

UC Davis
Pathology and Laboratory Medicine

Title

Foundations of Supervised Machine Learning in Clinical Predictions Research

Permalink

<https://escholarship.org/uc/item/26q4z1b7>

Authors

Fennell, Brandon
Rashidi, Hooman

Publication Date

2021

Data Availability

The data associated with this publication are not available for this reason: N/A

Foundations of Supervised Machine Learning in Clinical Predictions Research

Brandon Fennell and Dr. Hooman Rashidi
Department of Pathology and Laboratory Medicine
UC Davis School of Medicine

Introduction

Machine learning (ML) is an application of computational and statistical techniques to allow computers to learn and predict without explicit programming.¹ In recent years, with the increasing availability of large scale and low-cost computing power, ML capacity has expanded vastly and has begun to change how many industries operate.² The ability of machines to analyze large, complex datasets and to detect patterns beyond the scope of the human mind provides a powerful opportunity for application in a healthcare setting. ML has introduced new approaches to many dimensions of medicine including, but not limited to, Pathology,^{1,3} Radiology,⁴ drug development⁵, enhancing existing clinical predictive tools⁶, and the management of many diseases including cancer⁷ and autoimmune diseases⁸. Currently, ML remains in its infancy but has already started to make an impact in various healthcare disciplines.^{1,2}

This research project aimed to provide the foundational training and understanding of the modern approaches to ML and develop the skill set necessary to use available healthcare data to develop and deploy new ML models to assist in the delivery of future healthcare.

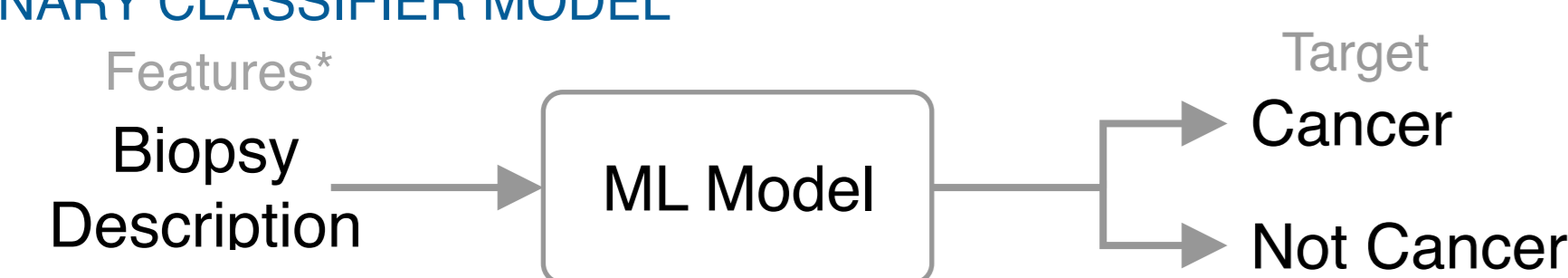
Learning Goals and Research Question

The components of this training were focused on the below four goals:

- Learn the foundational statistical concepts for evaluating ML models
- Learn the common approaches to supervised ML
- Learn the approaches to cross validation and understand how to define training and test data sets in the model building process (in tune with CRISP-DM)
- Learn how to optimize models using hyperparameter tuning

The exercises and applications were framed in the following research question: How do supervised binary classifier ML models perform relative to previously published breast cancer screening methods? The model is summarized in **FIGURE 1**.

FIGURE 1: BINARY CLASSIFIER MODEL



*features = ML term for attributes/data serving as variables that get mapped to the target (Not/Cancer)

Methods

The UCI Machine Learning Repository Breast Cancer Wisconsin Diagnostic Data Set was used for this project which is a publicly available online dataset with no patient identifiers made available through the ScikitLearn library.⁹ Scikit-learn 0.23 was used along with the Jupyter Notebook IDE to conduct this study.

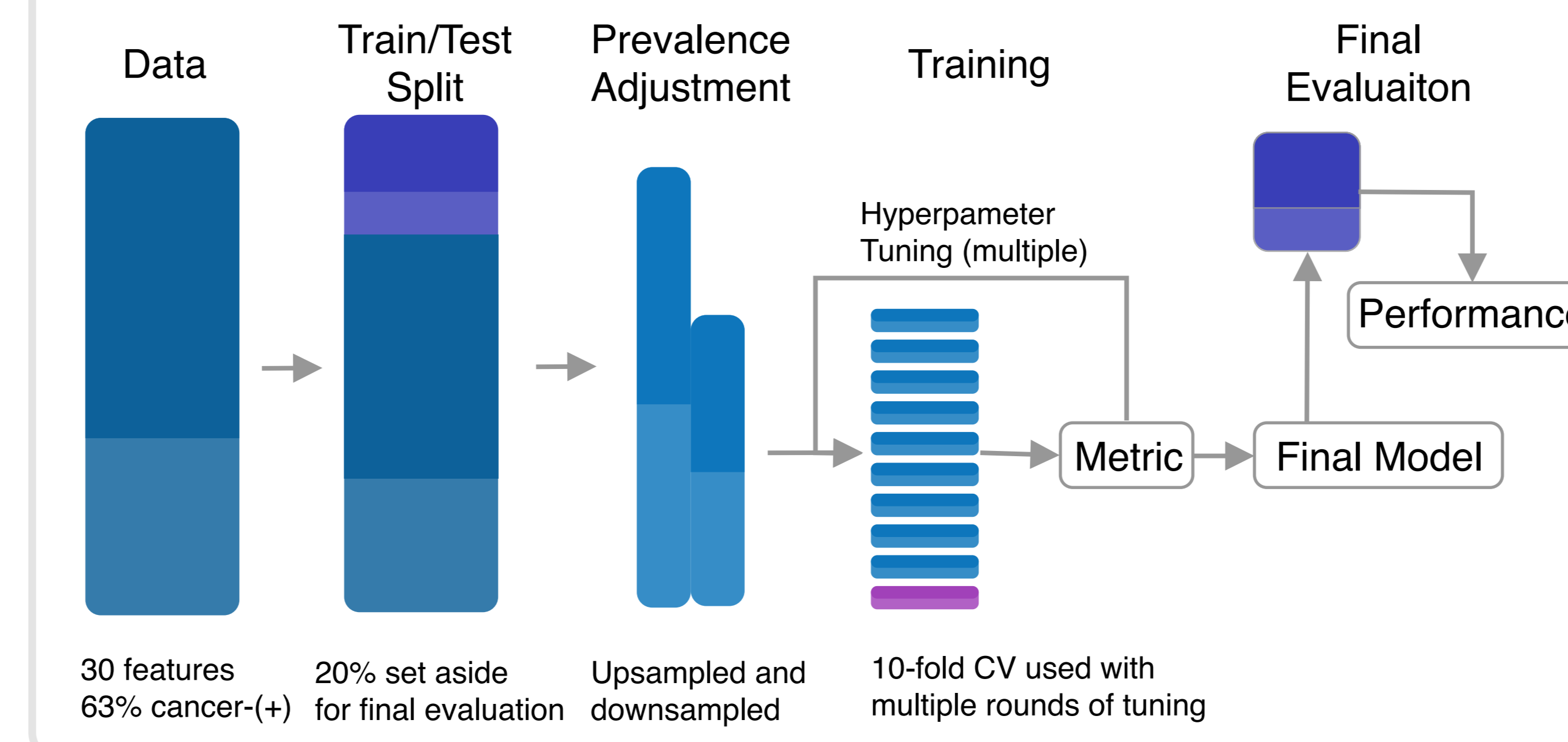
The model development process is summarized in **FIGURE 2**. The data set contained 30 features with a binary target variable (benign or malignant). 20% of the data was removed prior to training for use in final validation. In order to optimize the performance of the training algorithms and minimize initial bias, the training set prevalence was stratified to equalize the target classes (malignancy prevalence = 63% prior to equalization). The prevalence was equalized using both downsampling of the majority class (malignancy) and upsampling of the minority class (benign).

To identify and optimize the model, a grid search with 10-fold cross validation was employed for the below scikit-learn algorithms. The abbreviations used in figures/tables are provided in parentheses.

- Gaussian Naive Bayes (gnb)
- K-Nearest Neighbor (knn)
- Logistic Regression (log)
- Decision Tree (dt)
- Random Forest Ensemble (rf)
- Support Vector Classifier (svm)
- Multilayer Perceptron Neural Network (nn)

Methods Cont.

FIGURE 2: BINARY CLASSIFIER MODEL DEVELOPMENT AND EVALUATION



Pipelines (see **FIGURE 3**) were employed to scale the data for the appropriate ML approaches. The performance of each model was evaluated using mean accuracy of the 10-fold cross validation for each model. Multiple grid search trials were completed with modification of the pipelines and hyperparameters to improve specific model performance.

FIGURE 3: EXAMPLE PIPELINE CODE

```
def process_gs_pipeline(features, target, pipeline, gs_params):
    grid = GridSearchCV(pipeline, param_grid=gs_params, cv=10, verbose=0)
    model = grid.fit(features, target)
    return grid.cv_results_

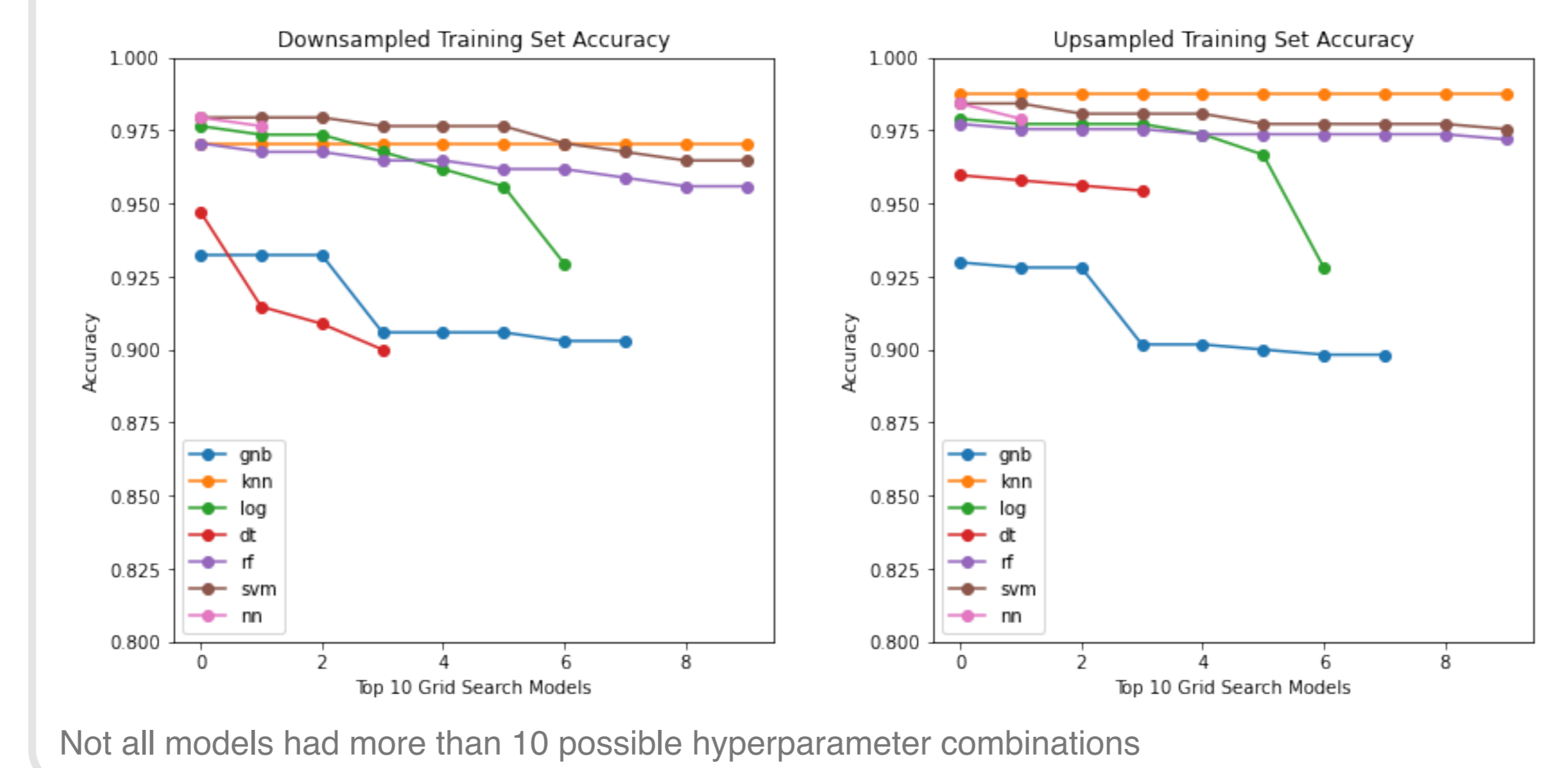
#KNN example pipeline and grid search parameters
knn_pipeline = Pipeline([
    ('scale', StandardScaler()),
    ('estimator', KNeighborsClassifier())
])
knn_param_grid = {
    'estimator__n_neighbors': range(1,20)[::2],
}
```

full code and writeup
<http://bit.ly/pd2021a>

Results: Model Training Metrics

FIGURE 4 shows the accuracy of the top performing models for each of the algorithms during training. Upsampling demonstrated a modest improvement in multiple algorithms which achieved comparable accuracy.

FIGURE 4 MODEL TRAINING PERFORMANCE BY APPROACH



Results: Model Performance

TABLE 1 shows the results of the final evaluation of the top models using 20% of the original data (previously unseen by the models). The precision of the K-Nearest Neighbor, Logistic Regression, Support Vector Classifier, and Neural Network models were all 0.99 with recall ranging from 0.96 to 0.92 with higher overall recalls observed for the upsampled training set. The Random Forest model showed comparable accuracy with slightly diminished precision relative to the other highest performing models.

TABLE 1: PERFORMANCE OF TUNED MODELS

	upsample			downsample		
	accuracy	precision	recall	accuracy	precision	recall
knn	0.96	0.99	0.96	0.95	0.99	0.93
svm	0.96	0.99	0.96	0.95	0.99	0.93
log	0.96	0.99	0.94	0.95	0.99	0.93
nn	0.95	0.99	0.93	0.95	0.99	0.93
rf	0.96	0.97	0.97	0.95	0.99	0.93
gnb	0.93	0.94	0.94	0.91	0.97	0.89
dt	0.92	0.94	0.93	0.93	0.94	0.94

Discussion

How does this model compare with existing methods? In a 2020 meta analysis by Yuan et al., breast cancer-screening via combined mammography and ultrasound had a clinical sensitivity (recall) of 0.96.¹⁰ A comparable precision metric was not available. None of the models were able to achieve comparable recall with this reported method.

With that in mind, it is also important to consider limitations. The clinical application of this model is limited on several fronts. First, the dataset employs biopsied tissue which has had numerous features that are not commonly evaluated in standard Pathology reports. Second, the generalizability and performance of this model is also unclear since an independent secondary dataset was not available to assess each model's true generalizability.

Future work in this area will continue to explore the core research question. Feature-based parameterization of images using classifier methods may not be able to enhance our current clinical screening or diagnosis of breast cancer. However, newer approaches using deep learning neural networks are showing exciting potential on the image itself rather than man-made extracted feature sets as shown in this study.³

Funding

Funding was generously provided by the UC Davis SOM Medical Student Research Selective program.



References

1. Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P. & Green, R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad. Pathol.* 6, 2374289519873088 (2019).
2. Deo, R. C. Machine Learning in Medicine. *Circulation* 132, 1920–1930 (2015).
3. Acs, B., Rantalainen, M. & Hartman, J. Artificial intelligence as the next step towards precision pathology. *J. Intern. Med.* joim.13030 (2020) doi:10.1111/joim.13030.
4. Strack, C., Seifert, R. & Kleesiek, J. [Artificial intelligence in hybrid imaging]. *Radiol.* (2020) doi:10.1007/s00117-020-00646-w.
5. Lin, X., Li, X. & Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* 25, 1375 (2020).
6. Rashidi, H. H. et al. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. *Sci. Rep.* 10, (2020).
7. Zhu, W., Xie, L., Han, J. & Guo, X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers* 12, 603 (2020).
8. Stafford, I. S. et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *Npj Digit. Med.* 3, 30 (2020).
9. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).
10. Yuan, W.-H.; Hsu, H.-C.; Chen, Y.-Y.; Wu, C.-H. Supplemental Breast Cancer-Screening Ultrasonography in Women with Dense Breasts: A Systematic Review and Meta-Analysis. *Br. J. Cancer* 2020, 1–16. <https://doi.org/10.1038/s41416-020-0928-1>.