Integrating LiDAR, Satellite Time-Series, and Ecological Features in a Machine Learning Pipeline for Enhanced Forest Carbon Monitoring

Samuele Capobianco[2], Selene Patani[2], Gonzalo Oton[1], Luca Caporaso[1,3], Matteo Piccardo[2], Valerio Abitabile[2], Alessandro Cescatti[1], Mirco Migliavacca[1]

(1) European Commission, Joint Research Centre, Ispra, Italy (2) Consultant with the European Commission, Joint Research Centre, Ispra, Italy
(3) National Research Council of Italy, Institute of BioEconomy, Rome, Italy

Introduction & Objectives

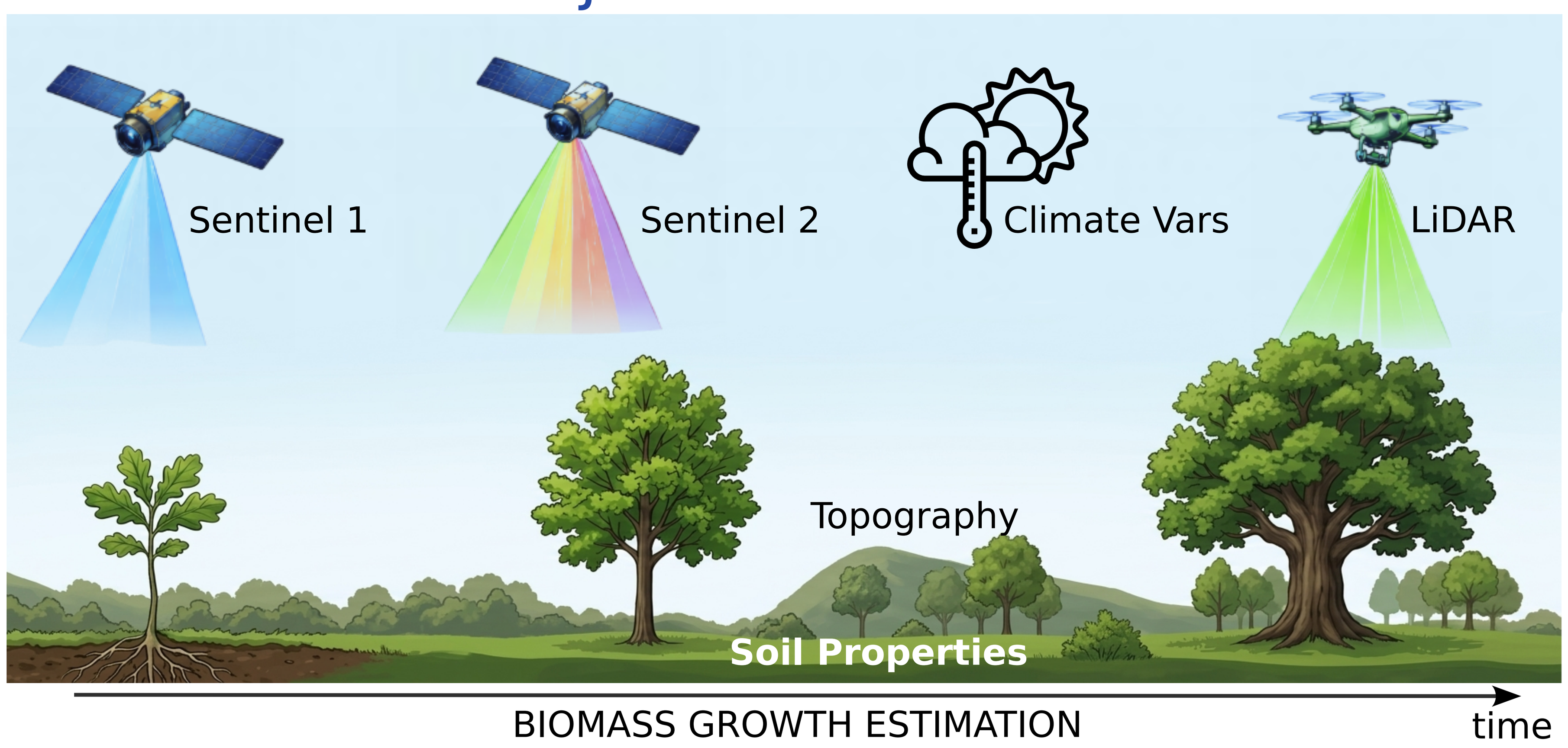


Figure 1. System overview with input data.

The European Union aims to enhance the forest carbon sink to achieve climate neutrality by 2050. Expanding this sink can be supported through nature-based solutions, such as carbon farming in forestry and agriculture, which require robust monitoring, reporting, and verification (MRV) systems.

This study presents a machine-deep-learning framework integrating multi-source data, i.e. ALS LiDAR, satellite, climate, soil, and topography, to estimate annual forest biomass increment at scale.

Data Driven Approach

Methodology

Forest growth is monitored by calculating biomass changes (ΔB) over specific time intervals. Net Annual Increment maps (ΔB/Δt), derived from multi-temporal airborne LiDAR, serve as the ground truth to train ML architectures. These models predict biomass growth by integrating satellite data with environmental variables, including soil, topography, forest types, climate, and disturbances.

ML Architectures

Each pixel is treated as an independent observation, emphasizing temporal information to estimate annual net biomass increment. Models are trained on 5% of sparse labeled points to evaluate their extrapolation capability. Models compared:

- **LightGBM** – Light Gradient Boosting Machine, a scalable gradient boosting decision tree model for structured data (Ke et al., 2017)
- **GANDALF** – Gated Adaptive Network for tabular data, a deep learning architecture with novel learned feature representation layers (Joseph & Raj, 2022)

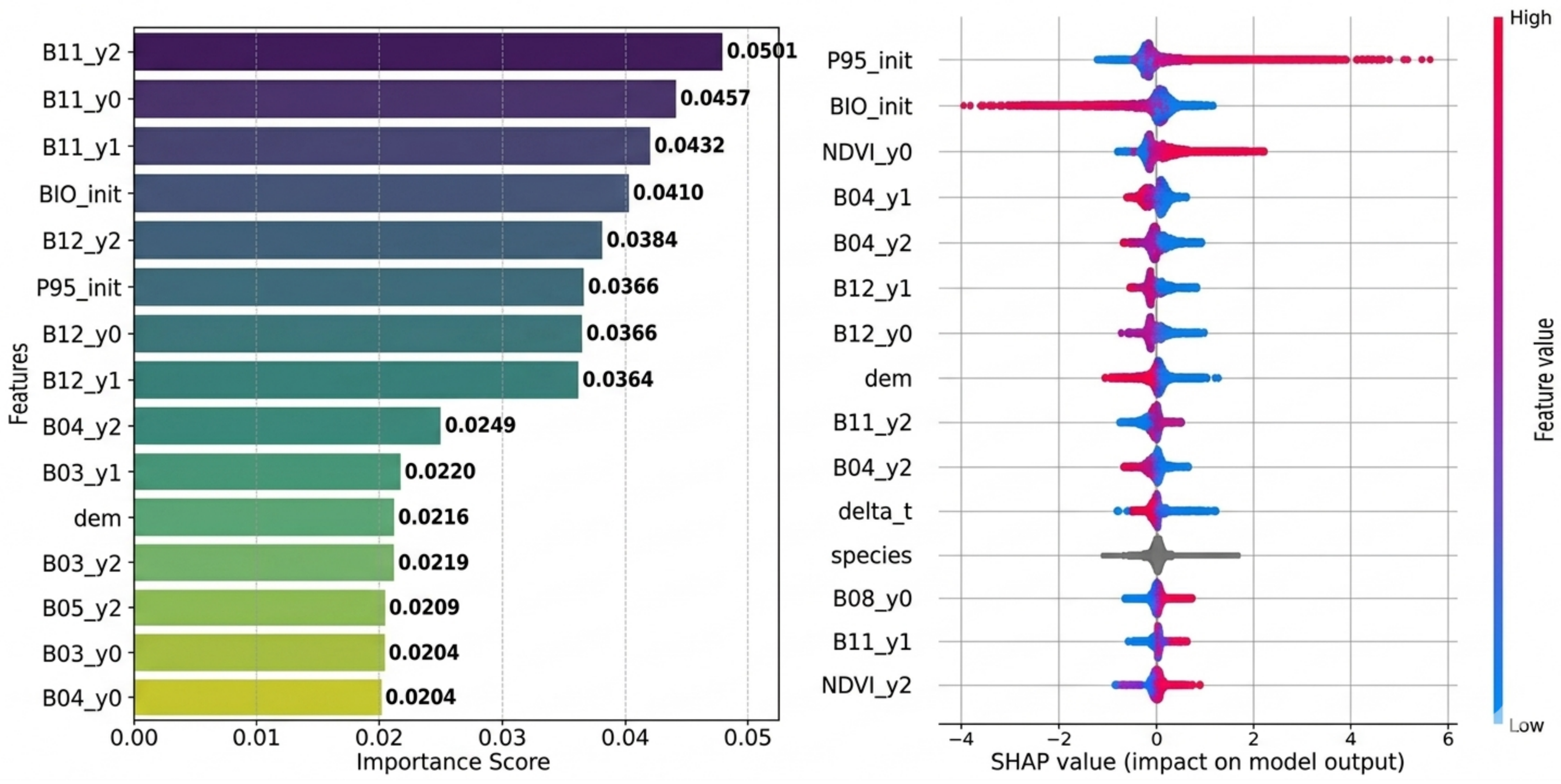


Figure 2. Global feature analysis for the proposed models.

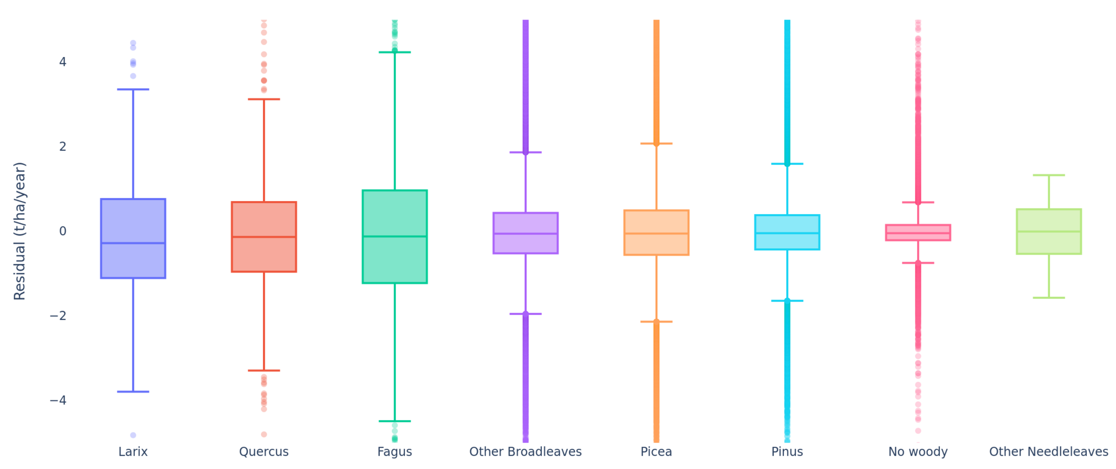
Experiments

Varying spatial resolution

SPATIAL RESOLUTION [M]	MODEL	RMSE	MAE	R^2
20	LightGBM	0.91	0.63	0.65
	GANDALF	0.93	0.66	0.63
100	LightGBM	0.53	0.37	0.79
	GANDALF	0.58	0.42	0.76
200	LightGBM	0.44	0.31	0.84
	GANDALF	0.48	0.35	0.80
2500	LightGBM	0.14	0.09	0.97
	GANDALF	0.21	0.15	0.93

Table 1. Metrics obtained using the proposed models.

Grouping Tree Species



SPECIES	MAE	RMSE	R^2	% DATA
No woody	0.29	0.60	0.43	4.68%
Pinus	0.57	0.79	0.64	46.76%
Other Broadleaves	0.70	1.04	0.66	11.61%
Picea	0.73	1.01	0.60	36.65%
Other Needleleaves	0.78	1.06	0.66	0.002%
Quercus	1.05	1.46	0.65	0.07%
Larix	1.17	1.57	0.53	0.04%
Fagus	1.44	1.94	0.65	0.18%

Figure 3. Boxplot per species and relative metrics with LightGBM (20 m).

Grouping Biomass Growth

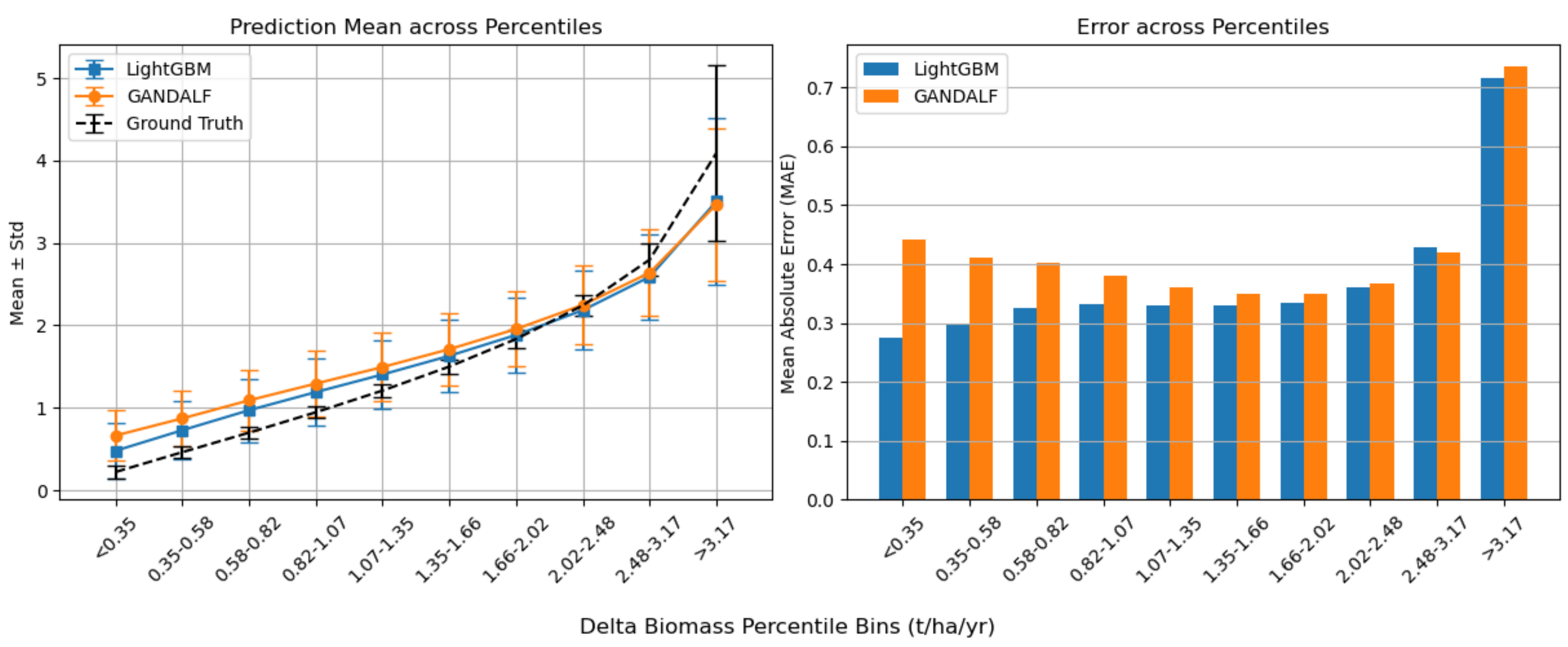


Figure 4. Comparison of prediction trends and MAE across ΔB/Δt percentiles.

Evaluation Maps

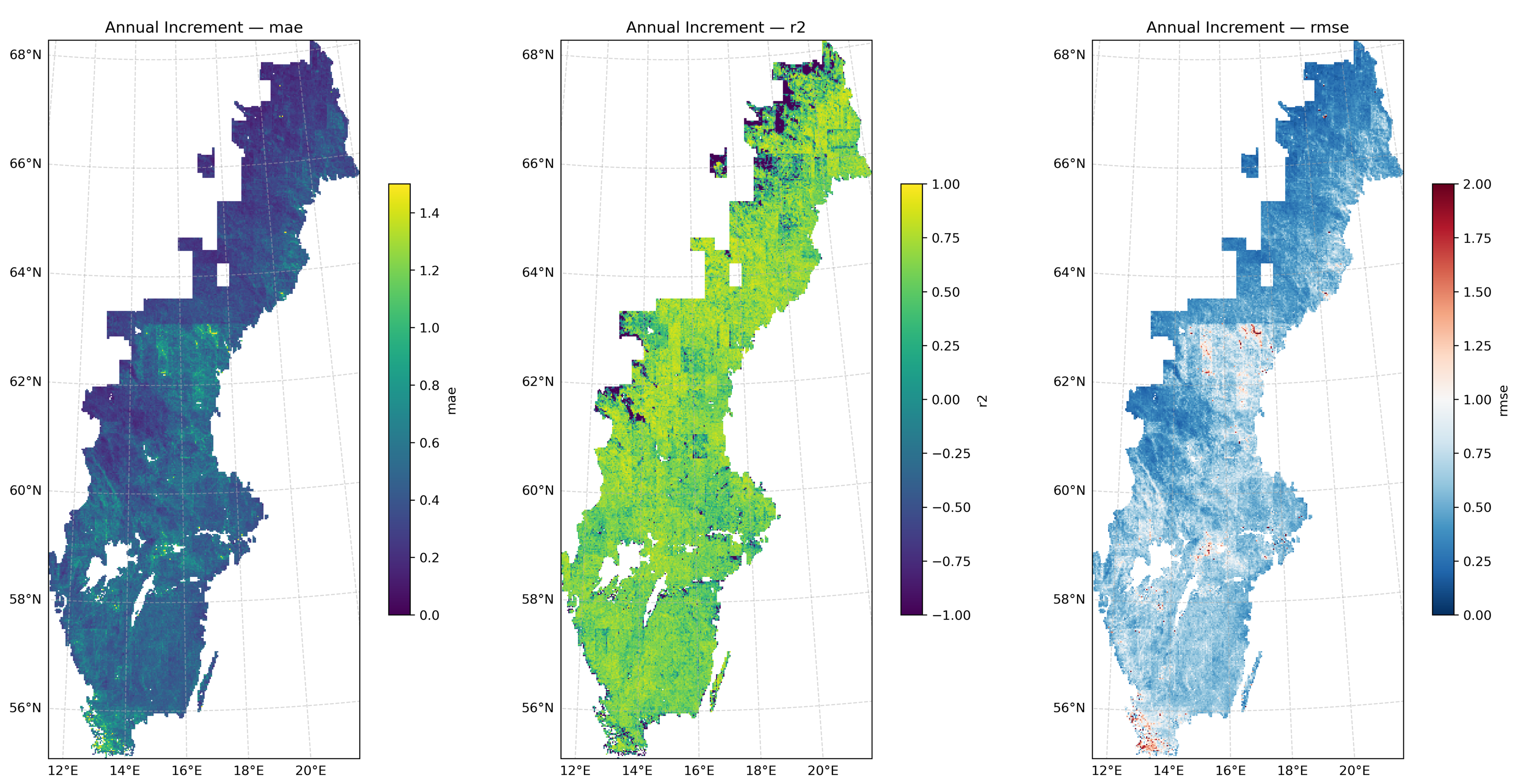


Figure 5. Geospatial evaluation of model performance (100 m).

Conclusions

- Both models capture overall forest growth, distinguishing low vs high growth;
- Uncertainty is consistently represented across percentiles;
- Performance varies by species; residuals remain unbiased (≈ 0).

Outlook

Future work will integrate Sentinel-1 data, adopt probabilistic uncertainty-aware predictions, and validate results against NFI data.

References

(Ke et al., 2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems 30.

(Joseph & Raj, 2022) GANDALF: Gated Adaptive Network for Deep Automated Learning of Features, arXiv.