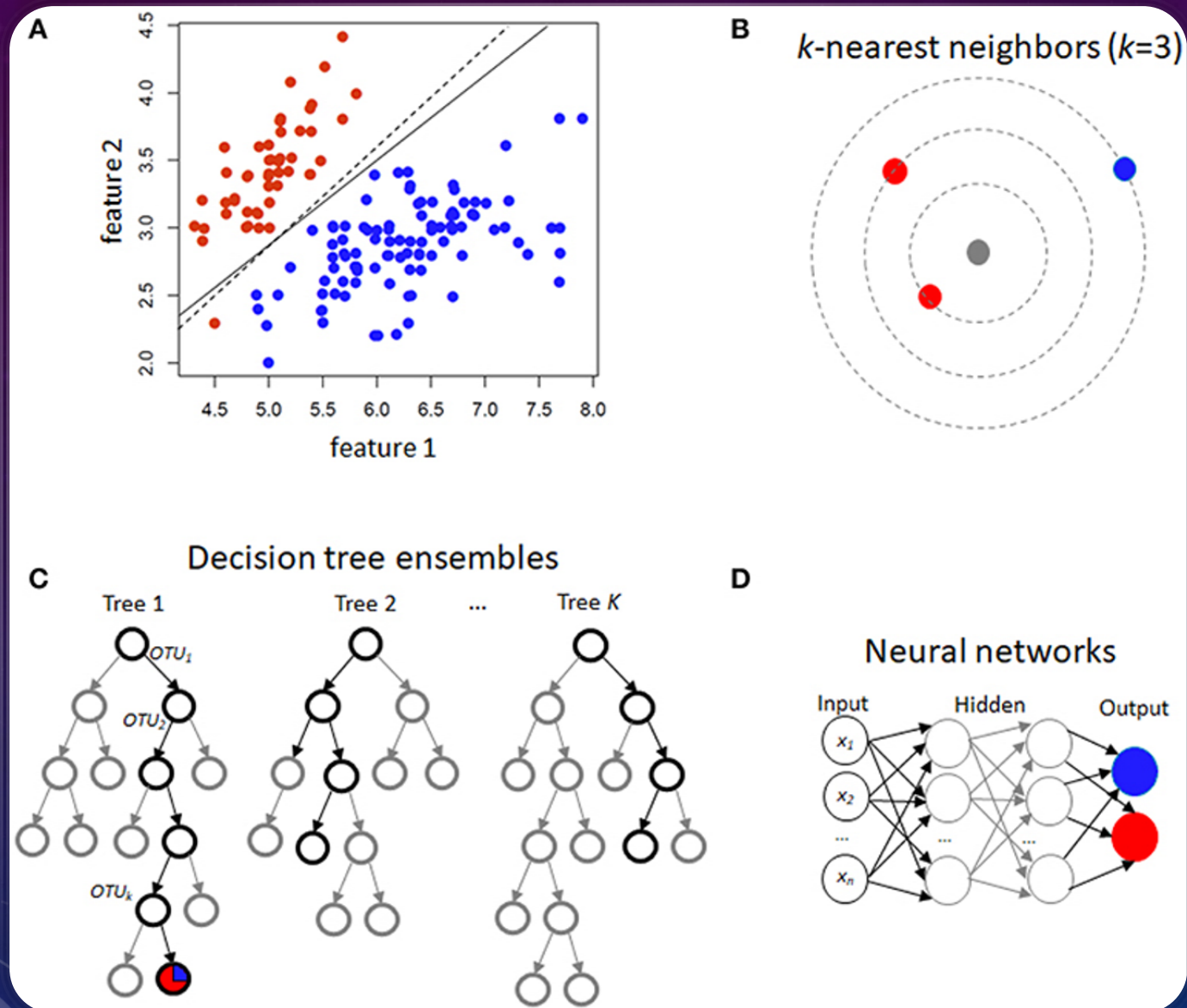


DISEASE PREDICTIONS FROM OPERATIONAL TAXONOMIC UNITS *A MACHINE LEARNING APPROACH*

TEAM 4

AGAZ WANI, ALEX DEAN, CHANG LI, NATHAN VAN BIDDER, PETER
RADULOVIC, YIBO DONG, GREG HERBERT, KRISTEN DOMINGUEZ, NGOC
TRAN, XIAOMING LIU

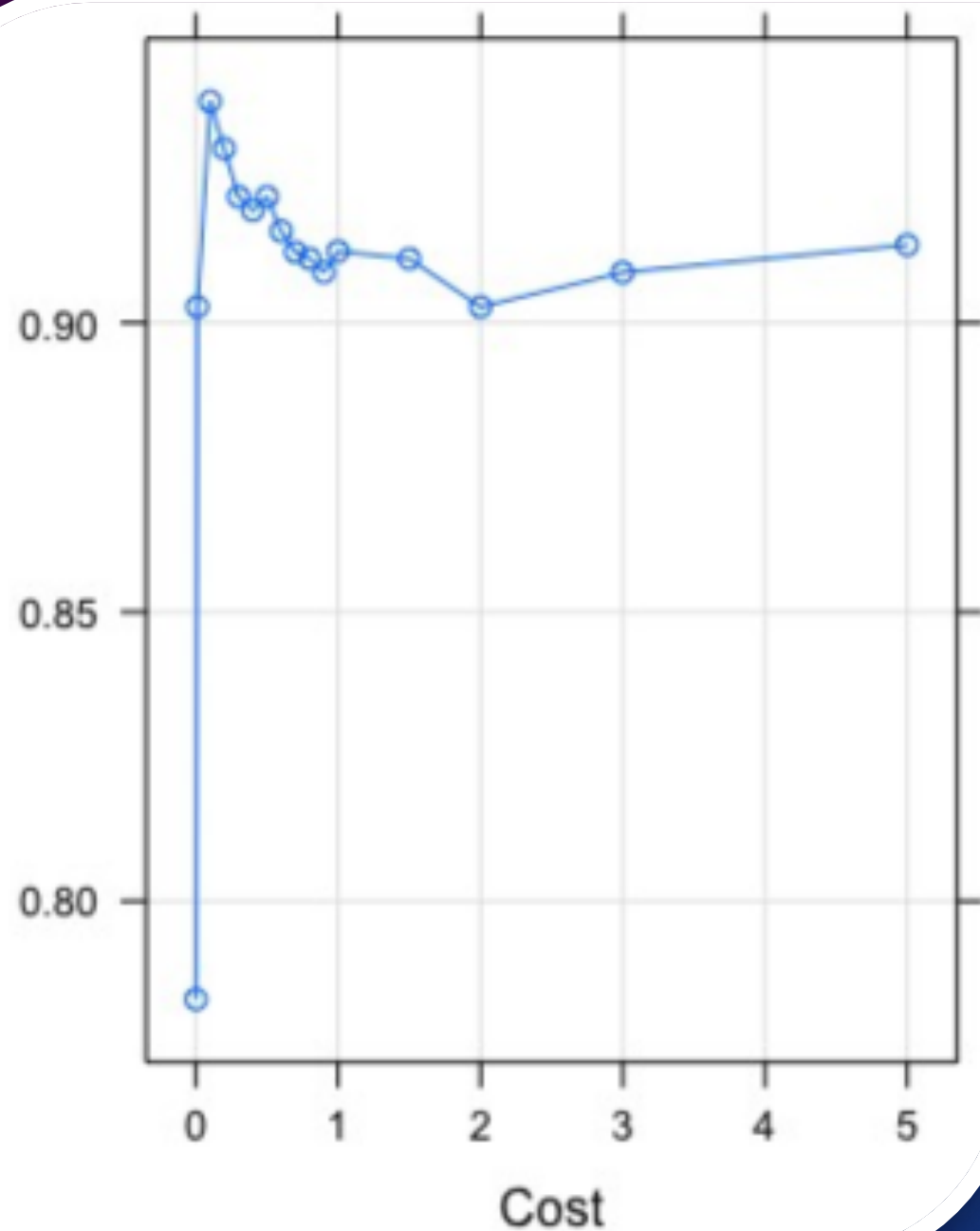


- Figure 1: Schematic illustration of several machine learning prediction methods using case/control (red/blue) status. For two features,
- **(A)** illustrates linear discrimination methods. The solid line shows the linear discriminant line corresponding to equally probable outcomes, while the dashed line shows the midpoint of the maximum-margin support vector machine.
- **(B)** For k -nearest neighbors, the gray point is predicted using an average of the neighbors (red, in this instance).
- **(C)** Decision tree ensembles include random forests, which average over bootstrapped trees, and boosted trees, where successive residuals are used for fitting. Trees may not extend to the level of individual observations, and modal or mean values in the terminal nodes are used for prediction.
- **(D)** A neural network with few hidden layers.

INITIAL MODEL DEVELOPMENT

- Add random small number to 0 values in OTU
- Combined features of OTU and age, family, ancestry, zygosity, twin type (No. of features = 5,468)
- Select minimum set of variables, conditioning on which the other variables will be irrelevant (No. of selected features = 26).
- 10-fold cross validation with linear Support Vector Machine to train the model on obese status (Overweightvs. Lean)

Accuracy (Repeated Cross-Validation)



Support Vector Machines with Linear Kernel

281 samples

26 predictor

2 classes: 'Lean', 'Overweight'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 253, 253, 253, 252, 253, 253, ...

Resampling results across tuning parameters:

C	Accuracy	Kappa
0.001	0.7830049	0.0000000
0.010	0.9027915	0.6579803
0.100	0.9383005	0.8033662
0.200	0.9301314	0.7795361
0.300	0.9218391	0.7538193
0.400	0.9194581	0.7493853
0.500	0.9218391	0.7578448
0.600	0.9158867	0.7433447
0.700	0.9122742	0.7340943
0.800	0.9111248	0.7301640
0.900	0.9087438	0.7229806
1.000	0.9123974	0.7381868
1.500	0.9110837	0.7345620
2.000	0.9027915	0.7108905
3.000	0.9087028	0.7263280
5.000	0.9135057	0.7457435

