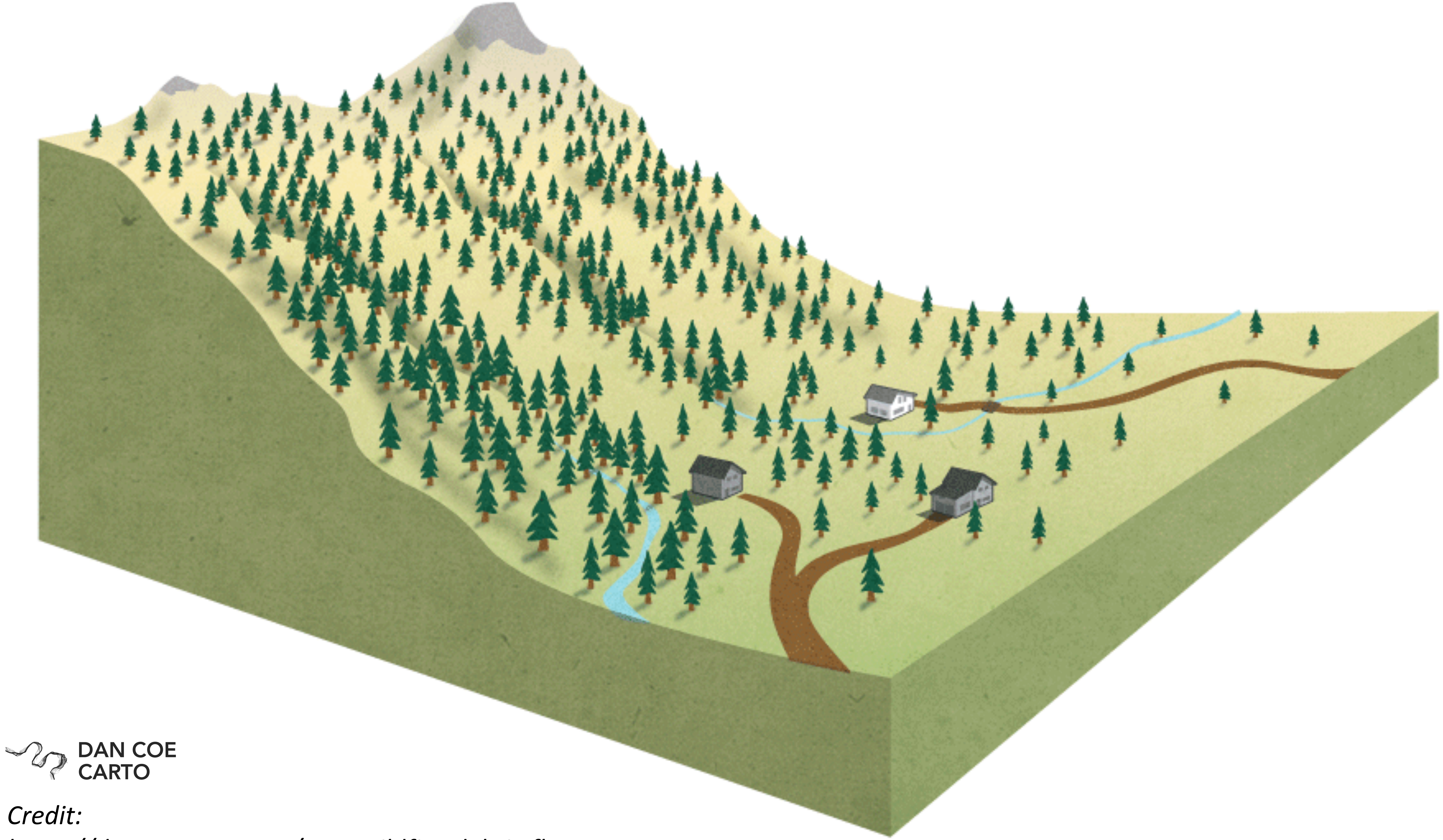
# Machine Learning for Improved Post-fire Debris Flow Likelihood Prediction

*Daniel Roten, Jessica Block, Daniel Crawl, Jenny Lee and Ilkay Altintas*

*IEEE Big Data 2022*
*December 17-20, Osaka, Japan*

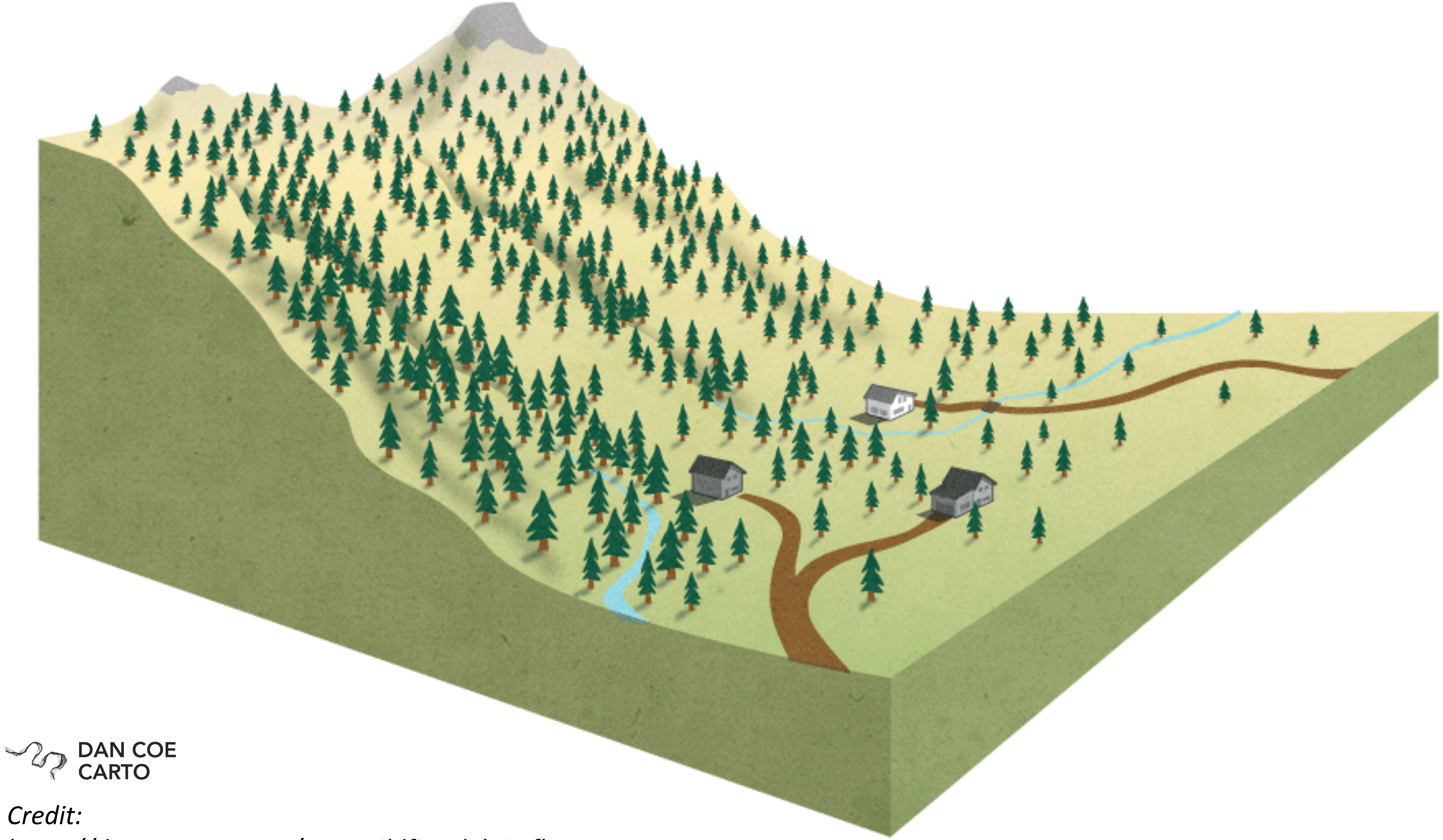
# Post-fire Debris Flows




*Credit: https://dancoecarto.com/post-wildfire-debris-flows*

- Debris flows are fast-moving masses of rock, soil and water.
- Areas impacted by wildfire are particularly prone to debris flows.
- Wildfires make soils water-repellent, increasing runoff.
- Loss of vegetation increases erosion

# Post-fire Debris Flows

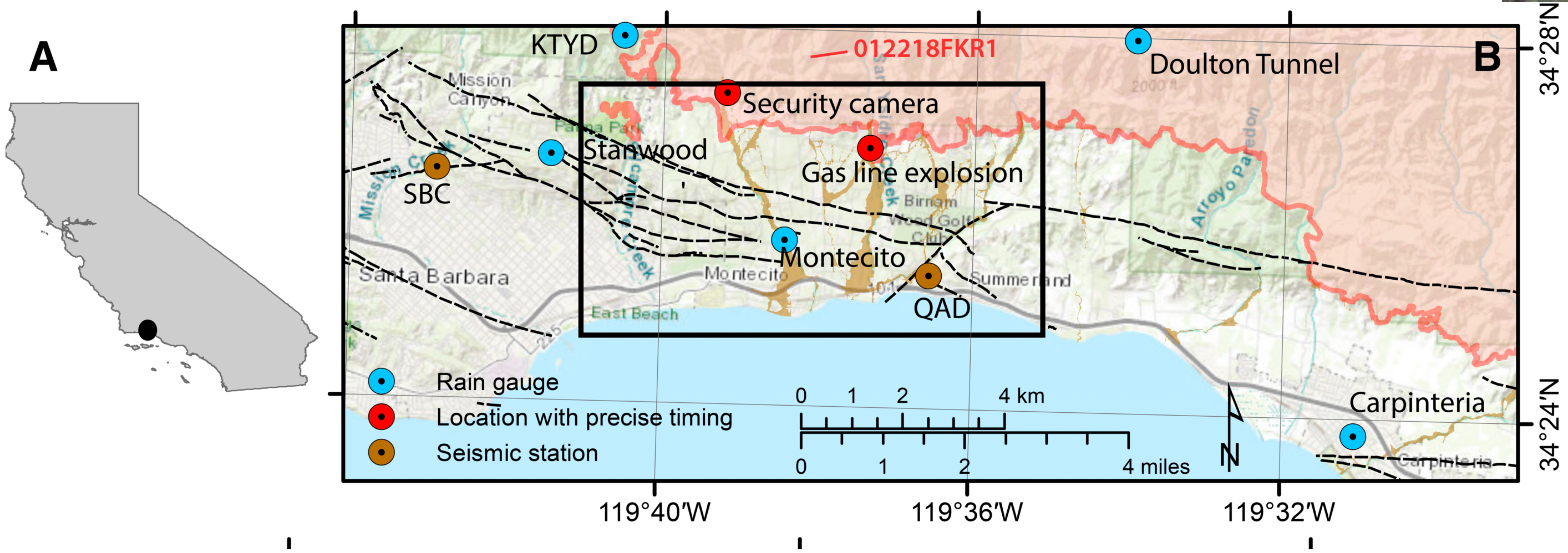



*Credit:*
*https://dancoecarto.com/post-wildfire-debris-flows*

- Debris flows are fast-moving masses of rock, soil and water.
- Areas impacted by wildfire are particularly prone to debris flows.
- Wildfires make soils water-repellent, increasing runoff.
- Loss of vegetation increases erosion

# 2018 Montecito Debris Flow Events

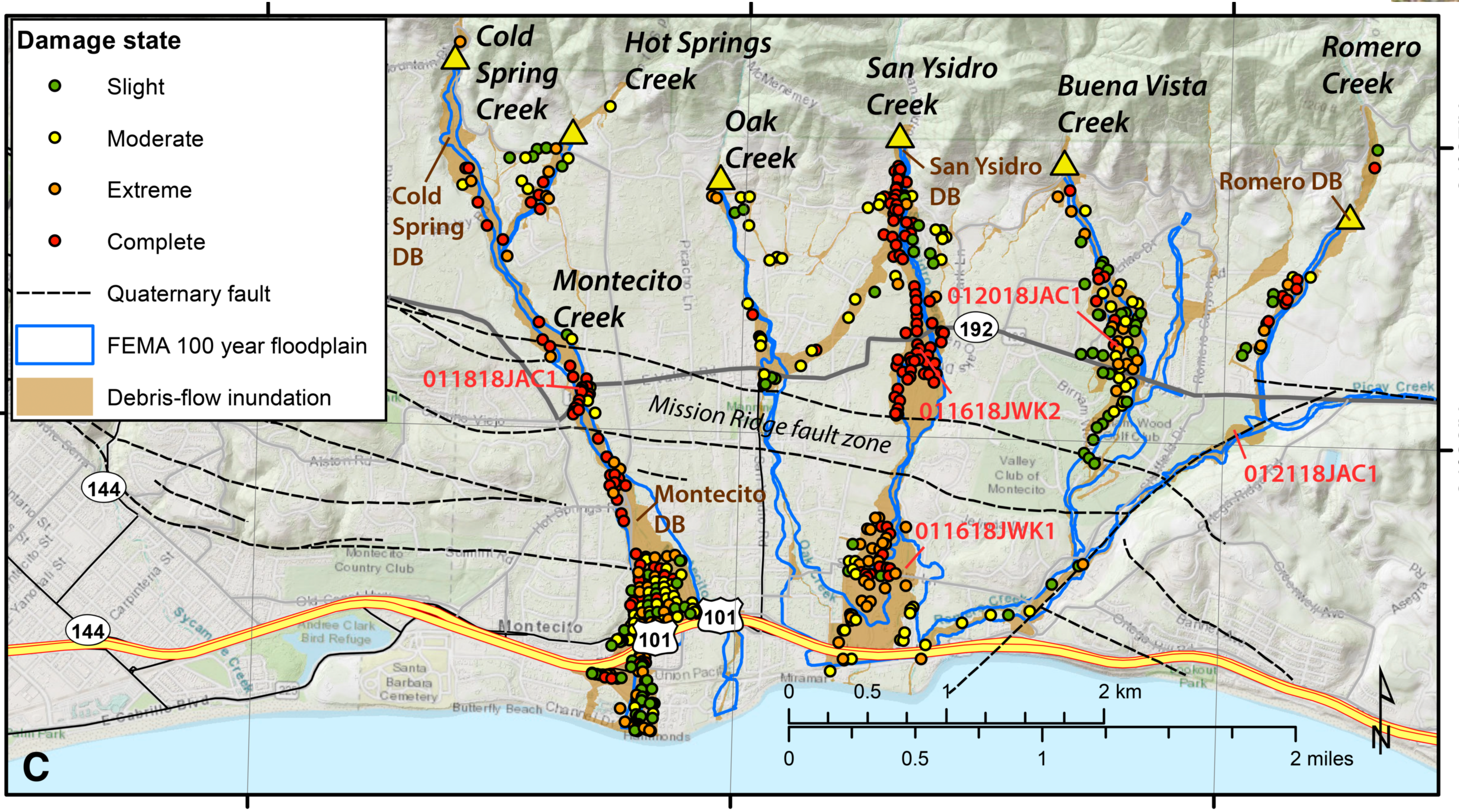
- Followed the 2017 Thomas fire during January, 2018 winter storm
- 23 fatalities
- Loss of 408 structures
- Evacuation order based on logistic regression debris flow likelihood model (Staley *et al*., 2016) likely reduced casualties




from Kean *et al.* (2019)

# 2018 Montecito Debris Flow Events

- Followed the 2017 Thomas fire during January, 2018 winter storm
- 23 fatalities
- Loss of 408 structures
- Evacuation order based on logistic regression debris flow likelihood model (Staley *et al*., 2016) likely reduced casualties



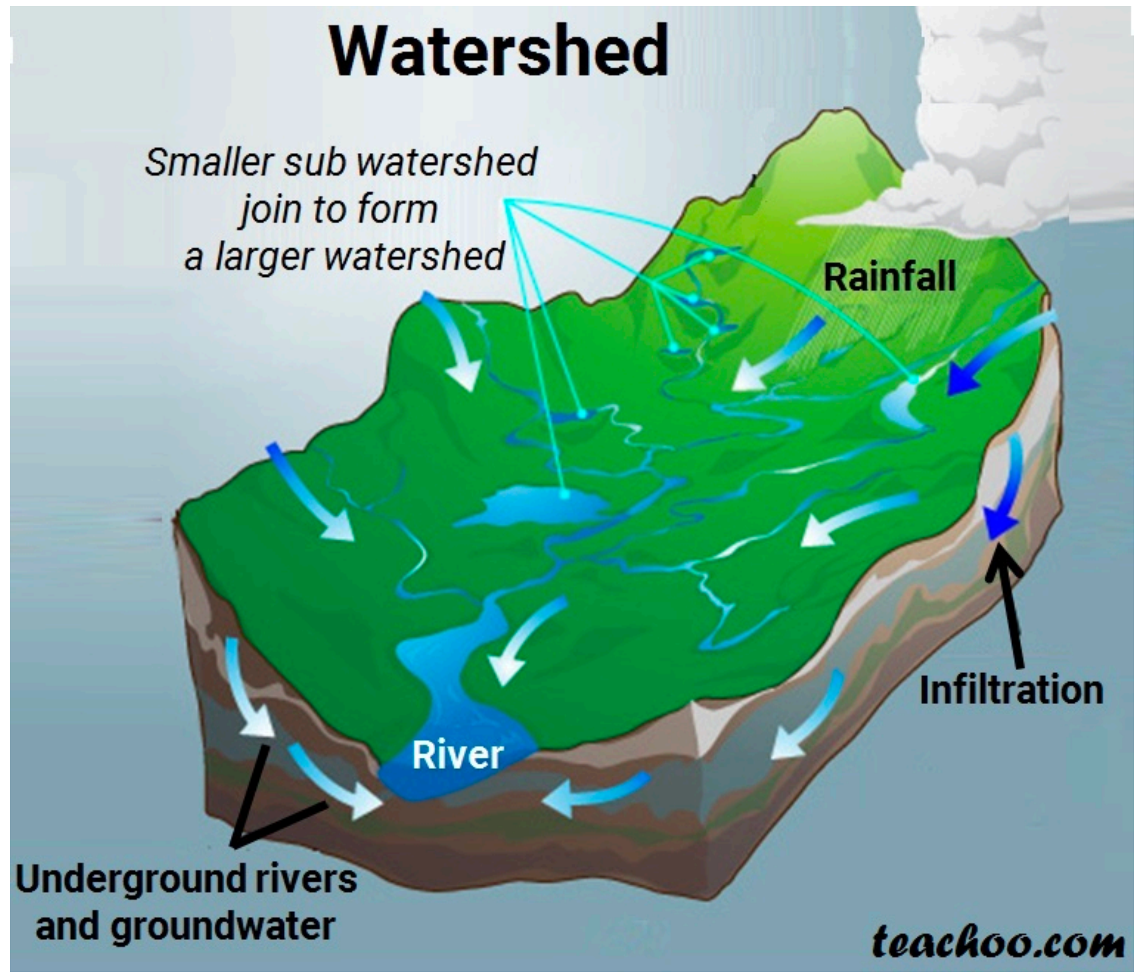

from Kean *et al.* (2019)

# Staley et al. (2016) Debris Flow Model and Data

## Debris-flow Likelihood Model

$$P_{\text{df}} = S(\beta + C_1 \cdot \mathbf{PropHM23} \cdot \mathbf{i15} + C_2 \cdot \mathbf{dNBR1000} \cdot \mathbf{i15} + C_3 \cdot \mathbf{KF} \cdot \mathbf{i15})$$

$$S(x) = \frac{1}{1 + e^x}$$

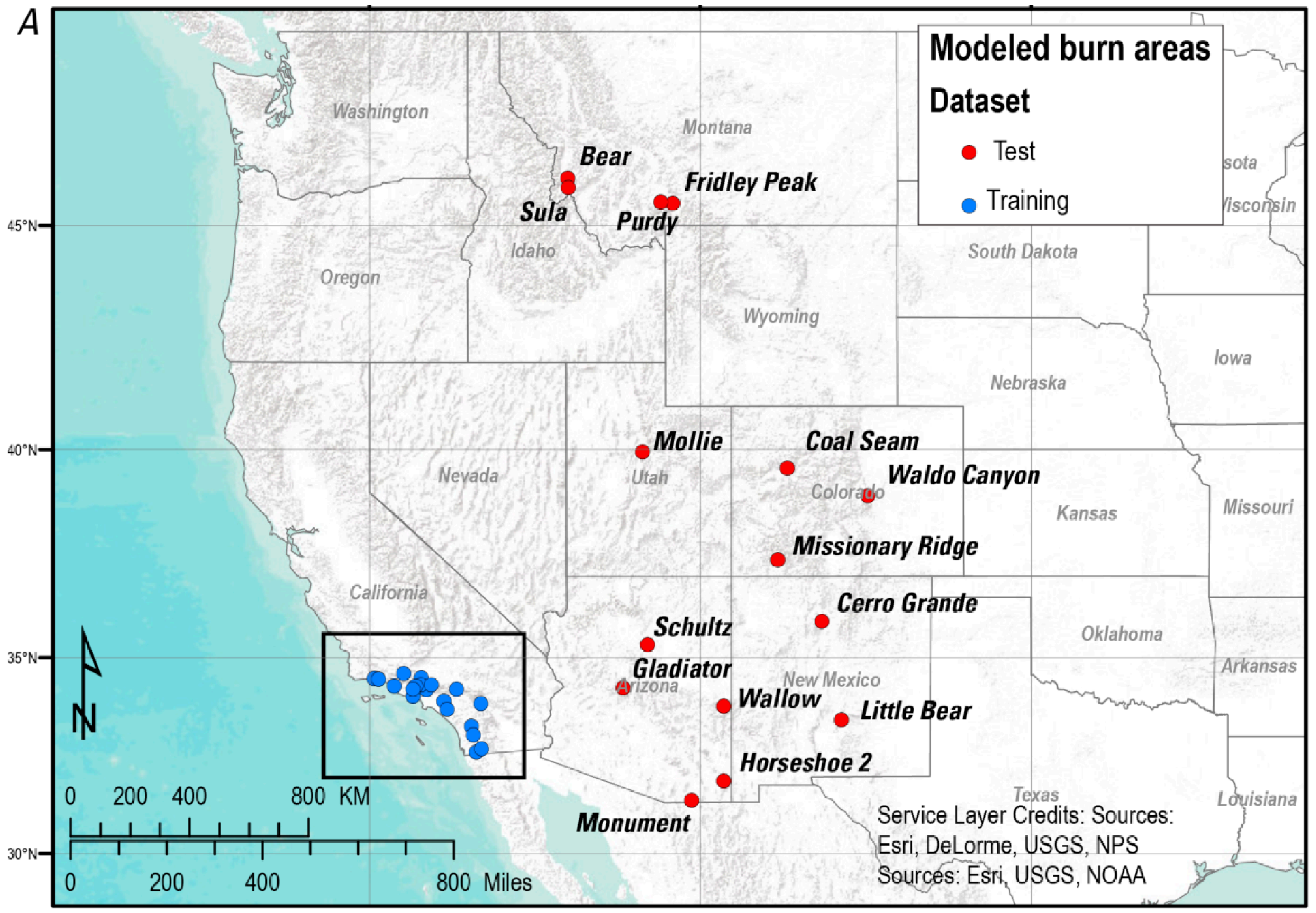



*Model features*

i15: peak 15-minute **rainfall intensity**
PropHM23: proportion of watershed w/ **slope** > 23 degrees and medium/high burn severity
dNBR1000: differential normalized **burn ratio** (/1000)
KF: **Soil** erodibility factor

## Debris-flow Data

- 1,216 complete debris flow records
- 716 debris flow sites (intermountain West of U.S. and southern California)
- Includes features not used in Staley (2016) model (e.g., storm duration and accumulation)

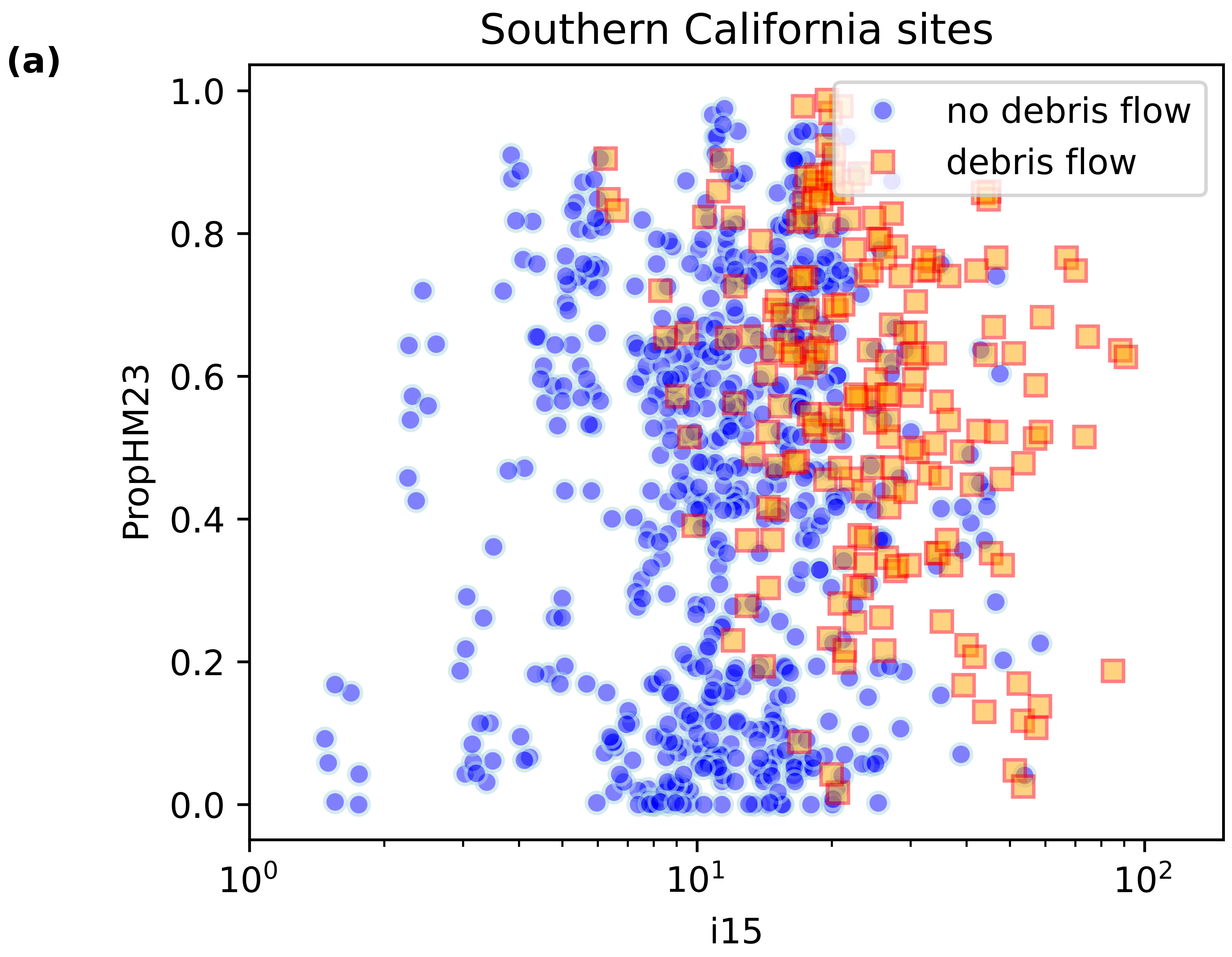


# Motivation

- Staley (2016) model overestimates debris flow likelihood for northern CA sites (Lindsay et al., 2020)
- Staley model was trained on southern California data only

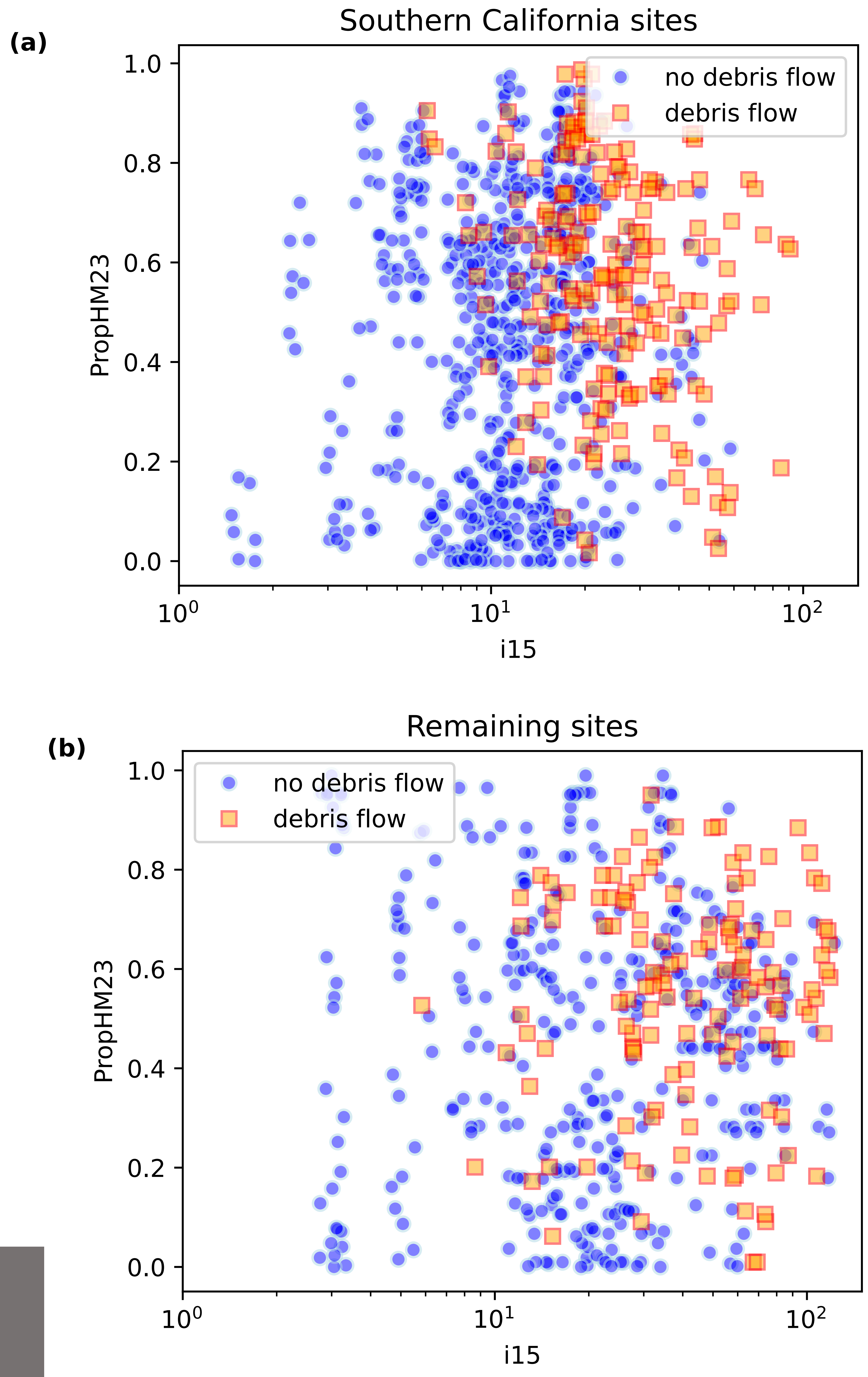
# Motivation

- Staley (2016) model overestimates debris flow likelihood for northern CA sites (Lindsay et al., 2020)
- Staley model was trained on southern California data only



# Motivation

- Staley (2016) model overestimates debris flow likelihood for northern CA sites (Lindsay et al., 2020)
- Staley model was trained on southern California data only

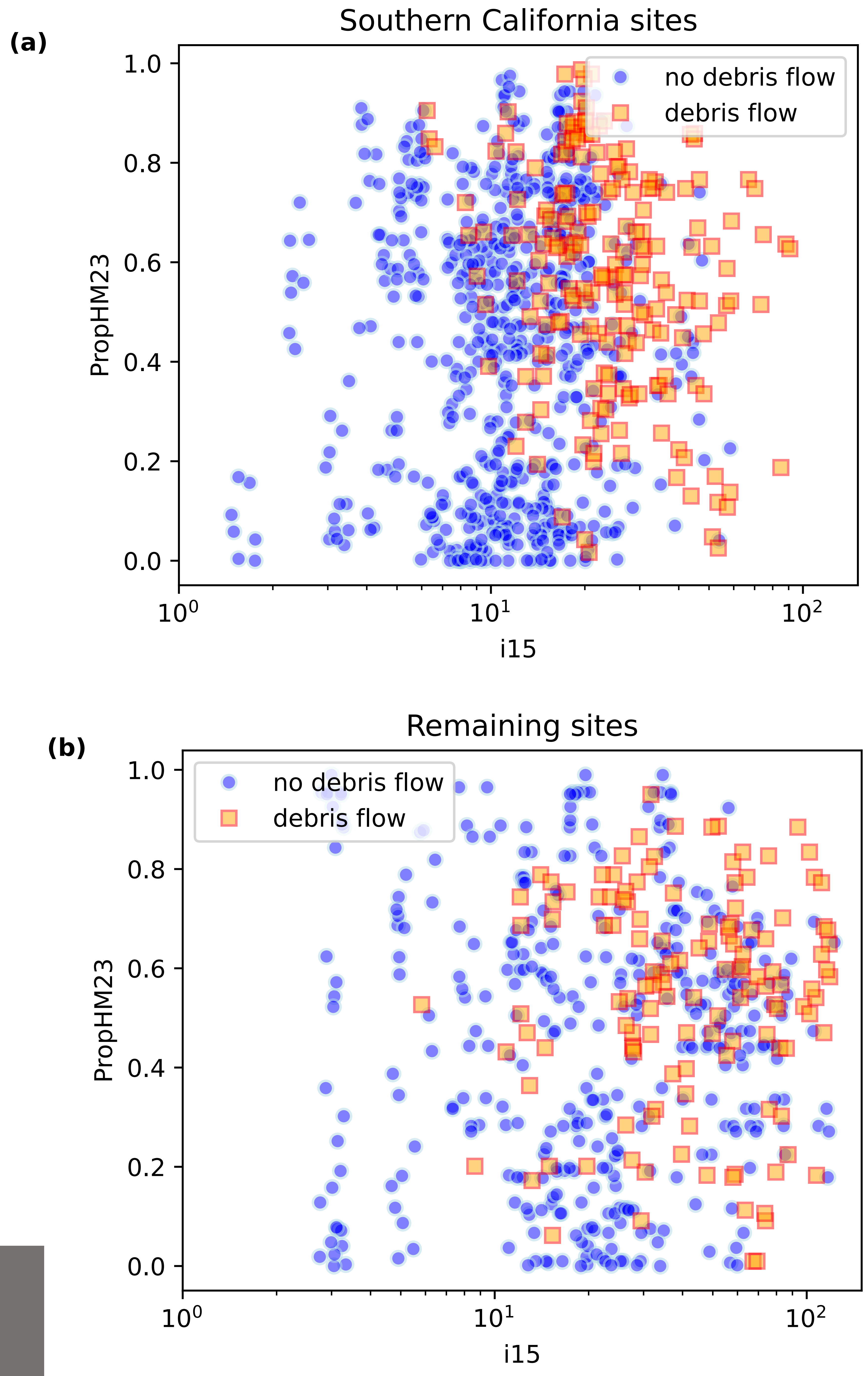


# Motivation

- Staley (2016) model overestimates debris flow likelihood for northern CA sites (Lindsay et al., 2020)
- Staley model was trained on southern California data only

# Scope of this Study

Improve debris-flow likelihood prediction by
- Using a random train-test split
- Introducing new features
- Training machine learning models beyond logistic regression (traditional and deep learning models)



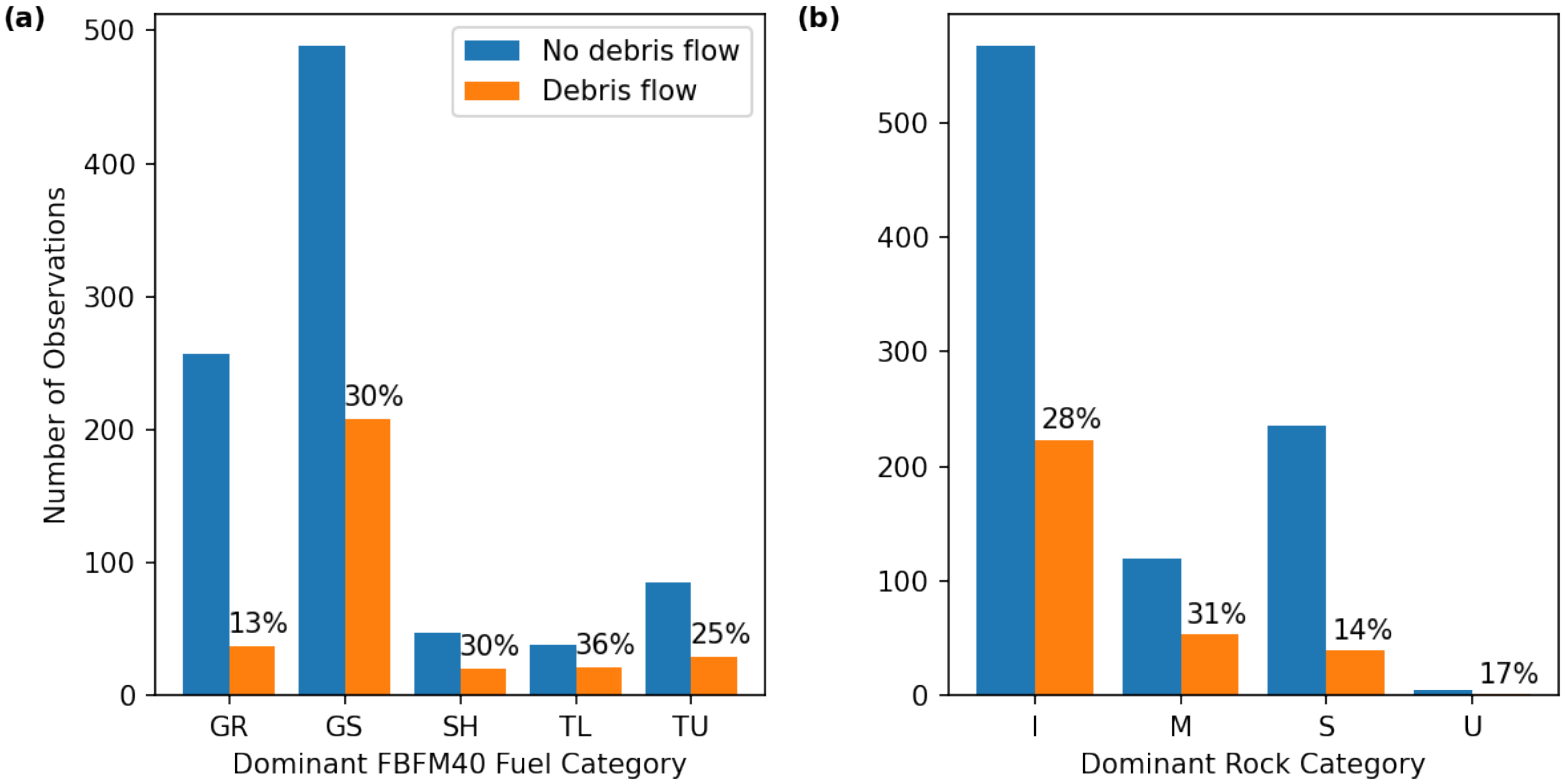
# Debris-flow Prediction as Big Data Problem

# Definition of Additional Features

Debris flow likelihood also depends on
- vegetation type
- burn severity
- type of underlying rock

Scott and Burgan Fire Behavior Fuel Model (FBFM40) parameters extracted from LANDFIRE dataset:
- Fraction of susceptible vegetation categories
- Fine fuel load

USGS geological map was queried for susceptible rock classes




LANDFIRE

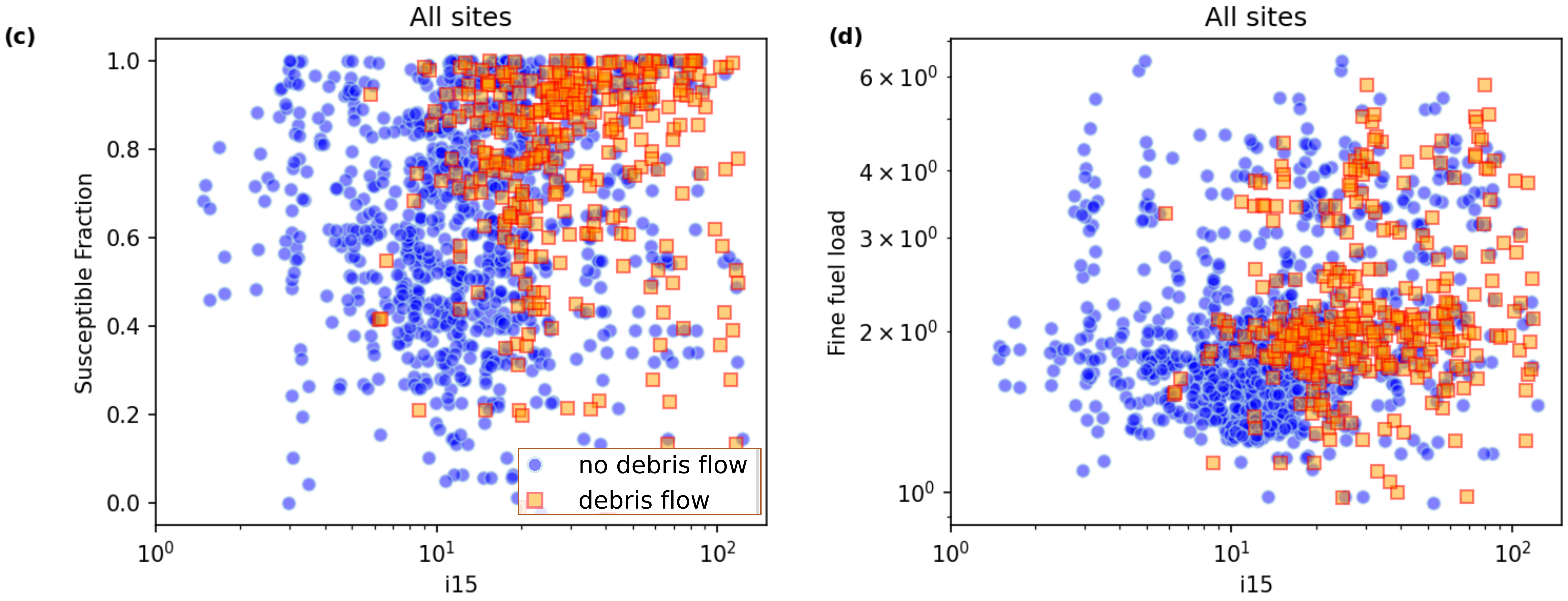
USGS science for a changing world

# Definition of Additional Features

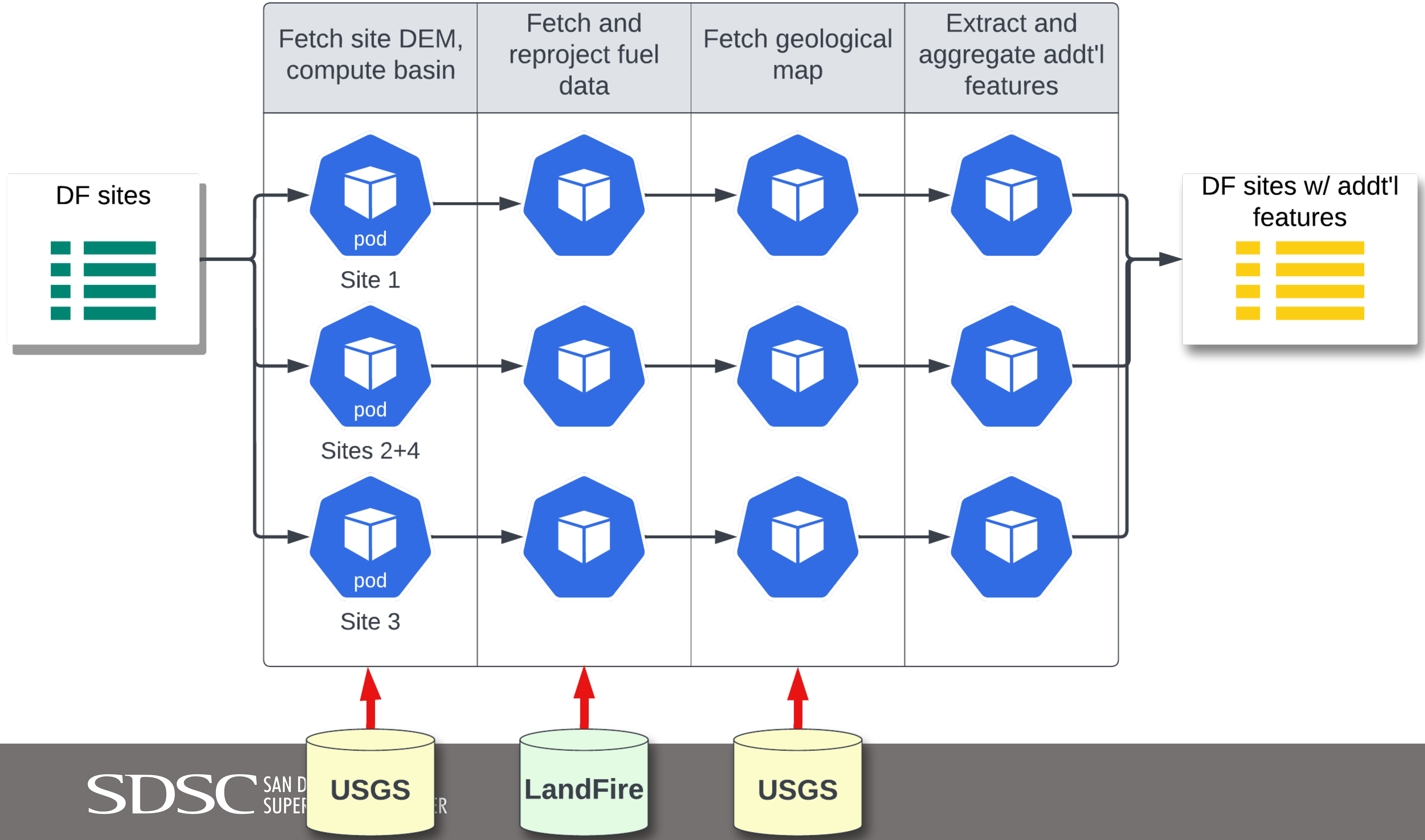
Debris flow likelihood also depends on
- vegetation type
- burn severity
- type of underlying rock

Scott and Burgan Fire Behavior Fuel Model (FBFM40) parameters extracted from LANDFIRE dataset:
- Fraction of susceptible vegetation categories
- Fine fuel load

USGS geological map was queried for susceptible rock classes



# Big Data Processing Pipeline




GeoPandas

xarray

dask

kubernetes

HELM

Nautilus

PRP PACIFIC RESEARCH PLATFORM

UC San Diego

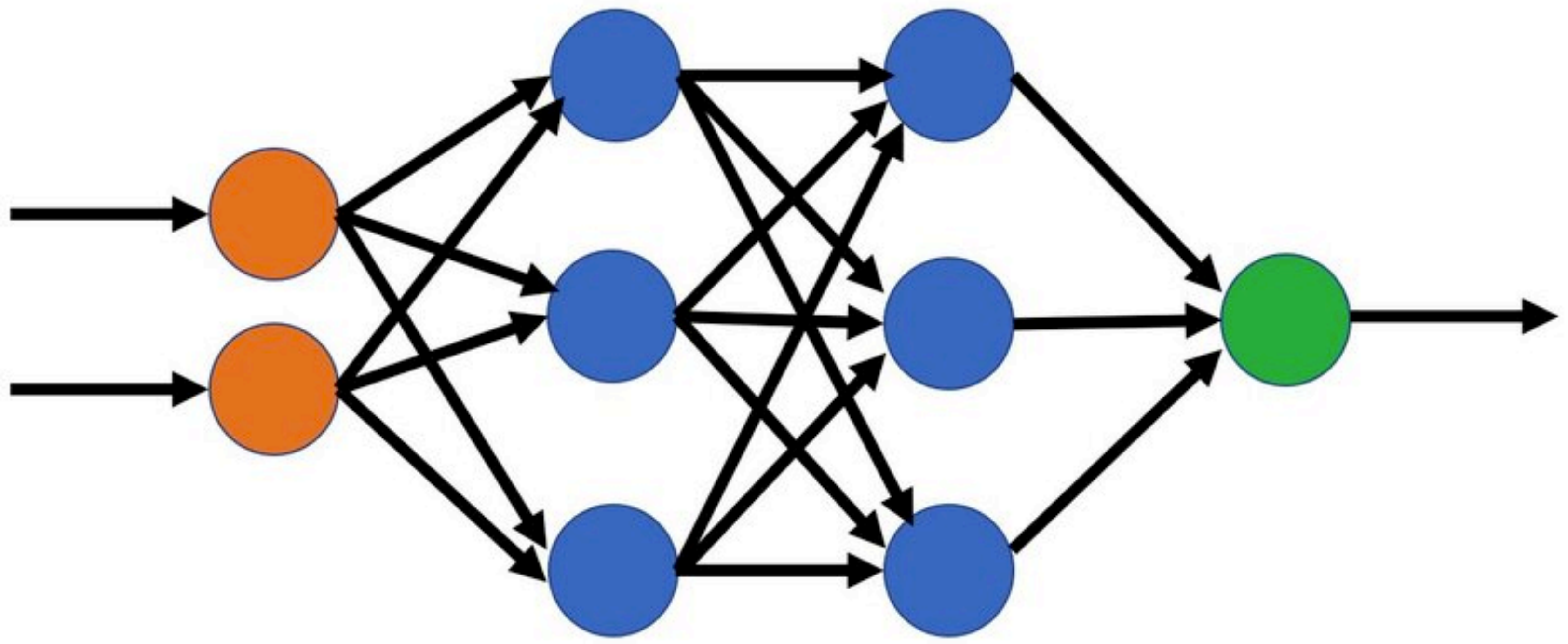
# ML Models Training and Testing

Traditional machine learning models (implemented using scikit-learn):

- Logistic regression
- Naive Bayes
- Support vector classifier (RBF kernel)
- K-nearest neighbors (k=5)
- Decision trees
- Random forest

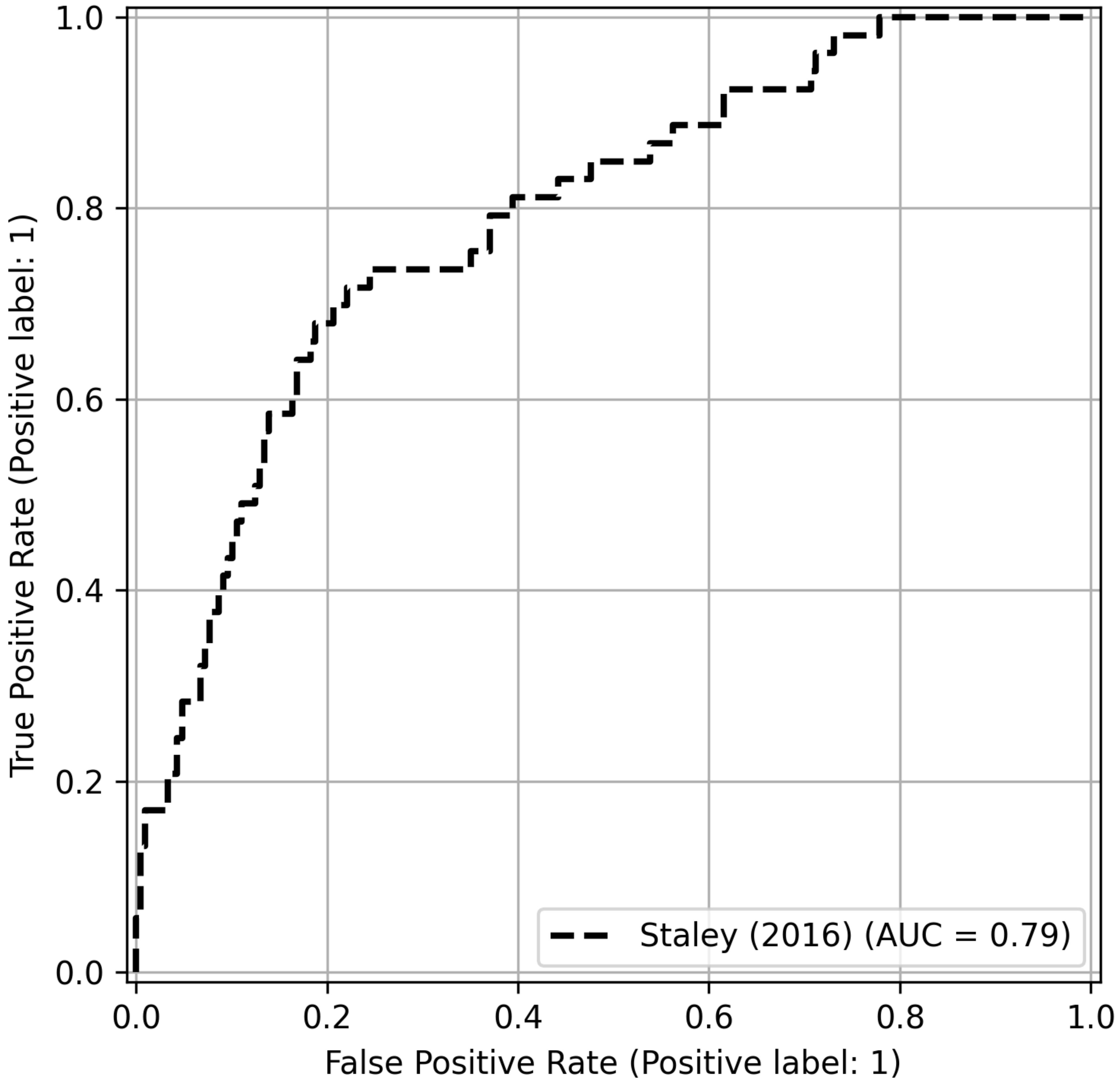
Fully connected, deep neural network (implemented using Keras w/ Tensorflow backend):

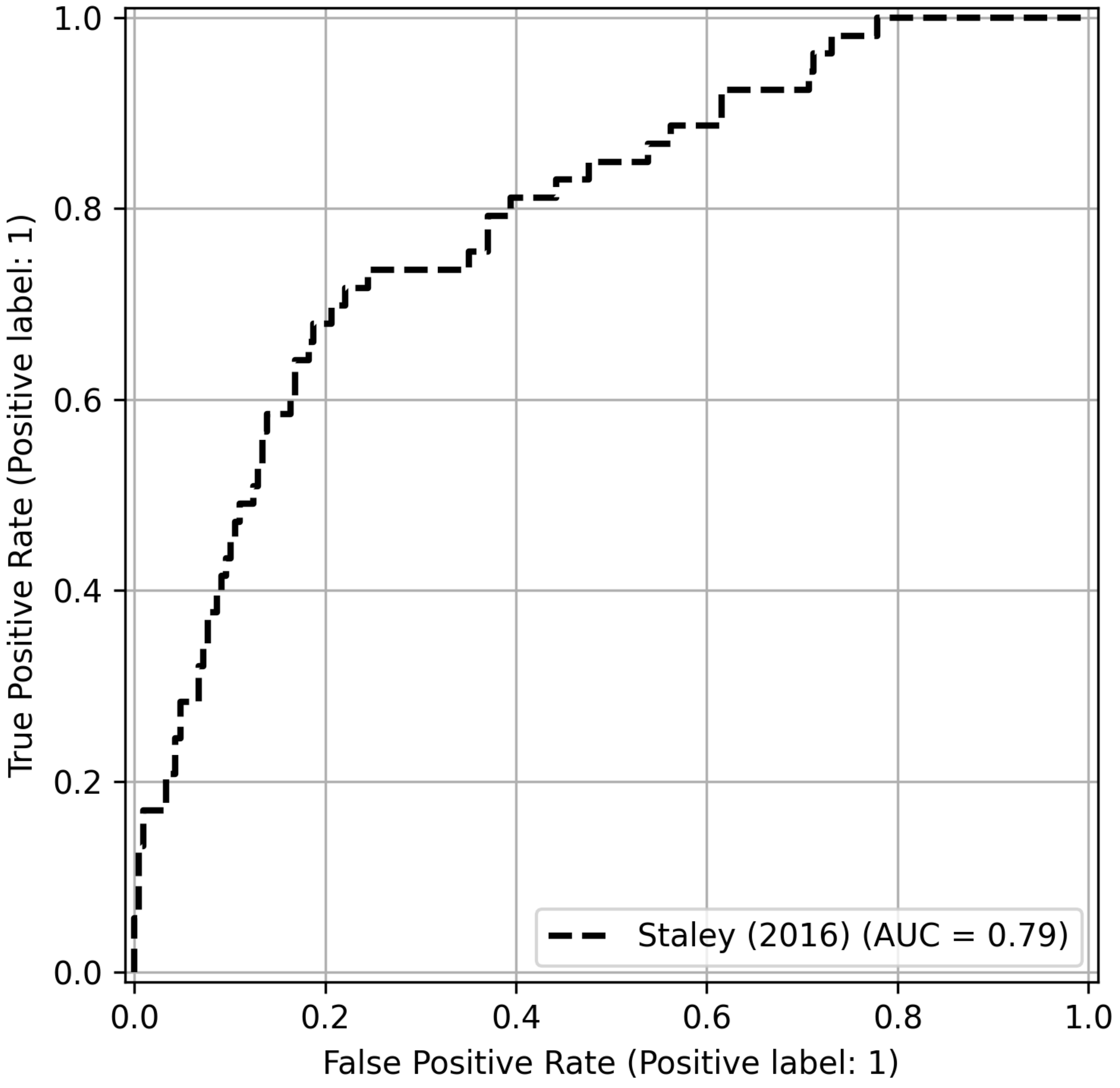
- 7 hidden layers for base features
- 11 hidden layers using additional features
- Drop-out rates between 0.2 and 0.3
- Weight initialization using Glorot (Xavier) method
- ReLU, tanh and sigmoid activation functions



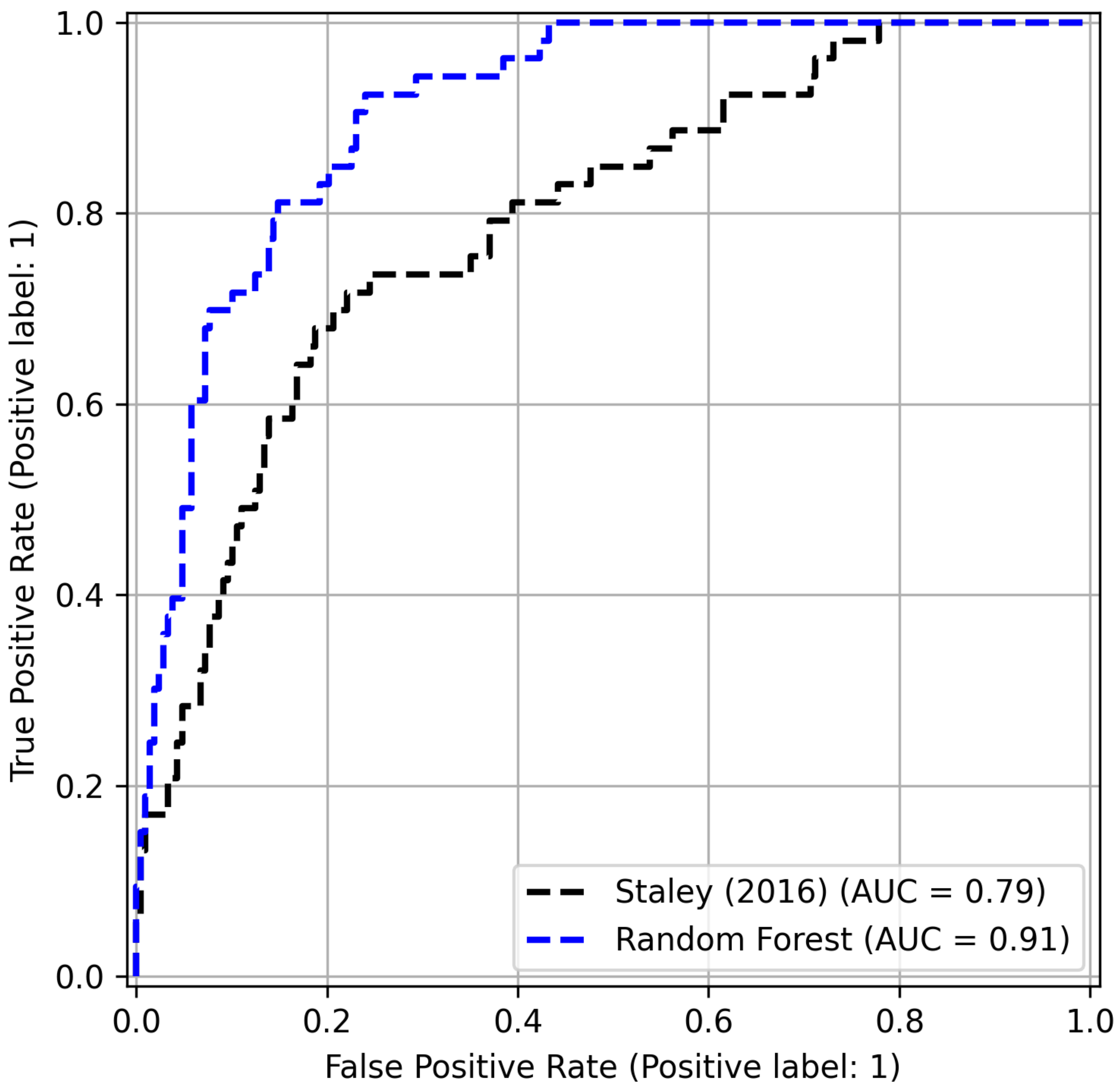
80-20% random train-test split (southern CA example):

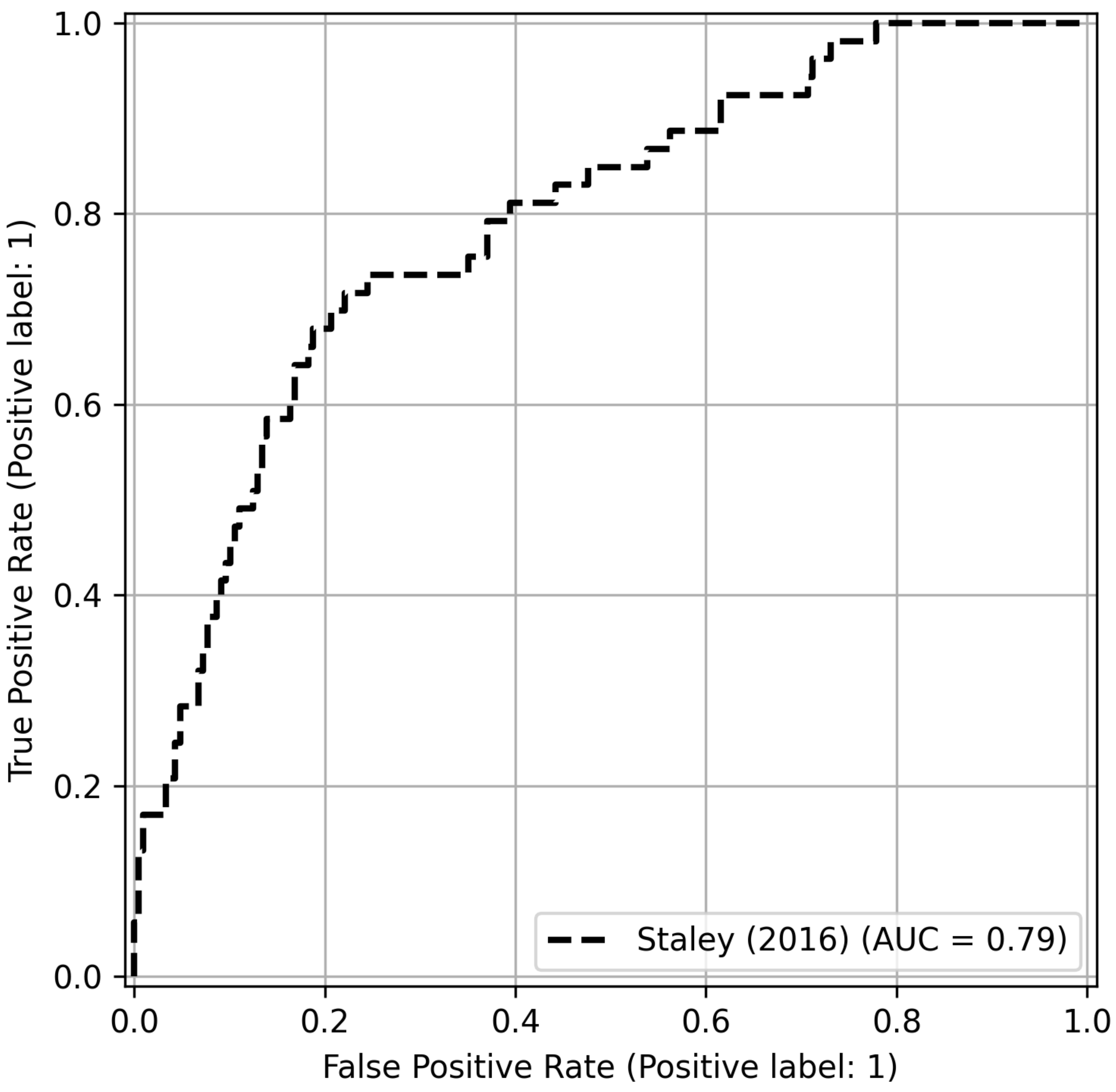
# Machine Learning Models and Performance



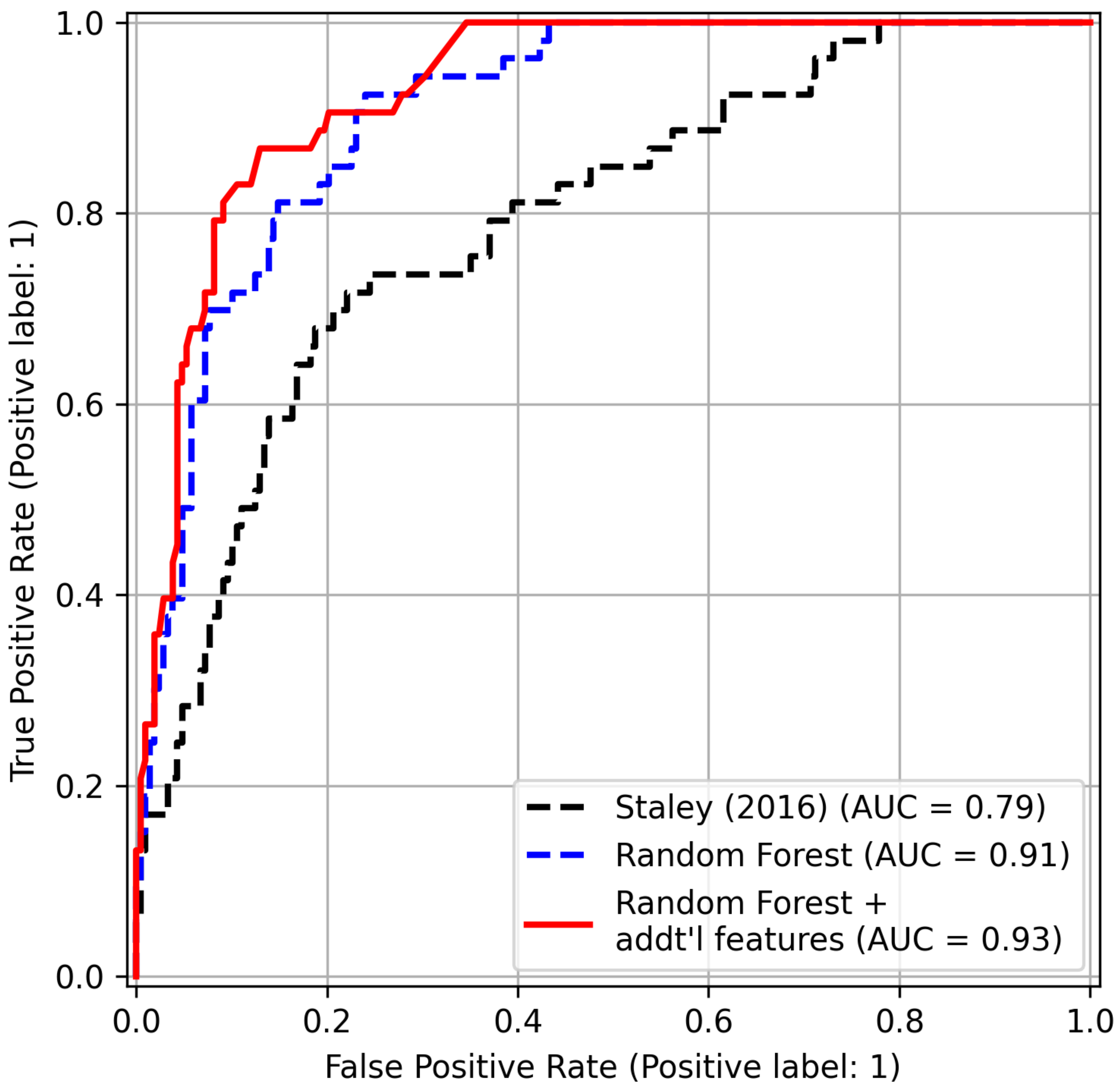



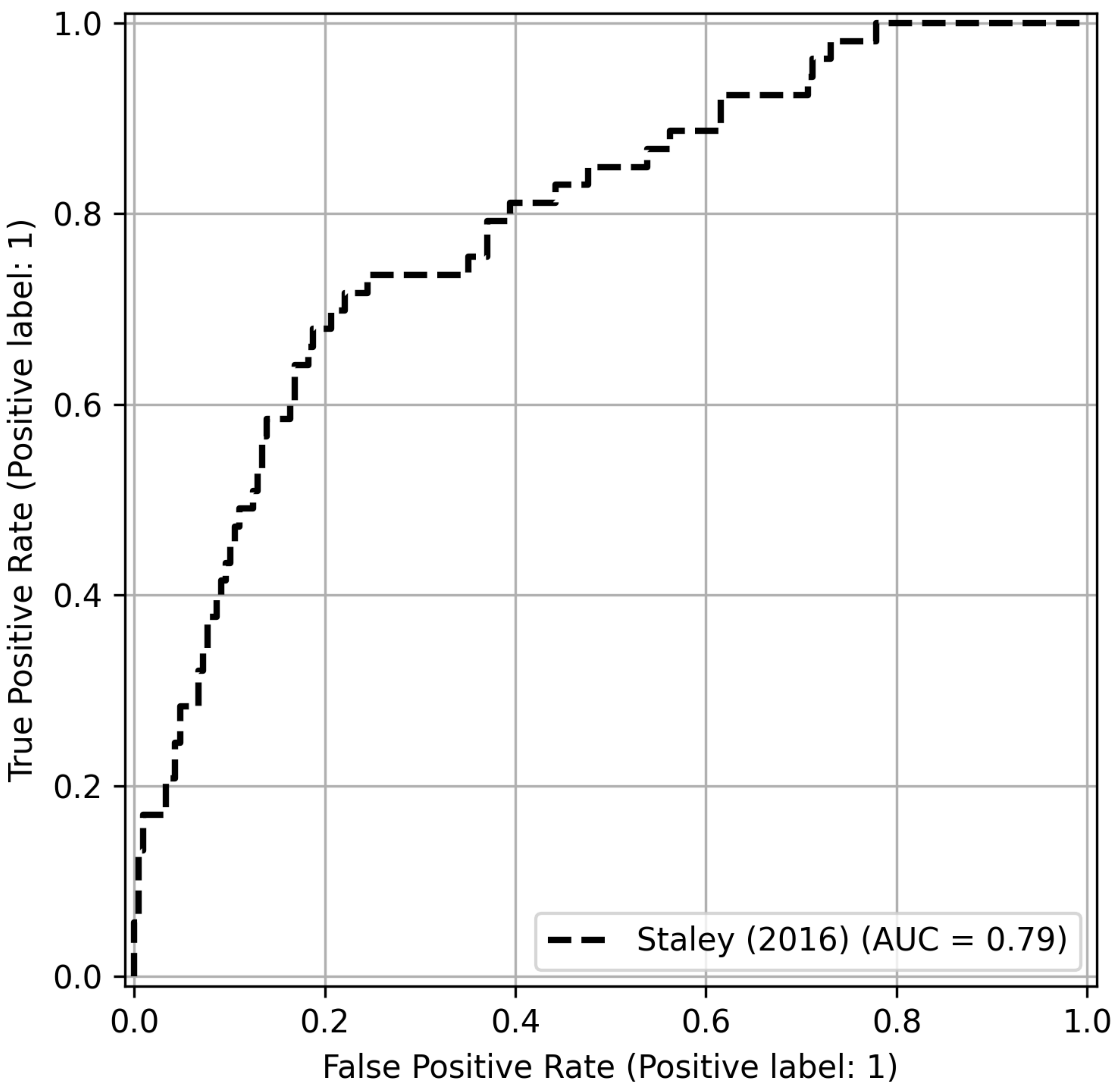
# Machine Learning Models and Performance



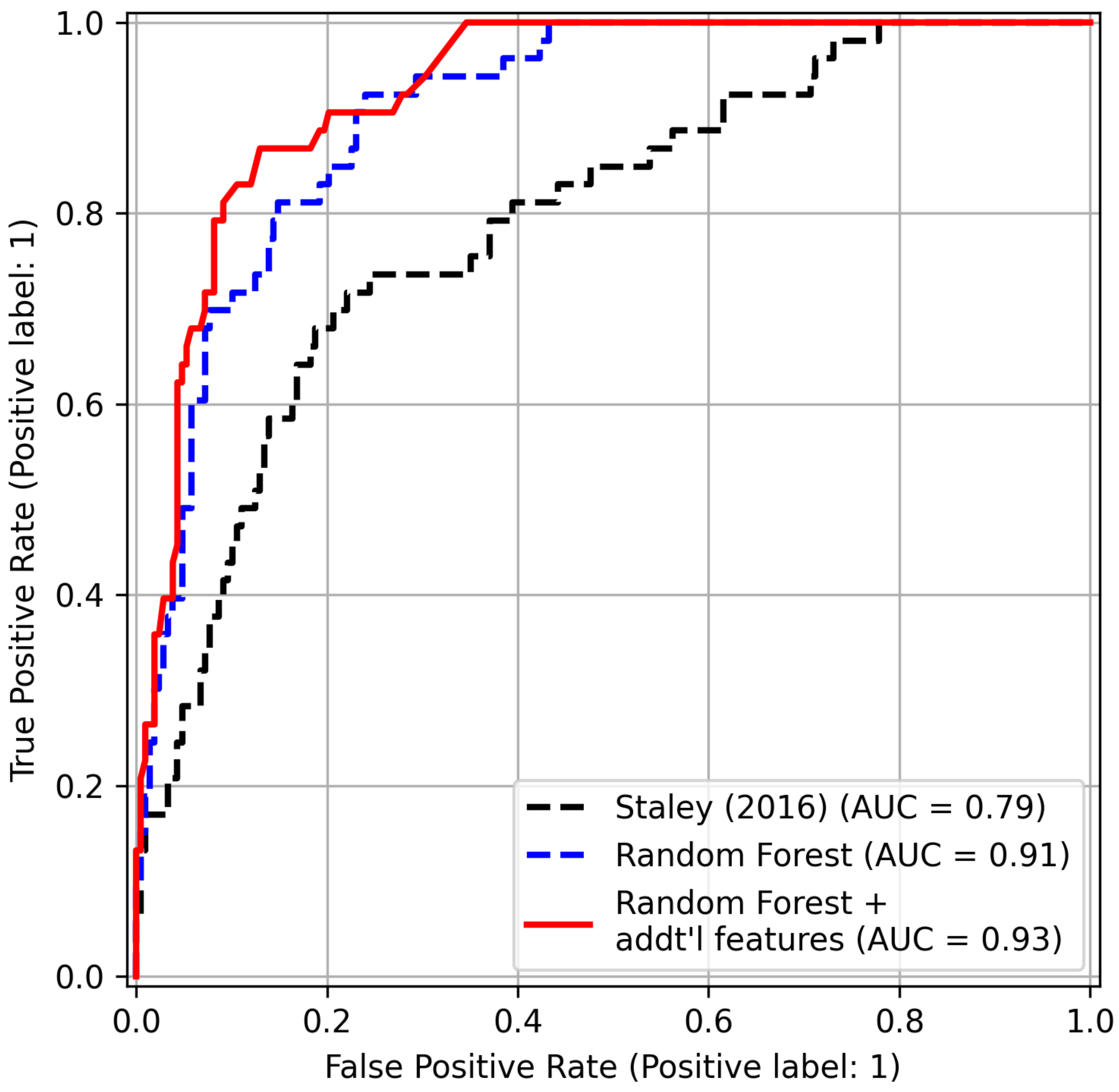



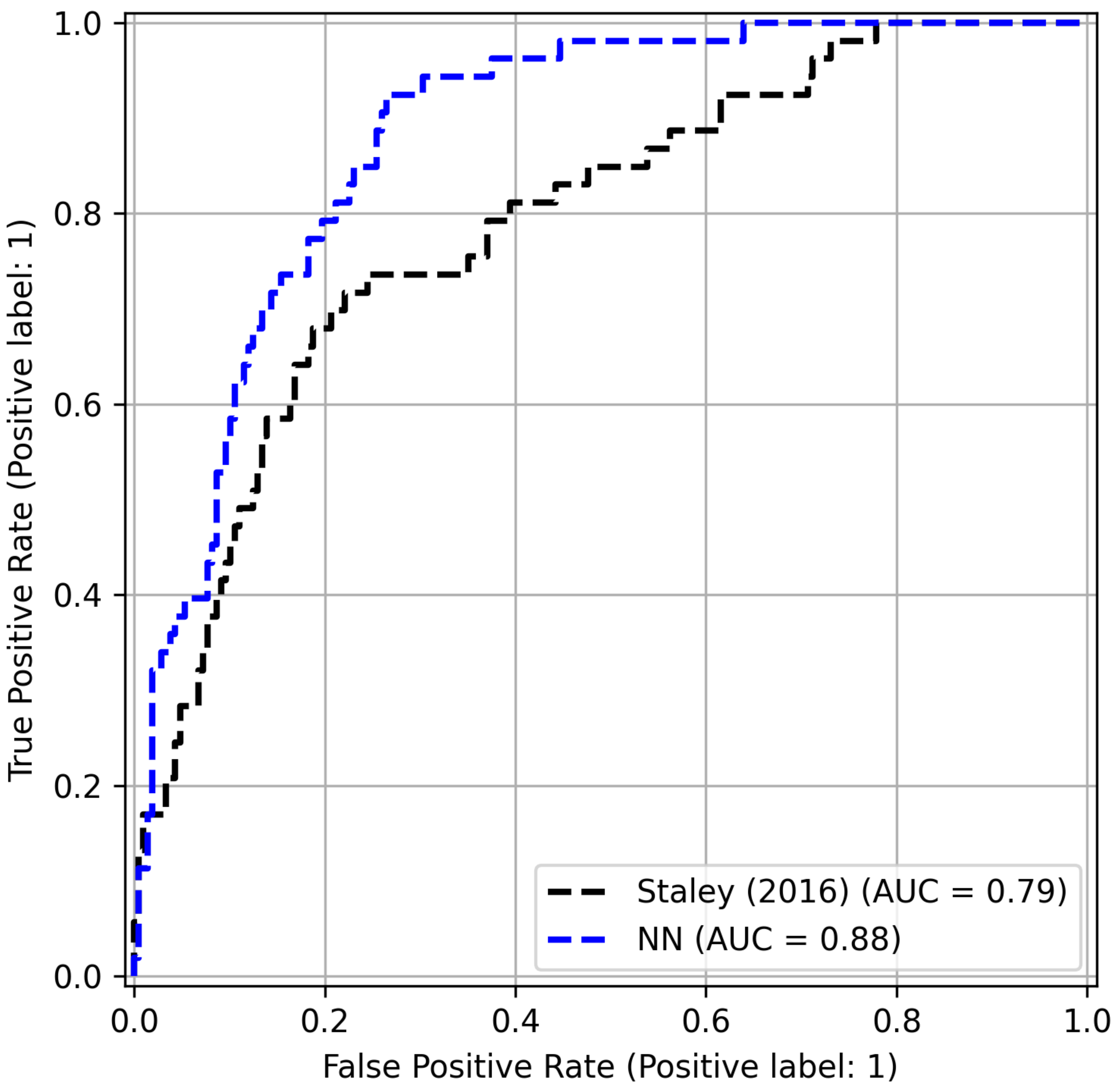
# Machine Learning Models and Performance



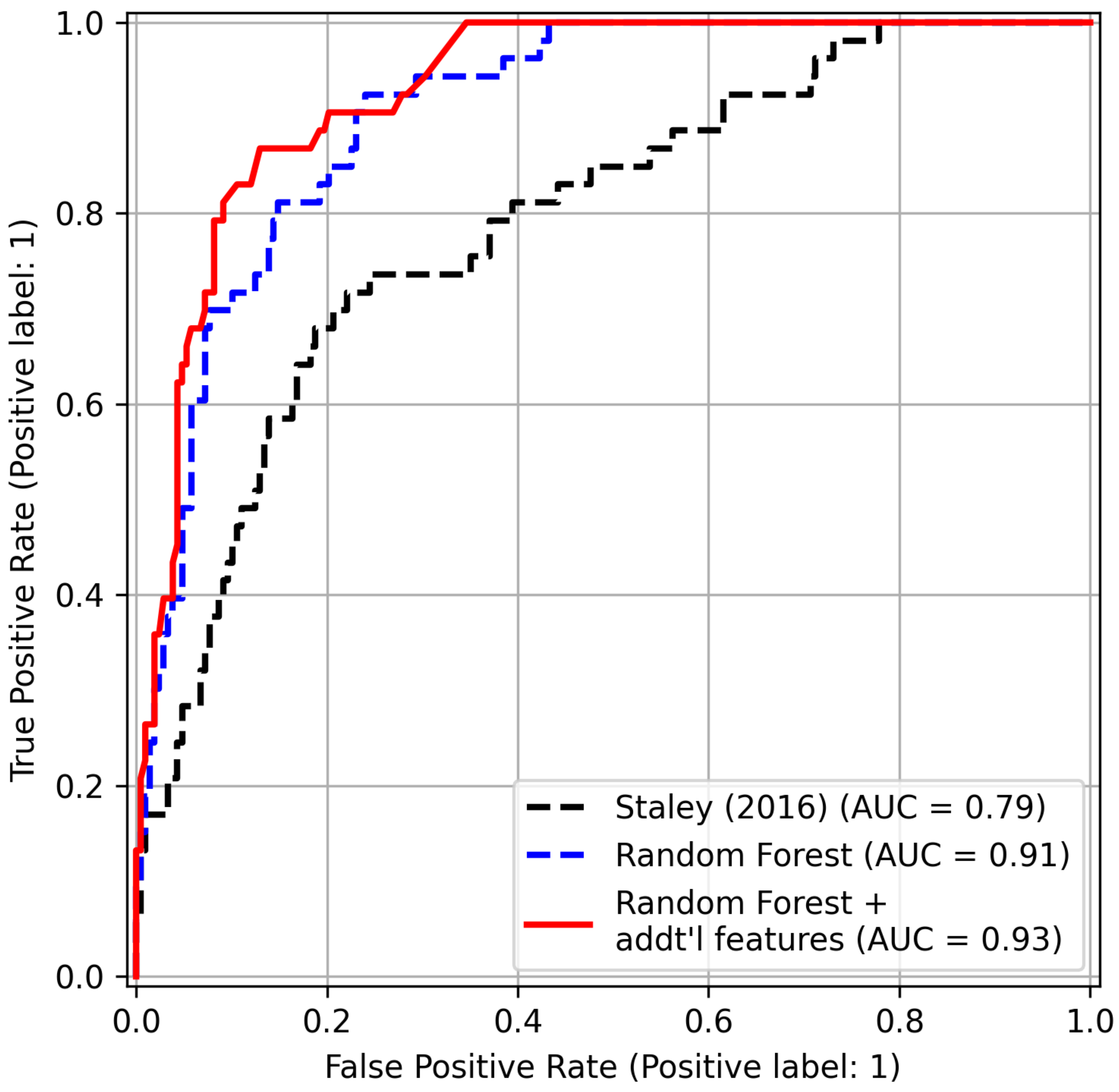



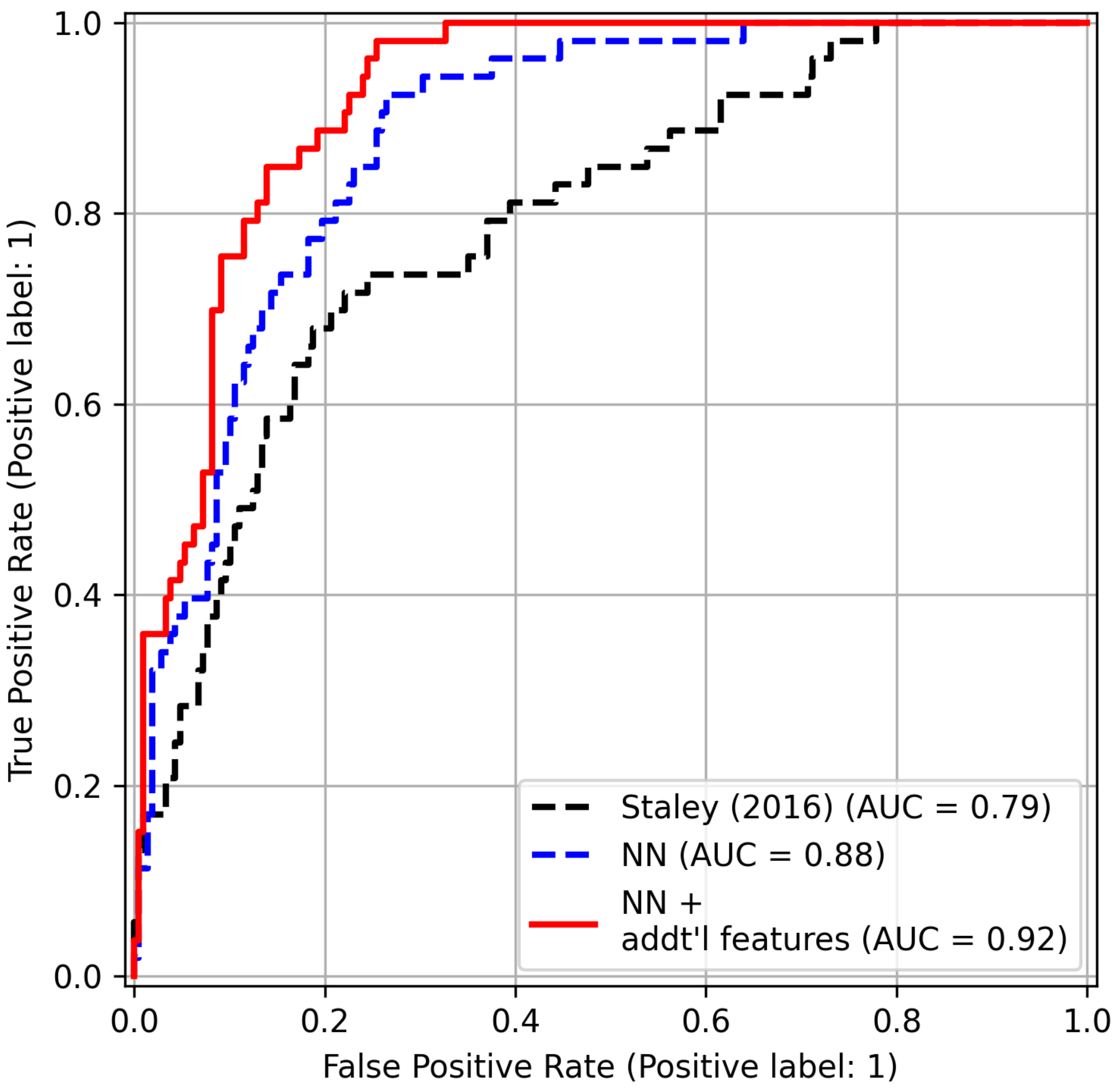
# Machine Learning Models and Performance



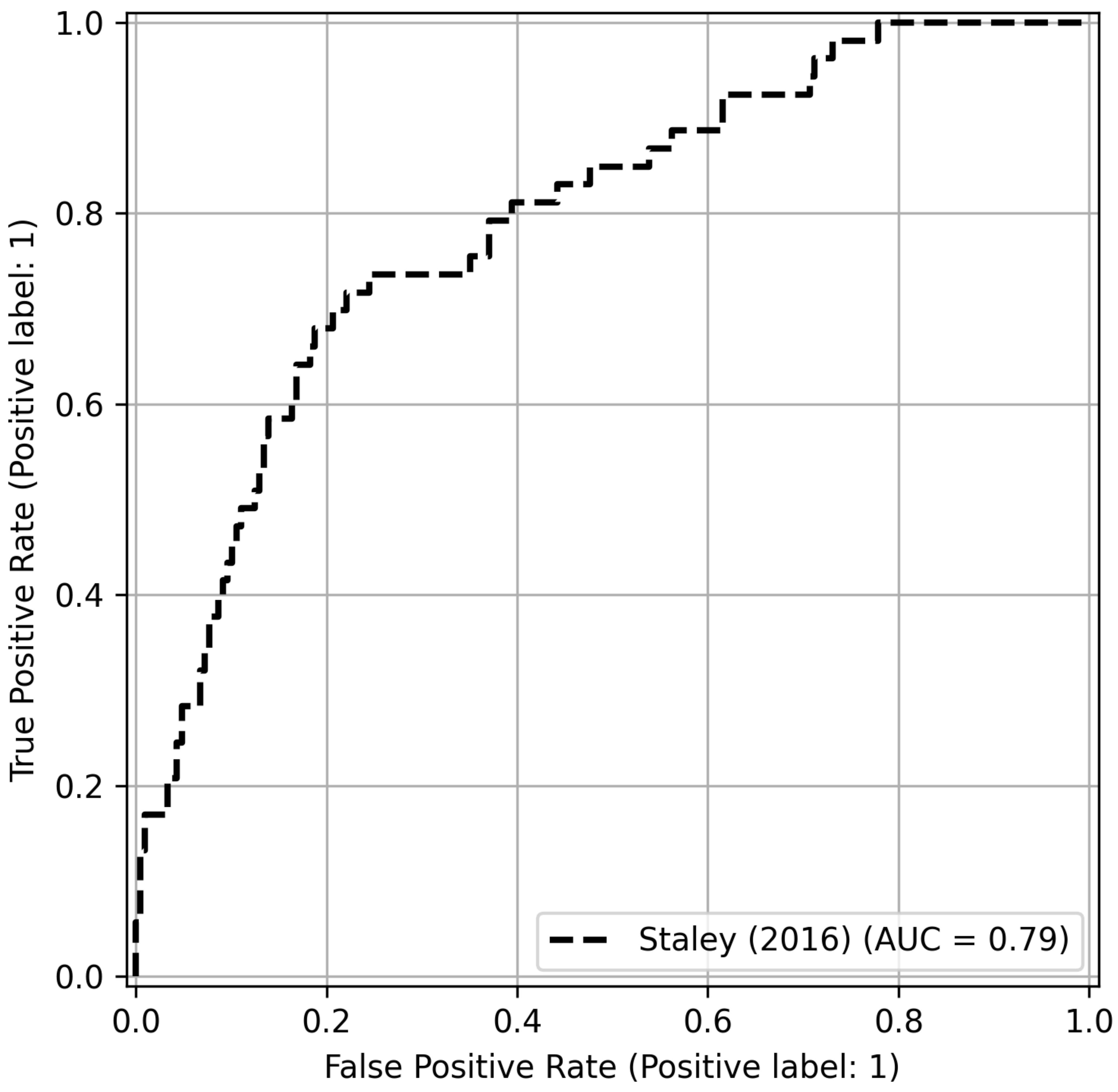



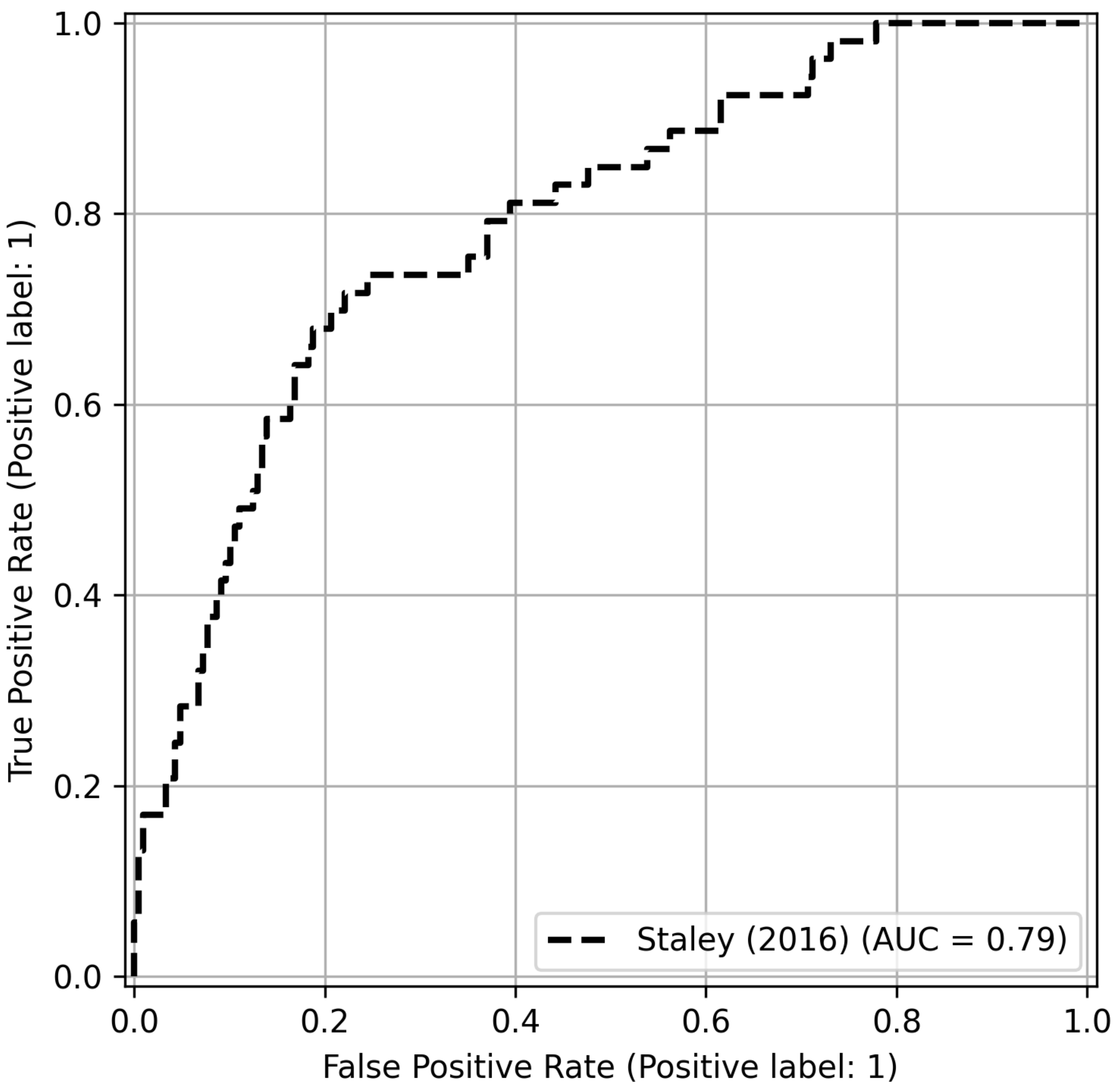
# Machine Learning Models and Performance



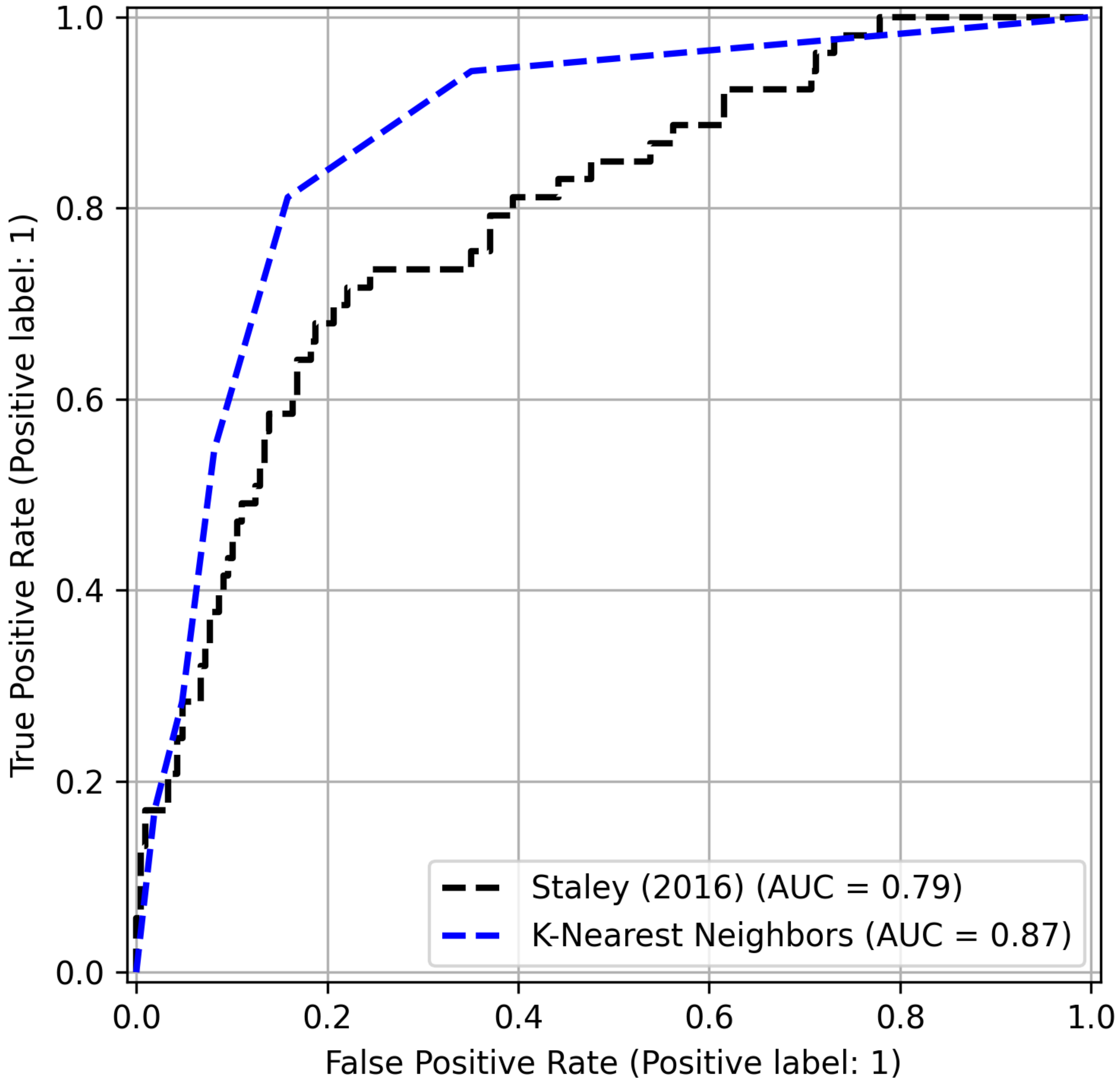



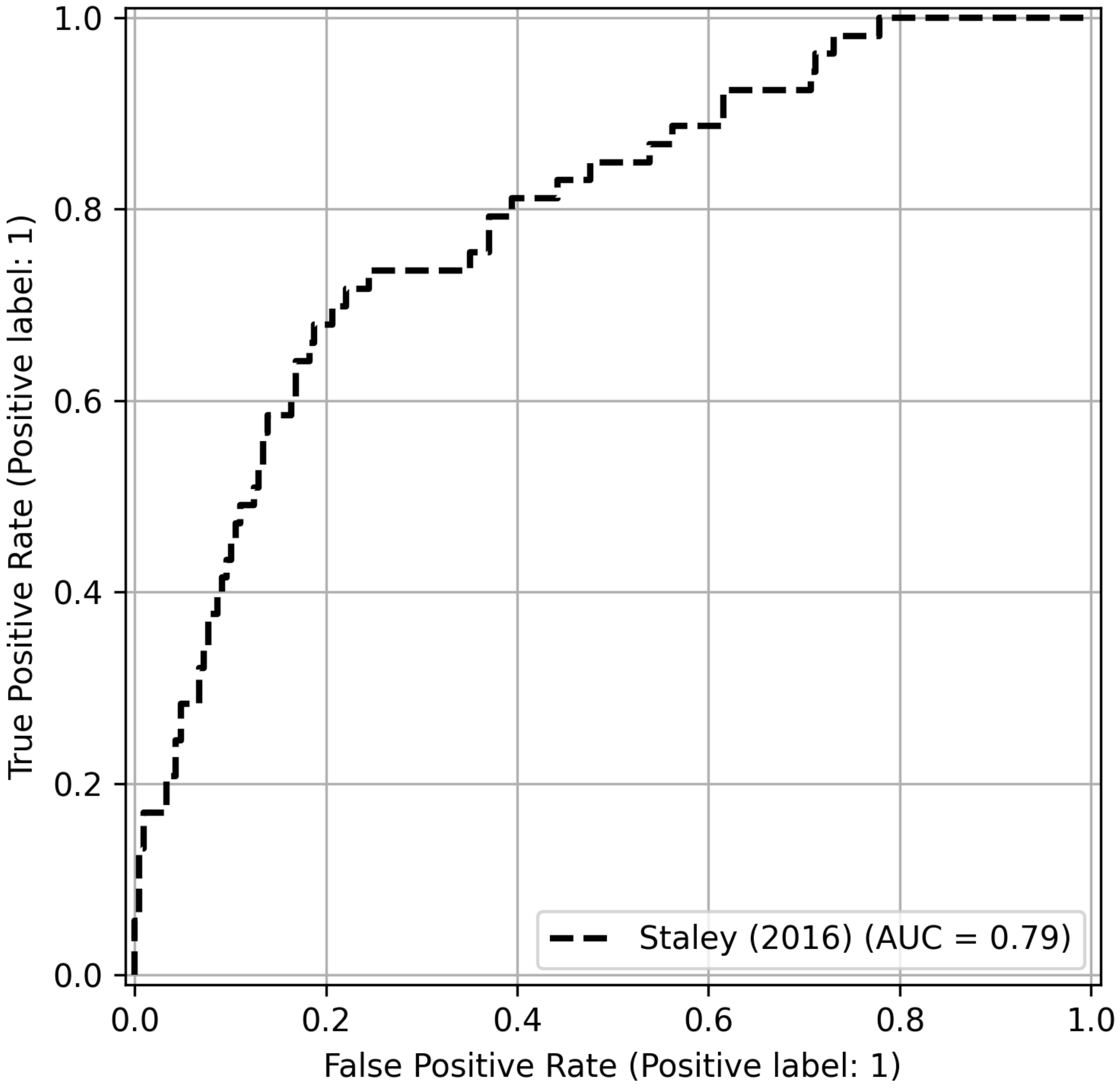
# Machine Learning Models and Performance



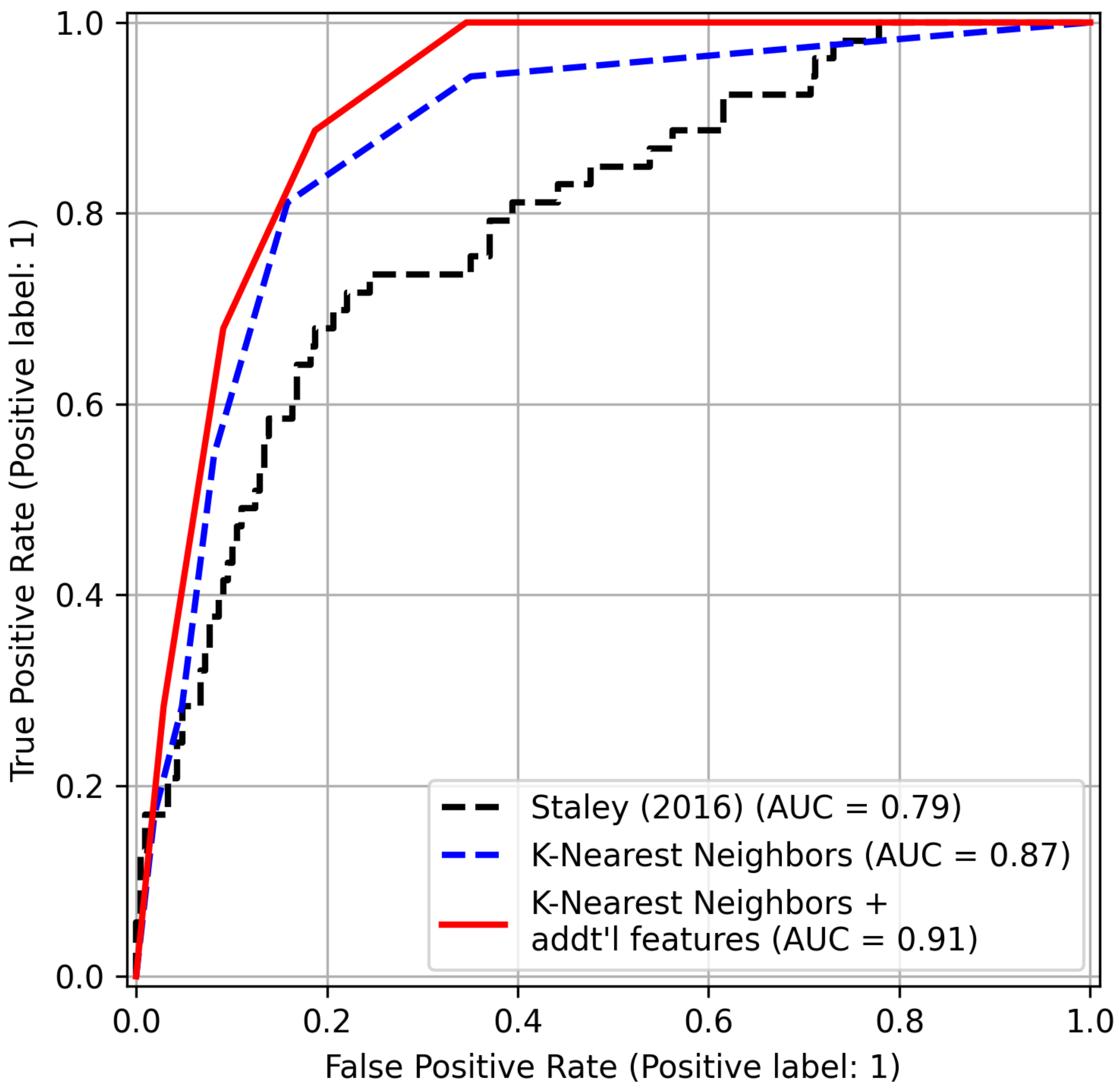



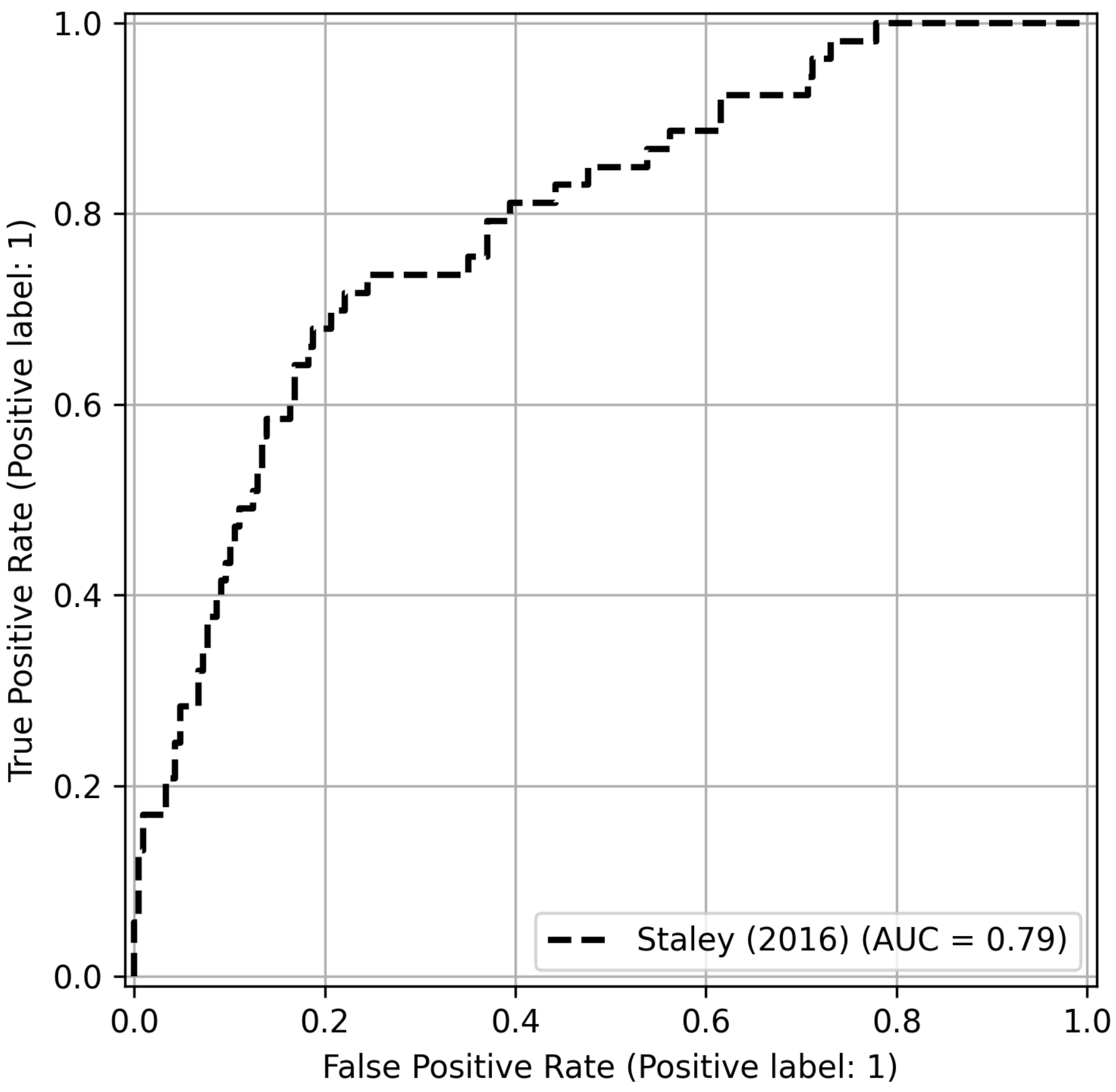
# Machine Learning Models and Performance



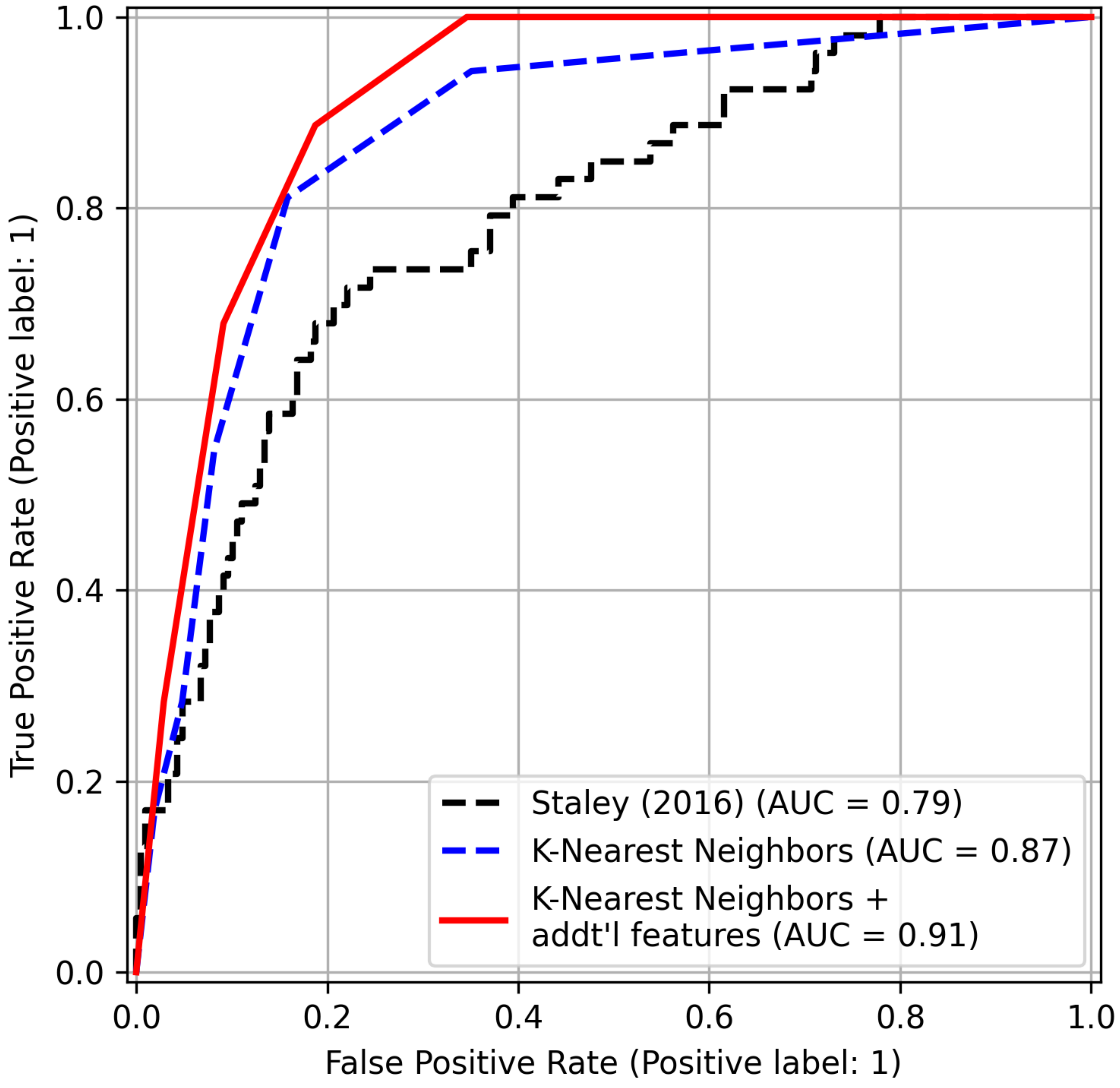



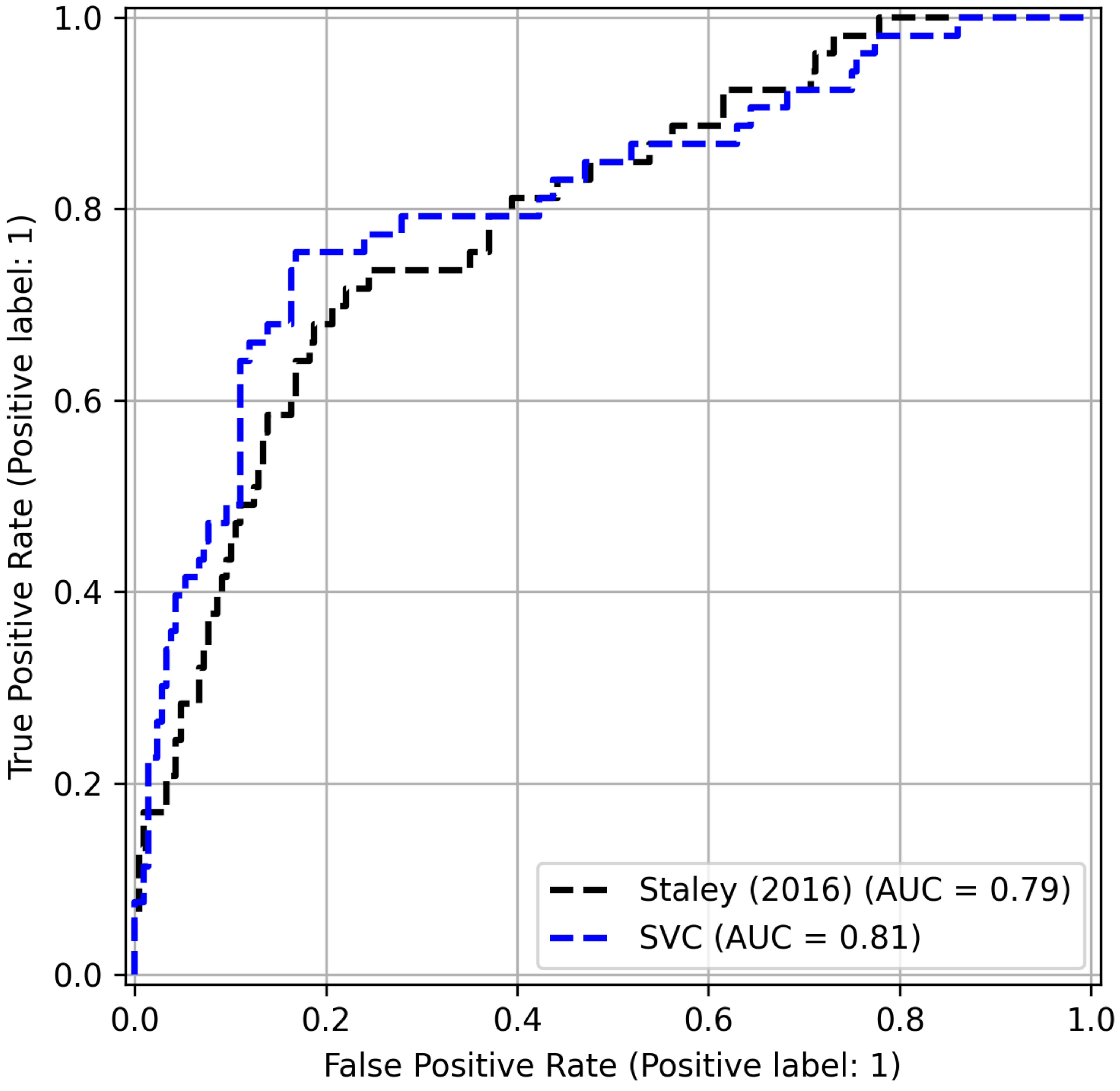
# Machine Learning Models and Performance



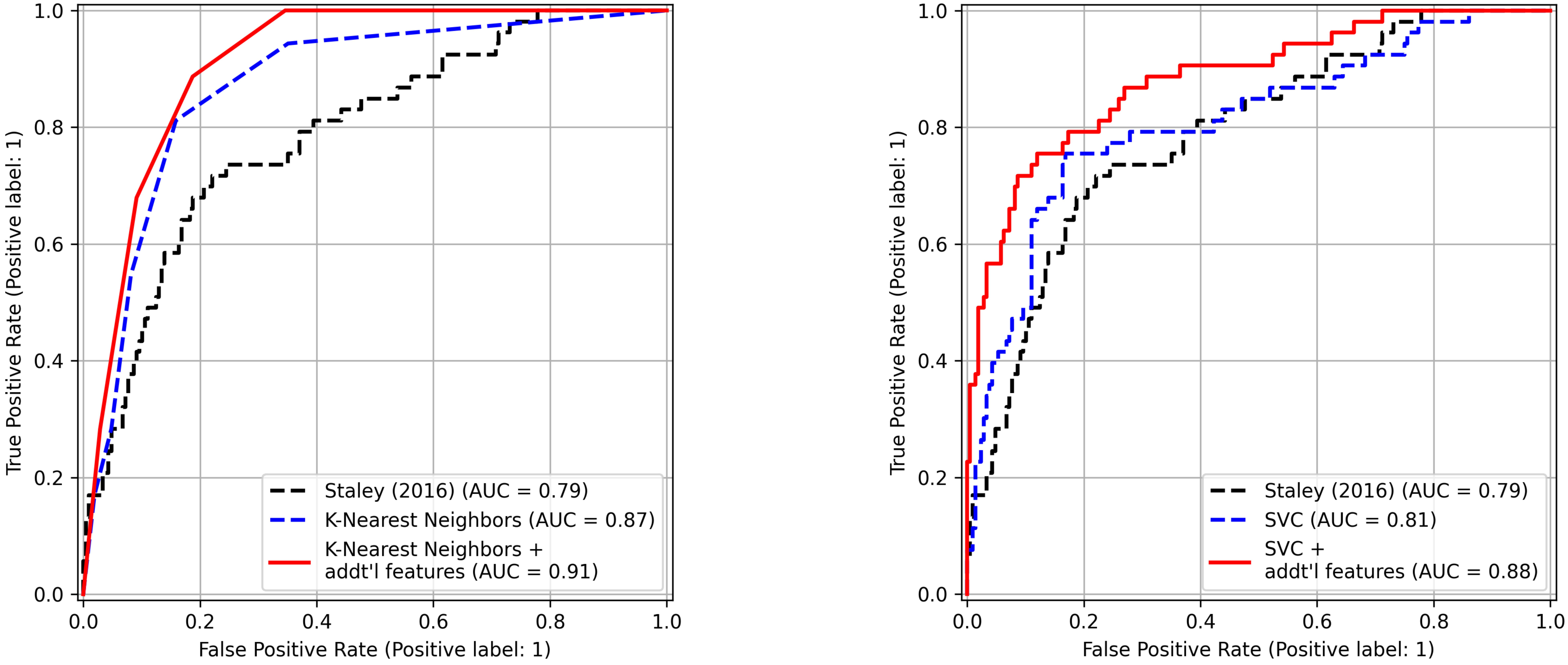



# Machine Learning Models and Performance



# Machine Learning Models and Performance



# Machine Learning Models and Performance

Performance metric: Jaccard index (Area under the curve, AUC)

| Method | base features | | w/ addt'l features | |
|---|---|---|---|---|
| | training | testing | training | testing |
| Staley16 | 0.39 (0.80) | 0.37 (0.79) | | |
| Logistic regr. | 0.35 (0.83) | 0.32 (0.86) | 0.35 (0.84) | 0.36 (0.87) |
| Naive Bayes | 0.24 (0.79) | 0.27 (0.82) | 0.27 (0.79) | 0.31 (0.84) |
| SVC | 0.70 (0.97) | 0.36 (0.81) | 0.75 (0.98) | 0.42 (0.88) |
| KNN | 0.59 (0.94) | 0.41 (0.87) | 0.61 (0.95) | 0.50 (0.91) |
| Random Forest | 0.97 (1.00) | 0.45 (0.91) | 1.00 (1.00) | **0.53 (0.93)** |
| Neural Network | 0.59 (0.93) | 0.37 (0.88) | 0.75 (0.98) | **0.54 (0.92)** |

$$\text{Jaccard index} = \frac{TP}{TP + FP + FN}$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

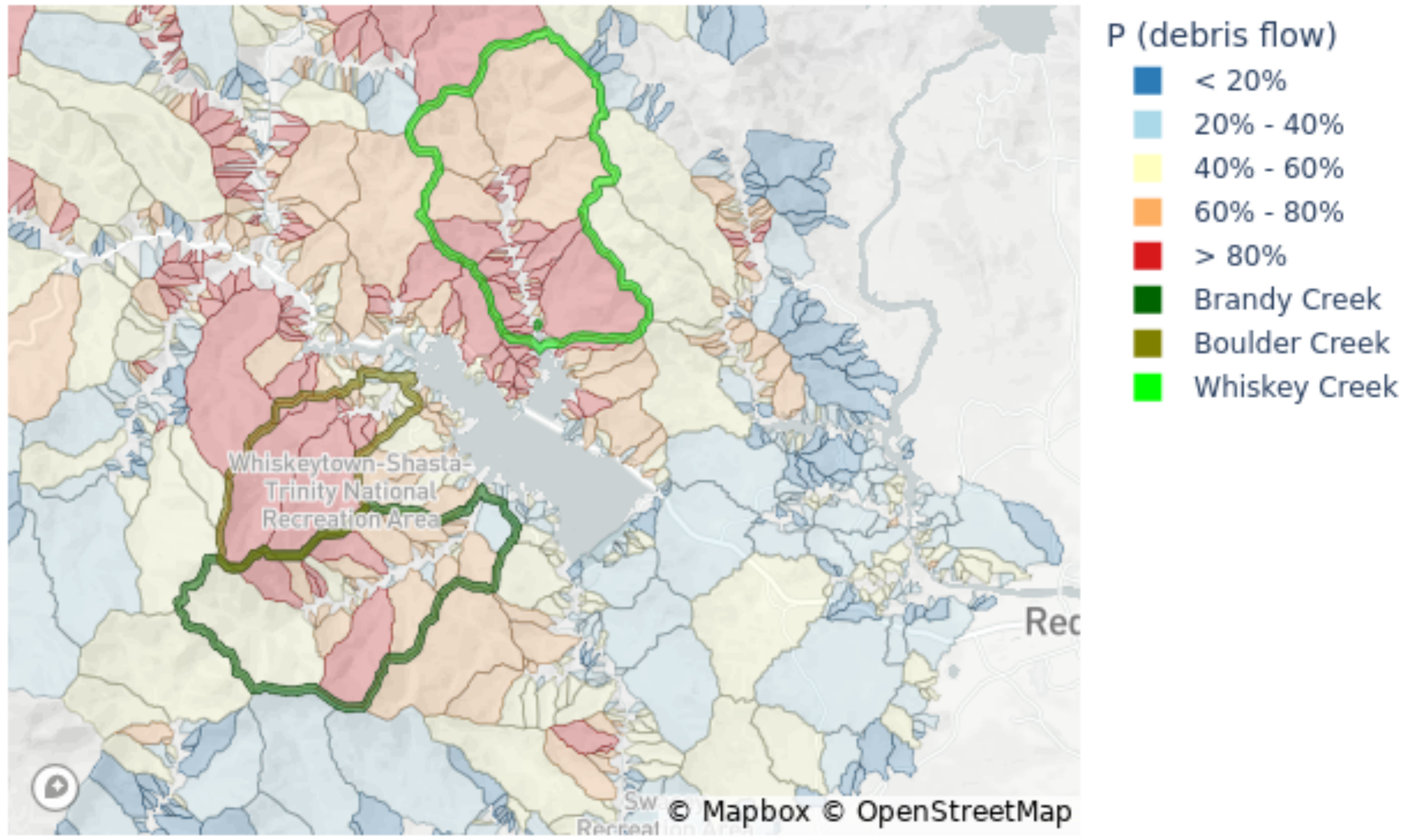

(a)
Staley (2016)
0.0
0.5
1.0
Debris flow Probability
(b)
Logistic Regression
(c)
K-nearest Neighbors
(d)
Support Vector Classifier
(e)
Random Forest
(f)
Neural Network
PropHM23
i15 (mm/h)
0.2
0.4
0.6
0.8
20
40
60
80
100

# Application to 2018 Carr Fire

East et al. (2021) monitored three creeks draining into Whiskeytown lake for two years after the fire and recorded no debris flows.

Predicted debris flow likelihood around Whiskeytown lake during a storm with
i15 = 24 mm/h, duration = 24 h, and accumulation = 20 mm.

Staley *et al.* (2016)

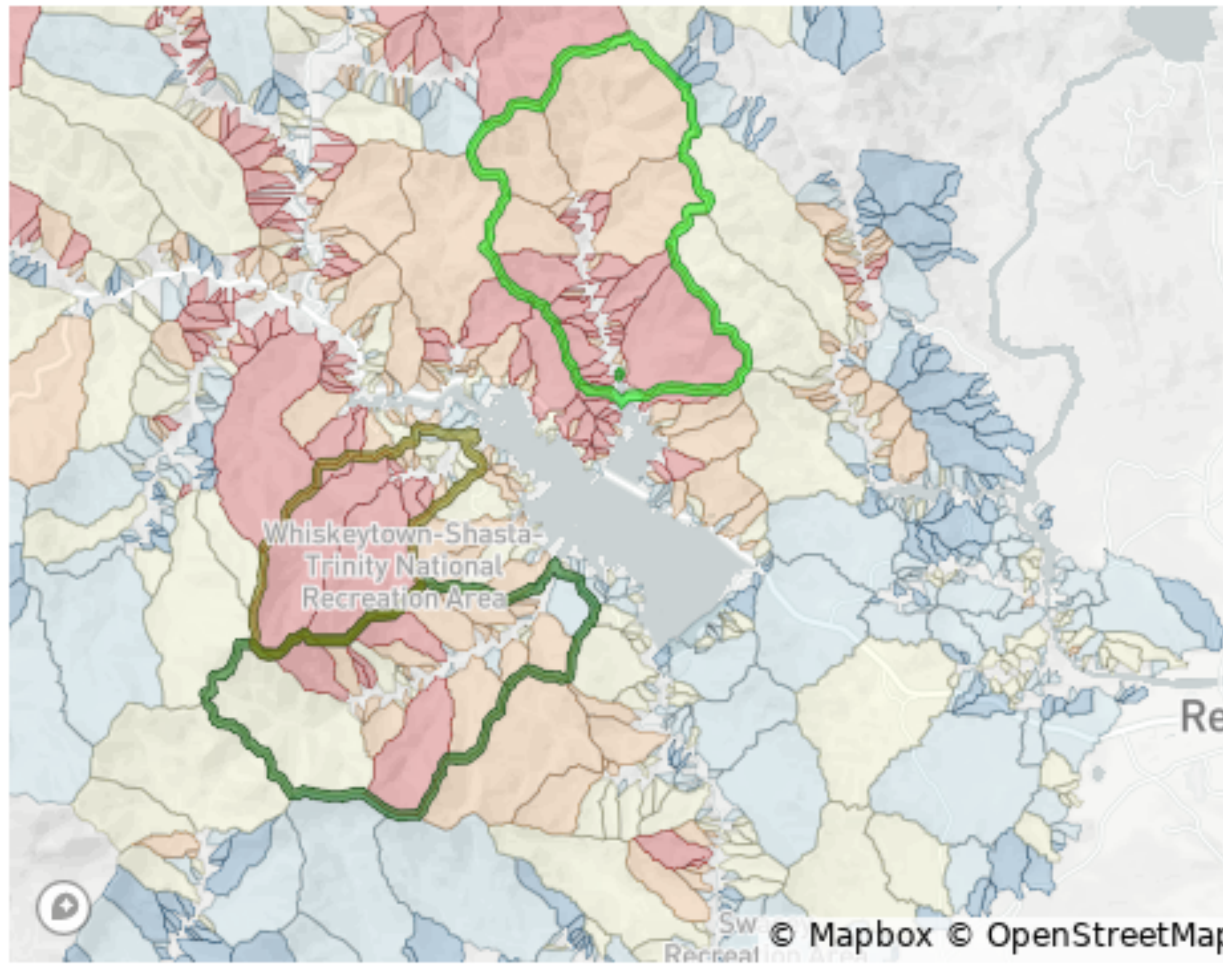


# Application to 2018 Carr Fire

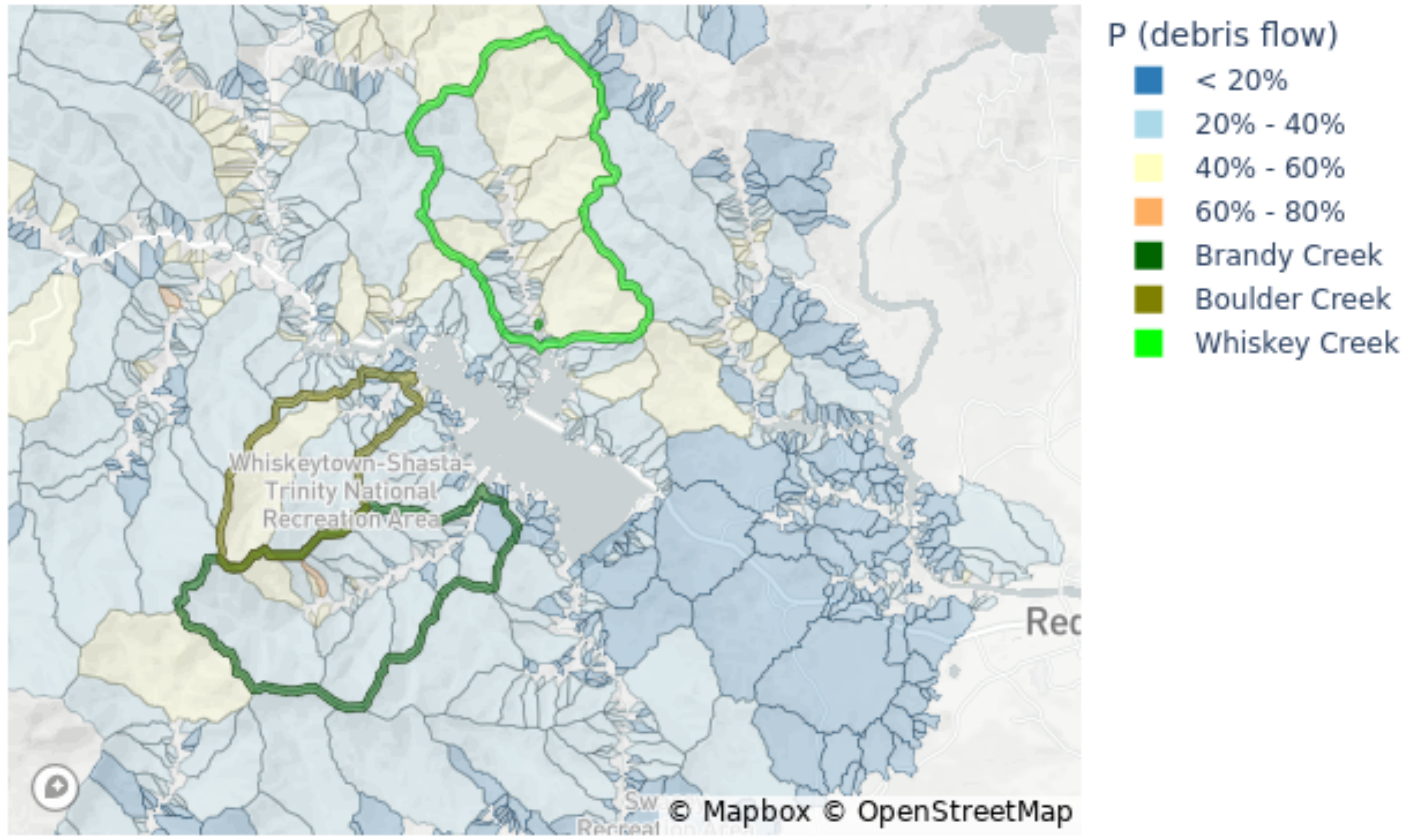
East et al. (2021) monitored three creeks draining into Whiskeytown lake for two years after the fire and recorded no debris flows.

Predicted debris flow likelihood around Whiskeytown lake during a storm with
i15 = 24 mm/h, duration = 24 h, and accumulation = 20 mm.

Staley *et al.* (2016)




Random Forest

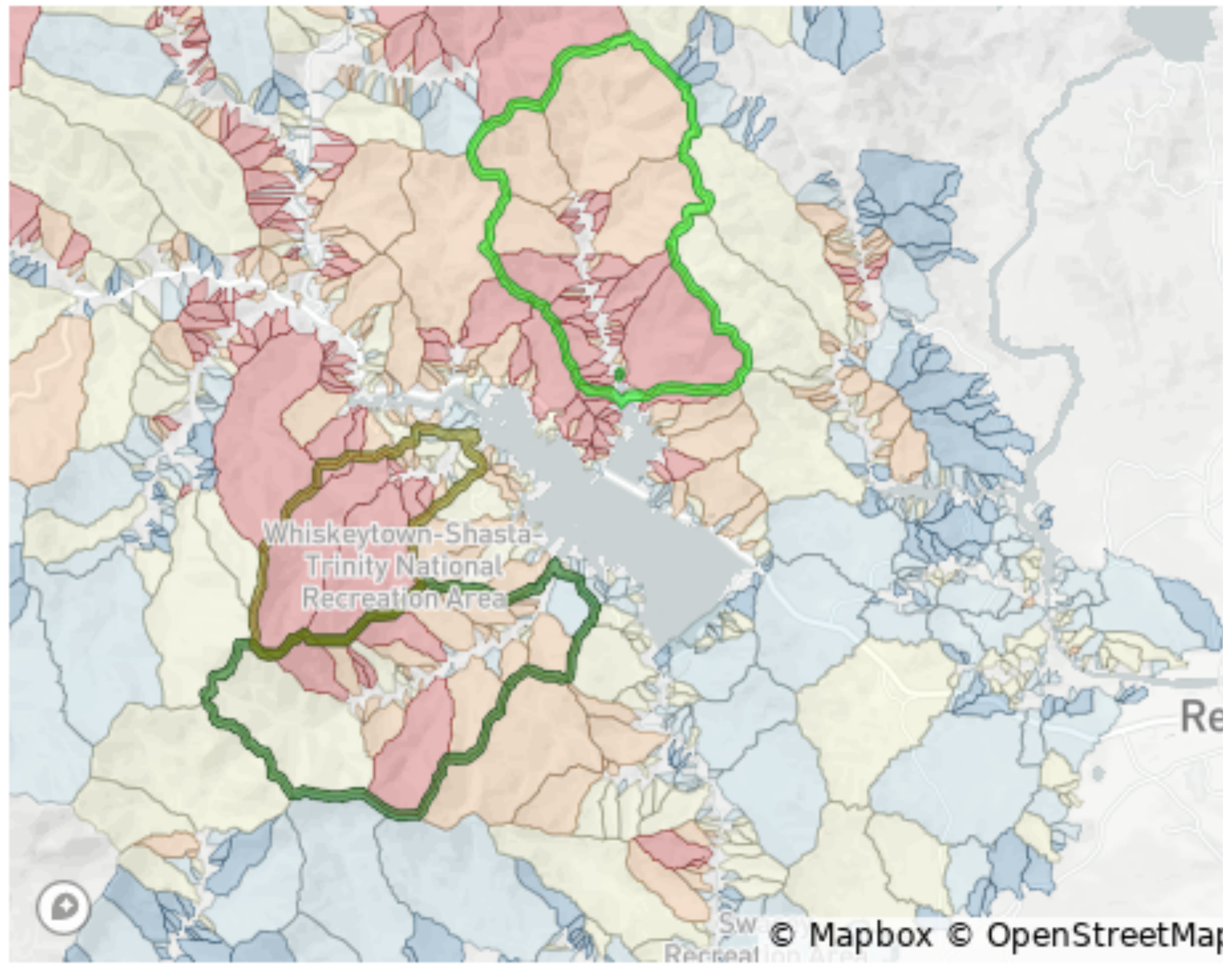


# Application to 2018 Carr Fire

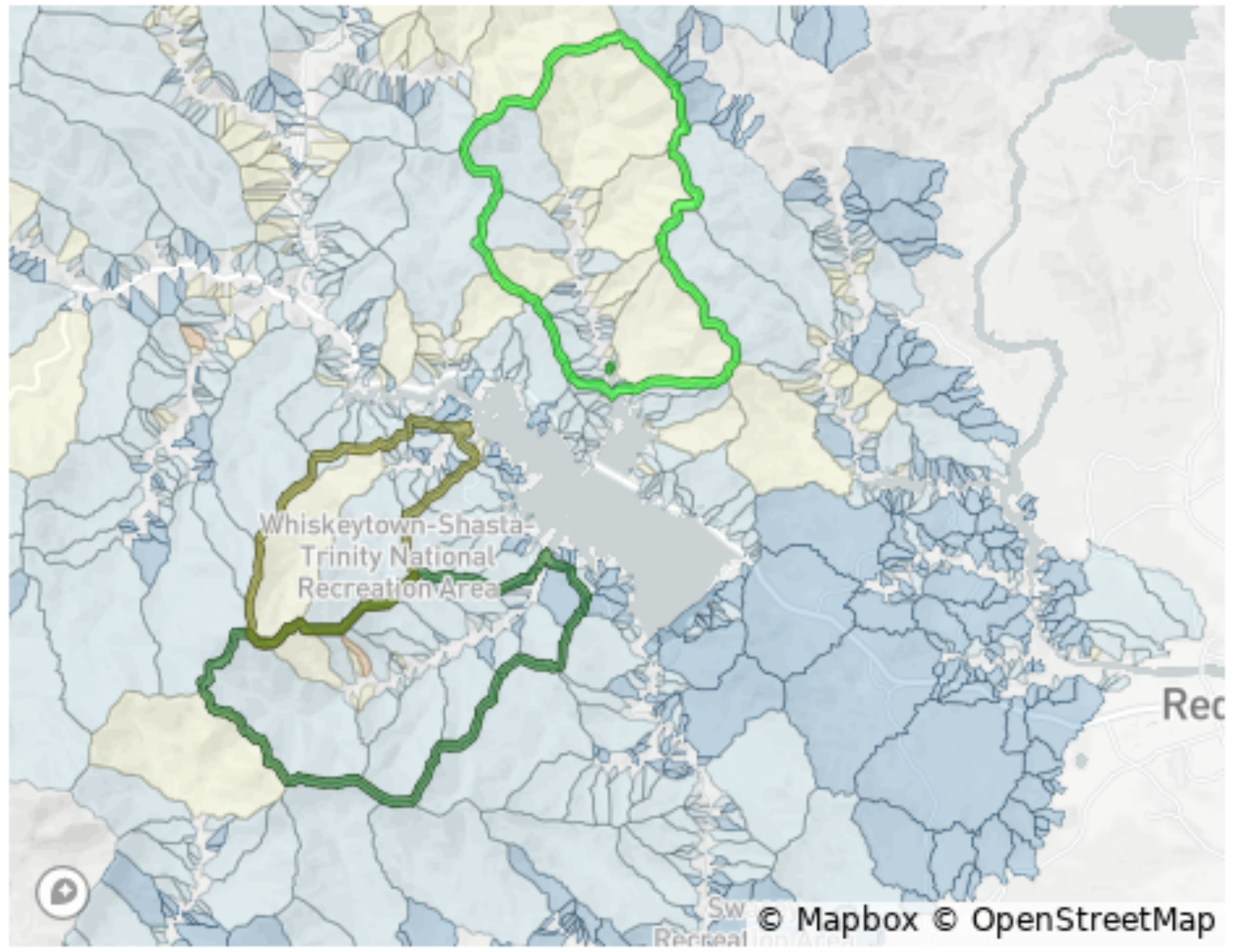
East et al. (2021) monitored three creeks draining into Whiskeytown lake for two years after the fire and recorded no debris flows.

Predicted debris flow likelihood around Whiskeytown lake during a storm with
i15 = 24 mm/h, duration = 24 h, and accumulation = 20 mm.

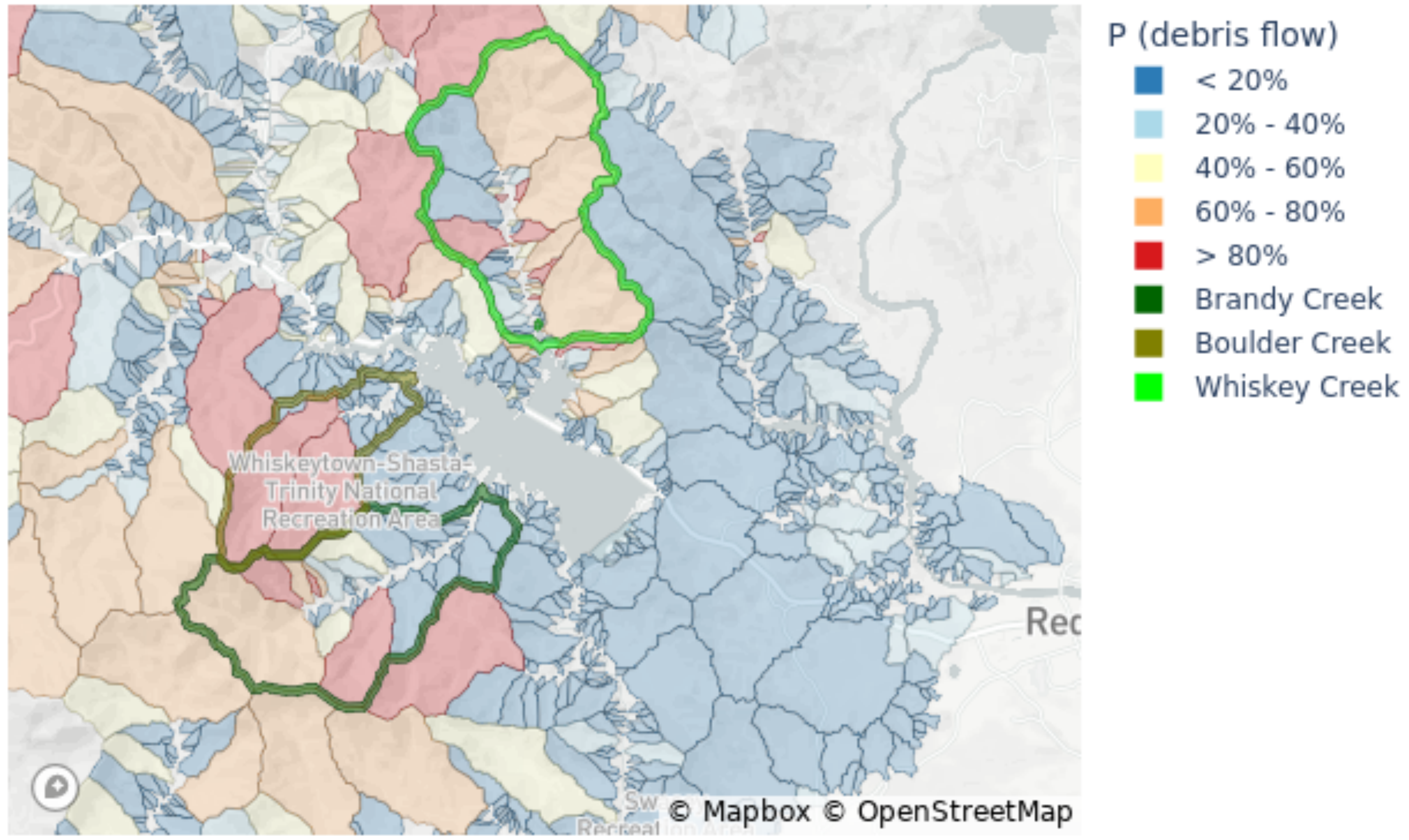
Staley *et al.* (2016)




Random Forest




Neural Network

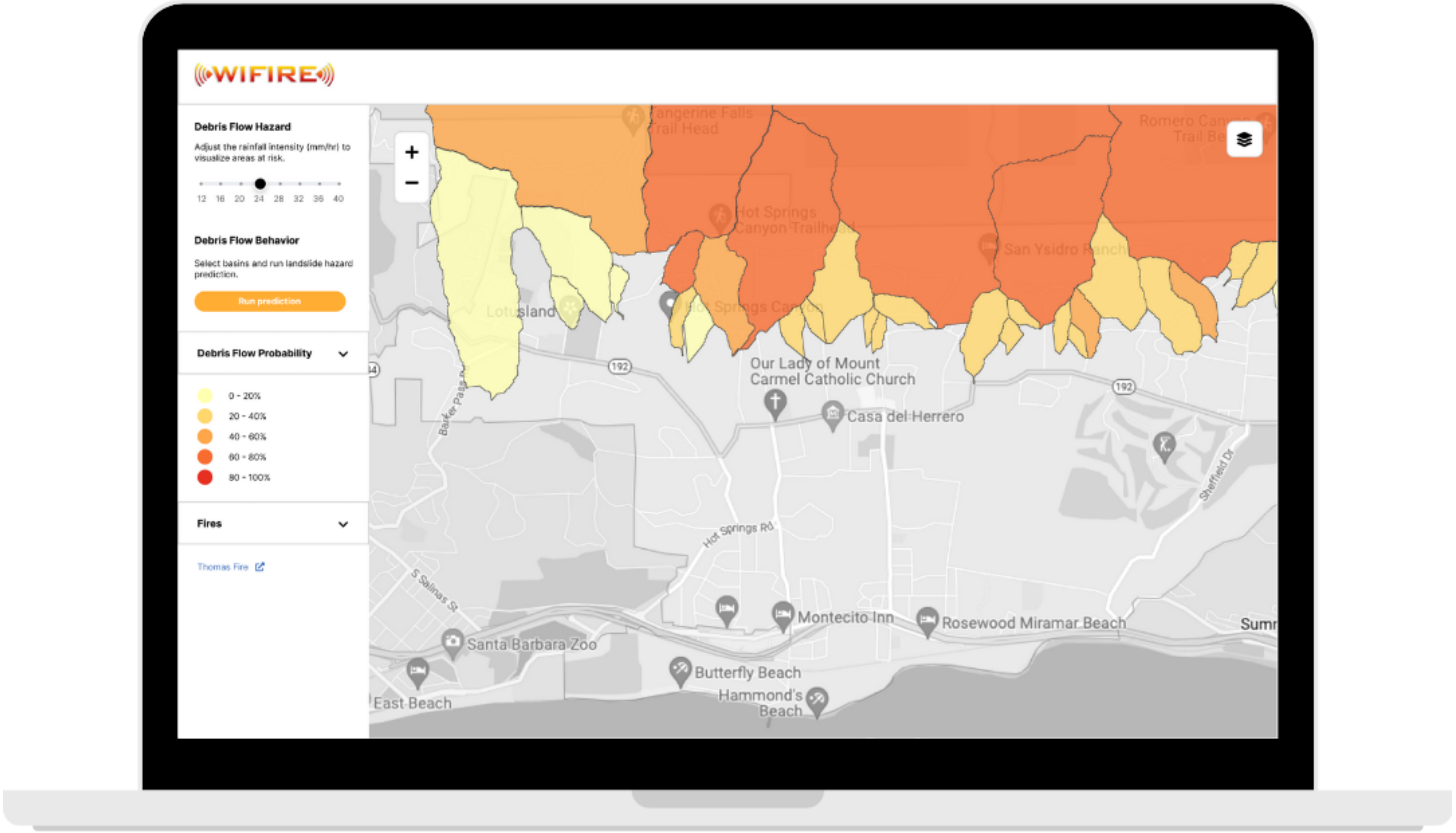


# Outlook: Real-time Debris Flow Prediction

An interactive web map displaying debris-flow likelihood for changing assumptions is under development.

Planned user controls:

- Storm parameters (i15, duration, accumulation)
- Choice of ML method
- Coupling storm parameters to live weather forecasts



# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.
- A scalable distributed **big data ingestion pipeline** for extraction of new features was developed.

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.
- A scalable distributed **big data ingestion pipeline** for extraction of new features was developed.
- Adding these additional features significantly improves the threat score and area under the ROC curve using both traditional ML methods and the fully connected NN.

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.
- A scalable distributed **big data ingestion pipeline** for extraction of new features was developed.
- Adding these additional features significantly improves the threat score and area under the ROC curve using both traditional ML methods and the fully connected NN.
- The random forest and NN with 11 hidden layers are our preferred models for debris flow prediction, with AUCs of 0.93 and 0.92, respectively (vs. 0.79 in Staley model).

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.
- A scalable distributed **big data ingestion pipeline** for extraction of new features was developed.
- Adding these additional features significantly improves the threat score and area under the ROC curve using both traditional ML methods and the fully connected NN.
- The random forest and NN with 11 hidden layers are our preferred models for debris flow prediction, with AUCs of 0.93 and 0.92, respectively (vs. 0.79 in Staley model).
- Both methods predict significantly lower debris-flow likelihoods in basins surrounding Whiskeytown Lake following the 2018 Carr Fire, bringing the predictions closer to observations.

# Summary and Conclusions

- Prediction of debris flow likelihood represents a **big data** problem due to the large variety and volume of partly time-dependent data sources.
- We have improved an existing debris flow likelihood model by using advanced traditional ML methods, deep learning and introducing new features.
- A scalable distributed **big data ingestion pipeline** for extraction of new features was developed.
- Adding these additional features significantly improves the threat score and area under the ROC curve using both traditional ML methods and the fully connected NN.
- The random forest and NN with 11 hidden layers are our preferred models for debris flow prediction, with AUCs of 0.93 and 0.92, respectively (vs. 0.79 in Staley model).
- Both methods predict significantly lower debris-flow likelihoods in basins surrounding Whiskeytown Lake following the 2018 Carr Fire, bringing the predictions closer to observations.
- An interactive web application is under development which will allow decision makers to consider these alternative debris-flow likelihood models when assessing the dangers posed by large precipitation events over fire-affected basins.

# Questions?

Email me at d1roten@ucsd.edu

# Acknowledgements

*The data processing pipeline was deployed on the Nautilus Kubernetes Cluster of the Pacific Research Platform using name-space "wifire-quicfire".*
*Thanks to Mai Nguyen for many helpful comments which helped to improve the manuscript.*
*Thanks to Abani Patra and Luke Mcguire for the successful collaboration that contributed to this research.*
*We thank the four anonymous reviewers for their helpful suggestions.*