At the start, the dataset had around 3,200 features divided into a training set, a test set and a blinded test set. As the initial step, the data was prepared so it would be clean and meet the needs for modeling. The redundant columns which held constant values taken out because they do not benefit the classification. Following this, records with no data (NaN) or infinite values were eliminated for the same reason. As a measure to solve possible multicollinearity, a correlation matrix was generated and features with a value higher than 0.95 were removed. The next step was to set a threshold for variance at 0.01, so any feature with less variability would be removed.

When this step was complete, the mutual information between every remaining feature and the target feature was valvulated. Features with a mutual information score of zero weren't kept, since they contained no information related to the target class. Those remaining features were then adjusted with the StandardScaler from scikit-learn.

Principal Component Analysis (PCA) was done to shrink the number of features without losing too much of the data's variation. A scree plot helped choose the number of main components, so then 50 were entered into the main model. The new data was then given as inputs to train Logistic Regression, Random Forest Classifier and Support Vector Machine (SVM). A 5-fold cross-validation grid search was used to adjust every model and accuracy was used to rate performance. For Logistic Regression, hyperparameters involved modifying regularization strength (C) and hyperparameters for Random Forest were adjusted by altering the numbers of estimators, the depth of the trees and the minimum number of samples in each split. For SVM, different values for C and both linear and RBF kernels were used.

The top models for every grid search were tried out on the test set with several statistics, mainly measuring accuracy, AUROC, sensitivity (recall), specificity and F1-score.

The calculated metrics are:

	Accuracy	AUROC	Sensitivity	Specificity	F1-Score
Logistic Regression	0.65	0.680213	0.33333	0.87910	0.4444
Random Forest	0.57	0.618021			0.271186
SVM	0.62	0.676519	0.214286	0.913793	0.321429