VisTabNet: Adapting Vision Transformers for Tabular Data

Ulvi Movsum-zada

05.12.2024

Why is tabular data important?

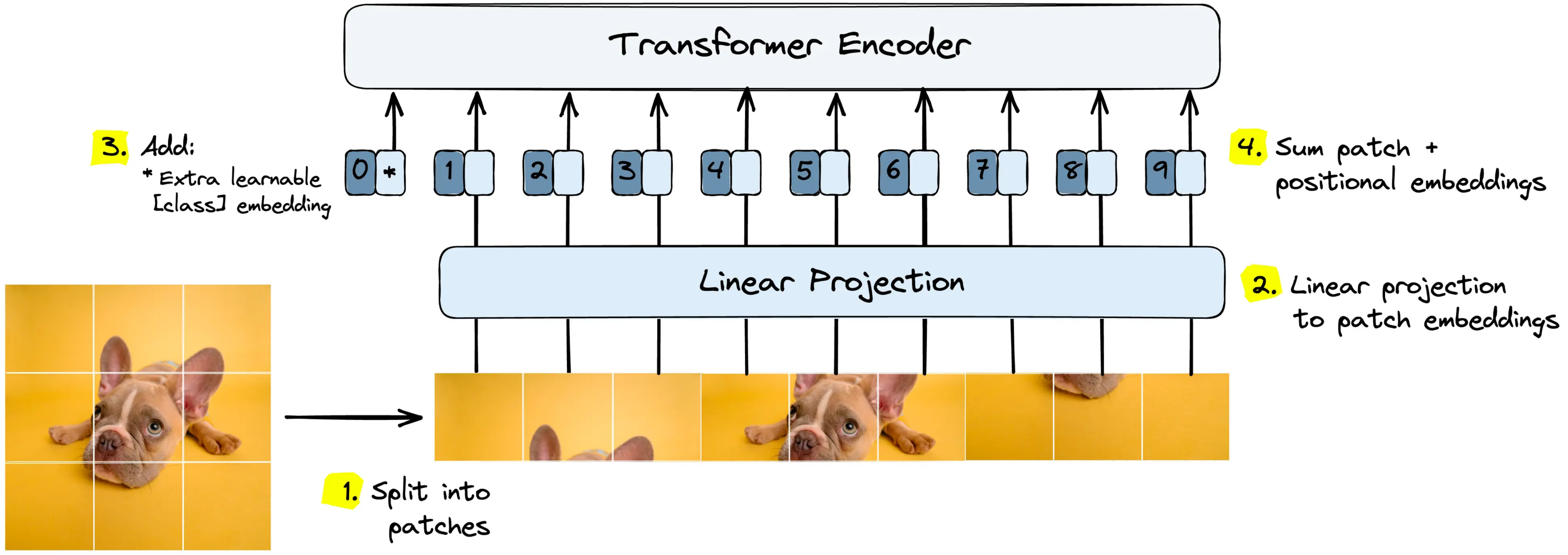
- ***P**revalent data type in domains like biology, physics, chemistry, finance, and industrial applications*
- **P**resents unique challenges due to its heterogeneity and small dataset sizes.

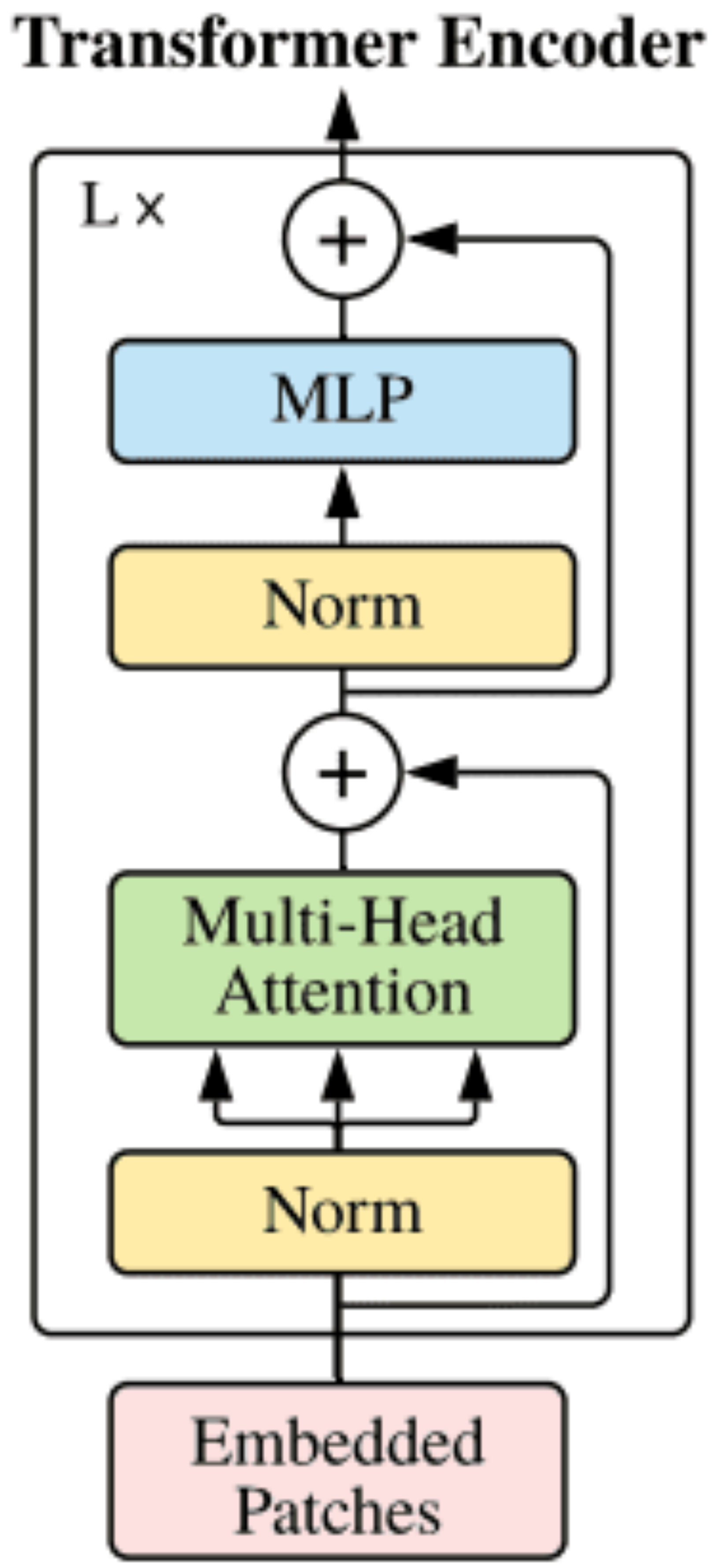
Datasets	**Image**	**Text**	**Tabular**
Kaggle	6799	3932	9983
Hugging Face	27699	24173	28829
Open ML	3540	2300	5400

Challenges with Existing Models

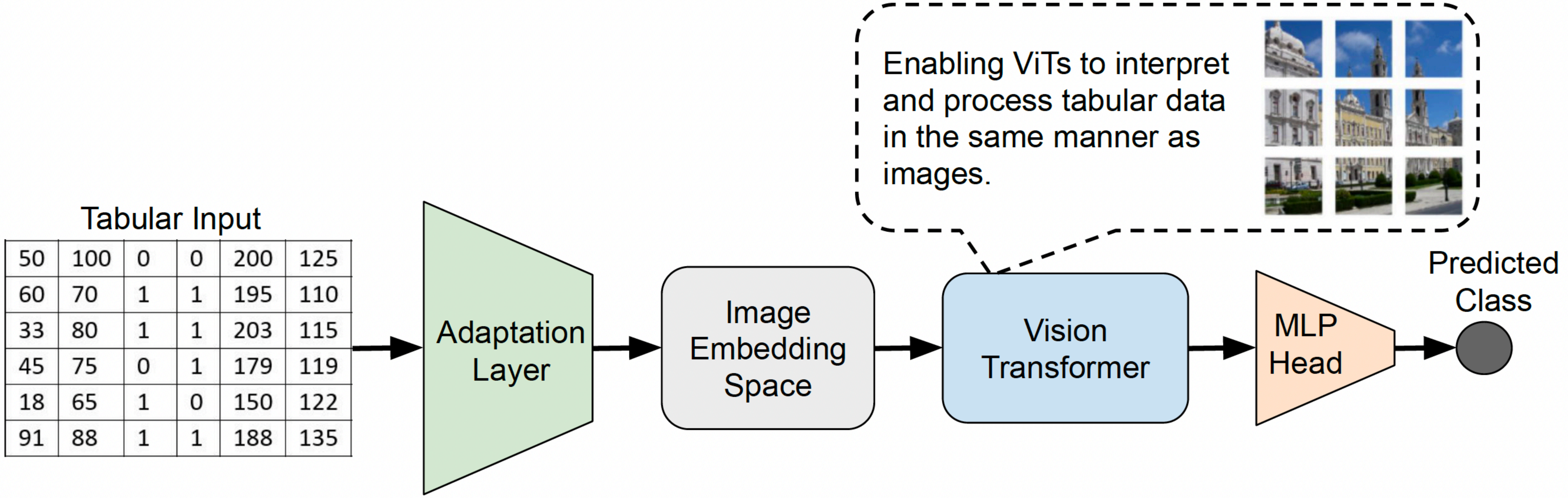
- Tree-Based Models are Hard to Beat: Ensemble models like **XGBoost, Random Forests, and Gradient Boosting Machines** have consistently outperformed deep learning models on tabular datasets, especially on small to medium-sized datasets.
- Difficulty in Handling Feature Types: Categorical data, Missing data
- Transfer Learning Challenges: Limited Pre-trained Models for Tabular Data, Lack of standardisation

Vision Transformer - ViT



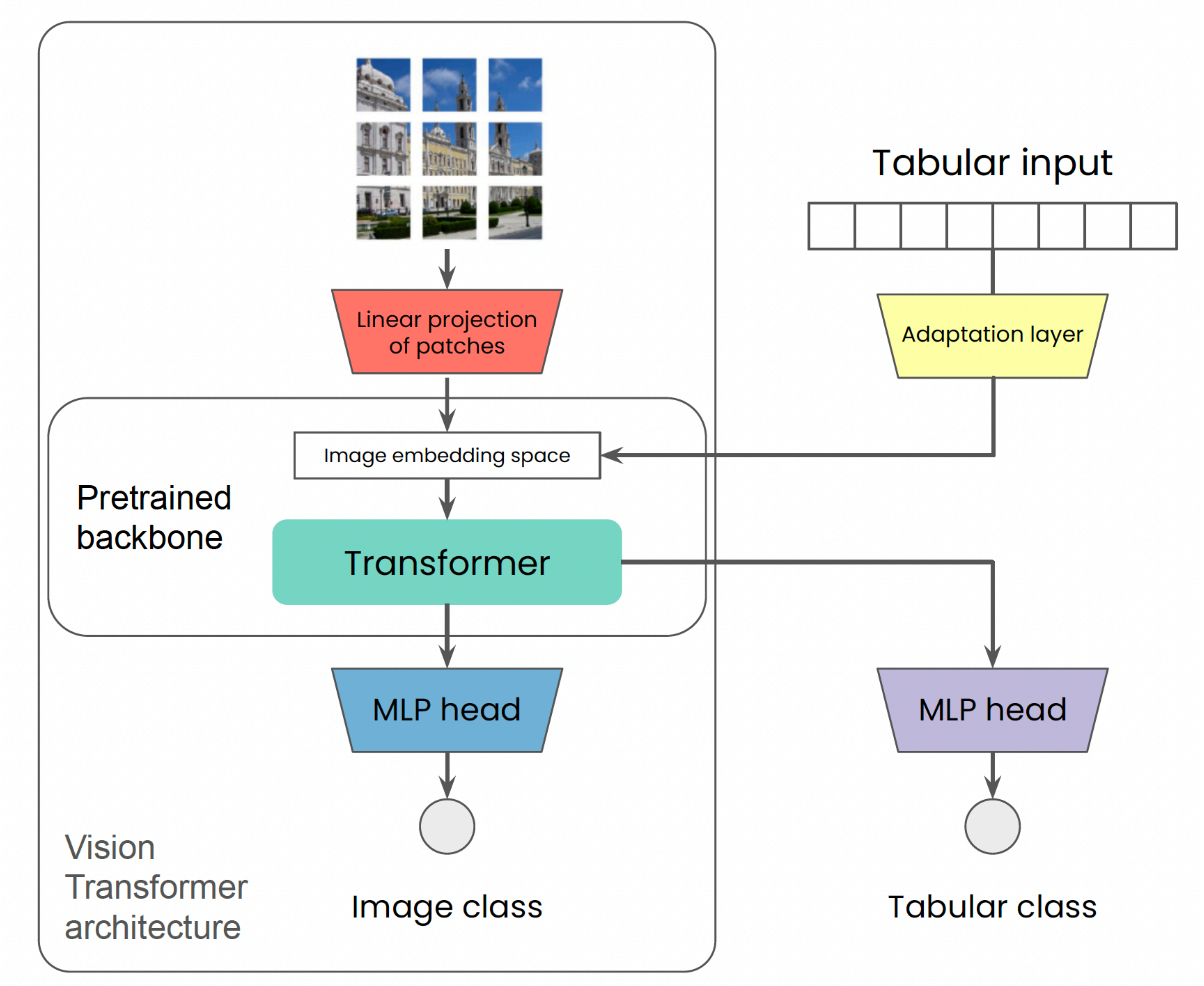


Overview of VisTabNet



Key components of the VisTabNet

- **Adaptation Layer**
- **Vision Transformer Backbone**
- **Cross-modal Transfer Learning**



Results and Benchmarks

Dataset	VisTabNet	RF	XGBoost	GB	LightGBM	ResNet	FT	NODE
Blood transf.	31.3 ± 7	22.0 ± 3	30.4 ± 4	30.4 ± 4	30.4 ± 4	**45.3 ± 6**	41.6 ± 6	28.5 ± 6
Wisconsin	**65.3 ± 5**	33.0 ± 3	30.6 ± 4	30.6 ± 4	30.6 ± 4	30.6 ± 5	31.7 ± 5	30.6 ± 2
Breast Cancer	91.1 ± 4	88.4 ± 2	80.7 ± 3	87.0 ± 3	89.6 ± 3	**97.3 ± 6**	94.6 ± 4	92.5 ± 18
Connectionist	**84.6 ± 5**	69.0 ± 3	76.2 ± 4	74.6 ± 4	63.6 ± 4	64.5 ± 7	37.7 ± 5	76.3 ± 4
Congr. Voting	91.5 ± 4	93.7 ± 2	91.7 ± 3	**95.7 ± 3**	90.3 ± 3	73.9 ± 6	79.9 ± 4	89.7 ± 2
Credit Approval	67.5 ± 1	74.1 ± 3	74.3 ± 4	71.1 ± 4	74.1 ± 4	65.9 ± 7	74.9 ± 5	**79.9 ± 5**
Cylinder bands	**45.0 ± 4**	44.3 ± 3	33.4 ± 4	33.4 ± 4	42.7 ± 4	43.7 ± 6	39.7 ± 6	44.4 ± 8
Dermatology	95.3 ± 1	**96.5 ± 2**	95.3 ± 3	93.1 ± 3	95.2 ± 3	84.9 ± 6	92.3 ± 4	91.1 ± 3
Ecoli	72.1 ± 5	76.2 ± 3	70.3 ± 4	68.3 ± 4	70.2 ± 4	87.1 ± 7	89.6 ± 5	**90.1 ± 4**
Glass	93.9 ± 4	93.8 ± 2	95.9 ± 3	95.9 ± 3	95.9 ± 3	64.6 ± 6	58.0 ± 4	**100.0 ± 0**
Haberman	**50.2 ± 6**	24.6 ± 3	27.8 ± 4	25.8 ± 4	30.4 ± 4	27.1 ± 7	40.1 ± 6	31.8 ± 12
Horse Colic	50.6 ± 5	**75.4 ± 3**	75.1 ± 4	75.1 ± 4	58.1 ± 4	43.1 ± 8	43.1 ± 5	57.4 ± 3
Ionosphere	87.7 ± 4	83.4 ± 2	79.4 ± 3	77.3 ± 3	69.6 ± 3	87.0 ± 6	**95.7 ± 4**	77.6 ± 19
Libras	**84.4 ± 3**	70.7 ± 3	66.9 ± 4	63.0 ± 4	70.7 ± 4	77.5 ± 7	59.7 ± 5	59.7 ± 5
Lymphography	70.7 ± 5	66.8 ± 3	47.7 ± 4	66.8 ± 4	41.4 ± 4	58.9 ± 7	42.7 ± 5	**72.1 ± 19**
Mammographic	60.1 ± 5	68.6 ± 3	72.6 ± 4	69.3 ± 4	70.9 ± 4	72.5 ± 6	**73.8 ± 5**	64.7 ± 12
Primary Tumor	**40.1 ± 6**	30.6 ± 3	34.6 ± 4	36.0 ± 4	35.2 ± 4	32.5 ± 7	39.1 ± 6	39.6 ± 9
Sonar	63.0 ± 5	63.0 ± 3	62.2 ± 4	63.0 ± 4	**68.8 ± 4**	36.0 ± 7	78.0 ± 5	60.1 ± 4
Statlog Australian	70.9 ± 5	71.8 ± 3	72.0 ± 4	73.5 ± 4	71.3 ± 4	67.5 ± 7	**74.9 ± 5**	60.8 ± 6
Statlog German	29.3 ± 6	**43.1 ± 3**	39.2 ± 4	39.2 ± 4	39.2 ± 4	41.0 ± 7	37.3 ± 6	42.5 ± 14
Statlog Heart	40.3 ± 5	55.4 ± 3	58.3 ± 4	52.4 ± 4	52.4 ± 4	62.3 ± 7	**78.0 ± 5**	43.7 ± 3
Vertebral	70.6 ± 5	**74.6 ± 3**	73.5 ± 4	58.7 ± 4	71.9 ± 4	67.6 ± 7	68.9 ± 5	65.7 ± 4
Zoo	94.3 ± 2	94.6 ± 2	94.6 ± 3	**100.0 ± 0**	94.6 ± 1	81.0 ± 6	81.0 ± 4	94.6 ± 6
Mean	**67.43**	65.81	64.47	64.36	63.35	61.38	63.14	64.93
Mean rank	**3.93**	4.04	4.39	4.91	4.87	5.17	4.24	4.43

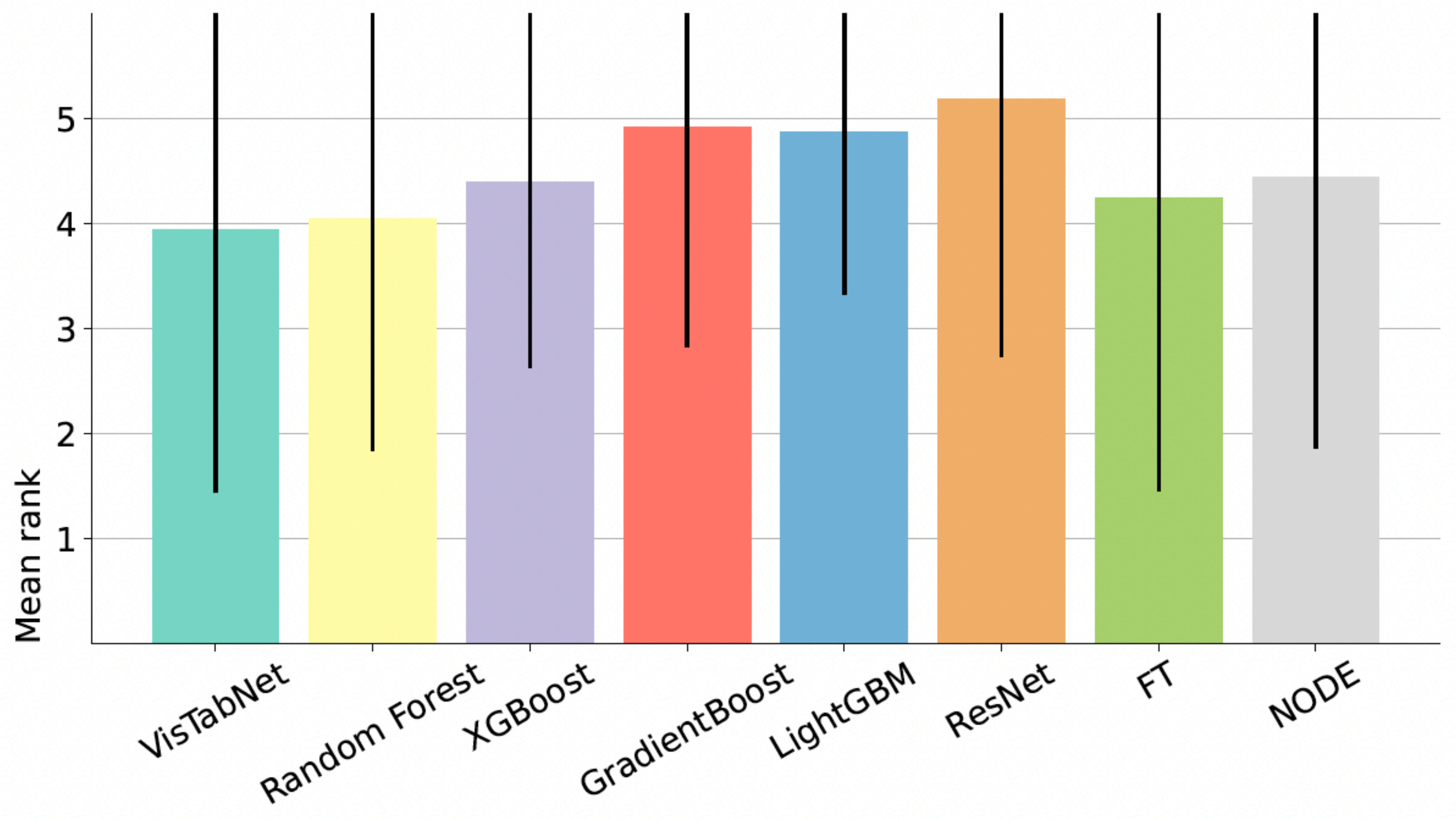


Figure : Comparison of average ranking with standard deviation as whiskers (the lower the better).

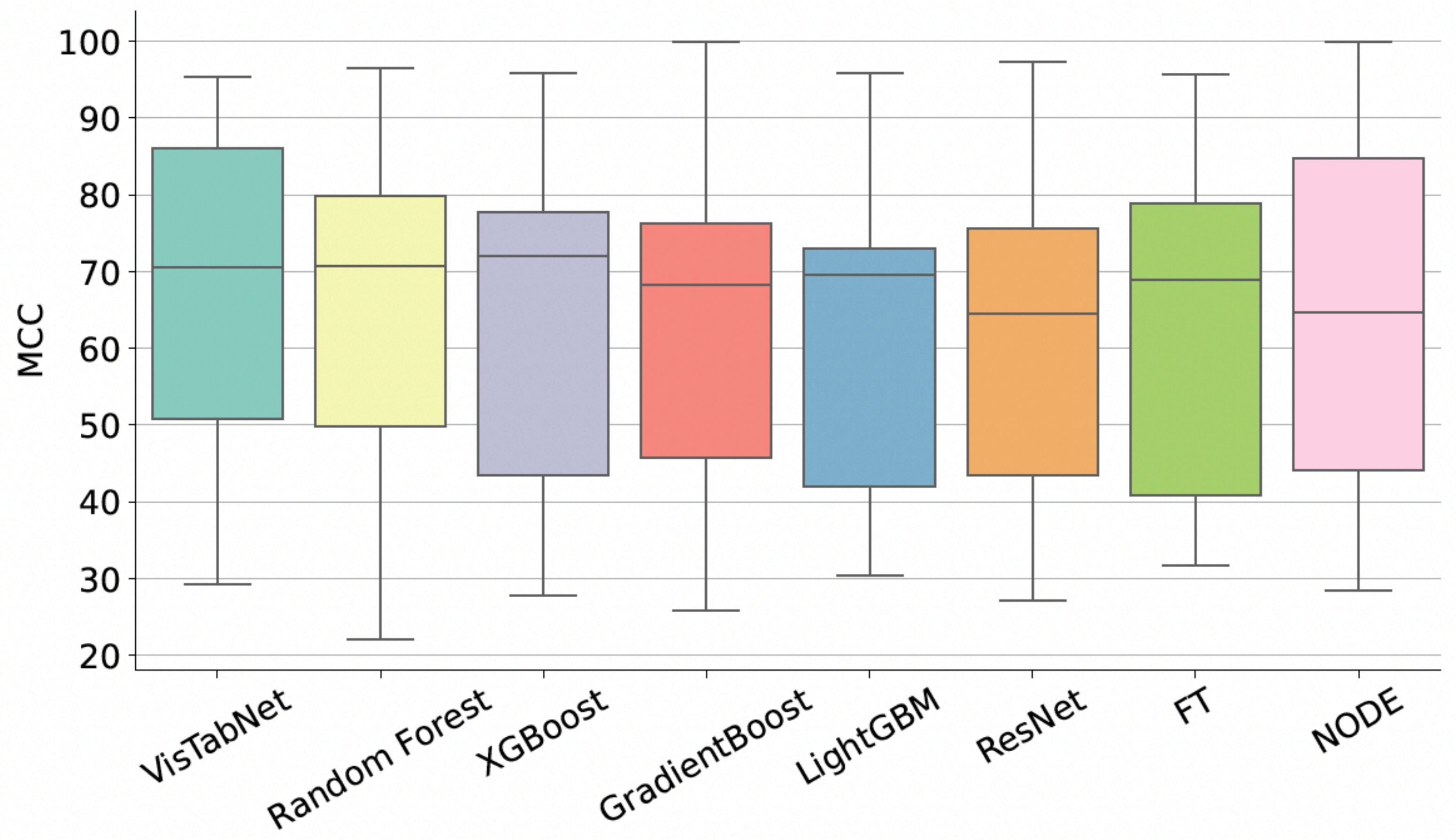


Figure : Comparison of the MCC score distributions (the higher the better).

Backbone selection

Table : Dependence of VisTabNet on the backbone size. VisTabNet consistently outperforms a regular neural network with an analogical number of trainable parameters by a large margin, indicating that the use of ViT is essential in achieving good performance.

Dataset	VisTabNet (B)	VisTabNet (B, fully trained)	VisTabNet (B, finetuned)	VisTabNet (L)	Dense (size of VisTabNet)
Dermatology	0.930	0.930	0.920	**0.957**	0.842
Libras	0.843	0.812	**0.853**	0.812	0.701
ZOO	**0.946**	0.838	0.891	0.838	0.733
Cylinder Bands	**0.426**	0.418	**0.426**	0.413	0.407
Credit approval	0.651	0.639	**0.665**	0.626	0.580
Volkert	**0.646**	0.631	**0.647**	0.639	0.621
Nomao	**0.745**	0.722	**0.745**	0.712	0.623

Few-shot transfer learning

- **Traditional Learning:** conventional convolutional neural networks directly on the limited MNIST dataset.
- **Fine-tuning from FashionMNIST:** pretraining a convolutional model on the FashionMNIST dataset before fine-tuning it on the constrained MNIST dataset.

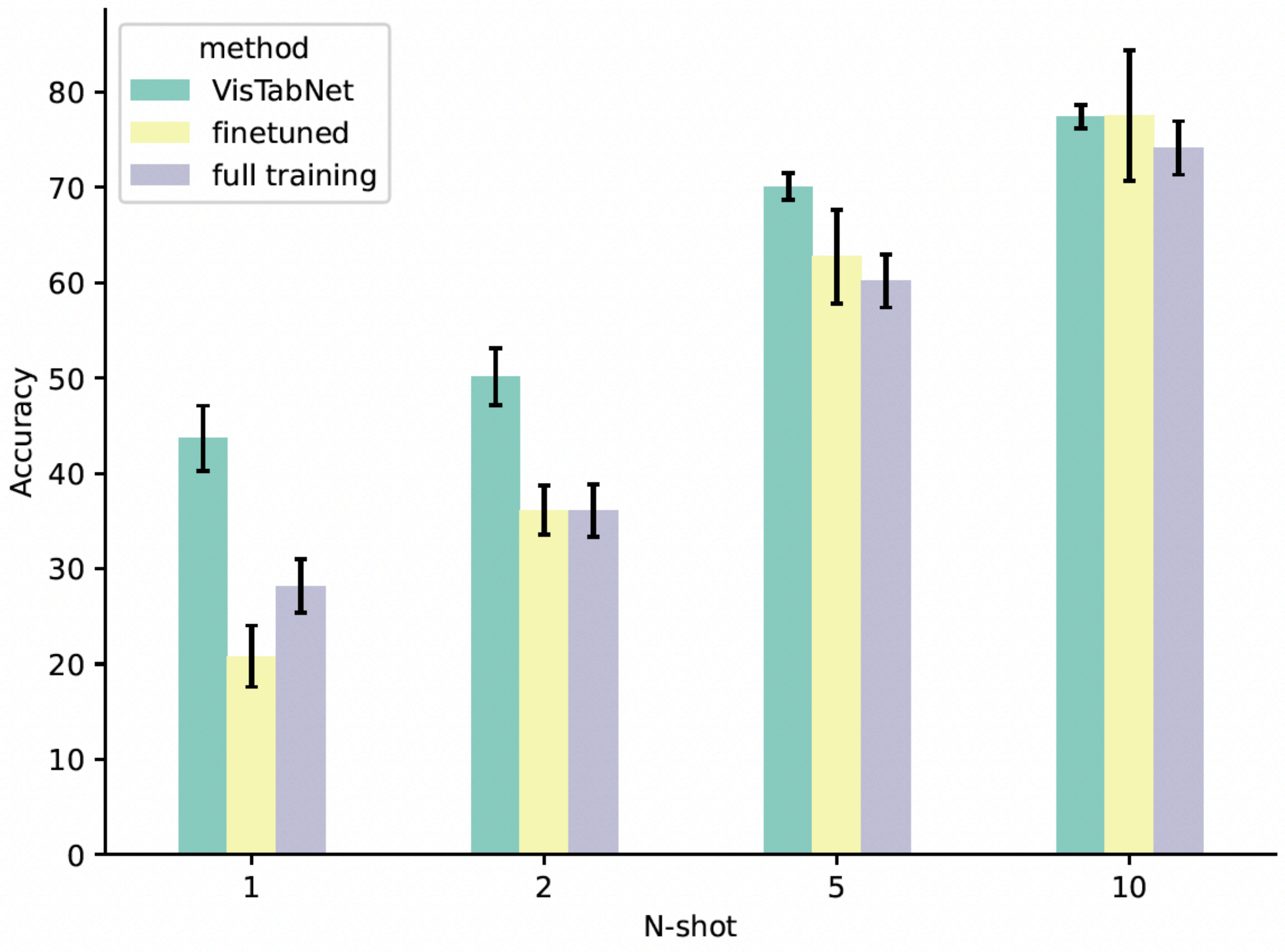


Figure : Results after learning on artificially limited number of samples. VisTabNet achieves significantly better scores in the few-shot setting, consistently outperforming other training methods up to 10-shot.

Summary of Contributions

- Cross-Modal Transfer Learning for Tabular Data
- Reduced Conceptual Cost
- State-of-the-Art Performance on Small Data
- Versatile Application of Vision Transformers

Future Directions

- Exploring Different Pre-Trained Models
- Optimising the Adaptation Process
- Incorporating Feature Engineering Techniques
- Broader Application Domains

Thank you!