Identifying Stellar Objects from Large Survey Catalogs Using Deep Neural Networks

Sungryong Hong[1], Eunbin Kim[2], Yujin Yang[1], Woong Sub Jeong[1], and Jongwan Ko[1]
[1]Korea Astronomy and Space Science Institute, [2]SI Analytics

Abstract

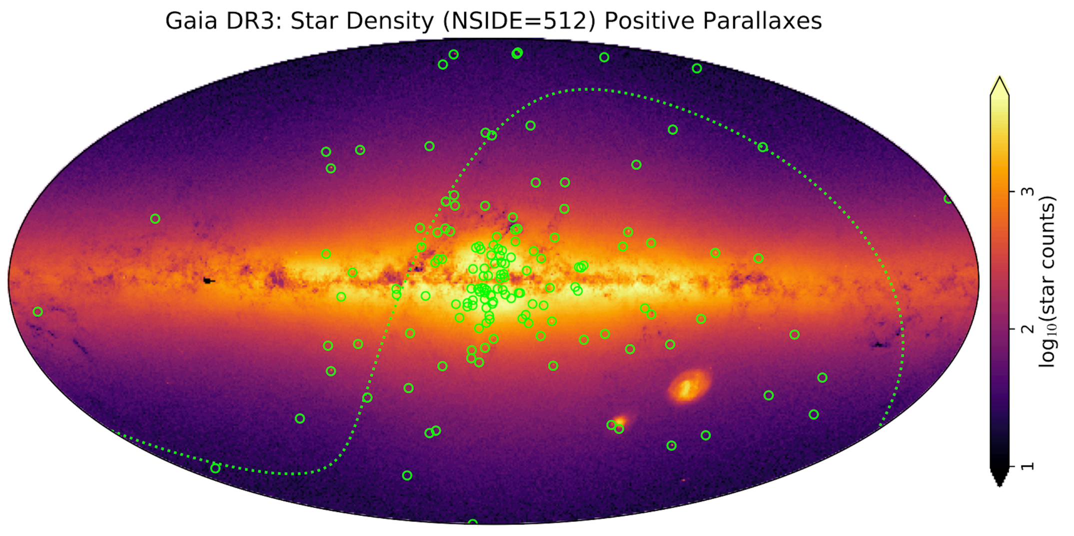
We explore the identification of stellar objects using deep neural networks applied to photometric tabular data from large open catalogs such as Gaia, Pan-STARRS, CatWISE, and the DESI Legacy Survey. In this poster, we present the results from three deep learning models —TabNet/IGNNet/NODE— and compare their performance with the well-established XGBoost model. Additionally, we discuss the interpretability of each deep learning model.

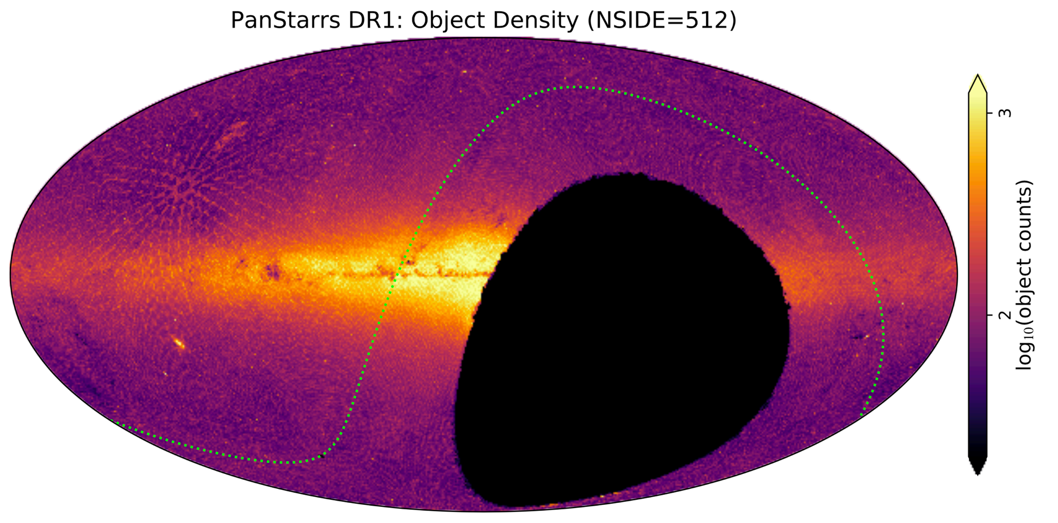
Large Surveys: Tabular Data

DESI Legacy Survey, PanSTARRS, Gaia, CatWISE

Table 1. Large Open Catalogs

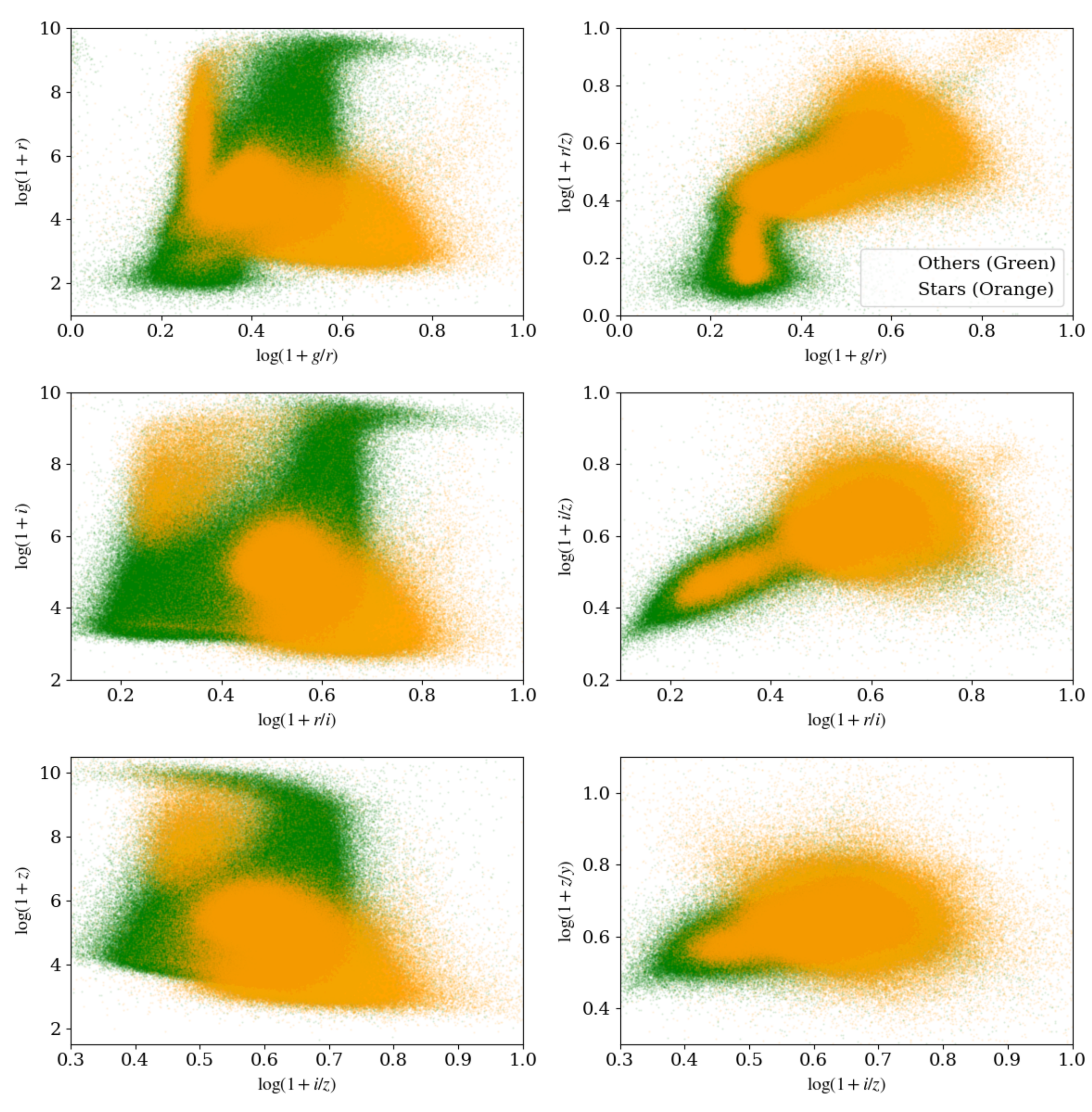
Surveys	Bands	N	Notes
DESI Legacy Survey	*grz*	3.15B	DR9+DR10
PanSTARRS	*grizy*	1.76B	DR1, Cleaned StackObject table
Gaia	G, G_{BP}, G_{RP}	1.81B	DR3, 0.1 arcsec resolution ($G < 20$)
CatWISE	$W1/W2$	1.81B	CatWISE2020, based on unWISE R5





Train Sample: Gaia Labels

Stars: classprob_dsc_combmod_star > 0.95 with Equally Balanced SubSampling

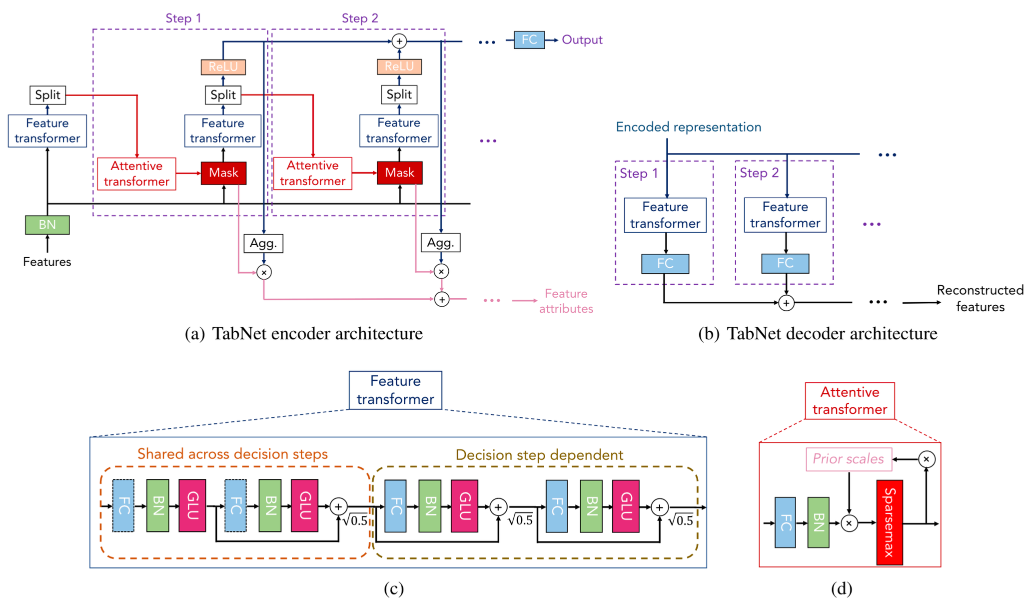


Input Features:
All Fluxes and Color Combinations from *g r i z y w1* and *w2* with log one plus function, such as log(1+g), log(1+g/r), log(1+r/g), log(1+r), log(1+r/z),log(1+z/r), … ; the 49 features in total.

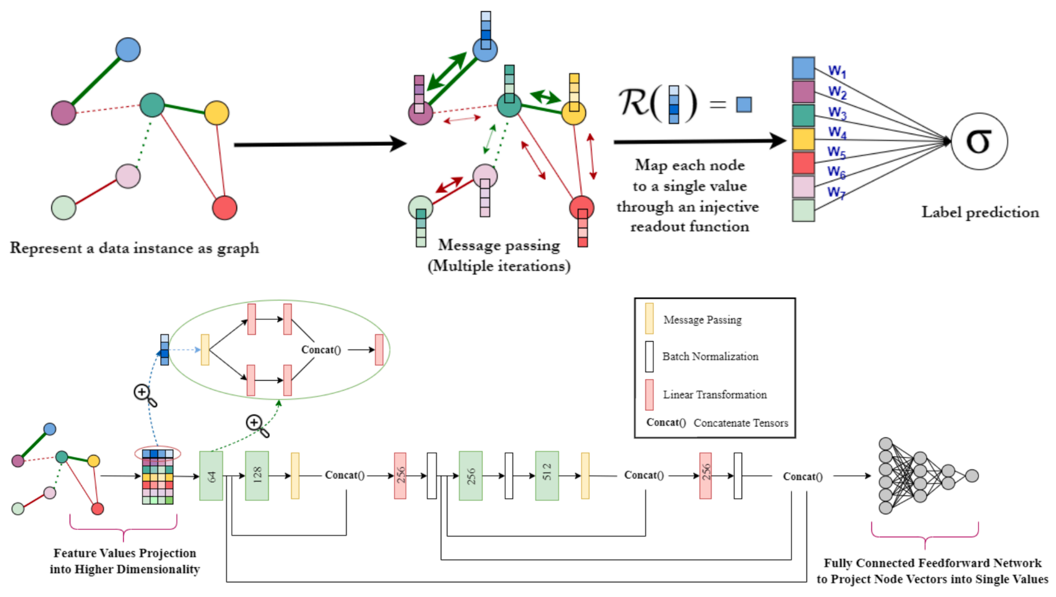
Deep Learning Models

TabNet: Attention Transformer Model
IGNNet: Graph Neural Network Model
NODE: Differentiable Tree Model
vs.
XGBoost: Most Popular ML Model for Tabular Data

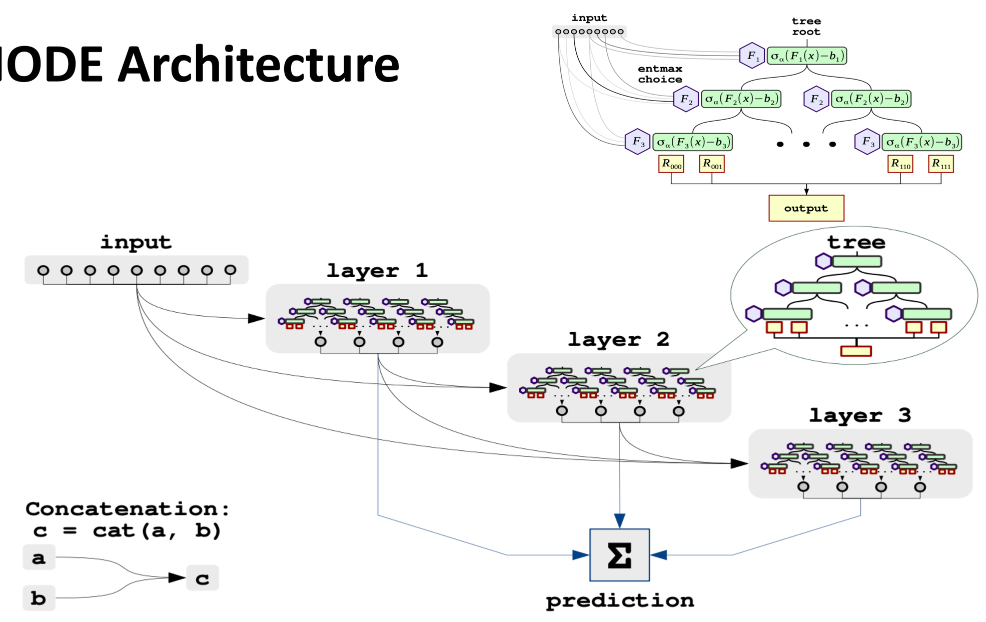
TabNet Architecture



IGNNet Architecture

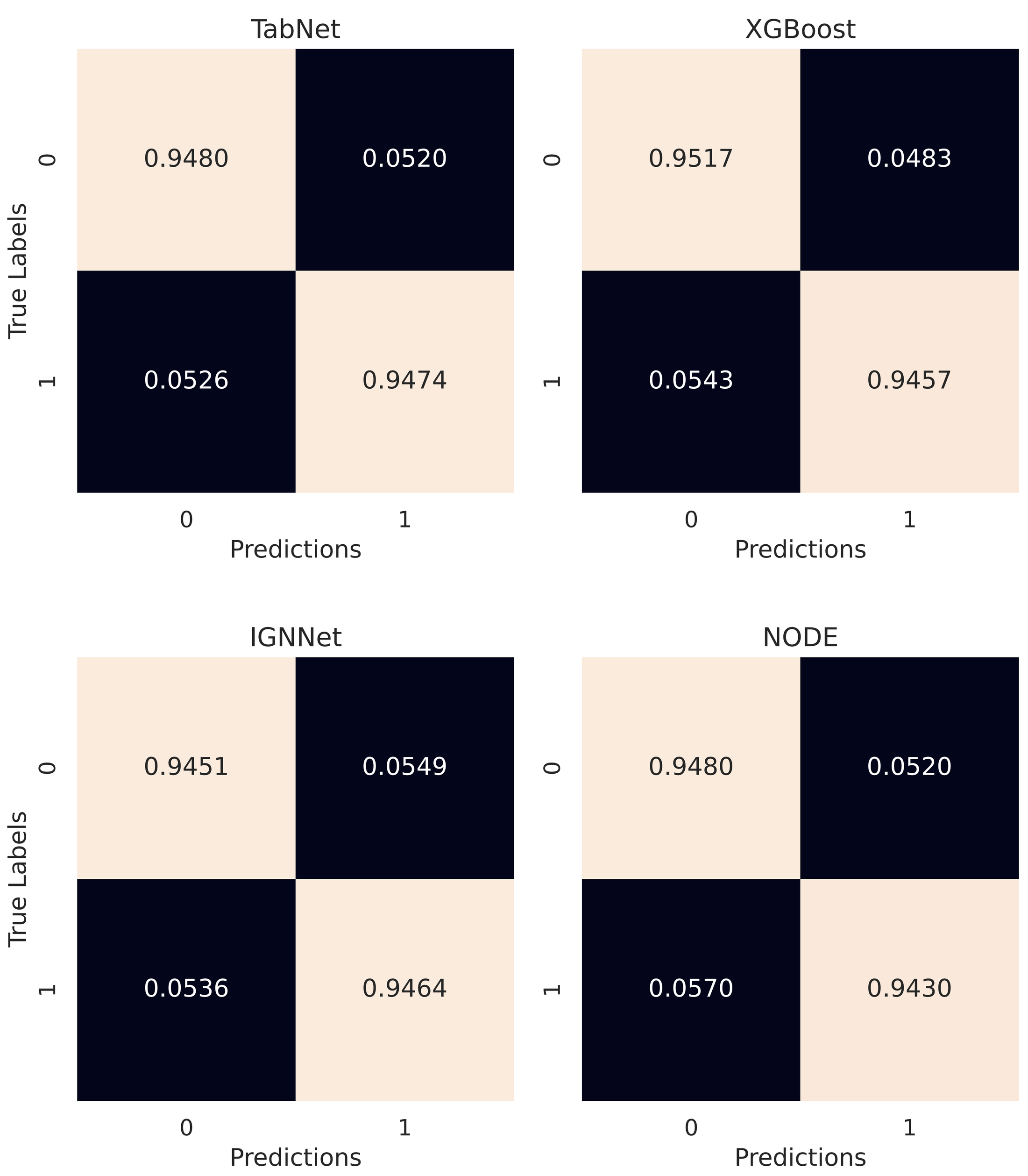


NODE Architecture



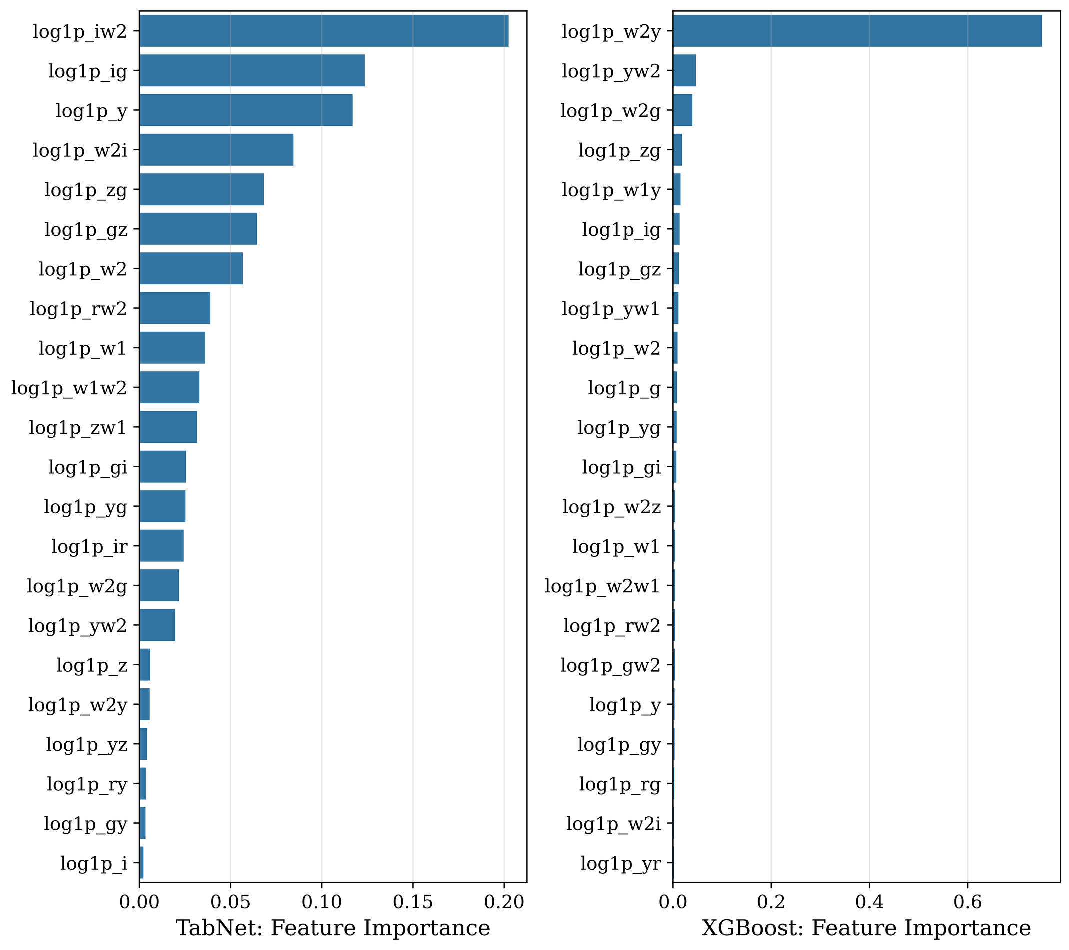
Results

Confusion Matrix



All ML/DL models demonstrate peak performance, suggesting that their capabilities exceed the requirements for this classification problem.

Global Importance



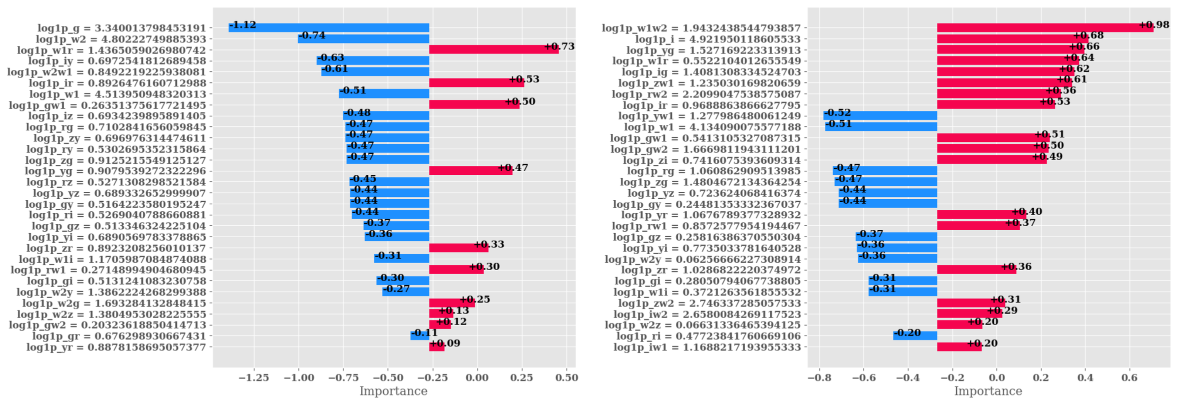
Overall redder bands such as *W1, W2, z,* and *y* are important features, as expected, for identifying stars. But the specific features favored by each model are different, suggesting that no single feature set is universally optimal for this classification problem.

Local Importance and Interpretable Deep Learning

XGBoost and *NODE* are (almost) *black-boxes* (or, less interpretable), while **TabNet** and **IGNNet** are inherently **interpretable** by their designs.

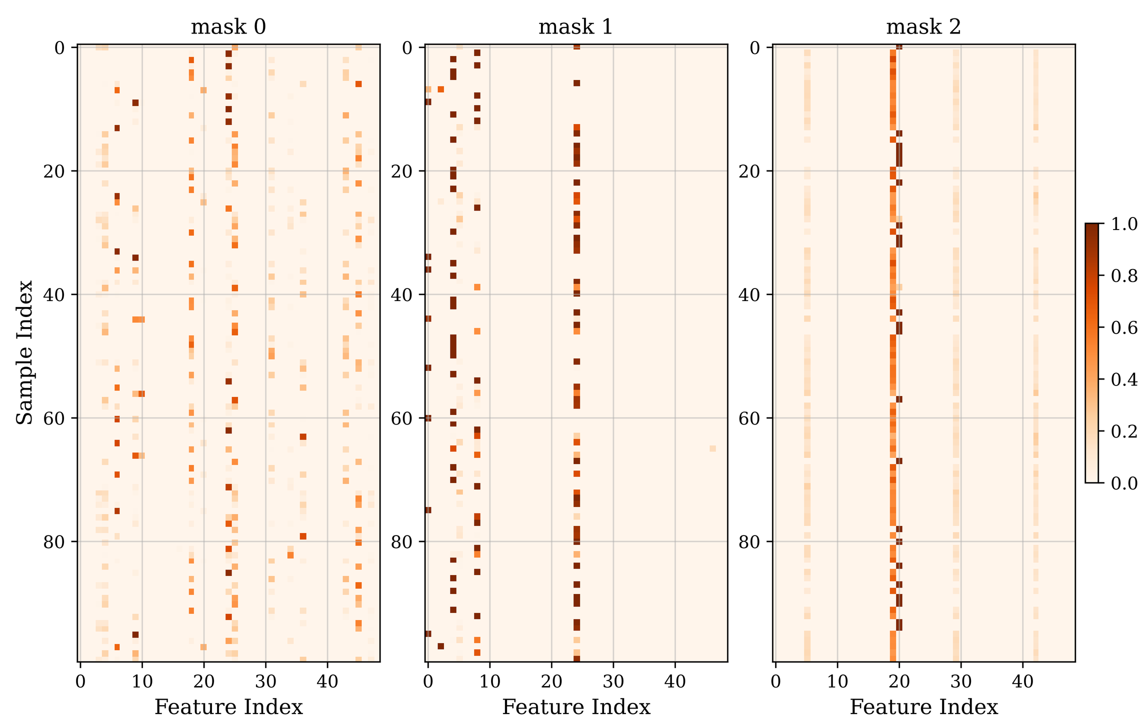
Interpretability of IGNNet

More red bars (star) vs. More blue bars (non-stellar object) for each object (hence, local importance).



Interpretability of TabNet

Attention weights of Mask 0,1,2 for each object (each row with 49 feature indices)



Discussion

The results demonstrate that all ML/DL models employed in this study achieved peak performance, indicating that their capabilities are more than sufficient for the given classification task.

While redder bands are consistently identified as important features for star identification, the specific features favored by individual models vary significantly, suggesting that there is no single optimal feature set for this problem.

With the extensive use of **tabular** and **point cloud data** in astronomy, it is important to focus on developing deep learning models optimized for these data types, alongside the well-established image-based models.

Contact

Sungryong Hong at Korea Astronomy and Space Science Institute
Email: shongscience@kasi.re.kr

References

1. Arik, S. O., & Pfister, T. 2020, TabNet: Attentive Interpretable Tabular Learning. https://arxiv.org/abs/1908.07442
2. Alkhatib, A., Ennadir, S., Bostr¨om, H., & Vazirgiannis, M. 2024, Interpretable Graph Neural Networks for Tabular Data. https://arxiv.org/abs/2308.08945
3. Popov, S., Morozov, S. & Babenko, A. 2019, Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. https://arxiv.org/abs/1909.06312