SPOT THE DISEASE : ANALYSIS AND PREDICTION OF MYOCARDIAL INFRACTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology Degree in
Information Technology
By

ASWIN J RICHARDS (37120011)
HARISH M (37120028)

DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI
SALAI, CHENNAI - 600 119

MARCH - 2021

SATHYABAMA UNIVERSITY

(Established under Section 3 of UGC Act, 1956)

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600119

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of who carried **ASWIN J RICHARDS (37120011)** and **HARISH M (37120028)** out the project entitled **SPOT THE DISEASE: ANALYSIS AND PREDICTION OF MYOCARDIAL INFRACTION USING MACHINE LEARNING** under my supervision from November 2020 to April 2021.

Internal Guide

Dr. SENDURU SRINIVASULU M.TECH., Ph.D.,

Head of the Department

Dr. R.SUBHASHINI M.E., Ph.D.,

Submitted for Viva voce Examination held on______________________

Internal Examiner **External Examiner**

DECLARATION

We, **ASWIN J RICHARDS** and **HARISH M** hereby declare that the Project Report entitled **ANALYSIS AND PREDICTION OF MYOCARDIAL INFRACTION USING MACHINE LEARNING** done by us under the guidance of **Dr.M.SENDURU SRINIVASULU .M.TECH.Ph.D.** (Internal) is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in **INFORMATION TECHNOLOGY**.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.SASIKALA M.E., Ph.D**, **Dean**, School of Computing, **Dr.SUBHASHINI M.E., Ph.D.,** Head of the Department of Information and Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide, **Dr. M.SENDURU SRINIVASULU M.TECH., Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

A myocardial infarction is an area of necrotic tissue in the heart resulting from a blockage or narrowing in the arteries supplying blood and oxygen to the heart. The restricted oxygen due to the restricted blood supply causes an ischemic stroke that can result in an infarction if the blood flow is not restored within a relatively short period of time. The blockage can be due to a thrombus, an embolus or an atheromatous stenosis of one or more arteries. Which arteries are problematic will determine which areas of the heart are affected (infarcted). These varying infarcts will produce different symptoms and outcomes. About one third will prove fatal.In the proposed system, Logistic regression, a machine learning regression algorithm is used to detect the myocardial infarction in a patient. The algorithm is trained using a structured medical dataset. By accurately detecting the myocardial infarction in a patient, this project is very significant in the medical field.

TABLE OF CONTENTS

	2.2.3 Yi Wang, Na Wang, Min Xu, Junxiong Yu, Chenchen Qin, Xiao Luo, Xin Yang, Tianfu Wang, Anhua Li, and Dong Ni*,” Deeply-Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound [Vol 0278-0062,2019]	15
	2.2.4 Jose M. Anton-Rodriguez , Peter Julyan, Ibrahim Djoukhadar, David Russell, D. Gareth Evans, Alan Jackson, and Julian C. Matthews,” Comparison of a Standard Resolution PET-CT Scanner With an HRRT Brain Scanner for Imaging Small Tumors Within the Head”, IEEE Transactions on radiation and plasma medical sciences, vol. 3, no. 4, july 2019	16
	2.2.6 Koyel Mandal, Rosy Sarmah, and Dhruba Kumar Bhattacharyya ”Biomarker Identification for Cancer Disease Using Biclustering Approach: An Empirical Study” [Vol: 1545-5963,2018]	18
	2.2.7 Julien Rouyer, Member, IEEE, Tony Cueva, Student Member, IEEE, Tamy Yamamoto, Alberto Portal, and Roberto Lavarello, Senior Member, IEEE,”In vivo Estimation of Attenuation and Backscatter Coefficients from Human Thyroids”,IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,Vol.0885-3010 (c) 2018.	19
	2.2.8 Koyel Mandal, Rosy Sarmah, and Dhruba Kumar Bhattacharyya, “Biomarker Identification for Cancer Disease Using Biclustering Approach: An Empirical Study”,	20

	IEEE/ACM Transactions on Computational Biology and Bioinformatics. [Vol: 2829931,2018]	
	2.2.9 Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao,"Multi-task Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition",IEEE Journal of Biomedical and Health Informatics,VOL. 14, NO. 8, AUGUST 2017.	21
	2.2.10 Xiangxiang Zheng ,Guodong Lv, Guoli Du, Zhengang Zhai, Jiaqing Mo and Xiaoyi Lv,"Rapid and Low-Cost Detection of Thyroid Dysfunction Using Raman Spectroscopy and an Improved Support Vector Machine",IEEE Photonics Journal, ol. 10, No. 6, December 2017.	22
	2.2.11 Nikhil S. Narayan, Pina Marziliano, Jeevendra Kanagalingam, MD and Christopher G.L. Hobbs, MD,"Speckle Patch Similarity for Echogenicity based Multi-Organ Segmentation in Ultrasound Images of the Thyroid Gland",IEEE Journal of Biomedical and Health Informatics,Vol.2168-2194,2016.	23
	2.2.12 Jennifer E. Rosen∗, Hyunsuk Suh, Nicholas J. Giordano, Ousama M. A'amar, Eladio Rodriguez-Diaz, Irving I. Bigio, and Stephanie L. Lee,"Preoperative Discrimination of Benign from Malignant Disease in Thyroid Nodules With Indeterminate Cytology Using Elastic Light-	24

CHAPTER -1

INTRODUCTION

1.1 GENERAL

Myocardial infarction is generalized ischemic stroke due to disturbance in the blood supply to myocardial infarction which can cause rapid loss of the heart function [1,2]. It can be divided into four periods according to duration of the illness, namely hyperacute phase, acute phase, subacute and chronic phase. Since magnetic resonance imaging (MRl) has the great advantage of imaging human soft tissue, radiologists employ T2 fluid attenuated inversion recovery (T2 FLAIR) technique, diffusion weighted imaging (DWI) and some other medical imaging techniques to determine in which phase myocardial infraction is Generally, myocardial infarction patients taken to hospital for clinic diagnosis are often in acute phase on account of rapid onset of infarction and rather short time of the hyperacute phase. In the hyperacute phase, cytotoxic heartedema appears around infarct and it causes mass effect in the following acute phase, which may increase the risk of disease progression. Therefore, treating cytotoxic heartedema in acute phase will improve the situation of blood supply and relieve the illness. If the size of edema can be estimated after separation of heartedema from infarct with the help of image segmentation technique, radiologist can judge the evolution situation of myocardial infarction easily.

Symptoms of myocardial infarction are determined by the parts of the hearta ffected. If the infarct is located in primary motor cortex, contralateral hemiparesis is said to occur. With brainstem localization, heart stem syndromes are typical: Wallenberg's syndrome, Weber's syndrome, Millard–Gubler syndrome, Benedikt syndrome or others.

Infarctions will result in weakness and loss of sensation on the opposite side of the body. Physical examination of the head area will reveal abnormal pupil dilation, light reaction and lack of eye movement on opposite side. If the infarction occurs on the left side heart, speech will be slurred. Reflexes may be aggravated as well.

1.2 TECHNOLOGY USED

1.2.1 Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: *The ability to learn*. Machine learning is actively being used today, perhaps in many more places than one would expect.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

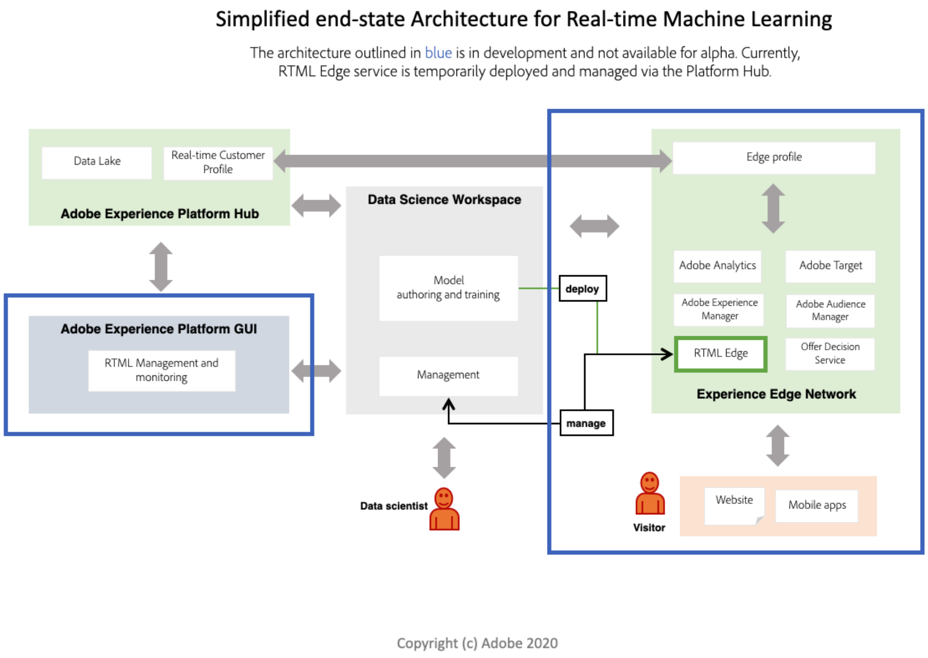

Simplified end-state Architecture for Real-time Machine Learning
The architecture outlined in blue is in development and not available for alpha. Currently, RTML Edge service is temporarily deployed and managed via the Platform Hub.
Data Lake
Real-time Customer Profile
Adobe Experience Platform Hub
Data Science Workspace
Model authoring and training
Management
Adobe Experience Platform GUI
RTML Management and monitoring
deploy
manage
Edge profile
Adobe Analytics
Adobe Target
Adobe Experience Manager
Adobe Audience Manager
RTML Edge
Offer Decision Service
Experience Edge Network
Data scientist
Visitor
Website
Mobile apps
Copyright (c) Adobe 2020

Figure 1.1 Machine learning architecture

1.2.1.1 Machine Learning approaches

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

Supervised learning:

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.

Supervised learning is where there are input variables (x) and an output variable (Y) and an algorithm is used to learn the mapping function from the input to the output.

$Y = f(X)$

The goal is to approximate the mapping function so well that when there is a new input data (x) that the output variables (Y) for that data can be predicted easily.

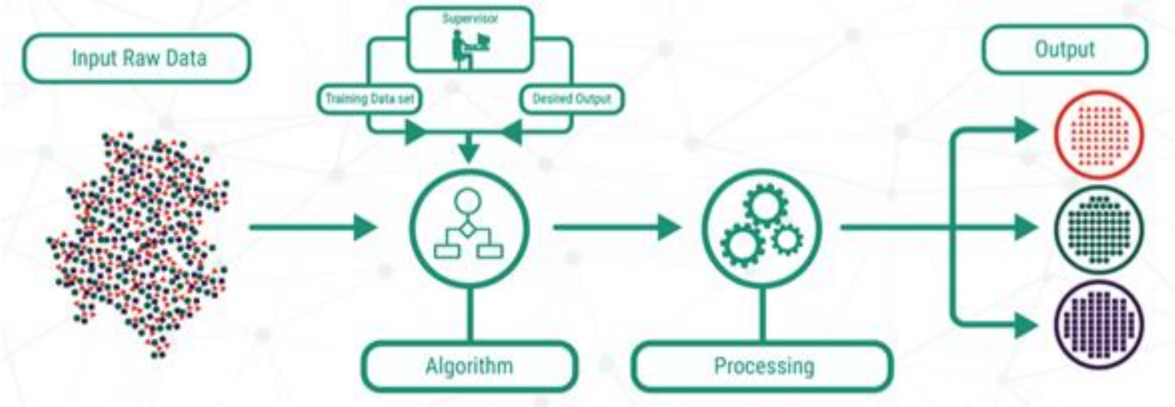


Figure 1.2 Supervised learning flowchart

Unsupervised learning:

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabeled data by itself.

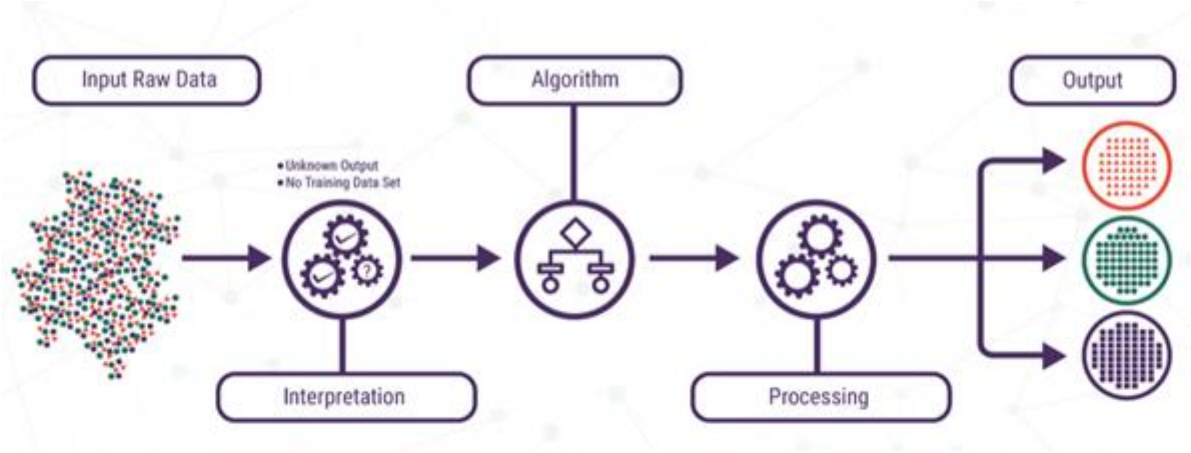


Figure 1.3 Unsupervised learning flowchart

Reinforcement learning:

A computer program interacts with a dynamic environment in which it must perform a certain goal. As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

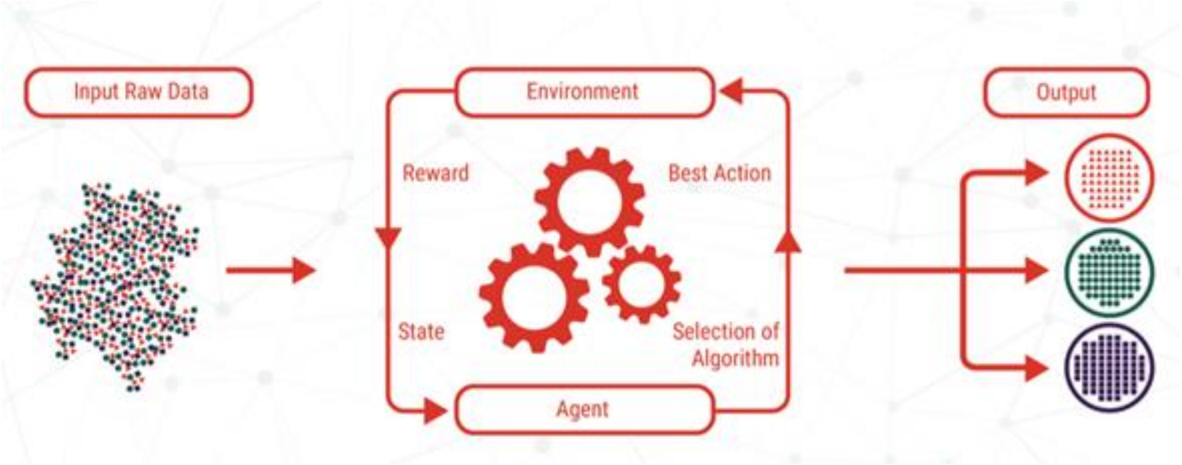


Figure 1.4 Reinforcement learning flowchart

1.3 OBJECTIVE:

- To effectively detect the myocardial infarction in a patient.

- To make use of logistic regression,a machine learning algorithm to make sure of improved accuracy and reliability.

CHAPTER - 2

LITERATURE REVIEW

2.1 INTRODUCTION:

The following shows survey did for myocardial infarction. The most popular of the existing techniques is been discussed as follows.

2.2.1 Avijit Sengupta , Kaushik Dutta , Theresa Beckie, and Sriram Chellappan,"Designing a Health Coach-Augmented mHealth System for the Secondary Prevention of Coronary Heart Disease Among Women"[2020]

While the number of women with Coronary Heart Disease requiring cardiac rehabilitation (CR) continues to increase, lack of access and other barriers to center-based CR present significant challenges. A mobile phone and wearable device-based technological system can offer home-based secondary prevention program for CHD. We have designed a women-specific digital health intervention system for use in a home-based CR program. The system that we have proposed combines human-expert intelligence with machine intelligence to enhance the decision-making capability of a health coach. We have evaluated the prototype of our system with eight women with CHD for over 13 weeks. The evaluation has shown a significant positive impact of preprogrammed intervention messages on participants' walking performance on the same and the next day. Since increased walking activities directly improve heart health, we can directly infer that our proposed system is beneficial as a tool for secondary prevention among women with CHD.

Pros:

The future scope of this system is in the future, especially for personalized and smart healthcare.

Cons:

This project does not give any real time implementation.

2.2.2 Beaumon, P. Onoma, M. Rimlinger , D. Broggio, P. Caldeira Ideias and D.Franck,"Age-specific experimental and computational calibration of thyroid in vivo monitoring",IEEE Transactions on Radiation and Plasma Medical Sciences,2019

Age-specific thyroid phantoms corresponding to 5, 10, 15 years-old and the adult case have been designed and manufactured with a 3D printer. Reference measurements of the counting efficiency have been carried out for thyroid in vivo monitoring of 131 I with all these phantoms. These measurements where performed for the emergency mobile units of IRSN. The full efficiency curve, between 29 and 1000 keV, was then obtained by Monte-Carlo calculations and validated by comparison of a large set of measurements. The obtained efficiency curves are consistent and show that the relative difference in efficiency between the adult and the children case are energy dependent. The developed thyroid phantoms enabled to obtain age specific calibration factors for emergency in vivo monitoring of children. Taking into account the size of thyroid for uptake measurement might be also useful in nuclear medicine department. Indeed, the treatment of benign thyroid disease, like Grave's disease, requires a personalized dosimetry and hence personalized thyroid retention function.

Pros:

The treatment of benign thyroid disease, like Grave's disease, requires a personalized dosimetry and hence personalized thyroid retention function.

Cons:

It is very poor in accuracy

2.2.3 Yi Wang, Na Wang, Min Xu, Junxiong Yu, Chenchen Qin, Xiao Luo, Xin Yang, Tianfu Wang, Anhua Li, and Dong Ni*,” Deeply-Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound,2019

This study presents a rapid and low-cost method to detect thyroid dysfunction using serum Raman spectroscopy combined with support vector machine (SVM). The serum samples taken from 34 thyroid dysfunction patients and 40 healthy volunteers were measured in this study. Tentative assignments of the Raman bands in the measured serum spectra suggested specific biomolecular changes between the groups. Principal component analysis (PCA) was used for feature extraction and reduced the dimension of high-dimension spectral data; then, SVM was employed to establish an effective discriminant model. To improve the efficiency and accuracy of the SVM discriminant model, we proposed artificial fish coupled with uniform design (AFUD) algorithm to optimize the SVM parameters. The average accuracy of 30 discriminant results reached 82.74%, and the average optimization time was 0.45 s. 40 normal thyroid function subjects and 34 abnormal thyroid function patients were analyzed by their serum Raman spectra. The experimental results showed that the profile and peak intensities of the serum spectra were very similar between the two groups, while the subtle differences imply that it was possible to preliminarily screen thyroid function patients through a powerful data analysis algorithm.

Pros:

The experimental results showed that the profile and peak intensities of the serum spectra were very similar between the two groups, while the subtle differences imply that it was possible to preliminarily screen thyroid function patients through a powerful data analysis algorithm.

Cons:

The (AFUD) algorithm needs longer time to train dataset.

2.2.4 Jose M. Anton-Rodriguez , Peter Julyan, Ibrahim Djoukhadar, David Russell, D. Gareth Evans, Alan Jackson, and Julian C. Matthews," Comparison of a Standard Resolution PET-CT Scanner With an HRRT Brain Scanner for Imaging Small Tumors Within the Head", IEEE Transactions on radiation and plasma medical sciences, 2019

We compared a conventional PET-CT scanner (Siemens Biograph TruePoint TrueV) with and without resolution modeling (RM) image reconstruction with a high resolution research tomograph (HRRT) in order to assess the utility of conventional scanners for brain scanning. A modified Esser phantom and 6 neurofibromatosis 2 (NF2) patients with vestibular schwannomas (VS) were scanned using both scanners. The phantom was filled with fluorine-18 (40 MBq, 4:1 contrast ratio) and scanned for 60 min on separate occasions. Patients were injected with ∼200 MBq of [18F] fluorodeoxyglucose (FDG) and [18F] fluorothymidine (FLT) on separate occasions and scanned for three consecutive 30 min periods moving between scanners. The HRRT images, although noisier, resulted in higher contrast recovery for the smallest cylindrical inserts in comparison to TrueV with and without RM. With the TrueV, higheruptake values were observed in VS lesions with both FDG and FLT which is consistent with greater spill-in from the brain for FDG and bone marrow for FLT. RM decreased measured FDG uptake. For large homogenous lesions the conventional TrueV gives similar or better results compared to the HRRT. For smaller lesions, the HRRT has benefit, with RM on the TrueV unable to restore parity, and with the potential for image artifacts.

Pros:

The HRRT has benefit, with RM on the TrueV unable to restore parity, and with the potential for image artifacts.

Cons:

The performance of HRRT is fairly poor compared to the other methods.

2.2.5 Shekoofeh Azizi, Sharareh Bayat, Pingkun Yan, Amir Tahmasebi, Jin Tae Kwak, Sheng Xu, Baris Turkbey, Peter Choyke, Peter Pinto, Bradford Wood, Parvin Mousavi*, Purang Abolmaesumi,” Deep Recurrent Neural Networks for Prostate Cancer Detection: Analysis of Temporal Enhanced Ultrasound IEEE Transactions on Medical Imaging “,2018

Temporal Enhanced Ultrasound (TeUS), comprising the analysis of variations in backscattered signals from a tissue over a sequence of ultrasound frames, has been previously proposed as a new paradigm for tissue characterization. In this manuscript, we propose to use deep Recurrent Neural Networks (RNN) to explicitly model the temporal information in TeUS. By investigating several RNN models, we demonstrate that Long Short-Term Memory (LSTM) networks achieve the highest accuracy in separating cancer from benign tissue in the prostate. We also present algorithms for in-depth analysis of LSTM networks. Our in vivo study includes data from 255 prostate biopsy cores of 157 patients. We achieve area under the curve, sensitivity, specificity, and accuracy of 0.96, 0.76, 0.98 and 0.93, respectively. Our result suggests that temporal modeling of TeUS using RNN can significantly improve cancer detection accuracy over previously presented works.

Pros:

The result suggests that temporal modeling of TeUS using RNN can significantly improve cancer detection accuracy over previously presented works.

Cons:

The results are not very reliable.

2.2.6 Koyel Mandal, Rosy Sarmah, and Dhruba Kumar Bhattacharyya ”Biomarker Identification for Cancer Disease Using Biclustering Approach: An Empirical Study”,2018

In clinical practice, an overwhelming majority of biopsied thyroid nodules are benign. Therefore, there is a need for a complementary and noninvasive imaging tool to provide clinically relevant diagnostic information about thyroid nodules to reduce the rate of unnecessary biopsies. The goal of this study was to evaluate the feasibility of utilizing comb-push ultrasound shear elastography (CUSE) to measure the mechanical properties (i.e., stiffness) of thyroid nodules and use this information to help classify nodules as benign or malignant. CUSE is a fast and robust 2-D shear elastography technique in which multiple laterally distributed acoustic radiation force beams are utilized simultaneously to produce shear waves. Unlike other shear elasticity imaging modalities, CUSE does not suffer from limited field of view (FOV) due to shear wave attenuation and can provide a large FOV at high frame rates. To evaluate the utility of CUSE in thyroid imaging, a preliminary study was performed on a group of five healthy volunteers and 10 patients with ultrasound-detected thyroid nodules prior to fine needle aspiration biopsy. The measured shear wave speeds in normal thyroid tissue and thyroid nodules were converted to Young’s modulus (E), indicating a measure of tissue stiffness. Our results indicate an increase in E for thyroid nodules compared to normal thyroid tissue. This increase was significantly higher in malignant nodules compared to benign.

Pros:

The results indicate an increase in E for thyroid nodules compared to normal thyroid tissue

Cons:

The comb-push ultrasound shear elastography (CUSE) has shorter push duration.

2.2.7 Julien Rouyer, Member, IEEE, Tony Cueva, Student Member, IEEE, Tamy Yamamoto, Alberto Portal, and Roberto Lavarello, Senior Member, IEEE,"In vivo Estimation of Attenuation and Backscatter Coefficients from Human Thyroids",IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 2015.

Fine-needle aspiration (FNA) remains the gold standard for the diagnosis of thyroid cancer. However, currently a large number of FNA biopsies result in negative or undetermined diagnosis, which suggests better non-invasive tools are needed for the clinical management of thyroid cancer. Spectral-based quantitative ultrasound (QUS) characterizations may offer a better diagnostic management as previously demonstrated in mouse cancer models ex vivo. As a first step towards understanding the potential of QUS markers for thyroid disease management, this paper deals with the spectral-based QUS estimation of healthy human thyroids in vivo. Twenty volunteers were inspected by a trained radiologist using two ultrasonic imaging systems, which allowed to acquire radiofrequency data spanning the 3 to 16 MHz frequency range. Estimates of attenuation coefficient slope (ACS) using the spectral logarithmic difference method had an average value of 1.69 dB/(cm.MHz) with a standard deviation of 0.28 dB/(cm.MHz). Estimates of backscatter coefficient (BSC) using the reference phantom method had an average value of 0.18 sr-1.cm-1 over the useful frequency range. The inter-subject variability when estimating BSCs was less than 1.5 dB over the analysis frequency range. Further, the effectiveness of three scattering models (i.e., fluid sphere, Gaussian, and exponential form factors) when fitting the experimentally estimated BSCs was assessed. Quantitative ultrasound characterization of the healthy human thyroid has been presented through the estimation of the ACSs and the BSCs over 20 volunteers inspected in a clinical context.

Pros:

Quantitative ultrasound characterization of the healthy human thyroid has been presented through the estimation of the ACSs and the BSCs over 20 volunteers inspected in a clinical context.

Cons:

It is a relatively recent and non-invasive method.

2.2.8 Chunyu Wang , Junling Guo, Ning Zhao , Yang Liu , Xiaoyan Liu, Guojun Liu , Maozu Guo,” A Cancer Survival Prediction Method Based on Graph Convolutional Network”,Vol:1536-124.2019

Cancer, as the most challenging part in the human disease history, has always been one of the main threats to human life and health. The high mortality of cancer is largely due to the complexity of cancer and the significant differences in clinical outcomes. Therefore, it will be significant to improve accuracy of cancer survival prediction, which has become one of the main fields of cancer research. Many calculation models for cancer survival prediction have been proposed at present, but most of them generate prediction models only by using single genomic data or clinical data. Multiple genomic data and clinical data have not been integrated yet to take a comprehensive consideration of cancers and predict their survival. Method: In order to effectively integrate multiple genomic data (including genetic expression, copy number alteration, DNA methylation and exon expression) and clinical data and apply them to predictive studies on cancer survival, similar network fusion algorithm (SNF) was proposed in this paper to integrate multiple genomic data and clinical data so as to generate sample similarity matrix, minredundancy and max-relevance algorithm (mRMR) was used to conduct feature selection of multiple genomic data and clinical data of cancer samples and generate sample feature matrix, and finally two matrixes were used for semi-supervised training through graph convolutional network (GCN) so as to obtain a cancer survival prediction method integrating multiple genomic data and clinical data based on graph convolutional network (GCGCN). Performance indexes of GCGCN model indicate that both multiple genomic data and clinical data play significant roles in the accurate survival time prediction of cancer patients. It is compared with existing survival prediction methods, and results show that cancer survival prediction method GCGCN which integrates multiple genomic data and clinical data has obviously superior prediction effect than existing survival prediction methods. All study results in this paper have verified effectiveness and superiority of GCGCN in the aspect of cancer survival prediction.

Pros:

Performance indexes of GCGCN model indicate that both multiple genomic data and clinical data play significant roles in the accurate survival time prediction of cancer patients. It is compared with existing survival prediction methods, and results show that cancer survival prediction method GCGCN which integrates multiple genomic data and clinical data has obviously superior prediction effect than existing survival prediction methods.

Cons:

The process is very time consuming, hence is a very inefficient process to consider.

2.2.9 Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao,"Multi-task Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition",IEEE Journal of Biomedical and Health Informatics, AUGUST 2017.

Thyroid ultrasonography is a widely-used clinical technique for nodule diagnosis in thyroid regions. However, it remains difficult to detect and recognize the nodules due to low contrast, high noise, and diverse appearance of nodules. In today's clinical practice, senior doctors could pinpoint nodules by analyzing global context features, local geometry structure, and intensity changes, which would require rich clinical experience accumulated from hundreds and thousands of nodule case studies. To alleviate doctors' tremendous labor in the diagnosis procedure, we advocate a machine learning approach to the detection and recognition tasks in this paper. In particular, we develop a multi-task cascade convolution neural network framework (MCCNN) to exploit the context information of thyroid nodules. It may be noted that, our framework is built upon a large number of clinically-confirmed thyroid ultrasound images with accurate and detailed ground truth labels. Other key advantages of our framework result from a multi-task cascade architecture, two stages of carefully-designed deep convolution networks in order to detect and recognize thyroid nodules in a pyramidal fashion, and capturing various intrinsic features in a global-to-local way. Within our framework, the potential regions of interest after initial detection are further fed to the spatial pyramid augmented CNNs to embed multi-scale discriminative information for fine grained thyroid recognition. The new learning architecture affords the detection and classification tasks to share commonly needed features, with an objective of better distinguishing benign nodules from malignant nodules, as well as the complex background.

Pros:

The new learning architecture affords the detection and classification tasks to share commonly needed features, with an objective of better distinguishing benign nodules from malignant nodules, as well as the complex background.

Cons:

Multi-task Cascade Convolution Neural Network framework (MCCNN) is very inefficient as the computation scales to the number of task

2.2.10 Xiangxiang Zheng ,Guodong Lv, Guoli Du, Zhengang Zhai, Jiaqing Mo and Xiaoyi Lv,"Rapid and Low-Cost Detection of Thyroid Dysfunction Using Raman Spectroscopy and an Improved Support Vector Machine",IEEE Photonics Journal, December 2017.

This study presents a rapid and low-cost method to detect thyroid dysfunction using serum Raman spectroscopy combined with support vector machine (SVM). The serum samples taken from 34 thyroid dysfunction patients and 40 healthy volunteers were measured in this study. Tentative assignments of the Raman bands in the measured serum spectra suggested specific biomolecular changes between the groups. Principal component analysis (PCA) was used for feature extraction and reduced the dimension of high-dimension spectral data; then, SVM was employed to establish an effective discriminant model. To improve the efficiency and accuracy of the SVM discriminant model, we proposed artificial fish coupled with uniform design (AFUD) algorithm to optimize the SVM parameters. The average accuracy of 30 discriminant results reached 82.74%, and the average optimization time was 0.45 s. 40 normal thyroid function subjects and 34 abnormal thyroid function patients were analyzed by their serum Raman spectra. The experimental results showed that the profile and peak intensities of the serum spectra were very similar between the two groups, while the subtle differences imply that it was possible to preliminarily screen thyroid function patients through a powerful data analysis algorithm.

Pros:

The experimental results showed that the profile and peak intensities of the serum spectra were very similar between the two groups, while the subtle differences imply that it was possible to preliminarily screen thyroid function patients through a powerful data analysis algorithm.

Cons:

The results are not upto the mark, hence this system is not recommended.

2.2.11 Nikhil S. Narayan, Pina Marziliano, Jeevendra Kanagalingam, MD and Christopher G.L. Hobbs, MD,"Speckle Patch Similarity for Echogenicity based Multi-Organ Segmentation in Ultrasound Images of the Thyroid Gland",IEEE Journal of Biomedical and Health Informatics,2016.

Ultrasound (US) imaging deals with forming a brightness image from the amplified back-scatter echo when an ultrasound wave is triggered at the region of interest. Imaging artefacts and speckles occur in the image as a consequence of back-scattering and subsequent amplification. We demonstrate the usefulness of speckle related pixels and imaging artefacts as sources of information to perform multi-organ segmentation in US images of the thyroid gland. The speckle related pixels are clustered based on a similarity constraint to quantize the image. The quantization results are used to locate useful anatomical landmarks that aid the detection of multiple organs in the image which are the thyroid gland, the carotid artery, the muscles and the trachea. The spatial locations of the carotid artery and the trachea are used to estimate the boundaries of the thyroid gland in transverse US scans. The SRR's are used to detect hyper-echoic anatomical landmarks which are used to detect and segment other organs using local phase based methods. The proposed method has an overall accuracy of over 92% and performs better than existing methods to segment both individual and multiple organs.

Pros:

The proposed method has an overall accuracy of over 92% and performs better than existing methods to segment both individual and multiple organs.

Cons:

The Ultrasound images are very risky in terms of the radiations emitted.

2.2.12 Jennifer E. Rosen∗ , Hyunsuk Suh, Nicholas J. Giordano, Ousama M. A’amar, Eladio Rodriguez-Diaz, Irving I. Bigio, and Stephanie L. Lee,”Preoperative Discrimination of Benign from Malignant Disease in Thyroid Nodules With Indeterminate Cytology Using Elastic Light-Scattering Spectroscopy”,IEEE Transactions On Biomedical Engineering, August 2016.

Thyroid nodules are common and often require fine needle aspiration biopsy (FNAB) to determine the presence of malignancy to direct therapy. Unfortunately, approximately 15–30% of thyroid nodules evaluated by FNAB are not clearly benign or malignant by cytology alone. These patients require surgery for the purpose of diagnosis alone; most of these nodules ultimately prove to be benign. Elastic light scattering spectroscopy (ESS) that measures the spectral differences between benign and malignant thyroid nodules has shown promise in improving preoperative determination of benign status of thyroid nodules. We describe the results of a large, prospective, blinded study validating the ESS algorithm in patients with thyroid nodules. An ESS system was used to acquire spectra from human thyroid tissue. Spectroscopic results were compared to the histopathology of the biopsy samples.

Pros:

The result shows differences between, malignant and benign thyroid.

Cons:

May result in optical dis alignment in this method, when using Elastic light scattering spectroscopy (ESS).

2.2.13 Micha Feigin, Daniel Freedman, and Brian W. Anthony,” A Deep Learning Framework for Single-Sided Sound Speed Inversion in Medical Ultrasound”,2016

Ultrasound elastography is gaining traction as an accessible and useful diagnostic tool for such things as cancer detection and differentiation and thyroid disease diagnostics. Unfortunately, state of the art shear wave imaging techniques, essential to promote this goal, are limited to highend ultrasound hardware due to high power requirements; are extremely sensitive to patient and sonographer motion, and generally, suffer from low frame rates. Motivated by research and theory showing that longitudinal wave sound speed carries similar diagnostic abilities to shear wave imaging, we present an alternative approach using single sided pressure-wave sound speed measurements from channel data. Methods: In this paper, we present a single-sided sound speed inversion solution using a fully convolutional deep neural network. We use simulations for training, allowing the generation of limitless ground truth data. Results: We show that it is possible to invert for longitudinal sound speed in soft tissue at high frame rates. We validate the method on simulated data. We present highly encouraging results on limited real data. Conclusion: Sound speed inversion on channel data has significant potential, made possible in real time with deep learning technologies. Significance: Specialized shear wave ultrasound systems remain inaccessible in many locations. longitudinal sound speed and deep learning technologies enable an alternative approach to diagnosis based on tissue elasticity. High frame rates are possible.

Pros:

Specialized shear wave ultrasound systems remain inaccessible in many locations. longitudinal sound speed and deep learning technologies enable an alternative approach to diagnosis based on tissue elasticity. High frame rates are possible.

Cons:

The diagnosis is only on based on tissue elasticity

2.2.14 Amir Mirbeik-Sabzevari, Student Member, IEEE, Negar Tavassolian, Senior Member, IEEE "Ultra-Wideband, Stable Normal and Cancer Skin Tissue Phantoms for Milli-meter-Wave Skin Cancer Imaging" ,2015

This work introduces new, stable, and broadband skin-equivalent semisolid phantoms for mimicking interactions of milli-meter waves with the human skin and skin cancer. Realistic skin phantoms serve as an invaluable tool for exploring the feasibility of new technologies and improving design concepts related to milli-meter-wave skin cancer detection methods. Normal and malignant skin tissues are separately mimicked by using appropriate mixtures of deionized water, oil, gelatin powder, formaldehyde, TX-150 (a gelling agent, widely referred to as 'super stuff'), and detergent. The dielectric properties of the phantoms are characterized over the frequency band of 0.5–50 GHz using a slim-form open-ended coaxial probe in conjunction with a milli-meter-wave vector network analyser. The measured permittivity results show excellent match with ex-vivo, fresh skin (both normal and malignant) permittivities determined in our prior work over the entire frequency range. This work results in the closest match among all phantoms reported in the literature to surrogate human skin tissues. The stability of dielectric properties over time is also investigated. The phantoms demonstrate long-term stability (up to 7 months was investigated). In addition, the penetration depth of milli-meter waves into normal and malignant skin phantoms is calculated. It is determined that milli-meter waves penetrate the human skin deep enough (0.6 mm on average at 50 GHz) to affect the majority of the epidermis and dermis skin structures

Pros:

It is determined that milli-meter waves penetrate the human skin deep enough (0.6 mm on average at 50 GHz) to affect the majority of the epidermis and dermis skin structures.

Cons:

The radiations emitted during the process may seem fatal in this process.

2.2.15 Yanbo Wang , Weikang Qian , Bo Yuan," A graphical model of smoking-induced global instability in lung cancer",2015

Smoking is the major cause of lung cancer and the leading cause of cancer-related death in the world. The most current view about lung cancer is no longer limited to individual genes being mutated by any carcinogenic insults from smoking. Instead, tumorigenesis is a phenotype conferred by many systematic and global alterations, leading to extensive heterogeneity and variation for both the genotypes and phenotypes of individual cancer cells. Thus, strategically it is foremost important to develop a methodology to capture any consistent and global alterations presumably shared by most of the cancerous cells for a given population. This is particularly true that almost all of the data collected from solid cancers (including lung cancers) are usually distant apart over a large span of temporal or even spatial contexts. Here we report a multiple non-Gaussian graphical model to reconstruct the gene interaction network using two previously published gene expression datasets. Our graphical model aims to selectively detect gross structural changes at the level of gene interaction networks. Our methodology is extensively validated, demonstrating good robustness, as well as the selectivity and specificity expected based on our biological insights. In summary, gene regulatory networks are still relatively stable during presumably the early stage of neoplastic transformation. But drastic structural differences can be found between lung cancer and its normal control, including the gain of functional modules for cellular proliferations such as EGFR and PDGFRA, as well as the lost of the important IL6 module, supporting their roles as potential drug targets. Interestingly, our method can also detect early modular changes, with the ALDH3A1 and its associated interactions being strongly implicated as a potential early marker, whose activations appear to alter LCN2 module as well as its interactions with the important TP53-MDM2 circuitry. Our strategy using the graphical model to reconstruct gene interaction work with biologically-inspired constraints exemplifies the importance and beauty of biology in developing any bio-computational approach.

Pros:

This method exemplifies the importance and beauty of biology in developing any bio-computational approach.

Cons:

The results seem to be not very accurate and are not very reliable.

2.3 DISADVANTAGES IN EXISTING SYSTEM

- The existing system focuses on identifying the heart rate in women
- Sensors are used to predict the condition of the patient where there is a possibility of sensor failure will lead to inaccurate prediction
- Ineffective in real time

CHAPTER - 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM:

While the number of women with Coronary Heart Disease requiring cardiac rehabilitation (CR) continues to increase, lack of access and other barriers to center-based CR present significant challenges. A mobile phone and wearable device-based technological system can offer home-based secondary prevention program for CHD. We have designed a women-specific digital health intervention system for use in a home-based CR program. The system that we have proposed combines human-expert intelligence with machine intelligence to enhance the decision-making capability of a health coach. We have evaluated the prototype of our system with eight women with CHD for over 13 weeks. The evaluation has shown a significant positive impact of preprogrammed intervention messages on participants' walking performance on the same and the next day. Since increased walking activities directly improve heart health, we can directly infer that our proposed system is beneficial as a tool for secondary prevention among women with CHD.

3.2 DISADVANTAGES OF EXISTING SYSTEM:

- The existing system focuses on identifying the heart rate in women
- Sensors are used to predict the condition of the patient where there is a possibility of sensor failure will lead to inaccurate prediction
- Ineffective in real time

3.3 PROPOSED SYSTEM:

A myocardial infarction (MI), commonly known as a heart attack, occurs when blood flow decreases or stops to a part of the heart, causing damage to the heart muscle.The most common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck or jaw. Often it occurs in the center or left side of the chest and lasts for more than a few minutes.The discomfort may occasionally feel like heartburn.Other symptoms may include shortness of breath, nausea, feeling faint, a cold sweat or feeling tired. About 30% of people have atypical symptoms.Women more often present without chest pain and instead have neck pain, arm pain or feel tired.In the proposed system, Logistic regression, a machine learning regression algorithm is used to detect the myocardial infarction in a patient. The algorithm is trained using a structured medical dataset. By accurately detecting the myocardial infarction in a patient, this project is very significant in the medical field.

3.4 ADVANTAGES:

- Accurate detection of myocardial infarction in a patient
- Makes use of logistic regression to identify the disease
- The logistic regression achieves good accuracy performs well when the dataset is linearly separable.
- It is very fast at classifying unknown records.

3.5 APPLICATIONS:

- Hopitals
- Medical test centers

CHAPTER 4

SYSTEM DESIGN

DETAILED DESIGN OF THE PROJECT:

This chapter describes the overall and the detailed architectural design. It also describes each module that is to be implemented along with Data Flow diagram.

4.1 ARCHITECTURE DIAGRAM:

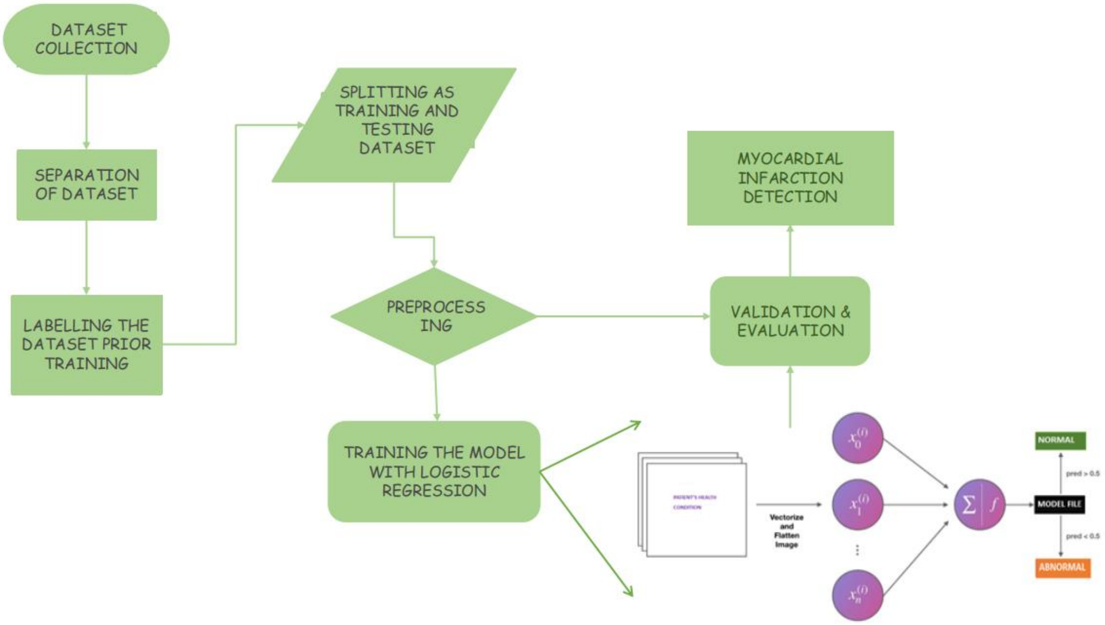

DATASET COLLECTION
SEPARATION OF DATASET
LABELLING THE DATASET PRIOR TRAINING
SPLITTING AS TRAINING AND TESTING DATASET
PREPROCESS ING
TRAINING THE MODEL WITH LOGISTIC REGRESSION
PATIENT'S HEALTH CONDITION
Vectorize and Flatten Image
$x_0^{(i)}$
$x_1^{(i)}$
$x_n^{(i)}$
Σ | f
MODEL FILE
pred > 0.5
NORMAL
pred < 0.5
ABNORMAL
VALIDATION & EVALUATION
MYOCARDIAL INFARCTION DETECTION

Figure 4.1 System Architecture

4.2 WORKING:

The aim of this project is to investigate and implement algorithms that could possibly identify and predict myocardial infarction in a patient. Data mining techniques and machine learning algorithms can be used for the prediction and detection of diseases .In this project the initial step will be collecting different datasets from internet on various medical test reports of the myocardial infarction from the internet which will be helpful in analysing the occurence of myocardial infarction then those dataset will be aligned accordingly. Then it will undergo a process called separation of datasets into training as well as testing where the training datasets will be used to train the model as well as testing will be used for evaluating the model. Then dataset pre-processing will be done which will align all the datasets into a specific category. There exist several regression algorithms in machine learning to develop a disease detection model such as random forest regressor algorithm . In the proposed system, logistic regression, a regression machine learning algorithm is used to train the dataset consisting of data on lab test results for myocardial infarction in a patient. By accurately predicting and detection the occurrence of myocardial infarction in a patient at a very early stages , this application could potentially save numerous lives.

4.3 MODULE DESCRIPTION:

- Medical dataset collection
- Splitting of Dataset
- Dataset clustering
- Data pre-processing
- Disease prediction using machine learning

4.3.1 MEDICAL DATASET COLLECTION MODULE:

A data set is a collection of data. Machine learning has become the go-to method for solving many challenging real-world problems. It’s definitely by far the best performing method for prediction tasks. These machine learning machines that have been working so well need fuel lots of fuel; that fuel is data. The more **labelled data** available, the better our model performs. The idea of more data leading to better performance has even been explored at a large-scale by Google with a dataset of 300 Million images!

When deploying a machine learning model in a real-world application**, data must be constantly fed** to continue improving its performance. And, in the machine learning era, data is very well arguably the most valuable resource. There are three steps of collecting data.

Classification. When an algorithm to answer binary yes-or-no questions or to make a multi-class classification (*grass, trees, or bushes*; *cats, dogs, or birds etc*.)

Regression. For an algorithm to yield some numeric value. For example, if you spend too much time coming up with the right price for your product since it depends on many factors, regression algorithms can aid in estimating this value.

Ranking. Some machine learning algorithms just rank objects by a number of features. Ranking is actively used to recommend movies in video streaming services or show the products that a customer might purchase with a high probability based on his or her previous search and purchase activities.

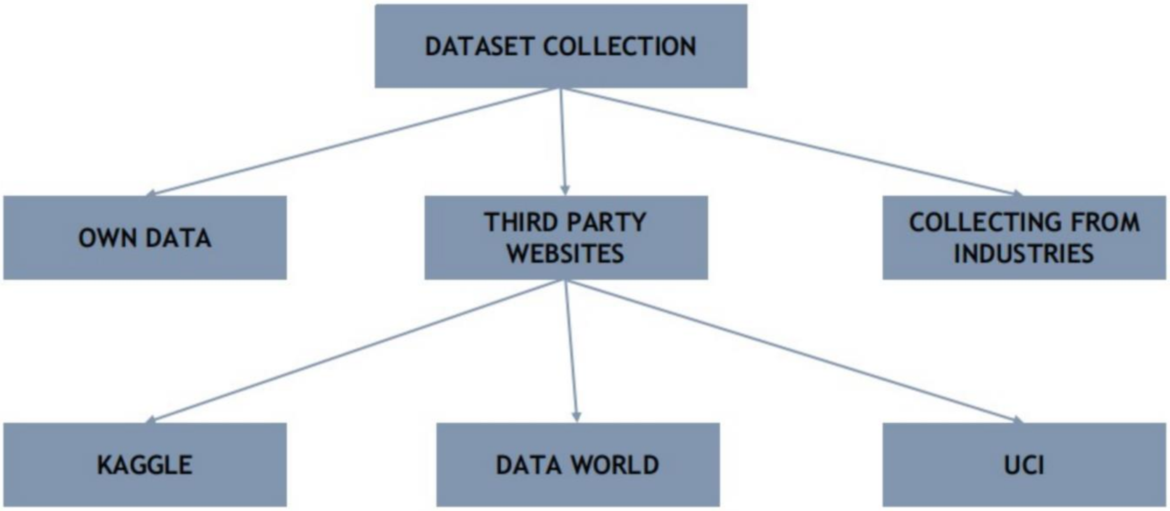


Figure 4.2 Dataset collection

4.3.2 SPLITTING OF DATASET

In machine learning ,any dataset is usually split into two : training data and test data. The output variable along with other variables are included in the training set . The model learns the data and tries to generate some pattern . The other part of the dataset serves as a test set to validate our model’s prediction. The scikit library has a function called train_test_split to divide our data . test_size is the parameter which gives us the percentage of data that should belong to the test set. train_size stores the remaining part as the training dataset , either of which should be specified. random_state acts as a random number generator . For our dataset , we split the training and testing set with 80 , 20 ratio the random state is passed as 0.

4.3.3 DATASET CLUSTERING

Dataset Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as *"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."*

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets. The clustering technique is commonly used for **statistical data analysis.**

4.3.4 DATA PRE-PROCESSING MODULE

Data pre-processing is a cleaning technique which is used to convert / transform the raw data into a clean and properly structured dataset suitable for further analysis. Data is usually collected and gathered from various sources , so it should be good enough and in some specific format before the model learns or gets trained with the data. This will help in achieving better and accurate results with valuable information. The basic steps in pre-processing involve filling up missing values and null values , getting rid of possible outliers and normalisation.

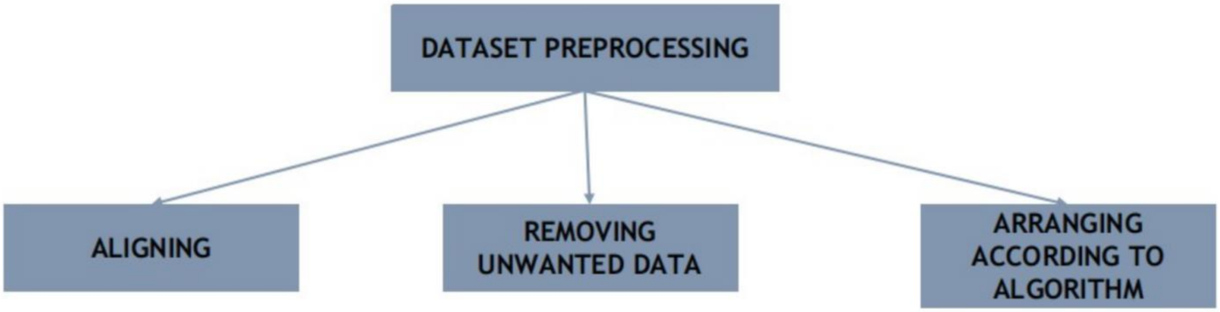


Figure 4.3 Dataset preprocessing

4.3.5 DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

In this project logistic regression a machine learning algorithm is used to predict if the patient is normal or is affected with myocardial infarction on the input medical data.

Logistic Regression statistical method is used for analyzing the dataset and produces a binary outcome. One or more autonomous variables may have consisted of the dataset. The result is determined by these variables that are dichotomous in nature. Which means only two results are possible. It is a specific category of regression and it is used in the best way to predict the binary and categorical output. Logistical Regression method is used to regulate the impact of numerous autonomous variables which are conferred at the same time. This method also predicts any one of the two independent categories of variables. Logistic regression designs the best-fitting function with the help of the maximum likelihood method in order to maximize the probability of classifying the recognized data into the proper division.The other various applications of logistic regression are forecast market trends, to find the success and failure rates in results, the true or false category in recruiting employees based on their performance in need of employment in a company, image categorization, health care and analyze a group of people affected by myocardial Infarction.

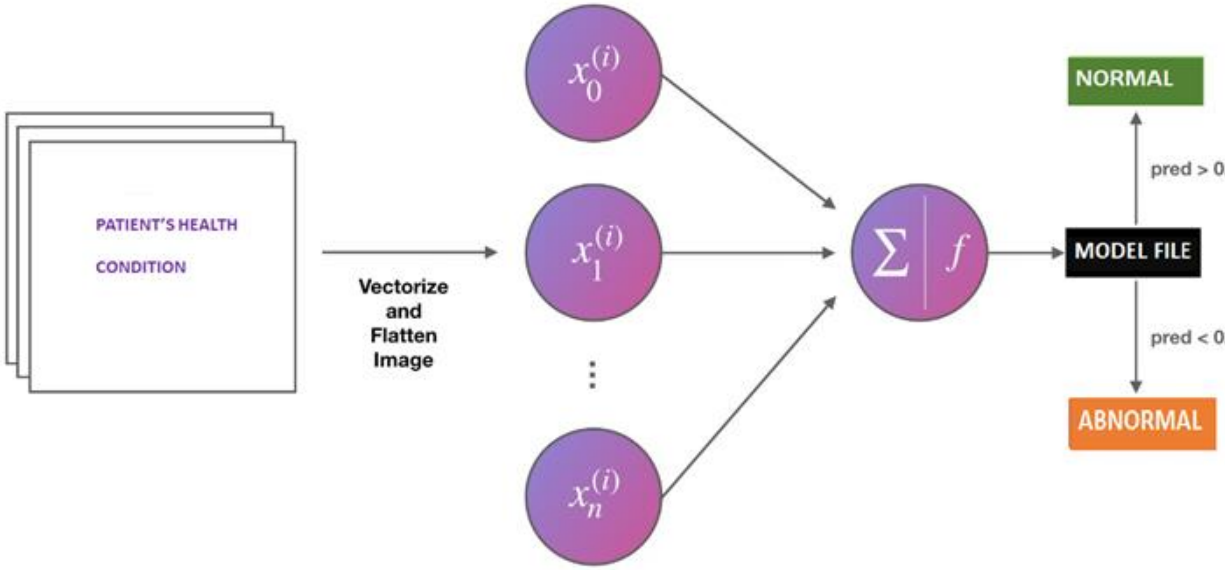


Figure 4.4 Logistic regression architecture

CHAPTER 5

SOFTWARE DESCRIPTION

5.1 Jupyter notebook

In this project the jupyter notebook is used as an IDE.

At some point, we all need to show our work. Most programming work is shared either as raw source code or as a compiled executable. The source code provides complete information, but in a way that's more "tell" than "show." The executable shows us what the software does, but even when shipped with the source code it can be difficult to grasp exactly how it works.

Imagine being able to view the code and execute it in the same UI, so that you could make changes to the code and view the results of those changes instantly, in real time? That's just what Jupyter Notebook offers.

Jupyter Notebook was created to make it easier to show one's programming work, and to let others join in. Jupyter Notebook allows you to combine code, comments, multimedia, and visualizations in an interactive document — called a notebook, naturally — that can be shared, re-used, and re-worked.

And because Jupyter Notebook runs via a web browser, the notebook itself could be hosted on your local machine or on a remote server.

5.2 Python:

In this project python is used as programming language.

In technical terms, Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options.

Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers.

Additionally, Python supports the use of modules and packages, which means that programs can be designed in a modular style and code can be reused across a variety of projects. Once you've developed a module or package you need, it can be scaled for use in other projects, and it's easy to import or export these modules.

One of the most promising benefits of Python is that both the standard library and the interpreter are available free of charge, in both binary and source form. There is no exclusivity either, as Python and all the necessary tools are available on all major platforms. Therefore, it is an enticing option for developers who don't want to worry about paying high development costs.

If this description of Python over your head, don't worry. You'll understand it soon enough. What you need to take away from this section is that Python is a programming language used to develop software on the web and in app form, including mobile. It's relatively easy to learn, and the necessary tools are available to all free of charge.

CHAPTER 6

RESULTS AND DISCUSSION

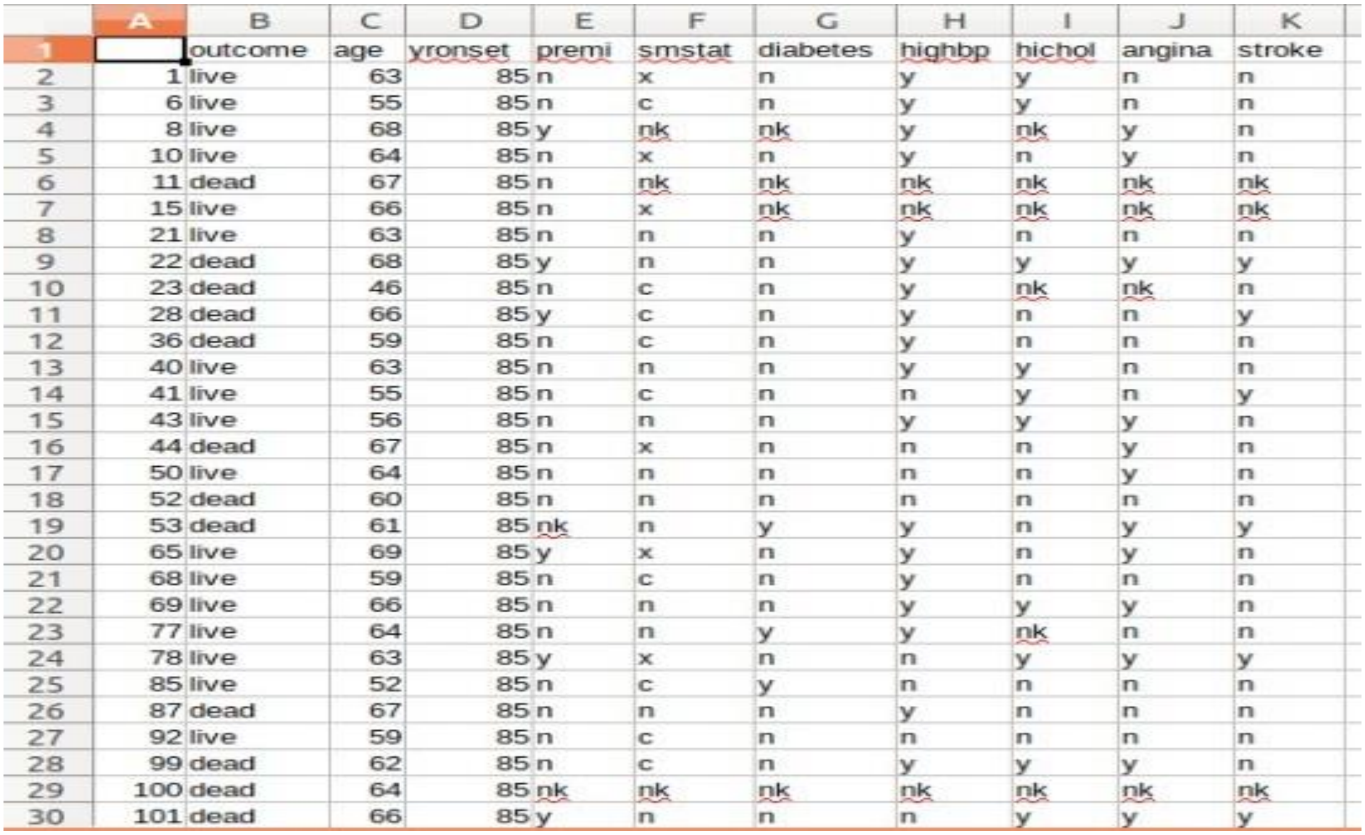
6.1 FINAL RESULTS OBTAINED

To begin with, testing of the trained model, we can split our project into modules of implementation that is done.

Dataset collection involves the process of collecting medical dataset of people affected from myocardial infarction

Various datasets were collected and one example among the collected dataset can be found below

The below screenshot shows a sample of dataset collected



	A	B	C	D	E	F	G	H	I	J	K
1		outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
2	1	live	63	85	n	x	n	y	y	n	n
3	6	live	55	85	n	c	n	y	y	n	n
4	8	live	68	85	y	nk	nk	y	nk	y	n
5	10	live	64	85	n	x	n	y	n	y	n
6	11	dead	67	85	n	nk	nk	nk	nk	nk	nk
7	15	live	66	85	n	x	nk	nk	nk	nk	nk
8	21	live	63	85	n	n	n	y	n	n	n
9	22	dead	68	85	y	n	n	y	y	y	y
10	23	dead	46	85	n	c	n	y	nk	nk	n
11	28	dead	66	85	y	c	n	y	n	n	y
12	36	dead	59	85	n	c	n	y	n	n	n
13	40	live	63	85	n	n	n	y	y	n	n
14	41	live	55	85	n	c	n	n	y	n	y
15	43	live	56	85	n	n	n	y	y	y	n
16	44	dead	67	85	n	x	n	n	n	y	n
17	50	live	64	85	n	n	n	n	n	y	n
18	52	dead	60	85	n	n	n	n	n	n	n
19	53	dead	61	85	nk	n	y	y	n	y	y
20	65	live	69	85	y	x	n	y	n	y	n
21	68	live	59	85	n	c	n	y	n	n	n
22	69	live	66	85	n	n	n	y	y	y	n
23	77	live	64	85	n	n	y	y	nk	n	n
24	78	live	63	85	y	x	n	n	y	y	y
25	85	live	52	85	n	c	y	n	n	n	n
26	87	dead	67	85	n	n	n	y	n	n	n
27	92	live	59	85	n	c	n	n	n	n	n
28	99	dead	62	85	n	c	n	y	y	y	n
29	100	dead	64	85	nk	nk	nk	nk	nk	nk	nk
30	101	dead	66	85	y	n	n	n	y	y	y

Figure 6.1 Dataset Collected

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.The below image shows the histogram of mycardial infarction disease onset age as per outcome:

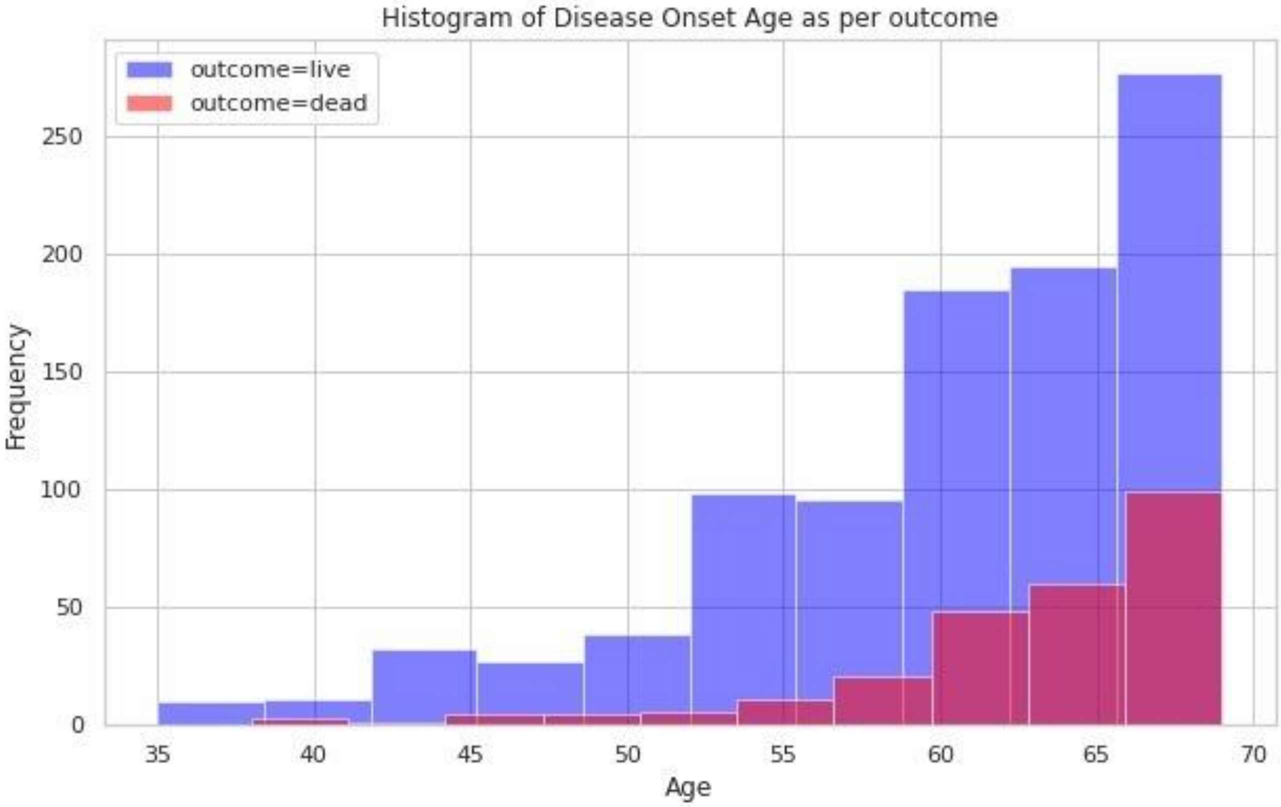
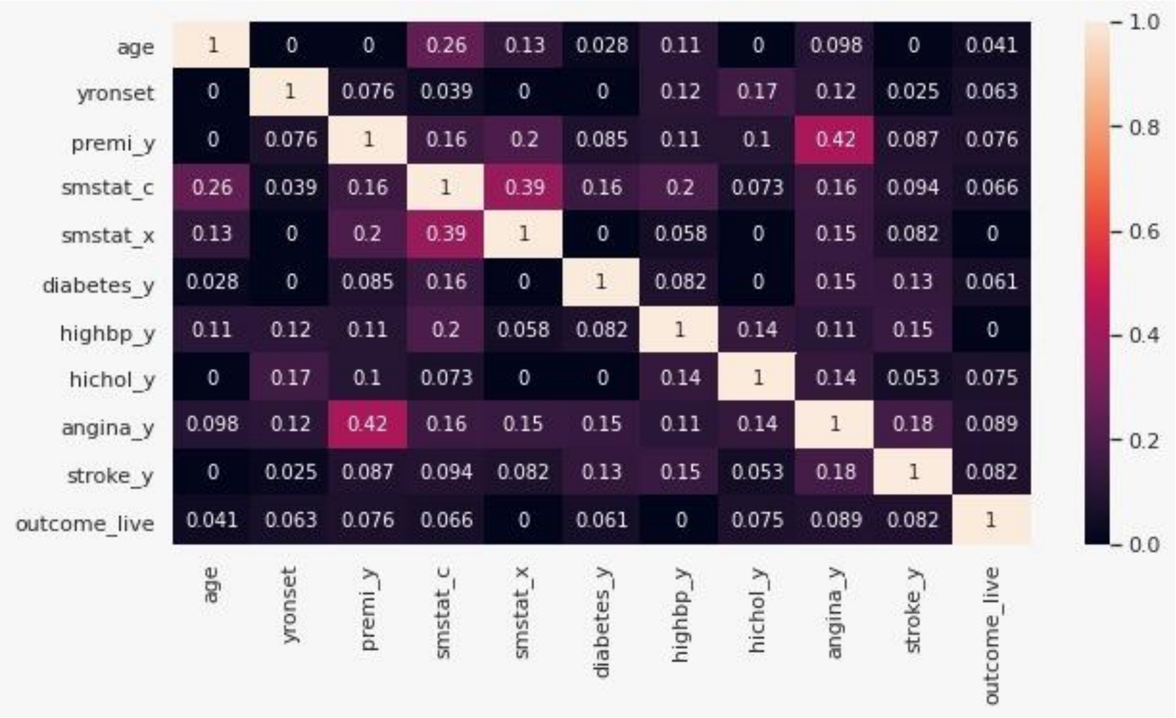


Figure 6.2: Histogram of Disease Onset Age

A heat map (or heatmap) is a graphical representation of data where values are depicted by color. Heat maps make it easy to visualize complex data and understand it at a glance.The below image shows the heat map of mycardial infarction disease onset age as per outcome:



	age	yronset	premi_y	smstat_c	smstat_x	diabetes_y	highbp_y	hichol_y	angina_y	stroke_y	outcome_live
age	1	0	0	0.26	0.13	0.028	0.11	0	0.098	0	0.041
yronset	0	1	0.076	0.039	0	0	0.12	0.17	0.12	0.025	0.063
premi_y	0	0.076	1	0.16	0.2	0.085	0.11	0.1	0.42	0.087	0.076
smstat_c	0.26	0.039	0.16	1	0.39	0.16	0.2	0.073	0.16	0.094	0.066
smstat_x	0.13	0	0.2	0.39	1	0	0.058	0	0.15	0.082	0
diabetes_y	0.028	0	0.085	0.16	0	1	0.082	0	0.15	0.13	0.061
highbp_y	0.11	0.12	0.11	0.2	0.058	0.082	1	0.14	0.11	0.15	0
hichol_y	0	0.17	0.1	0.073	0	0	0.14	1	0.14	0.053	0.075
angina_y	0.098	0.12	0.42	0.16	0.15	0.15	0.11	0.14	1	0.18	0.089
stroke_y	0	0.025	0.087	0.094	0.082	0.13	0.15	0.053	0.18	1	0.082
outcome_live	0.041	0.063	0.076	0.066	0	0.061	0	0.075	0.089	0.082	1

Figure 6.3:Heat map

Survival ratio is a part of survival analysis. It is the percentage of people in a study or treatment group still alive for a given period of time after diagnosis.The below image shows the survival ratio of people after diagnosis of myocardial infarction:

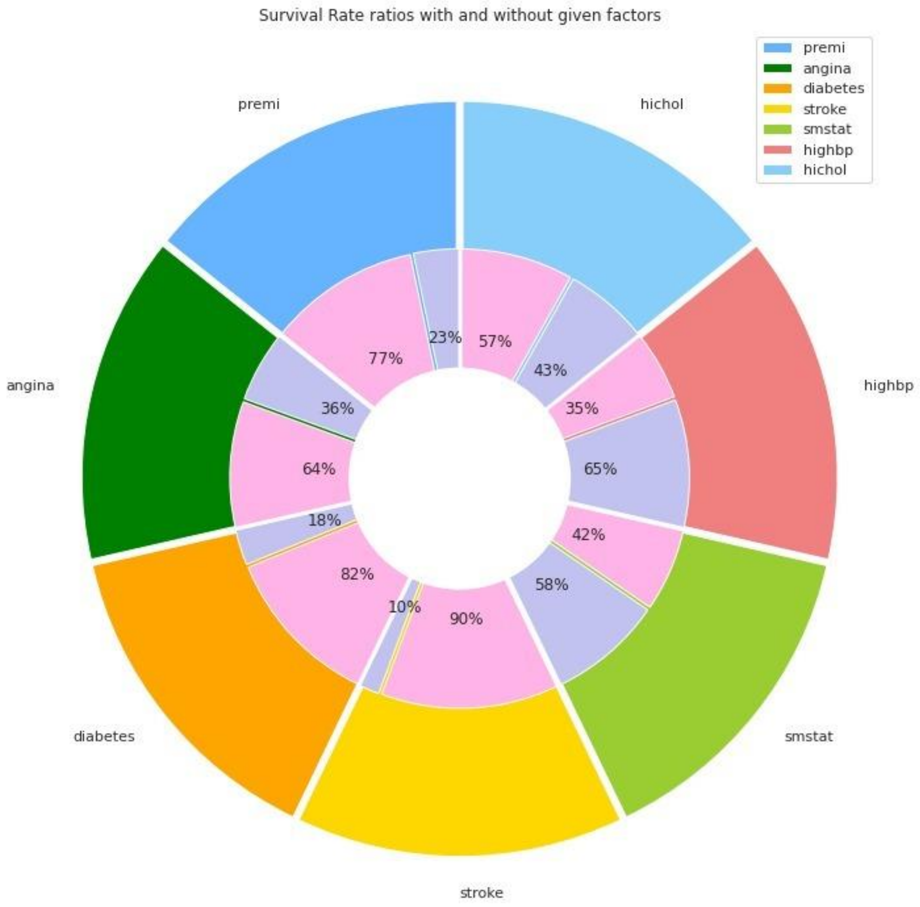


Figure 6.3:Survival Ratios

Mortality ratio is a part of mortality analysis. It is the percentage of people in a study or treatment group dead after diagnosis or treatment of a particular disease.The below image shows the mortalilty ratio of people after diagnosis of myocardial infarction:

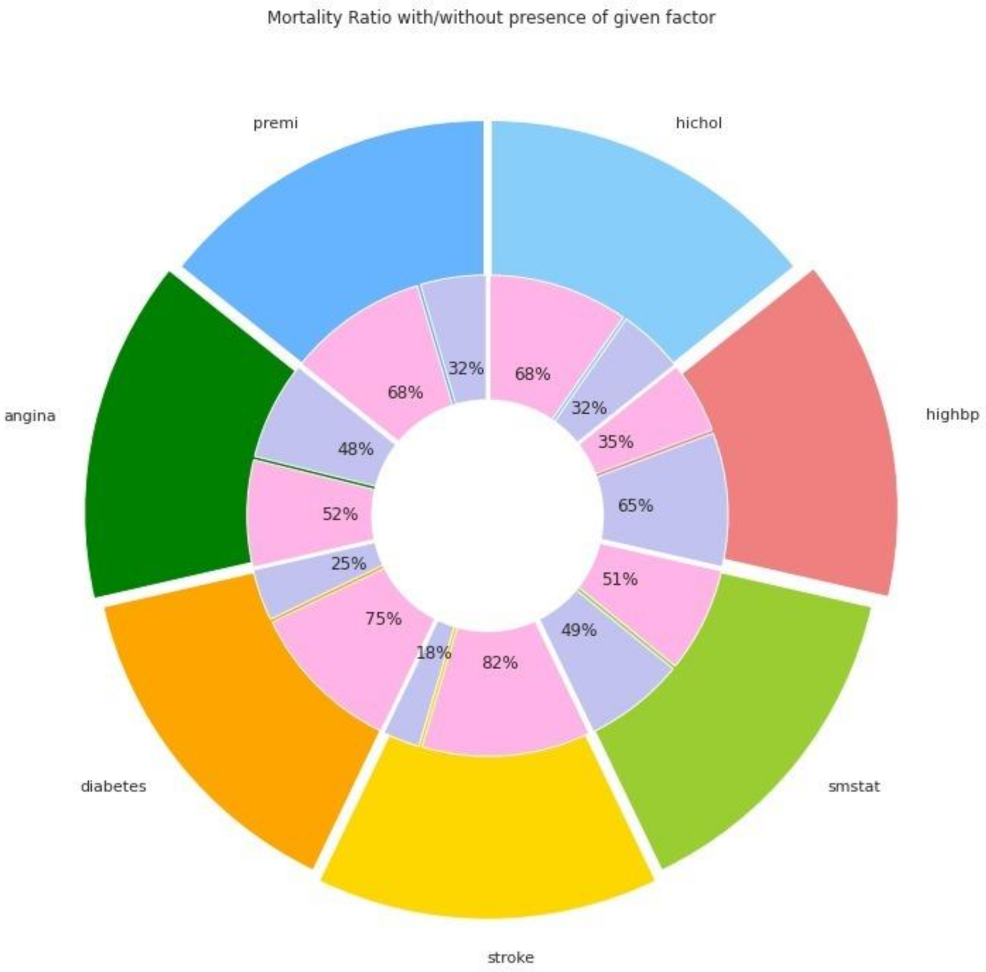


Figure 6.4: Mortality Ratio

SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.The below image shows the graph before the smote analysis is applied.

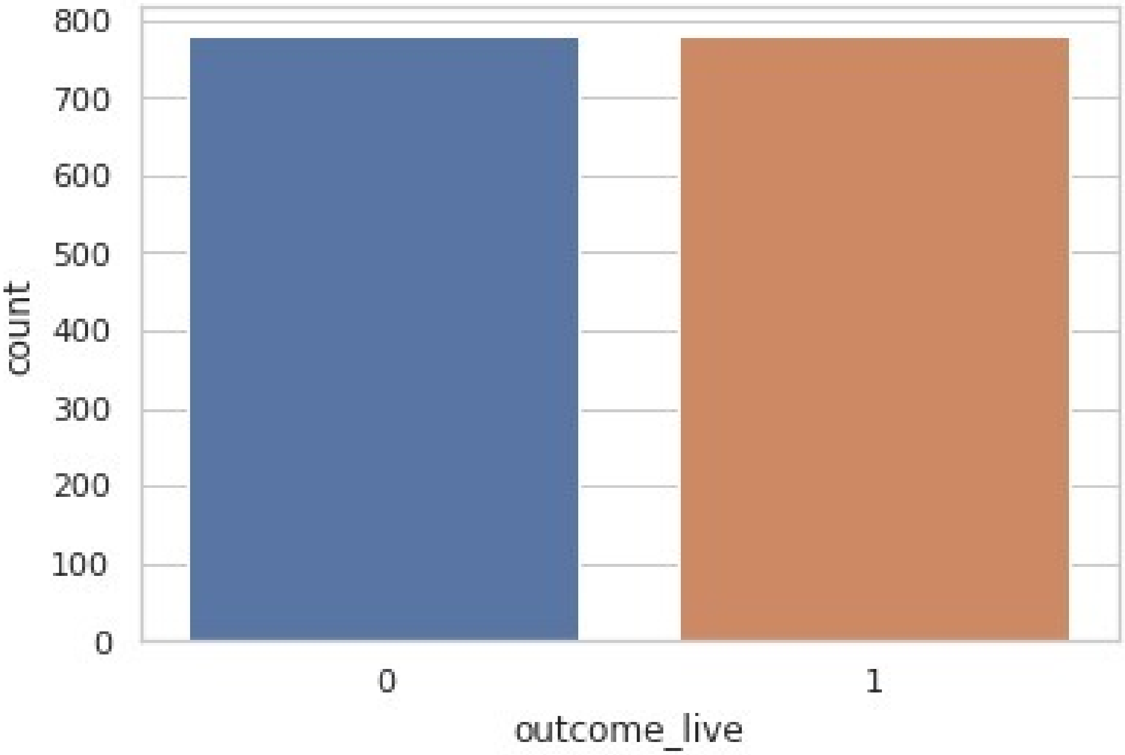


Figure 6.5:Before SMOTE Analysis

The below image shows the graph after the smote analysis is applied.

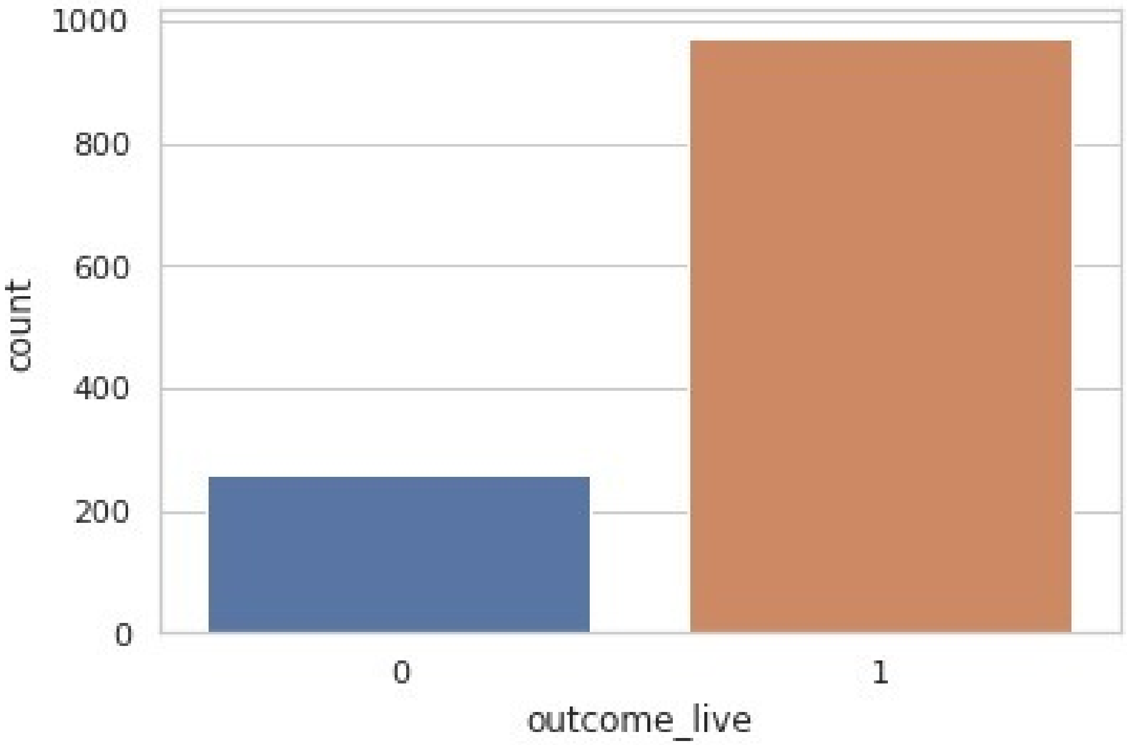
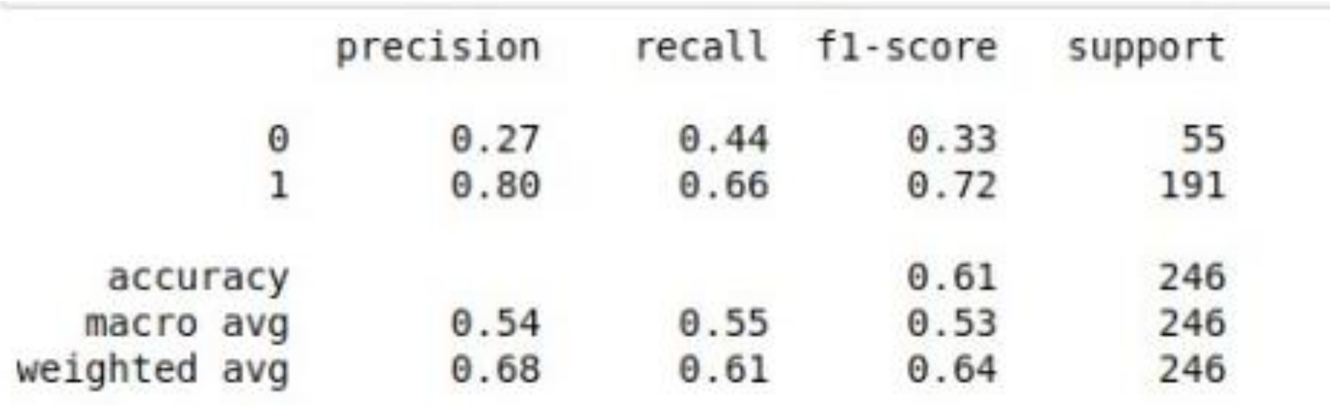


Figure 6.7:After SMOTE Analysis

The below image shows the accuracy achieved from the training process:



```
              precision    recall  f1-score   support

           0       0.27      0.44      0.33        55
           1       0.80      0.66      0.72       191

    accuracy                           0.61       246
   macro avg       0.54      0.55      0.53       246
weighted avg       0.68      0.61      0.64       246
```

Figure 6.8:Output Accuracy Obtained After Training Process

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.The below image shows the confusion matrix obtained for training with the logistic regression algorithm:

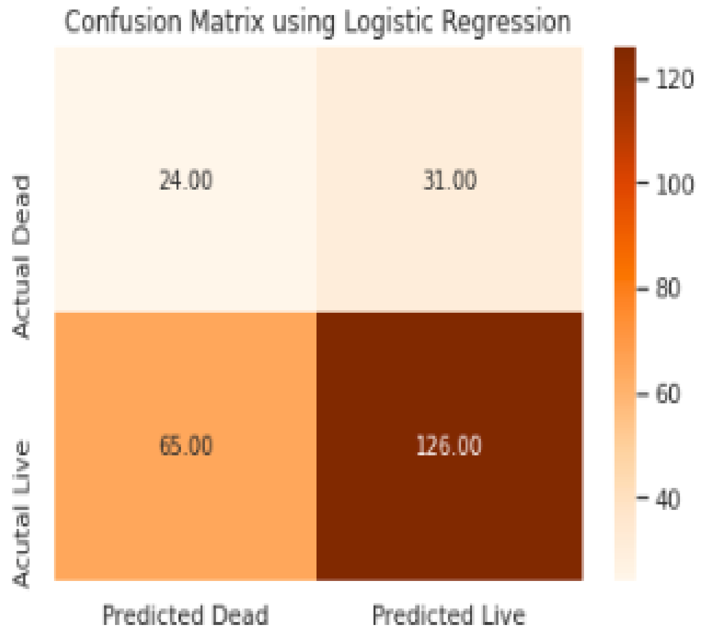


Figure 6.9:CONFUSION MATRIX FROM LOGISTIC REGRESSION

The below image shows the confusion matrix obtained for training with the decision tree classifier algorithm:

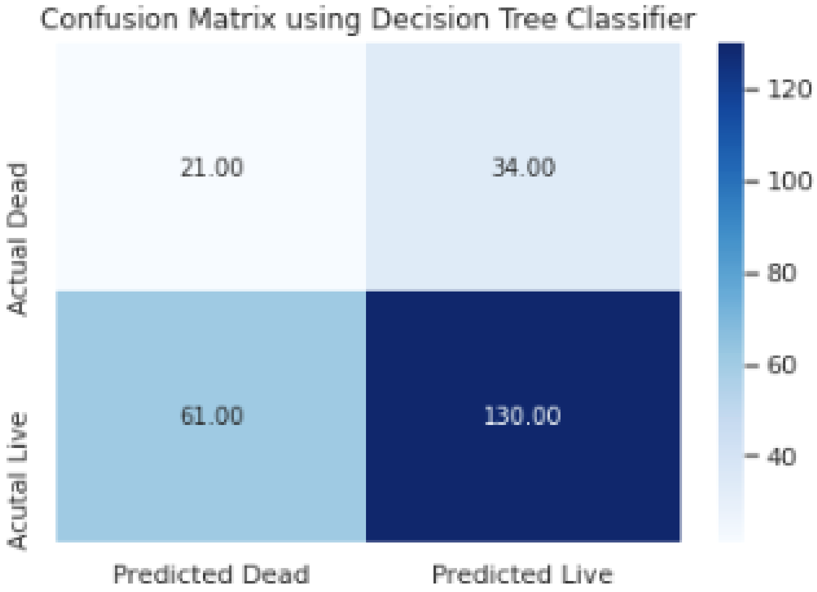


Figure 6.10:CONFUSION MATRIX FROM DECISION TREE CLASSIFIER

The below image shows the confusion matrix obtained for training with the random forest classifier algorithm:

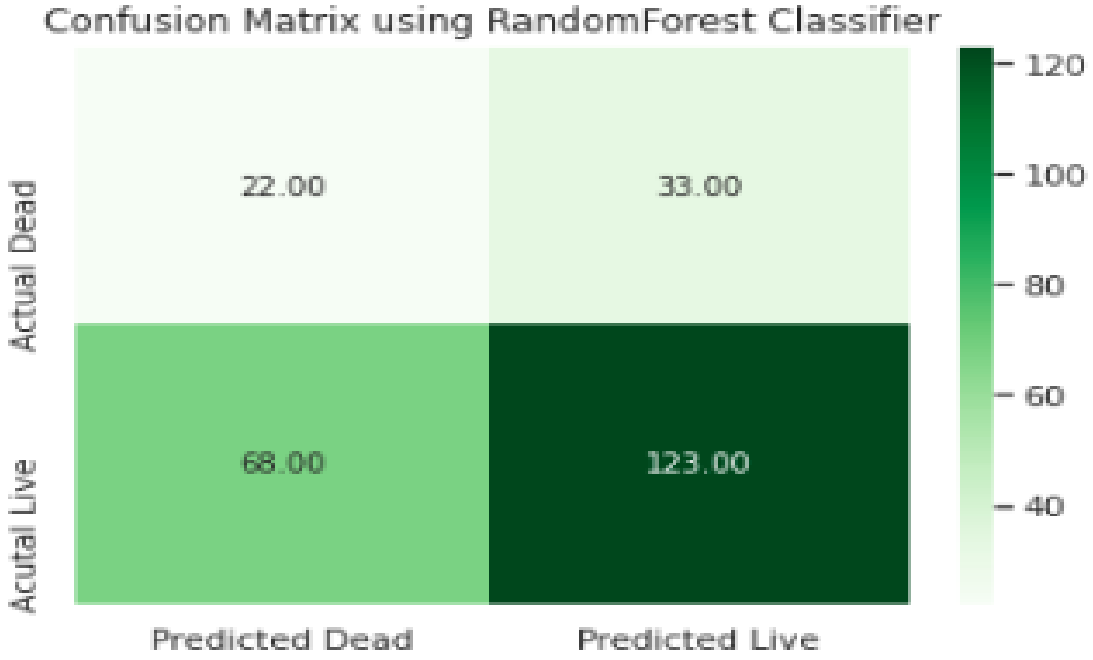
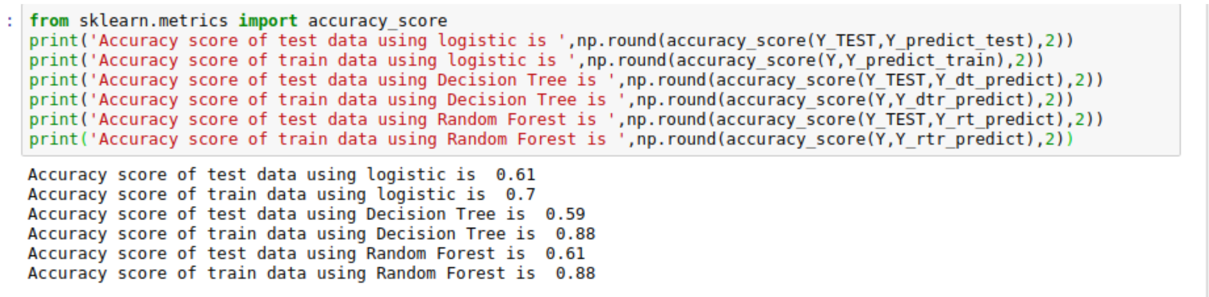


Figure 6.11:CONFUSION MATRIX FROM RANDOM FOREST CLASSIFIER

The below image shows the accuracies obtained from all three machine learning algorithms such as logistic regression,decision tree classifier and random forest classifier thereby giving an effective comparison:

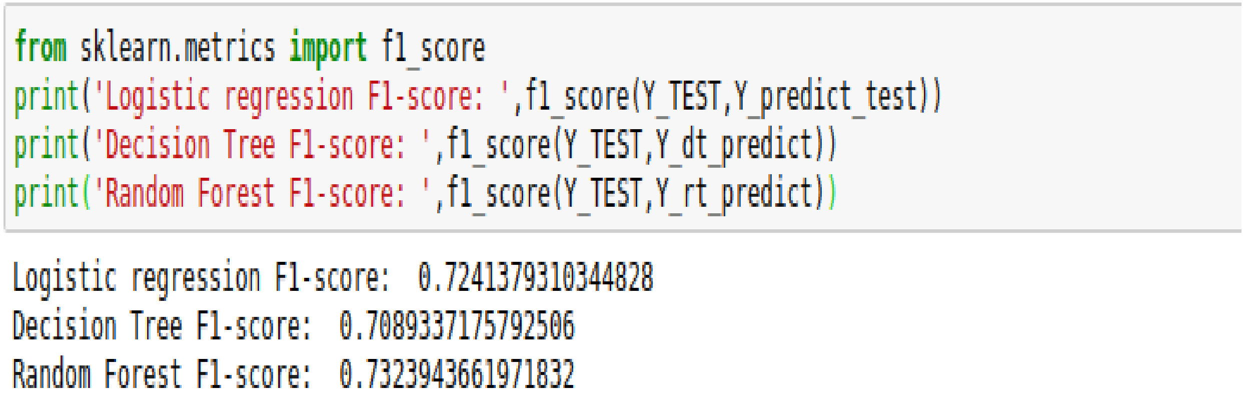


```
: from sklearn.metrics import accuracy_score
  print('Accuracy score of test data using logistic is ',np.round(accuracy_score(Y_TEST,Y_predict_test),2))
  print('Accuracy score of train data using logistic is ',np.round(accuracy_score(Y,Y_predict_train),2))
  print('Accuracy score of test data using Decision Tree is ',np.round(accuracy_score(Y_TEST,Y_dt_predict),2))
  print('Accuracy score of train data using Decision Tree is ',np.round(accuracy_score(Y,Y_dtr_predict),2))
  print('Accuracy score of test data using Random Forest is ',np.round(accuracy_score(Y_TEST,Y_rt_predict),2))
  print('Accuracy score of train data using Random Forest is ',np.round(accuracy_score(Y,Y_rtr_predict),2))

  Accuracy score of test data using logistic is  0.61
  Accuracy score of train data using logistic is  0.7
  Accuracy score of test data using Decision Tree is  0.59
  Accuracy score of train data using Decision Tree is  0.88
  Accuracy score of test data using Random Forest is  0.61
  Accuracy score of train data using Random Forest is  0.88
```

Figure 6.12:ACCURACY GENERATED FROM THE THREE ALGORITHMS

The below image shows the F1 score obtained from all three machine learning algorithms such as logistic regression,decision tree classifier and random forest classifier thereby giving an effective comparison:



```
from sklearn.metrics import f1_score
print('Logistic regression F1-score: ',f1_score(Y_TEST,Y_predict_test))
print('Decision Tree F1-score: ',f1_score(Y_TEST,Y_dt_predict))
print('Random Forest F1-score: ',f1_score(Y_TEST,Y_rt_predict))
```

```
Logistic regression F1-score:  0.7241379310344828
Decision Tree F1-score:  0.7089337175792506
Random Forest F1-score:  0.7323943661971832
```

Figure 6.13: F1 SCORE GENERATED FROM THE THREE ALGORITHMS

The below image shows the output graph obtained from all three training processes:

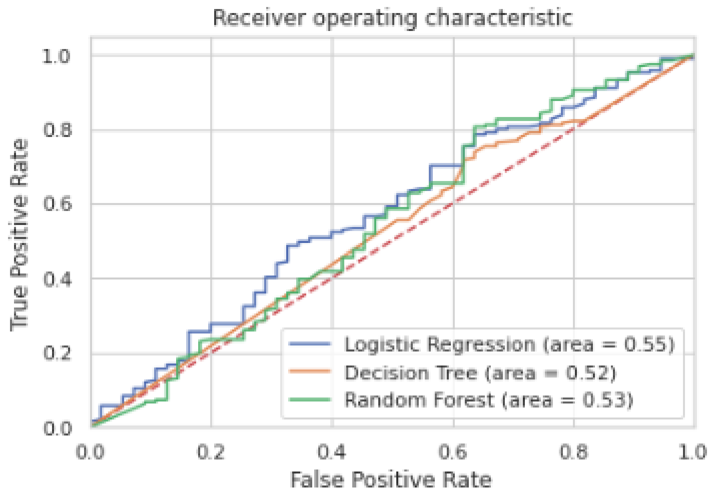


Figure 6.14:OUTPUT GRAPH OBTAINED FROM TRAINING PROCESS

Thus, from the above results and discussion, it is clear that we have efficiently made a project for effectively providing a comparison of training outputs obtained from using three machine learning algorithms. Thus, we have successfully implemented the scope of the project.

CHAPTER – 7

CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

This project is used to find the presence of myocardial infarction and provide prior measures to avoid the disease, giving a comparison of the accuracies achieved from various machine learning algorithms.This also help in providing efficient treatment in a most cheap way and eventually reduce the time required for finding the myocardial infarction in the current state, it is done manually which consumes more time and also involves human error rate. So, reduces the time required for manual classification and eliminates the human error rate by this project.

7.2 FUTURE WORK

In the coming future, we review the application of the myocardial infarction determine technology in the healthcare field and it can promote for detecting various types of cancer with more accuracy. In medical field they are more chance to develop or convert this project in many ways. Thus, this project has an efficient scope in coming future where manual predicting can be converted to computerized production in a cheap way.

REFERENCES

[1] Avijit Sengupta , Kaushik Dutta , Theresa Beckie, and Sriram Chellappan,”Designing a Health Coach-Augmented mHealth System for the Secondary Prevention of Coronary Heart Disease Among Women”[2020]

[2] Beaumon, P. Onoma, M. Rimlinger , D. Broggio, P. Caldeira Ideias and D.Franck,”Age-specific experimental and computational calibration of thyroid in vivo monitoring”,IEEE Transactions on Radiation and Plasma Medical Sciences, Vol: 2829931,2019

[3] Chunyu Wang , Junling Guo, Ning Zhao , Yang Liu , Xiaoyan Liu, Guojun Liu , Maozu Guo,” A Cancer Survival Prediction Method Based on Graph Convolutional Network”,Vol:1536-124.2019

[4] Hui Zhou, Kun Wang, Jie Tian, Member, IEEE,”Online Transfer Learning for Differential Diagnosis of Benign and Malignant Thyroid Nodules with Ultrasound Images”,IEEE Transactions on Biomedical Engineering,Vol.0018-9294,2020.

[5] Jennifer E. Rosen∗, Hyunsuk Suh, Nicholas J. Giordano, Ousama M. A’amar, Eladio Rodriguez-Diaz, Irving I. Bigio, and Stephanie L. Lee,”Preoperative Discrimination of Benign from Malignant Disease in Thyroid Nodules With Indeterminate Cytology Using Elastic Light-Scattering Spectroscopy”,IEEE Transactions On Biomedical Engineering, Vol. 61, No. 8, August 2016.

[6] Jose M. Anton-Rodriguez , Peter Julyan, Ibrahim Djoukhadar, David Russell, D. Gareth Evans, Alan Jackson, and Julian C. Matthews,” Comparison of a Standard Resolution PET-CT Scanner With an HRRT Brain Scanner for Imaging Small Tumors Within the Head”, IEEE Transactions on radiation and plasma medical sciences, vol. 3, no. 4, july 2019

[7] Julien Rouyer, Member, IEEE, Tony Cueva, Student Member, IEEE, Tamy Yamamoto, Alberto Portal, and Roberto Lavarello, Senior Member, IEEE,"In vivo Estimation of Attenuation and Backscatter Coefficients from Human Thyroids",IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,Vol.0885-3010 (c) 2018.

[8] Koyel Mandal, Rosy Sarmah, and Dhruba Kumar Bhattacharyya "Biomarker Identification for Cancer Disease Using Biclustering Approach: An Empirical Study" [Vol: 1545-5963,2018]

[9] Micha Feigin, Daniel Freedman, and Brian W. Anthony," A Deep Learning Framework for Single-Sided Sound Speed Inversion in Medical Ultrasound" 0018-9294 (c) 2016

[10] Nikhil S. Narayan, Pina Marziliano, Jeevendra Kanagalingam, MD and Christopher G.L. Hobbs, MD,"Speckle Patch Similarity for Echogenicity based Multi-Organ Segmentation in Ultrasound Images of the Thyroid Gland",IEEE Journal of Biomedical and Health Informatics,Vol.2168-2194,2016.

[11] Shekoofeh Azizi, Sharareh Bayat, Pingkun Yan, Amir Tahmasebi, Jin Tae Kwak, Sheng Xu, Baris Turkbey, Peter Choyke, Peter Pinto, Bradford Wood, Parvin Mousavi*, Purang Abolmaesumi," Deep Recurrent Neural Networks for Prostate Cancer Detection: Analysis of Temporal Enhanced Ultrasound IEEE Transactions on Medical Imaging "[Vol.no:0278-0062,2018]

[12] Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao,"Multi-task Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition",IEEE Journal of Biomedical and Health Informatics,VOL. 14, NO. 8, AUGUST 2017.

[13] Xiangxiang Zheng ,Guodong Lv, Guoli Du, Zhengang Zhai, Jiaqing Mo and Xiaoyi Lv,"Rapid and Low-Cost Detection of Thyroid Dysfunction Using Raman Spectroscopy

and an Improved Support Vector Machine”,IEEE Photonics Journal, ol. 10, No. 6, December 2017.

[14] Yanbo Wang , Weikang Qian , Bo Yuan,” A graphical model of smoking-induced global instability in lung cancer” [Vol: 1545-5963 (c) 2015]

[15] Yi Wang, Na Wang, Min Xu, Junxiong Yu, Chenchen Qin, Xiao Luo, Xin Yang, Tianfu Wang, Anhua Li, and Dong Ni*,” Deeply-Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound [Vol 0278-0062,2019]

APPENDIX:

```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "## Importing Libraries"
   ]
  },
  {
   "cell_type": "code",
 "execution_count": 2,
 "metadata": {},
 "outputs": [],
 "source": [
  "import numpy as np\n",
  "import pandas as pd\n",
  "import seaborn as sns\n",
  "import matplotlib.pyplot as plt\n",
  "%matplotlib inline\n",
  "sns.set(style=\"white\")\n",
  "sns.set(style=\"whitegrid\", color_codes=True)\n",
  "sns.set(style=\"white\")\n",
  "sns.set(style=\"whitegrid\", color_codes=True)"
 ]
},
{
 "cell_type": "code",
 "execution_count": 3,
 "metadata": {},
 "outputs": [],
  "source": [
   "data = pd.read_csv('dataset.csv',na_values='nk')"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 4,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div>\n",
      "<style scoped>\n",
      "    .dataframe tbody tr th:only-of-type {\n",
      "        vertical-align: middle;\n",
      "    }\n",
      "\n",
      "    .dataframe tbody tr th {\n",
      "        vertical-align: top;\n",
"    }\n",
"\n",
"    .dataframe thead th {\n",
"        text-align: right;\n",
"    }\n",
"</style>\n",
"<table border=\"1\" class=\"dataframe\">\n",
"  <thead>\n",
"    <tr style=\"text-align: right;\">\n",
"      <th></th>\n",
"      <th>Unnamed: 0</th>\n",
"      <th>outcome</th>\n",
"      <th>age</th>\n",
"      <th>yronset</th>\n",
"      <th>premi</th>\n",
"      <th>smstat</th>\n",
"      <th>diabetes</th>\n",
"      <th>highbp</th>\n",
"      <th>hichol</th>\n",
"      <th>angina</th>\n",
"      <th>stroke</th>\n",
"    </tr>\n",
"  </thead>\n",
"  <tbody>\n",
"    <tr>\n",
"      <th>0</th>\n",
"      <td>1</td>\n",
"      <td>live</td>\n",
"      <td>63</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>x</td>\n",
"      <td>n</td>\n",
"      <td>y</td>\n",
"      <td>y</td>\n",
"      <td>n</td>\n",
"      <td>n</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>1</th>\n",
"      <td>6</td>\n",
"      <td>live</td>\n",
"      <td>55</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>c</td>\n",
"      <td>n</td>\n",
"      <td>y</td>\n",
"      <td>y</td>\n",
"      <td>n</td>\n",
"      <td>n</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>2</th>\n",
"      <td>8</td>\n",
"      <td>live</td>\n",
"      <td>68</td>\n",
"      <td>85</td>\n",
"      <td>y</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>y</td>\n",
"      <td>NaN</td>\n",
"      <td>y</td>\n",
"      <td>n</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>3</th>\n",
"      <td>10</td>\n",
"      <td>live</td>\n",
"      <td>64</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>x</td>\n",
"      <td>n</td>\n",
"      <td>y</td>\n",
"      <td>n</td>\n",
"      <td>y</td>\n",
"      <td>n</td>\n",
"    </tr>\n",
 "    <tr>\n",
 "      <th>4</th>\n",
 "      <td>11</td>\n",
 "      <td>dead</td>\n",
 "      <td>67</td>\n",
 "      <td>85</td>\n",
 "      <td>n</td>\n",
 "      <td>NaN</td>\n",
 "      <td>NaN</td>\n",
 "      <td>NaN</td>\n",
 "      <td>NaN</td>\n",
 "      <td>NaN</td>\n",
 "      <td>NaN</td>\n",
 "    </tr>\n",
 "  </tbody>\n",
 "</table>\n",
 "</div>"
],
"text/plain": [
 "   Unnamed: 0 outcome  age  yronset premi smstat diabetes highbp hichol  \\\n",
     "0          1    live   63       85     n      x        n       y       y   \n",
     "1          6    live   55       85     n      c        n       y       y   \n",
     "2          8    live   68       85     y    NaN      NaN      y    NaN   \n",
     "3         10    live   64       85     n      x        n       y       n   \n",
     "4         11    dead   67       85     n    NaN      NaN    NaN    NaN   \n",
     "\n",
     "  angina stroke  \n",
     "0      n      n  \n",
     "1      n      n  \n",
     "2      y      n  \n",
     "3      y      n  \n",
     "4    NaN    NaN  "
    ]
   },
   "execution_count": 4,
   "metadata": {},
   "output_type": "execute_result"
  }
 ],
 "source": [
  "data.head(5)"
 ]
},
{
 "cell_type": "markdown",
 "metadata": {},
 "source": [
  "# Data Analysis"
 ]
},
{
 "cell_type": "code",
 "execution_count": 5,
 "metadata": {},
 "outputs": [],
 "source": [
  "data = data.drop(labels='Unnamed: 0',axis=1)"
 ]
},
{
```

```
"cell_type": "code",
"execution_count": 6,
"metadata": {},
"outputs": [
 {
  "name": "stdout",
  "output_type": "stream",
  "text": [
   "Number of live people in our data are 75.2123552123552%  \n",
   "Number of dead people in our data are 24.787644787644787%  \n"
  ]
 }
],
"source": [
 "data_live = len(data[data['outcome']=='live'])\n",
 "data_dead = len(data[data['outcome']=='dead'])\n",
 "data_live_p = (data_live/(data_dead+data_live))*100\n",
 "data_dead_p = (data_dead/(data_dead+data_live))*100\n",
 "print('Number of live people in our data are {}%  '.format((data_live_p)))\n",
 "print('Number of dead people in our data are {}%  '.format((data_dead_p)))"
```

```
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 7,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "<class 'pandas.core.frame.DataFrame'>\n",
     "RangeIndex: 1295 entries, 0 to 1294\n",
     "Data columns (total 10 columns):\n",
     " #   Column    Non-Null Count  Dtype \n",
     "---  ------    --------------  ----- \n",
     " 0   outcome   1295 non-null   object\n",
     " 1   age       1295 non-null   int64 \n",
     " 2   yronset   1295 non-null   int64 \n",
     " 3   premi     1239 non-null   object\n",
    " 4   smstat    1192 non-null   object\n",
    " 5   diabetes  1226 non-null   object\n",
    " 6   highbp    1219 non-null   object\n",
    " 7   hichol    1107 non-null   object\n",
    " 8   angina    1196 non-null   object\n",
    " 9   stroke    1216 non-null   object\n",
    "dtypes: int64(2), object(8)\n",
    "memory usage: 101.3+ KB\n"
   ]
  }
 ],
 "source": [
  "data.info()"
 ]
},
{
 "cell_type": "code",
 "execution_count": 8,
 "metadata": {},
 "outputs": [
{
 "data": {
  "text/html": [
   "<div>\n",
   "<style scoped>\n",
   "    .dataframe tbody tr th:only-of-type {\n",
   "        vertical-align: middle;\n",
   "    }\n",
   "\n",
   "    .dataframe tbody tr th {\n",
   "        vertical-align: top;\n",
   "    }\n",
   "\n",
   "    .dataframe thead th {\n",
   "        text-align: right;\n",
   "    }\n",
   "</style>\n",
   "<table border=\"1\" class=\"dataframe\">\n",
   "  <thead>\n",
   "    <tr style=\"text-align: right;\">\n",
"      <th></th>\n",
"      <th>outcome</th>\n",
"      <th>age</th>\n",
"      <th>yronset</th>\n",
"      <th>premi</th>\n",
"      <th>smstat</th>\n",
"      <th>diabetes</th>\n",
"      <th>highbp</th>\n",
"      <th>hichol</th>\n",
"      <th>angina</th>\n",
"      <th>stroke</th>\n",
"    </tr>\n",
"  </thead>\n",
"  <tbody>\n",
"    <tr>\n",
"      <th>4</th>\n",
"      <td>dead</td>\n",
"      <td>67</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>5</th>\n",
"      <td>live</td>\n",
"      <td>66</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>x</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>8</th>\n",
"      <td>dead</td>\n",
"      <td>46</td>\n",
"      <td>85</td>\n",
"      <td>n</td>\n",
"      <td>c</td>\n",
"      <td>n</td>\n",
"      <td>y</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>n</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>27</th>\n",
"      <td>dead</td>\n",
"      <td>64</td>\n",
"      <td>85</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>29</th>\n",
"      <td>dead</td>\n",
"      <td>69</td>\n",
"      <td>85</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"      <td>NaN</td>\n",
"    </tr>\n",
"  </tbody>\n",
    "</table>\n",
    "</div>"
   ],
   "text/plain": [
    "   outcome  age  yronset premi smstat diabetes highbp hichol angina stroke\n",
    "4     dead   67      85     n    NaN      NaN    NaN    NaN    NaN    NaN\n",
    "5     live   66      85     n      x      NaN    NaN    NaN    NaN    NaN\n",
    "8     dead   46      85     n      c        n      y    NaN    NaN      n\n",
    "27    dead   64      85   NaN    NaN      NaN    NaN    NaN    NaN    NaN\n",
    "29    dead   69      85   NaN    NaN      NaN    NaN    NaN    NaN    NaN"
   ]
  },
  "execution_count": 8,
  "metadata": {},
  "output_type": "execute_result"
 }
],
"source": [
 "data[data['angina'].isna()].head()"
]
```

```
},
{
 "cell_type": "code",
 "execution_count": 9,
 "metadata": {},
 "outputs": [],
 "source": [
  "data.dropna(thresh=4,inplace=True)\n",
  "data.dropna(thresh=5,inplace=True)\n",
  "data.dropna(thresh=6,inplace=True)\n",
  "data.dropna(thresh=7,inplace=True)"
 ]
},
{
 "cell_type": "code",
 "execution_count": 10,
 "metadata": {},
 "outputs": [
  {
   "data": {
```

```
"text/html": [
 "<div>\n",
 "<style scoped>\n",
 "    .dataframe tbody tr th:only-of-type {\n",
 "        vertical-align: middle;\n",
 "    }\n",
 "\n",
 "    .dataframe tbody tr th {\n",
 "        vertical-align: top;\n",
 "    }\n",
 "\n",
 "    .dataframe thead th {\n",
 "        text-align: right;\n",
 "    }\n",
 "</style>\n",
 "<table border=\"1\" class=\"dataframe\">\n",
 "  <thead>\n",
 "    <tr style=\"text-align: right;\">\n",
 "      <th></th>\n",
 "      <th>age</th>\n",
"      <th>yronset</th>\n",
"    </tr>\n",
"  </thead>\n",
"  <tbody>\n",
"    <tr>\n",
"      <th>count</th>\n",
"      <td>1229.000000</td>\n",
"      <td>1229.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>mean</th>\n",
"      <td>60.866558</td>\n",
"      <td>88.829129</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>std</th>\n",
"      <td>7.061855</td>\n",
"      <td>2.541531</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>min</th>\n",
"      <td>35.000000</td>\n",
"      <td>85.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>25%</th>\n",
"      <td>57.000000</td>\n",
"      <td>87.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>50%</th>\n",
"      <td>63.000000</td>\n",
"      <td>89.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>75%</th>\n",
"      <td>66.000000</td>\n",
"      <td>91.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
   "      <th>max</th>\n",
   "      <td>69.000000</td>\n",
   "      <td>93.000000</td>\n",
   "    </tr>\n",
   "  </tbody>\n",
   "</table>\n",
   "</div>"
  ],
  "text/plain": [
   "               age      yronset\n",
   "count  1229.000000  1229.000000\n",
   "mean     60.866558    88.829129\n",
   "std       7.061855     2.541531\n",
   "min      35.000000    85.000000\n",
   "25%      57.000000    87.000000\n",
   "50%      63.000000    89.000000\n",
   "75%      66.000000    91.000000\n",
   "max      69.000000    93.000000"
  ]
 },
   "execution_count": 10,
   "metadata": {},
   "output_type": "execute_result"
  }
 ],
 "source": [
  "data.describe()"
 ]
},
{
 "cell_type": "markdown",
 "metadata": {},
 "source": [
  "# Data visualisation"
 ]
},
{
 "cell_type": "code",
 "execution_count": 11,
 "metadata": {},
"outputs": [
 {
  "data": {
   "text/plain": [
    "Text(0, 0.5, 'Frequency')"
   ]
  },
  "execution_count": 11,
  "metadata": {},
  "output_type": "execute_result"
 },
```