OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12679858

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Envision: An Image Recognition App

Pushpa G*[1], Vishal Anil Bajpai[2], Naveen Krishna Bhat S[3], Shashank Devashetty[4], Mudit Kushwaha[5]

[1]Assistant Professor, Department of ISE, RV Institute of Technology and Management, Karnataka, India.
[2,3,4,5] Student, Department of ISE, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Pushpa, G., Bajpai, V. A., Bhat, N. K. S., Devashetty, S., Kushwaha, M. (2024). Envision: An Image Recognition App. *Indiana Journal of Multidisciplinary Research*, *4*(3), 205-211.

Abstract: **Objectives:** • Increasing Caption Quality: A major goal is to raise the standard and fluency of automatically generated captions. In order to do this, it may be necessary to investigate cutting-edge language modelling approaches, take context and coherence into account, and reduce faults like grammatical or descriptive errors. • Handling Ambiguity and Diversity: Addressing the difficulty of coming up with diverse and contextually relevant descriptions for pictures that could be interpreted in a variety of ways. Methods for capturing and displaying the many viewpoints, deciphering visual clues, and creating captions that accurately convey the intended meaning can all be the subject of research.
Findings: The papers highlight the importance of attention processes in image captioning by demonstrating how paying attention to particular areas of an image enhances the relevancy and calibre of the captions that are created. Datasets and evaluation metrics are covered in the reviews, which also offer insights into the advantages and disadvantages of popular evaluation metrics like BLEU, METEOR, CIDEr, and ROUGE. They also list well-known datasets that are frequently used for developing and testing image captioning algorithms, such as MS COCO, Flickr30K, and Visual Genome.
Keywords: Image captioning, Deep Learning, Accessibility, Visual Impairments

INTRODUCTION

An essential component of human cognition is the capacity to comprehend and describe the content of visuals. Humans naturally create evocative captions for photographs that include context, emotional cues, and visual features [1]. However, in the disciplines of computer vision and natural language processing, it.

The motivation behind image caption generation stems from the potential benefits and applications it holds. Here are a few key reasons why we have been driven to explore and advance this field:

- Enhanced Accessibility: Image captioning technology can help those with visual impairments have easier access to visual content. It enables visually challenged people to encounter and understand visual content that would otherwise be inaccessible to them by giving textual descriptions of the pictures.
- Image captioning can help in the effective interpretation and organization of content in an age of massive volumes of visual data. It is made simpler to search, retrieve, and organize photos based on their semantic information by automatically producing informative captions. Digital archiving, content management, and e-commerce can all profit from this.
- Deep learning advancements: The production of image captions serves as a difficult benchmark for investigating and pushing the limits of deep learning methods. Models are needed to comprehend intricate visual scenes, record minute details, recognise contextual connections, and produce coherent and insightful linguistic descriptions. The field of deep learning and its applications can benefit from improvements in image captioning.

BACKGROUND AND EXISTING SOLUTION

Related Work

- Real-Time Objects Recognition Approach for Assisting Blind People: This paper presents a novel approach utilizing two cameras integrated into glasses worn by blind individuals, along with GPS and ultrasonic sensors, to provide real-time information about their surroundings. Object detection algorithms, including the Speeded-Up Robust Features (SURF) method, are optimized for recognition of common objects such as faces, bicycles, and furniture. GPS coordinates are leveraged to categorize detected objects based on their locations, while the ultrasonic sensor detects obstacles at medium to long distances. However, the complexity of incorporating two cameras into glasses may pose challenges [2].

- Wearable Object Detection System for the Blind: This system introduces a wearable RFID device designed to assist blind individuals in identifying objects, particularly medications stored in cabinets. The device emits acoustic signals to indicate the proximity of desired objects and provides distance measurements using Received Strength Signal

***Corresponding Author:** Pushpa G

Indicator (RSSI) values. By measuring the movement of the antenna relative to the RFID tag, the device accurately assesses the distance to objects, streamlining the search process [3].

- Smart Obstacle Detector for Blind Person: This system focuses on detecting and classifying obstacles in front of blind individuals using a combination of MATLAB signal processing and video recording. The output is provided audibly and through vibration feedback, with a vibrating motor connected to an ultrasonic sensor. While effective, the system's reliance on bulky equipment, including a camcorder and sensor-equipped stick, limits its practicality and portability [4].

Applications

Image Recognition and computer vision: CNNs are widely used for popular types of images including object recognition, image segmentation, and image segmentation. Applications include famous faces, autonomous vehicles, and clinical image analysis. Natural Language Processing (NLP): CNN and RNN [5] are used for NLP responsibilities along with language translation, sentiment analysis and text age. Applications include chat bots, language translation services and speech recognition systems. Health care: Deep learning [6] is applied to healthcare for scientific image analysis, disease prognosis, general medical advice, and drug discovery. financial: AI/ML is used in finance for fraud detection algorithmic trading, credit scoring, and chance assessment. retail [7]: AI/ML is used in retail for customized advertising, demand forecasting, stock management and customer segmentation. Vehicle Use: Deep learning is used in autonomous vehicles for object recognition, navigation, navigation planning and decision making. Construction: AI/ML is used in manufacturing for predictive maintenance, priority management, supply chain optimization, and robotics. agriculture: AI/ML is used in agriculture for crop management, yield prediction, pest detection, and agricultural precision. entertainment: AI/ML is used for leisure content recommendations, customized advertising, and content generation. security: AI/ML is used in security for facial recognition, behavioral analysis, coincidence detection and Cybersecurity.

This study [7] suggests a collaborative deep learning architecture that combines LSTM and GRU models for dynamic stock market forecasting. It also recommends further research to encompass a wider range of stock data and media sentiments, as well as optimizing the representation of financial news sentiments to enhance predictive accuracy.

METHODOLOGY

Blind People Problems

Blind individuals encounter a host of daily challenges due to the lack of visual information. From manoeuvring through physical spaces independently to accessing printed materials like books and menus, each task presents its own set of obstacles. Recognizing objects and interpreting facial expressions without sight can be particularly daunting, as can accurately reading text on screens. However, there's hope on the horizon with the advent of image recognition technology. Thanks to object recognition algorithms, blind individuals can now receive auditory cues about their surroundings, making navigation and interaction more manageable. Optical character recognition (OCR) further enhances accessibility by converting printed text into a digital format, opening doors to various materials. Detailed scene descriptions and facial recognition technology also play crucial roles in facilitating easier identification and interaction. While integrating image recognition into digital platforms holds promise, ensuring its accuracy, reliability, and user-friendliness is paramount. Additionally, addressing privacy and ethical concerns surrounding facial recognition is vital to protecting the dignity and rights of blind individuals [15].

Recommended techniques

a) Data Acquisition and Preprocessing

Data acquisition involves the systematic collection and storage of data from diverse sources, a process analogous to gathering essential ingredients for a recipe. It commences with the deployment of sensors or instruments to capture requisite data, followed by signal conditioning to refine raw signals for clarity and coherence. Subsequently, analog signals are digitized for computer processing and storage, with the recorded data typically archived in databases or files, either in real-time or through batch processing. Supplementary steps may entail transmitting data for remote monitoring or analysis. Following acquisition, rigorous data validation and quality assurance protocols are enacted to ensure precision and reliability. Finally, the acquired data undergoes thorough processing and analysis, often employing techniques such as statistical analysis or machine learning, to unearth valuable insights. Integral to scientific research, industrial monitoring, and healthcare, data acquisition facilitates informed decision-making and knowledge generation.

Preprocessing serves as the preparatory groundwork for meaningful data analysis, resembling the meticulous arrangement of ingredients before culinary endeavours commence. It encompasses the initial stages of cleaning, transforming, and organizing raw data to optimize subsequent analysis. This preparatory phase aims to rectify errors or inconsistencies, refine data for clarity and comprehension, and identify pertinent features for focused examination. Analogous to the methodical arrangement of a room before hosting guests, preprocessing ensures a structured and coherent dataset by addressing missing data elements, translating categorical information into analytically relevant formats, and simplifying complex datasets for ease of

***Corresponding Author:** Pushpa G

handling. Additionally, preprocessing entails segmenting data into manageable subsets for training, validation, and testing, ensuring the integrity and reliability of subsequent analyses. Despite its understated role, preprocessing plays a pivotal role in ensuring the accuracy, reliability, and actionable nature of insights derived from data analysis.

b) Algorithm

i) CNN-RNN

At the heart of image captioning algorithms lies the CNN-RNN architecture, which integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to craft descriptive textual captions for images [6]. This approach has attracted significant attention for its capacity to understand complex visual scenes and generate coherent verbal descriptions. Within this architecture, CNNs take on a crucial role by extracting advanced features from input images [6]. These features encapsulate the visual content of the images, providing a comprehensive representation for generating captions.

In contrast, RNNs, including variants like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), serve as decoders within the CNN-RNN structure, excelling at capturing sequential patterns and generating linguistic sequences [6]. The RNN component leverages the representation of CNN-extracted features to sequentially generate words or tokens forming the final caption.

The synergy between CNNs and RNNs offers several advantages for image captioning. CNNs excel at capturing visual features, facilitating the recognition of salient features within images to produce semantically aligned captions [6]. This capability enhances the overall coherence and relevance of the descriptions, contributing to the effectiveness of the CNN-RNN architecture in image captioning tasks.

The inherent recurrent nature of RNNs facilitates the creation of captions with temporal coherence and contextual understanding [6]. By incorporating feedback loops that consider previously generated words, RNNs iteratively refine the caption generation process, ensuring that each word harmonizes with its predecessors and contributes cohesively to the overall narrative flow.

Recent advancements in the CNN-RNN framework have notably augmented its performance and applicability in image captioning tasks [5]. For instance, Hoxha *et al.* introduced a novel CNN-RNN framework tailored specifically for remote sensing image captioning, demonstrating superior captioning accuracy and robustness. These innovations underscore the adaptability and versatility of the CNN-RNN algorithm across diverse domains and modalities.

In summary, the CNN-RNN algorithm stands as a robust framework in image captioning, leveraging the strengths of CNNs for visual feature extraction and RNNs for sequential language generation. As research progresses and refinements are made, the CNN-RNN architecture remains at the forefront of advancing image captioning performance, offering opportunities for future developments to enhance multimedia comprehension and accessibility.

Character Ist ics	CNN (Convolutional Neural Network)	RNN (Recurrent Neural Network)
Architecture	Multilayered structure comprising convolutional, pooling, and fully connected layers	Recurrent connections allow feedback loops, enabling sequential processing.
Input Type	Typically used for processing grid-like data such as images or videos	Suited for sequential data like time series, text, or speech
Memory	Lacks explicit memory, focuses on hierarchical feature extraction	Possesses internal memory, capable of retaining information over time
Data Dependency	Local dependencies within receptive fields, capturing spatial hierarchies	Temporal dependencies across sequences, retaining context over time
Application s	Image and video recognition, object detection, image classification	Speech recognition, language translation, time series prediction
Training	Often requires large datasets for effective feature learning	Slow and complex training, prone to vanishing or exploding

***Corresponding Author:** Pushpa G

Gradients	Extraction for visual tasks, scalable architecture	Understanding and sequence modeling, precise predictions

ii) Deep Learning

Deep learning, with a focus on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has emerged as a dominant method for image captioning tasks. These algorithms excel in autonomously learning hierarchical data representations, making them well-suited for tasks requiring intricate visual and linguistic understanding.[6]

In image captioning, deep learning algorithms offer distinct advantages over traditional methods. CNNs, for instance, efficiently extract hierarchical features from images, capturing both detailed visual elements and overarching semantic concepts.[6]

When integrated with RNNs, deep learning models enable end-to-end training for image captioning, optimizing the visual feature extractor (CNN) and language generator (RNN) simultaneously. This approach enhances the integration between visual and linguistic modalities, resulting in more coherent and contextually relevant.

There are two main models in CNN:

i) VGG: VGG-16, a deep convolutional neural network introduced at the ILSVRC 2014[8] Conference by the Visual Graphics Group at the University of Oxford, revolutionized image classification tasks. It surpassed the previous standard set by AlexNet and quickly gained popularity among researchers and industry professionals.

The architecture of VGG-16 includes 16 layers, consisting mainly of convolutional layers followed by max-pooling layers. The network concludes with three fully connected layers and a softmax classifier for output. Its simplicity and effectiveness made it a benchmark for image classification tasks.

Here is the architecture of VGG-16 [9]

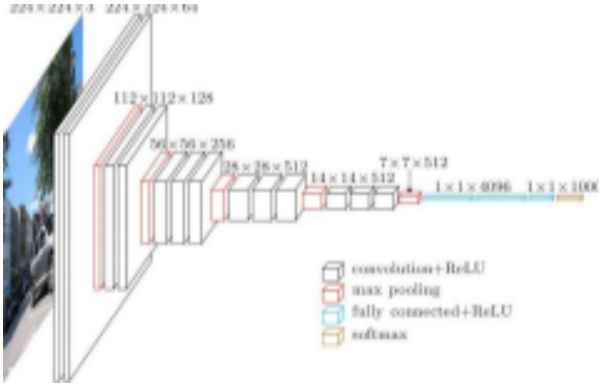

Figure 1: VGG16 Architecture

ii)Resnet: Recent advancements, such as Hoxha *et al*.'s CNN-RNN framework tailored for remote sensing image captioning, have further improved model performance and generalization capabilities. Moreover, deep learning algorithms exhibit promise in diverse image domains, from natural scenes to medical images, due to their ability to learn abstract representations from extensive datasets.[5]

In summary, deep learning, particularly CNNs and RNNs, represents a cutting-edge approach for image captioning. Their capacity to autonomously learn hierarchical representations, coupled with end-to-end training and adaptability to various domains, makes them invaluable for generating descriptive image captions.[6]

***c)* Recommended Models**

ResNet, short for Residual Network [8], addresses the vanishing gradient problem encountered in deep neural networks. As more layers are added, the gradient diminishes, leading to difficulties in training. ResNet introduces skip connections, also known as identity mappings, to alleviate this issue. These connections allow the gradient to flow more easily by skipping over certain layers, enabling the training of very deep networks with hundreds or thousands of layers.

By utilizing skip connections, ResNet enables the training of very deep neural networks, surpassing the limitations of previous architectures. It has become a cornerstone in computer vision applications, achieving state-of-the-art performance in tasks such as image classification and object detection.

ResNet provides an innovative solution to the vanishing gradient problem, known as “skip connections”. ResNet stacks multiple identity mappings (convolutional layers that do nothing at first), skips those layers, and reuses the activations of the previous layer. Skipping speeds up initial training by compressing the network into fewer layers.

One of the most Popular RNN model is:

LSTM (Long Short -Term Memory): Long short-term memory (LSTM) is a significant advancement over traditional recurrent neural networks (RNNs). It addresses the vanishing and exploding gradient problem by introducing memory blocks, which differ from the standard RNN units. Additionally, LSTM [10] includes a cell state to preserve long-term states, a key distinction from RNNs. This design enables an LSTM network to retain and connect previous information to current data.

***Corresponding Author:** Pushpa G

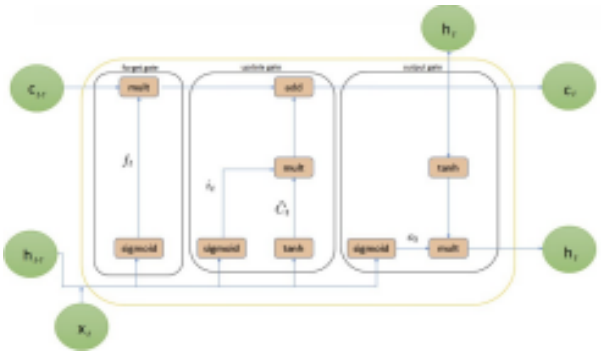

Figure 2: Long short-term memory internal structure

When selecting a model for image recognition, various models with differing accuracies are available. However, in [11], we propose a novel CNN-RNN architecture that improves image recognition accuracy. Despite its benefits, this model also comes with challenges such as increased complexity, which can lead to reduced training speed and a higher risk of overfitting. To address these challenges, we incorporate regularization techniques to control model complexity.

Our approach involves extracting features from the CNN layer using the RNN network. This innovative method combines the strengths of both CNNs and RNNs, enabling more comprehensive feature extraction and enhancing overall recognition accuracy A case study from [10] supports the findings of [11], this includes the tests on X-ray images of lungs.

There were 3 classes given for labelling of the images, these are pneumonia, Normal and COVID-19. Table 1 summarizes the overall accuracy, specificity, sensitivity, and F1-score for each case of the CNN architecture, with a visual representation shown in Fig. 3. The highest values for specificity, sensitivity, and F1-score were achieved in the normal cases, whereas lower sensitivity values were observed in the pneumonia cases.

Table 1: Performance of the CNN network [10].

Class	Accuracy (%)	Specificity (%)	Sensitivity (%)	F1-Score (%)
COVID-19	98.5	98.2	99.0	97.7
Pneumonia	98.6	99.7	96.4	97.8
Normal 99.9	99.8	100.0	99.8	99.8

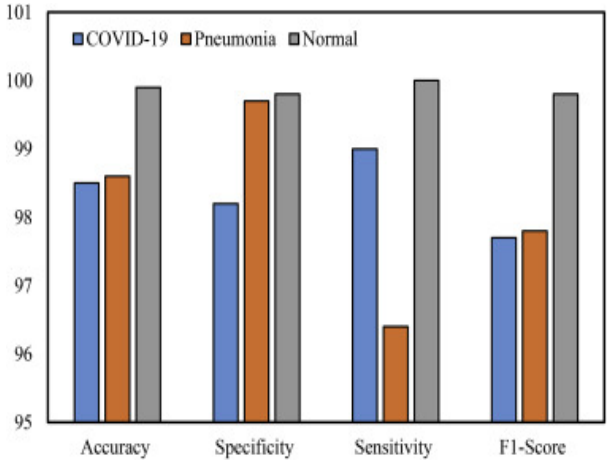

Figure 3: The graphical representation of CNN network.

Table 2 and Figure 4 present the performance metrics of each class for the developed CNN-LSTM network. The normal cases achieved the highest sensitivity and F1-score, while the pneumonia cases showed lower sensitivity values.

Table 2: Performance of the CNN-LSTM network [10]

Class	Accuracy (%)	Specificity (%)	Sensitivity (%)	F1-score (%)
COVID-19	99.2	99.2	99.3	98.9
Pneumonia	99.2	99.8	98.0	98.8
Normal	99.8	99.7	100.0	99.7

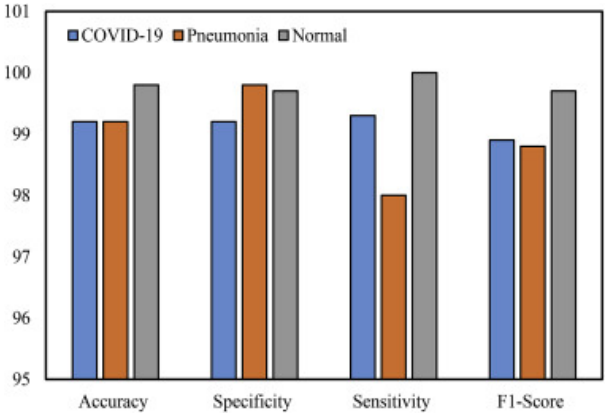

Figure 4: The graphical representation of the CNN-LSTM network.

RESULTS

The CNN-RNN architecture stands out as a top contender for cutting-edge image recognition software. Its allure lies not just in its effectiveness but also in its adaptability and ease of integration, especially within mobile applications [12]. By harnessing Convolutional Neural Networks (CNNs) for extracting visual features and Recurrent Neural Networks (RNNs) for generating sequential language, the CNN-RNN model showcases impressive capabilities in comprehending intricate visual contexts and crafting coherent textual descriptions. However, while the CNN-RNN framework displays promise, further exploration of various extraction techniques and combinations is crucial to enhance its performance and versatility across different domains.

***Corresponding Author:** Pushpa G

The insights provided by the blog [12] offer valuable guidance in selecting the most suitable model for specific use cases [17] and provide extensive resources for delving deeper into image classification concepts using TensorFlow Lite. As research advances and breakthroughs emerge, the CNN-RNN architecture evolves, heralding innovative solutions in image recognition and beyond [18][19].

Future, the CNN-RNN combination architecture can be powerful tool field of Internet of medical things like digital imaging, feature extraction and data fusion and temporal context, and even it can be extended in the field of Context aware computing [20].

CONCLUSION

While the CNN-RNN architecture stands out as a frontrunner in image captioning, further research and refinement are necessary to optimize its performance and versatility across diverse domains. Collaborative efforts, as suggested in the study, offer potential for advancing the field, with opportunities for future developments to enrich multimedia comprehension and accessibility. The study underscores the crucial role of attention processes in image captioning, highlighting their capacity to improve the relevance and quality of captions by focusing on specific image areas.

Furthermore, the study emphasizes the critical need for thorough evaluation of datasets and metrics employed in evaluating image captioning algorithms. It delves into the nuances of prominent evaluation metrics such as BLEU, METEOR, CIDEr, and ROUGE, illuminating both their strengths and limitations. In addressing the obstacles faced by visually impaired individuals, the research investigates the role of object recognition and optical character recognition technologies in aiding daily tasks. This exploration underscores the transformative potential of image recognition technology in improving accessibility, offering auditory cues and digitizing printed text to enhance the user experience. Furthermore, the study identifies the CNN-RNN architecture as pivotal in image captioning, acknowledging its effectiveness in comprehending intricate visual scenes and producing coherent textual descriptions.

ACKNOWLEDGEMENTS

We thank **Dr. Latha C A**, Head of Department of Information Science and Engineering, RV Institute of Technology and Management, Bengaluru, for her encouragement.

We would like to extend our gratitude to **Dr. Jayapal R,** Principal, RV Institute of Technology and Management, Bengaluru, for facilitating us to present the Survey Paper.

We would like to extend our gratitude to the **Management, RV Institute of Technology and Management**, Bengaluru, for providing all the facilities to present the Survey Paper.

REFERENCES

1. Karjala, D. S., & Johansen, L. (2019). Noise reduction and feature extraction in CNN for improved facial emotion recognition. *IET Image Processing*, *13*(13), 2559-2566. https://doi.org/10.1049/iet-ipr.2019.1270
2. Zraqou, J., Alkhadour, W., & Siam, M. (2017). Real-time objects recognition approach for assisting blind people. *Multimedia Tools and Applications*, *76*(1), 31-40. https://doi.org/10.1007/s11042-016-3668-9
3. Dionisi, A., Sardini, E., & Serpelloni, M. (2012). Wearable object detection system for the blind. *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 1255-1258. https://doi.org/10.1109/I2MTC.2012.6229180
4. Daniyal, D., Ahmed, F., Ahmed, H., Shaikh, E., & Shamshad, A. (2014). Smart obstacle detector for blind person. *Journal of Biomedical Engineering and Medical Imaging*, *1*(3), 31-40. https://doi.org/10.14738/jbemi.13.245
5. Hoxha, G., Melgani, F., & Slaghenauffi, J. (2020). A new CNN-RNN framework for remote sensing image captioning. *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, 1-4. https://doi.org/10.1109/M2GARSS47143.2020.9105191
6. Hrga, I., & Ivašić-Kos, M. (2019). Deep image captioning: An overview. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 995-1000. https://doi.org/10.23919/MIPRO.2019.8756821
7. Shahi, T. B., Shrestha, A., Neupane, A., & Guo, W. (2020). Stock price forecasting with deep learning: A comparative study. *Mathematics*, *8*(9), 1441. https://doi.org/10.3390/math8091441
8. Gu, S., & Ding, L. (2018). A complex-valued VGG network based deep learning algorithm for image recognition. *2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*, 340-343. https://doi.org/10.1109/ICICIP.2018.8606702
9. Analytics Vidhya. (2020). Top 4 pre-trained models for image classification with Python code. Retrieved from https://www.analyticsvidhya.com/blog/2020/08/top-4-pre-trained-models-for-image-classification-with-python-code
10. Islam, M. Z., Islam, M. M., & Asraf, A. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using

***Corresponding Author:** Pushpa G

X-ray images. *Informatics in Medicine Unlocked, 20*, 100360. https://doi.org/10.1016/j.imu.2020.100360

11. Yin, Q., Zhang, R., & Shao, X. (2019). CNN and RNN mixed model for image classification. *MATEC Web of Conferences, 277*, 02001. https://doi.org/10.1051/matecconf/201927702001
12. TensorFlow. (n.d.). Image classification with TensorFlow Lite model. Retrieved from https://www.tensorflow.org/lite/examples/image_classification/overview
13. Kaggle. (n.d.). Flickr8k dataset. Retrieved from https://www.kaggle.com/datasets/adityajn105/flickr8k
14. Deeplearning.ai. (n.d.). Retrieved from https://www.deeplearning.ai
15. Brady, E., Morris, M. R., Zhong, Y., White, S., & Bigham, J. P. (2013). Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 2117-2126). Association for Computing Machinery. https://doi.org/10.1145/2470654.2481291
16. Saleem, A., Asif, K. H., Ali, A., Awan, S. M., & Alghamdi, M. A. (2014). Pre-processing methods of data mining. *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 451-456. https://doi.org/10.1109/UCC.2014.57
17. Pushpa, G., Mankame, D. P., Patil, B., & Patil, S. (2023). Telehealth interpretation of COVID-19 patients using artificial intelligence. *Journal of Data Acquisition and Processing, 38*(3), 7224-7230. ISSN 1004-9037
18. Mankame, D. P., Patil, B., Pushpa, G., & Patil, S. (2023). New-fangled internet of things architecture for real-time heart attack menace prediction. *Journal of Data Acquisition and Processing, 38*(3), 7205-7213. ISSN 1004-9037
19. Pushpa, G., & Mankame, D. P. (2023). Threshold alarm algorithm for in-patient monitoring system. *International Journal of Engineering Research & Technology, 9*(6). http://dx.doi.org/10.17577/IJERTV9IS060430
20. Chandraprabha, K. S., Chittragi, N. B., Pushpa, G., & Venkataraman, P. (2016). Context-based in-patient monitoring system. In *2nd International Conference on Theoretical Computing and Communication Technology (iCATccT-2016)*, IEEE.

***Corresponding Author:** Pushpa G