Artificial Intelligence & Machine Learning (623.504, 23S)

Proposal: **Detecting Black Hole Candidates Using Machine Learning**

Walter Telsnig[1,2]

[1]Alpen-Adria-Universität Klagenfurt,Universitätsstraße 65-67, 9020 Klagenfurt, Austria

[2]Email: wtelsnig@edu.aau.at

April 28, 2025

Abstract

This project proposal outlines the use of Machine Learning (ML) techniques to assist in identifying black hole candidates from astronomical survey data. It discusses the motivation, learning task, related work, the planned approach, and risk management strategies. Starting with foundational models such as linear regression will help mitigate early implementation risks and provide a baseline for model performance and will provide a baseline for model performance. As the project progresses, more complex and powerful machine learning models, such as ensemble methods or neural networks or physical-informed networks, will be tested.

1 Motivation

Ever since my first encounter with science fiction and, later on, through computer games, black holes have radiated something obscure and deeply fascinating. There's something about these cosmic mysteries that instantly drew me in — the idea that parts of the universe could defy what we think we know about time, space, and physics. Black holes twist spacetime so extremely that they challenge our classical understanding of gravity, thermodynamics, and even quantum theory. What amazes me most is the thought that time itself behaves differently near them — almost as if the rules of nature break down. And while they might sound like science fiction, black holes are very real. We can see their presence indirectly through the way they affect nearby stars, the intense radiation from their accretion disks, and even through the ripples in spacetime known as gravitational waves.

What really drives this project is my curiosity about how modern computational tools — especially Machine Learning — can help me explore one of the most exciting frontiers in astrophysics. Traditional ways of detecting black holes, like

tracking X-ray emissions or analyzing how nearby stars move, are powerful but also incredibly resource-heavy. They demand a lot of telescope time, expert analysis, and manual effort. With the massive amount of data we're now collecting from space, it's becoming clear that we need smarter, more scalable ways to handle it. That's where ML comes in. These techniques can quickly scan through huge datasets, spot patterns we might miss, and highlight unusual signals that could point to new black hole candidates — all with minimal human input. It opens the door to discoveries we might never make otherwise.

The goal of this project is thus to apply ML techniques to assist in the identification of black hole candidates from astronomical survey data. By automating parts of the detection process, we can not only speed up discovery but also reduce observational biases and potentially reveal new classes of astrophysical phenomena. My hypothesis is that the physical properties of astronomical objects, such as redshift, luminosity, and spectral features, contain sufficient information to enable ML models to reliably distinguish black hole candidates from other celestial objects.

This project merges my personal interest in the extreme physics of black holes with my interest in using machine learning to solve complex scientific challenges, aiming to contribute meaningfully to both fields.

2 Learning task

2.1 Training experience

Initially, training data will be sourced from real astronomical datasets, such as the Sloan Digital Sky Survey (SDSS) catalog [1], which includes a wide range of objects and labeled sources (e.g., quasars, stars, galaxies) for million of galaxies. The NASA/IPAC Extragalactic Database (NED) [2] provides information on galaxy dynamics and black hole masses. The Event Horizon Telescope (EHT) & Chandra X-ray Observatory [3] contains X-ray and radio data related to black hole environments. Finally, the Supermassive Black Hole Mass Database [4] is a curated list of black hole mass measurements from the literature and will be used to check our estimates.

Black hole candidates can be inferred from quasars and active galactic nuclei (AGN) categories. Should issues arise with data availability, data quality, or preprocessing complexities, the project will pivot to using simulated datasets. Example data entries include attributes like redshift (z), spectral line features (e.g., H-alpha emissions), brightness/magnitude in different filters, and radio/X-ray flux (where available).

2.2 Learning task

The core task is a classification problem: given a set of features describing an astronomical object, predict whether it is a potential black hole candidate.

Learning steps include preprocessing, training initial classification models (e.g.,

traditional regression models, random forest [5], support vector machine [6], shallow neural networks [7]), and evaluating performance. In [8], Chainukun et al. showed that even small neural network approached outperform traditional linear regression methods. Advanced methods like physics-informed neural networks (PINNs) [9] or ensemble approaches will be explored if time permits.

Following an initial and non-exhaustive review of the literature, I identified several astrophysical properties that are commonly used to predict black hole mass [10, 11]. The following overview summarizes the key features that emerged during this early-stage research and will be refined and expanded as the project progresses. To predict the mass of a black hole, a range of astrophysical properties can be utilized as input features. These features can be broadly categorized into galaxy properties and emission or spectroscopic characteristics.

Galaxy properties offer several important predictors. Stellar velocity dispersion (σ) measures the speed at which stars move near the galactic center and provides insight into the gravitational influence of the central black hole. Bulge luminosity, representing the brightness of the galaxy's central region, has been shown to correlate with black hole mass. The total mass of the host galaxy, encompassing both visible and dark matter, is another significant feature. Additionally, the effective radius, describing the size of the galactic bulge, and the galaxy's metallicity, referring to the abundance of heavy elements, are relevant properties that can indirectly inform black hole mass estimates.

Emission and spectroscopic features also provide valuable information. High-energy X-ray and radio emissions are often associated with accreting black holes and serve as important observational signatures. The broad-line region width, describing the extent of spectral emission lines from fast-moving gas near the black hole, offers a direct probe of the black hole's gravitational field. Furthermore, the Eddington ratio, defined as the ratio of the black hole's observed luminosity to its theoretical Eddington limit, reflects the level of accretion activity and provides additional context for mass estimation.

2.3 Performance measure

Performance will be evaluated using:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **ROC Curve and AUC**

Interpretation: A precision of 0.7 implies that 70% of predicted black hole candidates are true candidates, which may be acceptable depending on the follow-up costs and time remaining.

3 Related work

Recent advances in machine learning have demonstrated significant success in classifying astrophysical objects and extracting hidden patterns from complex data sets. A notable study is "AstroNet: Deep Learning for Classifying Astronomical Objects" by Shallue and Vanderburg (2018) [12], where convolutional neural networks (CNNs) were applied to time-series photometric data from the Kepler space telescope to detect exoplanet candidates with high precision. Although their focus was primarily on exoplanet discovery, the underlying principle of automatically learning representations from astrophysical data directly applies to the present problem of black hole mass estimation. In adapting their methodology, instead of using time-series flux curves, this project will use static tabular data (e.g., stellar velocity dispersions, galaxy luminosities) as model inputs, with potential domain-specific data augmentations (e.g., noise modeling or feature perturbation) to improve generalization.

Similarly, research by Domínguez Sánchez et al. (2018) [13]demonstrated the effectiveness of deep learning in morphological galaxy classification using imaging data, showing that neural networks can uncover subtle structural differences between galaxies that correlate with underlying physical properties. Although their inputs were images rather than catalog features, the success of deep models in extracting latent astrophysical features supports the feasibility of learning black hole mass predictors directly from observables.

Beyond deep learning, classical machine learning techniques have also been successfully applied to astrophysical problems. For instance, Ball and Brunner (2010) [14] reviewed a range of supervised methods such as random forests and support vector machines (SVMs) in the context of astronomical classification and regression tasks. Given the structured nature of the tabular data available for black hole studies, ensemble methods such as random forests or gradient boosting (e.g., XGBoost) offer strong baseline models that can be compared against deep learning approaches.

Overall, the emerging trend across the astrophysics community is the replacement of manual feature engineering with automated, data-driven discovery through machine learning. Following this paradigm, the present work seeks to leverage machine learning models — ranging from interpretable ensemble methods to more expressive deep neural networks — to predict black hole masses from galaxy properties, drawing inspiration from successful methodologies in related astrophysical tasks.

4 Plan

To keep things clear and manageable, I'll follow a step-by-step approach for this project — starting from data collection all the way to building something interactive at the end. Since time is limited, I'll begin with simple, tried-and-tested models like linear regression, which are easier to understand and quick to run. The process

kicks off with gathering and cleaning the data — making sure the inputs are solid and filtering out anything that doesn't fit well. Then I'll explore the data a bit to get a feel for how the different features relate to each other. After that, I'll start training some machine learning models, checking how well they perform using common evaluation methods. I plan to build things up gradually, adding complexity only when needed — kind of like leveling up as I go. I also want to take a look at what the models are actually "thinking," so I'll include some analysis on how they make their predictions. Once everything's working, I'll wrap things up by comparing the different approaches and if the addition workload was worth the time spent.

The following subsections outline each stage in detail:

1. **Data Collection & Preprocessing**
    - Download and clean datasets (e.g., remove missing values, normalize features).
    - Feature engineering (log transformations, standardization).
    - Data augmentation (e.g., generating synthetic data from known distributions).
2. **Exploratory Data Analysis (EDA)**
    - Visualize feature correlations (e.g., velocity dispersion vs. black hole mass).
    - Plot histograms and distributions of galaxy properties.
3. **Model Training & Validation**
    - Split dataset into training and test sets.
    - Train different ML models and compare performances.
    - Use cross-validation and hyperparameter tuning (e.g., GridSearchCV for Random Forest).
4. **Evaluation Metrics**
    - Mean Absolute Error (MAE) – Measures absolute differences in predictions.
    - Root Mean Squared Error (RMSE) – Penalizes larger prediction errors.
    - R^2 Score – Indicates how well the model explains variance in the data.
5. **Interpretability & Results Analysis**
    - Feature importance analysis (e.g., SHAP values).
    - Compare ML-predicted masses to observed values in astrophysical studies.
6. **Deployment & Visualization**

5 Risk management

To keep things on track and avoid running into problems too early, I'm taking an iterative approach to building the models. I'll start with something simple and easy to understand — like linear regression — to make sure the data looks good and the basic idea works. Starting simple also helps avoid technical issues that could slow things down, especially with limited time. Once I've got a basic version up and running, I'll move on to trying more advanced models, like ensemble methods or maybe even neural networks, to see if they can improve the predictions. This step-by-step approach helps make sure I always have something working, while still leaving room to explore better options later on.

Risks and mitigations include:

- **Data quality issues**: fallback to simulated datasets.
- **Model convergence problems**: start with simpler models.
- **Computational limitations**: use smaller datasets or lightweight models.
- **Time constraints**: refine baseline models if needed.

5.1 File management and folder structure

```
blackhole-mass-estimation/
├── data/
│   ├── raw/
│   └── processed/
├── notebooks/
│   └── 01_simple_regression.ipynb
├── src/
│   ├── data_loader.py
│   ├── model.py
│   └── utils.py
├── models/
├── outputs/
│   ├── figures/
│   └── metrics/
├── requirements.txt
├── README.md
└── .gitignore
```

References

[1] Andrés Almeida, Scott F Anderson, Maria Argudo-Fernández, Carles Badenes, Kat Barger, Jorge K Barrera-Ballesteros, Chad F Bender, Erika Benitez, Felipe Besser, Jonathan C Bird, et al. The eighteenth data release of the sloan digital sky surveys: Targeting and first spectra from sdss-v. *The Astrophysical Journal Supplement Series*, 267(2):44, 2023.

[2] George Helou, BF Madore, M Schmitz, MD Bicay, X Wu, and J Bennett. The nasa/ipac extragalactic database. *Databases & On-Line Data in Astronomy*, pages 89–106, 1991.

[3] Kazunori Akiyama, Juan Carlos Algaba, Antxon Alberdi, Walter Alef, Richard Anantua, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, et al. First m87 event horizon telescope results. viii. magnetic field structure near the event horizon. *The Astrophysical Journal Letters*, 910(1):L13, 2021.

[4] Misty C Bentz and Sarah Katz. The agn black hole mass database. *Publications of the Astronomical Society of the Pacific*, 127(947):67, 2015.

[5] PM Plewa. Random forest classification of stars in the galactic centre. *Monthly Notices of the Royal Astronomical Society*, 476(3):3974–3980, 2018.

[6] Ilya N Pashchenko, Kirill V Sokolovsky, and Panagiotis Gavras. Machine learning search for variable stars. *Monthly Notices of the Royal Astronomical Society*, 475(2):2326–2343, 2018.

[7] Kaushal Sharma, Ajit Kembhavi, Aniruddha Kembhavi, T Sivarani, Sheelu Abraham, and Kaustubh Vaghmare. Application of convolutional neural networks for stellar spectral classification. *Monthly Notices of the Royal Astronomical Society*, 491(2):2280–2300, 2020.

[8] P Chainakun, I Fongkaew, S Hancock, and Andrew J Young. Predicting the black hole mass and correlations in x-ray reverberating agns using neural networks. *Monthly Notices of the Royal Astronomical Society*, 513(1):648–660, 2022.

[9] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

[10] Cosimo Bambi. Astrophysical black holes: a review. *arXiv preprint arXiv:1906.03871*, 2019.

[11] Gustavo E Romero and Gabriela S Vila. *Introduction to black hole astrophysics*, volume 876. Springer, 2013.

[12] 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(2):94, 2018.

[13] H Domínguez Sánchez, M Huertas-Company, M Bernardi, D Tuccillo, and J L Fischer. Deep learning for galaxy surface brightness profile fitting. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, 2018.

[14] Nicholas M Ball and Robert J Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, 2010.