GROUP 18

DISEASE PREDICTION WEBSITE

A

Project Report

Submitted for the partial fulfillment

of B.Tech Degree

in

COMPUTER SCIENCE & ENGINEERING

By

Shivansh Verma(1805210050)

Yashika Tyagi(1805210065)

Shubhi Agarwal(1900520109005)

Under the Supervision of

Dr Manish Gaur

Dr Jasvant Kumar

Department of Computer Science and Engineering

Institute of Engineering and Technology , Lucknow

Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh

Content

Declaration

We hereby solemnly declare that the report titled **“Disease Prediction Website”** is prepared and completed by us under the supervision and guidance of Dr. Manish Gaur and Dr. Jasvant Kumar, it contains no material hitherto published by some other person which to a substantial error has been accepted by any university.

We also confirm that the report is only prepared for our academic requirement, not for any other purpose. I hereby warrant that the work we have presented does not breach any existing copyright.

Submitted by: - Date: 26th May 2022

1. Name: Shivansh Verma
   Roll No.:1805210050
   Branch: CSE

   Signature:

2. Name: Yashika Tyagi
   Roll No.:1805210065
   Branch: CSE

   Signature:

3. Name: Shubhi Agarwal
   Roll No.:1900520109005
   Branch: CSE

   Signature:

Certificate

This is to certify that the project report entitled “Disease Prediction Website” presented by Shivansh Verma, Yashika Tyagi and Shubhi Agarwal in the partial fulfillment for the award of Bachelor of Technology in Computer Science and Engineering, is a record of work carried out by them under my supervision and guidance at the Department of Computer Science and Engineering at Institute of Engineering and Technology, Lucknow.

It is also certified that this project has not been submitted at any other Institute for the awardof any other degrees to the best of my knowledge.

Dr. Manish Gaur

Department of Computer Science and Engineering

Institute of Engineering and Technology, Lucknow

Dr Jasvant Kumar

Department of Computer Science and Engineering

Institute of Engineering and Technology, Lucknow

Acknowledgement

This work would not have been possible without the expert support of Dr. Manish Gaur and Dr. Jasvant Kumar. We are especially indebted to Dr. Pawan Kumar Tiwari who have been supportive of our project goals and who worked actively to provide us with the protected academic time to pursue those goals. We are grateful to all of those with whom we have had the pleasure to work during this projects. Each of the members of our CSE department has provided us extensive personal and professional guidance and taught us a great deal about both scientific research and life in general. We would especially like to thank Dr. D.S Yadav, the HOD of CSE department ,as our teacher and mentor, he has taught us more than we could ever give him credit for . He has shown us, by his example, what a good teacher (and person) should be.

Nobody has been more important to us in the pursuit of this project than the members of our respective families. We would like to thank our parents, whose love and guidance are with us in whatever we pursue. They are the ultimate role models.

Abstract

There has been an unprecedented increase in the incidence of diabetes and skin diseases worldwide.

Diabetes is one of the most common non-communicable diseases that endangers human health. It has become a major health problem worldwide. According to a study by the World Health Organization (WHO), the number of diabetic patients will rise to 552 million by 2030, which means that by 2030, 1 in 10 adults will be diabetic. 9% of adults aged 18+. The machine-based learning method is suggested for setting, starting point, and diabetes forecast. Mechanical learning stages have been used to diagnose diabetes, with Random Forest (RF), Vector Machine (SVM), Naive bayes (NB).

Dermatitis among humans has become a common disease, with millions suffering from various skin ailments. Often, these diseases have subtle dangers that not only lead to insecurity and depression but also lead to increased risk of skin cancer. Medical professionals and advanced instruments are needed to diagnose these skin diseases due to the lack of a visible solution in the images of skin diseases. The proposed framework incorporates in-depth learning strategies such as CNN architecture and a predefined models called VGG16. Dataset for seven diseases pictures has been taken to classify skin diseases which includes diseases such as Acne, Hair Loss, Melanoma, Poison Ivy and other contact diseases ,Nail fungus ,Vitiligo, Warts Molluscum and other viral infections. The database was expanded by adding cut-and-burn images, classified as skin diseases by most existing systems. The use of Deep Learning algorithms has reduced the need for human activity, such as extracting features and rebuilding data for segmentation purposes.

The purpose of the project is to create early awareness about diseases and timely treatment of problems which might later prove to be fatal. The problem predicts diabetes chances based on symptoms and skin disease based on photo uploaded.

The project can be further extended to include more disease prediction to reach more audience.

List of Figures

List of Tables

Chapter 1

Introduction

The skin is one of the most important and fastest growing tissues in the human body. The burden of dermatology is regarded as a multidisciplinary concept that understands the psychological, social and economic significance of dermatitis in patients and their homes and communities. It is a filth that happens to people of all ages. The skin often breaks down because it affects the body. There are more than 3000 skin diseases. A good-looking spoiler disease will have a major impact and can cause severe pain and chronic damage. Many chronic skin conditions, as well as atopic eczema, psoriasis, vitiligo and leg ulcers, which are not fatal at the moment, can be found as a major skin problem that includes physical, emotional and economic consequences. On the other hand, skin cancer can be dangerous and its problem associated with our time. One of the most common diseases among people worldwide is dermatitis. Basal cell carcinoma (BCC), melanoma, intraepithelial carcinoma, and squamous cell carcinoma are examples of skin cancer (SCC). The incidence of skin cancer is currently much higher than other new types of lung and breast cancer. Many skin diseases have symptoms that can take a long time to heal as they may be months away. As a result, computer-based diagnostics are initiated because they can produce results in less time with greater accuracy than human analysis using laboratory procedures. In-depth study is the most widely used technology in diagnosing skin diseases. In-depth learning models will use targeted data to identify and evaluate features in unspecified data patterns, leading to significant performance even for low-cost calculation models. This study provides a solid way to accurately diagnose skin diseases using diagnostic methods that reduce the cost of diagnosis. This has led researchers to consider the use of an in-depth research model to differentiate skin diseases based on the image of the affected region.

Diabetes is a common complaint and is often diagnosed by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic disorders when a person has low blood sugar, either due to an insulin-induced insulin or immune system. they do not turn right back. on insulin, or both. This increases the concentration of glucose in the blood. Most cases of diabetes can be classified into two categories, type 1 and type 2, although some conditions are more difficult to differentiate.Many problems arise when diabetes is left untreated. Therefore, it not only complains but also creates colorful conditions such as heart disease, blindness, kidney complaints, etc. Diabetes has been one of the leading causes of complaints and deaths on behalf of society in the largest countries. According to a report by the International Diabetes Federation, this figure is expected to rise to 642 million by 2040, so early detection and the perception of people with diabetes are critical to diagnosing and treating diabetes. Diabetes data analysis is challenging because physician data are unstructured, inconsistent, built-in, and complex in nature. The use of machine learning techniques in diabetes testing is an important way to use large amounts of diabetes-related information to capture information. It also helps people to diagnose diabetes directly. The purpose of this study was to compare the performance analysis of Random Forest (RF) models, Support Vector Machine (SVM), Naive bayes (NB) for the classification of diabetes mellitus.

1.1 Purpose of the project

Researchers have concluded that in-depth study algorithms are effective in diagnosing skin diseases. The aim of the study was to use a deep neural network algorithm to differentiate common skin diseases and predict diabetes. Researchers have developed an algorithm from the Google Net Inception V3 package. They adjust the storage layer to add their data sets using transfer read. It had promising results with 86.54% accuracy using the database. We suggest that we examine the aforementioned skin diseases by harvesting images found on professional and publicly accessible websites, the atlas of dermatology photographs and taking them personally and classifying each image in the appropriate category of dermatology by transmitting the VGG16 study model.

We also aim to learn to predict diabetes using machine learning strategies such as SVM, Naïve Bayes etc. Symptom-related questions are asked on the web app to distinguish people with diabetes and non-diabetics.

Typically, this study aims to design a system for predicting skin and diabetes predictors in web applications that will differentiate between different diseases using a pre-trained neural network network model and a Machine Learning model in the specified data field.

Chapter 2

Literature Review

There is a problem with the change in the rate of error rate in the database, caused by a change in the size of the data used in different skin cancer tests. Therefore, a lack of a standard database can lead to serious problems; average error values are considered in most tests. In addition, data collection of multiple studies depends on each study, resulting in unnecessary effort and time. When the actual class is marked by hand and compared with the class predicted to calculate the matric of one parameter, the pixels are lost when the background is cut into a skin cancer image using Adobe Photoshop . At this point, the process influences the outcomes of all the faithful parameter groups (matriculants, relationships, and behaviors), which are considered controversial. High reliability and low level of complexity cannot be achieved simultaneously, which is reflected in the training program and is influenced by conflicts between different levels, leading to major challenges . The mechanism used for the detection of one skin lesion may not work to detect others . Many different courses and test sets have been used to evaluate the proposed methods. In addition, in the training and testing parameters, different researchers are interested in different areas. This lack of consistency in all papers makes the correct comparison impossible at all . Although these references in the literature have been widely criticized, studies continue to use them to evaluate the application to skin cancer andother imaging fields.

The data used for testing is usually very small to allow a credible statement about system performance to be made. While it is not possible to collect a fair amount of relevant data online at this age of information, this information, with significant uncertainty, obviously cannot meet the requirements of independent and uniform distribution, which is one of the essential requirements for in-depth learning to be. used successfully. For certain rare and minor ailments, only a limited number of photographs are available for training. To date, a large number of algorithms have been shown to discriminate against small groups, which could lead to a significant gap in the health service between "the rich" and the "poor". Many aspects of the training process are required using in-depth learning strategies. In addition, although the in-depth learning process has been used successfully in other professions, skin-enhanced models work only for specific dedicated diseases and do not work under normal conditions. Diagnosis of dermatology is a complex process, in addition to image recognition, it should be supplemented with other techniques such as blushing, sniffing, temperature changes, and microscope.

We have used pre trained VGG16 model .The model’s last three dense layers are customized according to our classification of diseases. The website is simple to use and increases user interactivity.

In Jobeda Jamal Khanam's research, the Indian Diabetes Pima (PID) dataset is collected by the UCI Machine Learning Repository, which is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). In the PID dataset, all patients are female and are at least 21 years old. The dataset contains information on 768 patients and their corresponding nine unique attributes. The nine attributes used for diabetes prediction are Pregnancy, BMI, Insulin Level, Age, Blood Pressure, Skin Thickness, Glucose, Diabetes Genealogical Function, and Outcome. The "result" attribute is treated as a dependent or target variable and the remaining eight attributes are treated as independent variables / characteristics. The diabetes "result" attribute consists of a binary value where 0 means non-diabetes and 1 implies diabetes.

In our research, we used data mining and machine learning algorithms to predict whether a patient has diabetes or not with enhanced accuracy by asking questions with the user based on their symptoms. The dataset used is from a hospital of Bangladesh in which team of doctors asked people about their symptoms and also mentioned their category of diabetes or not via test. The dataset seemed perfect for application of machine learning algorithms ,hence we applied random forest to predict the category of user.

.

Chapter 3

Methodology

3.1 Problem Definition

Diabetes and skin diseases are among the most neglected/poorly diagnosed diseases prevalent in majority population of the globe around.

Hence we have built a website using which people showing probable symptoms can at least be made aware about their present health status.

3.2 Methodology for skin disease detection

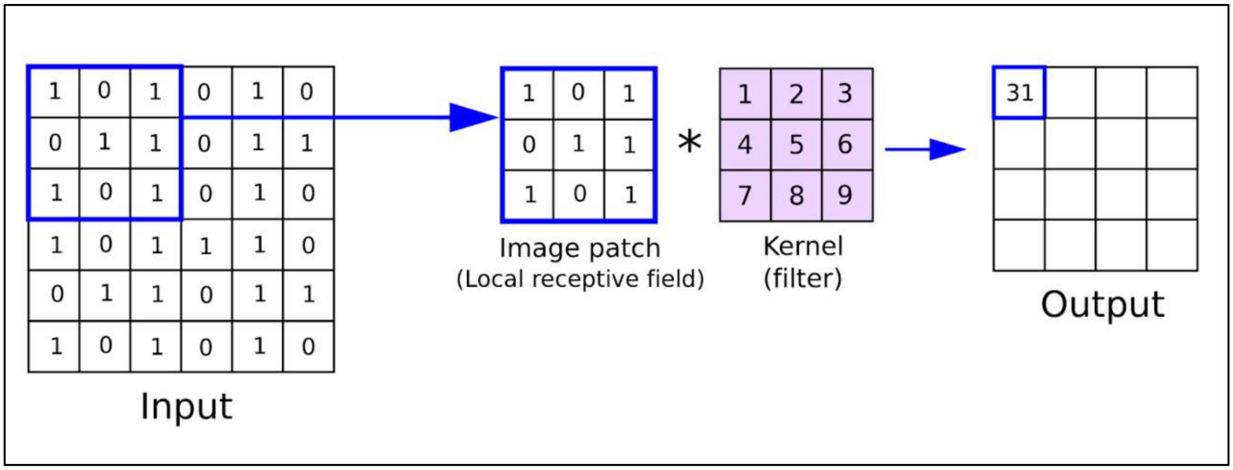
3.2.1 Classification:

The CNNs structure is based on the data we provide in our test we train and test CNN using a Dermnet database which is modified by us using images from personal lives and internet containing only 8 classes randomly selected 2500 of them as a training set and 1000 of them as pre-training confirmation using image.net pretrained VGG16 models.
After training use ImageNet trained models VGG16 to ensure-accuracy was found at more than 84%.

3.2.2 CNN:

In deep learning, the convolutional neural network (CNN, or ConvNet) is a component of the Artificial neural network (ANN), and is widely used in image and image analysis. They have many applications in image and video recognition, complimentary programs, photo classification, image classification, medical image analysis, natural language processing, brain and computer communication and time series.

CNN is a variant of the Multilayer Perceptron. Multilayer perceptrons are fully connected networks, i.e., each neuron in a single layer is connected to all neurons in the next layer. The "full connection" of these networks makes them prone to data overload. Common ways to do so, or to prevent excessive immersion, include: punitive restrictions during training (such as weight loss) or interconnection (skipping links, dropping out of school, etc.) CNN takes a different approach: they take advantage of hierarchical pattern in data . and incorporates complex growth patterns using small and simple patterns labeled in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.



1	0	1	0	1	0
0	1	1	0	1	1
1	0	1	0	1	0
1	0	1	1	1	0
0	1	1	0	1	1
1	0	1	0	1	0

1	0	1
0	1	1
1	0	1

1	2	3
4	5	6
7	8	9

31			

Figure 3.1 CNN Explanation

Strides defines the movement of the sludge; if you set stride = 1, which is the dereliction, the kernel takes one step at a time. Commonly, the sludge size is decrease than the input facts, and the sort of addition applied between the sludge and the enter records pattern of the sludge size is the fleck product. A fleck product is an element-sensible addition between the sludge weights and the sludge size pattern of the input facts, blended into a single cost. The sludge length is designed lower than the enter information as it permits the identical set of sludge weights to be multiplied multiple times by the enter matrix at specific locales inside the photograph.

.

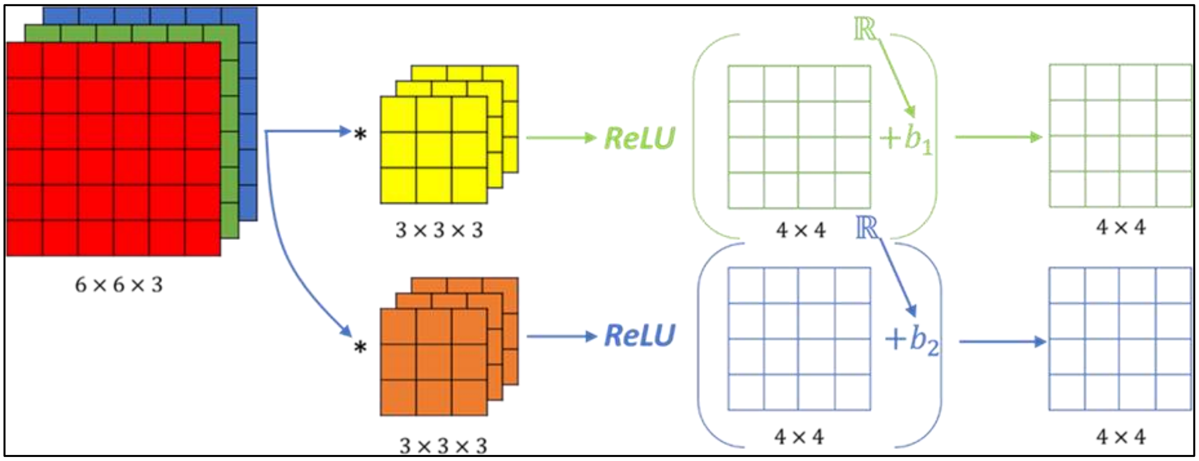


Figure 3.2 Stride

3.2.3 Maxpooling:

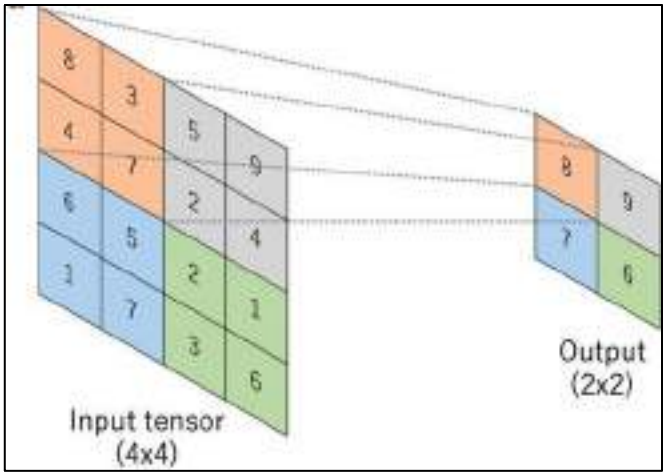


Figure 3.3 MaxPolling

Max pooling is a complication approach wherein the kernel excerpts the maximum figure from the range it's country miles convolving. Max Pooling is absolutely telling the Convolutional Neural community that we're suitable to handiest transmit this statistics indeed as this is the widest statistics breadth available. Maxpooling on a 4 * 4 channel with a 2 * 2 kernel and a step of 2 How we fold with a 2 * 2kernel.However, the channel has four values 8, 3, If we test the number one 2 * 2 set that the kernel focuses on. MaxPooling selects the most figure of this pool, which is"8". Right then, in our environment, we're suitable to produce a kernel that amplifies the cat s eye image to such an extent that indeed after maximum pooling, the dominant statistics isn't lost. Now whilst Max Pooling flowers my pixels, 25 of the remaining pixels is enough to keep the statistics about the cat. So there can be a channel or function chart as a way to comprise the cat's eye statistics, anyhow of what, to lessen pixels to seventy five. In another case, we're suitable to say that we're giving away statistics we don't need via constructing kernels that make it doable to gain the critical statistics through Max Pooling.

3.2.4 Residual Block:

In order to solve the problem of the vanishing/exploding gradient, this architecture introduced the concept called Residual Network.

In this network we use a technique called *skip connections* **.** The skip connection skips training from a few layers and connects directly to the output. The approach behind this network is instead of layers learn the underlying mapping, we allow network fit the residual mapping. So, instead of say H(x), initial mapping, let the network fit, $F(x) := H(x) – x$ which gives

$H(x) := F(x) + x$.

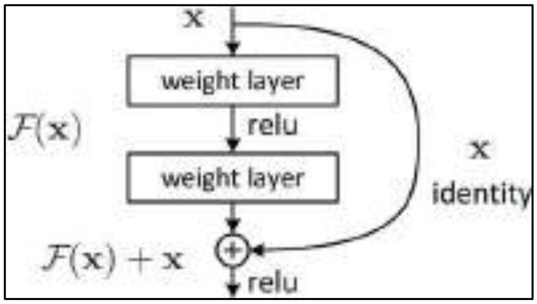


Figure 3.4 Residual Block

3.2.5 VGG16:

VGG stands for Visual Geometry Group; it is a standard deep Convolutional Neural Network (CNN) architecture with multiple layers. The "deep" refers to the number of layers with VGG-16 or VGG-19 consisting of 16 and 19 convolutional layers.

The VGG architecture is the basis of ground-breaking object recognition models. Developed as a deep neural network, the VGGNet also surpasses baselines on many tasks and datasets beyond ImageNet. Moreover, it is now still one of the most popular image recognition architectures.

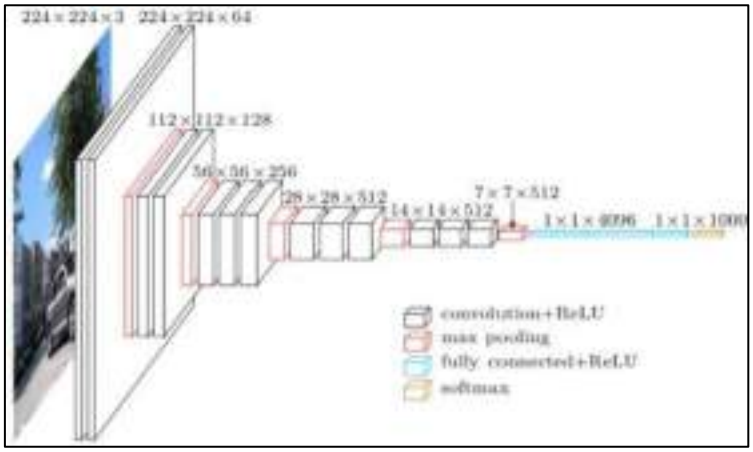


Figure 3.5 VGG16 Architecture

The accuracy of VGG16 on our dataset is 86%. VGG16 is CNN model with 13 convolutional layers and 3 fully connected layers . VGG16 is already trained on imagenet dataset. We have used pretrained VGG16 model via transfer learning . Last layers of our specific problem is added and model is then used to predict the input into 8 classes.

We can fine tune the parameters of the layers to get the desired output.

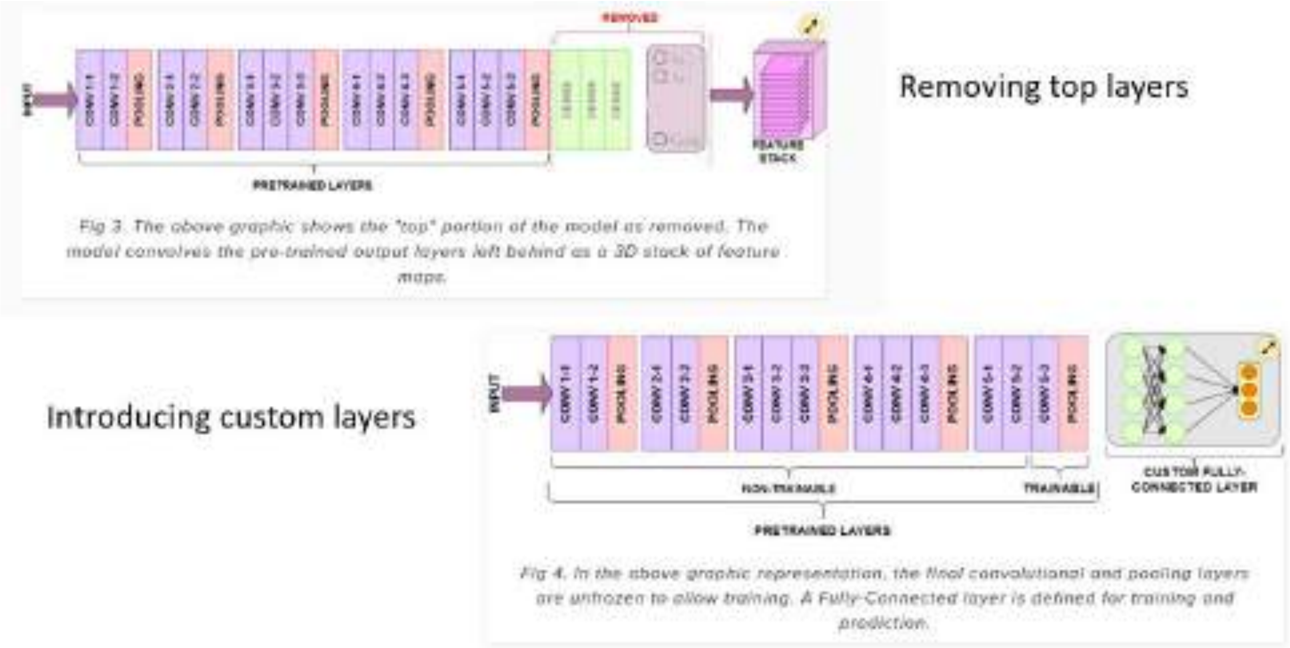


Fig 3. The above graphic shows the "top" portion of the model as removed. The model convolves the pre-trained output layers left behind as a 3D stack of feature maps.

Fig 4. In the above graphic representation, the final convolutional and pooling layers are unfrozen to allow training. A Fully-Connected layer is defined for training and prediction.

Figure 3.6 VGG Customization

Code

```
for layer in vgg_model.layers[:15]:
layer.trainable = False
x = vgg_model.output
x = Flatten()(x) # Flatten dimensions to for use in FC layers
x = Dense(512, activation='relu')(x)
x = Dropout(0.5)(x) # Dropout layer to reduce overfitting
x = Dense(256, activation='relu')(x)
x = Dense(8, activation='softmax')(x) # Softmax for multiclass
transfer_model = Model(inputs=vgg_model.input, outputs=x)
for i, layer in enumerate(transfer_model.layers):
print(i, layer.name, layer.trainable)
```

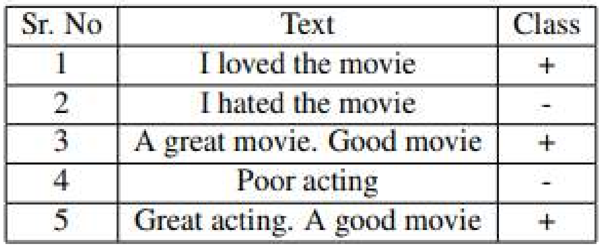
3.3 Methodology for diabetes prediction:

The dataset was tried on three different algorithms and the algorithms with most accurate results is selected to predict the final answer. Naïve Bayes was able to achieve 86% of accuracy but Random forest attained maximum accuracy of 94%.

3.3.1 Naive Bayes:

Naive bayes is wide utilized ai classifier and probabilistic calculation essential uses of naive bayes region unit to channel spam order archives and so on the component feed into the model is independent of each unique that is renascent the value of any of the other component utilized inside the calculation naive bayes enjoys vital benefit is that we've an adapted to confront measure prepared to coded up to foresee the yield ongoing speedy it's only climbable and old calculation is also a most reasonable alternative for planet applications that region unit needed to answer to client as by and by as feasible Let's take an Example: You have a set of reviews and classification

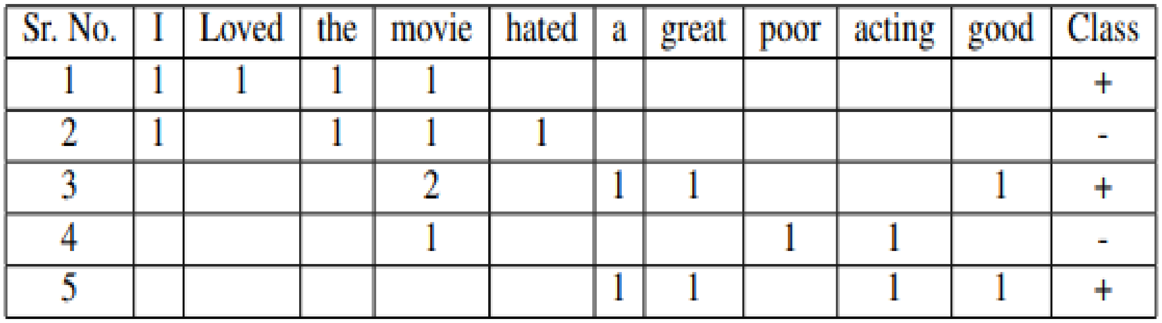
Table 3.1: Naive bayes



Sr. No	Text	Class
1	I loved the movie	+
2	I hated the movie	-
3	A great movie. Good movie	+
4	Poor acting	-
5	Great acting. A good movie	+

Above table define movie review with sentiment data. In on top of table there's a text column that is input and there are 10 unique words that are: - "I, loved, the, movie, hated, a, great, poor, acting, good". categories contain the sentiment information that's negative and positive. which define that movie review is negative or positive based on 10 unique words. Then we've got to convert above data table into features and based on that we tend to get sentiment output. 1st we have to convert it into matrix type and also have to find how many times has that word come back. In table 3.2 unique words are comeback that is "I" word is continual in initial and second review then "Loved" word is repeated in exactly initial review and so on. in class column there are total 5 categories that's mixture of positive and negative classes in this there are 3 positive categories and a couple of negative categories.

Table 3.2: Naive bayes



Sr. No.	I	Loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							+
2	1		1	1	1						-
3				2		1	1			1	+
4				1				1	1		-
5						1	1		1	1	+

Then we've to count all the positive unique words. Then we've to calculate the likelihood against positive category therefore the probability is 3/5. Then we computing p(I) before that there's a formula that we've to refer that'sWe add 1 in each probability thus the chance, like P(class | text) can never be zero. we are trying to determine if a data row should be classified as negative or positive. due to these we are able to ignore the divisor. therefore we have to calculate the probabilities of every classification and therefore the probabilities of every feature falling into each classification. now we have to place values in our equation In P(I | +) nk 1 that is I is occurred in initial document just once. This method is comparable for negative text also. however vocabulary count is constant in both the cases. Now we have to train our classifier, for that there's one formula that is Vnb = argmax(summation of all the words occur in sentence) That is, suppose there's a sentence like "I hated the poor acting" then 1st we've to classify all words that is therein sentence and then we have to get the likelihood to get the class that is positive or negative. thus we grab all distinctive words and then place it in above equation.

If Vj = +; p(+)P(I|+) P(I|+) P(I|+) P(I|+) P(I|+) = 6.03x10-7

If Vj = -; p(-)P(I|-) P(I|-) P(I|-) P(I|-) P(I|-) = 1.22x10-5

So get the probability of having this sentence in positive and negative. For positive classification we get 6.03x10-7 and for negative classification we get 1.22x10-5.Now, we have to determine whether or not sentence is classed into positive or negative. so, the answer is negative. because of 10-5. In negative values the min negative range is greater than the most negative number so that sentence get classified into negative class.

3.3.2 Random Forest:

Each decision tree has a high variance, but when we combine them all together in parallel, the resulting variance is low as each decision tree is trained perfectly on that particular given sample and therefore the output does not depend on one decision tree but on multiple decision trees . A random forest is an ensemble technique that can perform both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.

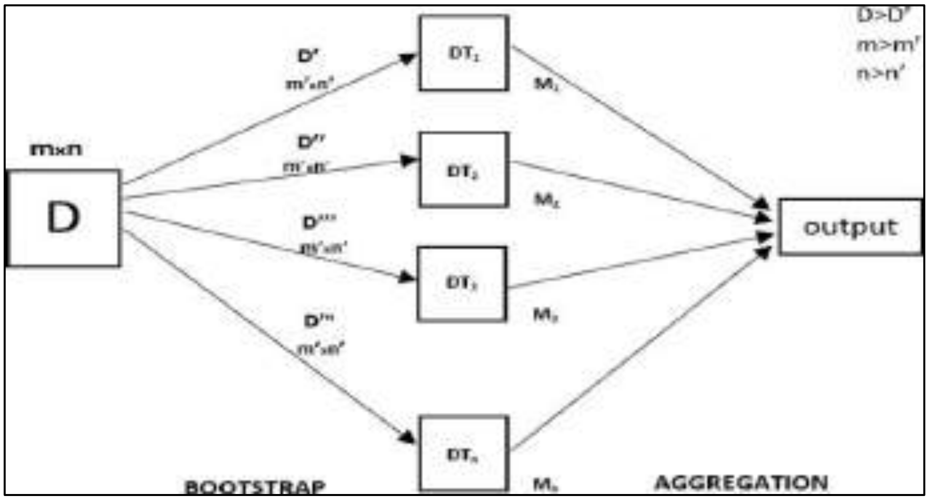


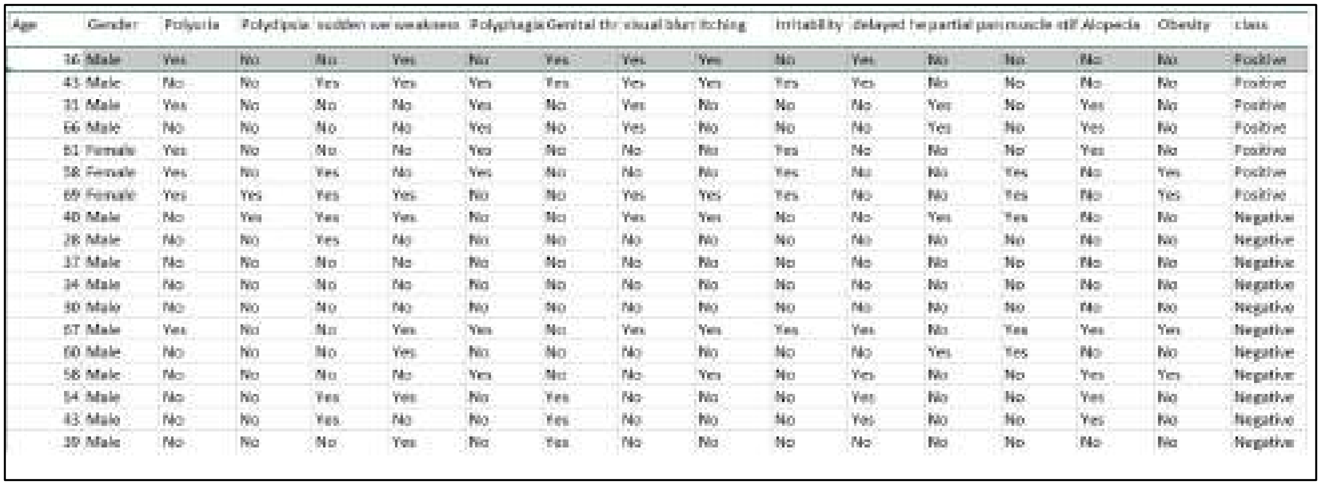
Figure 3.7 Random Forest

The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

The accuracy which was achieved in Random Forest was 94%. Random forest divides the dataset into small datasets and the apply decision tree algorithm on each dataset and then takes the answer which is predicted by most of the trees. The diabetic dataset has binary classes for each symptom whether yes or no. We have total 14 attribute in our dataset.

Decision tree algo calculates information gain for each attribute and takes the maximum as the root node. The algorithm is quite efficient and is easy to implement in Python.



Age	Gender	Polyuria	Polydipsia	sudden we	weakness	Polyphagia	Genital thr	visual blurr	Itching	Irritability	delayed he	partial par	muscle stif	Alopecia	Obesity	class
36	Male	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	No	Positive
43	Male	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Positive
33	Male	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Positive
66	Male	No	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Positive
61	Female	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	Positive
58	Female	Yes	No	Yes	No	Yes	No	No	No	Yes	No	No	Yes	No	Yes	Positive
69	Female	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Positive
40	Male	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Negative
28	Male	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Negative
37	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
34	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
30	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative
67	Male	Yes	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Negative
60	Male	No	No	No	Yes	No	No	No	No	No	No	Yes	Yes	No	No	Negative
56	Male	No	No	No	No	Yes	No	No	Yes	No	Yes	No	No	Yes	Yes	Negative
54	Male	No	No	Yes	Yes	No	Yes	No	No	No	Yes	No	No	Yes	No	Negative
43	Male	No	No	Yes	No	No	Yes	No	No	No	Yes	No	No	Yes	No	Negative
39	Male	No	No	No	Yes	No	Yes	No	No	No	No	No	No	No	No	Negative

Figure 3.8 Dataset for Diabetes

Step 1: Calculate Information Gain for each Attribute

Entrophy(dataset) = -(7/18)log(7/18)-(11/18)log(11/18)=0.29021 S[+7,-11]

Value(gender)=Male, Female

Entrophy(male)=-(4/15)log(4/15)-(11/15)log(11/15)=0.251 S(male)[+4.-11]

Entrophy(female)=-(3/3)log(3/3)-(0/3)log(0/3)=0 S(female)[+3,-0]

Gain(gender)=Entrophy(dataset)-∑|Sv|/|S|*Entrophy(Sv) =0.081

Similarly we can calculate Information gain of all attributes:

Gain(polyuria)=0.16
Gain(polydipsia)=0.023
Gain(sudden Weight loss)=0.09
Gain(visual Blurring)= 0.12
Gain(Irritability)=0.04
Gain(Partial Paresis)=0.202
Gain(Weakness)=0.079
Gain(Genital trush)=0.075

Gain(Delayed hearing)=0.01
Gain(Itching)=0.036
Gain(Muscle Stiffness)=0.092
Gain(Alopecia)=0.054
Gain(Obesity)=0.102

Now the attribute with maximum gain is made root node. Here, Polyurea is made the root node.

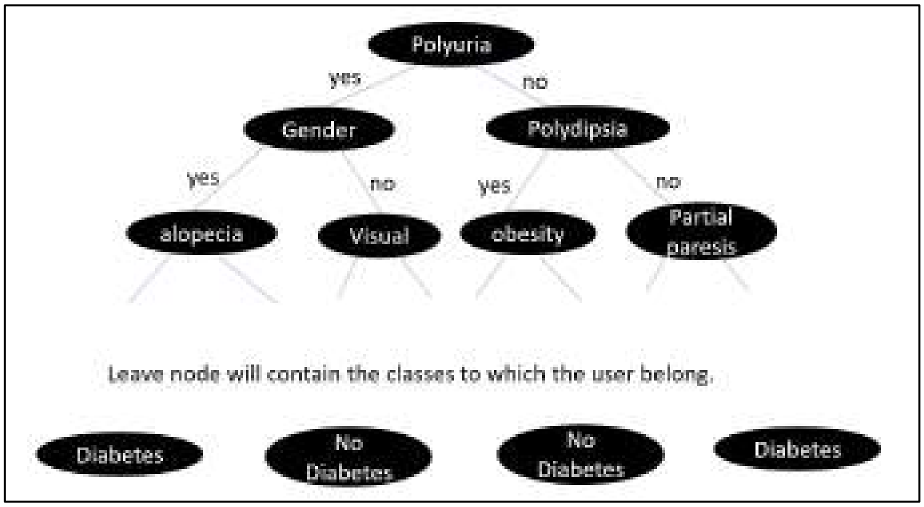


Figure 3.9 Decision Tree

Table 3.3: Accuracy

	Naives Bayes	**Random Forest**
ACCURACY	86%	94%

Chapter 4

Hardware and Software

4.1 Hardware Requirement:

- Windows: Microsoft Windows 8/7 4 GB RAM

4.2 Software Specification:

- Windows Operating System.
- MySQL
- Python
- Flask
- Anaconda ,Jupyter, Spyder

4.3 Technologies Used:-

4.3.1 MySQL:

Mysql is prestigious as world's most and large utilized relational database. It is a open source database. It is highly compatible with large and small scale applications. MySQL makes it really easy to write and inset queries. It can be easily coupled with most of the available technologies. SQL queries can be used to easily get the job done.

4.3.2 Python:

Python could likewise be a taken item organized basic level language with dynamic derivation its straightforward level in-created information structures got together with unique organization and dynamic restricting sort it outrageously interesting for speedy application advancement what's more on be utilized as a pre piece or glue language to relate existing components on pythons clear direct to be told accentuation highlights quality by then decreases the cost of program fixes python maintains modules and packs that moves program quality and code utilize the python go-between and what's more the escalated standard library are offered in give or combined sort to nothing of charge for each and every fundamental stage and wish to be uninhibitedly spread of programmers fall stricken with python because of the misrepresented strength it gives since there is no aggregation step the special stepped area test-investigate cycle is unfathomably expedient work python programs is basic a bug or unfortunate information won't ever cause a division deformity taking everything into account once the interpreter discovers a blunder it raises an extraordinary case once the program doesn't get the exception the go-between prints a stack follow a stock level program licenses assessment of local and world elements examination of self-emphatic enunciations setting breakpoints wandering through the code a line at a rapidly on the program is written in python itself vouching for pythons smart power barring generally the quick in view of right a program is to incorporate a few print clarifications to the accessibility the quick modify test-explore cycle makes this simple philosophy dreadfully amazing.

4.3.1 Flask:

A Flask is Web Application Framework that is built with Flexibility and Speed in the Mind .Flaskis Built in Python , which many data Scientist are familiar with . Flask takes care of the Environment and Project setup involved in web Applications Allowing the Developer to focus on their application rather than thinking about HTTP ,routing , dataset etc. Flask allow Data Scientist to create simple Single page Applications and help or look into if they want to create Products for Consumers .Flask is a micro web framework written in Python. it is classified as a microframework because it does not require special tools or libraries. It is not the database abstraction layer, form validation, or other components where legacy third-party libraries provide common functions. However, Flask supports extensions that will add application functionality as if they were implemented in Flask itself. There are extensions for object relational mappers, form validation, load management, various open authentication technologies, and a number of other common framework-related tools

Flask was created by Armin Ronacher of Pocoo, a worldwide group of Python enthusiasts formed in 2004. According to Ronacher, the thought was originally an April Fool's joke that was popular enough to turn into a meaningful application. When Ronacher and Georg Brandl created a bulletin board system written in Python, the Pocoo Werkzeug and Jinja projects were developed. Flask has become popular with Python enthusiasts. As of October 2020, its second most stars on GitHub among Python web development frameworks, only slightly behind Django, and was voted the most popular web framework within the Python Developers Survey 2018.

These are some Important features of the Flask:

1. It is a Development Server

2. Debugger

3.RESTful request dispatching

4. Unicode Based

5. Flask have google app engine Compatibility

4.4 Installation Process

STEP 1: Installation of anaconda

Anaconda Navigator is a desktop graphical user interface included in Anaconda that allows you to launch applications and easily manage conda packages, environments and channels without the need to use command line commands.

https://anaconda.org/anaconda/anaconda navigator#:~:text=Anaconda%20Navigator%20is%20a%20desktop,to%20use%20command%20line%20commands.

STEP 2: Creating a new virtual environment in anaconda
https://www.geeksforgeeks.org/set-up-virtual-environment-for-python-using-anaconda/

STEP 3: Installation of jupyter text editor

STEP 4: Installation of spyder

STEP 5: Installation of MySQL

STEP 6: Installation of SQLyog

Chapter 5

Design and Implementation

5.1 Software development Life Cycle

The entire project spanned for duration of 6 months. In order to effectively design and develop a cost-effective model the Waterfall model was practiced.

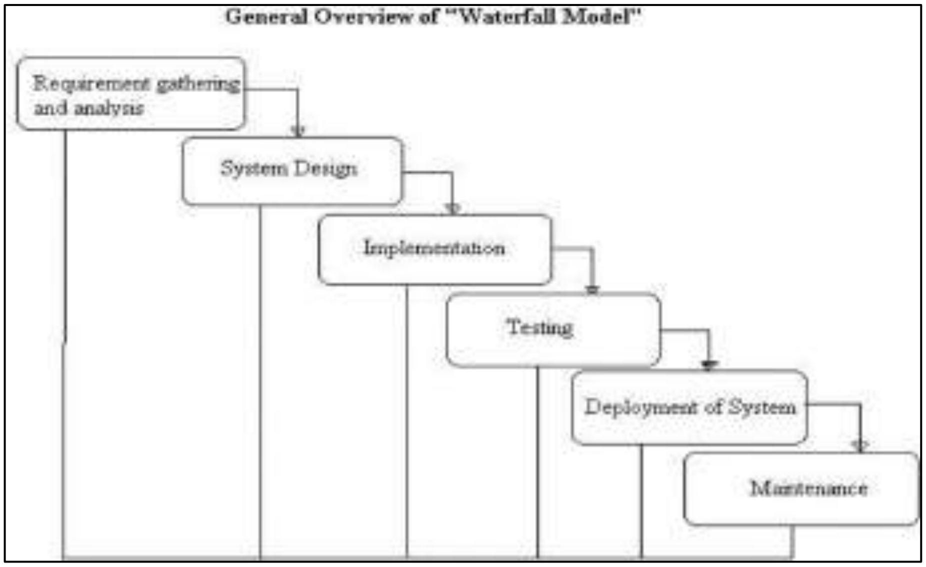

General Overview of "Waterfall Model"
Requirement gathering and analysis
System Design
Implementation
Testing
Deployment of System
Maintenance

Figure 5.1 Waterfall Model

5.2 Functional requirements :

1. System should have sufficient internet to fetch the data from the server.
2. The system will acquire correct results for particular disease.
3. System should be able to match required configurations.
4. Database should be updated by the latest values.

5.3 Non-functional requirements :

1. The reliability of the product will be dependent on the accuracy of the data.
2. Site is hands on or friendly so that customers can view / use it easily.
3. The processing speed of prediction of algorithm should be less than a minute.

5.4 ER -Diagram:

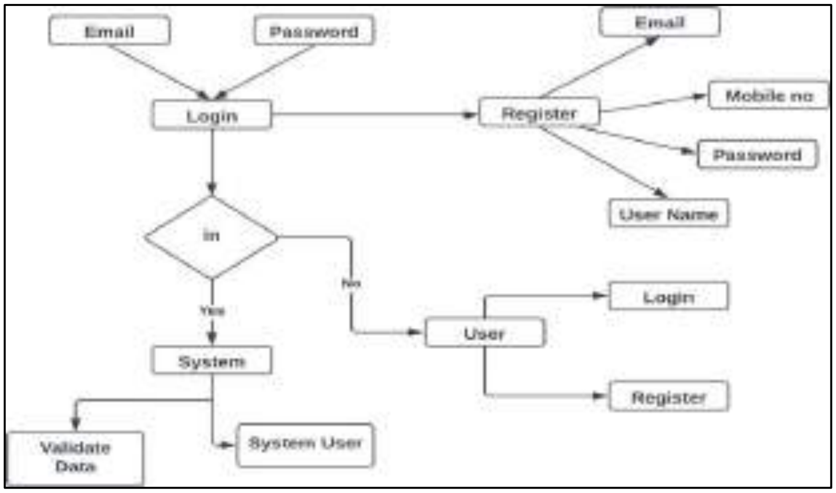


Figure 5.2 ER Diagram

5.5 Flowchart:

5.5.1 Skin Disease flowchart:

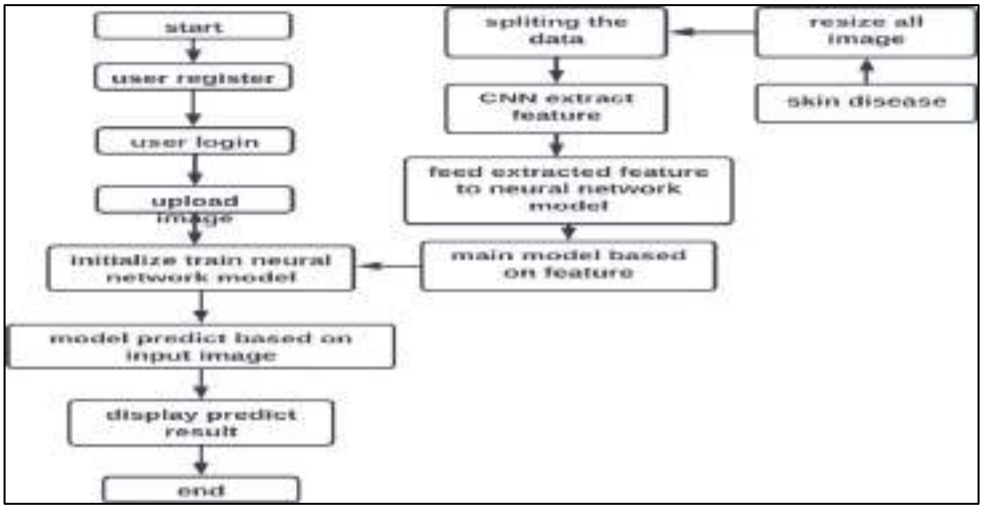

start
user register
user login
upload image
initialize train neural network model
model predict based on input image
display predict result
end
spliting the data
resize all image
CNN extract feature
skin disease
feed extracted feature to neural network model
main model based on feature

Figure 5.3 Flowchart for Skin Disease

5.5.2 Diabetes flowchart:

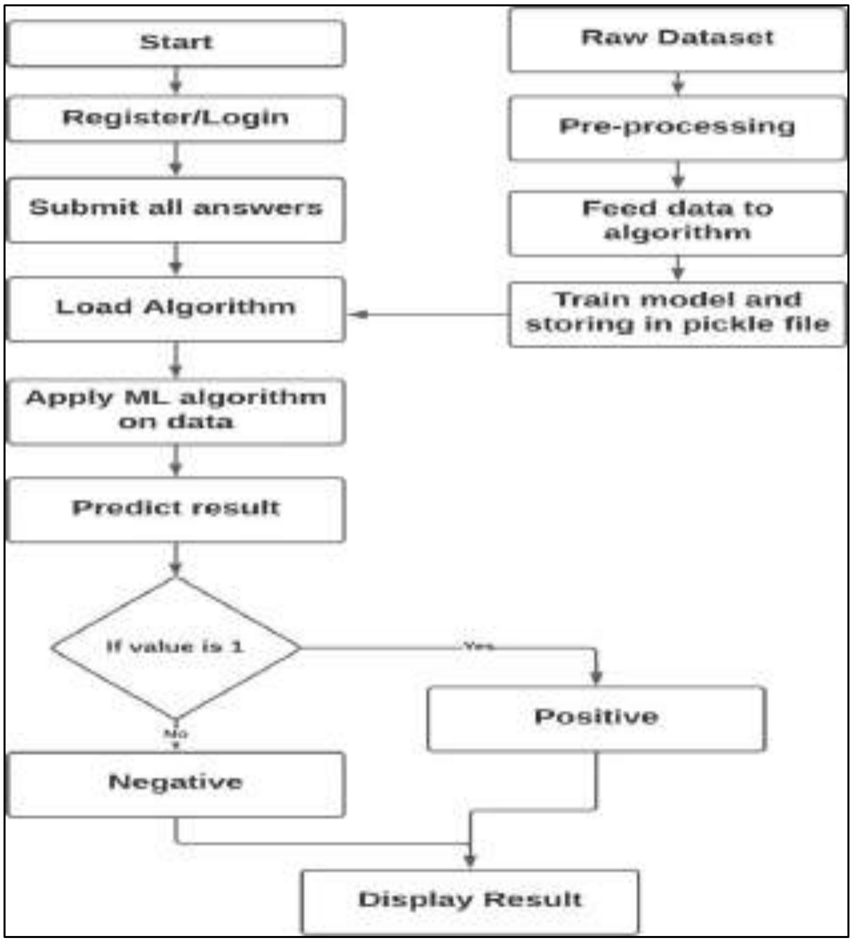


Figure 5.4 Flowchart for Diabetes Prediction

5.6 Activity Diagram:

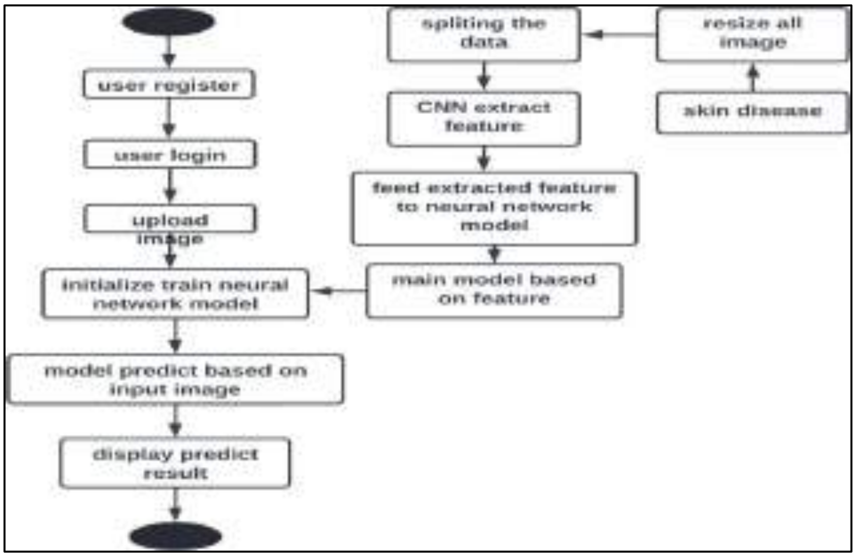

user register
user login
upload image
initialize train neural network model
model predict based on input image
display predict result
spliting the data
resize all image
skin disease
CNN extract feature
feed extracted feature to neural network model
main model based on feature

Figure 5.5 Activity Diagram

5.7 Class Diagram:

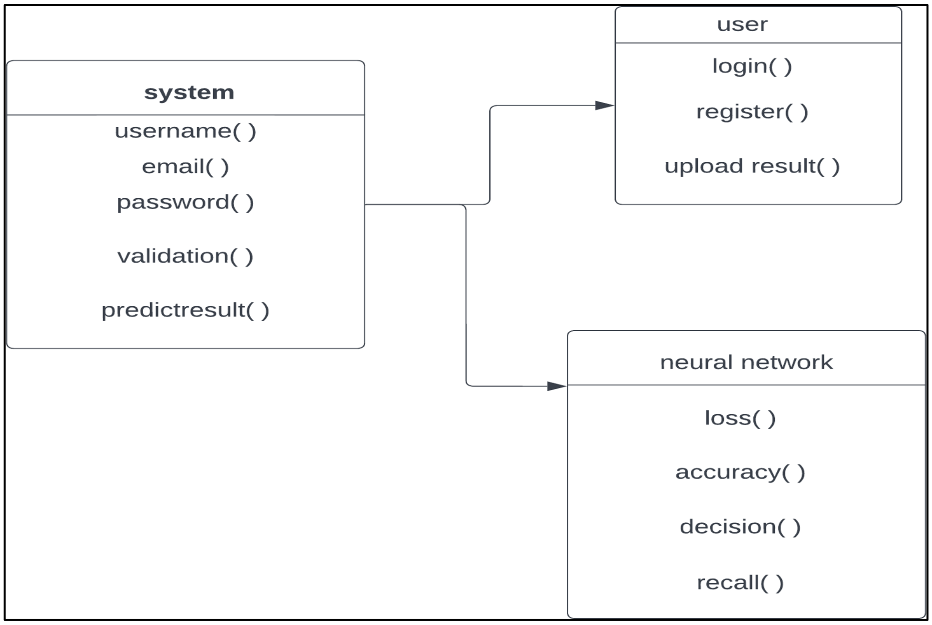


Figure 5.6 Class Diagram

Chapter 6

Experimentation Results

6.1. Experiment Result 1

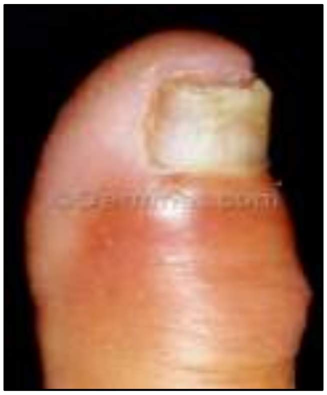
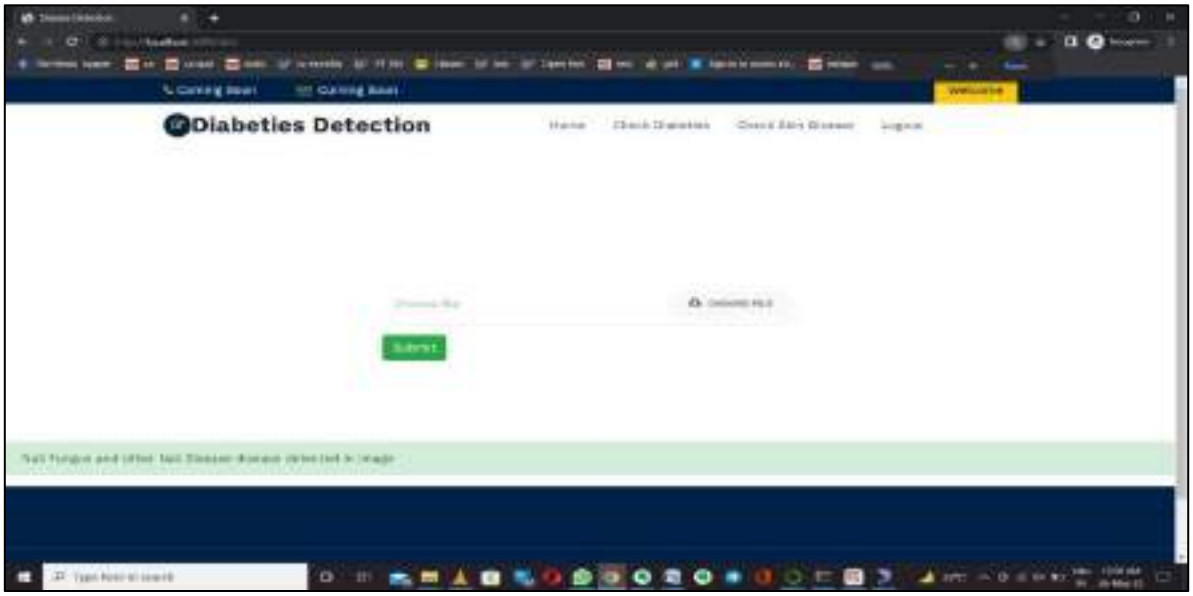


Figure 6.1.1 Input Image



Nail Fungus and other Nail Disease disease detected in image

Figure 6.1.2 Predicted disease

Actual disease

Nail Fungus and other Nail Disease

Result: Predicted diseases matches output disease

6.2. Experiment Result 2

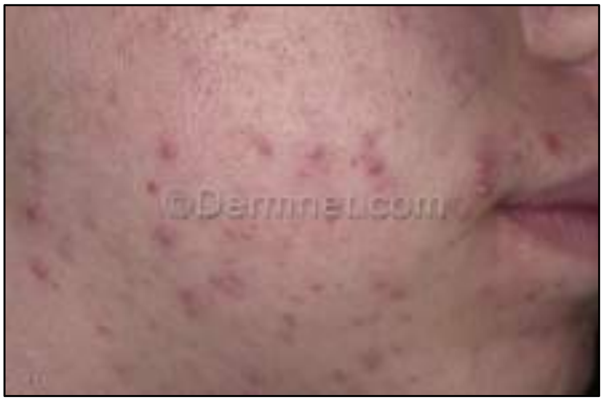
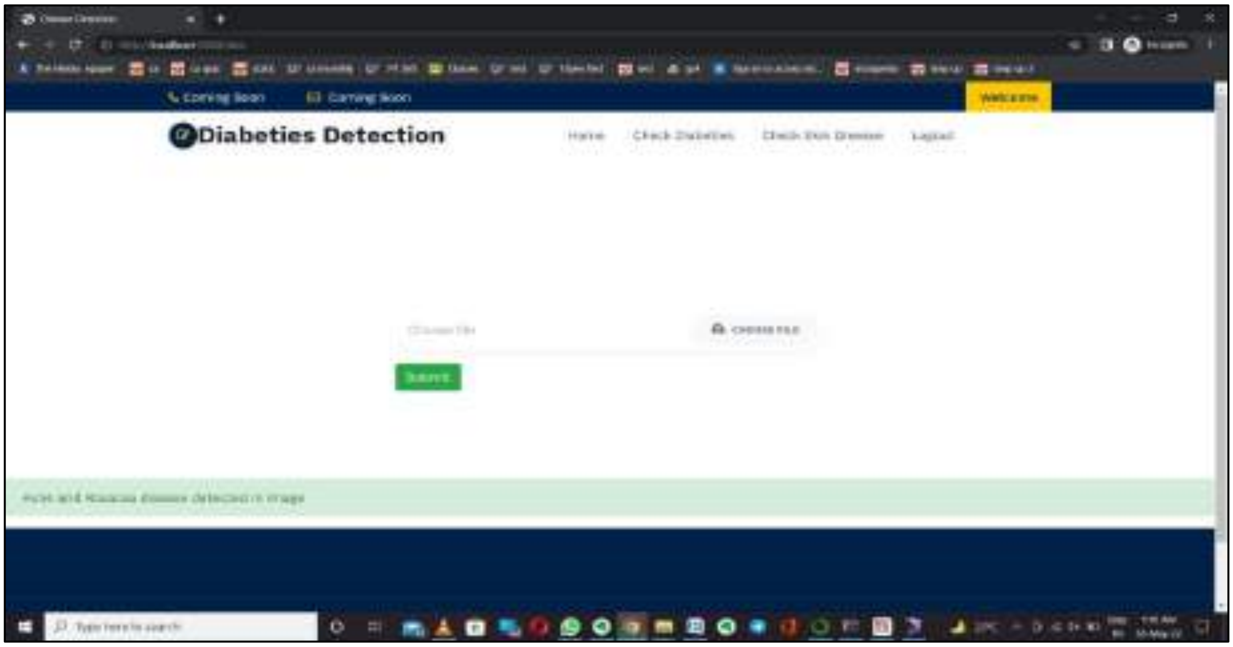


Figure 6.2.1 Input Image



Acne and Rosacea disease detected in image

Figure 6.2.2 Predicted Disease

Actual disease

Acne and Rosacea disease

Result: Predicted diseases matches actual disease

6.3. Experiment Result 3

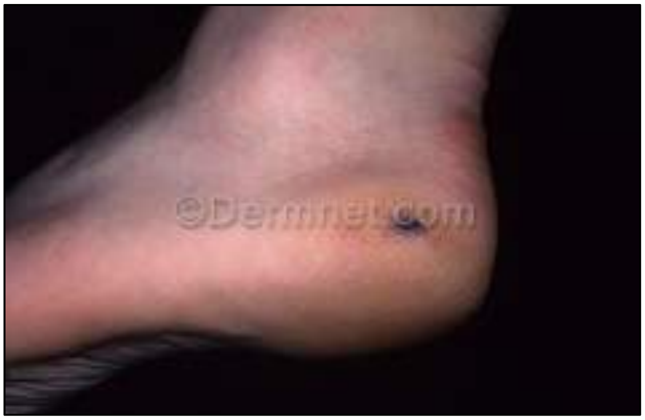
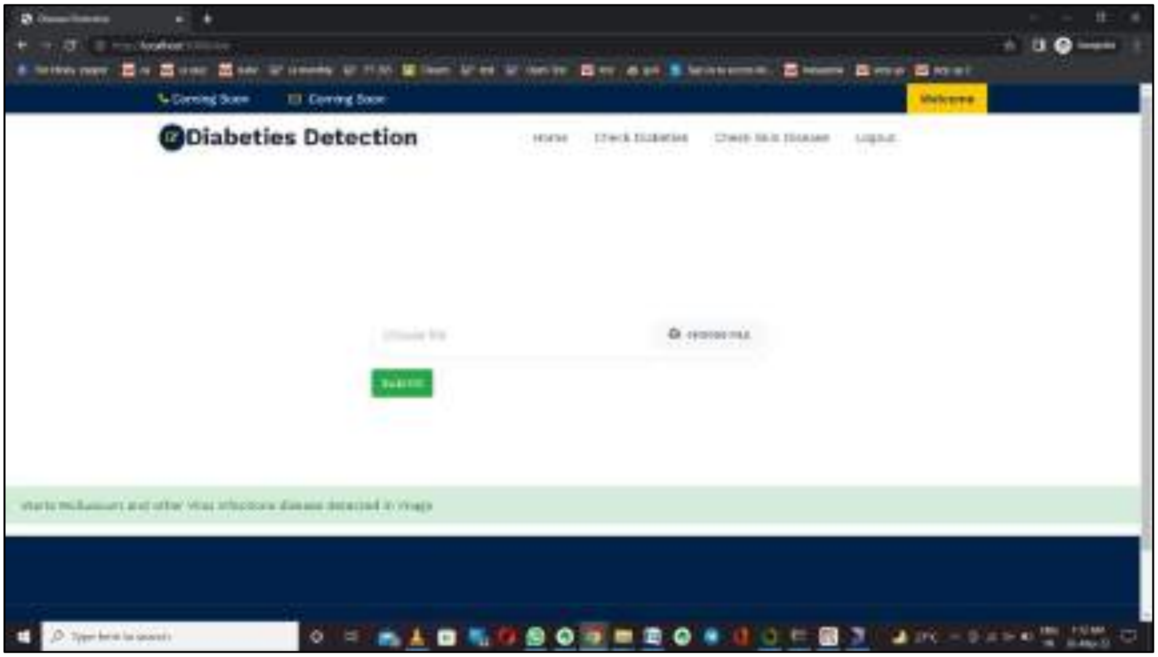


Figure 6.3.1 Input Image



Warts Molluscum and other Viral Infections disease detected in image

Figure 6.3.2 Predicted disease

Actual disease

Warts Molluscum and other Viral Infections disease

Result: Predicted diseases matches Actual disease

6.4. Experiment Result 4

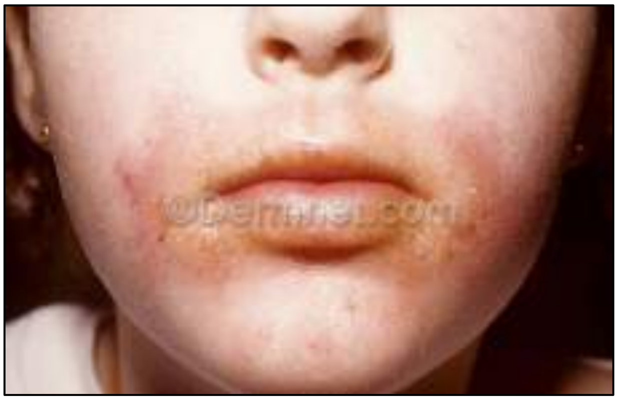
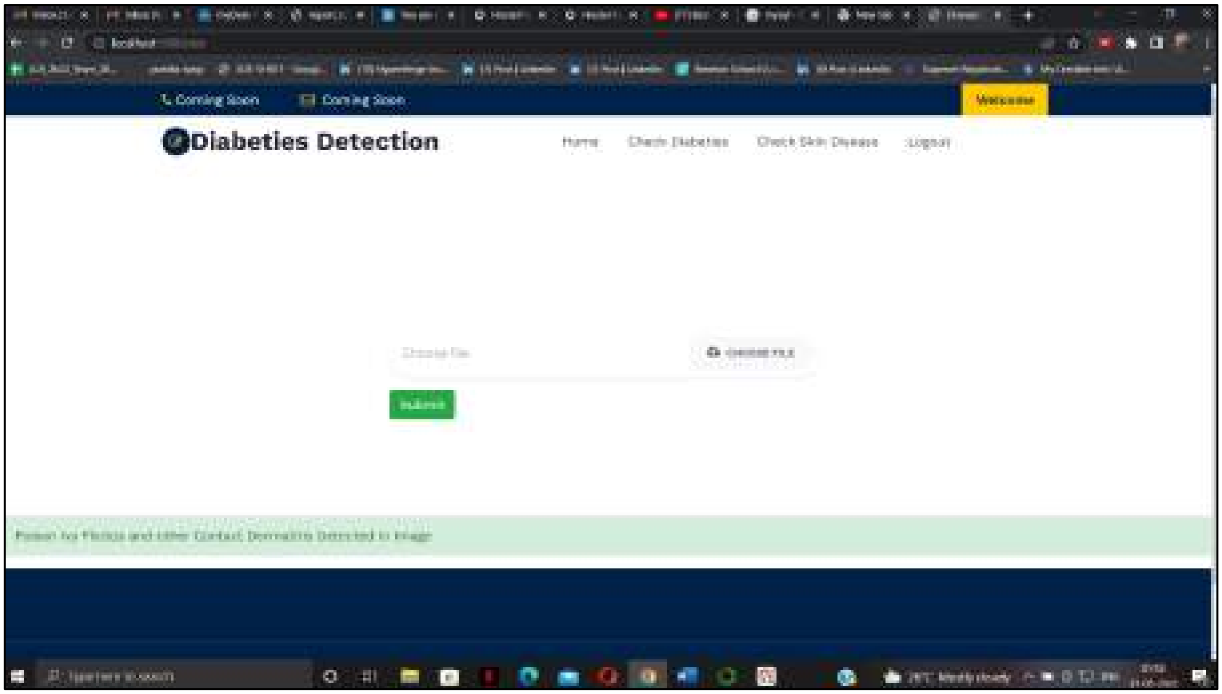


Figure 6.4.1 Input Image



Poison Ivy Photos and other Contact Dermatitis Detected in Image

Figure 6.4.2 Predicted Disease

Actual disease
Poison ivy and other contact Dermatitis.

Result: Predicted diseases matches actual disease

6.5 Experiment Result 5

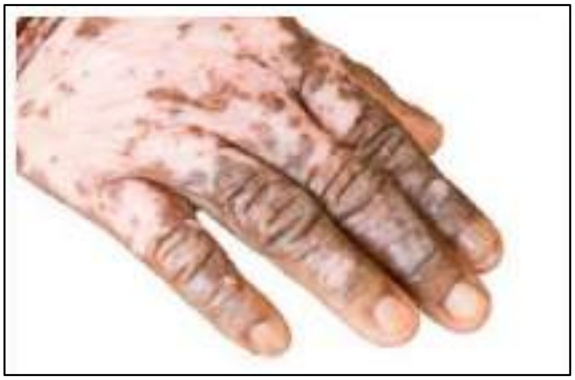
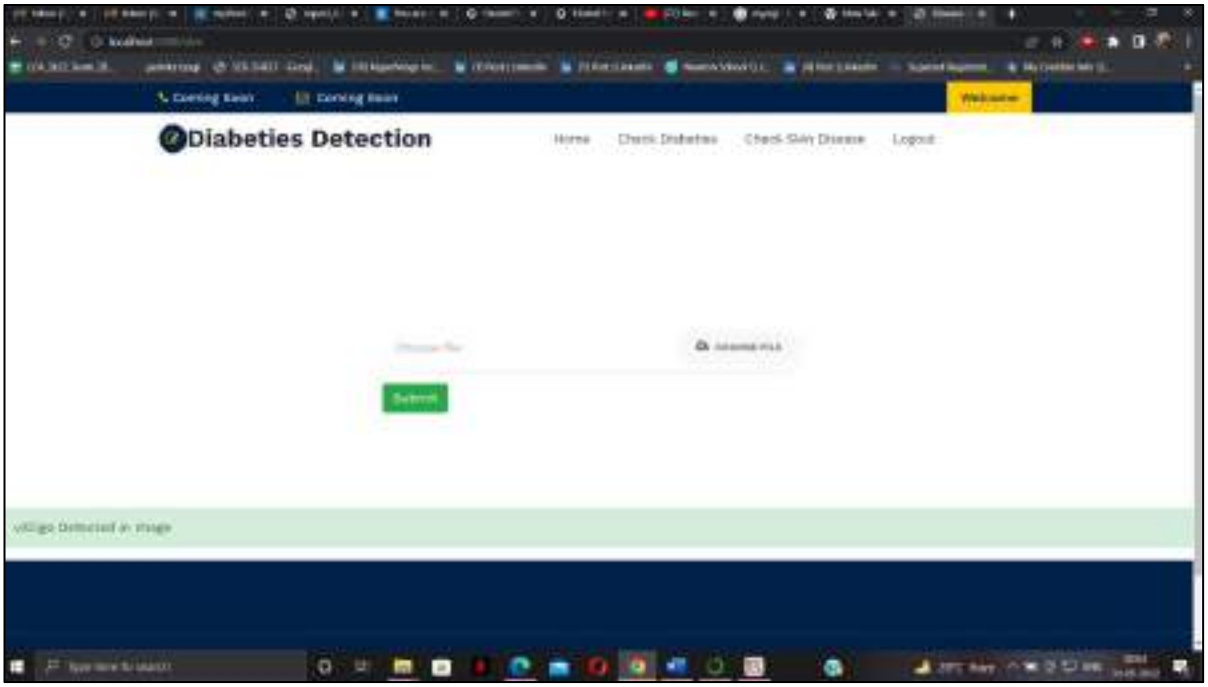


Figure 6.5.1 Input Image



vitiligo Detected in image

Figure 6.5.2 Predicted Image

Actual disease
Vitiligo

Result: Predicted diseases matches actual disease

6.6 Experiment Result 6

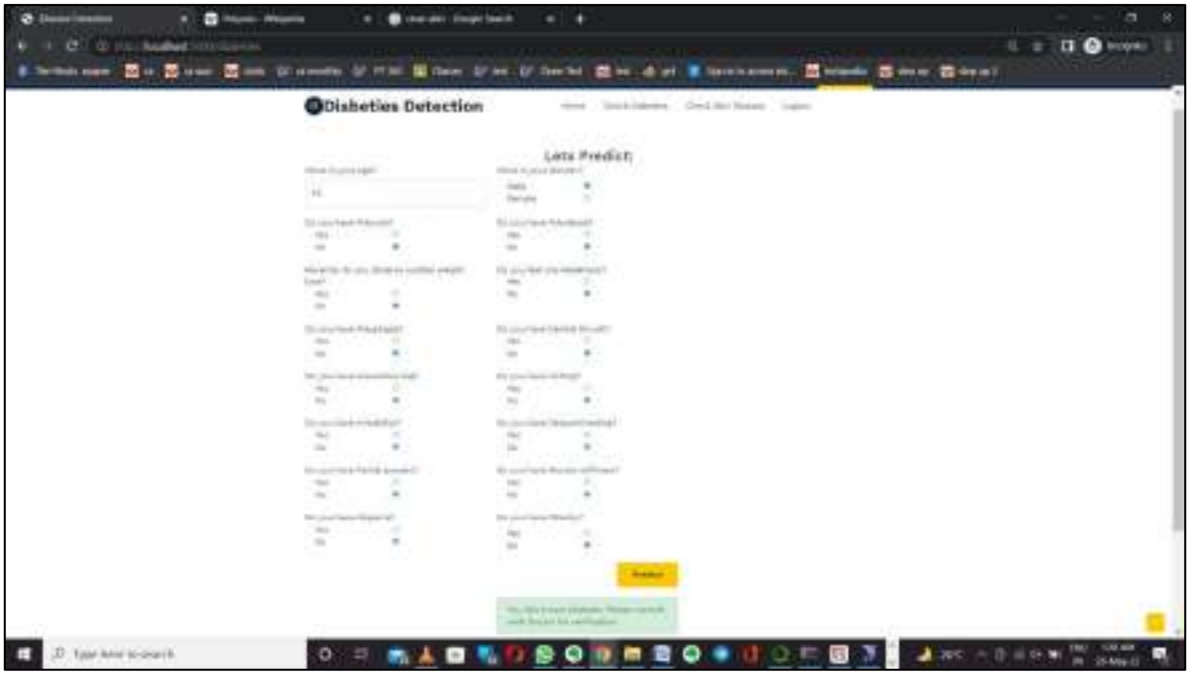


Figure 6.6.1 Diabetes is not Detected

6.7 Experiment Result 7

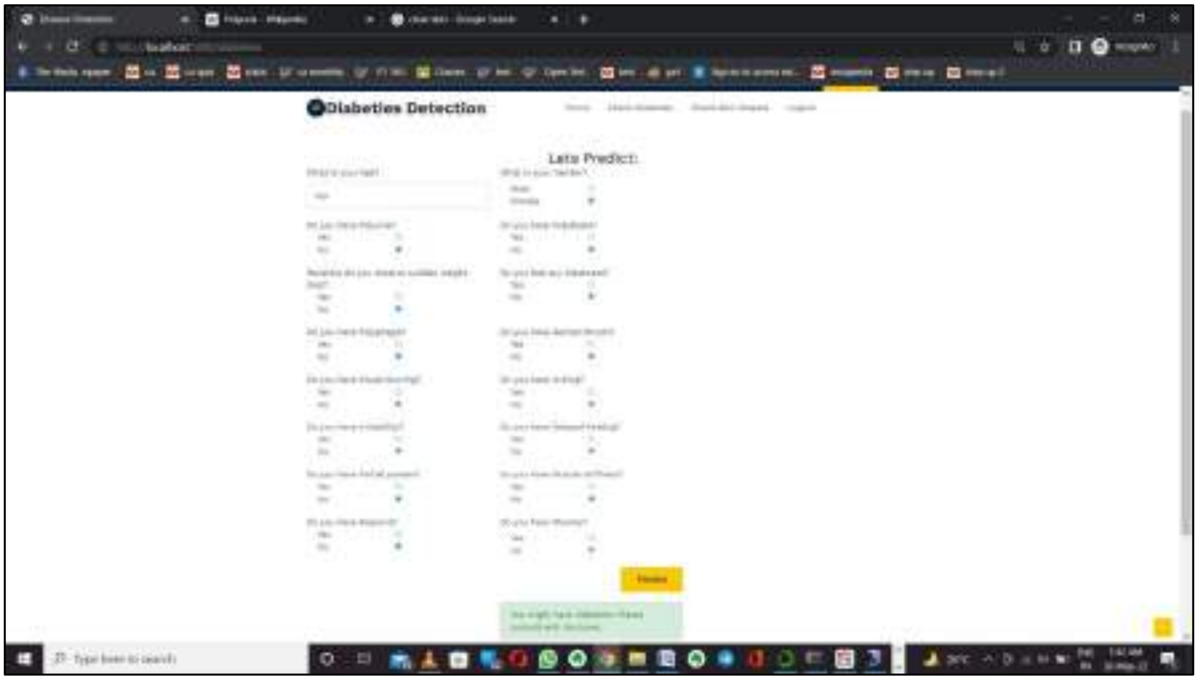


Figure 6.7.1 Diabetes is Detected

Chapter 8

Result and Conclusion

The proposed system is to predict diseases based on the disease selected.

If skin disease is selected model classify the given skin image as ‘Not a Skin disease” or one of the seven diseases namely Acne, Hair loss, Melanoma, Poison Ivy and other contact diseases, Nail fungus, Vitiligo, Warts Molluscum and other viral infections using deep learning techniques. The dataset is divided into training and testing and deep learning models are build using CNN and a pretrained networks called VGG16 .The accuracy of the model is about 86%.

If diabetes disease is selected model predicts the category of non-diabetic and diabetic patients by using the answers to the questions given as input by user. This model uses machine learning technique to predict the correct class. The accuracy level is more than 94%.

Conclusion:

The potential benefits of in-depth dermatological study solutions are enormous and have an unparalleled benefit in reducing duplicate activity of dermatologists and pressure on accurate access to medical services computer science and medicine with the continuous development of the above fields in-depth learning is rapidly developing and has become enticing. The attention of many countries enabled affordable software solutions that can quickly collect and reasonably process large data and hardware that can accomplish what people can do is clear that in-depth learning to diagnose skin disease is a viable option in the future. The ability to create a general skin classification system has been investigated using CNN, VGG16 and Inceptionv3. CNN did much better than training data but not test data. Best accuracy can be achieved by providing a training set with more flexibility and also by increasing its size. It has also been found that VGG16 has provided much better accuracy compared to other networks in the diagnosis of skin diseases. Our proposed method is designed for a specific skin tone. Future research is needed to evaluate the effect of different skin tones on the function of the lesion and the classification system.

Diabetes is vital health hassle in human society. This paper has summarised kingdom of art techniques and to be had techniques for predication of this sickness. Deep studying an rising region of Machine Learning showed a few promising bring about different area of clinical diagnose with excessive accuracy. It continues to be an open area waiting to get applied in Diabetes predication. Some strategies of deep studying has been discussed which may be implemented for Diabetes predication, alongside pioneer machine getting to know algorithms. An analytical assessment has been completed for locating out best available algorithm for clinical dataset. In future our purpose is to carry ahead the work of temporal scientific dataset, wherein dataset varies with time and retraining of dataset is needed.

Future Work:

In future we aim to enhance the applicability of our website by including more number of diseases for example, related to cancer, TB, and a lot of worlds neglected tropical diseases including meningitis, diphtheria etc. The logic can be developed in the form of app as with app the target audience would also increase.

REFRENCES

1. G.H. Dayan, et al. Recent resurgence of mumps in the United States. New England Journal of Medicine, vol. 358, no. 15, pages 1580–1589, 2008. https://doi.org/10.1056/NEJMoa0706589

2. S. Jowett and T. Ryan. Skin disease and disability: analysis of the impact of skin conditions, vol. 20, no. 4, pp. 425-429, 1985.

3. H. Zeng, H. Lui, C. MacAulay, B. Palcic, D. I. McLean. Resources and methods related to skin diagnostic programs, U.S. Patent No. 6,069,689, 2000.

4. Y. Demchenko, C. De Laat, and P. Membrey. Explaining the structural components of the Big Data Ecosystem, 2014 International Conference on Collaboration Technologies and Systems (CTS), pp. 104-112, 2014.

5. X. Jin, B. W. Wah, X. Cheng, and Y. Wang. Significance and challenges of big data research, Big Data Research vol. 2, no. 2, pages 59-64, 2015.

6. P. Jain, N. Pathak, P. Tapashetti, and A. S. Umesh. Confidentiality keeping data decision analysis tree based on single sample selection and decomposition, 2013 9th International Conference on Information Security and Security (IAS), pp. 91-95, 2013. https://doi.org/10.1109/ISIAS. 2013.6947739 Jessica Velasco et al., International Journal of Advanced Styles in Computer Science and Engineering, 8 (5), September - October 2019, 2632-2637 2637

7. J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang. Big health data, IEEE journal biomedical and health informatics, vol. 19, no. 4, pages 1193-1208, 2014.

8. A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton. Big data: management change, Harvard Business Review, vol. 90, when. 10, pages 60-68, 2012.

9. H. Kantarjian and P. P. Yu. Artificial intelligence, big data, and cancer, JAMA oncology, vol. 1, no. 5, pages 573-574, 2015.

10. T. B. Murdoch, and A. S. Detsky. The Inevitable Use of Big Data in Health Care, Jama, vol. 309, no. 13, pages 1351-1352, 2013.

11. A. L. Beam, and I. S. Kohane. Big data and machine learning in health care, Jama, vol. 319, no. 13, pages 1317-1318, 2018.

12. D. D. Miller and E. W. Brown. Artificial Intelligence in the medical profession: the answer to the question ?, The American Journal of Medicine, vol. 131, no. 2, pages 129-133, 2018.

13. S. Jha no-E. J. Topol. Practicing artificial intelligence: radiologists and pathologists as experts, Jama, vol. 316, no. 22, pages 2353-2354, 2016.

14. L. C. De Guzman, R. P. C. Maglaque, V. M. B. Torres, S. P. A. Zapido, and M. O. Cordel. Design and Examination of Multi-model, Multi-level Artificial Neural Network for Eczema Skin Lesion Detection, at the Third International Conference on Artificial Intelligence, Modeling and Simulation (AIMS), pages 42-47, 2015.

15. L. K. Tolentino, R. M. Aragon, W. R. Tibayan, A. Alvisor, P. G. Palisoc, and G. Terte. Diagnosis of Surrounding Fingernails Through Artificial Neural Network, Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-4, pages 181-188, 2018.

16. J. Velasco, J. Rojas, J. P. Ramos, H. M. Muaña and K. L. Salazar. Lifetime Assessment Tool Using Language Analysis Based Successful Image Analysis, International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 3, pages 451-457, 2019. https://doi.org/10.30534/ijatcse/2019/19832019 17. P. S. Ramesh, S. Arivalagan and P. Sudhakar. Analysis of Alternative Diagnosis Alzheimer's Disease and Its Functions, International Journal of Advanced Styles in Computer Science and Engineering, vol. 8, no. 3, pages 755-757, 2019. https://doi.org/10.30534/ijatcse/2019/65832019 18. A. R. Bayot, L. A. M. Samoy, N. A. D. Santiago, M. R. Tomelden, and. Cruel diagnosis of basal cell carcinoma using image processing and neural artificial network, DLSU Engineering Journal, vol. 1, no. 1, pp. 70-79, 2007. 19. J. Zambales and P. A. Abu. Modification of skin lesions using imaging techniques and neural network, in Proceedings of the 17th Philippine

17.Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In Contemporary Computing and Informatics (IC31), 2016 2nd International Conference on (pp. 347-352). IEEE.

18.Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.

19.https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/

20.https://biomedpharmajournal.org/vol11no3/automated-skin-disease-identification-using-deep-learning-algorithm/