

A Study of Social Media Data and Data Mining Techniques

Mr. Prasad A. Joshi, Mrs. Swati N. Bhamare

Assistant Professor, Department of Computer Science & I.T., JET's Z. B. Patil College, Dhule-424002(M.S.)

ABSTRACT

Artificial Neural Networks (ANNs) has highly interconnected elements (neurons) which unanimously work to solve the specific problems. Recently ANNs are involved in the areas like image and speech recognition, character & pattern recognition with statistical analysis and data modeling for solving the problems related to forecasting & classification. In this paper, we are focusing on learning process of a neural network.

Keywords: ANNs, Neurons, Pattern recognition, Supervised and unsupervised learning.

I. INTRODUCTION

The concept of ANN is basically introduced from the subject of biology where neural network plays a important and key role in human body. In human body work is done with the help of neural network. Neural Network is just a web of inter connected neurons which are millions and millions in number. With the help of this interconnected neurons all the parallel processing is done in human body and the human body is the best example of Parallel Processing . [2]

II. HOW THE HUMAN BRAIN LEARNS

In the human brain, a typical neuron collects signals from others through a host of fine structures called dendrites. The neuron sends out spikes of electrical activity through a long, thin stand known as an axon, which splits into thousands of branches. At the end of each branch, a structure called a synapse converts the activity from the axon into electrical effects that inhibit or excite activity from the axon into electrical effects that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.[3]

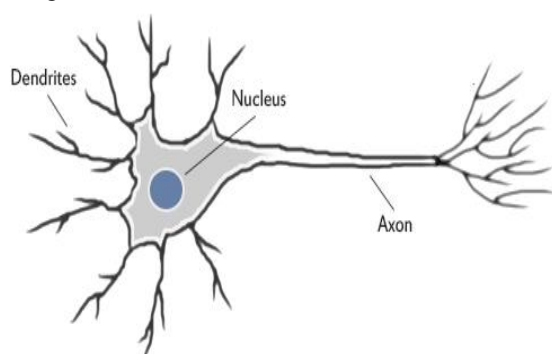


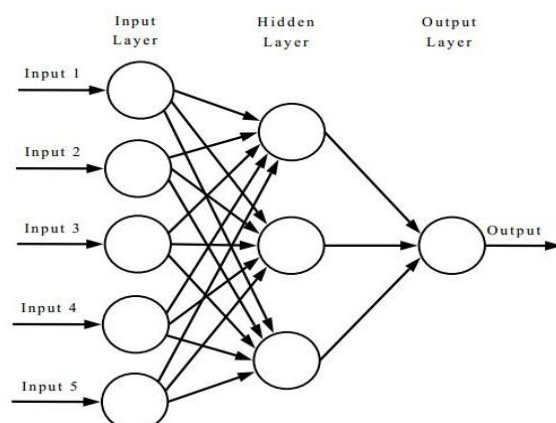
Fig: A biological Neuron

III. WHAT IS ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information.

The neural network consist of three layers, called as input layer, hidden layer, and output layer. The input nodes take in information, in the form which can be numerically expressed. The information is presented as activation values, where each node is given a number, the higher the number, the greater the activation. This information is then passed throughout the network. Based on the connection strengths (weights), inhibition or excitation, and transfer functions, the activation value is passed from node to node. Each of the nodes sums the activation values it receives; it then modifies the value based on its transfer function.

The activation flows through the network, through hidden layers, until it reaches the output nodes. The output nodes then reflect the input in a meaningful way to the outside world.[4]



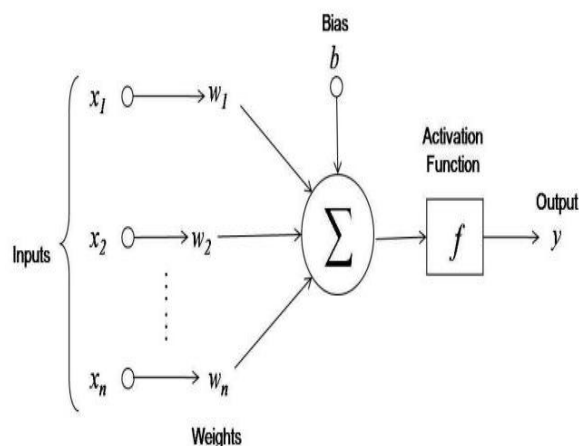
Fig[a]: Multilayered Artificial Neural Network

Why ANN?

The ANN has given many best and excellent characteristics to brain of human being which are not present in modern computers which are : [2]

- 1) Learning Ability
- 2) Generalization Ability
- 3) Adaptive Learning
- 4) Fault Tolerance
- 5) Self Organization

This is a flow diagram of active nodes used in hidden and output layers of neural network. Each input is multiplied by a weight and then summed. This produces a single value that is passed through an activation function.



IV. TYPES OF NEURAL NETWORKS

There are different types of neural networks, but they are generally classified into feed-forward and feed-back networks.

A feed-forward network is a non-recurrent network which contains inputs, outputs, and hidden layers; the signals can only travel in one direction. Input data is passed onto a layer of processing elements where it performs calculations. Each processing element makes its computation based upon a weighted sum of its inputs. The new calculated values then become the new input values that feed the next layer. This process continues until it has gone through all the layers and determines the output. A threshold transfer function is sometimes used to quantify the output of a neuron in the output layer. Feed-forward networks include Perceptron (linear and non-linear) and Radial Basis Function networks. Feed-forward networks are often used in data mining. [4]

A feed-back network has feed-back paths meaning they can have signals traveling in both directions using loops. All possible connections between neurons are allowed. Since loops are present in this type of network, it becomes a non-linear dynamic system which changes continuously until it reaches a state of equilibrium. Feed-back

networks are often used in associative memories and optimization problems where the network looks for the best arrangement of interconnected factors.[4]

The Learning Process:

All learning methods used for neural networks can be classified into two major categories:

Supervised learning:

In this, every input pattern that is used to train the network is associated with an output pattern, which is the target or the desired pattern. A teacher is assumed to be present during the learning process, when a comparison is made between the network's computed output and the correct expected output, to determine the error. The error can then be used to change network parameters, which result in an improvement in performance. [5]

Unsupervised learning:

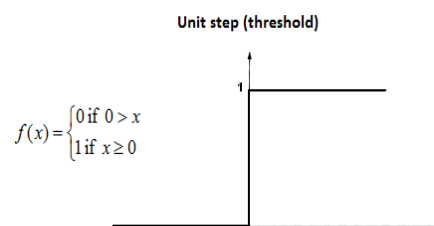
In this learning method, the target output is not presented to the network. It is as if there is no teacher to present the desired patterns and hence, the system learns of its own by discovering and adapting to structural features in the input patterns. [5]

V. TRANSFER (ACTIVATION) FUNCTIONS

The behavior of an ANN (Artificial Neural Network) depends on both the weights and the input-output function (transfer function) that is specified for the units. Four types of transfer functions are commonly used, Unit step (threshold), sigmoid, piecewise linear, and Gaussian.

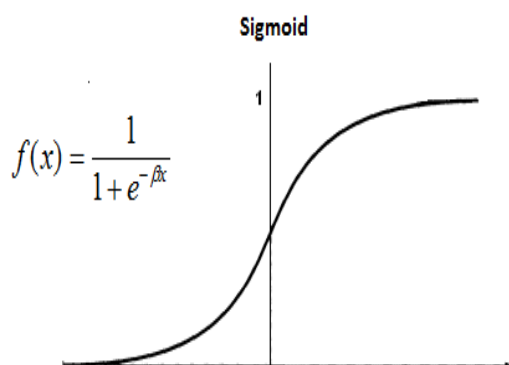
1) Unit step (threshold):

The output is set at one of two levels, depending on whether the total input is greater than or less than some threshold value.[4]



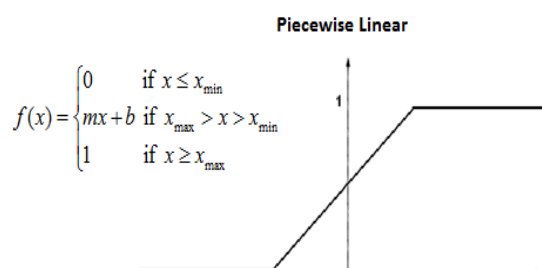
2) Sigmoid:

The sigmoid function consists of 2 functions, logistic and tangential. The values of logistic function range from 0 and 1 and -1 to +1 for tangential function.[4]



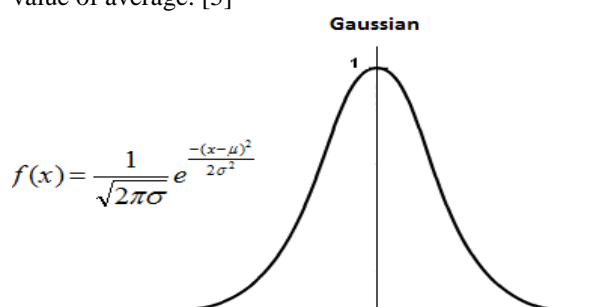
3) Piecewise Linear:

The output is proportional to the total weighted output.[4].



4) Gaussian:

Gaussian functions are bell-shaped curves that are continuous. The node output (high/low) is interpreted in terms of class membership (1/0), depending on how close the net input is to a chosen value of average. [5]



We can teach a three-layer network to perform a particular task by using the following procedure:

1. We present the network with training examples, which consist of a pattern of activities for the input units together with the desired pattern of activities for the output units.
2. We determine how closely the actual output of the network matches the desired output.
3. We change the weight of each connection so that the network produces a better approximation of the desired output.

VI. APPLICATIONS OF NEURAL NETWORKS

Image (data) Compression:

Neural networks can perform real-time compression and decompression of data. These networks convert eight bits of data by reducing to three bits and then reverse again to eight bits.[6]

Pattern Recognition:

Many example acknowledgment applications are being used like, a framework that can distinguish bombs in baggage at airplane terminals by recognizing from little differences and examples from inside particular sensor's yields, a back-proliferation neural system which can segregate between a genuine and a false heart assault, a system which can examine and furthermore read the PAP smears and so on. Many robotized quality control applications are presently being used, which are based on pattern recognition. [6]

Language Processing:

These applications include text-to-speech change, acoustic input for machines, automatic language translation, secure voice keyed locks, automatic transcription, aids for the deaf and the physically disabled which respond to voice commands and natural language processing. [6]

Character Recognition:

By using Neural network based products one can identify the hand printed characters through a scanner. It is 97% accurate for numbers, and up to 92% for alphabetical characters. Quantum Neural Network software package (Qnspec) is available for recognizing characters, including cursive characters. [6]

Signal Processing:

Neural networks are also capable of filtering out electronic noise. [6]

Financial: Banks, credit card companies and lending institutions deal with many decisions that are not clear-cut. They involve learning and statistical trends. Neural networks are now trained on the data from past decisions and being used in decision making.[6]

Conclusion

This paper's content will motivate the researcher to work with ease. The techniques given are applicable for almost all types of research areas related to ANN. Neural networks do not perform miracles. But if used sensibly they can produce some amazing results.

REFERENCES

- [1]. https://www.inf.ed.ac.uk/teaching/courses/nlu/reading/Gurney_et_al.pdf (page no. 17)
- [2]. International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Volume 2, Issue 10, October 2012 ISSN: 2277 128X Available online at: www.ijarcsse.com
- [3]. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
- [4]. http://chemeng.utoronto.ca/~datamining/dmc/artificial_neural_network.htm
- [5]. http://www.iasri.res.in/ebook/EBADAT/5-Modeling%20and%20Forecasting%20Techniques%20in%20Agriculture/5-ANN_GKJHA_2007.pdf
- [6]. <http://shodhganga.inflibnet.ac.in/bitstream/10603/48>

Social media is an extraordinary way of communication among the society. It makes a easy way for the individuals to participate in online world by sharing, posting, commenting, blogging, liking, tweeting the information in different formats (text, audio, video, photos, mails, multimedia, etc) in any place at any time. There are various components which has made the study of social media data increase tremendous significance by analysts. Data mining is an interdisciplinary subfield of computer science. In data mining we extract the information from a huge data set and apply the suitable techniques to find knowledge & patterns for further use. Data mining is an integral part of many related fields including statistics, machine learning, pattern recognition, database systems, visualization, data warehouse, and information retrieval [1].

VII. SOCIAL MEDIA AND ITS CHARACTERISTICS

Social media is defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchanges of user-generated content.[2] It gives users in the society an easy-to-use way for communication and provides network on an extraordinary scale and at rates never seen in traditional media. The ubiquity of social media keeps on developing exponentially causing in an evolution of wikis, social news, microblogging, media sharing, etc. Following table shows different types of social medias along with their characteristics and type of data used by those social sites given in below Table 1.

Table 1: Social Medias & their characteristics

Social Medias	Characteristics of Social Media	Popular Social Networking Sites	Format of Data
Wikis	It allows multiple users to develop Web pages by providing editing environment	Wikipedia, Wikitravel, Wikihow	Text
Social news	Users can share and selects news stories and articles	Digg, Slashdot, Reddit	Link or Text
Microblogging	Same as blogs but with limited content.	Twitter, Tumblr, Plurk	Text, Pictures, Emoticons
Media sharing	It Shares variety of media on the Web	YouTube, Flickr, UstreamTV	video, audio, and photo
Blogging	Maintained by users or community	Huffington Post Business Insider, Engadget	Mainly text, but multimedia is allowed.
Online social networking	Allows to connect with real-world friends and associates online. Users interact with each other through status updates, comments, media sharing, messages, etc.	Facebook, Myspace, LinkedIn	pictures, videos, music websites, articles, blogs, groups, etc.
Social bookmarking	Allow users to bookmark Web Content for storage, organization, and sharing	Delicious, StumbleUpon	In the form of links

Social Media Data: This data can be gathered with the help of traditional networks and social networks. Traditional networks include newspaper, News Channels, Movies, Radio, TV, FAX and television, etc. While social network sites are discussed in the above table 1.

The popularity of social media continues to grow exponentially, in which user are added at each second; following pie chart(Figure 1) represents active users (in millions) on different social networking sites as of 2017. As we can see quantity of active users, thereby the data generated by the active users is also huge. Some of the statistics of data generation on these sites is shown in Table 2:

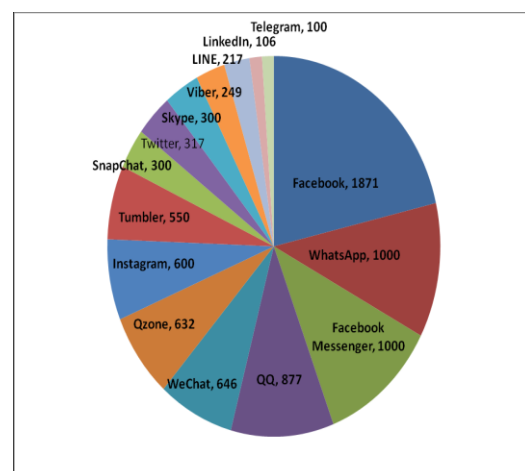


Figure. 1: Quantity of Active Users (in millions) on Popular Social Networking Sites as of 2017 [3]

Table 2: Data Generated by Social Networking Sites [4]

Social Networking Sites	Data Generated Per Minute of The Day
FACEBOOK	4,166,467 Post
TWITTER	347222 Tweets
YOUTUBE	300 User Upload Videos
INSTAGRAM	1,736,111 User Like Photos
APPELE	51000 User Download App
NETFLIX	77,160 Subscribers
REDDIT	18,327 User Vote
AMAZON	4,310 Visitors
VINE	1041666 Play Videos
TINDER	590278 User Swipe Times
SNAPCHAT	284722 USER Share Snaps
BUZZFEED	34150 User View Video
SKYPE	110040 User Make Calls
UBER	694 Passenger Take Rides

VIII. CHALLENGES IN SOCIAL MEDIA DATA

Unfortunately Social media data is variant in nature from the traditional data because it is gathered from different social media sites and in different formats. Apart from the fact that it is huge in size, it has some novel challenges which are elaborated below:

Big Data: Social media data is certainly big and huge. Whenever we zooms into individuals little amount of data for an individual appears. Obtaining Sufficient Samples: Data can be collected through application programming interfaces (APIs) from social media sites. Only a limited amount of data can be obtained daily. Also how to rely on samples? How can we ensure that our findings are of true patterns that can benefit our research or business development?

Noise Removal Fallacy: Noise in social media data means irrelevant post, tweets, messages, anything. But removing the noise can make the data useless because of inherently linked nature of social media data. It complicates the task of noise removal and requires approaching noise-removal differently in comparison with attribute-value data.

Evaluation Dilemma: Traditional data mining often use a portion of the dataset for testing. But with social media data, traditional test data may not be viable. Evaluating patterns from social media data can cause apparently undefeatable challenge because without credible evaluation, how can we guarantee the validity of the patterns?

Unstructured: User generated content in social media is often highly unstructured. Nowadays more and more users use their mobiles to publish content such as updating statuses in Facebook, sending

tweets in Twitter and commenting on posts, which results in 1st short texts and 2nd ty-pos and spacing errors occurring very frequently [5].

Incomplete: Users' attributes are predictable with their personal data [6]. To address such privacy concerns, social media services often allow their users to use their profile settings to mark their personal data such as demographic profiles, status updates, lists of friends, videos, photos, and interactions on posts, invisible to others.

Missing Values: Those values that are missing in instances. For example, individuals may avoid filling profile information on social media sites, like as their age, location, or hobbies. To solve this problem, we can (i) Remove instances that have missing values, (ii) Guess missing values or (iii) Ignore missing values.

Outliers: are instances that are considerably different from other instances in the dataset. E.g: While measuring the likes an individual have for any post. In that case a celebrity will be considered as outlier because he can get maximum likes because of its popularity as compare to individual.

Due to this nature of social media data; we need to pre-process it.

Data mining: It is a process used to discover actionable knowledge in large-scale data. It is the analysis step of the knowledge discovery in databases process, or KDD.[7] It describes the typical process of extracting useful information from raw data. The KDD process broadly consists of the following tasks: data pre-processing, data mining, and post processing.

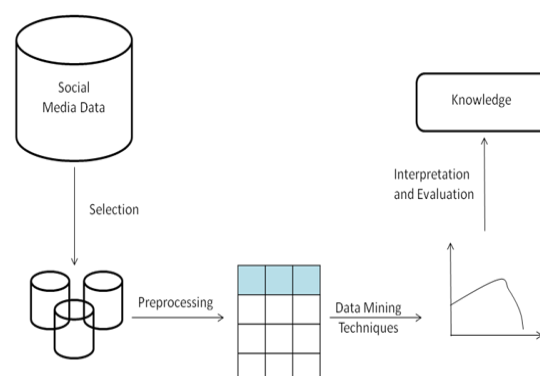


Figure. 2: Data Mining Process

Pre-processing of Social Data: After resolving the challenges we need to pre-process data before applying to data mining techniques. Typical data pre-processing tasks are as follows:

Aggregation: It is performed when multiple features are combined into single feature or when the scale of feature is changed.

Discretization: It is the process of transforming quantitative data to qualitative data.

Feature Selection: Mostly all of the features are not useful, because some of them have less computational power or some are irrelevant hence the selection is made which will increase the performance of data mining technique.

Feature Extraction: In contrast to feature selection, feature extraction converts the current set of features to a new set of features that can perform the data mining task better[8]. A transformation is performed on the data, and a new set of features is extracted.

Sampling: Social media data is big in nature, so to process the complete data is costly. Hence sampling is needed. In sampling, a small random subset of instances are selected and processed instead of the whole data. The selection should be made in such a way that results obtained on the samples are close to obtained on the complete dataset.

Data mining Techniques: Now we can apply the appropriate techniques to the data. Techniques can be divided into Supervised Learning and Unsupervised Learning.

Supervised learning: In this a given data set is typically divided into two parts: training and testing data sets with known class labels.[9] Supervised algorithms build models from the training data and use the learned models for prediction. It can be divided into classification and regression.

IX. CLASSIFICATION

In this given data is categorized into different classes according to a classification model. When the class attribute is discrete, it is called classification.

Popular classification algorithms are - Decision Tree, Naive Bayes classifier, K-nearest neighbour.

Decision tree: It has nodes that forms directed tree with a node called root that has no incoming edges. The technique is used with review sets of positive and negative as leaf nodes. The most significant review formed the root of the tree.

Naive Bayes classifier: It is used for large data set. Naïve Bayes counts the occurrences of values and combinations of values in historical data in order to use conditional probabilities. It is regarded as one of the three mostly employed supervised learning technique.[10]

K-nearest neighbour: It utilizes the neighbors of an instance to perform classification. In particular, it uses the k nearest instances, called neighbors, to perform classification. To determine the neighbors of an instance, we need to measure its distance to all other instances based on some distance metric. Commonly, Euclidean distance is employed

Regression: In regression the object to be predicted is continuous rather than discrete.

Unsupervised Learning: Unsupervised division of objects into groups of similar objects is called unsupervised learning. Clustering is the form of unsupervised learning.

Clustering: It can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used

X. CONCLUSION

This Paper Focus on complex information hidden in social media data. With the help of above mentioned techniques one can simplify his work to generate desired data and consistent performance. This paper motivates and stimulates researchers to review the various data mining techniques for their further work in this area.

REFERENCES

- [1]. J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2011.
- [2]. A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons* 53(1):59{68, 2010.
- [3]. <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> visited on 28/2/2016
- [4]. <http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/> visited on 28/2/2016
- [5]. D. Kim, D. Kim, E. Hwang, and S. Rho. Twitter trends: a spatio-temporal trend detection and related keywords recommendation scheme. *Multimedia Systems*, 2014.
- [6]. M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.
- [7]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008
- [8]. Social Media Mining: An Introduction By Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, Pg No:142
- [9]. Mining Social Media: A Brief Introduction: Pritam Gundecha Huan Liu
- [10]. Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques

for social media analysis. arXiv preprint
arXiv:1312.4617.

REFERENCES

- [11]. Jagadeesh. D. Pujari Reduced Color and Texture features based Identification and Classification of Affected and Normal fruits' images International Journal of Agricultural and Food Science -2013.
- [12]. Jundare Manisha. A. Jundare Pallavi T. Jundare Pragati V. Prof. C.S.Aryan, Plant Disease Detection and its Treatment using Image Processing ,International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering- 2016.
- [13]. SUVARNA KANAKARADDI ANALYSIS AND GRADING OF PATHOGENIC DISEASE OF CHILLI FRUIT USING IMAGE PROCESSING International Conference on Advances in Engineering & Technology- 2014.
- [14]. Smita Naikwadi, Niket Amoda, ADVANCES IN IMAGE PROCESSING FOR DETECTION OF PLANT DISEASES, International Journal of Application or Innovation in Engineering & Management - 2013.
- [15]. K. Jagan Mohan, Recognition of Paddy Plant Diseases Based on Histogram Oriented Gradient Features International Journal of Advanced Research in Computer and Communication Engineering -2016.
- [16]. Sujeet Varshney, A Novel Approach for the Detection of Plant Diseases, International Journal of Computer Science and Mobile Computing -2016.
- [17]. Vyshnavi.G.K, Sirpa.M. Chandramoorthy, Padmapriya. Healthy and Unhealthy Plant Leaf Identification and Classification Using Hierarchical Clustering, International Research Journal of Engineering and Technology