THE PREDICTION OF THE ACCURACY PERCENTAGE OF IMAGE CAPTION GENERATOR USING CNN TO HAVE ENHANCED ACCURACY (94%) WHEN COMPARED TO THE LSTM (78%)

Sai Teja. N.R[1], Geetha. R[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To perform an automated image caption generator using a convolutional neural network compared with Long Short-Term Memory.
Material and Methods: Automated Image caption generator performed using convolutional neural network (N=10) and long short term memory (N=10) with the split size of training and testing dataset 70% and 30% using G-power setting parameters:(α=0.05 and power=0.85) respectively**.**
Results: (CNN) convolutional neural network (94%) as the better accuracy compared to long short term memory accuracy(78%) and attained the significance value 0.651 (Two-tailed, p>0.05).
Conclusion: convolutional neural network achieved significantly better classification than Long Short Term Memory for generating a description of the image.

Keywords: Deep Learning, Image Caption Generator, Convolutional Neural Network, Long Short Term Memory, Novel Caption Generation, Encoder-Decoder, Classifier.

[1]Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode - 602105.
[2*]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode - 602105.

1. Introduction

Image caption generator is a crucial challenge in AI that bridges the gap between computer vision and natural language processing (Loganathan et al. 2020). It's a difficult undertaking to automatically describe the content of an image using properly constructed English sentences (Sharma et al., n.d.). Many sophisticated models for extracting visual information from photos based on the visual categorization and encoder-decoder of objects in the images have been created (Dehaqi, Seydi, and Madadi 2021). In most cases, the visual recognition processes followed are difficult in terms of processing complexity and achieving the requisite accuracy (Han and Choi 2020). Novel caption generation has a variety of applications, including, virtual assistants, encoder-decoder, image indexing, accessibility for visually impaired people, social media and a variety of other natural language processing applications (Kesavan, Muley, and Kolhekar 2019).

There were many distinct performances of LSTM and simple CNN. Around 108 related papers were found in IEEE Xplore and 185 were found in the Science Direct database. Many Python libraries were utilized in the development, including Keras, which included a VCG net for object recognition, and TensorFlow (Peng et al. 2019) which was created by Google and is used to build deep learning neural networks by performing CNN and LSTM algorithms. (Julakanti 2021) tested different encoder-decoder models to evaluate how each one affects caption prediction, as well as to demonstrate various use cases on our system (Iwamura et al. 2021). For image caption generators, provide a unique parallel-fusion CNN LSTM architecture. The proposed method achieves a significant improvement in performance and efficiency. present a novel caption generation survey (Banda 2021).

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy et al. 2022).Categorizing image captioning approaches into distinct groups based on the strategy used in each method. (Verma et al. 2021) was very much helpful in understanding how to perform image caption with an image dataset. Improving features extraction and efficiency of CNN classifier was discussed clearly. The Long Short Term Memory classifier to train this data showed better results in the novel caption generation (Loganathan et al. 2020). The literature review identified from the existing system has less accuracy. The flaw of this clear result is that it necessitates the presence of a large-scale corpus, which is not available for many languages. The aim of this study is to improve the accuracy of classification by incorporating CNN and comparing its performance with LSTM by encoder-decoder models. The proposed model improves classifiers to distinguish objects more efficiently with the help of image caption using deep learning techniques.

2. Materials and Methods

The study setting of the proposed work was conducted in the DBMS Laboratory, Saveetha School of Engineering in guidance with faculty. To perform this research two groups were taken. Group 1 is the Convolutional Neural Network and group 2 is the Long Short Term Memory shown in Table 1. The Sample size was calculated using clinical analysis by keeping G power fixed with 80%, 740 sample sizes estimated per group, totally 1098, 94% confidence, pretest power 80%, and enrolment ratio 1 and the maximum accepted error is fixed as 0.05, the accuracy of two classifiers Convolutional Neural Network and Long short term memory was compared. Independent variables are image, vocabulary, preprocessed words, description length, and variables in images. Dependent variables are images and objects are Independent ((Loganathan et al. 2020)).

The two groups that used CNN and LSTM algorithms were performed by taking 8000 images and 5 different captions for each image as a dataset. A collection of images was used in the encoder-decoder model, with about 549 images with descriptions of novel captions generated. These captions were extracted using convolutional neural networks and preprocessing was performed. The first group in this paper is the CNN algorithm which performs classification by forming groups of every different class in the data. CNN classifier takes k groups as input size and tries to do classification with the k groups. Significance value p=0.651. The proposed work is designed and implemented with the help of google colab platform. The platform to assess deep learning was Windows 10 OS. The Hardware configuration was an Intel corei7 processor with a RAM size of 8GB. The system sort used was 64-bit. For the implementation of code, the python programming language was used. As for code execution, the

dataset is worked behind to perform an output process for accuracy.

Convolutional Neural Network

Convolutional Neural Network (CNN) is a Deep Learning method that takes an input image and assigns relevance (learnable weights and biases) to various aspects/objects in the image, allowing it to distinguish between them. Image categorization is one of the most often used uses of this architecture. Several convolutional layers, as well as nonlinear and pooling layers, make up the neural network. Depicts a high-level picture of our model. We employed a discriminative model that optimizes the likelihood of the right description given the image, following the method. Our model is formulated as in equation 1.

$$\theta^{*}= \arg\max \theta \sum_{(I,S)} \sum_{t=0}^{N} \log p(St \mid I, S0, ...,S_{t-1}; \theta) \quad (1)$$

The first summation is made up of pairs of images I and their proper transcriptions S. The sum for the second summation is over all of the words S_t in S, where N is the length of S. Its worth noting that the second summation shows the sentence's probability in relation to the combined probability of its words.

Pseudocode for Convolutional Neural Network

INPUT: Training Dataset
OUTPUT: Classifier accuracy

Step 1: Algorithm Parallel-CNN
Step 2: Input: d: dataset,
1: dataset true labels,
W: Word2Vce matrix
Step3: Output: score of Parallel-CNN trained model on
Test dataset
Step 4: Let f be the features 2d matrix
Step 5: for I in dataset do
Step 6: let
f_i Be the feature set matrix of sample i
Step 7: for j in I do
Step 8: $V_j \leftarrow vectorize_{(j,w)}$
Step 9: append V_j to f_i
Step 10: append f_i to f
Step 11: $f_{train}, f_{test}, I_{train}, I_{test}$ ←split feature set and labels
Step 12: M←Parallel-CNN (f_{train} , I_{train})
Step 13: Score←evaluate (i, I_{test}, M)
Step 14: Return score

Long Short-Term Memory

is a type of RNN that can deal with vanishing and exploding gradients as well as extended dependencies. A memory cell and different gates govern the input, output, and memory behaviors in an LSTM. With input gate, input modulation gate $a^{(t)}$ output gate $Ux^{(t)}$, and forgetting gate$Wh^{(t-1)}$we use an LSTM. is the number of hidden units. The LSTM may carry out relevant information throughout the processing of inputs, and it can discard non-related information using a forget gate equation 2.

$$a^{(t)} = Wh^{(t-1)} + Ux^{(t)} \quad (2)$$

Pseudocode for Long Short-Term Memory

INPUT: Training Dataset
OUTPUT: Classifier accuracy

Step 1: Generate five descriptions for each image.
Step 2: Get the data values and extract them.
Step 3: Find the dependent and independent attributes and divide them.
Step 4: Adjust the attributes so that there will be a loss function between them.
Step 5: Finally make the regularization of the penalties for the loss function calculated.
Step 6: Return the predicted class.
Step 7: End the program.

Statistical Analysis

The statistical analysis is done using IBM SPSS statistical analysis tool with version 26. Independent Sample T-test analysis was performed by using the Machine learning models and evaluated the quality of the study. In the Statistical package for the social sciences, SPSS version 26 software tool was used for statistical analysis. The dataset is prepared using the 10 samples from each of the algorithms and the total samples is 20. Group id is given 1 for CNN Classifier and 2 for LSTM. An independent sample T-test was conducted for accuracy. Standard deviation, standard mean errors were also calculated using the SPSS software tool. The significance values of proposed and existing algorithms contain group statistical values of proposed and existing algorithms

3. Results

The proposed CNN algorithm and LSTM were run at different times in Anaconda Navigator with a sample size of 10 using. Table 2 represents the predicted accuracy of image caption and recognition of novel caption generation using encoder-decoder models. These 10 data samples are used for each algorithm along with their loss values to calculate statistical values that can be used for comparison. From the results, it is observed that the mean accuracy of the CNN algorithm was 94% and LSTM was 78%. Table 3 represents mean accuracy values for CNN and

LSTM. The Mean value of CNN is better when compared with the LSTM with a standard deviation of 2.57388 and 3.27763 respectively. Table 4 shows the Independent sample T-test data of CNN and LSTM with the significance value obtained is 0.651 (Two-tailed, $p>0.05$).

Figure 1 denotes the comparison of CNN and LSTM in terms of mean accuracy and loss. The group statistics value along with mean, standard deviation, and standard error mean for the two algorithms are also specified in deep learning. The graphical representation of comparative analysis, means loss between two algorithms of CNN and LSTM are classified. This indicates that Convolutional Neural Networks are significantly better with 94% accuracy when compared with Long Short Term Memory classified accuracy of 78%. The Standard Deviation Error Bars are +/- 1 SD as given in Fig. 1.

4. Discussion

In the given study, the significance value obtained is 0.615 because, of a large number of datasets with fewer parameters (Two-tailed, $p>0.05$). which implies that CNN appears to be better than LSTM using the encoder-decoder model. The accuracy analysis of the CNN classifier is analyzed as 94% whereas the accuracy of the LSTM classifier is 78%. A Comparative previous assessment of CNN over LSTM is depicted in this paper (Kesavan, Muley, and Kolhekar 2019). This clearly indicates that CNN appears to be a better classifier when compared to the LSTM classifier (Bai and An 2018). This work shows a comparative accuracy analysis between CNN and LSTM in Which CNN shows an accuracy of 94% and LSTM shows an accuracy of 78%.

In deep learning, CNN is said to be a type of artificial neural network that generates captions of the given novel image using the previously saved datasets (Yang, Zhang, and Cai 2020). The connection between the two hidden layers is done by CNN. The output layer can get data from the past and future states simultaneously. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information (Yang, Zhang, and Cai 2020). Similar reach examines cutting-edge algorithms for object detection and recognition, attribute prediction, and image caption production. To create picture captioning, an encoder-decoder architecture with a visual attention mechanism is used. The encoder is based on CNN, while the decoder is based on the visual attention module. The opposite finding of the reach transforms the problem into a retrieval task. It also creates a database based on picture and text features. These applications include recommendations in editing applications, Novel Caption generation usage in virtual assistants, encoder-decoder, image indexing, visually impaired persons, for social media, and several other natural language processing applications. It is helpful in generating a caption of the image (Sharma et al., n.d.) (Iwamura et al. 2021).

The limitations of this study are that it takes a very long time to train a convolutional neural network, especially with large datasets in deep learning (Sahasrabuddhe, n.d.). The more independent and dependent variables there are, the better the accuracy of using the Encoder decoder model. The Future scope of this study is that the system should be expanded to include a larger number of images with lesser time consumption in training the data set. As a result of characteristics like these, accuracy and exact precision numbers can be increased (Lee, Eum, and Kwon 2020).

5. Conclusion

In this research work, the prediction of the accuracy percentage of image caption generator using CNN to have enhanced accuracy (94%) When compared to the LSTM (78%). Accuracy estimation for various image caption generators has been successfully calculated for the Images. The main focus was on the algorithmic substance of various attention processes, as well as a summary of how they are used. Conclude that we have succeeded in creating an encoder-decoder model that is a major improvement above all other image caption generators previously available. Accurate descriptions of accurate calculations for each Image can be done using this model.

Declarations

Conflicts of Interests

No conflict of interest in this manuscript.

Authors Contribution

Author ST was involved in data collection, data analysis, and manuscript writing. Author RG was involved in conceptualization, data validation, and critical reviews of manuscripts.

Acknowledgment

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical

Sciences (formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Infysec Solution, Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102142.

Bai, Shuang, and Shan An. 2018. "A Survey on Automatic Image Caption Generation." Neurocomputing. https://doi.org/10.1016/j.neucom.2018.05.080.

Banda, Anish. 2021. "Image Captioning Using CNN and LSTM." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.37846.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." Bioresource Technology 351 (May): 126956.

Dehaqi, Ali Mollaahmadi, Vahid Seydi, and Yeganeh Madadi. 2021. "Adversarial Image Caption Generator Network." SN Computer Science. https://doi.org/10.1007/s42979-021-00486-y.

Han, Seung-Ho, and Ho-Jin Choi. 2020. "Domain-Specific Image Caption Generator with Semantic Ontology." 2020 IEEE International Conference on Big Data and Smart Computing (BigComp). https://doi.org/10.1109/bigcomp48618.2020.00-12.

Iwamura, Kiyohiko, Jun Younes Louhi Kasahara, Alessandro Moro, Atsushi Yamashita, and Hajime Asama. 2021. "Image Captioning Using Motion-CNN with Object Detection." Sensors 21 (4). https://doi.org/10.3390/s21041270.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." Environmental Research. https://doi.org/10.1016/j.envres.2022.113095.

Julakanti, Vaibhav. 2021. "Image Caption Generator Using CNN-LSTM Deep Neural Network." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.35663.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." Materials Chemistry and Physics. https://doi.org/10.1016/j.matchemphys.2022.125900.

Kesavan, Varsha, Vaidehi Muley, and Megha Kolhekar. 2019. "Deep Learning Based Automatic Image Caption Generation." 2019 Global Conference for Advancement in Technology (GCAT). https://doi.org/10.1109/gcat47503.2019.8978293.

Lee, Hyungtae, Sungmin Eum, and Heesung Kwon. 2020. "ME R-CNN: Multi-Expert R-CNN for Object Detection." IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2019.2938879.

Loganathan, K., R. Sarath Kumar, V. Nagaraj, and Tegil J. John. 2020. "CNN & LSTM Using Python for Automatic Image Captioning." Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.10.624.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." Journal of Energy Storage. https://doi.org/10.1016/j.est.2022.104422.

Peng, Yuqing, Xuan Liu, Weihua Wang, Xiaosong Zhao, and Ming Wei. 2019. "Image Caption Model of Double LSTM with Scene Factors." Image and Vision Computing. https://doi.org/10.1016/j.imavis.2019.03.003.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102102.

Sahasrabuddhe, Yash. n.d. "Fake Malware Classification with CNN via Image Conversion." https://doi.org/10.31979/etd.q8vd-npfh.

Sharma, Grishma, Priyanka Kalena, Nishi Malde, Aromal Nair, and Saurabh Parkar. n.d. "Visual Image Caption Generator Using Deep Learning." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3368837.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022. "Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." Computational Intelligence and Neuroscience 2022 (February): 5906797.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" Fuel. https://doi.org/10.1016/j.fuel.2022.123494.

Verma, Akash, Harshit Saxena, Mugdha Jaiswal, and Poonam Tanwar. 2021. "Intelligence Embedded Image Caption Generator Using LSTM Based RNN Model." 2021 6th International Conference on Communication and Electronics Systems (ICCES). https://doi.org/10.1109/icces51350.2021.9489253.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." International Journal of Biological Macromolecules 209 (Pt A): 951–62.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Yang, Xu, Hanwang Zhang, and Jianfei Cai. 2020. "Auto-Encoding and Distilling Scene Graphs for Image Captioning." IEEE Transactions on Pattern Analysis and Machine Intelligence PP (December). https://doi.org/10.1109/TPAMI.2020.3042192.

TABLES AND FIGURES

Table 1. Group, Accuracy, and Loss value uses 8 columns with 8 width data for image caption generator.

SI.NO	Name	Type	Width	Decimal	Columns	Measure	Role
1	Group	Numeric	8	2	8	Nominal	Input
2	Accuracy	Numeric	8	2	8	Scale	Input

3	Loss	Numeric	8	2	8	Scale	Input

Table 2. Accuracy and Loss Analysis of Convolution neural network and Long short term memory.

S.No	**GROUPS**	**ACCURACY**	**LOSS**
1	CNN	94.89	5.11
		94.42	5.58
		91.33	8.67
		93.00	7.00
		93.94	6.06
		93.42	6.58
		89.85	10.15
		93.21	6.79
		89.12	10.88
		87.12	12.88
2	LSTM	78.74	21.26
		78.12	21.88
		77.12	22.88
		75.54	24.46
		74.16	25.84
		70.00	30.00

		68.85	31.15
		74.67	25.33
		76.35	23.65
		76.65	23.35

Table 3. Group Statistical Analysis of CNN and LSTM. Mean, Standard Deviation and Standard Error Mean are obtained for 10 samples. CNN has higher mean accuracy and lower mean loss when compared to LSTM.

Name	**GROUP**	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
ACCURACY	CNN	10	92.0300	2.57388	.81393
	LSTM	10	75.0200	3.27763	1.03648
LOSS	CNN	10	7.9700	2.57388	.81393
	LSTM	10	24.9800	3.27763	1.03648

Table 4. Independent Sample T-test: CNN is insignificantly better than LSTM with p value 0.651 (Two tailed, p>0.05)

Name	**Variances**	**F**	**Sig.**	**t**	**df**	**Sig (2-tailed)**	**Mean Diffencene**	**Std.Error differenc e**	**Lower**	**Upper**
ACCURACY	Equal variances assumed	.212	.651	12.907	18	.000	17.01000	1.31787	14.24126	19.77874
	Equal Variances not assumed	.212	.651	12.907	17.042	.000	17.01000	1.31787	14.23006	19.78994
LOSS	Equal variances assumed	.212	.651	-12.907	18	.000	-17.01000	1.31787	-19.77874	-14.24126
	Equal Variances	-	-	-12.90	17.042	.000	-17.01000	1.31787	-19.7899	-14.2300

	not assumed			7					4	6

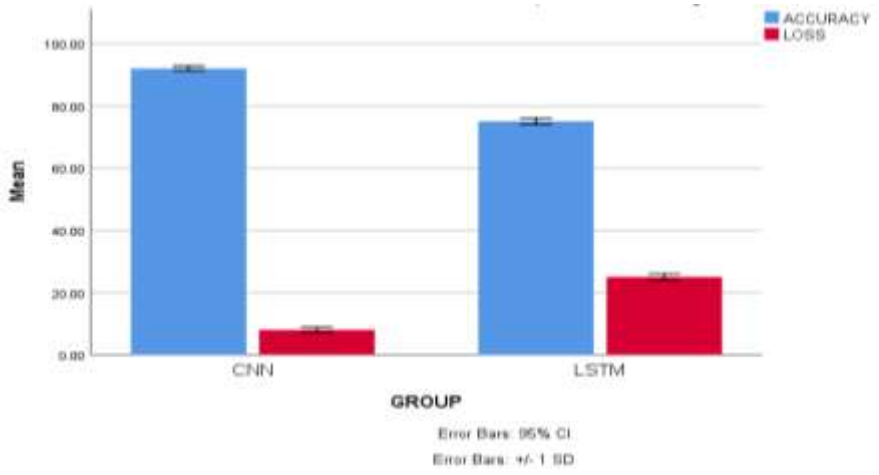

Fig. 1. Comparison of CNN and LSTM Classifier in terms of mean accuracy and loss. The mean accuracy of CNN is better than LSTM Classifier; the Standard deviation of CNN is slightly better than LSTM. X Axis: CNN Vs LSTM Classifier and Y-Axis: Mean accuracy of detection ± 1 SD

GROUPS	ACCURACY	LOSS
YOLO	80	0.20
KNN	91	0.09
CNN	90	0.10
RESNET	89	0.11
SVM	88	0.12

GROUP	N	Mean	Std.Devianation	Std.Error Mean
YOLO	10	80.0300	2.57388	0.81393
KNN	10	90.0200	3.27763	1.03648
CNN	10	91.9700	2.57388	0.81393
RESNET	10	89.9800	3.27763	1.03648
SVM	10	16.8540	1.2451	0.1547