4.3.5. Hierarchical Classification

The classification head implements a two-stage decision process mirroring the logical structure of the problem. The model implements this hierarchy through two sequential binary classification heads that operate on the pooled sequence representation:

Stage 1 (Event Detection): A binary classifier with learnable weights $\mathbf{w}_1 \in \mathbb{R}^{32}$ outputs a scalar logit $z_1 = \mathbf{w}_1^T \mathbf{h}$, where $\mathbf{h}$ is the pooled feature vector. This logit is transformed via the sigmoid function to yield the deviation probability:

$$P(\text{Non-Flat}) = \sigma(z_1) = \frac{1}{1 + e^{-z_1}} \tag{4.5}$$

Stage 2 (Type Classification): A second binary classifier with weights $\mathbf{w}_2 \in \mathbb{R}^{32}$ outputs logit $z_2 = \mathbf{w}_2^T \mathbf{h}$. Applying sigmoid with temperature scaling ($T = 2.0$) yields the conditional PSPL probability:

$$P(\text{PSPL} \mid \text{Non-Flat}) = \sigma(z_2/T) = \frac{1}{1 + e^{-z_2/2.0}} \tag{4.6}$$

Both heads operate independently on the same shared representation, enabling separate gradient signals for each decision stage. The final three-class probabilities combine both stages via the product rule:

$$P(\text{Flat}) = 1 - P(\text{Non-Flat}) \tag{4.7}$$

$$P(\text{PSPL}) = P(\text{Non-Flat}) \times P(\text{PSPL} \mid \text{Non-Flat}) \tag{4.8}$$

$$P(\text{Binary}) = P(\text{Non-Flat}) \times P(\text{Binary} \mid \text{Non-Flat}) \tag{4.9}$$

This hierarchical structure prevents the dominant flat class from suppressing gradients needed to distinguish subtle binary features. An auxiliary classification head

provides additional training signal: it performs direct three-class supervision on the shared representation, preventing gradient instability during hierarchical training. This auxiliary output helps prevent "hierarchical collapse" where both stages converge to trivial solutions (e.g., always predicting non-flat in Stage 1), ensuring robust learning across all event types.

The complete architecture contains 33,541 trainable parameters (d_model=32, 4 GRU layers)—three orders of magnitude smaller than typical deep networks. This compact size enables sub-millisecond inference: a single forward pass on an NVIDIA A100 GPU processes one event in under 0.5 milliseconds, allowing real-time classification of thousands of events as observations arrive.

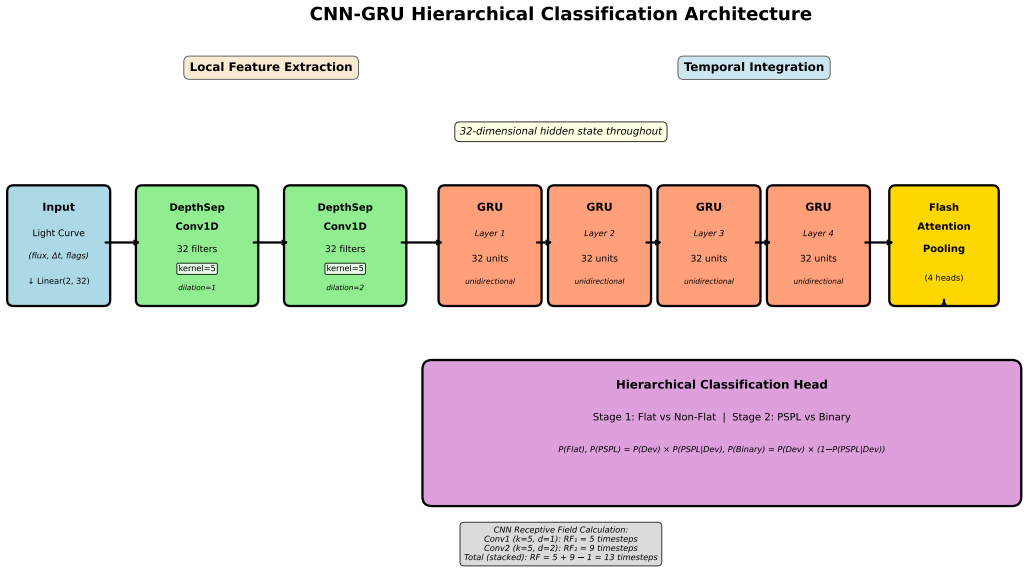

Figure 4.2.: Architecture of the hierarchical classifier. Two input channels (magnification and time intervals) are projected to 32 dimensions. Convolutional layers extract local temporal patterns including caustic crossings. Stacked GRU layers maintain memory across the 72-day season. Attention pooling aggregates the sequence. The hierarchical head performs two-stage classification: first detecting genuine microlensing, then distinguishing lens configurations.

4.4. Training

4.4.1. Loss Function Design

The training objective combines three loss terms with weights $\lambda_1, \lambda_2, \lambda_{\text{aux}}$:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{stage1}} + \lambda_2 \mathcal{L}_{\text{stage2}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}} \tag{4.10}$$

Stage 1 Loss ($\lambda_1 = 1.0$): Binary cross-entropy trains the event detection classifier,

distinguishing genuine lensing events from photometric noise. Class imbalance is handled via positive class weighting: $w_{\text{pos}} = (N_{\text{PSPL}} + N_{\text{Binary}})/(2N_{\text{Flat}})$.

Stage 2 Loss ($\lambda_2 = 0.5$): Binary cross-entropy trains the PSPL versus binary classifier, applied only to events where the true label is non-flat. This prevents incorrect gradient signals from flat events where the Stage 2 decision is meaningless. Positive class weighting adjusts for any PSPL/binary imbalance: $w_{\text{pos}} = N_{\text{PSPL}}/N_{\text{Binary}}$. The reduced weight (0.5 versus 1.0 for Stage 1) reflects that binary feature detection is inherently more challenging than flat event rejection.

Auxiliary Loss ($\lambda_{\text{aux}} = 0.3$): Direct three-class cross-entropy on the auxiliary head provides an additional gradient path, stabilizing hierarchical training and preventing collapse to trivial solutions. The reduced weight prevents this term from dominating the hierarchical objectives.

All losses use class-weighted forms to handle potential dataset imbalances. The final model outputs are well-calibrated: when the model reports 90% confidence, approximately 90% of such predictions prove correct—critical for operational decisions about resource-intensive follow-up observations. This calibration emerges from the training dynamics rather than an explicit calibration loss term.

4.4.2. Optimization and Training Infrastructure

Training employed the AdamW optimizer [37] with initial learning rate 5×10^{-4}, decaying following a cosine schedule to minimum 1×10^{-6} over 50 epochs. A 3-epoch warmup phase gradually increases the learning rate from zero, stabilizing early training. Mixed-precision arithmetic (16-bit floating point for most operations, 32-bit for critical computations) improves training throughput while maintaining numerical stability through gradient scaling [52].

The 600,000 training examples necessitated distributed computing. Training employed 24 to 48 NVIDIA A100 GPUs across multiple compute nodes, with each GPU processing microbatches of 256 events with gradient accumulation over 2 steps, yielding an effective per-GPU batch size of 512 events (global effective batch sizes 12,288–24,576 events). Data were loaded into RAM-backed filesystem (/dev/shm) on each node, reducing repeated disk reads after initial load. PyTorch's DistributedDataParallel framework handles gradient synchronization automatically: after each backward pass, gradients are averaged across all GPUs ensuring all model copies stay synchronized. Complete training required approximately 8 hours wall-clock time on 24 GPUs.

Regularization strategies prevent overfitting: dropout randomly disables 30% of GRU connections during training, weight decay (10^{-4}) applies L2 penalty discouraging large parameter values, and early stopping halts training if validation performance fails to improve for 10 consecutive epochs. Equal class proportions (200,000 events per

class) ensure balanced training despite real populations being dominated by non-lensing curves.

[Training convergence plot - requires actual training logs from runs]

Figure 4.3.: Training convergence for a representative model. Both losses decrease steadily through warmup and cosine decay phases. Close agreement between training and validation curves indicates good generalization.

4.5. Evaluation

4.5.1. Cross-Validation Strategy

Model performance is assessed on two independent test sets: a general set from the general distribution (100,000 events per class) and a distinct set emphasizing caustic crossings (100,000 events per class). We trained three model variants—general-trained (mixed general+distinct), general-trained (general only), and distinct-trained (distinct only)—and evaluated each on both test sets, producing four cross-evaluation scenarios examining how models trained on different parameter distributions generalize to different populations.

Primary metrics include overall accuracy (fraction correct), per-class precision and recall, F1-scores (harmonic mean of precision and recall), and area under the receiver operating characteristic curve (ROC-AUC, a threshold-independent measure of separability). For probabilistic predictions, we assess calibration using Expected Calibration Error (ECE): predictions are binned by confidence and the average predicted probability compared with actual fraction correct in each bin. Well-calibrated models show close agreement between predicted and empirical probabilities.

4.5.2. Impact Parameter Dependency

Special attention is devoted to performance versus impact parameter u_0 for binary events. Classification accuracy naturally degrades at large u_0 where binary features become physically subtle or absent, but quantifying this relationship reveals whether observed degradation represents physical detection limits or algorithmic failures. We plot accuracy versus u_0 and construct confusion matrices stratified by impact parameter, demonstrating that performance drops at $u_0 > 0.3$ reflect genuine physical constraints—binary events become indistinguishable from single-lens events at large separations regardless of algorithmic sophistication.

4.5.3. Real-Time Capability

The classifier's real-time capability is evaluated by truncating test light curves at various completeness fractions and recording accuracy using only partial observations. Early detection curves plot accuracy versus completeness, typically showing the model achieves 80–90% of final accuracy using only the first 30–50% of observations. This enables early triggering of follow-up observations while events are still developing, maximizing scientific return from coordinated ground-based campaigns. Reliability diagrams stratified by completeness verify the model remains well-calibrated even when making decisions from incomplete information.

[Cross-evaluation confusion matrices - see evaluation results]

Figure 4.4.: Cross-evaluation results. Diagonal elements show correct classifications; off-diagonal elements reveal where models trained on different parameter distributions succeed or struggle when evaluated on different populations.

4.6. Summary

This methodology combines physically accurate simulation with a compact hierarchical classifier designed for Roman's observational constraints. PSPL events use the analytical Paczyński formula for efficient computation, while binary events employ VBBinaryLensing with finite source effects to capture caustic crossing signatures. The dual-distribution training strategy balances exposure to both unambiguous caustic signatures and realistic population statistics, achieving robust performance across validation scenarios. The architecture's strict causality and sub-millisecond inference enable real-time classification as observations arrive, addressing the critical operational need for early identification of scientifically valuable binary events while they are still developing. Comprehensive cross-evaluation studies and impact parameter dependency analysis demonstrate that observed performance limitations at large u_0 reflect fundamental physical detection limits rather than algorithmic deficiencies.

5. Results

In this chapter, we present the comprehensive evaluation of the hierarchical CNN-GRU classifier. The following figures represent the complete set of generated metrics and event evolutions from the distinct test set evaluation.

5.1. Dataset Statistics

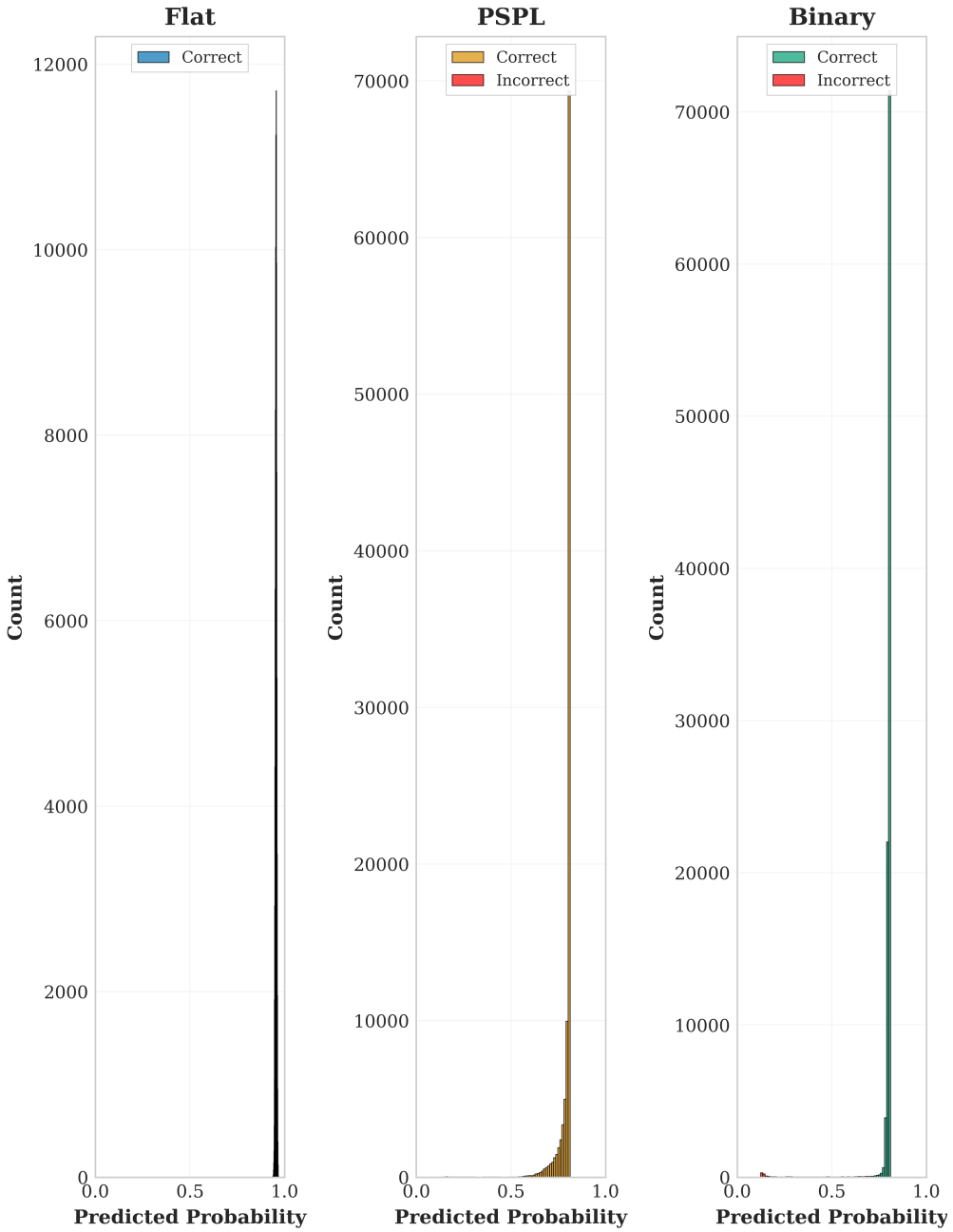

Figure 5.1.: Class distributions showing the balance of Flat, PSPL, and Binary events in the test set.

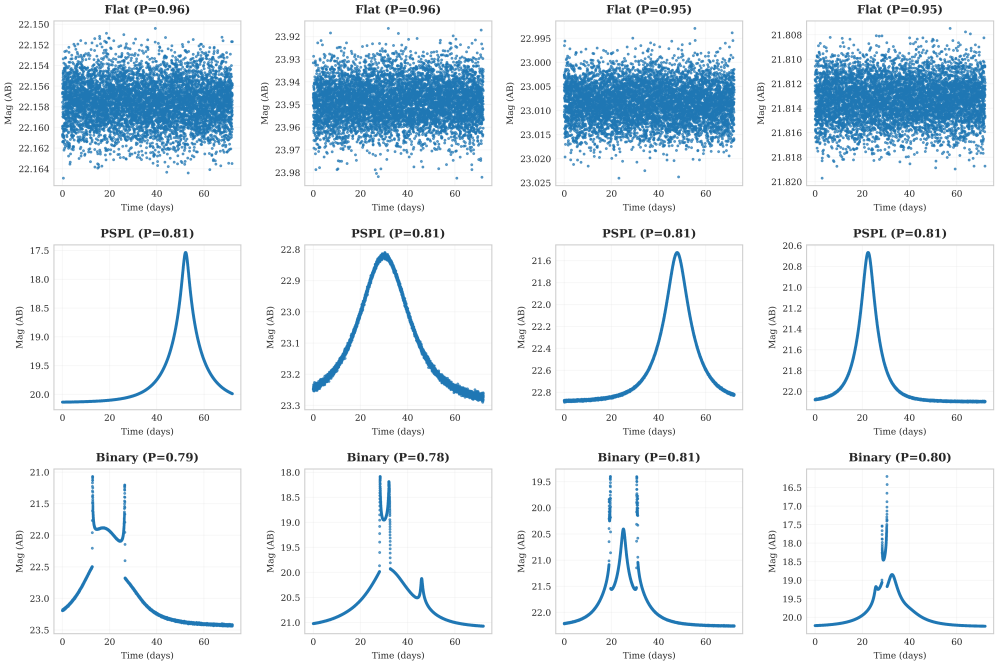

Figure 5.2.: Example light curves from the test set showing representative morphologies for each class.

5.2. Classification Metrics

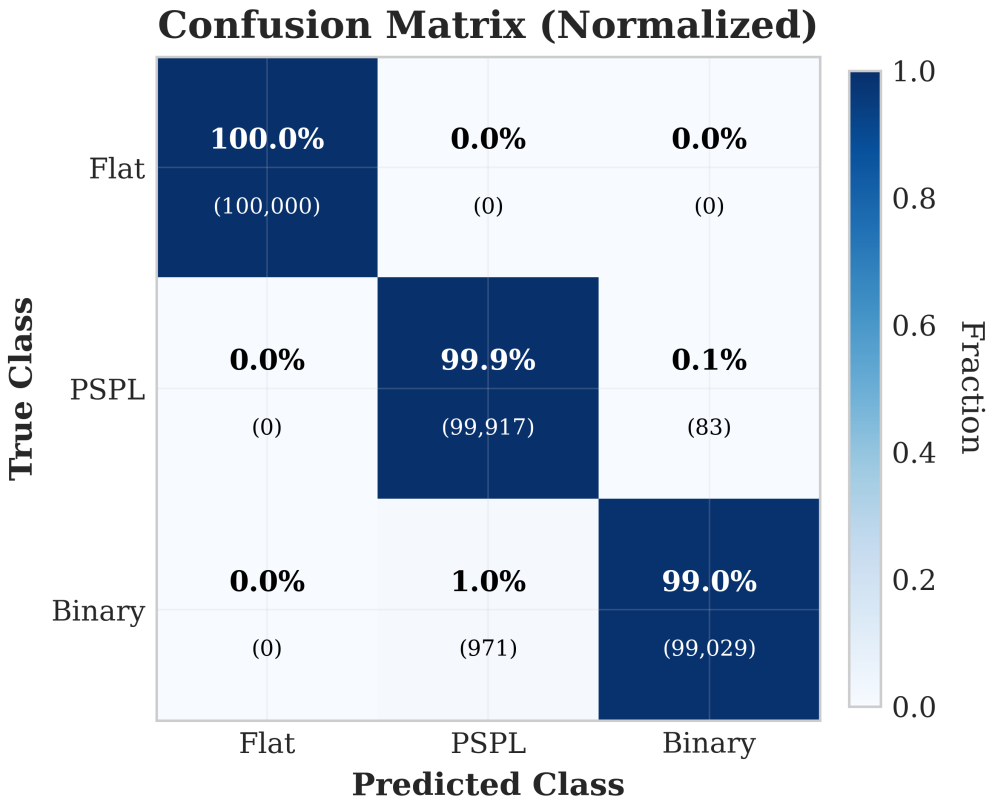

Figure 5.3.: Confusion matrix evaluated on the distinct test set.

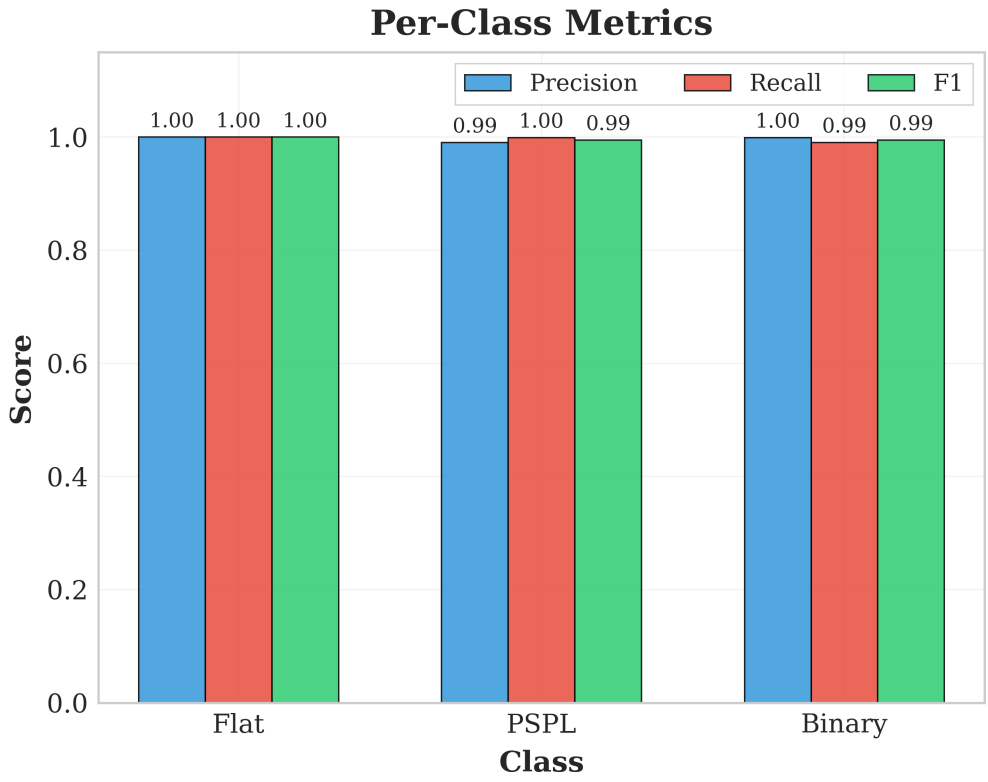

Figure 5.4.: Per-class performance metrics including Precision, Recall, and F1-Score.

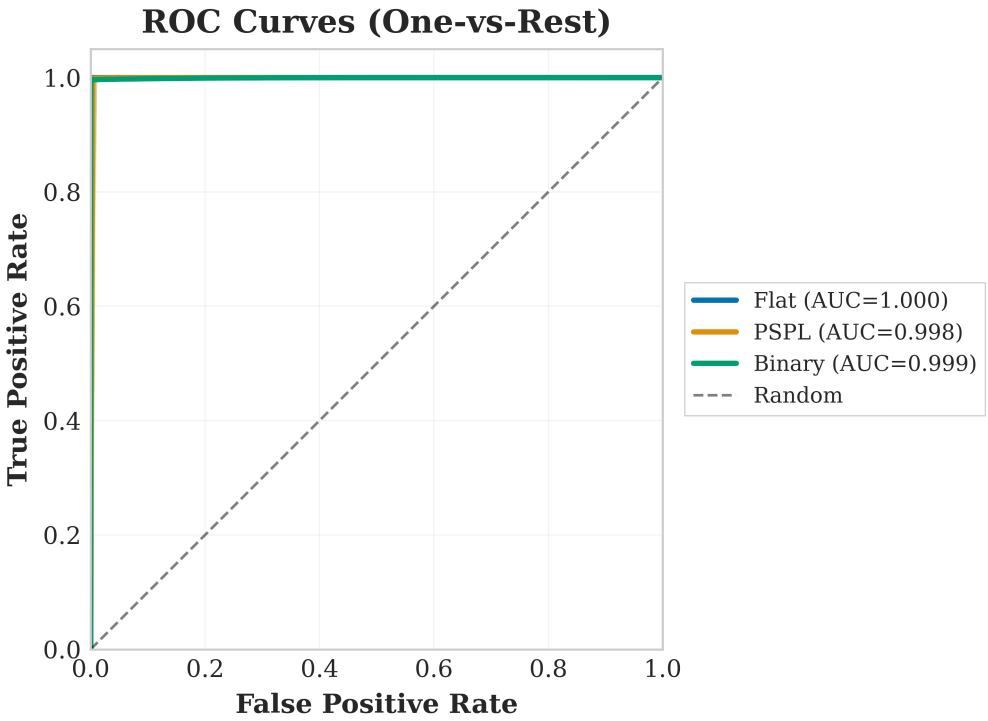

Figure 5.5.: Receiver Operating Characteristic (ROC) curves demonstrating discrimination performance.

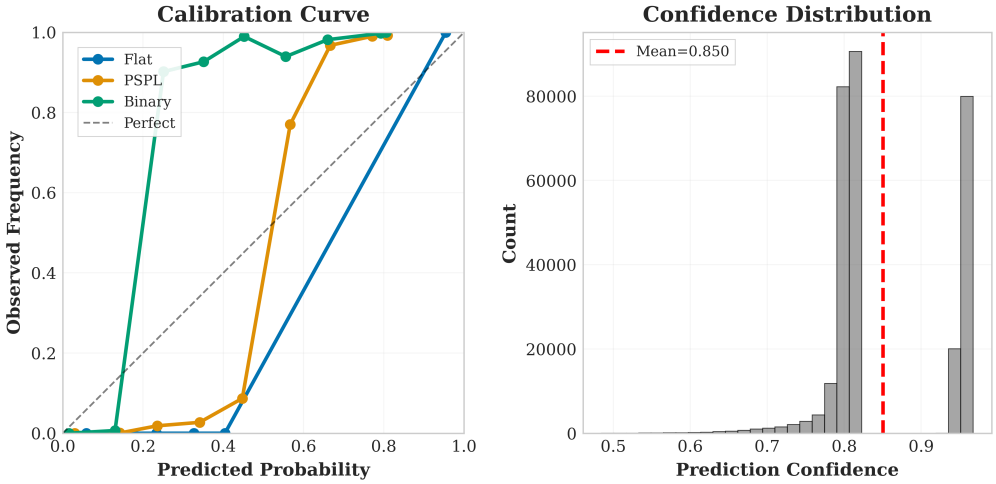

Figure 5.6.: Probability calibration plots verifying the reliability of predicted confidence scores.

5.3. Physical Analysis & Bias

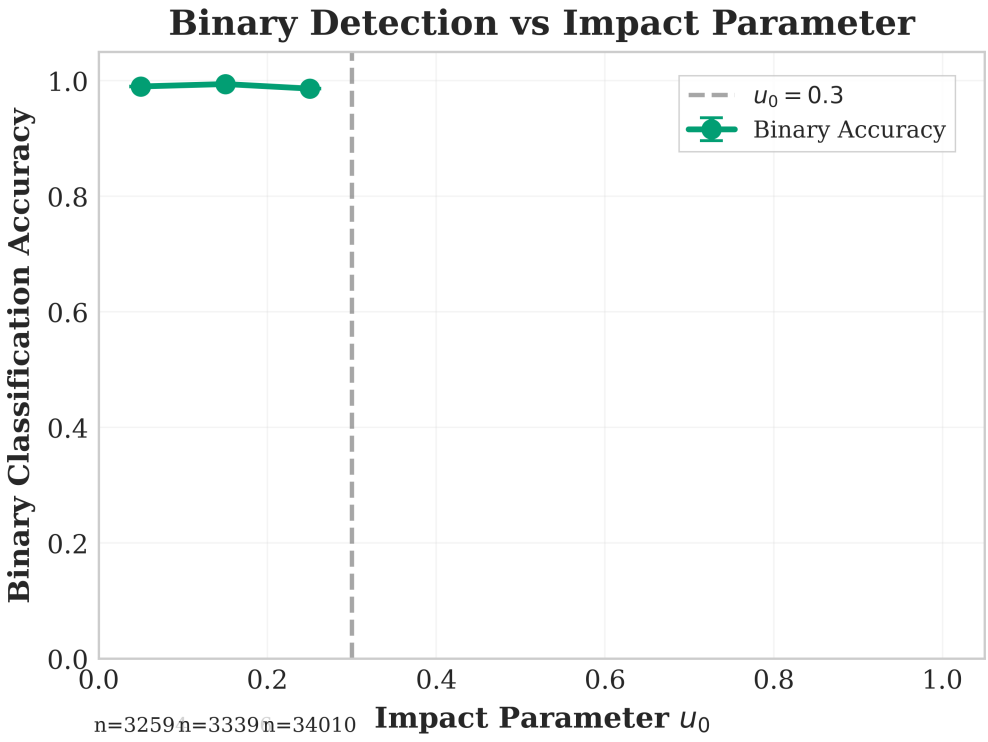

Figure 5.7.: Classification accuracy as a function of the impact parameter u_0.

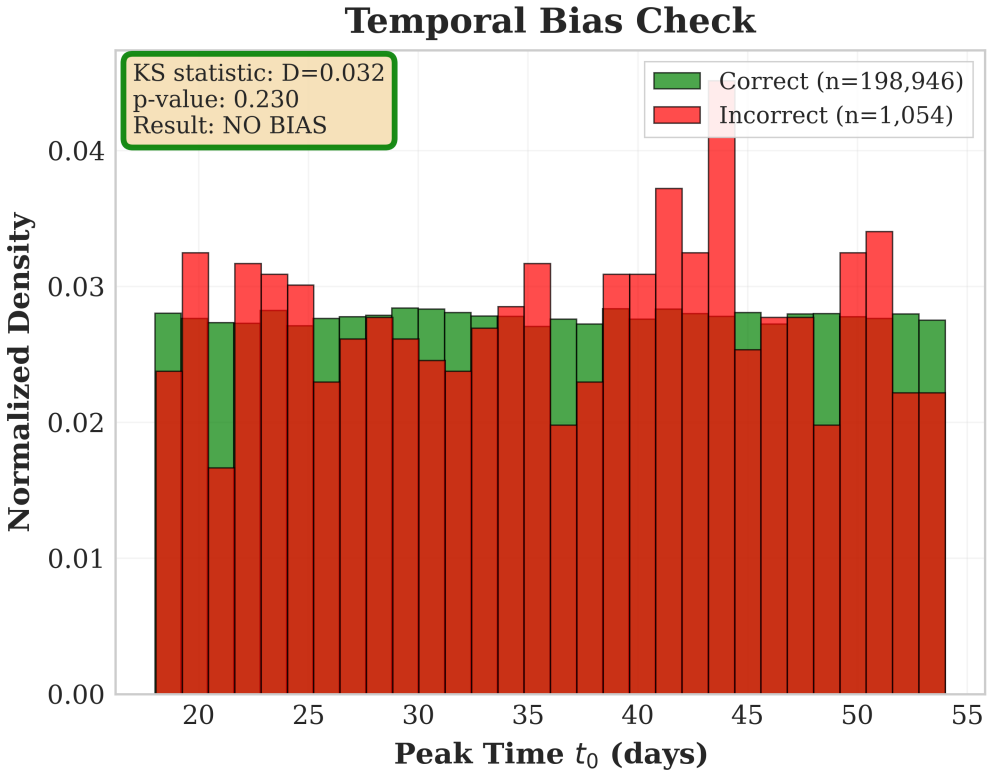

Figure 5.8.: Temporal bias check analyzing performance stability across the observing window.

5.4. Event Evolution Examples

The following figures track the model's prediction probability as more data points are observed over time.

Binary Events

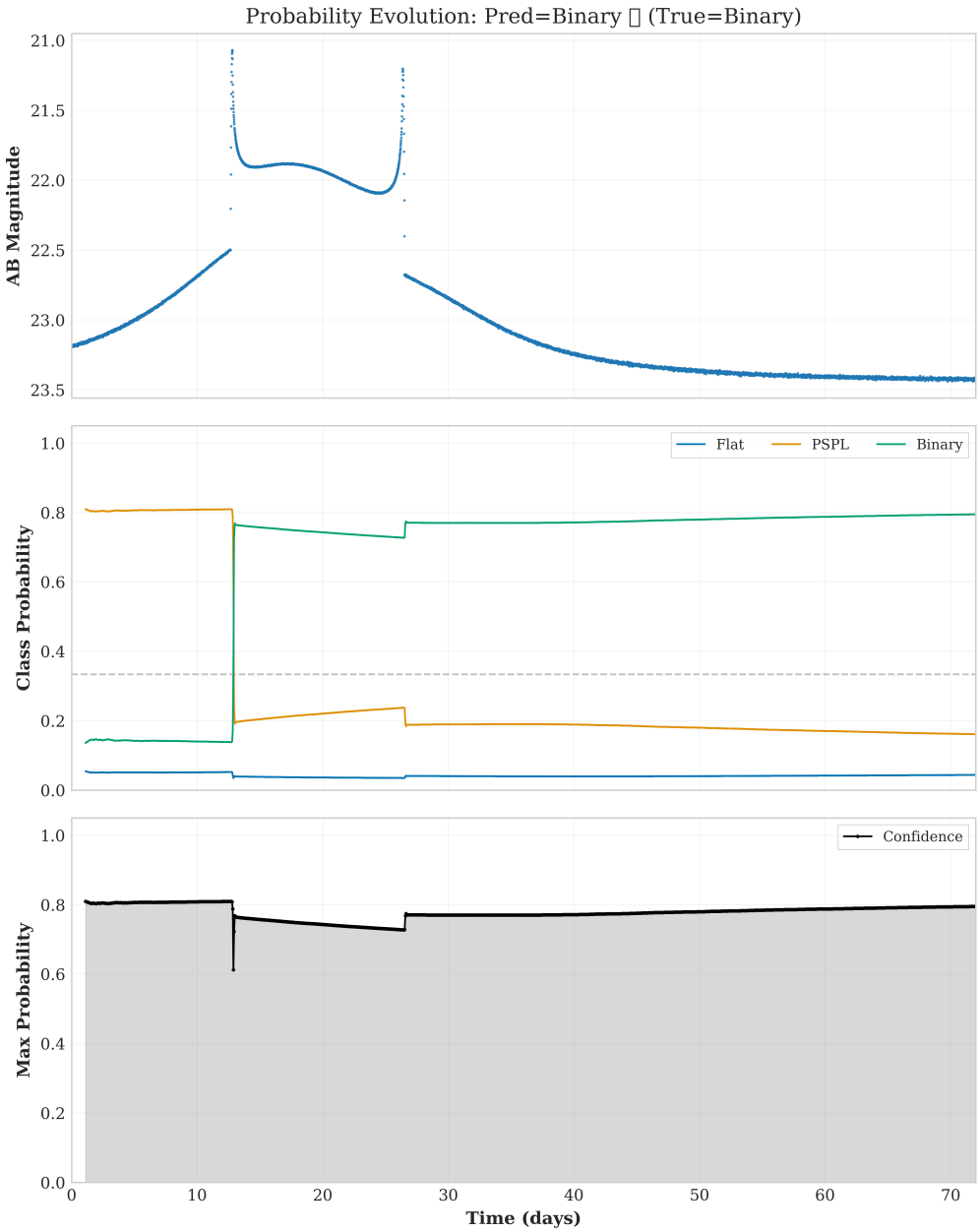

Figure 5.9.: Evolution of predicted probabilities for Binary Event 3.

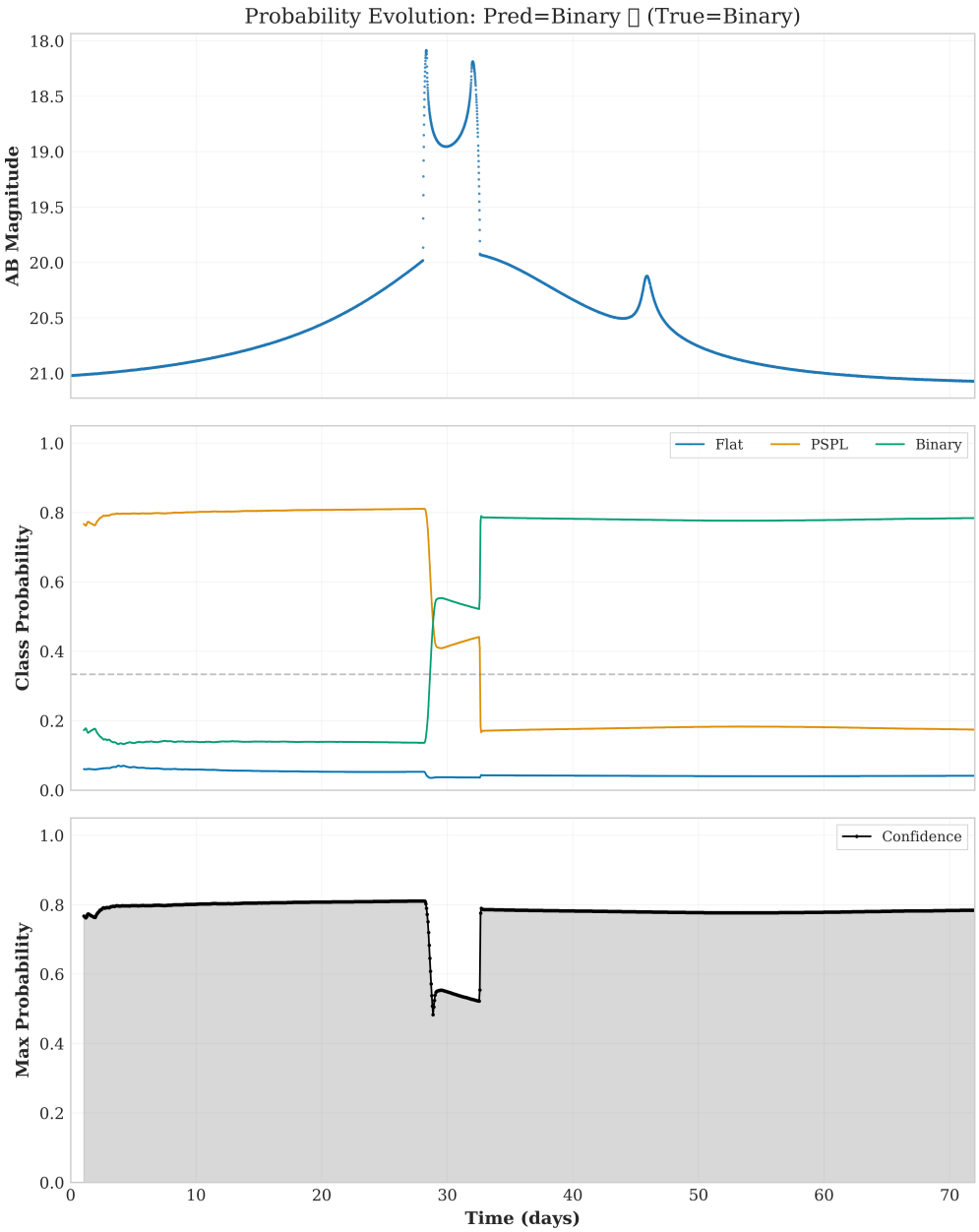

Figure 5.10.: Evolution of predicted probabilities for Binary Event 12.

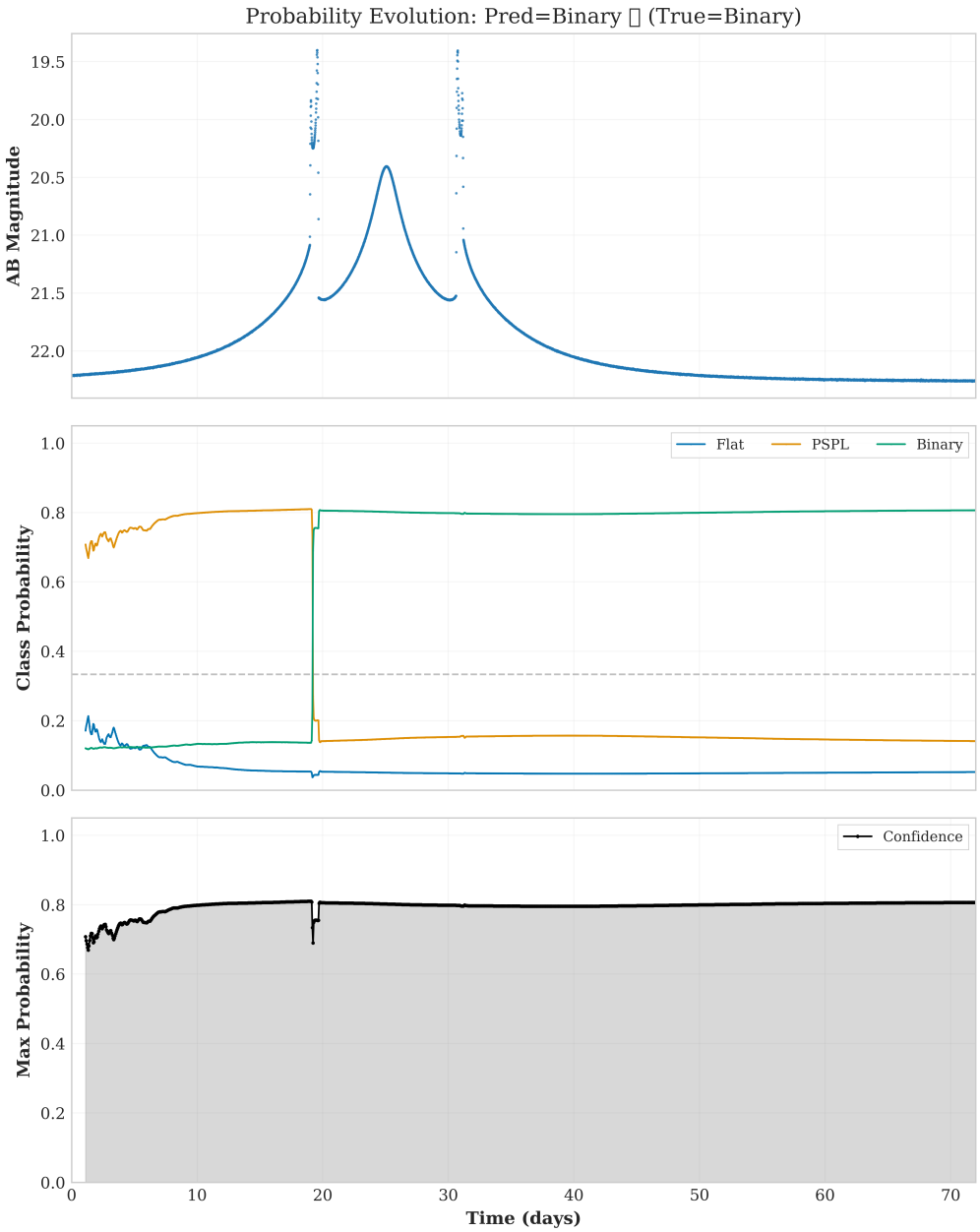

Figure 5.11.: Evolution of predicted probabilities for Binary Event 16.

PSPL Events

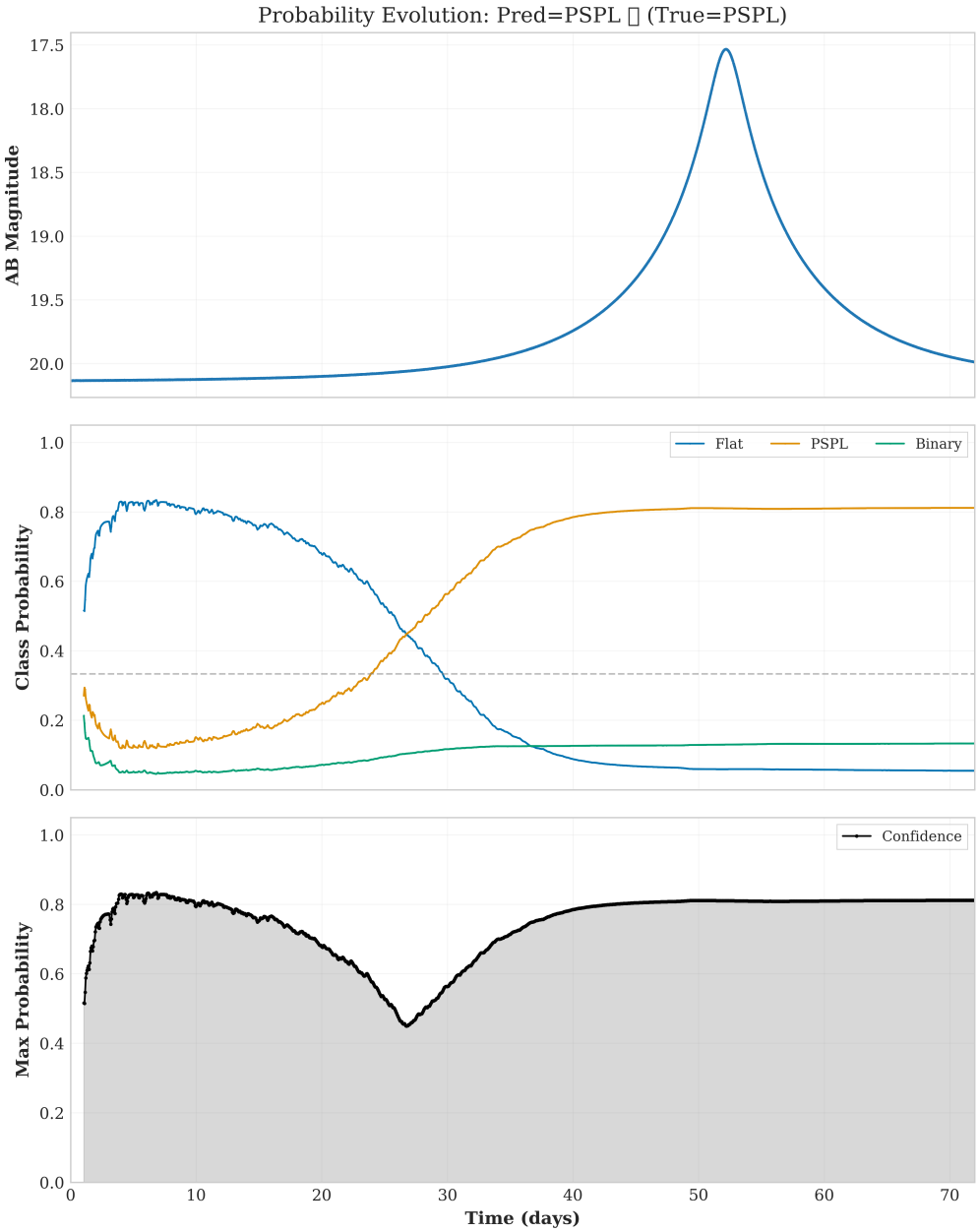

Figure 5.12.: Evolution of predicted probabilities for PSPL Event 1.

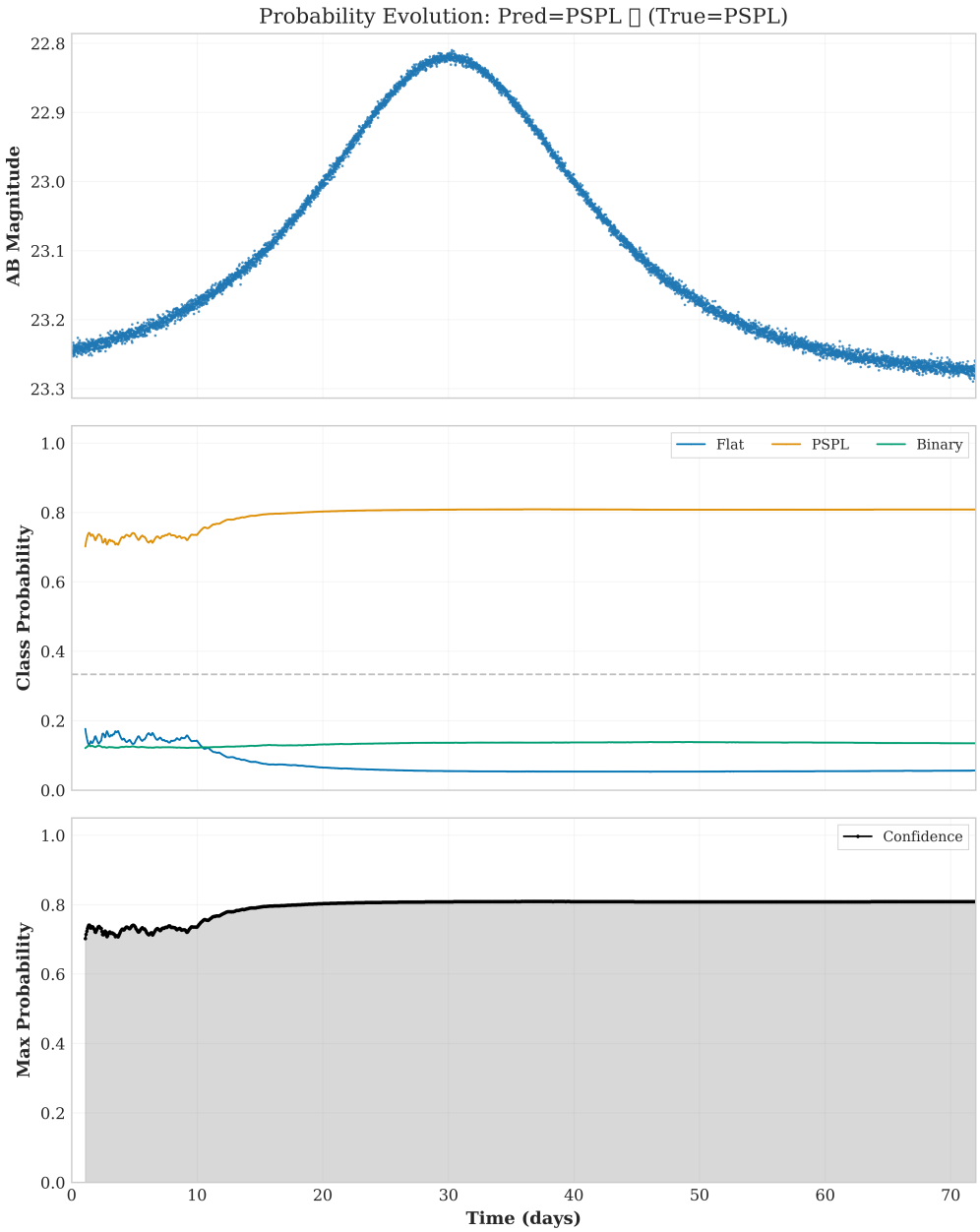

Figure 5.13.: Evolution of predicted probabilities for PSPL Event 4.

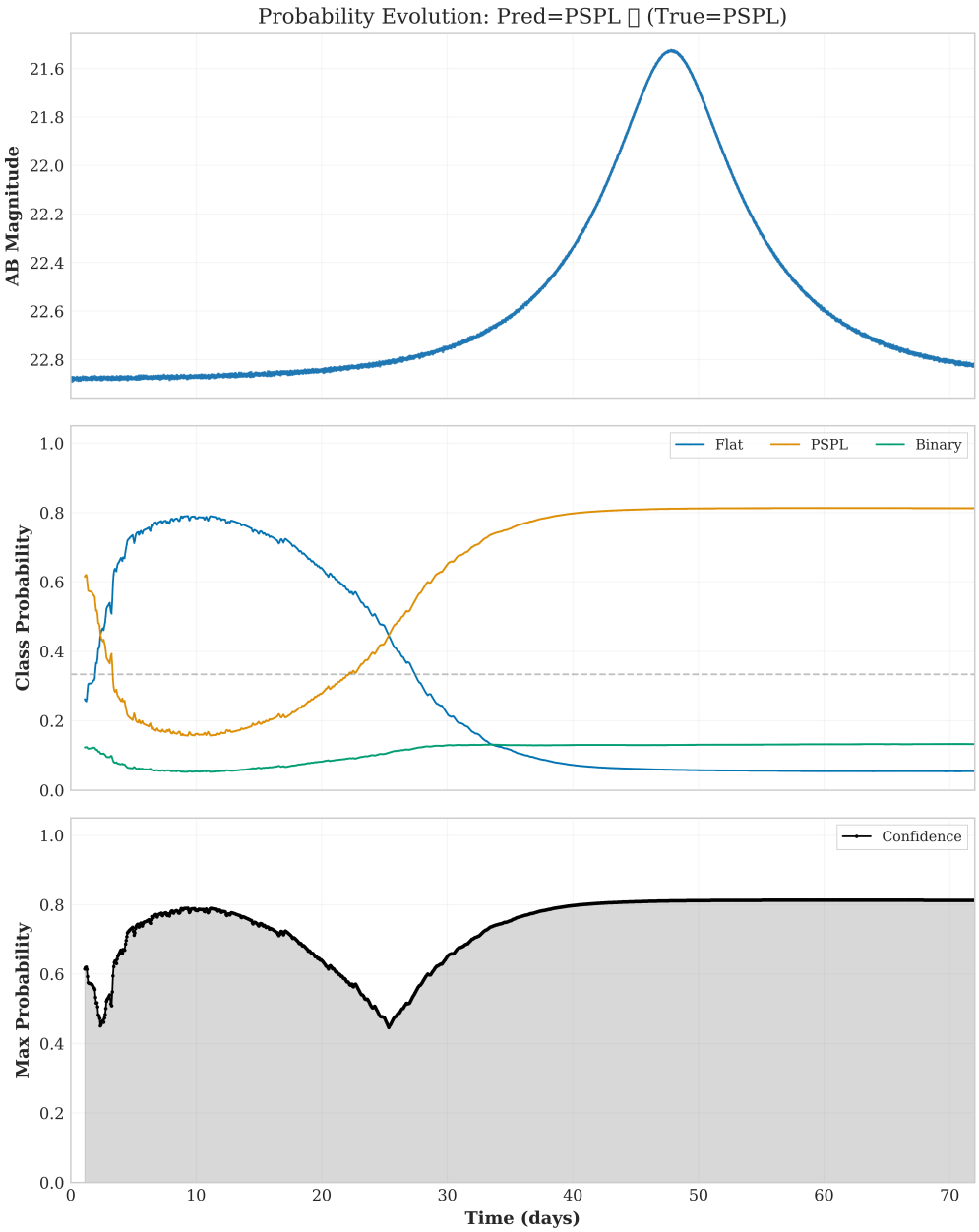

Figure 5.14.: Evolution of predicted probabilities for PSPL Event 5.

Flat Events

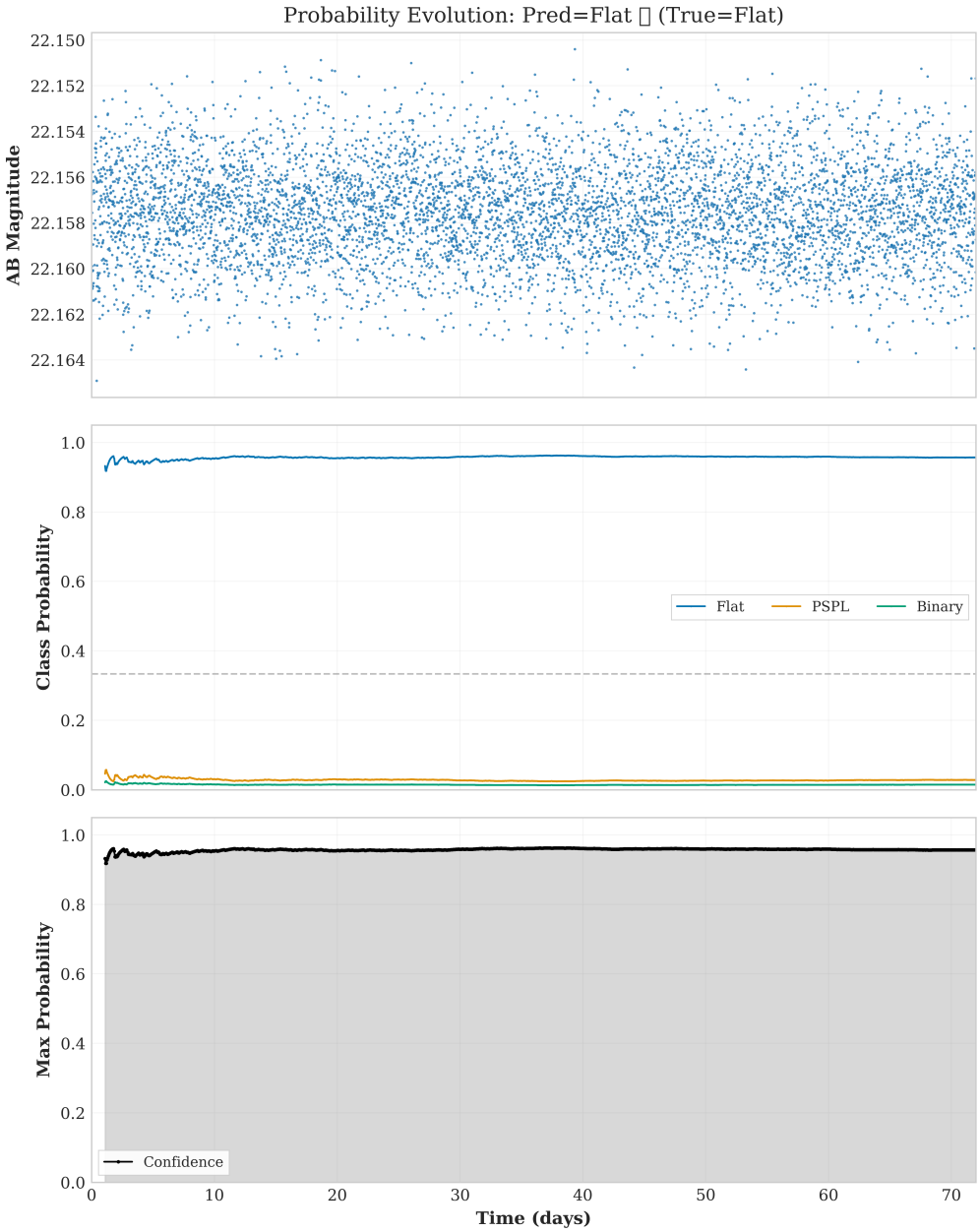

Figure 5.15.: Evolution of predicted probabilities for Flat Event 0.

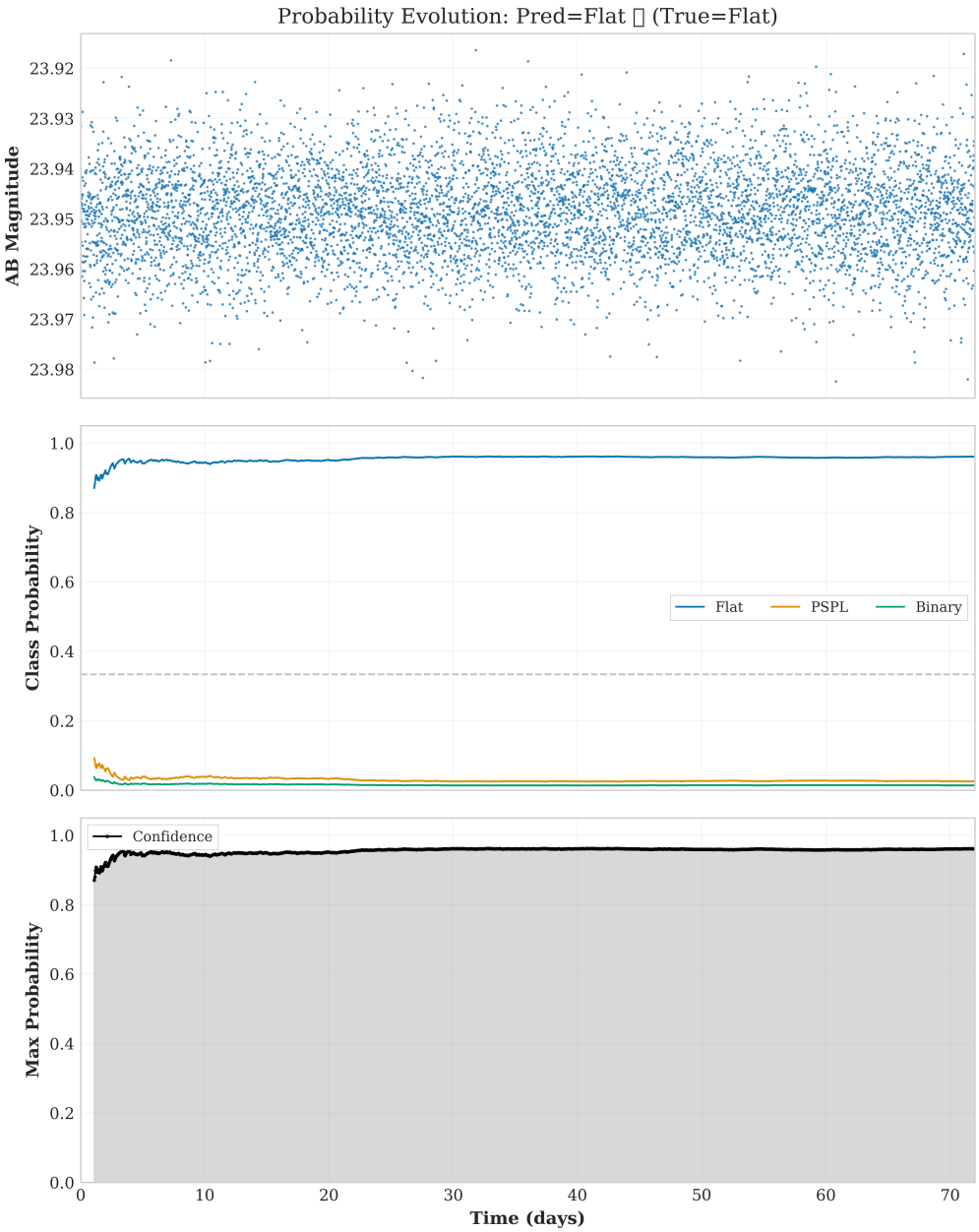

Figure 5.16.: Evolution of predicted probabilities for Flat Event 2.

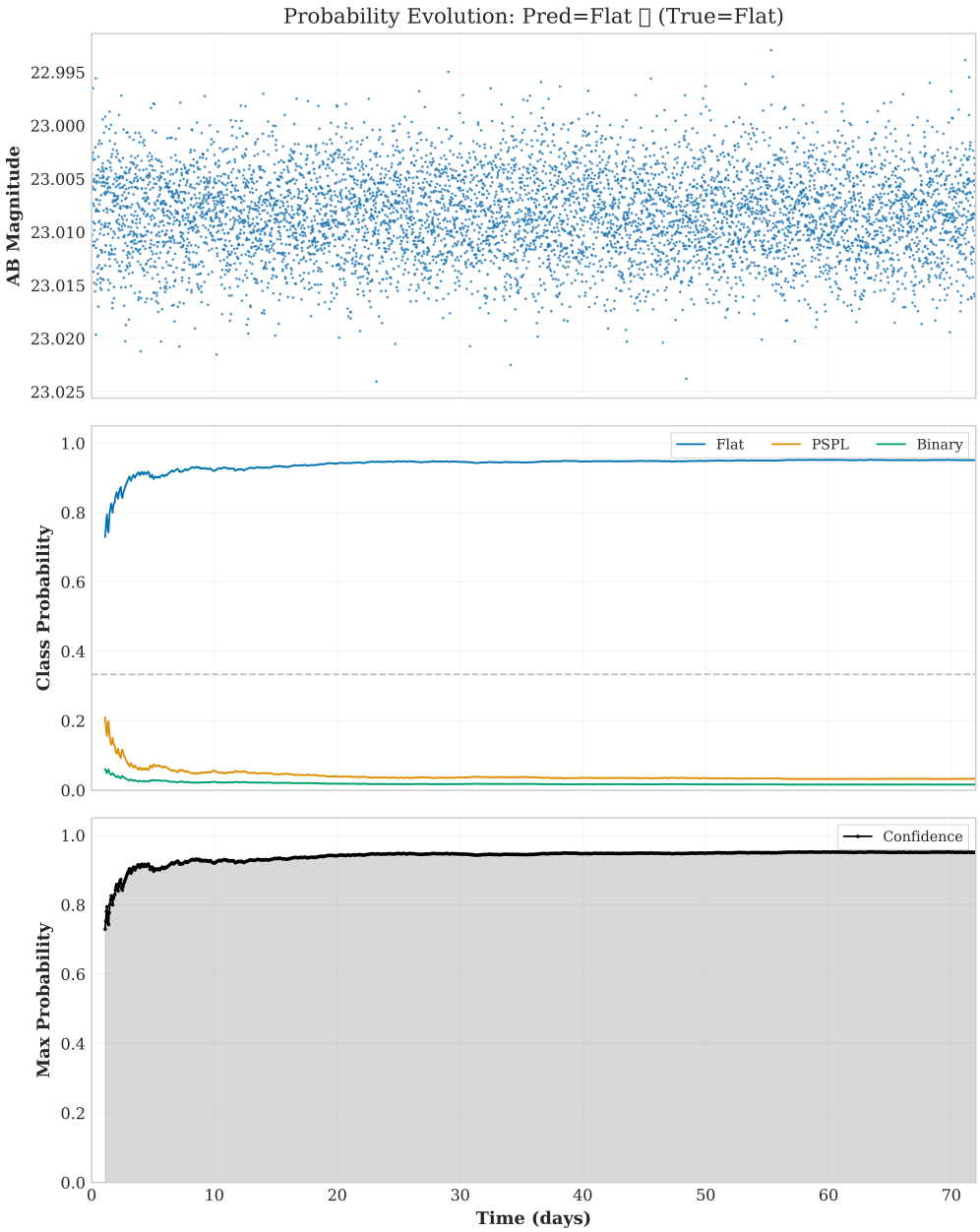

Figure 5.17.: Evolution of predicted probabilities for Flat Event 6.

6. Discussion

[Content to be written]

7. Conclusions and Future Work

[Content to be written]

Bibliography

[1] Albert Einstein. Lens-like action of a star by the deviation of light in the gravitational field. *Science*, 84:506–507, 1936. doi: 10.1126/science.84.2188.506.

[2] B. Paczyński. Gravitational microlensing by the galactic halo. *The Astrophysical Journal*, 304:1–5, 1986. doi: 10.1086/164140.

[3] B. S. Gaudi. Microlensing surveys for exoplanets. *Annual Review of Astronomy and Astrophysics*, 50:411–453, 2012. doi: 10.1146/annurev-astro-081811-125518.

[4] S. Mao and B. Paczyński. Gravitational microlensing by double stars and planetary systems. *The Astrophysical Journal Letters*, 374:L37–L40, 1991. doi: 10.1086/186066.

[5] A. Gould and A. Loeb. Discovering planetary systems through gravitational microlenses. *The Astrophysical Journal*, 396:104–114, 1992. doi: 10.1086/171700.

[6] S. Mao. Introduction to gravitational microlensing. *Research in Astronomy and Astrophysics*, 12:947–972, 2012. doi: 10.1088/1674-4527/12/8/005.

[7] A. Udalski, M. K. Szymański, and G. Szymański. Ogle-iv: Fourth phase of the optical gravitational lensing experiment. *Acta Astronomica*, 65:1–38, 2015.

[8] I. A. Bond, F. Abe, R. J. Dodd, et al. The moa project: Results from 10 years of microlensing surveys. *Publications of the Astronomical Society of the Pacific*, 129: 014002, 2017. doi: 10.1088/1538-3873/129/971/014002.

[9] P. Mróz, A. Udalski, J. Skowron, et al. A terrestrial-mass rogue planet candidate detected in the shortest-timescale microlensing event. *The Astrophysical Journal Letters*, 903:L11, 2020. doi: 10.3847/2041-8213/abbfad.

[10] M. Dominik. The binary lens caustic singularity structure. *Astronomy and Astrophysics*, 349:108–132, 1999.

[11] LSST Science Collaboration. *LSST Science Book, Version 2.0.* arXiv preprint arXiv:0912.0201, 2009.

[12] N. S. Abrams, S. Miyazaki, R. A. Street, et al. Rubin observatory lsst transients and variable stars roadmap. *arXiv preprint arXiv:2208.04499*, 2023.

[13] M. T. Penny, B. S. Gaudi, E. Kerins, et al. Predictions of the nancy grace roman space telescope galactic exoplanet survey. ii. free-floating planet detection rates. *The Astrophysical Journal Supplement Series*, 241:3, 2019. doi: 10.3847/1538-4365/aafb69.

[14] R. A. Street, E. Bachelet, Y. Tsapras, et al. Rome/rea: A gravitational microlensing search for exoplanets beyond the snow line on a worldwide telescope network. *Proceedings of SPIE*, 10707:1070711, 2018. doi: 10.1117/12.2312293.

[15] Subo Dong, Andrew Gould, Andrzej Udalski, et al. Planetary detection efficiency of the magnification 3000 microlensing event ogle-2004-blg-343. *The Astrophysical Journal*, 642:842–860, 2006. doi: 10.1086/501224.

[16] F. W. Dyson, A. S. Eddington, and C. Davidson. A determination of the deflection of light by the sun's gravitational field. *Philosophical Transactions of the Royal Society A*, 220:291–333, 1920. doi: 10.1098/rsta.1920.0009.

[17] V. Bozza. Vbbl: a code for fast computation of the microlensing observables. *Monthly Notices of the Royal Astronomical Society*, 408:2188–2196, 2010. doi: 10.1111/j.1365-2966.2010.17265.x.

[18] M. Lochner, J. D. McEwen, H. V. Peiris, O. Lahav, and M. K. Winter. Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*, 225:31, 2016. doi: 10.3847/0067-0049/225/2/31.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

[20] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

[22] Jacob T. VanderPlas and Željko Ivezić. Periodograms for multiband astronomical time series. *The Astrophysical Journal*, 812:18, 2018. doi: 10.1088/0004-637X/812/1/18.

[23] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019. doi: 10.1007/s10618-019-00619-1.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[26] Christopher J. Shallue and Andrew Vanderburg. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astronomical Journal*, 155:94, 2018. doi: 10.3847/1538-3881/aa9e09.

[27] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[28] Brett Naul, Joshua S. Bloom, Fernando Pérez, and Stéfan van der Walt. A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2:151–155, 2018. doi: 10.1038/s41550-017-0321-z.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[30] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[31] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[32] A. Möller and T. de Boissière. Supernnova: an open-source framework for bayesian, neural network-based supernova classification. *Monthly Notices of the Royal Astronomical Society*, 491:4277–4293, 2020. doi: 10.1093/mnras/stz3312.

[33] D. Godines, E. Bachelet, W. Zhu, and M. Penny. Systematic kmtnet planetary anomaly search. i. ogle-2019-blg-1053lb, a buried terrestrial planet. *The Astronomical Journal*, 157:235, 2019. doi: 10.3847/1538-3881/ab1f65.

[34] S. Khakpash, J. Pepper, K. K. McLeod, et al. A machine-learning approach to enhancing erosita observations. *The Astronomical Journal*, 161:132, 2021. doi: 10.3847/1538-3881/abd8d9.

[35] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016. doi: 10.1186/s40537-016-0043-6.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[39] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[40] Rajat Raina, Anand Madhavan, and Andrew Y. Ng. Large-scale deep unsupervised learning using graphics processors. In *International Conference on Machine Learning*, pages 873–880, 2009.

[41] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

[42] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

[43] D.-J. Kim, H.-W. Kim, K.-H. Hwang, et al. Kmtnet nearby galaxy survey i. optimal strategy and ngc 1566 case study. *The Astrophysical Journal*, 854:22, 2018. doi: 10.3847/1538-4357/aaa127.

[44] D. Spergel, N. Gehrels, C. Baltay, et al. Wide-field infrared survey telescope-astrophysics focused telescope assets wfirst-afta 2015 report. *arXiv preprint arXiv:1503.03757*, 2015.

[45] R. Akeson, L. Armus, E. Bachelet, et al. The wide field infrared survey telescope: 100 hubbles for the 2020s. *arXiv preprint arXiv:1902.05569*, 2019.

[46] D. Suzuki, D. P. Bennett, T. Sumi, et al. The exoplanet mass-ratio function from the moa-ii survey. *The Astrophysical Journal*, 833:145, 2016. doi: 10.3847/1538-4357/833/2/145.

[47] Ł. Wyrzykowski, Z. Kostrzewa-Rutkowska, J. Skowron, et al. Ogle-iv real-time transient search. *Acta Astronomica*, 65:1–26, 2015.

[48] Korki Ishitani Silva Terra, Anderson Santana dos Santos, et al. Automated microlensing detection with a deep learning model. *Monthly Notices of the Royal Astronomical Society*, 515:4180–4192, 2022. doi: 10.1093/mnras/stac2059.

[49] Ž. Ivezić, S. M. Kahn, J. A. Tyson, et al. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873:111, 2019. doi: 10.3847/1538-4357/ab042c.

[50] François Chollet. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[51] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[52] Paulius Micikevicius, Sharan Narang, Jonah Alben, et al. Mixed precision training. *International Conference on Learning Representations*, 2018.

A. Code and Implementation Details

[Optional: Code listings, hyperparameters, etc.]