Mobile Machine Learning Model for Cardiovascular Disease Prediction

Kunal Singh Lohiya - 191IT128
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Emai: kunalsingh.191it128@nitk.edu.in

Atul Kumar Singh - 191IT106
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Emai: aksxy.191it106@nitk.edu.in

Maheshwari Mihir Premjibhai - 191IT129
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Emai: mihirpm.191it129@nitk.edu.in

***Abstract*—This paper is about Mobile Machine Learning Model for Cardiovascular Disease Prediction, a system designed for mobile devices that facilitates prediction of cardiovascular disease(CVD). This Model is deployed in as android application which is easily accessible to laymen. It accepts information about eleven features which are considered for creating this model. Instead of consulting health care professionals, this system analyse the user feed and predict using the Mobile Machine Learning Model deployed in as an android application. This model classifies a patient as 'continued risk' or 'no risk' for cardiovascular disease (CVD). The model is trained using dataset of 70,000 patients. This type of Application is useful in the current time of global pandemic where there is risk on visiting hospital for the people having heart disease and low immunity and if this application predicts 'continued risk' then one can consult physicain and could get the medical support at the early stage.**

***Keywords*—Health Care, Cardiovascular Disease, Mobile Machine Learning model, TensorFlow Lite**

I. Introduction

In today's fast growing world people are ignoring their health. Especially in the urban areas, the reason being the fast lifestyle, the availability of hospitals nearby and the cost of routine checkups.

"India has one of the highest burdens of cardiovascular disease (CVD) worldwide. The annual number of deaths from CVD in India is projected to rise from 2.26 million (1990) to 4.77 million (2020). Coronary heart disease prevalence rates in India have been estimated over the past several decades and have ranged from 1.6% to 7.4% in rural populations and from 1% to 13.2% in urban populations." [7]

We need to eliminate some of these factors, and give people a handy way to keep track of their health. We propose a mobile machine learning model to give them updates on their health risks within seconds.We have used a dataset with datas of nearly 70,000 people to give predict results. [6]

The dataset contains all the factors require to predict accurately if one has any heart related disease. All the factors in the dataset are such that they can be easliy avaliable at house, like, weight, height, blood pressure, glucose level,etc. Cost of Digital Blood Pressure and Digital Glucose level measurement machine is around thousand rupees in india. Also it is said that cholestrol should be checked every four to six years for good health. Moreover, people who have heart disease or diabetes or who have a family history of high cholesterol, need to get their cholesterol checked more often.

II. Literature Survey

[1]In the proposed research paper data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values and classification of attributes for prediction and decision making at different levels. The performance is obtain by classification, accuracy, sensitivity and specificity analysis. This project paper proposes a prediction model to predict whether a people have a heart disease or not and to provide an awareness for health. This is done by comparing the accuracy of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

[2] In this paper they present Mobile Machine Learning Model for Monitoring Cardiovascular Disease, a system designed specifically for mobile devices that facilitates monitoring of cardiovascular disease. This model uses wearable sensors to collect observable trends of vital signs contextualized with data from clinical databases. this model pridect patient as "continued risk" or "no longer at risk" for CVD. Using a SVM the system monitors features extracted from wearable sensors and clinical databases. main components of this system architecture are Input, Data Processing, Machine Learning, Decision Making and Output. Data Processing, Machine Learning and Decision Making are all performed on the mobile device

[3] This paper presents the overview of machine learning techniques in classification of diabetes and cardiovascular diseases using Artificial Neural Networks and Bayesian Networks .The most commonly used type of ANN is multilayer feedforward neural network with Levenberg-Marquardt learning algorithm.and the most commonly used type of BN is NaIve Bayesian network which shown the highest accuracy values for classification of diabetes and CVD and the calculation of mean accuracy of observed networks has shown better results using ANN.

[4] In this paper Data mining techniques used. The outcomes of this system provide the chances of having heart disease in terms of percentage. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using two main Machine Learning Algorithm namely Decision Tree Algorithm and Naive Bayes Algorithm which shows the best algorithm among these two in terms of accuracy level of heart disease.

[5] In the proposed paper,systems like the expulsion of noise data, evacuation of missing information, filling default esteems if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model was obtained by using method by classification, accuracy, sensitivity and specificity analysis. this model Support Vector Machine, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

III. Problem Statement

Implementing a mobile machine learning model to detect cardiovascular diseases in individuals based on certain features and factors.

A. Objectives

- Making a model which predicts accurately & training on dataset.
- Converting Model to mobile compatible and achieve low latency model.
- Deploying this mobile machine learning model in an android application which can be made easily accessible to users.

IV. Proposed Work

Basically A Neural Network Model is implemented using Keras Sequential model with 2 hidden layers(dense with activation ReLU (Rectified Linear Unit max(x,0)) and dense layer with activation Sigmoid $(1/(1+e^x))$)). This model is converted to Mobile Machine Learning Model using Tensorflow developer tool called TensorFlow Lite. This Mobile Machine Learning model is used in Android Studio to make an Android Mobile Application, which is easily accessible to the users.

A. Dataset Processing

There is a need to pre-process the dataset to make it suitable for model training and converting it to mobile machine learning model.[6] Kaggle Dataset — "Cardiovascular Disease Dataset" is used to train model, Initially dataset has 13 attributes out of which target attribute is the output(0 — Absence of CVD Disease and 1 — Presence of CVD Disease) and one is id which is not useful for prediction and is dropped from dataset.So, size of dataset is 70,000 i.e. data of 70K people, with 12 features in each data row. Pandas python library is used to read, and process the dataset.

1) Following are the attribues in the Dataset and the processing applied to the dataset:

- **Age** in number of days (int)
  Age is converted to number of years by dividing year by 365 days
- **Gender** in 0-men, 1-women
  Gender was 1-Women and 2-Men which is converted to binary zeros and ones by getting remainder on dividing by 2
- **Height** int cm (int)
- **Weight** in kg (float)
- **ap_hi** (int)
  Systolic Blood Pressure(in mmHg), In dataset there were few negative values, so such erroneous data were removed
- **ap_lo** (int)
  Diastolic Blood Pressure(in mmHg), In dataset there were few negative values, so such erroneous data were removed
- **cholestrol**
  Cholestrol Level: 1 - normal, 2 - above normal, 3 - well above normal
- **gluc**
  Glucose Level: 1 - normal, 2 - above normal, 3 - well above normal
- **smoke**
  Whether one Smokes or not, 0: NO, 1: Yes
- **alco**
  Whether one Consumes Alcohol, 0: NO, 1: Yes
- **Active**(exercise): 0: NO, 1: Yes
- **cardio:**
  whether one is Heart diseased or not 0: Safe, 1: Danger

Thus after preprocessing dataset, size of dataset is 69992 (as 8 erroneous data were removed) with 12 features. target column is removed and saved into another array for calculating accuracy. Hence, there are 11 Input attributes and 1 output, which is final prediction value(0 or 1).

B. Splitting Dataset

Dataset is splitted into Training and Test sets in ratio of 7:3 (70% Training data & 30% Test data) using train_test_split method in Scikit-learn python library. Now, the size of Train data is 48,994 and Test data is 20,998.

Sr.No	Features	Values	Type
1	Age	Years	INTEGER
2	Gender	0-Men,1-woman	INTEGER
3	Height	Cm	INTEGER
4	Weight	Kg	FLOAT
5	Systolic BP (ap_hi)	mm Hg	INTEGER
6	Diastolic BP (ap_lo)	mm Hg	INTEGER
7	Cholestrol	1-Normal,2-Above Normal,3-Well above normal	INTEGER
8	Glucose	1-Normal,2-Above Normal,3-Well above normal	INTEGER
9	Smoke	0-No,1-Yes	BINARY
10	Alcohol	0-No,1-Yes	BINARY
11	Physical Activity	0-No,1-Yes	BINARY
12	Target	0-No,1-Yes	BINARY

Fig. 1: Dataset Description

C. Training the model

A Problem in machine learning, where the categories are predefined, and is used to categorize new probabilistic observations into two categories is termed as Statistical **Binary Classification**. This is **Supervised learning** method for machine learning.
Cardiovascular Disease Prediction can also be called **Statistical Binary Classification** Problem. As the prediction made is into two classes namely "NO risk for cardiovascular disease" & "Continued risk for cardiovascular disease". The Standard Machine learning Model for binary classification are implemented using Scikit-learn library's algorithms and a Neural Network Model is implemented using Tensorflow Keras. The accuracy scores of models are compared and the best fit model among all is used.

D. Machine Learning Algorithms for Binary Classification

Following Standard Binary Classification Models are implemented using the Scikit-learn library's algorithms.

1) ***Logistic Regression****:* Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.The dependent variable is binary in nature having data coded as either 1 or 0 . Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc

2) ***Naive Bayes****:* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem . Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$=Probability of hypothesis A on the observed event B

$P(B|A)$=Probability of the evidence given that the probability of a hypothesis is true.

P(A)=Probability of hypothesis before observing the evidence.

P(B)=Probability of Evidence.

3) ***SVM(Support Vector Machine)****:* SVM technique plots a hyperplane for every attributes which was present in the dataset. Classification is performed by identifying the hyperplane that divides one class with the other class. It builds a model which assigns new example to the other, making it a non-probabilistic binary linear classifier.

4) ***K Nearest Neighbour****:* K-NN algorithm compares the new data and available data and puts the new data into a similar available category . K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

5) ***Decision Tree****:* Decision Tree algorithm is a supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes.The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

6) ***Random Forest****:* Random forest is a machine learning algorithm which is used for classification and regression. It creates decision trees for each attribute. It corrects the overfitting to their training set. It also avoids the missing values, outliers by data analysis, data pre-processing. It is a kind of machine learning method where the weak models are combined to form a dynamic model.

7) ***Neural Network****:* For Creating Mobile Machine Learning model we need a model with low-latency and a small binary size which is provied by Tensorflow Lite model. Thus A Neural Network Model is made using Keras.

A Keras Sequential model with 2 hidden layers is trained on the dataset.

- Input to this model is 11 attributes of the dataset.

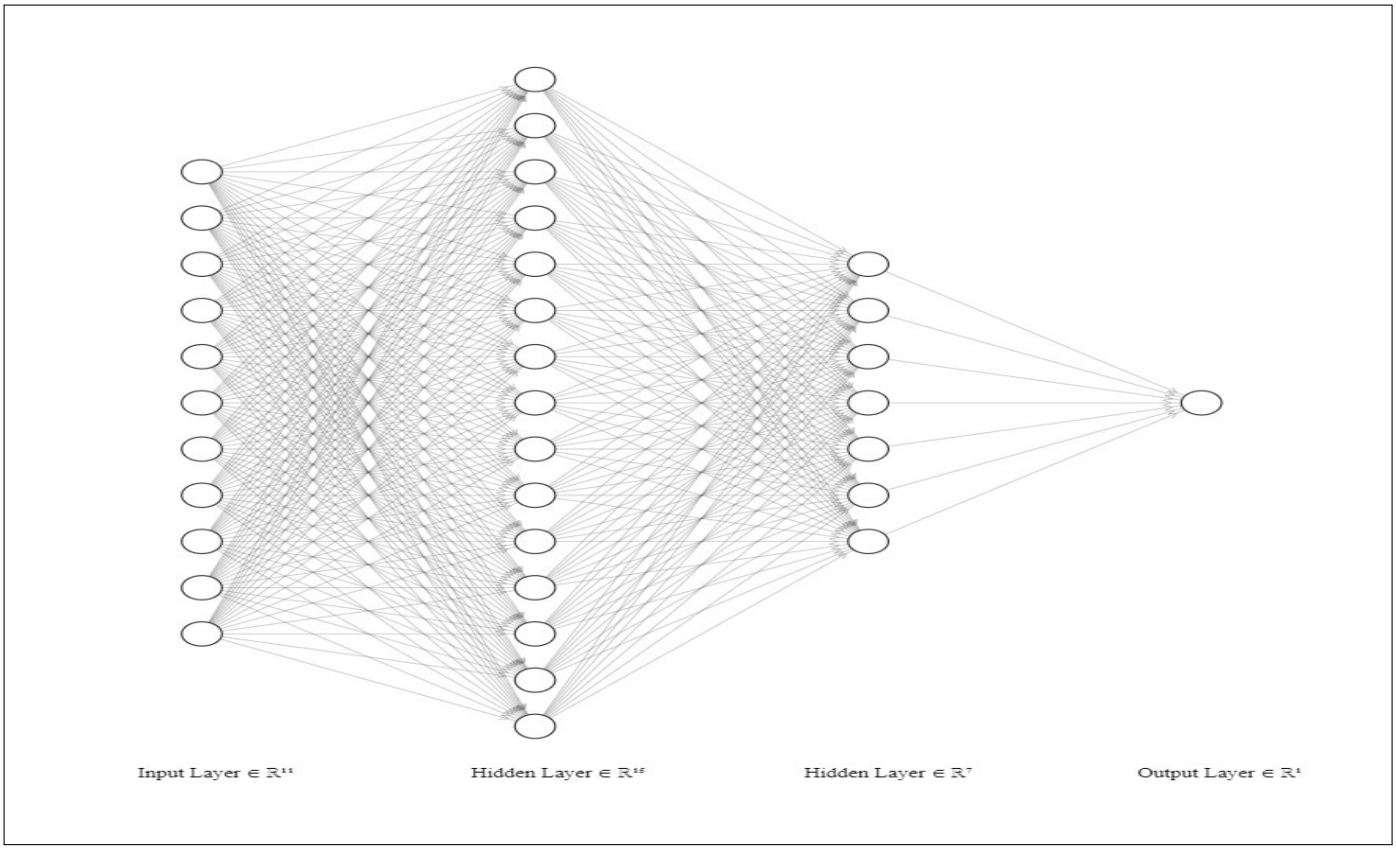

Fig. 2: Neural Network Architecture

- There are 2 hidden Layers in the model
  - Firstly a Dense Layer with activation ReLU (relu(x) = max(0,x)) and output dimension 15
  - Then a Dense Layer with activation ReLU and output dimension 7.
  - Finally a Dense Layer with activation Sigmoid(to get output in range of 0 to 1) and output dimension 1.
- Model Compilation with 'binary_crossentropy' loss function which is best for Binary classification and 'adam' optimizer.
- Fitting the model with 300 epochs and batch size of 22.
- Predicting on Test data and find accuracy score using accuracy_score from Scikit-learn library's. Accuracy on Test data is 72.72%. which is better the other standard binary classification Algorithms.

E. Converting Keras Model to Mobile Machine Learning Model

Keras Neural Network Model is converted to Mobile Machine Learning Model using Tensorflow developer tool called TensorFlow Lite.

TensorFlow Lite: Tensorflow provides set of developer tools called TensorFlow Lite which helps running Tensorflow models on mobile, IOT and embedded devices. It makes possible for machine learning model to run on device with low latency and effeciently.

F. Creating Android Application

To made this Model available to the laymen, there is a need to have GUI and Android Mobile Application is the best option. A Mobile Machine Learning Model is achieved using the TensorFlow Lite converter which converts Keras Model to tensorflow lite model and is saved as model.tflite file. This model file can be used in Android Studio to create an android application, which can be easily accessible to the users.

V. Experimentation and Results

Comparing the Accuracy Score of the Binary Classification Machine Learning Model, which are calculated using accuracy_score from Scikit-learn library.
Following is the Accuracy Score comparison Graph of all the Algorithms

S.NO	Algorithm	Accuracy
1	Logisitic Regression	71.18 %
2	Naive Bayes	59.00 %
3	Support vector machine	71.96 %
4	K Nearest Neighbor	71.35 %
5	Decision Tree	63.35 %
6	Random Forest Classifier	70.63 %
7	Neural Network	72.72 %

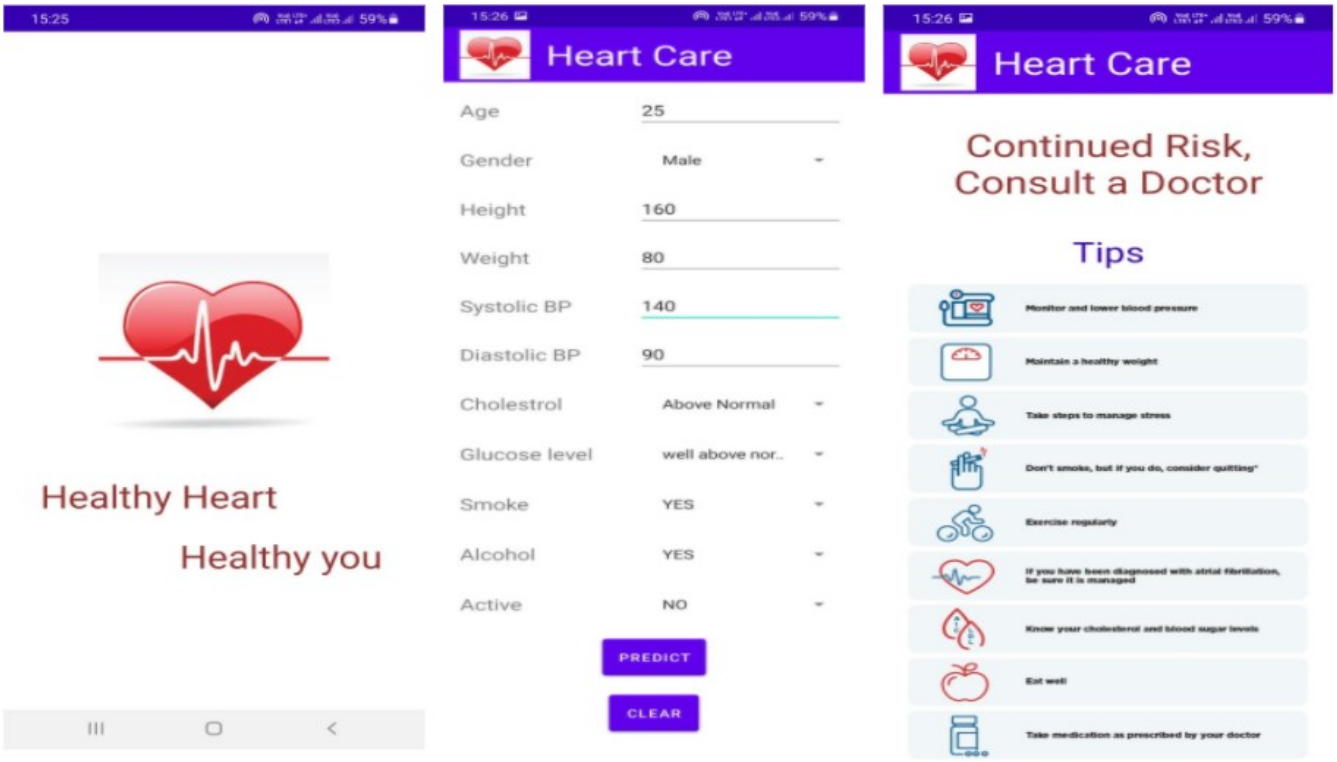

Fig. 3: Final Outcome

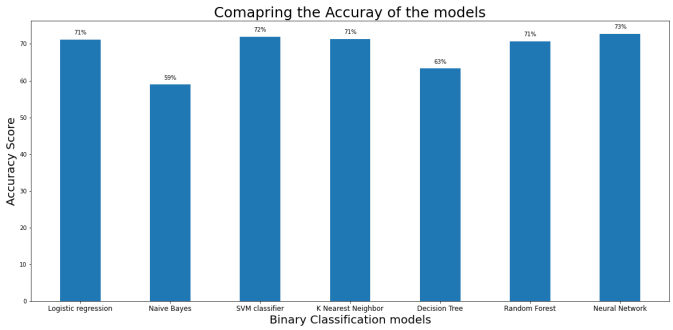

Fig. 4: Comparison of Accuracy

VI. Conclusion

Thus, we have achieved a fine working mobile machine learning model which predicts quite accurately. But the UI of the android can be made better and indulging for real world use. We would also like to add embedded senors to predict the real time ECG and Blood Pressure of the patients to have more realistic results.

Acknowledgment

We thank our mentors Dr. Anand Kumar M. and Ms. Trupti Chandak to guide us through the projects and giving the insights of the project.We sincerly thank you for giving the right guidance and an opportunity to work on a project which gives hands on experience with the real life problems.

References

[1] https://ieeexplore.ieee.org/abstract/document/8550857
[2] https://www.sciencedirect.com/science/article/pii/S1877050915024928
[3] https://ieeexplore.ieee.org/abstract/document/7977152
[4] https://ieeexplore.ieee.org/abstract/document/8741465
[5] https://www.ijeat.org/wp-content/uploads/papers/v9i3/B3986129219.pdf
[6] Kaggle Dataset—Cardiovasculare Dataset, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio$_train.csv$
[7] https://www.ncbi.nlm.nih.gov/