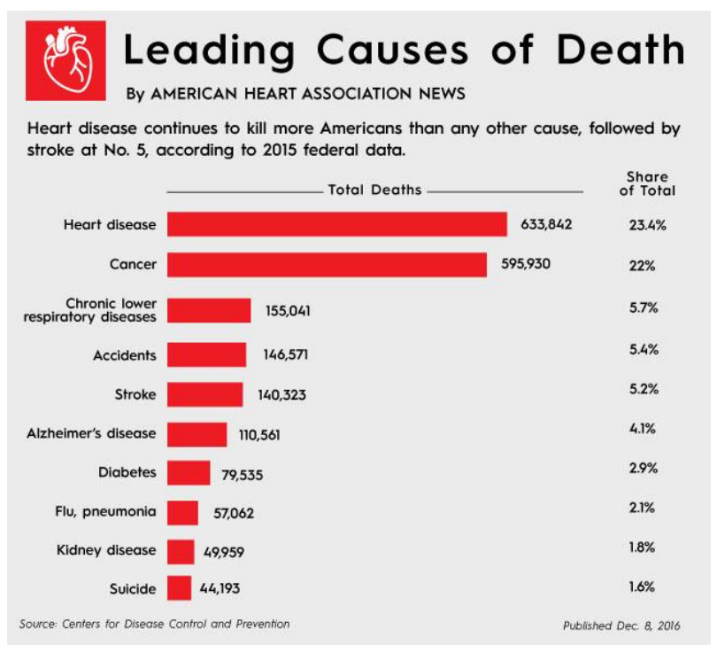
A NOVEL APPROACH TO THE DIAGNOSIS OF HEART DISEASE USING MACHINE LEARNING AND DEEP NEURAL NETWORKS

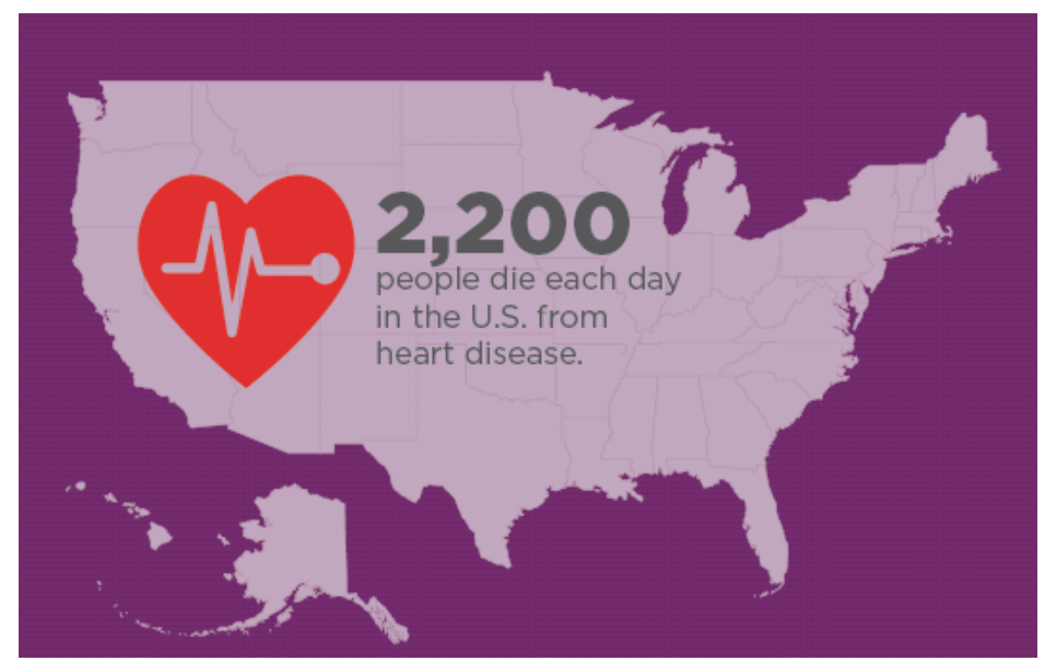
SAHITHI ANKIREDDY

JAMES B. CONANT HIGH SCHOOL

PURPOSE

- Heart Disease is the leading cause of death for both men and women. With the number of heart disease patients expected to rise in the near future, it is vital to find a solution.
- About 610,000 people die of heart disease in the US every year
- Early detection could be the difference between life and death for some





RESEARCH PROBLEM/GOAL

- Currently there is a only 1 out 3 heart disease cases are misdiagnosed
- The use of Artificial Intelligence (AI) and more specifically Machine Learning (ML) and Deep Neural Networks (DNN) can mitigate the possibility of human error while increasing prediction accuracy rates.
- Goal is to develop a system that assists in a fast, accurate, and accessible way to diagnose heart disease.

DESIGN REQUIREMENTS

- A higher accuracy prediction rate of at least 75% or more for both the DNN and ML model.
- A final application of which medical care professionals can easily use in everyday patient visits.

BACKGROUND

- ML is a general data analysis technique that uses computational models and methods to "learn" information directly from data without rule based programming. Subset of ML includes DNNs.
- 2 types of data: supervised and unsupervised, labeled or unlabeled data respectively
- Supervised learning uses two main techniques: classification and regression.
- Classification algorithms produce discrete responses or binary data. There are only two outputs, usually in the form of "yes or no" or "1 or 0" .
- Regression algorithms predict continuous responses.

DATA / FEATURES

- A dataset provided by the Cleveland Clinic Foundation was used for the ML and to train the DNN.
- Dataset contains 75 total attributes for 303 patients and 14 attributes out of the 75 were chosen.

	Attributes	Information
1	age	age in years
2	sex	1=male,0 =female
3	cp (chest pain type)	1= typical angina, 2= atypical angina, 3= non-anginal pain, 4= asymptomatic
4	trestbps (resting blood pressure)	in mm Hg on admission to the hospital
5	chol (cholestrol)	serum cholestoral in mg/dl
6	fbs (fasting bloood sugar)	1 = true; 0 = false
7	restecg (resting electrocardiograpic results)	0= normal; 1= having ST-T wave abnormality; 2=showing probable or definite left ventricular hypertrophy
8	thalach (maximum heart rate achieved)	
9	exang (exercise included angina)	1 = yes; 0 = no
10	oldpeak (ST depression induced by exercise relative to rest)	
11	slope (the slope of the peak exercise ST segment)	1= upsloping ; 2= flat 3= downsloping
12	ca (number of major vessels colored by fluoroscopy)	
13	thal (thallium heart scan results)	3 = normal; 6 = fixed defect; 7 = reversable defect
14	num (patient diagnosis of heart disease and predicted atttribute)	0=unlikely to obtain heart disease, 1=likely to obtain heart disease

METHODS

1. READING DATA : Data set was split into 2 sets: *trainData* with 236 records and *testData with 61 records.* Each set was further separated into the 13 attributes and diagnosis

2. DEFINE MODEL :
RandomForest classifier algorithm for ML model

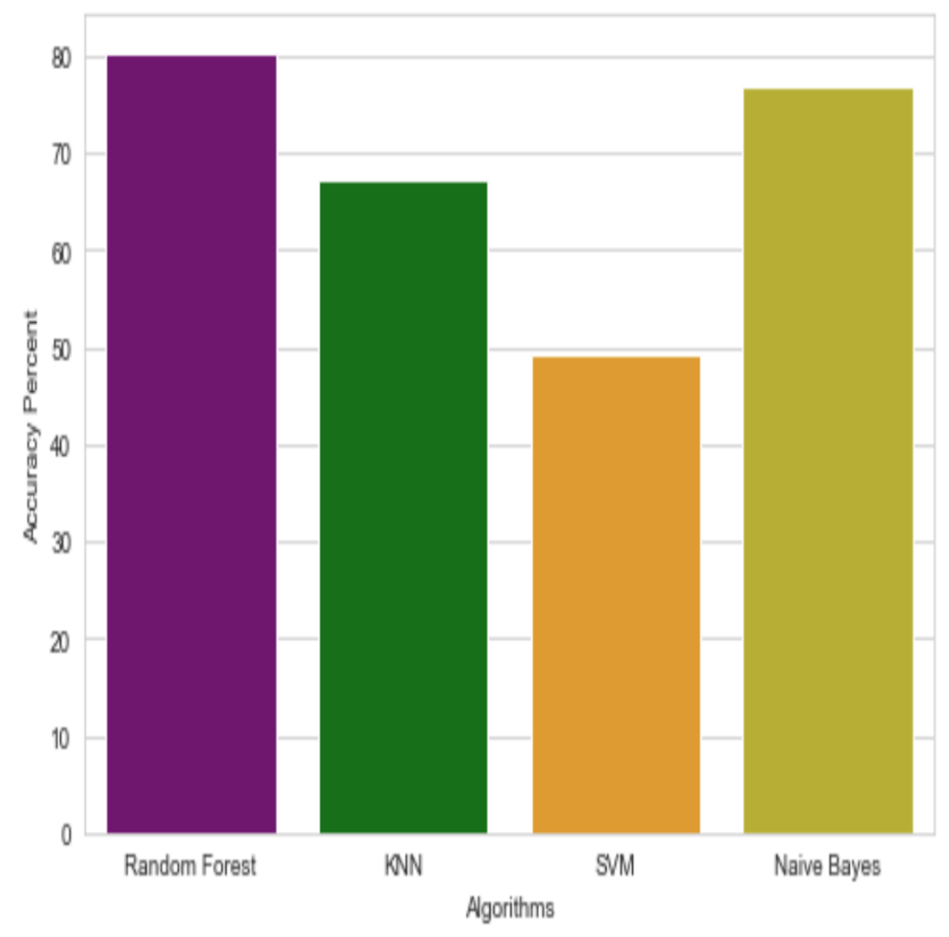
3. COMPILE MODEL: Model is being prepared for performing and predicting the last value (heart disease diagnosis, 1 or 0).

4. FIT MODEL: Model is trained or fitted on the loaded data by calling the fit() function of the model for ML. For the DNN, the epochs represent when an entire dataset is passed forward and backward through the neural network

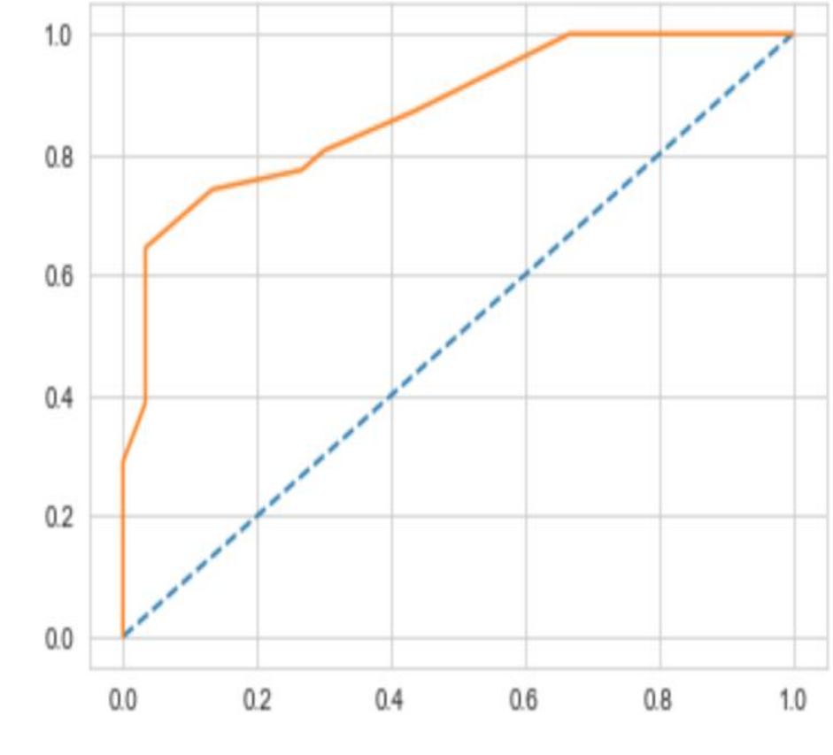
5.EVALUATE MODEL: Model in ML and the DNN have been trained. Now, the performance of the model or network will be evaluated via the test data.

DEVELOPMENT / MACHINE LEARNING ALGORITHMS

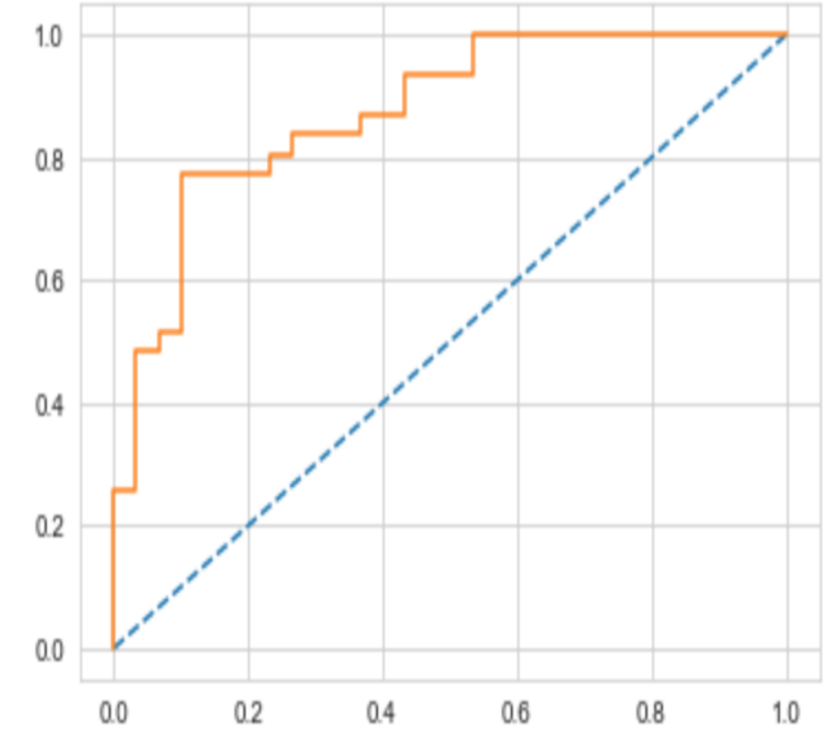
- Accuracy scores were produced for each of KNN, SVM, RF and Naive-Bayes Algorithms and as shown Random Forest had highest accuracy score



- ROC graphs were established for RF, KNN, Naive-Bayes with the AUC score. RF and Naïve-Bayes are Shown Below respectively.

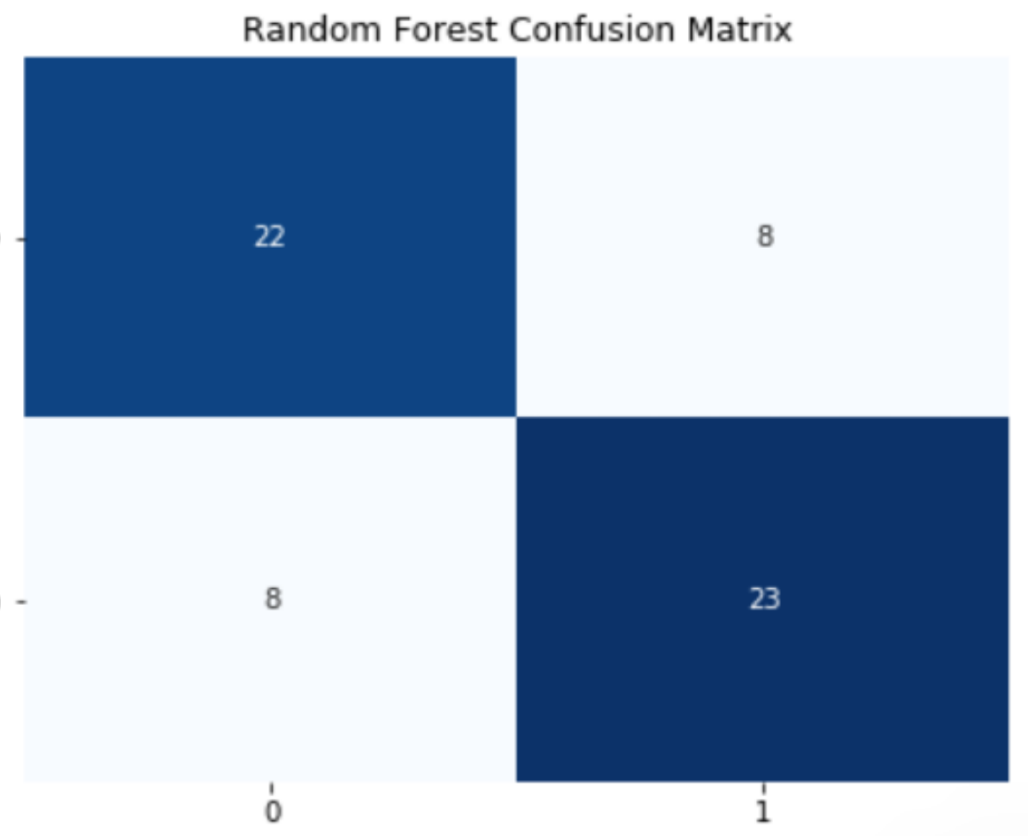


```
0.8715053763440861
```

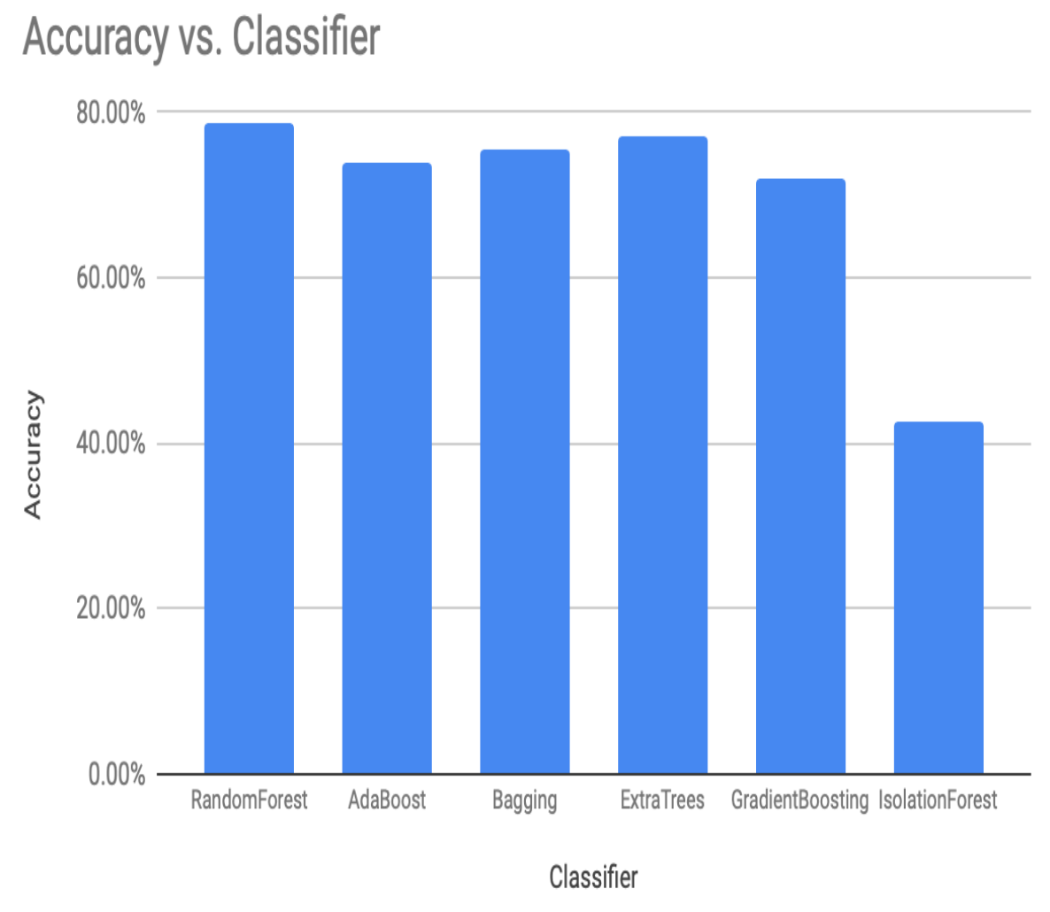


```
0.8741935483870967
```

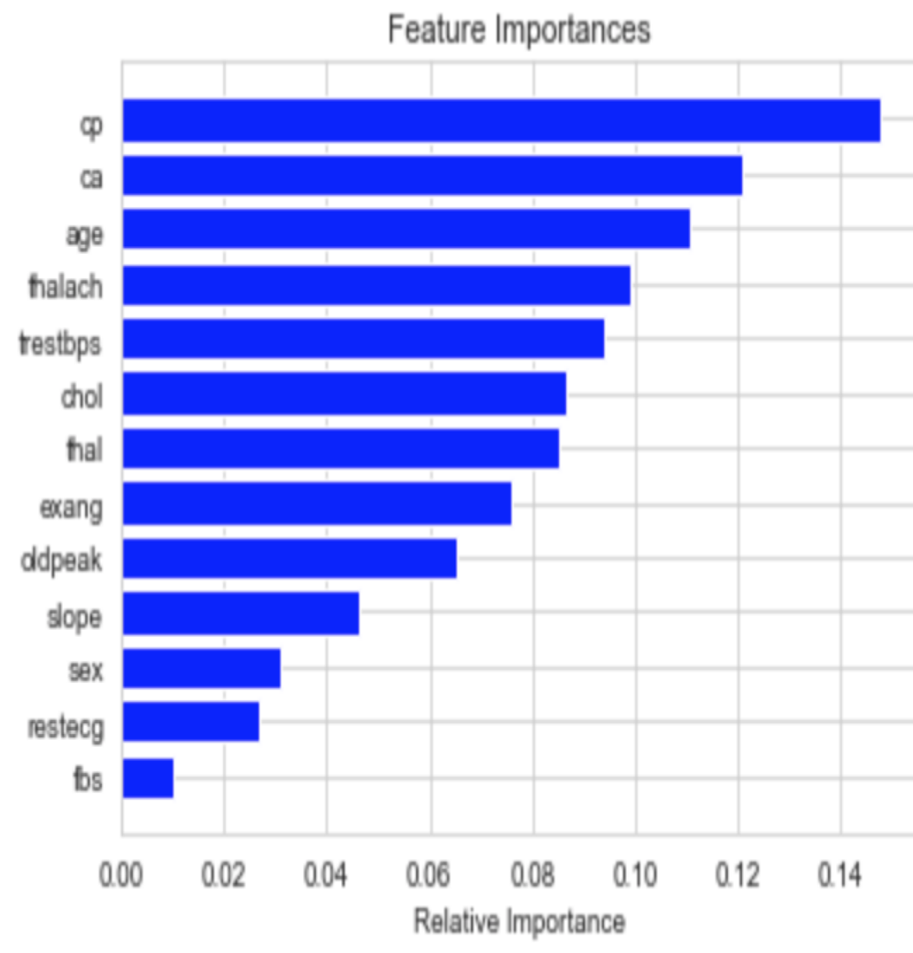
- Confusion matrices were created for the 4 algorithms, SVM,RF, NB, and KNN. RF is shown below



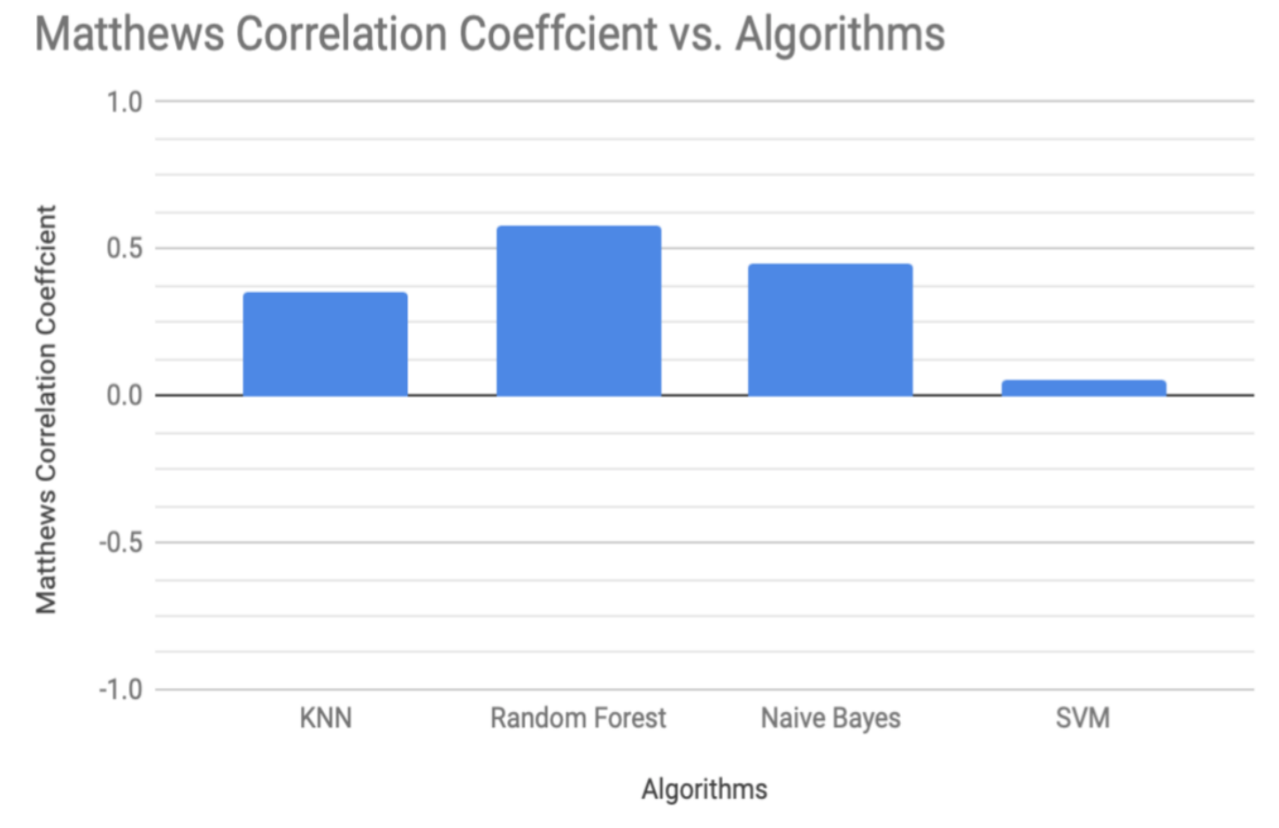
- All the ensemble classification methods in scikit learn were also tested and evaluated via accuracy score.

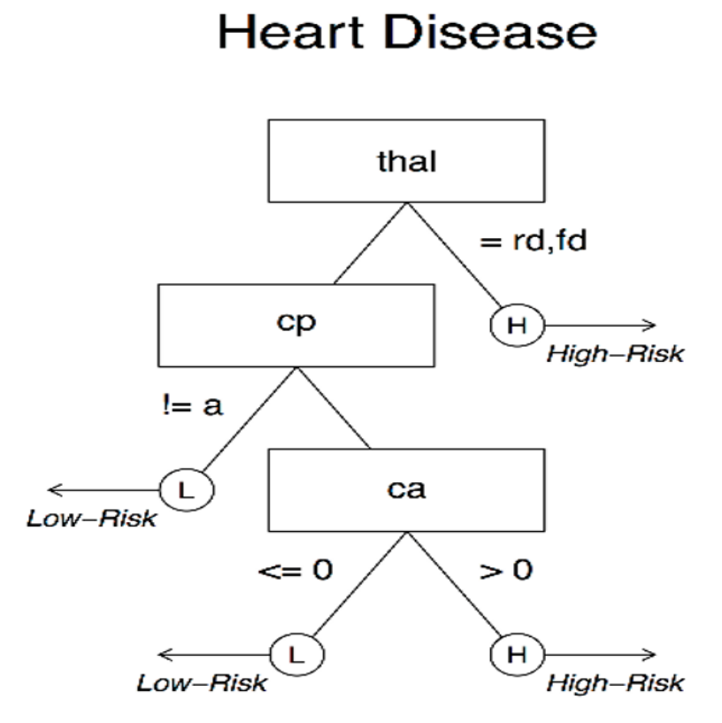


- A feature importance graph was created for the ML model, displaying the most important attributes in order, as shown below in figure



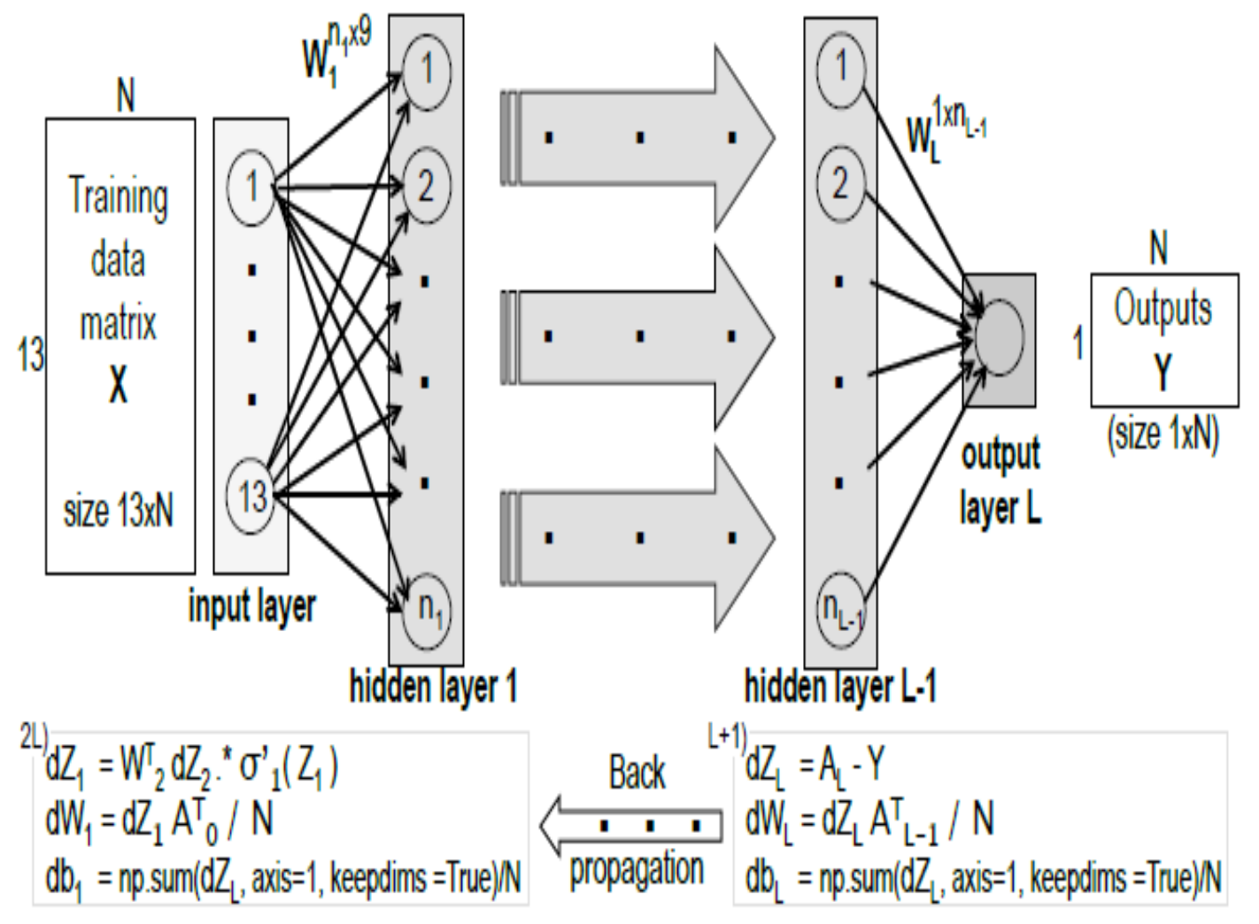
- A numerical way to view confusion matrices, as shown below through the Matthews Correlation Coefficient (MCC). MCC is a correlation coefficient between target and predictions.





Depiction of Random Forest Decision Tree

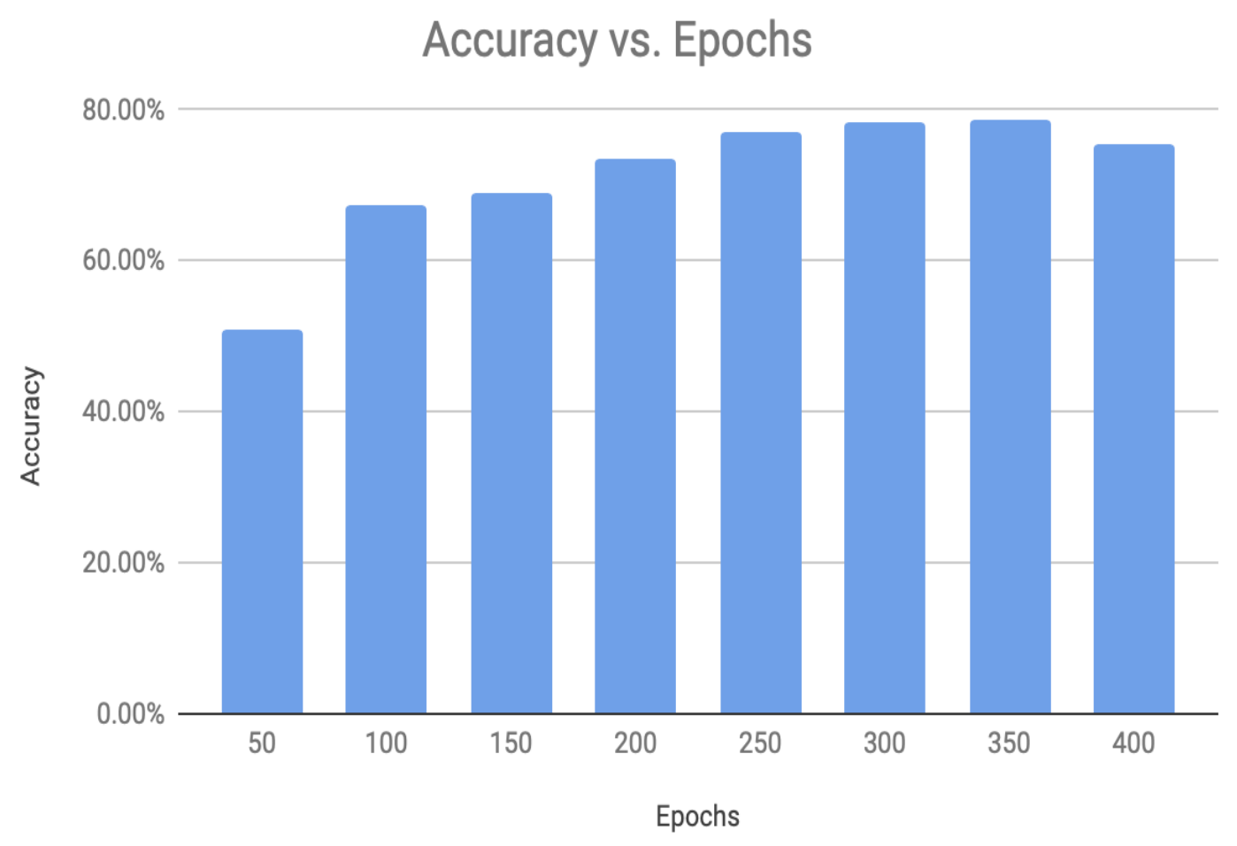
DEVELOPMENT / DEEP NEURAL NETWORKS



Depiction of Deep Neural Network Architecture

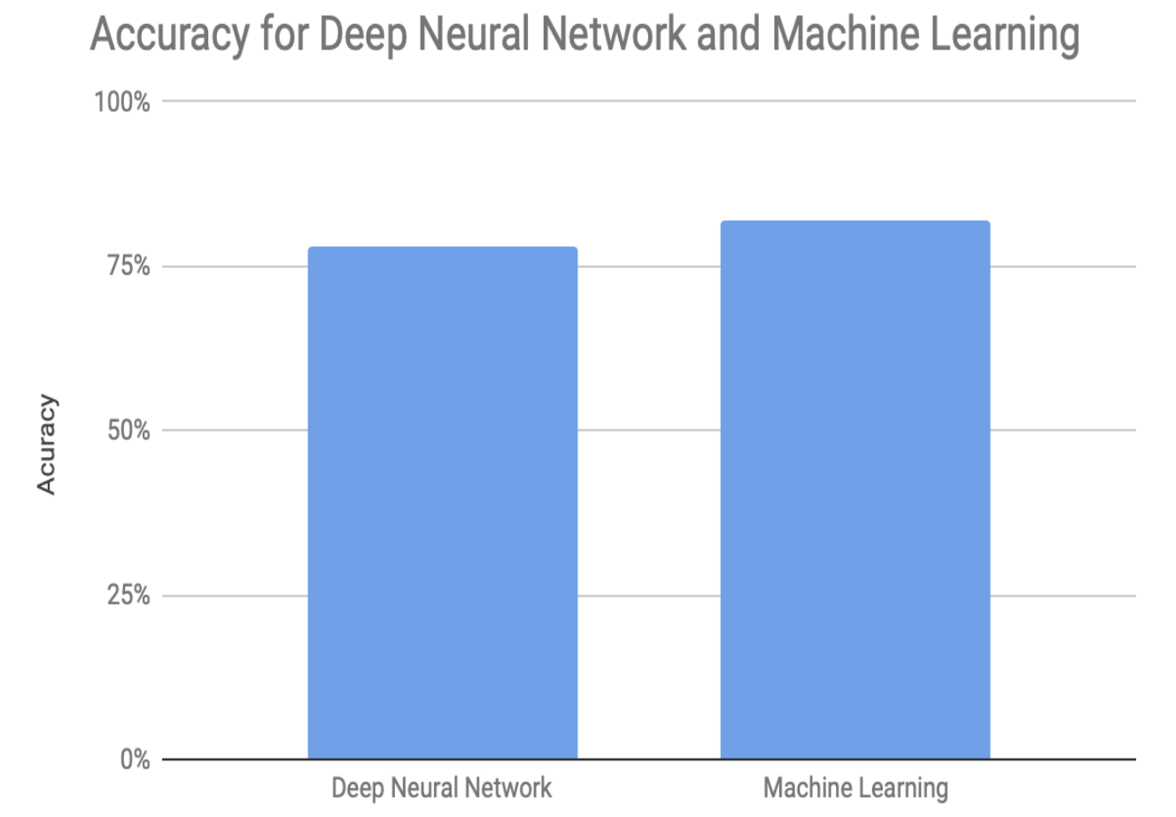
- Input data = 13 features
- 2 Hidden Layers
- Hidden layer 1= 8 Neurons
- Hidden layer 2 = 5 Neurons

- To better meet the performance criteria, changes to the epochs were made to the DNN and, as tested, up till a certain period, it resulted in much higher accuracy rates of the neural network.



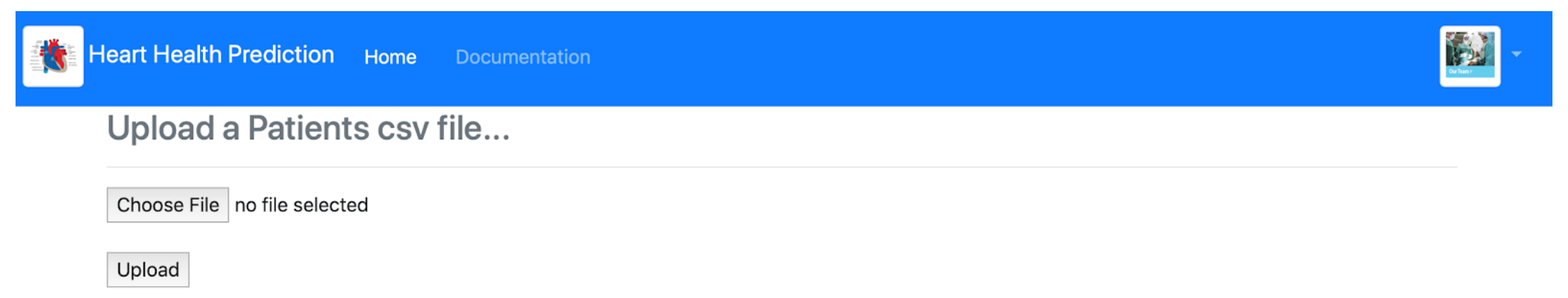
RESULTS

- Machine learning model resulted in an 81.79% accuracy and though K-fold cross validation, a mean accuracy of 82.13%. Thus it can be concluded that ML algorithm has accuracy rate of roughly 82%, as the values are very similar.
- DNN is trained with 350 epochs with a batch size of 8. The activations "relu" and "sigmoid"' were used for the layers of the DNN and resulted in 78.69% accuracy.



APPLICATION

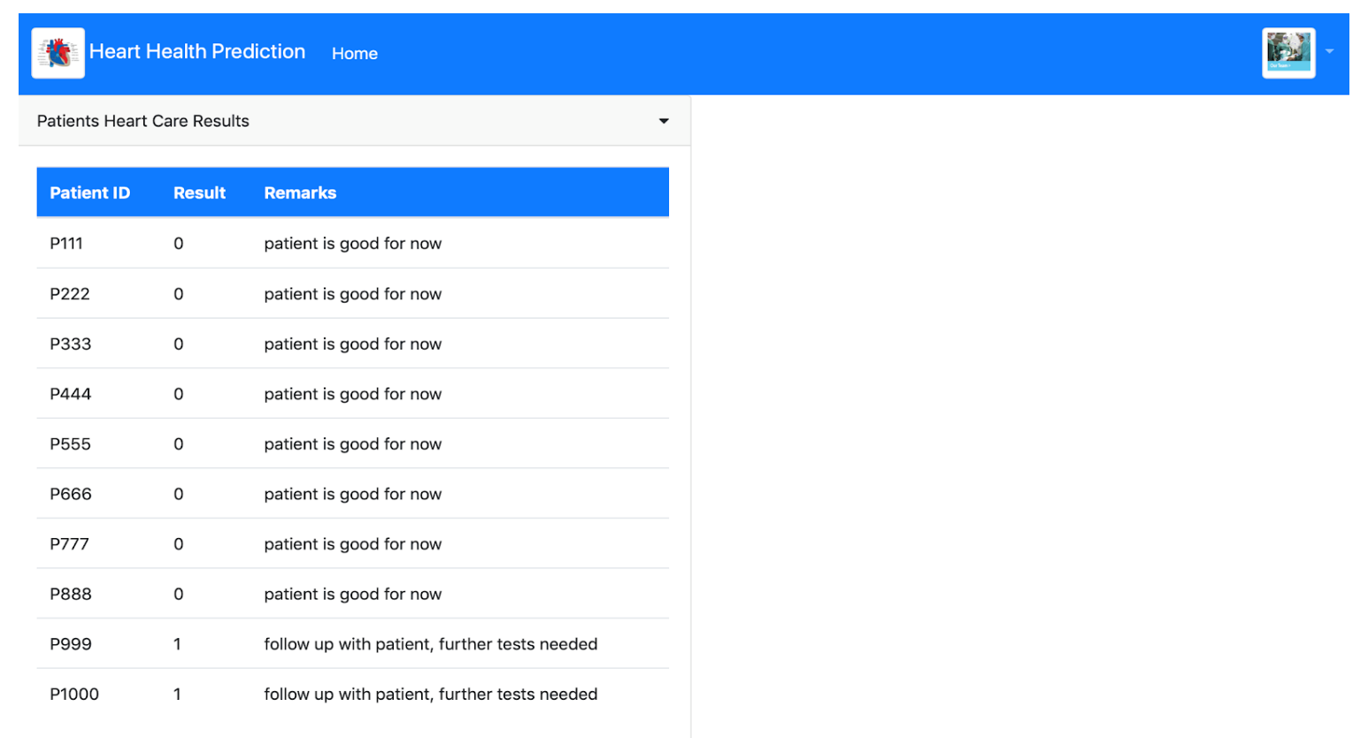
- The application is developed for medical professionals, by allowing them to upload patient data files as is in the format of a csv, and then getting immediate results on the diagnosis for each patient.
- First, the model was exported using the scikit learn machine learning API. The application was created using Python Flask REST API, and the user interface was developed using Bootstrap.
- When uploading the CSV file of patient data, it's put into a location defined in the program. Once saved to the correct location, the CSV file is passed to the ML model for diagnosis of the heart disease.
- Additionally, the patient ID column was initially removed as it's not needed for the diagnosis through the model, and at the end was later appended when displaying the results.



Upload a Patients csv file...

no file selected

Depiction of Input Screen



Patients Heart Care Results

Patient ID	Result	Remarks
P111	0	patient is good for now
P222	0	patient is good for now
P333	0	patient is good for now
P444	0	patient is good for now
P555	0	patient is good for now
P666	0	patient is good for now
P777	0	patient is good for now
P888	0	patient is good for now
P999	1	follow up with patient, further tests needed
P1000	1	follow up with patient, further tests needed

Depiction of output screen

CONCLUSION

- DNNs are known to be more successful than ML algorithm, however throughout the entirety of this project the ML model has always performed higher than the DNN, making the initial prediction incorrect.
- This is because neural networks require vast amounts of data to be successful in its learning or training. Due to the limitation of publicly available heart disease sets, in this case, there wasn't as much data, therefore why the DNN preformed lower than the ML model.

FUTURE STEPS

- Creating synthetic data to create more successful models to increase the data for training, as Cleveland dataset was not enough data to train a completely successful DNN
- Trying other ML algorithms and DNN models to determine which truly performs the best for this particular research topic. More specifically trying the Naïve-Bayes model, as during testing, Naive-Bayes performed very close in level when compared with Random Forest.

REFERENCES

- (n.d.). Retrieved January 5, 2019, from http://archive.ics.uci.edu/ml/datasets/Heart Disease
- (2016, August 30). Retrieved January 10, 2019, from https://www.nhs.uk/news/heart-and-lungs/one-in-three-heart-attack-cases-misdiagnosed/
- Anaconda. (n.d.). Retrieved January 1, 2019, from https://www.anaconda.com
- A Beginner's Guide to Neural Networks and Deep Learning. (n.d.). Retrieved January 12, 2019, from https://skymind.ai/wiki/neural-network