Skin Cancer Detection

Using

Convolutional Neural Network

and

Ensemble Modeling

Under the Guidance of

**Dr. Agilandeeswari L**

(Associate Professor, School of Information Technology and Engineering)

For the Course

**ITE1015 – Soft Computing**

For Winter Semester 2020-21

**Project Team:**

Aashish Bansal 19BIT0346

Perumalla Sasank 19BIT0338

Keerthi Yasasvi 19BIT0335

Table of Contents

[Abstract 4](#_Toc68295804)

[Ketwords 4](#_Toc68295805)

[Introduction 4](#_Toc68295806)

[Dataset - **ISIC2017: Skin Lesion Analysis Towards Melanoma Detection** 5](#_Toc68295807)

[Abstract of Dataset 5](#_Toc68295808)

[Background 5](#_Toc68295809)

[Melanoma 5](#_Toc68295810)

[Dermoscopy 5](#_Toc68295811)

[About the Diseases 6](#_Toc68295812)

[Motivation 6](#_Toc68295813)

[General Architecture 7](#_Toc68295814)

[Comparative study on various subtitles: 7](#_Toc68295815)

[Literature Survey 7](#_Toc68295816)

[Grouping 7](#_Toc68295817)

[Future Scope 7](#_Toc68295818)

[Summary on Literature Survey based on the General Arhitecture Processes 8](#_Toc68295819)

[Data Selection 8](#_Toc68295820)

[Exploratory Data Analysis 8](#_Toc68295821)

[Checking the Types of Data 8](#_Toc68295822)

[Finding the Outliers 8](#_Toc68295823)

[Data Visualization 8](#_Toc68295824)

[Data Pre-Processing 8](#_Toc68295825)

[Splitting the Data 8](#_Toc68295826)

[Checking for Missing Values 8](#_Toc68295827)

[Checking Categorical Features 8](#_Toc68295828)

[Normalizing Dataset 9](#_Toc68295829)

[Feature Transformation 9](#_Toc68295830)

[Why do we need Feature Transformation and Scaling? 9](#_Toc68295831)

[Feature Transformations used in the Models 9](#_Toc68295832)

[MaxAbsScalar 9](#_Toc68295833)

[Robust Scalar 9](#_Toc68295834)

[Unit Vector Scaler 9](#_Toc68295835)

[Feature Selection 10](#_Toc68295836)

[Principal Component Analysis 10](#_Toc68295837)

[Linear Discriminant Analysis 10](#_Toc68295838)

[Model Selection 11](#_Toc68295839)

[Model Training 11](#_Toc68295840)

[Model Evaluation 11](#_Toc68295841)

[Conclusions: 11](#_Toc68295842)

[References: 11](#_Toc68295843)

[Image acquisition 12](#_Toc68295844)

[Image Pre-processing: 12](#_Toc68295845)

[Feature Extraction 12](#_Toc68295846)

[Classification: 12](#_Toc68295847)

[Network Architecture 12](#_Toc68295848)

[Pre-processing 12](#_Toc68295849)

[Data Preprocessing 13](#_Toc68295850)

[Read Image 13](#_Toc68295851)

[Resize Image 13](#_Toc68295852)

[Splitting the Data 13](#_Toc68295853)

[Image Preprocessing 13](#_Toc68295854)

[Convert into Array and then into Tensors 13](#_Toc68295855)

[Building 13](#_Toc68295856)

[Neural Network 13](#_Toc68295857)

[Architecture and Process Flow Model for the Proposed Model 14](#_Toc68295858)

[Methodology and System Design 16](#_Toc68295859)

[Image Processing 16](#_Toc68295860)

[Convolutional Neural Network 16](#_Toc68295861)

[Proposed Simple Process Flow Diagram 16](#_Toc68295862)

Skin Cancer Detection using CNN

General Architecture

# Abstract

The project is a CNN trained model which can predict whether the patient has a suffering from Cancer or not by checking the images of the infected areas on the body. The model has been trained on a variety of images through which it predicts the required.

In this project, the image file of the patient is upload into a software, which is GUI-based interface, developed with the help of Tkinter, and it consists of the model saved as a file and the software uses that to analyse the image and give the prediction which can help doctors to start with the medication way faster instead of waiting for the laboratory reports for the confirmation.

So basically,

* Skin cancer is an abnormal growth of skin cells. Most skin cancers are caused by exposure to ultraviolet (UV) light. When the skin is not protected, UV rays from sunlight or tanning beds can damage and alter skin's DNA that leads to the cancer.
* Deep learning model has been built to classify and identify the binary diagnostic group of melanocytic images obtained through dermoscopy.
* Based on the model, disease detection through dermal cell images has been investigated, and classifications on dermal cell images have been performed.

# Ketwords

* Model
* Convolutional Neural Network
* Cancer
* Malignant
* Benign
* Detection
* Tkinter
* Software
* Analysis
* Uploading
* Training
* Testing
* Validation
* Prediction
* MobileNet
* Inception
* Xception

# Introduction

* The significant growth of medical images and techniques requires comprehensive and exhaustive efforts from a medical professional who is susceptible to human error and the result can also vary widely among various experts.
* In this project, we have used the above stated idea behind disease detection, to develop a system using convolutional neural network (CNN) that will help in detection of a particular disease.
* The system has been made user-friendly with the help of GUI, so that it can be used not only by the medical professionals but also by the population at large.

# Dataset - **ISIC2017: Skin Lesion Analysis Towards Melanoma Detection**

## Abstract of Dataset

The goal of the challenge is to help participants develop image analysis tools to enable the automated diagnosis of melanoma from dermoscopic images. Image analysis of skin lesions is composed of 3 parts:

* Part 1: Lesion Segmentation
* Part 2: Detection and Localization of Visual Dermoscopic Features/Patterns
* Part 3: Disease Classification

This challenge provides training data (150 images) and blind held-out test dataset (~600 images) will be provided for participants to generate and submit automated results.

## Background

### Melanoma

Skin cancer is a major public health problem, with over 5 million newly diagnosed cases in the United States each year. Melanoma is the deadliest form of skin cancer, responsible for over 9,000 deaths each year.

### Dermoscopy

As pigmented lesions occurring on the surface of the skin, melanoma is amenable to early detection by expert visual inspection. It is also amenable to automated detection with image analysis. Given the widespread availability of high-resolution cameras, algorithms that can improve our ability to screen and detect troublesome lesions can be of great value. As a result, many centers have begun their own research efforts on automated analysis. However, a centralized, coordinated, and comparative effort across institutions has yet to be implemented.

Dermoscopy is an imaging technique that eliminates the surface reflection of skin. By removing surface reflection, visualization of deeper levels of skin is enhanced. Prior research has shown that when used by expert dermatologists, dermoscopy provides improved diagnostic accuracy, in comparison to standard photography. As inexpensive consumer dermatoscope attachments for smart phones are beginning to reach the market, the opportunity for automated dermoscopic assessment algorithms to positively influence patient care increases.

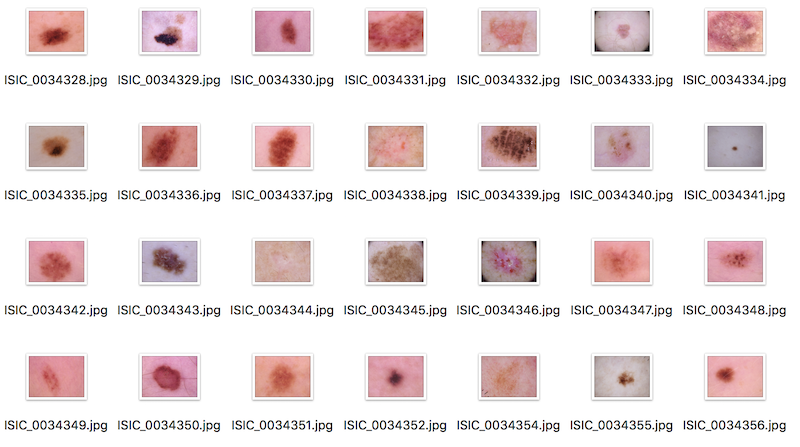


Figure 1: Sample Dataset Images

# About the Diseases

Skin cancer is the most prevalent type of cancer. Melanoma, specifically, is responsible for 75% of skin cancer deaths, despite being the least common skin cancer. The American Cancer Society estimates over 100,000 new melanoma cases will be diagnosed in 2020. It's also expected that almost 7,000 people will die from the disease. As with other cancers, early and accurate detection—potentially aided by data science—can make treatment more effective.

Currently, dermatologists evaluate every one of a patient's moles to identify outlier lesions or “ugly ducklings” that are most likely to be melanoma. Existing AI approaches have not adequately considered this clinical frame of reference. Dermatologists could enhance their diagnostic accuracy if detection algorithms take into account “contextual” images within the same patient to determine which images represent a melanoma. If successful, classifiers would be more accurate and could better support dermatological clinic work.

As the leading healthcare organization for informatics in medical imaging, the [Society for Imaging Informatics in Medicine (SIIM)](https://siim.org/)'s mission is to advance medical imaging informatics through education, research, and innovation in a multi-disciplinary community. SIIM is joined by the [International Skin Imaging Collaboration (ISIC)](https://www.isic-archive.com/), an international effort to improve melanoma diagnosis. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions.

In this competition, you’ll identify melanoma in images of skin lesions. In particular, you’ll use images within the same patient and determine which are likely to represent a melanoma. Using patient-level contextual information may help the development of image analysis tools, which could better support clinical dermatologists.

Melanoma is a deadly disease, but if caught early, most melanomas can be cured with minor surgery. Image analysis tools that automate the diagnosis of melanoma will improve dermatologists' diagnostic accuracy. Better detection of melanoma has the opportunity to positively impact millions of people.

# Motivation

* Disease detection plays a very important role in the process of diagnosis. Therefore, the motivation lies in accurate classification and detection of the diseases based on medical images.
* The main aim is to minimize the chances of error that might happen due to the doctor's misjudgement.
* Developing a system that will not only help in detecting the diseases efficiently but will also save the time and effort of the medical practitioners.
* This will also save the patients from running to the doctor to get their medical reports verified.

# General Architecture

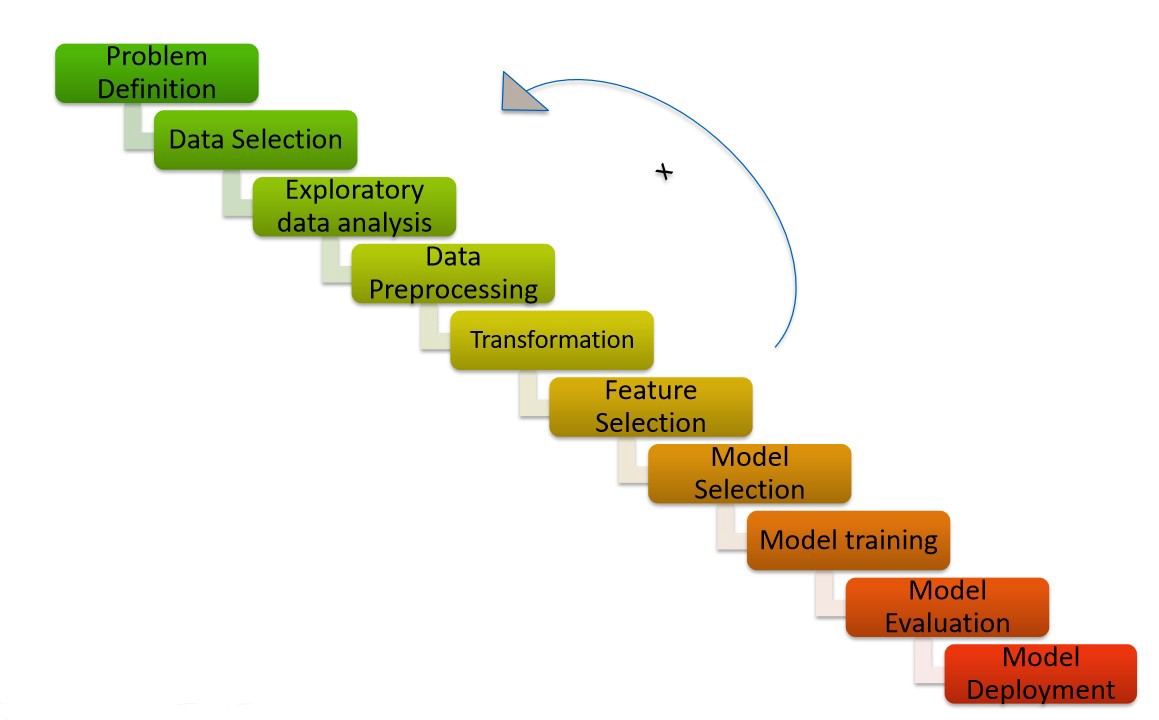
(Analytics Vidya, 2019)

Figure 2: General Architecture

# Comparative study on various subtitles:

## Literature Survey

The Literature Survey done for the accomplishment of the project is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Authors &Year | Methodology or Techniques used | Advantages | Issues | Metrics used | Pros | Cons |
| May-20 | CNN, AlexNet, ResNet-18, VGG16, SVM, Black-hat filter, Inpaint Algorithm, Median Filter, Otsu’s Methodology | SVM Accuracy = 86.21%, ResNet Accuracy = 87% | Accuracy Original Data = 80%, Accuracy Augmented Data = 98.61% | ReLU, CNN with Data Augmentation = 88.87%, CNN without Data Augmentation = 78.96% | Good Accuracy | Small Dataset |
| 2020 | CNN, Inception-v3, Keras, TensorFlow, DCNN, LeakyReLU, Adamax optimizer, TPR is similar to the positive predictive value | 0.86 AUROC for BKL | 0.78 AUROC for MEL | Accuracy |  |  |
| 2020 | MVSM classifier, CNN, feature extraction, GLCM, SVM, ABCD | dataset which consists of eight different classes is compressed into 800 images and applied, the accuracy achieved is about 96.25%. | accuracy is lowered if minute amounts of foreign elements are found on the sample | Accuracy | Eight Classes help specify the disease for specific medication | High accuracy on a very small specific training dataset |
| 2020 | CNN SENet154, WSL, Adam, weighted loss-entropy | Efficient Architecture | Not much improved with ensemble strategy | EfficientNet, SENet (T1 = 67.2%, T2 = 70.0%), ResNeXt WSL (T1 = 65.9%, T2 = 68.1%) | Transfer Learning | Not implemented properly with small dataset, parameter tuning required |
| 2020 | GLCM, HOG, GAC | Feature extraction for early detection | Not enough/adequate dataset | ABCD Rule, SVM Classifier, Accuracy, Sensitivity, Specificity using KNN |  |  |
| 2019 | Multiclass SVM, AlexNet, ReLU | Accuracy – 94.016% | Model used is a pre-trained model, robust | GOPS, L1D miss rate |  |  |
| 2019 | CNN, pooling layers, dense network, SVM | AlexNet, VGG16, ResNet-18 | Deep Network but Small Dataset – 3000 images | Accuracy = 74% |  |  |
| Apr-19 | CNN, pooling layer, dense network | Accuracy – 89.5% | Time consuming | Accuracy = 89.5%, Recall = 0.84, Specificity, Precision = 0.8325, F-measure = 0.8325 |  |  |
| Mar-19 | CNN, Inception V2 Net, K-means Cluster, Max-pooling, Sonification Algorithms | No. of K-means Epochs = 100 | F2-score +ve Prediction = 59.9%, High Sensitivity, Low Specificity | F2-score = 81.8%, Sensitivity = 91.7%, Specificity = 41.8%, Precision = 57.3% |  |  |
| 2019 | CNN, McNemar Test, ResNet50, Bonferroni Correction | MATLAB | Small dataset (11,444), training may be inefficient, class imbalance | Accuracy = 100%, detection rate = 100% |  |  |
| 2018-2019 | CNN, VGG16, ImageNet | Accuracy – 92.5% | F1-score = 0.77, VGG16 Accuracy = 78% | Random Forest = 65.9%, XGBoost = 65.15%, SVM = 65.86%, ReLU, Sigmoid | Max-pooling fetches maximum pixel | Low F-score |
| 2018 | MatConvNet & GoogLeNet Inception V3 CNN, GoogLeNet, AlexNet, ResNet, VGGNet, Simple Majority Voting, SMP | 1.28 million NATURAL images 500 epochs, pre-trained models used, MatConvNet provides pre-trained CNN models and some functions to create and initialize new neural networks | Limited computational resources, time-consuming procedures | GoogLeNet Error Rate = 0.1, ResNet Error Rate= 0.02, AlexNet Error Rate = approx. 0.001, VGGNet Error Rate = approx. 0.001 | High Results for all the evaluation metrics | Too many evaluation metrics and parameters used |
| 2018 | CNN, ImageNet,  AlexNet, VGG ,GoogLeNet, ResNet | Big dataset | There is a risk of overfitting the neural network | Accuracy,Image classification | advantageous for the decision making of dermatologists | There is a risk of overfitting the neural network |
| 2020 | skin lesions using CNNs, Alexnet. deep learning with PNASNet-5 | large dataset with 4 classsifications | Less accuracy | Good dataset | efficient in computing time, consume less memory |  |
| 2018 | CNN, AlexNet, deep learning, AlexNet, VGG ,GoogLeNet, ResNet, Inception V3 | Public dataset and ISIB-2016,2017 | Time-consuming,and errors | Large variety, dataset | advantageous  for  the  decision  making  of dermatologists | Time-consuming. In addition, errors and the loss of informationin the first processing steps have a very strong influence on theclassification quality |
| 2020 | deep convolutional neural network, computer image analysis algorithms,CNN  , GoogLeNet Inception V3, AlexNet Deep Learning CNN | More variants from ISIC | Accuracy of less than 75% | Accuracy but small dataset | metric area under the curve of 99.77% was observed. | This would consume time and the patient may advance to later stage |
| 2019 | Machine learning, | Faster identification | Less accuracy | Accuracy,time | Got accuracy of min  85% | highly  complex  and  expensive  diagnosis  with  difficulties  and subjectivity of human  interpretation |
| 2017 | deep neural networks, Deep convolutional neural networks (CNNs) | dataset of 129,450 clinical images of Malignant and benign | less variants | CNN; melanoma;  skin cancer; image preprocessing | Large dataset and accuracy | Less variants of Malignant and benign |
| 2020 | deep learning,CNN, AlexNet and VGG-16, including VGG-Net, ResNet50, InceptionV3, Xception, and DenseNet121 | 70 % images were used for training and 30% used for testing | Less accuracy | Good dataset | 70 % images were used for training and 30% used for testing | accuracy of 65% to 75%,time consuming |
| 2017 | SVM, CNN, MobileNet | High accuracy | Small dataset | High accuracy | High accuracies in most cases | High Cost of pre trained models which are required |
| 2018 | CNN, GLCM, deep learning, CNN, ResNet,InceptionV2 | Trained on many variants | Small dataset | Accuracy but small dataset | Accuracy increases with bigger dataset | Bigger dataset required ,Time consuming process |
| 2020 | CNN, SVM, KNN, Naïve Bayes, and neural network | 97.8 % of Accuracy | High rate of overfitting and misidentification | accuracy | obtained is 97.8 % of Accuracy and 0.94 Area under Curve using SVM classifiers and additionally the Sensitivity obtained was 86.2 % and Specificity obtained was 85 % using KNN. | High rate of overfitting and misidentification |
| 2017 | GANs, CNN, AlexNet, StyleGANs, InceptionV3-StyleGANs, ResNet50- StyleGANs, VGG16BN- StyleGANs | size of 600×600 as input dataset, | sets the weight coefficient w in the SoftMax loss function | Accuracy | Model Automatically learns the feature representations required for the corresponding detection or classification tasks through the dataset, and has a good performance in many applications | Proposed DCGANs, which have clear structural constraints and indicate that they have weak credibility for unsupervised learning and that they are generalized most of the time. |
| 2018 | CNN, SciKit, Keras, TensorFlow, OpenCV, ReLU | 90% accuracy, Convolution maintains the spatial interrelation of the pixels, values of the pixels ranging from 0 - 255 i.e., 256 pixels. | Rectified Linear Unit is a non-linear operation. ReLU acts on an elementary level. | Accuracy | With a large dataset accuracy can be increased to 90% | Average’s and accuracy of 70% on standard publicly available dataset and time consuming |
| 2019 | AlexNet, Ordinary CNN, VGG-16, LIN, Inception-v3, and ResNet. Lévy flight, ReLU | size of input images in the input is considered 28×28 pixel. | doesn’t give the best global solution | Accuracy | 97% accuracy | Imbalance of training and testing dataset |
| 2019 | STM32, ROC, CNN, ReLU, NLSC | Accuracy - 99%,  F1-Score - 99% | computing and index loss, poor lesion skin discrimination specificity | Accuracy | This methodology, based on “visual” investigation by the dermatologist and/or oncologist, has the advantage of not being invasive and quite easy to perform | Several approaches proposed in scientific literature increase sensitivity of the pipelines to the disadvantage of ‘specificity’ or vice versa. |
| 2019 | CNN, Feature Extraction, HSV format | Accuracy of 98%. for melanoma skin cancer detection and 93% for melanoma type, TPR of 94.25%, FPR of 3.56%, and EP of 4%, average accuracy of 91.66% | high error rates, 25.6% Caucasian error and 23.2 Xanthous error, validation loss of 57.56% | Accuracy | achieved TPR of 94.25%, FPR of 3.56%, and EP of 4% | With a small datasets an accuracy of 74.76% and validation loss of 57.56% is aquired |
| 2019 | CNN, keras, AlexNet, VGG16, SGD optimiser, | trained on more than 126k images, higher image augmentation (24x) and image resolution (1k), the same performances can be achieved using less than 5000 images, no impact of image resize filters | Experiments at 277x277 pixel resolution, Experiments without transfer learning | Accuracy | 98% specificity | 73% sensitivity and Jaccard Index of 0.69. |
| 2019 | CNN, grad-CAM, TensorFlow, Inception-ResNet-v2, DenseNet121, Xception | consists of 150,223 clinical images from 543 different skin diseases, achieved an accuracy of 87.25 ± 2.24% on the dermoscopic images for four common skin diseases, including SK, BCC, psoriasis and melanocytic nevus. | highest average precision (77.0%) | Accuracy | achieved 92.9%, 89.2%, and 84.3% recalls for the LE, BCC, and SK, respectively, | mean recall and precision reached 77.0% and 70.8%. |
| 2019 | CNN, keras, TensorFlow, Inception V3, ResNet50, VGG16, MobileNet and InceptionResnet | 7 types of skin lesion diseases identification namely: Benign Keratosis, Dermatofibroma, Vascular Lesion, Melanoma, Melanocytic Nevus, Basal Cell Carcinoma and Actinic Keratosis., InceptionResnet achieved an average accuracy of 91%, Accuracies of 90 and 91% | low F1 score | Accuracy | This model is advantageous over feed-forward neural networks which cannot understand translation invariance | Low F1 Scores |

## Grouping

### oN THE bASIS OF dATA uSED

The groups for this which can be formed are:

* Dataset is classified into classes and stored separately beforehand
  + Data is classified as:
    - Malignant or Benign
    - 7 classes of Skin Cancer
    - 3 classes of keratosis
* Dataset is not classified into classes and instead CSV file is provided

### oN THE bASIS OF dATA aNALYSIS tECHNIQUE

Some the analysis methods used are:

* ABCDE Rule
* Data Augmentation
* Normal scanning and cropping and photoshop
* CNN (common)
* Deep Learning Pipeline (Morphological Analysis) (common)
* Biopsy, Histopathological Testing, Dermoscopic Assessment
* StyleGANS (common)
* GLCM, SVM

### On the basis of Data Pre-processing Technique

Some of the pre-processing methods can be grouped as:

* Dull Razor Method (common)
* Transfer Learning (Most Common)
* Hyper-parameters for Image Augmentation and CNNs (common)
  + CNN to Binary Output
* Adam, SGD, RMSprop
* DCNN
* ANN, SVM, Naïve-Bayes Algorithm
* One-Hot Encoding
* Characteristic Curve
* Noise Removal
* Segmentation
* Resize, Feature Extraction, Classification (Common)

### On the Basis of Feature Selection Techniques

Some of the feature selection methods can be grouped as:

* Classification
  + Binary
  + SoftMax classifier
* CNN, Pooling Layer (common)
  + Max pooling (most common)
  + Sum pooling
  + Average pooling
* Autoencoders
  + Stacked Deep Autoencoders
* No. of Hidden layers in Dense and Sparse network
* Multilayer Perceptron
* Algorithms
  + SGNN
  + Genetic
  + Skin Lesion Segmentation’
  + GLCM

### On the Basis of Model Training Method

Some of the training methods can be grouped as:

* MVSM
* MCNN
* CNN (Most Common)
* Transfer Learning: (used as a group of 2-3 with CNN)
  + Inception (V1,V3)
  + ResNet
  + ResNet50
  + VGG16BN
  + VGG16
  + VGG
  + MobileNet
  + GoogLeNet
  + AlexNet
* Deep Pipeline
* SGD

### On the Basis of Model Evaluation Method

Some of the evaluation methods can be grouped as:

* Precision & Recall
* F-score
* Accuracy (Most Common)
* Random Forest, XGBoost, SVM
* Specificity & Sensitivity
* Confusion Matrix
* ABCDE Criteria
* Rare:
  + Jaccard similarity coefficient (JSC)
  + geometric mean (G-mean)
  + Matthew’s correlation coefficient (MCC)
  + Cohen’s kappa score (CKS)
  + AUROC
  + precision-recall curve (PR-AUC)
  + evaluation time

# Future Scope

* Implementation of various other algorithms and using several optimization techniques. Also, more data will be collected in order to recognize the features more accurately.
* Major attention will be given to increase the accuracy such that our proposed system can be used to detect a large number of chronic and critical diseases.
* When these enhancements are done, the system can be integrated with an android application to make it more convenient and easily portable. This will allow people from all strata to use it effectively even if they do not have a personal computer.

# Summary on Literature Survey based on the General Arhitecture Processes

## Data Selection

The data selection is done on the basis of the amount of data and the type of data which is available. The data could be in the form of images, scanned reports, tf records, dcm files, etc. Based on the algorithm selected and the kind of data available, the model will be built. The data obtained can be collected via survey or from public databases.

## Exploratory Data Analysis

### Checking the Types of Data

To find what all columns it contains, of what types and if they contain any value in it or not, with the help of functions.

### Finding the Outliers

An outlier is a piece of data that is an abnormal distance from the other points. In other words, it’s data that lies outside the other values in the set. These points can be found by plotting the entire data.

### Data Visualization

Using this data, we can:

* Analyse individual columns
* Check for missing values
* Perform variable analysis
* Check condition column
* Check quality column
* Plot between different variables and targets

## Data Pre-Processing

### Splitting the Data

It is very important because your model needs to be evaluated before it has been deployed. And that evaluation needs to be done on unseen data because when it is deployed, all incoming data is unseen. The main idea behind the train test split is to convert the original data into training and testing data. For most of the articles which have been analysed, the data has been split into training and testing data in the range of ration of 75% to 25% (this is an approximate range provided considering all the research papers analysed), respectively.

### Checking for Missing Values

If your data set is full of NaNs and garbage values, then surely your model will perform on garbage too. So, taking care of such values is important and it mostly done using the Simple Imputer method.

### Checking Categorical Features

The most common methods used for this are:

* Label Encoding
* One-hot Encoding

### Normalizing Dataset

The models mostly use the following methods of normalization for the data:

* Standard Scaler
* Variance before Standard Scaler
* Variance after Standard Scaler

## Feature Transformation

Feature pre-processing is one of the most crucial steps in building a Machine learning model. Too few features and your model won’t have much to learn from. Too many features and we might be feeding unnecessary information to the model. Not only this, but the values in each of the features need to be considered as well.

### Why do we need Feature Transformation and Scaling?

Oftentimes, we have datasets in which different columns have different units – like one column can be in kilograms, while another column can be in centimeters. Furthermore, we can have columns like income which can range from 20,000 to 100,000, and even more; while an age column which can range from 0 to 100(at the most). Thus, Income is about 1,000 times larger than age.

But how can we be sure that the model treats both these variables equally? When we feed these features to the model as is, there is every chance that the income will influence the result more due to its larger value. But this doesn’t necessarily mean it is more important as a predictor. So, to give importance to both Age, and Income, we need feature scaling.

### Feature Transformations used in the Models

#### MaxAbsScalar

In simplest terms, the MaxAbs scaler takes the absolute maximum value of each column and divides each value in the column by the maximum value.

Thus, it first takes the absolute value of each value in the column and then takes the maximum value out of those. This operation scales the data between the range [-1, 1].

#### Robust Scalar

If you have noticed in the scalers we used so far, each of them was using values like the mean, maximum and minimum values of the columns. All these values are sensitive to outliers. If there are too many outliers in the data, they will influence the mean and the max value or the min value. Thus, even if we scale this data using the above methods, we cannot guarantee a balanced data with a normal distribution.

The Robust Scaler, as the name suggests is not sensitive to outliers. This scaler-

removes the median from the data

scales the data by the Interquartile Range(IQR)

Are you familiar with the Inter-Quartile Range? It is nothing but the difference between the first and third quartile of the variable.

#### Unit Vector Scaler

Normalization is the process of scaling individual samples to have unit norm. The most interesting part is that unlike the other scalers which work on the individual column values, the Normalizer works on the rows! Each row of the data frame with at least one non-zero component is rescaled independently of other samples so that its norm (l1, l2, or inf) equals one.

Just like MinMax Scaler, the Normalizer also converts the values between 0 and 1, and between -1 to 1 when there are negative values in our data.

However, there is a difference in the way it does so.

* If we are using L1 norm, the values in each column are converted so that the sum of their absolute values along the row = 1
* If we are using L2 norm, the values in each column are first squared and added so that the sum of their absolute values along the row = 1

## Feature Selection

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modelling your data are:

* **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
* **Improves Accuracy**: Less misleading data means modelling accuracy improves.
* **Reduces Training Time**: Less data means that algorithms train faster.

The methods used for Feature Selection are:

* Principal Component Analysis
* Linear Discriminant Analysis

### Principal Component Analysis

Principal Components Analysis is a way of recognizing patterns in data, and expressing the data in such a manner as to focus their differences and similarities. Subsequently patterns in data may be complex to discover in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The key advantage of PCA is that once we have found the patterns in the data, and you compress the data, i.e. by reducing the number of dimensions, without much loss of information. This technique used in image compression.

### Linear Discriminant Analysis

There are many possible techniques for classification of data. Principal Component Analysis and Linear Discriminant Analysis are commonly used techniques for dimensionality reduction and data classification. Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This technique maximizes the proportion of between-class variance to the within-class variance in any specific data set in that way promising maximal separability. The Linear Discriminant Analysis is used for classification issues such as speech recognition. The key difference among LDA and PCA is that PCA perform feature classification and LDA works for data classification. The shape and location of the inventive data sets changes when transformed to a different space in PCA, on the other hand LDA doesn’t change the location but only attempts to offer more class separability and induce a decision region among the given classes. This technique also supports to better recognize the distribution of the feature data.

## Model Selection

Some of the models which were deployed are:

* Simple Convolutional Neural Network Models
* Transfer Learning Models
* Ensemble Models
* Simple K-Means Model
* Generative Automotive Networks

## Model Training

The models deployed one of the following training techniques:

* Infected Area Detection
* Image Classification
* Instance Segmentation

## Model Evaluation

Some of the most common evaluation methods are:

* Accuracy
* Sensitivity
* Specificity
* Recall
* Precision
* F-measure

Some of the rare evaluation methods used are:

* CNN with/without Data Augmentation
* ResNeXt WSL
* ABCD Rule
* GOPS
* L1D Miss Rate
* XGBoost
* GoogLeNet/ResNet/AlexNet/VGGNet Error Rate

# Conclusions:

* A “health discernment system” has been proposed for medical image classification that will work in real-life scenarios.
* The proposed method is based on ***Convolutional Neural Network*** architecture.
* Different sub-models pertaining to the two diseases (skin cancer: Melanoma, Benign) have been designed using convolutional neural network (CNN) and they have all been tested separately.

# References:

* ResearchGate.com
* ScienceDirect.com
* GeeksforGeeks.com
* TutorialsPoint.com
* cs.stanford.edu
* Springer.com
* towardsdatascience.com
* academictorrents.com
* kaagle.com