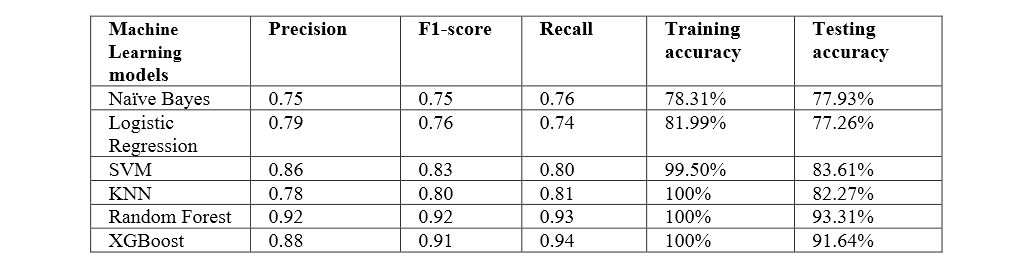
Jonathan Swanson, Jack Schnair, Ishan Patel, Zack Sabri, Vishal Patel

**Section 1: Paper Review: Nikhil Bora**

Nikhil Bora’s study titled “Using Machine Learning to Predict Heart Disease” assesses the usability of various different machine learning networks in predicting whether or not a patient has heart disease based on various numerical features. The networks that Bora investigated include: Logistic Regression, Naïve Bayes, Support Vector Machines, K Nearest Neighbor, eXtreme Gradient Boosting, and Random Forests. In his study, he determined the Random Forest was the most performant in predicting whether or not a patient will have heart disease, as shown below.



“Using Machine Learning to Predict Heart Failure”, p16

This is significant work in this field, because it provides meaningful results indicating that machine learning is indeed applicable in this area, and it is significant to us as it provides us with a wealth of knowledge we plan to expand upon in our project. Starting with the Random Forest algorithm, it is a supervised machine learning algorithm that creates several decision trees that it uses to then vote on and decide what the prediction is for that row of data. This idea of having multiple decision trees lends itself well to the idea of ensemble learning. Ensemble learning is a methodology that uses the results of multiple trained machine learning networks to come to a conclusion, rather than just using a single network. Intuitively, this seems as though it should lend itself to better performance, and it typically does. Using multiple networks allows for error to be reduced, as the bias that one network develops towards a certain set of features may not be present in other networks, which helps reduce mistakes made that were based upon that bias.

Nikhil Bora’s study lays a solid foundation for other researchers to build off of and expand upon his ideas. Using the fact that we know Random Forest (along with XG-Boost) to be highly performant in this area of numerical classification of heart disease data, one could expand upon this work by changing the targets of this algorithm to attempt to predict more detailed outcomes. In our case, this means predicting whether or not someone’s heart disease will be reversible or not. Knowing whether or not one’s heart disease will be treatable can help people make the best decisions for themself, and being able to accomplish this with machine learning would lead to a fast and accurate response to this question.

**Section 2: Paper Review: Fahmy et al.**

For the purposes of this project, numerical data is far more effective than image data. This is due to the nature of heart disease indicators statistics rather than how someone looks.

Nikhil Bora’s study yields upwards of a 93% accuracy in assessing heart disease with numerical data. In contrast, a study done with cardiac imaging, using many of the same techniques but with a CNN as well, yields accuracies in the 78 - 86% range. Additionally, studies such as one done by Tufts Medical Center using real patient data have assessed risk with accuracies in the 70’s to 80%.

Some of the primary statistics we use to learn from include age, sex, cholestorol, resting ECG, and heart rate. Of all of those, age and sex could supposedly be found from images of the patients, but it’s more efficient to just get the numbers. Working with numbers also allows us to come up with guidelines for a healthy lifestyle. If we only had images as data, we have little hard data to go off of for determining what causes heart disease. On the other hand, with numerical data we can see what a healthy range is for things like cholesterol and resting ECG. This way, people can try to live by the numbers that correlate to low risk for heart disease. The model will also run faster working with numbers instead of images, allowing for us to reasonably add more depth and layers for a better accuracy. Because we’re using numerical data, we will not be using a CNN (convolutional neural network). Instead we will simply use an ANN (artificial neural network). Using a combination of the superior data that Tufts has and the effective techniques that Nikhil Bora employed, we can effectively increase risk detection accuracy.

**Section 3: Project Plan**

Our project is to predict whether a patient has heart disease or not. If the patient does have heart disease, we want to predict whether it is reversible or not. Our goal is to classify heart diseases into 3 categories: doesn’t exist, reversible, or not reversible. We will use numerical data from Kaggle and utilize ensemble learning to create a model that predicts the type of heart diseases. Each member in our group will create their own model and run it to get an accuracy. We will then combine all the models (5 total) to create a much more accurate model.

Nikhil Bora’s study shows that random forests give the best accuracy. Random forests are based on ensemble learning. We will apply this ideology and use our 5 models to create the most accurate model. Multiple models will reduce errors. Bora’s study shows that random forests gave 100% training accuracy and 93% testing accuracy. Random forests model was the most accurate out of all the other ones Bora tested. We will try to achieve a success rate greater than 93%.

This experiment will help us determine how different models help classify heart diseases and whether machine learning is a viable option for determining heart disease. By using ensemble learning, we will determine whether using multiple models is better than using a single model. This will help future researchers determine whether they should use ensemble learning or not for their problem.

**Section 4: Project Methodology**

The heart disease dataset will be downloaded from the kaggle database. Rows of data that have null values will be dropped. Features and target labels will be extracted from the csv file using the pandas library. Data will be visualized using matplotlib. Training and testing values will be extracted from the dataframe and the data will be split into training and testing data via scikit-learn's train\_test\_split method. The data will be normalized so that the gradient descent algorithm can converge more efficiently, and so that features with relatively larger values do not influence the model's decision making more than they do in comparison to the other features as they may or may not be more indicative of heart disease just because their numerical value is higher on average than a different feature. Five Artificial Neural Network models will be created, one by each team member. Each team member will explore a multitude of different model structures, optimization algorithms, learning rates, training techniques and number of epochs. These five fully trained models will be saved using built in Pytorch methods. After model completion from each member, we will utilize a technique that was inspired by one of the fundamental ideas behind the Random Forests that performed very well on this particular dataset, that idea is ensemble learning. The five unique models from each team member will be placed into a python array, predictions will be made by each model on a specific testing datapoint, for each prediction, ensemble voting will be enacted, and the classification that is seen the most in the prediction array will be the determined classification of our ensemble model. The accuracy will be calculated by creating a prediction array from the ensemble models predictions and comparing them against the ground truth labels of the test data array. A confusion matrix will be used as a prediction analysis tool to better understand how our model is performing on specific classifications. Further work would be to utilize sequential boosting algorithms to reweight wrongly classified data.

**Section 5: Conclusion**

To conclude, we want to achieve a success rate better than 93% which is doable since we will be using ensemble learning which will allow us to accurately analyze the 5 models each team member is going to make. The work given by Nikhil Bora will be a leading example for us to complete this project and it will give us a ton of insight as to how to code certain models. Once each model is trained using a different number of epochs, it will give us an indication at how many epochs our training models perform best. Overall, we will be able to create a model that will allow us to predict whether a patient has heart disease or doesn't exist and whether or not it's reversible or not.