IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 2, April 2019
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Automatic Colorization of Black and White Images using Deep Learning

[1] Sindhuja Kotala; [2] Srividya Tirumalasetti; [3] Vudaru Nemitha; [4] Swapna Munigala

[1] CSE Department, Osmania University- Stanley college of Engineering and Technology for Women, B.E III Year, Hyderabad, Telangana 500036, India

[2] CSE Department, Osmania University- Stanley college of Engineering and Technology for Women, B.E III Year, Hyderabad, Telangana 500036, India

[3] CSE Department, Osmania University- Stanley college of Engineering and Technology for Women, B.E III Year, Hyderabad, Telangana 500036, India

[4] CSE Department, Osmania University- Stanley college of Engineering and Technology for Women, Asst Prof CSE Dept, Hyderabad, Telangana 500036, India

Abstract - The main aim of our paper is to give an overall idea about how a grayscale image can be converted into a colorful image with the colorization problem and how it can further be used to color a video. To achieve artifact-free quality this generally requires manual reconciliation and therefore considered as a very strenuous problem . A cautious selection of colorful allusion images are generally required for the process. Far from the preceding methods, this paper aims at a high grade fully unmanned colorization method and also attempt to apply this concept to images obtained from video sequences. The recent achievements in deep learning approaches is the inspiration behind this paper, that focuses on reformulation of the problem of colorization so as we can employ the deep learning approaches promptly and that this technique can be applied on to the videos. Our proposed method is a fully automated process. To our best apprehension, no prevailing papers or research studies label this issue of using deep learning techniques to colorize videos.

Keywords **- Deep learning, Convoluted neural networks, Machine learning.**

1. Introduction

A gizmo recognized as Ferranti Mark 1 used an algorithm successfully to the game master checkers. Newell and Simon developed an algorithmic program to fathom mathematical issues. In addition, the LISP programming language was developed by John McCarthy in 50s, which later became significant in machine learning. In the year of 1960, researchers started developing algorithms which has the capability to elucidate mathematical complications and geometrical postulates. In late 1960s, scientists worked on Machine Vision Learning and developing robots for machine learning. WABOT-1 was the first 'intelligent' humanoid robot, which was built in Japan in the year of 1972.According to John McCarthy, the father of AI "The science and engineering of making intelligent machines, especially intelligent computer programs" is called as Artificial Intelligence. It is a way of making a computer, a computer-controlled robot, or a software that can think intelligently[15]. The main aim of Artificial Intelligence is to make a robot think like a human. It can be accomplished by studying how a human brain thinks, and how humans learn, decide, and work when they are trying to solve a critical problem.

The goals used for creating AI are, first to create an expert system and then implementing human intelligence in machines. The contributions to AI are in various fields, such as, maths, biology, neuroscience, psychology, computer science, philosophy, sociology, etc.,. AI knowledge has few unwelcomed properties in the real world. They are, it has a huge volume which is next to unimaginable, knowledge is not well-organized and the knowledge constantly keeps changing[1]. To use this knowledge efficiently and organise, we use AI techniques. To correct errors it is ought to be easily modifiable. Though the techniques are incomplete or inaccurate, they should be functional in numerous situations. The complex programs equipped with AI need to be executed at an elevated speed by using AI techniques.

AI has many applications. AI is used in many games which include chess, poker, missionaries and cannibals, water jug problem, 8 puzzled game, etc,. Natural language processing is a major application which makes it possible

to interact with the computer by the personage in his language. Expert systems are those which use AI to solve critical problems and understand situations that would take a human a lot of experience and expertise to complete a task. Few intelligent systems can not only recognise the speech but also can handle accents, slang words, change in human's pitch due to flu,etc[15]. Handwriting recognition software uses AI that can scrutinize the text written by personage. It can then transmute it into modifiable text form.

Machine learning is a sub division of artificial intelligence (AI) which helps the systems to assimilate viscerally & enhance from incident without being programmed by a human. It focuses on developing the computer programs which can entrance data and utilize it as a tool to learn.The learning process begins with observing the data in order to look for repeated design and make better resolutions in the future. Examples of data observations are direct experience, instruction, etc. The main goal is to have no person interference or aid and that the computers learn automatically and adjust actions accordingly.We got self-driving cars, effective web search, practical speech recognition, and a hugely upgraded comprehensiveness of the human ordination with the help of machine learning, in the past decade. Applications of machine learning are, vision processing, language processing, forecasting, pattern recognition, games, data mining, expert systems, robotics, etc,. Machine learning algorithms are, supervised machine learning, unsupervised machine learning, semi-supervised machine learning, reinforcement machine learning.[16]

Deep learning is an existing function of AI that works similarly like a human brain. Example, it processes the data and creates patterns for the use in decision making. Machine learning is the superset of deep learning. Deep learning has networks which are capable of learning independently form of data that is unlabeled. Deep neural learning is also called as Deep neural network. Deep learning is a set of algorithms which is utilized in machine learning which can be used to replica high level premonition in data by the usage of model architecture. It is a specific approach which is used for constructing and instructing neural networks which are considered to be highly promising decision making nodes. An algorithm is recognised to be deep if the output is obtained when a given input is made to pass through a series of non linear transformations [14]. In contrast, most of the contemporary machine learning algorithms are evaluated to be "shallow". As the input can only get through hardly any stages of subroutine calling. In data, using deep learning we can remove manual identification of features. Deep Learning is evaluated to be as a traditional learning. High performance hardware is needed by deep learning. New features are created by deep learning itself. More time is taken by deep learning to train.

Deep learning training process includes few stages. They are, artificial neural networks prompt few binary true/false questions, withdrawing numerical values from data blocks, classification of data as stated to the received answers and labelling of data[1]. Characteristics of deep learning are, deep learning always looks for meaning, it focuses on the concepts and arguments which are centrally needed to solve the problem, deep learning helps in active interaction, helps us to differentiate among altercation and proof. We can also form connections between various modules using deep learning. Through deep learning we can relate new and previous knowledge. Deep learning links real life to course content.

The advantages of deep learning are, it provides top level accomplishment on issues that surpasses other solutions in several domains significantly, like speech language, vision, playing games, etc., it reduces one of the major time occupying areas of machine learning practice which is feature engineering, it is a model that can be reshaped to new problems that are comparatively easily. The disadvantages of deep learning are, continuous input data management, the training process in deep learning is based on examining huge quantity of data, the deep learning algorithms are to be adapted by the data scientists such that neural networks can handle high amounts of continual input data, ensuring conclusion transparency, another drawbacks of deep learning software is that it is impotent of supplying reasons for reaching a certain decision[14].

Deep learning technology demands high resources. It requires high-performance and more powerful GPUs, large amounts of space to store the data that is used to teach the models, so on. Unlike the traditional machine learning, this technology takes more time to be trained[16]. Though deep learning has all the above mentioned challenges,it is still being used because it has been discovering new improved methods of unstructured big data analytics day-by-day. Many organizations and businesses gain significant benefits through deep learning. Implementations of deep learning are, it automatically adds sound to silent movies or videos, it can perform automatic machine translation, it can classify objects and detects photographs, it can generate handwriting and text automatically, it can also generate captions for images, it can create chatbots and can also recognise pictures of the similar person.

2. Literary Survey

Though convolutional neural networks (often shortened to ConvNets, CNNs, dCNNs) was not entirely new

technology, it had taken the world by storm since 2012. CNNs now developed which are utilized in different digital applications were actually first observed in living organisms. In the 1950s & 1960s, Hubel & Wiesel have worked on cat and monkey which showed that their visual cortexes contained neurons that separately reciprocate to small areas of the visual field. The visual stimuli in the area of visual space affects the firing of a single neuron provided their eyes are not moving. This is its receptive field. Likely and intercepting receptive regions have been observed in the neighbouring cells. Each hemisphere in the cortex represents the contralateral visual field. This lead to the introduction of neocognitron, delay in the time of neural networks and trainable weights. All these formed basics to the first ever documented commercial use of CNNs which dates back to 1998, with LeNet-5. LeNet-5 was designed by LeCun et al. Years of researching CNNs, made it possible to recognise text character based on 32x32 pixel images. Since then, CNNs have dominated this area of computer vision and surpassed results obtainable by other machine learning methods[4]. But the large scale application of CNNs were not possible until a decade later. In 2012, when the CNN based model entry of Alex Krizhevsky et al and their AlexNet in a ImageNet Large Scale Visual Recognition Challenge won that year's image classification challenge by a significant margin, it took most of the computer vision research community by surprise and lit a huge amount of interest. This event made CNN a staple in the computer vision field and many others. Countless problems, such as image recognition, facial recognition, video sequence tracking, automatic image segmentation, handwriting to text conversion, natural language processing have been made possible to solve easily with the help of CNN. CNNs have proven to be functioning as automatic data encoders i.e., they can learn very complex mappings of inputs to outputs from huge amounts of data[5]. The 16 layers deep VGG-16 and the 22 layers deep GoogLeNet models which were introduced in the year 2014, have excelled the projected human error in the image classification task on the ImageNet dataset.

In 2015, A year later, Microsoft Research of Asia has introduced an alternative architectural approach to traditional (plain) convolutional networks, called residual networks (shortened to ResNets), achieving state-of-the-art accuracy on ImageNet classification. This architecture had allowed the team to significantly increase the depth of their networks, the best performing model consisted of 152 layers.

3. Objective

While previous colorization techniques focused on converting grayscale to colored ones, our paper aims to project a methodology to convert a continuous series of grayscale images(frames) of a video and then assemble them together to form a colored video. To make the process faster and easier a separate and huge dataset is allocated through which the learning efficiency of CNNs also increase.

4. Related Work

Our project was inspired by Ryan Dahl's CNN based system which automatically colorizing images. His system relies on several layers of ImageNet-trained from VGG16, which will be integrating with a system which is an auto encoder which has residual connections which helps in merging intermediate outputs that are generated by the encoding part of the network comprising the VGG16 layers, with those generated by the latter decoding part of the network[3].The inspiration behind the residual connections are the ResNet system built by He et al that won the 2015 ImageNet challenge. As the connections are used to link upstream network edges with downstream network edges, they allow more swift doorgifte of gradients with the help of the system, so that training convergence time gets reduced and it also enables more reliably in training more deeper networks. Veritably, Dahl reports on a much larger scale decreases training loss on each training iteration with the help of his most recent system when compared to an earlier variant that could not utilize residual connections.As per the results, Dahl's system carries out extremely well to make it realistic[8]. We nonetheless discern that in many cases, the images which are produced by the system are mainly sepia-toned and muted in color. Image colorization is formulated as a problem of regression by Dahl but the training goal to be minimized is a sum of Euclidean distances between each pixel's that are blurred color channel values in the target image as well as predicted image. While regression does seem to be properly suited to the task as it shows continuous nature of color spaces, practically, an approach based on classification may work better. To understand why, consider a pixel which exists in a flower petal across multiple images that are identical, save for the color of the flower petals. The taken pixel can take various tones or colors of red, yellow, blue, and many more. With a system which is regression-based that uses an `2 loss function, the value of the predicted pixel minimizes the loss for this particular pixel is the mean pixel value. Accordingly, the predicted pixel ends up being an unattractive, subdued mixture of the possible colors or tones. Taking this scenario into consideration, we hypothesize that a system which is regression-based would tend to generate images which are desaturated and impure in color tonality, especially for the objects that which take on many colors in the real world and is the reason behind lack of

punchiness in color in the images colorized by Dahl's system[10].

5. Methodology

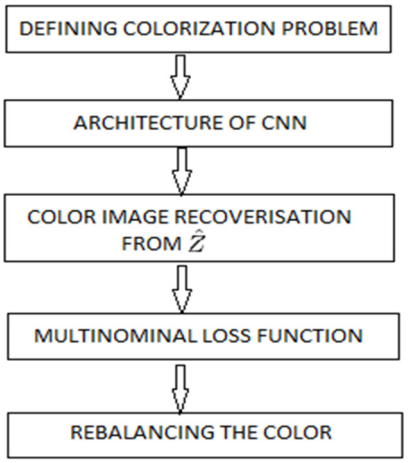

Fig. 1 Proposed flow of the methodology

5.1. The colorization problem

The colorization problem can be defined in two ways, RGB color space and CIE color space.

5.1.1 In RGB color space black and white images can be entitled as grids of pixels and the value each pixel ranges between 0 and 255 where 0 indicates black and 255 indicate white. The values of pixel correlates to its brightness.

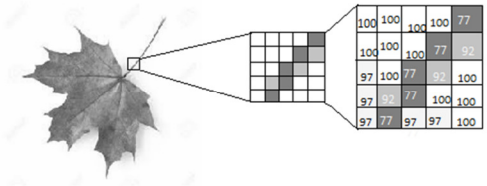

Fig. 2 Representing pixel values of a part of a black and white image

Ared layer, a green layer, and a blue layer are the three layers that a RGB color images generally consist. For example, consider a green leaf that is split into three channels on a white background[2]. One may think that the leaf is present only in the green channel but as shown in the below diagram, the leaf is present in all the 3 channels of a RGB color space, which determines the brightness along with the color.

Fig. 3 Presence of an image in all three channels of a RGB color space

We need an equal distribution of all colors to achieve the white color. A brighter color of green can be achieved by adding an equal amount of red and blue. Thus, using the three layers, color and contrast of a colored image can be encoded as shown in the below figure.

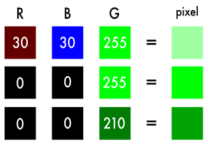

Fig. 4 Different combinations of RGB values and their results

We need a neural network which establishes a link between an input value and output value i.e., that links grayscale images with colored ones. In conclusion, we need a CNN with features that links a grid of grayscale values to the three color grids[2].

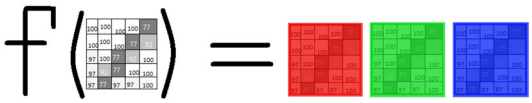

Fig. 5 Image showing functional relation between Black and white and it's respective colored image

5.1.2 Let's now look at the CIE Lab color space of colorization problem. The similarity between RGB and CIE color space is that they both are 3-channel color spaces, but unlike the RGB color space, in CIE color space only 2 channels namely a(red-green channel) and b(blue-yellow channel) store the encoded color information and the L(lightness) channel stores the information about intensity encoding.

As shown in the below figure, the entire ab space is quantized into 313 bins. This quantization makes the calculations easier i.e., we will simply find a bin number between 0 and 312 instead of finding the a and b values for every pixel[6]. The black and white picture(or frame of a video) already has a L value which ranges between 0 and 255 and the value of ab channel ranging between 0 and 313 needs to be known. Now the color prediction task can be viewed as selecting a bin from 313 classes for every gray pixel which already has its L value thus turning it into

a monic polynomial classification problem.

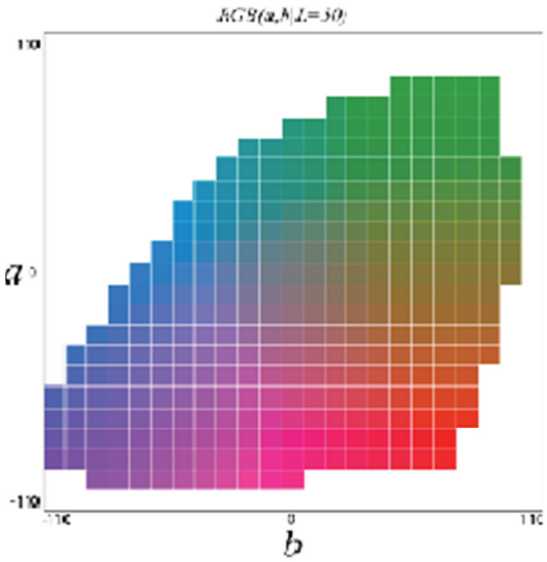

Fig. 6 Quantized colors in ab space

5.2 CNN Architecture for Colorization

The architecture of CNN is a VGG-style network with multiple convolutional blocks which is proposed by Zhang et al. Each block generally consists of 3 parts of layers, the first part has two or three internal convolutional layers, second layer consists of Rectified Linear Unit (ReLU) and the third layer is a Batch Normalization layer. This architecture has no pooling layers[11].

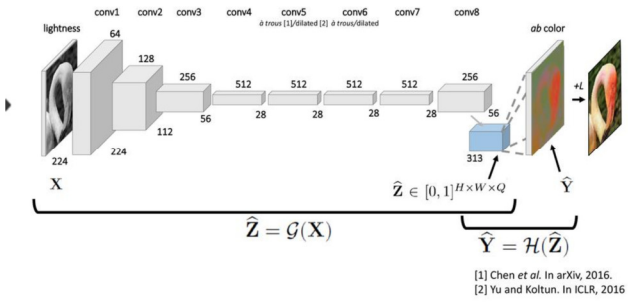

Fig. 7 CNN architecture for Colorization.

Let 'X' be the input image. X should be a image that is rescaled to 224x224. X is made to pass through the above neural network and gets transformed into 'Z^'. This transformation can be represented with 'G' and can mathematically be written as

$$Z^{\wedge} = G(X) \quad (1)$$

HxWxQ is the dimensions of Z^, where H(=56) and W(=56) represents the height and width of the output that is produced in the last convolution layer. Z contains a vector of Q(=313) value for each of the HxW pixels. Q is a value which entitles the probability of the pixel being part of that class. The goal of our paper is to find a single pair of ab channel values for each probability distribution $Z^{\wedge}_{h,w}$.

5.3 Color image recoverisation from Z^

Group of distributions in Z^ from the resized input image X is shown in the above shown figure. Now we need to recover a single ab value pair from each distribution in Z^.

We might simply take the mean of the distribution and choose the ab pair corresponding to the nearest quantized bin center. The obtained distribution is not Gaussian, i.e., the mean of the distribution corresponds to an unnatural desaturated color. Consider the color of the sky as an example, which is sometimes blue and sometimes orange-yellow. The distribution of colors of the sky is bimodal. Either blue or yellow will result in a plausible coloring, while coloring the sky but the average of blue and yellow results gray.

We may use the mode of the distribution to get either blue or yellow sky to get vibrant colors, but it sometimes breaks the spatial consistency. The solution is to interpolate between the mean and mode estimates to obtain a quantity called the *annealed-mean*. Temperature(T) was used as a parameter to control the degree of interpolation. A final value of T=0.38 is used as a trade-off between the two extremes.

The ab pair corresponding to the annealed-mean of the distribution $Z^{\wedge}_{h,w}$ is represented in $Y_{h,w}$, which can be written as a transformation of the original distribution $Z^{\wedge}_{h,w}$

$$Y = H(Z^{\wedge}) \quad (2)$$

As the image passing through the CNN, is resized to 56×56, the predicted ab image, Y, also has the dimension 56×56. It is upsampled to the original image size to obtain the colour image and then added to the lightness channel, L, to produce the final color image.

5.4 Multinomial Loss Function with Color Rebalancing

Loss function is used to train the Neural Networks and is used to minimize the loss over the training set. The output of the CNN is Z^ given an input image X . We need to transform all color images in the training set to their corresponding Z values. Mathematically, we simply want to invert the mappings H

$$Z = H^{-1}(Y) \quad (3)$$

In an output image **Y**, for every pixel **Y** we can simply find the nearest abbin and represent $\mathbf{Z}_{h,w}$ as a one-hot vector, in which we assign 0 to all the far 312 bins and 1 to the nearest ab bin. But for a better result, the 5-nearest neighbors are considered and a Gaussian distribution is used to compute the distribution $\mathbf{Z}_{h,w}$ depending on the distance from the ground truth. We may be use the standard cross-entropy loss to compare the ground truth **Z** and the estimate **Z^** using

$$L(\hat{Z}, Z) = -\frac{1}{HW}\sum_{h,w}\sum_{q} Z_{h,w,q}\log(\hat{Z}_{h,w,q}) \quad (4)$$

The dataset that we use(ImageNet) posses the color distribution that has heavy color around gray line. As a result very dull colors are produced when the above loss function is used[13].

5.5 Rebalancing the color

The loss function is changed, as follows, to produce vibrant colors

$$L(\hat{Z}, Z) = -\frac{1}{HW}\sum_{h,w} v(Z_{h,w})\sum_{q} Z_{h,w,q}\log(\hat{Z}_{h,w,q}) \quad (5)$$

We may need to rebalance the loss occured due to the color class rarity to make the output vibrant with more colors. Such rebalancing term used here is **V(.)**.

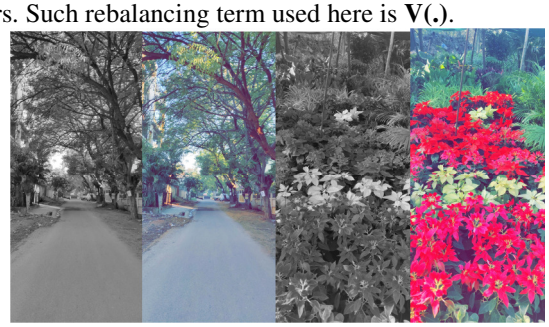

Fig 8. Results of colorization

6. Conclusion and Future Work

6.1 Conclusion

We have presented a method of fully automatic colorization of unique greyscale images combining state-of-the-art CNN techniques[5]. Using color representation and the right loss function, we have represented that the method is capable of producing a plausible and vibrant colorization of certain parts of images that has properties which may be applied to video sequences also. Our model does very well with the animals like cats and dogs because the dataset we chose i.e., ImageNet consists large amount of pictures of these animals[12]. Even the outdoor scenes turnout very good with our model. The model also captures notion of sunset and paints it orange. The model produces plausible images even with the sketches. Finally, the model also works well with the black and white videos in most of the cases.

6.2 Future Work

A betterment in the video colorization could still be made with the large datasets that are not available now. The model works very well with the image colorization and video colorization in most of the cases. Even though, the video colorization can still be brushed.

References

[1] Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification." *ACM Transactions on Graphics (TOG)* 35.4 (2016): 110.

[2] Cheng, Zezhou, Qingxiong Yang, and Bin Sheng. "Deep colorization." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[3] Larsson, Gustav, Michael Maire, and Gregory Shakhnarovich. "Learning representations for automatic colorization." *European Conference on Computer Vision*. Springer, Cham, 2016.

[4] Li, Bo, et al. "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[5] Zhang, Richard, et al. "Real-time user-guided image colorization with learned deep priors." *arXiv preprint arXiv:1705.02999* (2017).

[6] Deshpande, Aditya, Jason Rock, and David Forsyth. "Learning large-scale automatic image colorization." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[7] Limmer, Matthias, and Hendrik PA Lensch. "Infrared colorization using deep convolutional neural networks." *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 2016.

[8] Deshpande, Aditya, Jason Rock, and David Forsyth. "Learning large-scale automatic image colorization." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super resolution. In ECCV, pages 184–199. Springer, 2014.

[10] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep learning. arXiv preprint arXiv:1412.7725, 2014.

[11] Charpiat, Guillaume, Matthias Hofmann, and Bernhard Schölkopf. "Automatic image colorization via multimodal predictions." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2008.

[12] Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." *European*

Conference on Computer Vision. Springer, Cham, 2016.

[13] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436.

[14] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[15] Marr, David. "Artificial intelligence—a personal view." *Artificial Intelligence* 9.1 (1977): 37-48.

[16] Nasrabadi, Nasser M. "Pattern recognition and machine learning." *Journal of electronic imaging* 16.4 (2007): 049901.

First Author Sindhuja Kotala is currently pursuing B.E III-Year, CSE at Stanley college of Engineering and Technologies affiliated to Osmania University. Her area of interests are Artificial Intelligence, Machine Learning , Deep Learning, Predictive Analysis.

Second Author Srividya Tirumalasetti is currently pursuing B.E III-Year, CSE at Stanley college of Engineering and Technologies affiliated to Osmania University. Her area of interests are Artificial Intelligence, Machine Learning.

Third Author Vudaru Nemitha is currently pursuing B.E III-Year, CSE at Stanley college of Engineering and Technologies affiliated to Osmania University.