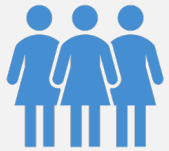
A Novel Approach to the Diagnosis of Heart Disease using Machine Learning and Deep Neural Networks

Sahithi Ankireddy

James B. Conant High School, Hoffman Estates, IL

PROBLEM

610,000 people die from heart disease every year, 1 in 4 deaths



1 out 3 heart disease cases are misdiagnosed

Common diagnostic procedures can be dangerous

The use of Machine Learning (ML) techniques and Deep Neural Networks (DNN) can mitigate the possibility of human error while increasing prediction accuracy rates.

Develop accurate Machine Learning and Deep Neural Network algorithms to create a application for assisted heart disease diagnosis. The ML model will be compared with the DNN and the most accurate model is used for the application.

SOLUTION

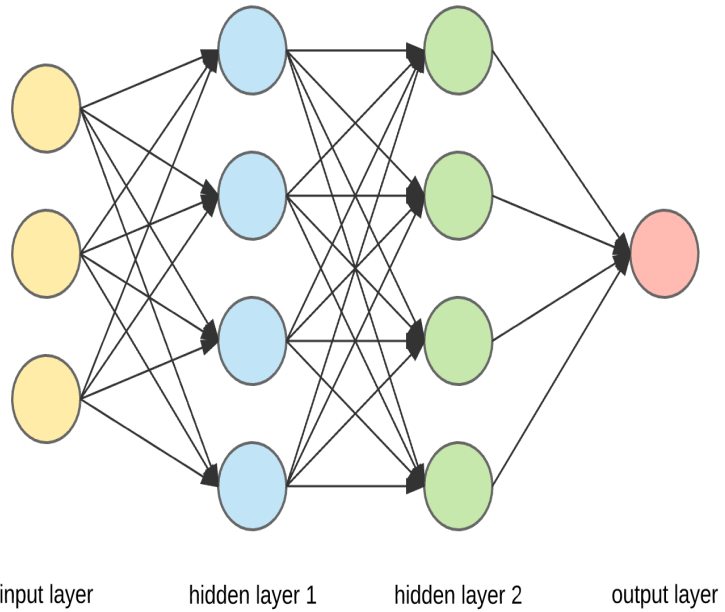
BACKGROUND

Machine Learning is a general data analysis technique that uses computational models and methods to "learn" information directly from data without using a preset formula or rule based programming.

A subset of machine learning includes deep learning or the training of a deep neural network. Deep Neural Networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns.

Data enters the input layer, it's processed in the hidden layers through a mathematical function (activation function), and predicted in output layer

Hyperparameter tuning: hyperparameters cannot be directly learned from the regular training process, they are fixed before training begins, thus this method finds the most effective combination of hyperparameters for the model.



DATA SET

- 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

OVERVIEW OF PROJECT

Dataset

RF | KNN | NB | SVM | DNN

Comparison ML and DNN algorithms

Algorithm with higher accuracy prediction rate

Input File

Tool

Patient / Doctor input

Patient ID	Result	Remarks
P111	0	patient is good for now
P222	0	patient is good for now
P999	1	follow up with patient, further tests needed
P1000	1	follow up with patient, further tests needed

METHODS FOR MACHINE LEARNING MODELS

1. Split dataset into 2/3 training and 1/3 testing + Rescaled data

2. Grid search optimization – hyperparameter tuning

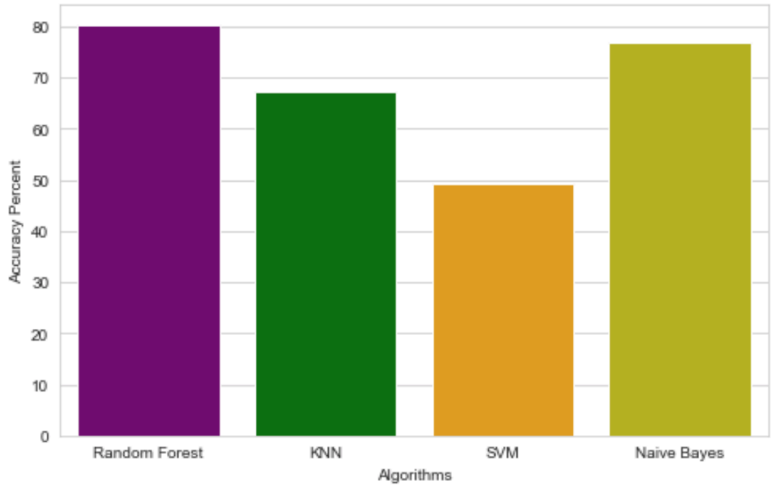
3. Trained KNN, NB, RF, SVM

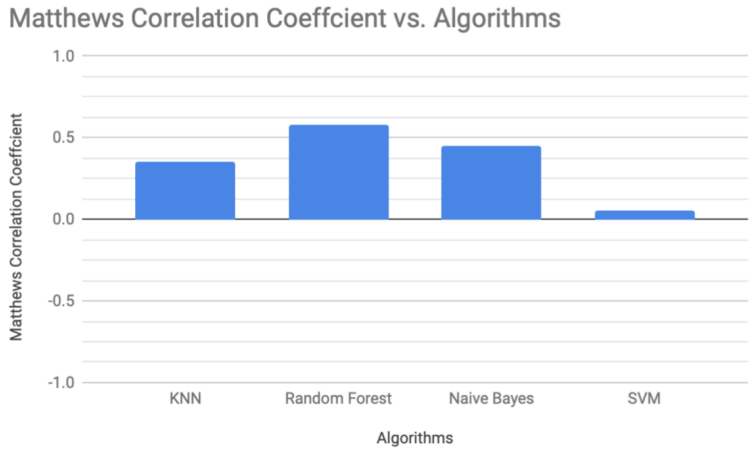
4. K Fold Cross Validation: divides and rearranges input test data

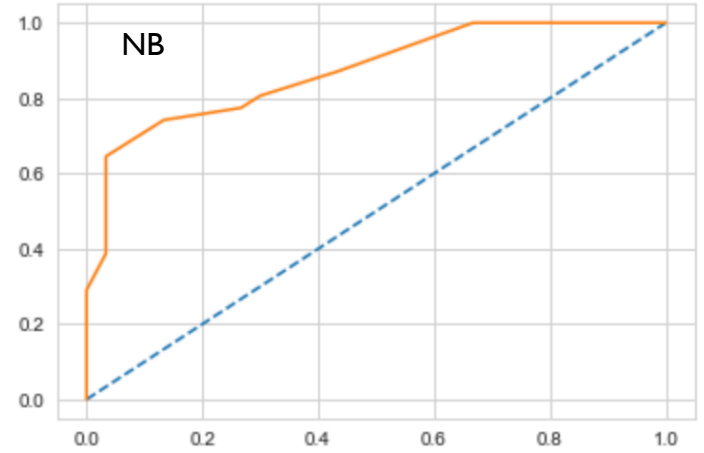
Matthews Correlation Coefficient: measures performance without data specific bias

ROC/AUC: probability curve and AUC measures degree of separability

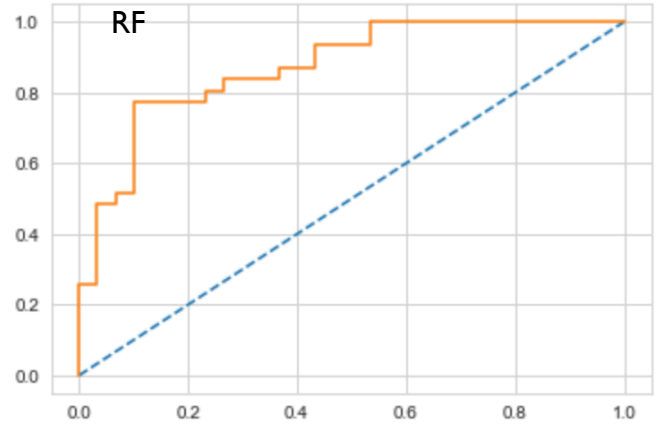
$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



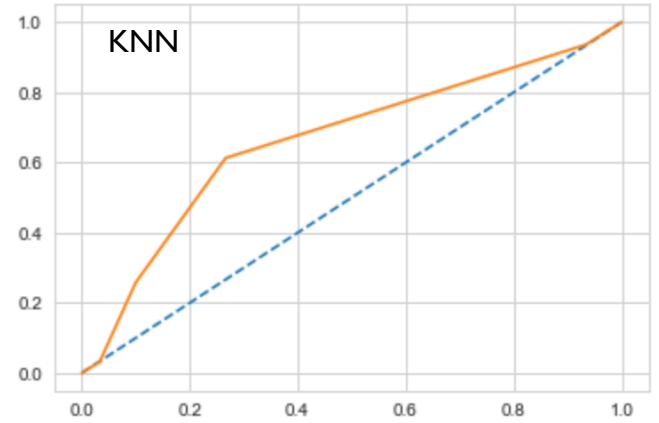




0.8715053763440861



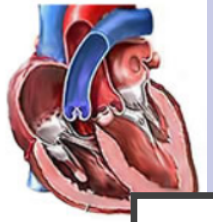
0.8741935483870967



0.6634408602150538

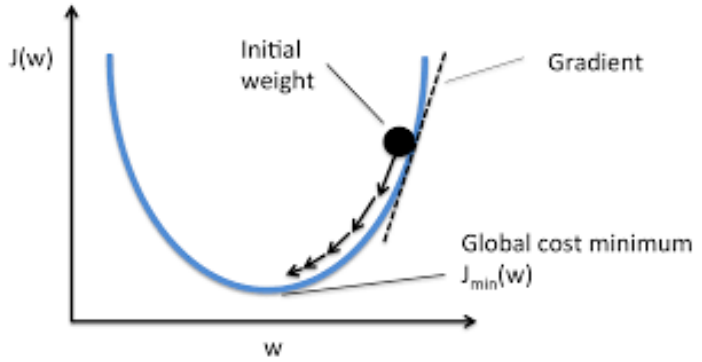
Best ML Model: Random Forest
83%

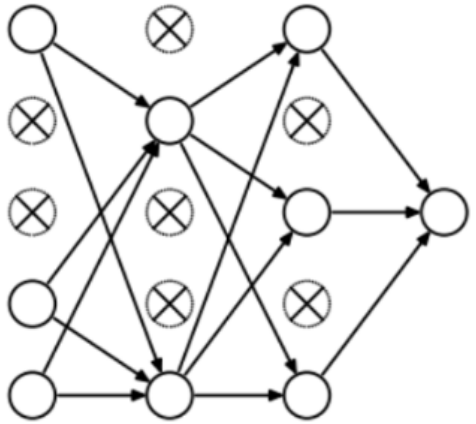
MACHINE LEARNING ALGORITHMS RESULTS



DEEP NEURAL NETWORK

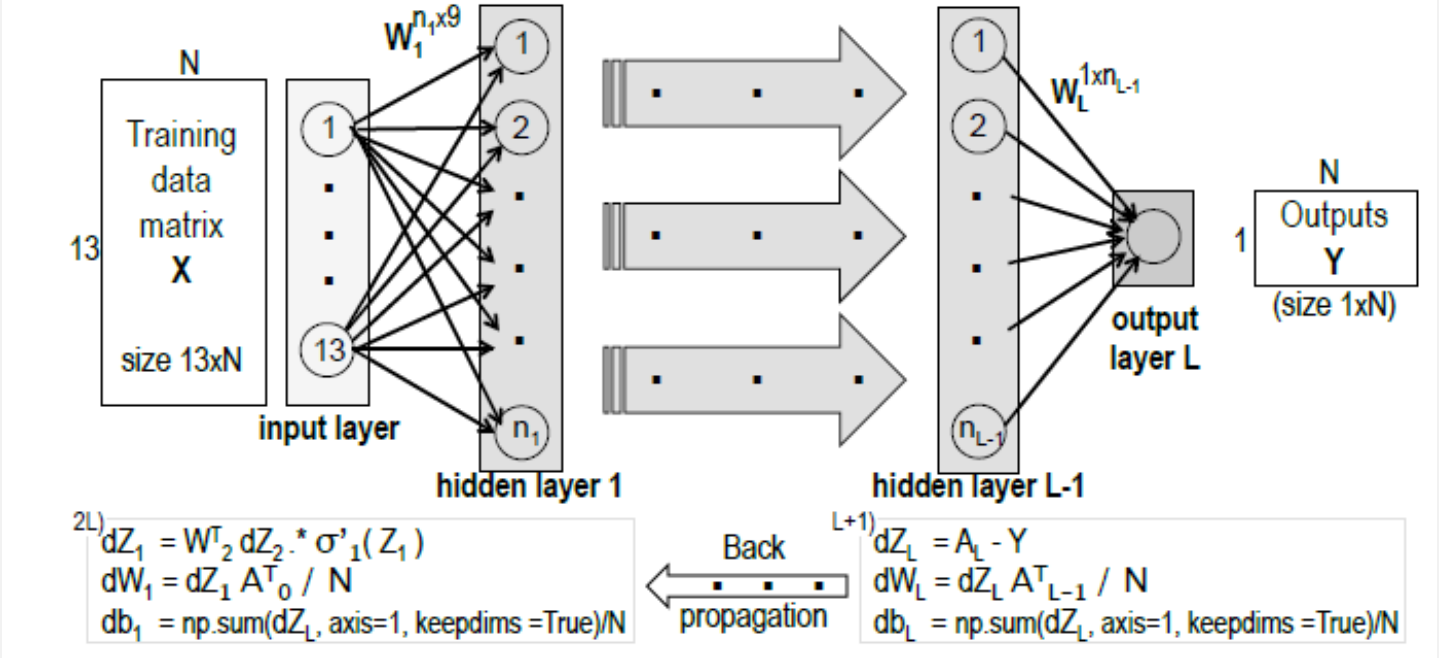
- Split dataset into 2/3 training and 1/3 testing
- Gradient Descent Optimization: repeatedly finds parameters and each iteration reduces cost function
- Dropout Regularization: randomly kills certain neurons at each pass, forces network to learn different paths
- Evaluation through Kfold Cross Validation and Mathews Correlation Coefficient





DEEP NEURAL NETWORK RESULTS

- Input data = 13 features
- 350 Epochs
- Batch Size: 8
- 7 Layer Network
- 94% accuracy after K-Fold Cross Validation
- MCC score of 0.91



COMPARISON OF ML AND DNN

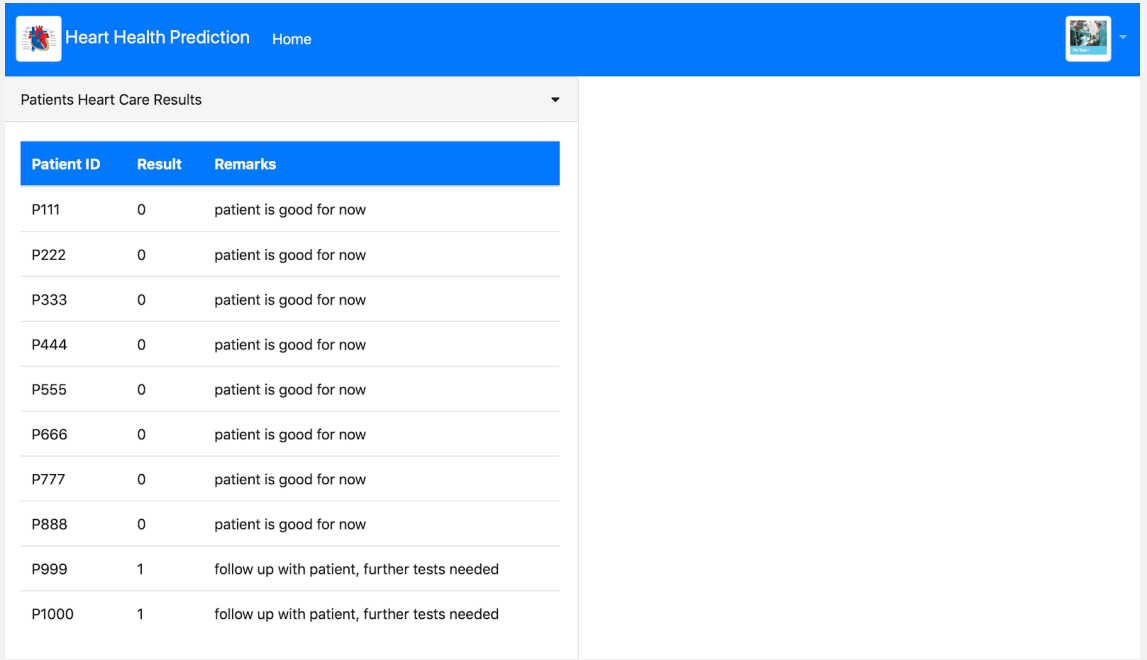
ML

- 83 percent accuracy after Kfold Cross validation
- 0.6 MCC Score

DNN

- 94 percent accuracy after Kfold Cross Validation
- 0.91 MCC Score

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	PID
42	1	2	120	295	0	0	162	0	0	1	0	3	P111
41	1	2	110	235	0	0	153	0	0	1	0	3	P222
41	0	2	126	306	0	0	163	0	0	1	0	3	P333
49	0	4	130	269	0	0	163	0	0	1	0	3	P444
61	1	1	134	234	0	0	145	0	2.6	2	2	3	P555
60	0	3	120	178	1	0	96	0	0	1	0	3	P666
67	1	4	120	237	0	0	71	0	1	2	0	3	P777
58	1	4	100	234	0	0	156	0	0.1	1	1	7	P888
47	1	4	110	275	0	2	118	1	1	2	1	3	P999
52	1	4	125	212	0	0	168	0	1	1	2	7	P1000



HEART DISEASE DIAGNOSIS TOOL

- DNN performed better than Machine Learning model with 94% accuracy (model used for tool)
- Web-Application using Python Flask
- Doctor input CSV file from lab test with patient information
- File sent into application and results outputted for each patient

FUTURE STEPS

Improving DNN to 96% or above accuracy rate by getting more data

Working on IOS app

Working on connecting with hospitals to get application tested by cardiologists