



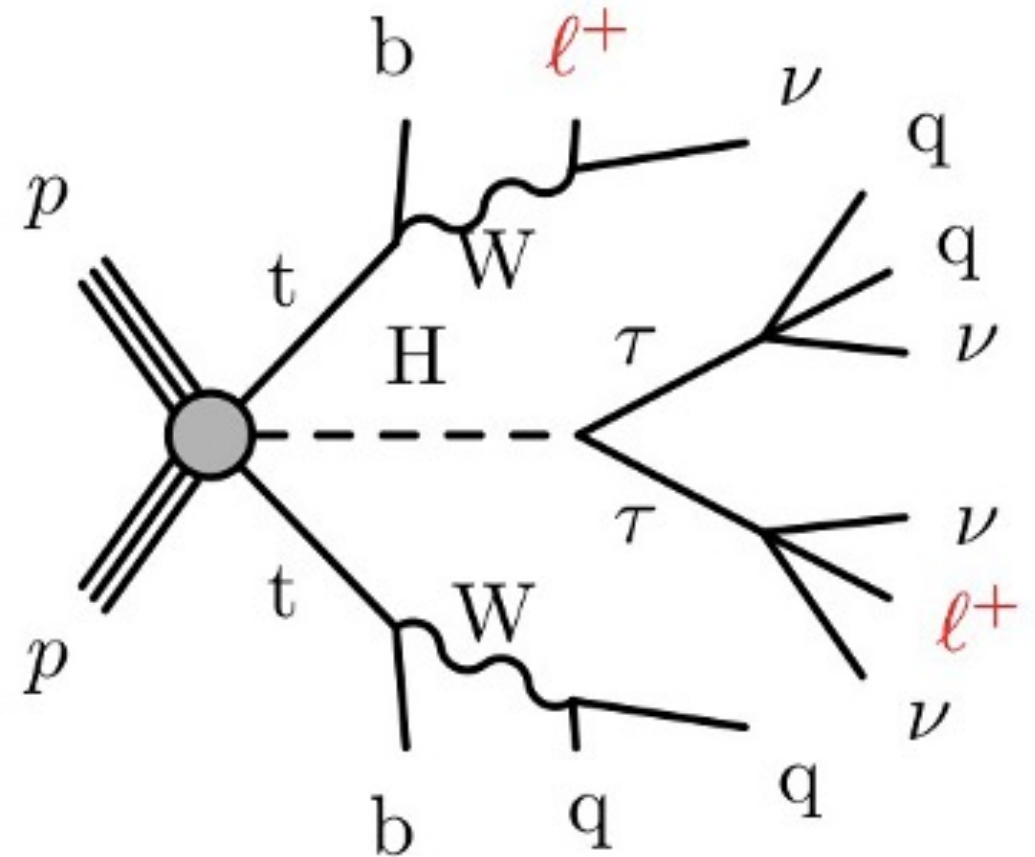
# Summer Student Project: Machine Learning Classification of tth Events at the ATLAS Experiment

Severin König, Supervisor: André Sopczak

03.08.2022

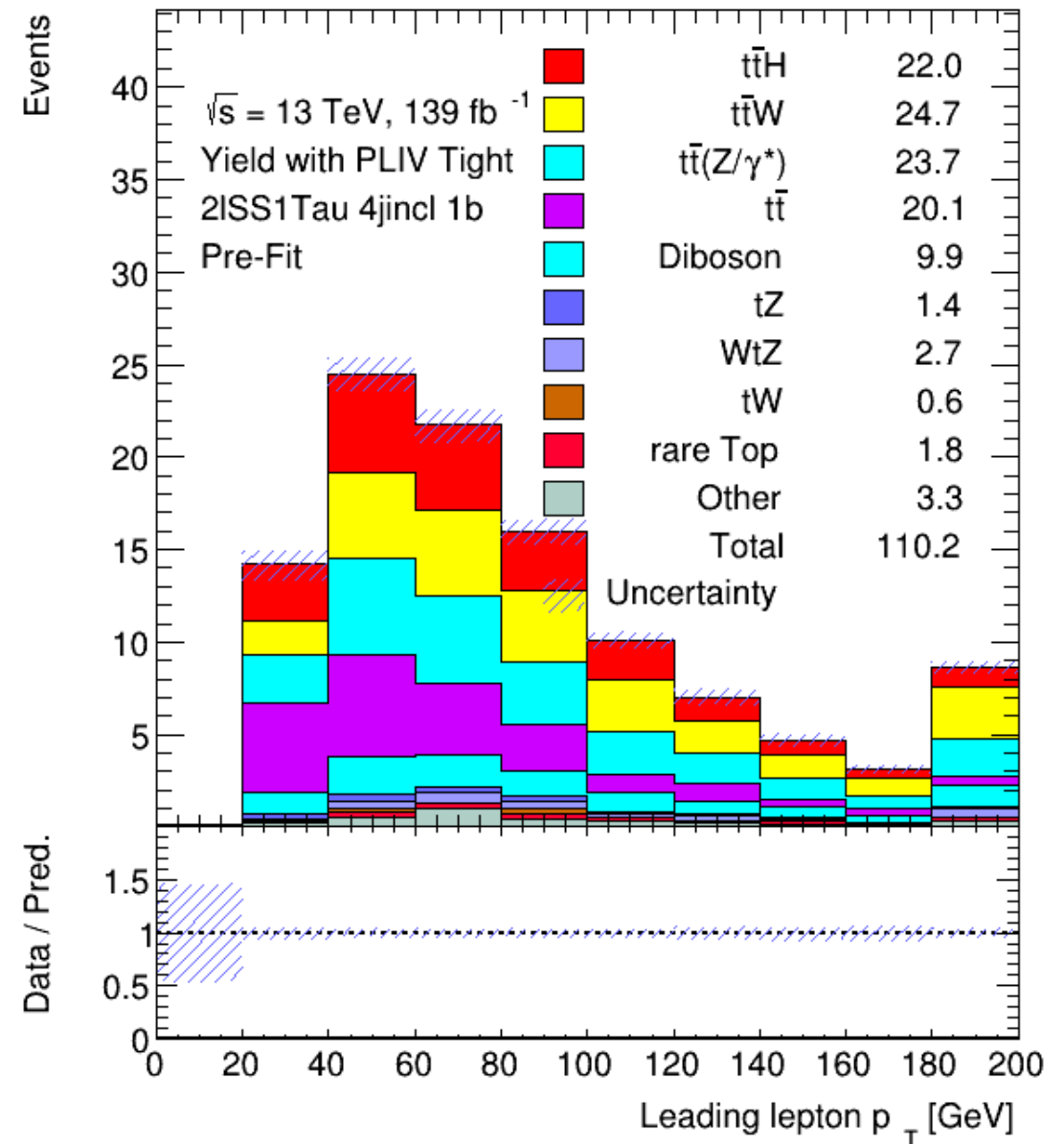
# Theoretical Background

- **Process of Interest: ttH**
- **Important process for study of the Higgs boson at CERN**
- **Both top quarks and Higgs have several decay modes.**
- **Focus on 2LSS1Tau channel which requires two same-sign light leptons and a hadronically decaying tau**
- **Task: machine learning optimization following up on the Masters thesis [1]**



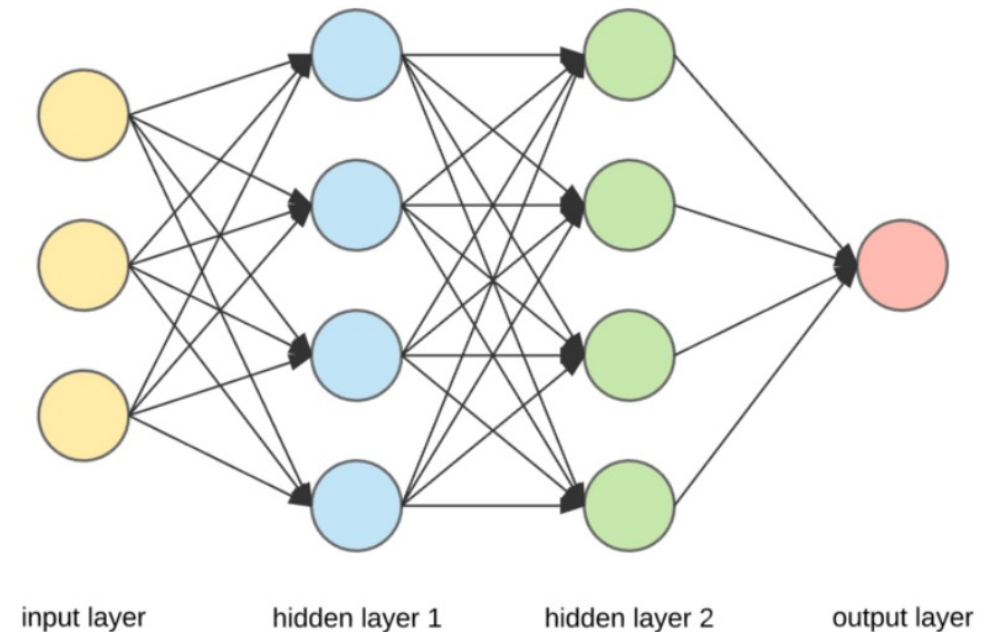
# Theoretical Background

- Large number of background reactions require machine learning to separate ttH signal
- TRexFitter [2] used to illustrate preselected events as input for the machine learning



# Neural Network Approach

- **Commonly used machine learning algorithm: Multi-Layer Perceptron(MLP)**
- **Choose Network Architecture, i.e. number of nodes and layers, and other hyperparameters**
- **Train with simulated input data using weights to scale to expected events**
- **Network output: probability score (0-1) for signal and background classes**
- **Alternative method also tested: Gradient boosting**

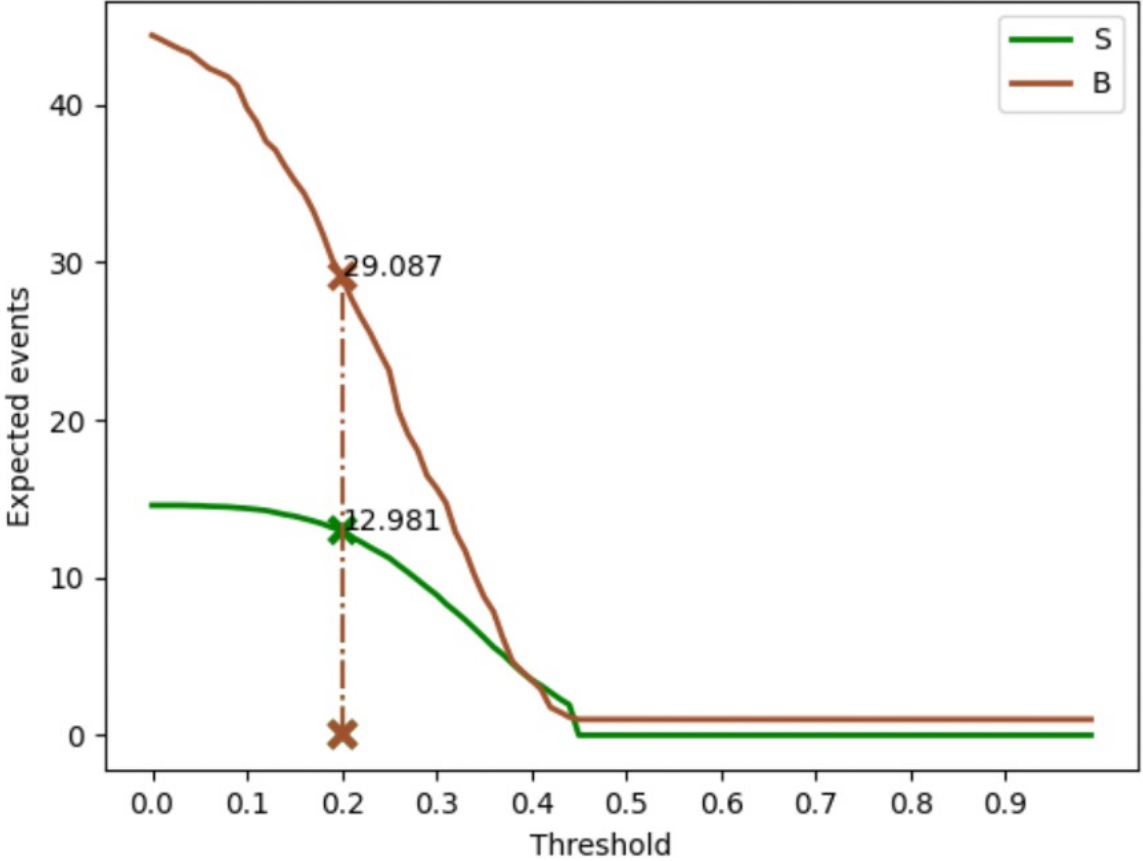


# Results: Optimal threshold search

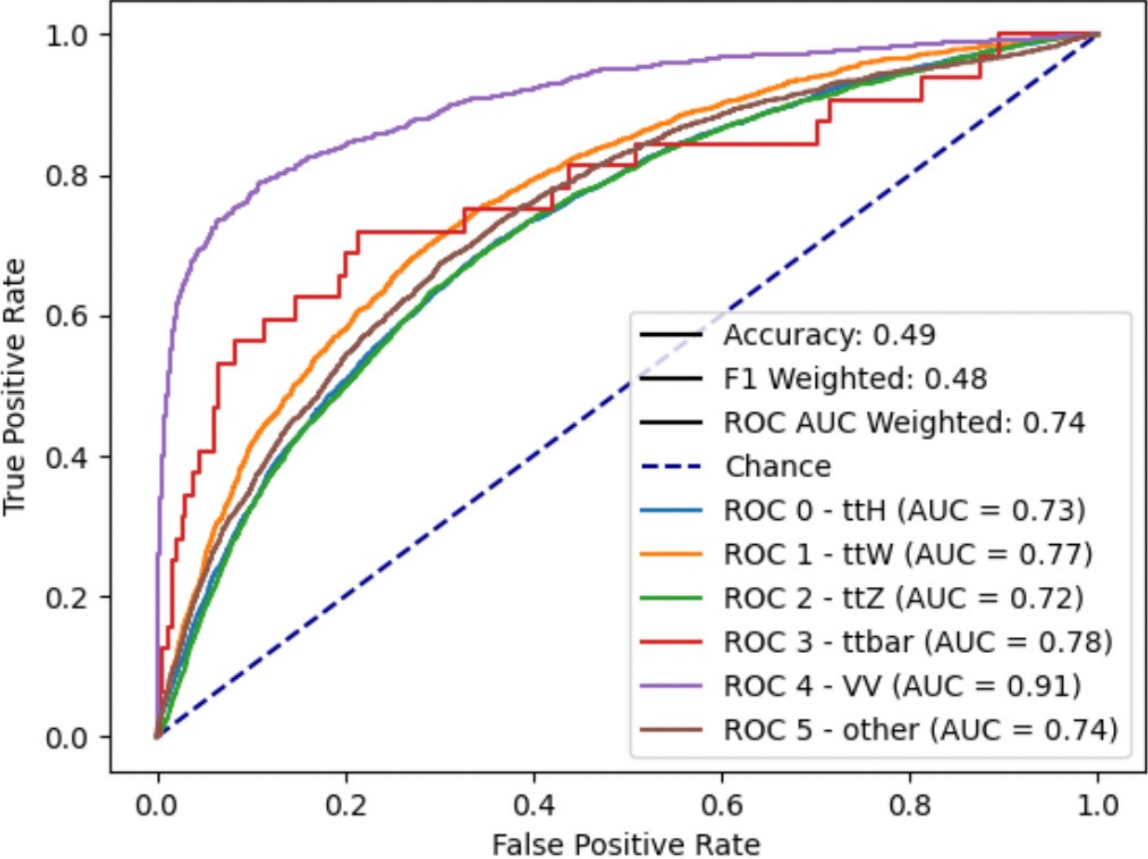
Signal/Background vs threshold

ROC curve

S & B to Threshold characteristics

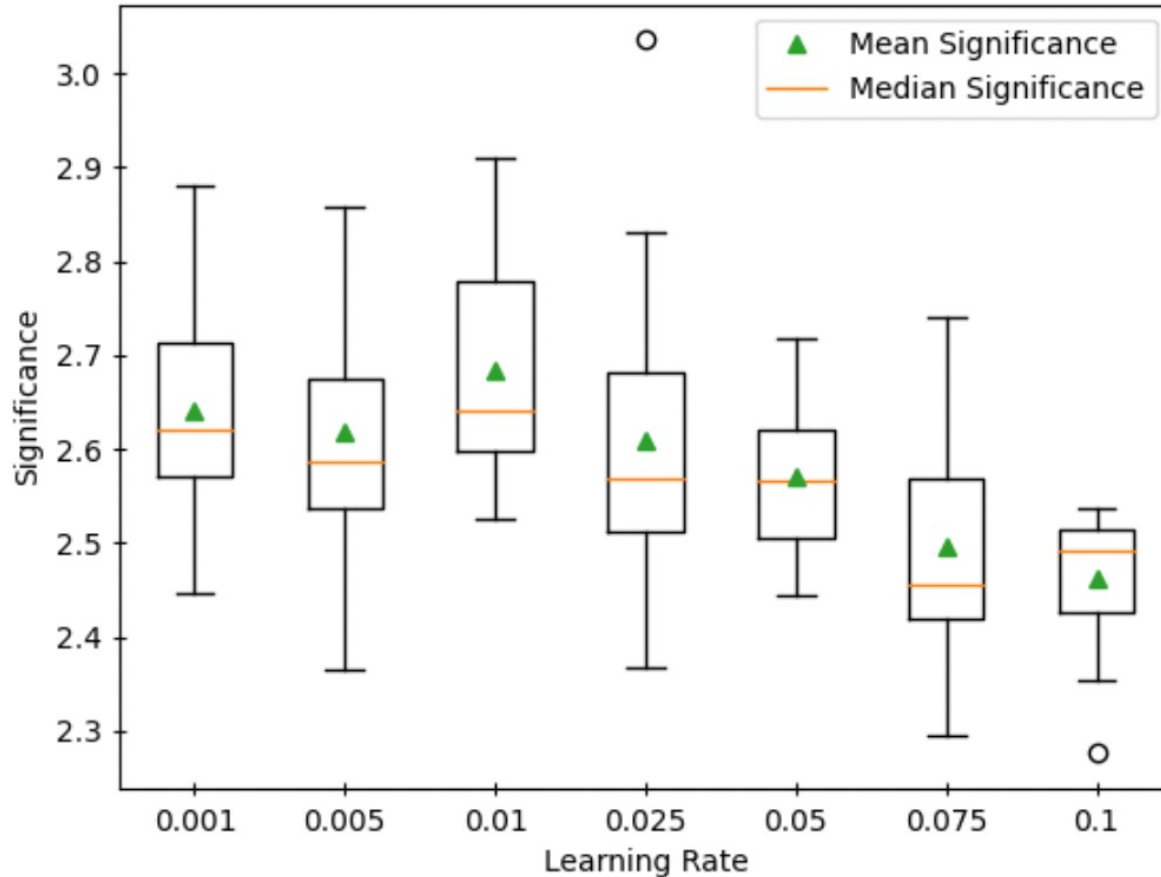


Model Evaluation



# Results: Hyperparameter Optimization

Significance vs Learning Rate



- Significance approximated as

$$\frac{S}{\sqrt{B}}, \text{ where } S = \text{no. of signal events,}$$

**B = no. of background events**

- Optimization also performed for other parameters

# Summary

- **Study of ttH production with machine learning**
- **Optimization of machine learning with different preselections**
- **Goal: increasing sensitivity for ttH events by reducing background events**

- **References:**

[1]: Presperín, Jan. *Machine Learning for ttH Mechanism Higgs Boson Detection from CERN ATLAS Data*. May 2022.

[2]: TRExFitter: <https://trexfitter-docs.web.cern.ch/trexfitter-docs/>

# Supplement



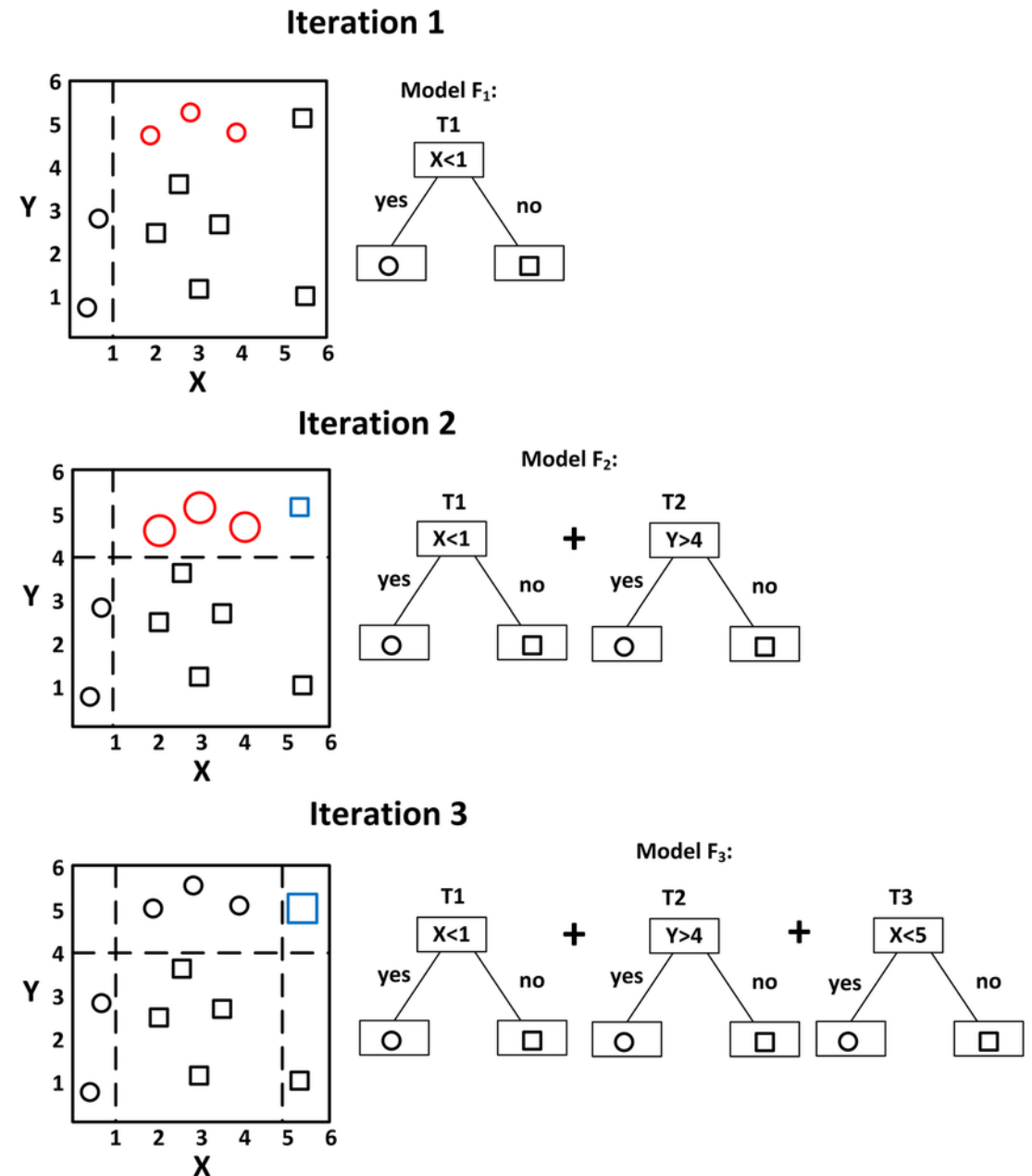
# Confusion Matrix

CM - test (0.32 threshold)

Actual \ Predicted	tFH	tFW	tFZ	tF	W	other	
tFH	2.67 11.72%	1.03 4.52%	0.57 2.50%	0.06 0.26%	0.1 0.44%	0.1 0.44%	4 58.84% 41.06%
tFW	0.69 3.03%	3.37 14.79%	0.72 3.16%	0.05 0.22%	0.2 0.88%	0.06 0.26%	5 66.21% 33.79%
tFZ	1.2 5.27%	1.99 8.73%	1.3 5.70%	0.08 0.35%	0.13 0.57%	0.07 0.31%	4 37.25% 72.75%
tF	0.53 2.33%	1.27 5.57%	0.73 3.20%	1.21 5.31%	0.28 1.23%		4 30.10% 69.90%
W	0.1 0.44%	0.31 1.36%	0.09 0.39%	0.08 0.35%	1.36 5.97%	0.07 0.31%	2 67.66% 32.34%
other	0.69 3.03%	0.69 3.03%	0.14 0.61%	0.04 0.18%	0.61 2.68%	0.2 0.88%	2 8.44% 91.56%
	5 45.41% 54.59%	8 38.91% 61.09%	3 36.62% 63.38%	1 79.61% 20.39%	2 50.75% 49.25%	0 40.00% 60.00%	22 44.36% 55.64%
	tFH	tFW	tFZ	tF	W	other	

# Gradient Boosting

- Alternative approach to machine learning algorithm
- Based on minimizing loss function by combining multiple "weak learners" which correct each others mistakes
- In analogy with MLP analysis, hyperparameter optimization is performed
- Image credit: Zhang et. al. *Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning*. July 2018.



# Results: Feature Importance

