Dean Taylor

STAT 517

15 November 2018

Literature Review: Support vector machines with feature selection in breast cancer diagnosis

Breast cancer is the second leading cause of death among women, even though it is one of the most curable forms of cancers, with treatments vastly improving the past decade. However, there is still a problem with identifying cancerous tissue in the early stages, with even the most experience surgeons finding difficulty in catching the cancerous cells in their earliest stages. One solution for diagnosing breast cancer in its earliest stages are using a support vector machine (SVM) to diagnosis early forms of cancer.

Support vectors (SV) define the boundaries between two classes, for example, in breast cancer researchers want to distinguish cancerous cells (malignant) and healthy cells (benign), using a set of features. However, designing an SVM is not without its limitations as well. One such problem is choosing the features to use in the predictive model, too many features often leads to complex models that fail with unknowns, while too few features fails with high variation. In this study researchers set out to use SVM feature selection to optimize predictions of cells collected as either malignant or benign.

At the time of publication, the authors state most other models using the UC Irvine breast cancer data set achieve prediction accuracy between 95% to 98%, but the author plans to optimize the feature selection and model parameters for greater predictive accuracy.

Using a linear SVM the goal is to maximize the hyperplane, the area between the two classes, using the closest points of each class, these are the support vectors. However, a linear SVM is not always possible, instead a non-linear SVM is used, where the SVM maps the vectors on a high dimensional plane, then the SVM sets out to maximize the hyperplane between the two classes.

SVM is uses a set of training vectors that belong to two classes, first we will review linear SVM, and separate them by a linear line. For binary classification we have the classic problem of :

Where x is the data point and its corresponding y. Linear SVM finds the optimal separating margin by vector w and the bias b.

A good SVM also wants to minimize

The hyperplane that stratifies these constraints and optimizations is determined by point called support vectors (SV). The SVs are a small subset of the training data. However, in many cases we cannot use linear SVM to separate our points, so instead we need to use non-linear SVM.

In non-linear SVM the principles remain the same, we want to separate two classes from each other, but in this case, we need to map the input vectors in high dimensional space defined as

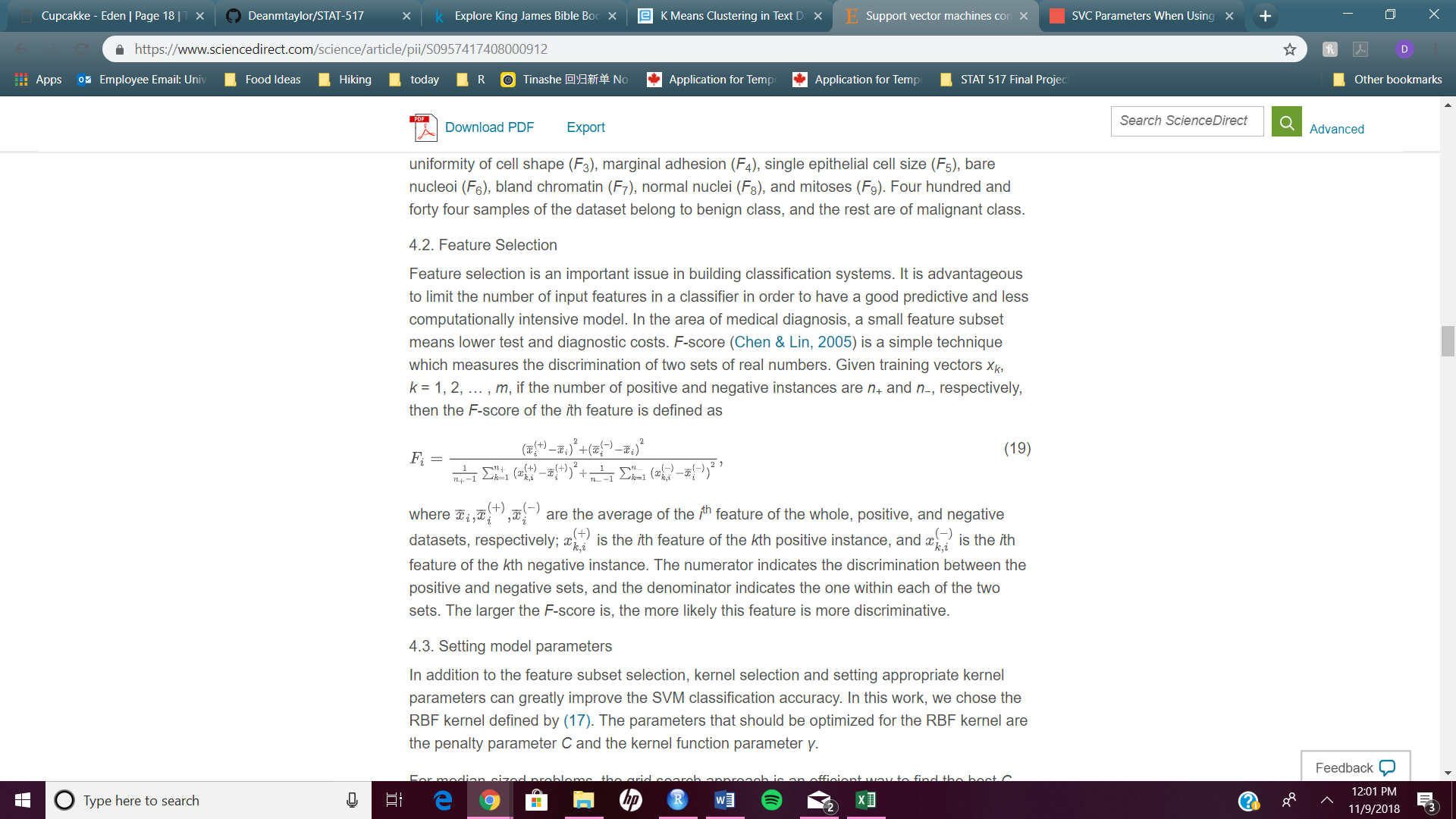
We use a kernel function, sometimes referred to as the kernel trick, to map the input into high dimensional space. The two most commonly used functions are the polynomial function and radial basis function (RBF)

RBF:

Polynomial function:

To select the proper features for their model the author sets out to calculate the F-score of each feature, the larger the F-score the more likely the feature is discriminative. The author then ordered the features based on the F-score, higher scores being the best, and then created models with increasing features based on the F-score. In terms of medical implications feature selection also helps doctors pick which features they should screen for, screening for many features takes more time and costs more money. So, the authors hopes that through feature selection doctors can only screen for the most useful features.

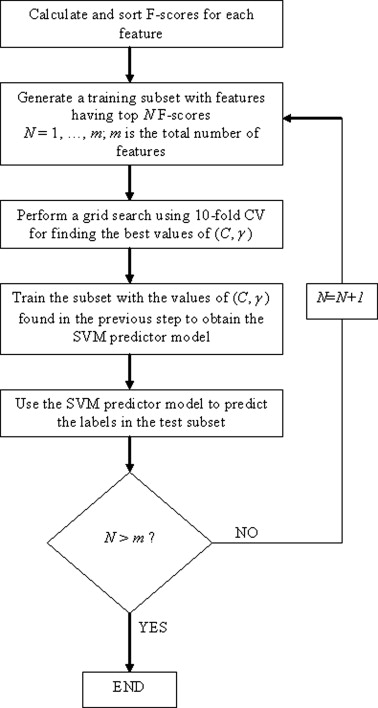
The author uses the following formula to calculate the F-score:



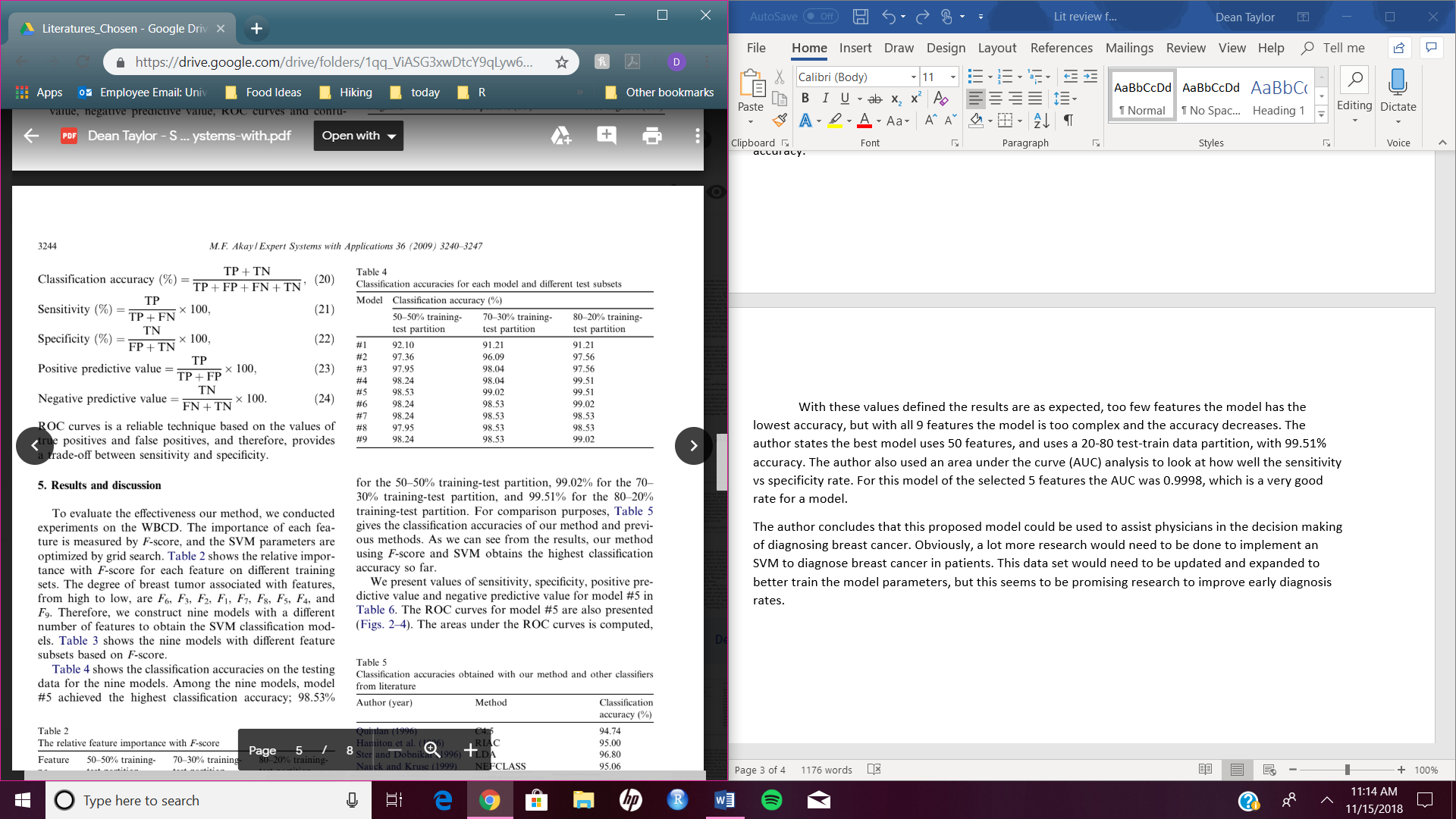
The numerator indicates the difference between the positive and negative set, while the denominator indicated one within each of the sets. The higher the F-score the more discriminative the feature, these F-scores will be used later to build the models.

One important part of building an SVM is optimizing the values of C and , these two parameters determine how strictly the model it going to classify the points, and the number of points that determine the support vector. C is the penalty parameter, so a large C aims to classify more the training sets correctly, while a low C increases the probability of outliers being misclassified. So, depending on the data set C needs to be optimized to better classify the training set. The term is the kernel function, meaning it adjusts the curvature of the decision boundary. With a high the boundary will be relatively smooth, while a low will make the decision boundary more curved and create pockets of classifications.

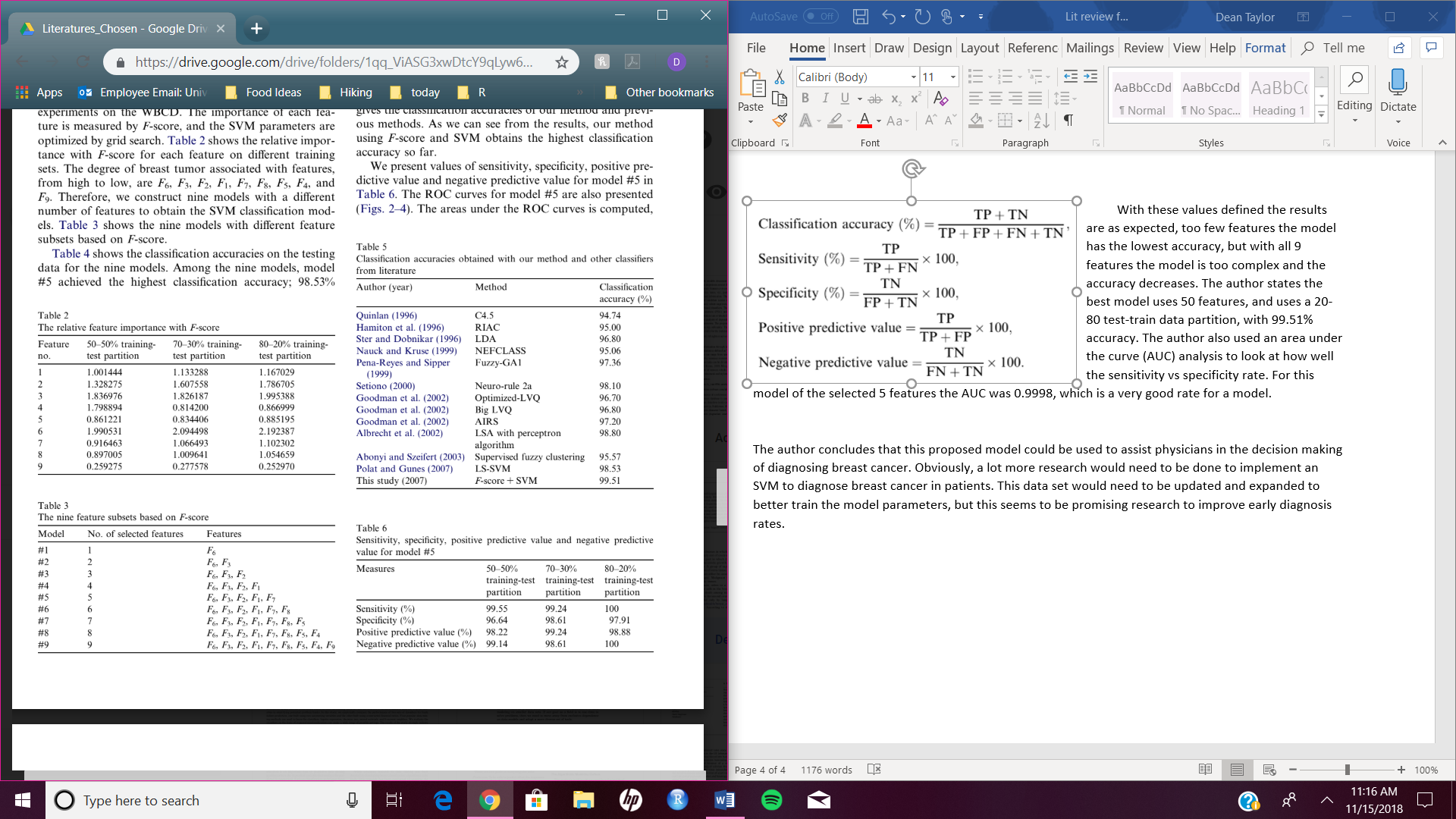
The author states for this example they are using grid search paired with 10-fold cross validation to optimize and C. Grid search works by using a pair of values for and C, then the one with the best cross validation score is chosen. So cross validation creates training sets based on the subsets of the dataset *D,* to create a training set =*D-.* Then, the pair of and C that leads to the highest classification rate based on k-fold CV is chosen.

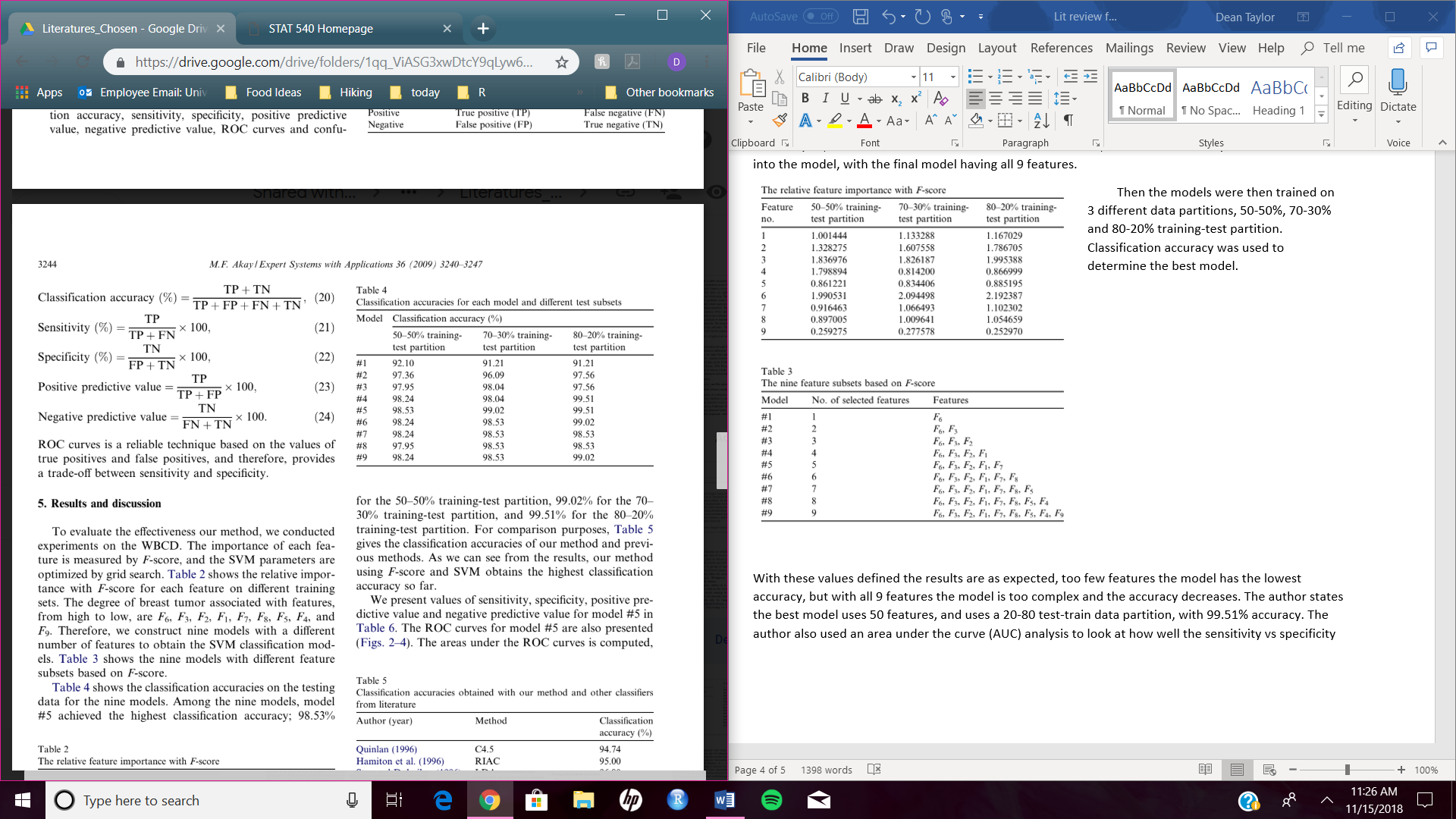
Figure 2. The author uses this figure to explain the flow of model creations. The first step of model creation is calculating the f-score based in the above equation. Next, the the training set is generated with the top F-score feature. Using 10-fold cross validation the C and are optimized. This first model is then training and predicts the labels of the test subset. For the first model the N = 1, but m = 9, so N<m, the program continues to build models until N>m.

The author than sets test parameters for these models, with 50-50 test-train, 30-70 test-train, and 20-80 test-train partitions of the data set. The author defines the parameters to define the best model sensitivity being rate of true positives, specificity being the rate of true negatives, and classification accuracy.

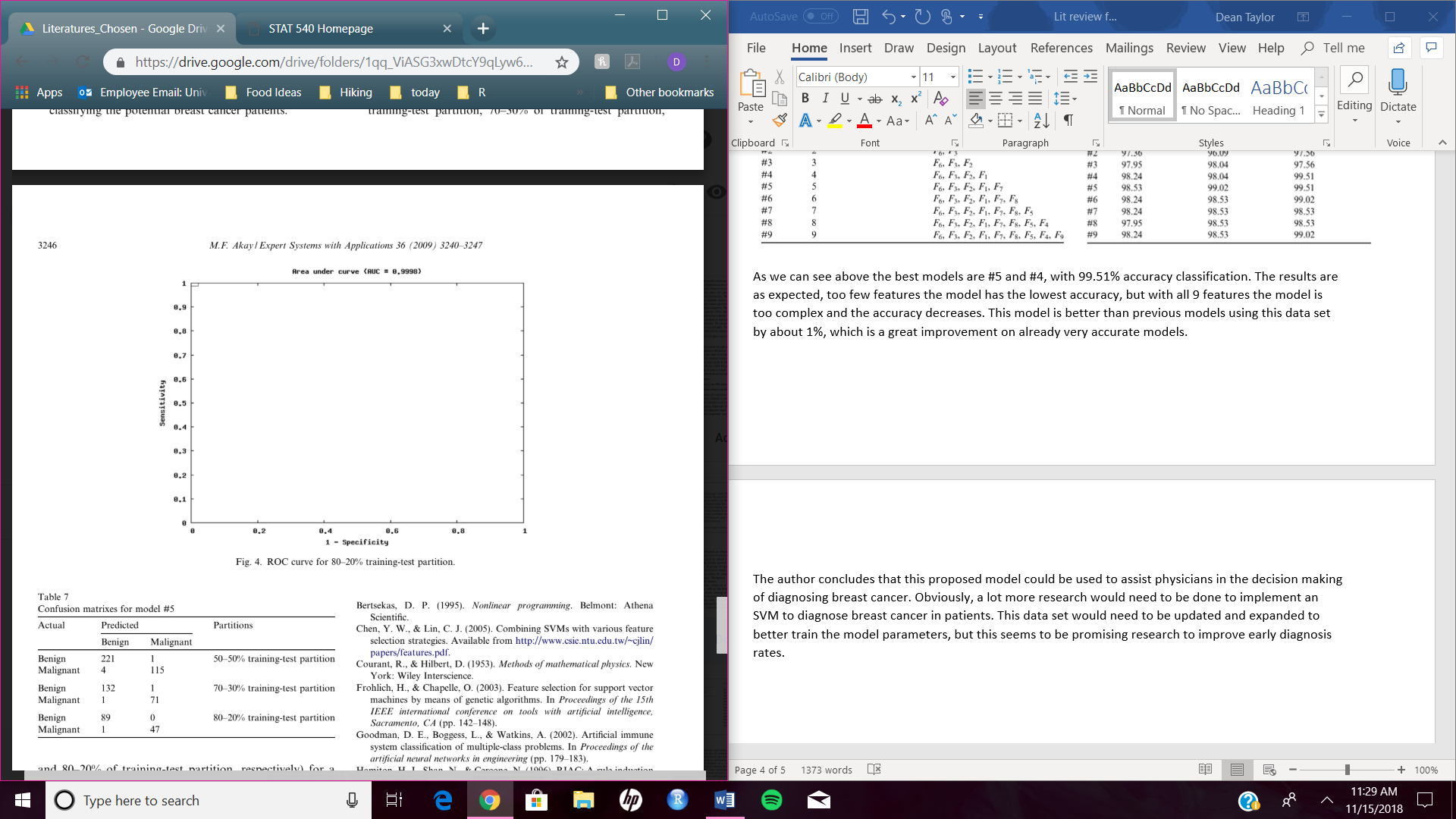
 These measures for model performance are common in predictive modeling. Classification accuracy is best for models that do not have biased classification, and for this data set it is not too biased in one direction. Sensitivity tells us the rate of true positives, while specificity tells us the rate of true negatives. In this breast cancer data set we would rather have higher sensitivity than specificity, false negatives are telling someone that has cancer that they are false, which could complicate treatment down the road. It would be better to assign someone without cancer as cancerous and do further tests, this would not risk their life. The author also uses ROC curves to get the area under the curve (AUC) for the sensitivity vs specificity.

The models were built in order of the calculated F-score, which can be seen below. The features are ordered from highest F-score to lowest, F6, F3, F2, F1, F7, F8, F5, F4, and F9. So, the first model built used only F6, then the second model was built with F6 and F3, and so on until every feature was added into the model, with the final model having all 9 features.

 Then the models were then trained on 3 different data partitions, 50-50%, 70-30% and 80-20% training-test partition. Classification accuracy was used to determine the best model.



As we can see above the best models are #5 and #4, with 99.51% accuracy classification. The results are as expected, too few features the model has the lowest accuracy, but with all 9 features the model is too complex and the accuracy decreases. This model is better than previous models using this data set by about 1%, which is a great improvement on already very accurate models.



As we can see from the ROC curve and AUC = 0.9998, the model does a very good job as classifying between cancerous tissue and benign tissue, using only 5 of the features in this data set. The model only misclassified one cancerous point as non-cancerous, ideally, we would rather all malignant points to be classified as such, so further work would be needed.

The author concludes that this proposed model could be used to assist physicians in the decision making of diagnosing breast cancer. Obviously, a lot more research would need to be done to implement an SVM to diagnose breast cancer in patients. The model is currently better than doctors themselves with classifying cancerous tumors, with doctor’s accuracy rate around 70%, this tool could easily assist doctors. This data set would need to be updated and expanded to better train the model parameters, but this seems to be promising research to improve early diagnosis rates.