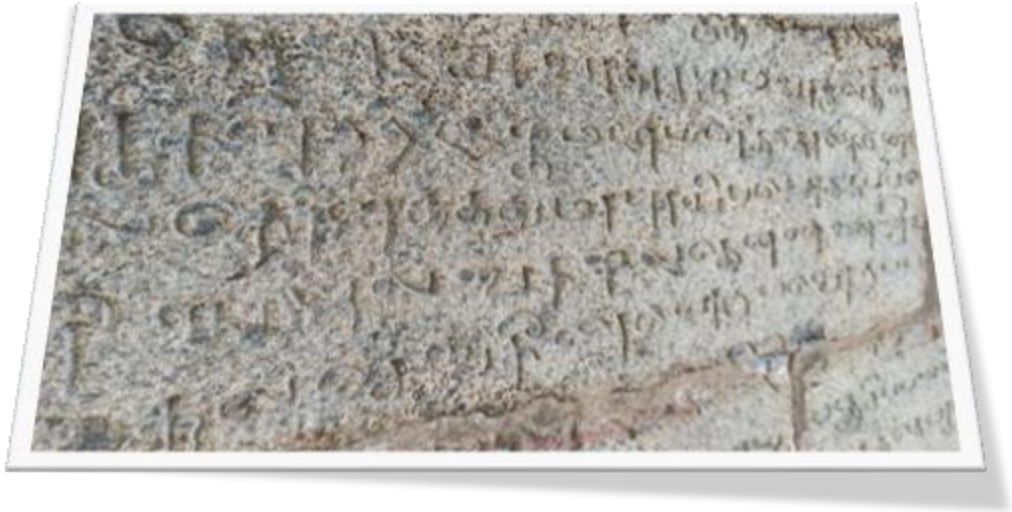
# CONVERTING THAMIZHI SCRIPTS INTO MODERN TAMIL LETTERS USING CHARACTER RECOGNITION TECHNIQUE

- *Kishorekanna S*

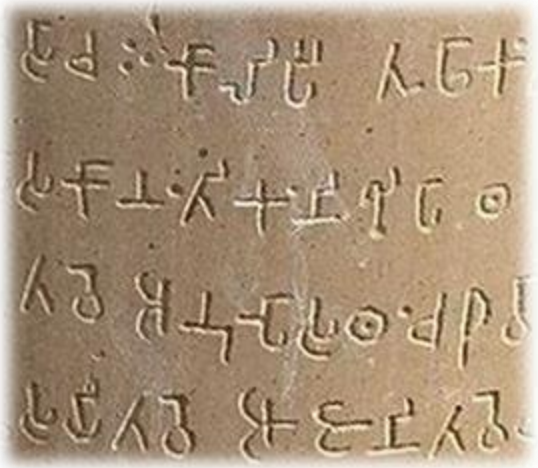
**Table Of Contents:**


- Introduction
- Objective of the project
- Challenge in Existing System
- Contributions
- Dataset Description
- Proposed System
- System Design
- Results and Discussion
- Conclusion and Future Work
- References

# INTRODCUTION

- Ancient Tamil letters are known as Thamizhi or Thamizh Brahmi.
- Thamizhi scripts date from the 3rd century BC.
- They are found in caves, stone beds, potsherds, jar burials, coins, seals, and rings.
- Epigraphers are responsible for modernizing, translating, and inscribing inscriptions on pillars, stones, and rocks in caves.



Thamizhi Script


Grandha Brahmi Script

Vattezhuthu Script

# INTRODCUTION

## DETAILS ABOUT DOMAIN

### ❖ Deep learning:

- Deep learning is a class of machine learning that uses multiple layers to progressively extract higher-level features from the raw input.
- Speech Recognition , Self-Driving Car , Object detection are the some examples of deep learning.

### ❖ Character recognition:

- Character recognition is the process of detecting and processing a character or words and storing it in textual format .

# OBJECTIVE OF THE PROJECT

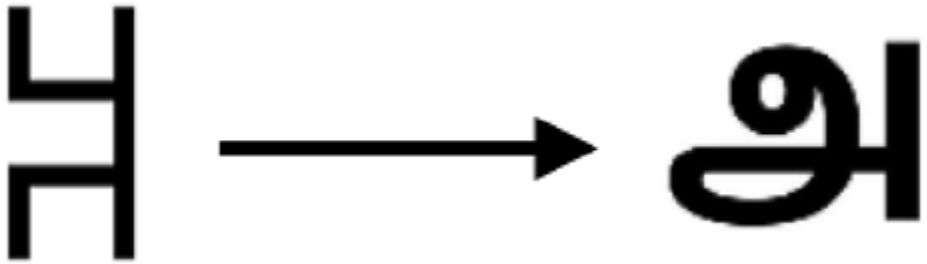
- The proposed system aims to convert the user given Thamizhi scripts into modern Tamil letters, which is in the textual format.
- As the initial process the letter "அ" is alone used for conversion.

| INPUT | PROCESSING | OUTPUT |
| --- | --- | --- |
| Input a Thamizhi letter into the system for recognition. | The machine collects input from the user and compares it against pre-trained models. | The output will be the corresponding Tamil character for the given Thamizhi character. |

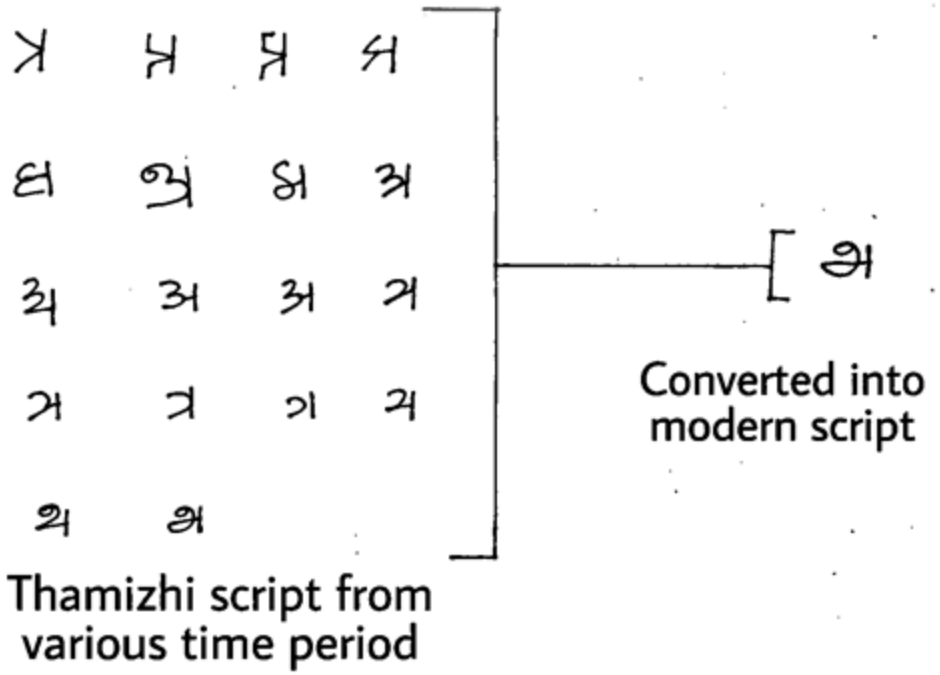
# CHALLENGE IN EXISTING SYSTEM

- Thamizhi script has varied across different time periods from the 3rd century BC to the 20th century AD. There are nearly 18 types of Thamizhi scripts, each evolving gradually over the years. Most existing models are trained using a single type of script from a particular timeline.\

- In the provided image, the letter "அ" has evolved from its Thamizhi form (dating back to the 2nd century) into its modern Tamil counterpart.



# SOLUTION FOR THE CHALLENGE IN EXISTING SYSTEM




- The dataset developed for the proposed system includes various types of Thamizhi scripts.
- This allows the system to predict letters from any time period and convert them into modern script.

# CONTRIBUTIONS

- In most Thamizhi projects, only accuracy has been evaluated without actual conversion being performed.
- This proposed work of conversion will serve as a reference for future projects.
- There is currently no available dataset for Thamizhi scripts.
- Therefore, a proper dataset has been developed and will be published online.
- This will benefit upcoming researchers and developers.
- This work will enable laymen to convert Thamizhi scripts found globally without needing an epigrapher's help.

# DATASET DESCRIPTION

- The proposed dataset for the system consists of the collection of 12 vowels (Uyir Ezhuthukkal) of the Tamil language.
- **Only one vowel "அ" is used for the deployment of the system, which consists of 15 classes (types).**
- These collected letters belong to various time periods, with the maximum possible characters collected for each vowel.
- These scripts are not in circulation; hence, the dataset is created manually (hand-written).


*The above image contains 18 types of "அ" from various time periods. Each letter consists of different types of the same letter.*

# DATASET DESCRIPTION

| Character | அ | ஆ | இ | ஈ | உ | ஊ |
|---|---|---|---|---|---|---|
| No of Types | 18 | 15 | 14 | 13 | 13 | 13 |
| Character | எ | ஏ | ஐ | ஒ | ஓ | ஔ |
| No of Types | 9 | 11 | 8 | 12 | 12 | 3 |

Types of  Tamil vowels individually collected.

# DATASET DESCRIPTION

## Data Augmentation:

- Dataset augmentation is a method used to artificially expand the size of a dataset. In this system, the image dataset undergoes augmentation, a process known as image augmentation.
- The dataset was manually created using handwritten characters. Due to insufficient data, image augmentation was applied with various filters.
- As a result, the dataset's size was increased fivefold through image augmentation.

| Image Augmentation | | | | | |
|---|---|---|---|---|---|
| **Filter** | No filter | Grayscale | Sepia | Contour | Negative |
| **Output** |  |  |  |  |  |

Image Augmentation under various filters

# PROPOSED SYSTEM - METHOD

## ❖ Algorithm Applied:

- CNN:
  - A Convolutional Neural Network (CNN) is a algorithm of deep learning that is specifically designed to process pixel data.
  - CNN is widely used in computer vision for many visual applications such as image classification, video classification.

## ❖ Dataset Used:

- Customized Dataset:
  - Hand Written characters.

# PROPOSED SYSTEM – FRAMEWORK

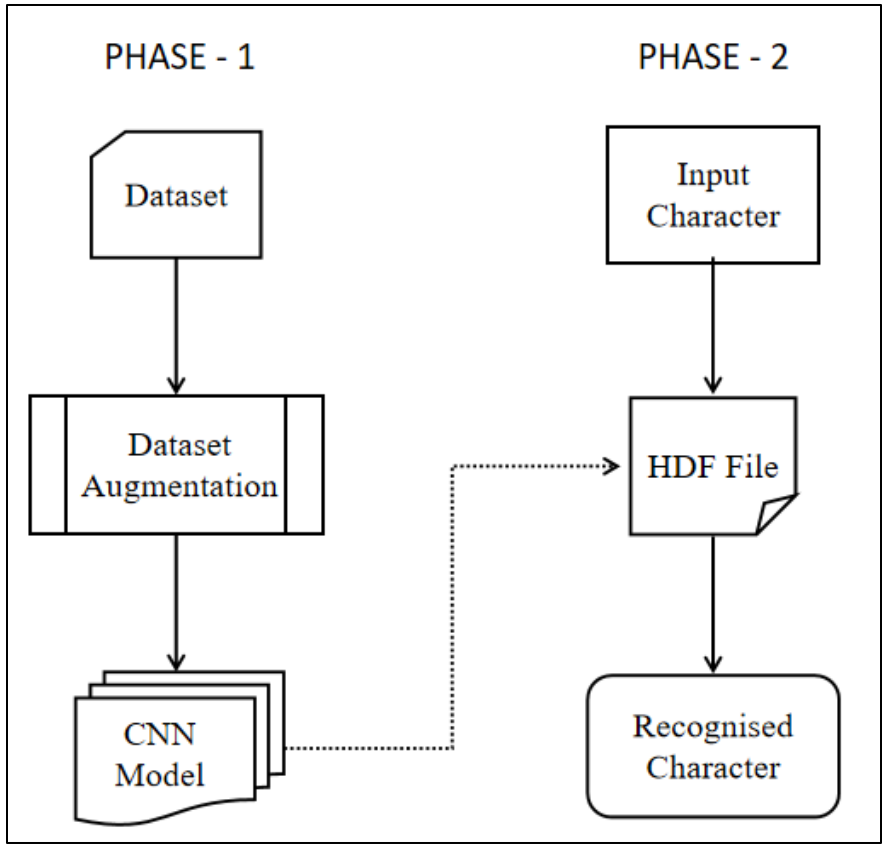



**Fig. Framework of the Proposed System**

| PHASE 1 | |
|---|---|
| **Dataset** | Collection of Thamizhi characters. |
| **Data Augmentation** | The amount of data is artificially increased. |
| **CNN Model** | Convolution Neural Network model is created and saved as an HDF file. |

| PHASE 2 | |
|---|---|
| **Input Character** | Providing the machine with a new or test input. |
| **HDF File** | Extracts the result of the CNN model and saves temporarily. |
| **Recognized Character** | Display the predicted character as an output. |

# SYSTEM DESIGN

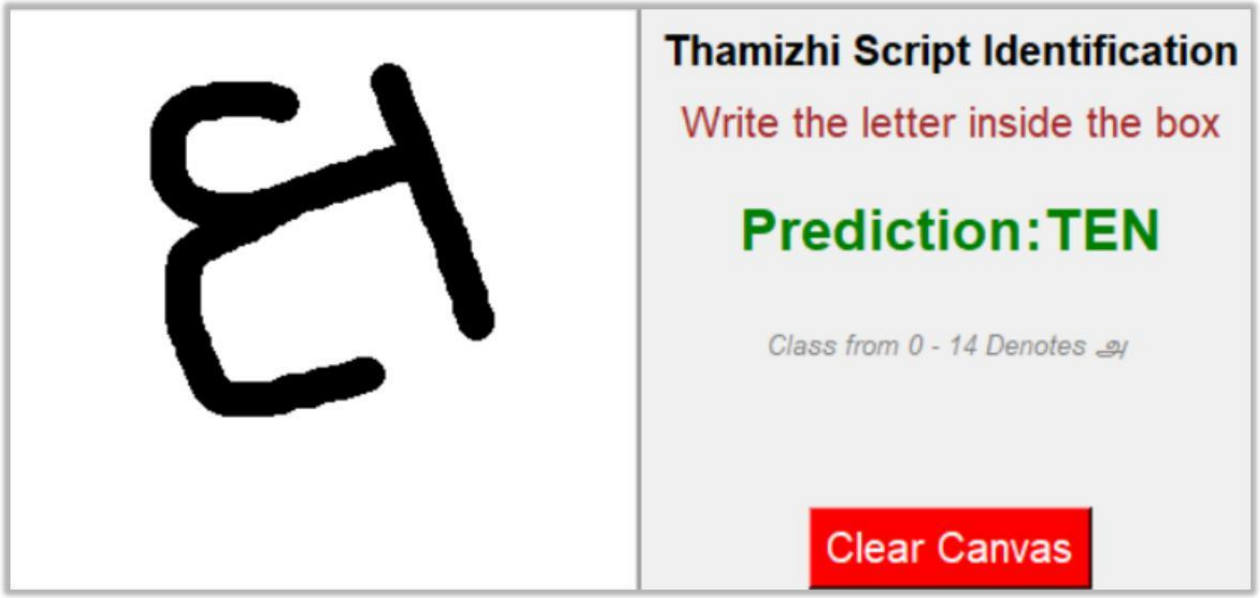
## Hardware:

- System: Intel i3 Processor.
- Hard Disk: 500 GB.
- Ram: 4 GB.
- Graphics Card: 2 GB.
- Input Devices: Keyboard, Mouse.

## Software:

- Operating system: Windows 10 / 11.
- Coding Language: Python.
- IDE: Jupyter Notebook.
- Packages: NumPy, OpenCV, Tensorflow, Sklearn, Tkinter, PIL, H5py.

# RESULTS AND DISCUSSION

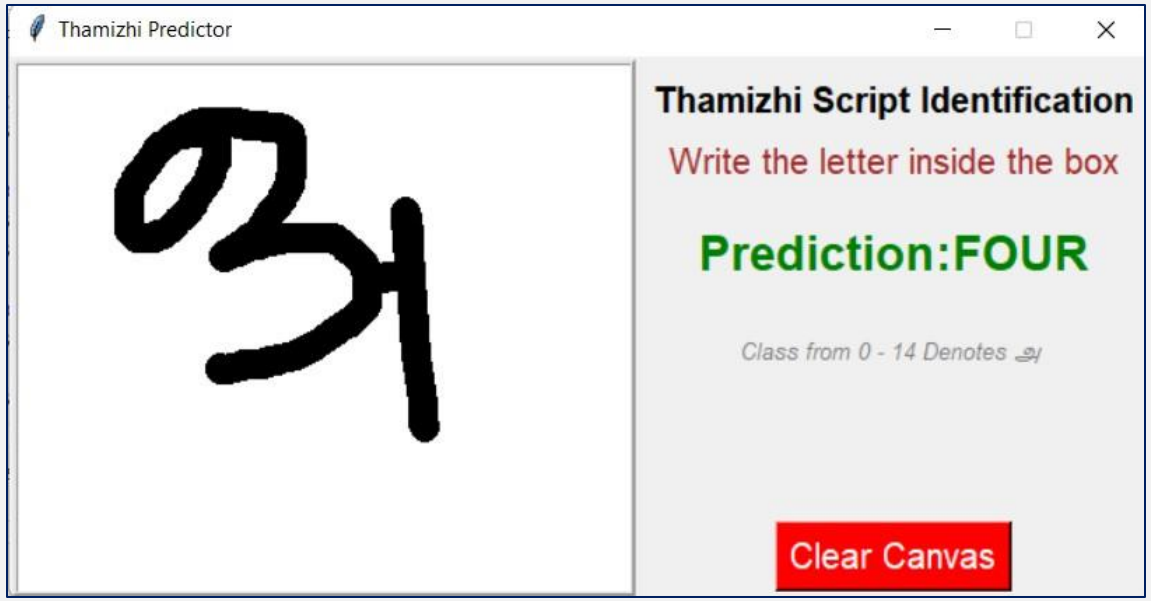



The letter "அ" is classified into 15 distinct types, organized into 15 classes.

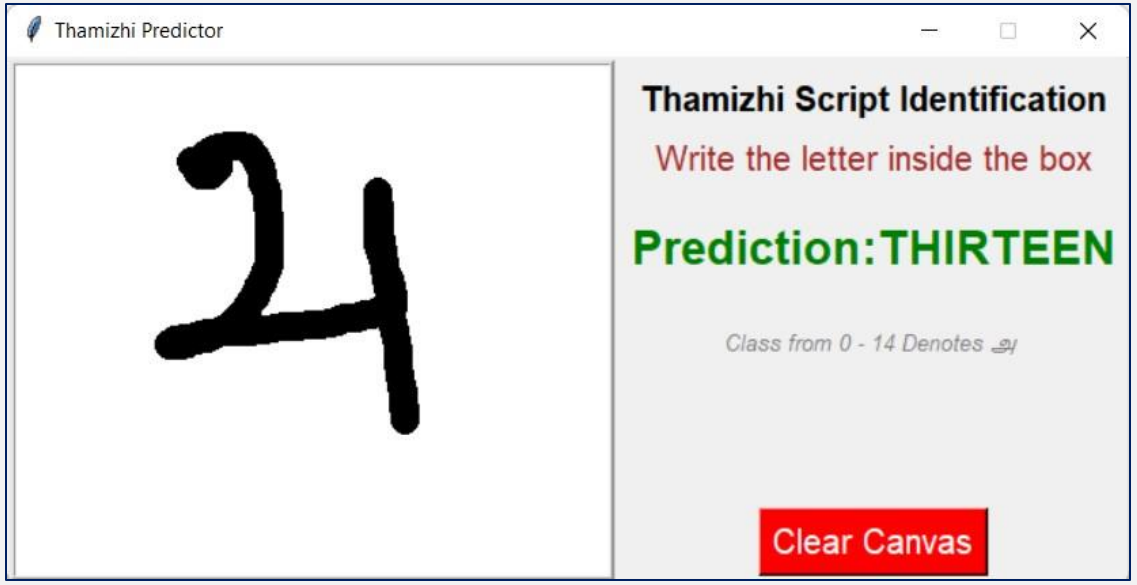
**The user-drawn input is matched with the structure of the specified class by comparing it against the pre-trained model.**

# RESULTS AND DISCUSSION

## SCREENSHOTS OF RESULT




The character belongs to 4th class.




The character belongs to 13th class.

# CONCLUSION AND FUTURE WORK

## FUTURE WORK

- The proposed system was initially developed to convert the letter "அ" only. Future enhancements aim to expand its functionality to support all 247 Tamil characters.

- Currently, the developed dataset includes only the vowels of the Tamil language, comprising 12 letters. Future iterations will extend the dataset to encompass the remaining 235 letters.

- At present, the system accepts user input via a drawpad. In upcoming versions, cameras will be integrated to enable real-time conversion and detection.

- This model has the potential to detect Tamil-Brahmi inscriptions and scriptures. When combined with advanced image recognition techniques and algorithms, it could even interpret inscriptions from rock carvings and ancient manuscripts.

# REFERENCES


1. Devi, P. Dharani, and V. Sathiyapriya. ”Brahmi Script Recognition System using Deep Learning Techniques.” 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2021.
2. Gautam, Neha, Soo See Chai, and Jais Jose. “Recognition of Brahmi Words by Using Deep Convolutional Neural Network.” (2020).
3. Subadivya, S., et al. “Tamil-Brahmi Script Character Recognition System Using Deep Learning Technique.” (2020).
4. Wijerathna, KASA Nilupuli, et al. “Recognition and translation of Ancient Brahmi Letters using deep learning and NLP.” 2019 International Conference on Advancements in Computing (ICAC). IEEE, 2019.
5. Dhivya, S., and G. Usha Devi. “TAMIZHI: Historical Tamil Brahmi Handwritten Data.” Sustainable Communication Networks and Application. Springer, Singapore, 2021. 585-592.
6. Rahman, Md Mahbubar, et al. “Bangla handwritten character recognition using convolutional neural network.” International Journal of Image, Graphics and Signal Processing 7.8 (2015): 42.
7. Maitra, Durjoy Sen, Ujjwal Bhattacharya, and Swapan K. Parui. “CNN based common approach to handwritten character recognition of multiple scripts.” 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015.
8. Prakash, A. Arun, et al. “Tamil Handwritten Character Recognition Using ConvNet Model.” (2019).
9. Chowdhury, Rumman Rashid, et al. “Bangla handwritten character recognition using convolutional neural network with data augmentation.” 2019 Joint 8th International Conference on Informatics, Electronics Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision Pattern Recognition (icIVPR). IEEE, 2019.
10. Baldominos, Alejandro, Yago Saez, and Pedro Isasi. ”A survey of handwritten character recognition with mnist and emnist.” Applied Sciences 9.15 (2019): 3169.

# REFERENCES


11. Bora, Mayur Bhargab, et al. “Handwritten character recognition from images using cnn-ecoc.” Procedia Computer Science 167 (2020): 2403-2409.
12. Truong, Quang Vinh, Hoai Duy Le, and Nguyen Thanh Nhan. “Vietnamese handwritten character recognition using convolutional neural network.” IAES International Journal of Artificial Intelligence 9.2 (2020): 276.
13. Acharya, Minal, Priti Chouhan, and Asmita Deshmukh. “Scan. it-Text Recognition, Translation and Conversion.” 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.
14. Suriya, S., et al. “Computational Linguistics-Based Tamil Character Recognition System for Text to Speech Conversion.” Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches (2021): 129-153.
15. Dessai, Brijeshwar, and Amit Patil. “A deep learning approach for optical character recognition of handwritten Devanagari script.” 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). Vol. 1. IEEE, 2019.
16. Deore, Shalaka Prasad, and Albert Pravin. “Devanagari Handwritten Character Recognition using fine-tuned Deep Convolutional Neural Network on trivial data.”cSadhana 45.1 (2020): 1-13.
17. Memon, Jamshed, et al. “Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR).” IEEE Access 8 (2020): 142642-142668.
18. Narang, Sonika Rani, Munish Kumar, and Manish Kumar Jindal. “DeepNetDevanagari: a deep learning model for Devanagari ancient character recognition.” Multimedia Tools and Applications 80.13 (2021): 20671-20686.
19. Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. “Thamizhimorph: A morphological parser for the Tamil language.” Machine Translation 35.1 (2021): 37-70.

# REFERENCES


11. Bora, Mayur Bhargab, et al. “Handwritten character recognition from images using cnn-ecoc.” Procedia Computer Science 167 (2020): 2403-2409.
12. Truong, Quang Vinh, Hoai Duy Le, and Nguyen Thanh Nhan. “Vietnamese handwritten character recognition using convolutional neural network.” IAES International Journal of Artificial Intelligence 9.2 (2020): 276.
13. Acharya, Minal, Priti Chouhan, and Asmita Deshmukh. “Scan. it-Text Recognition, Translation and Conversion.” 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.
14. Suriya, S., et al. “Computational Linguistics-Based Tamil Character Recognition System for Text to Speech Conversion.” Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches (2021): 129-153.
15. Dessai, Brijeshwar, and Amit Patil. “A deep learning approach for optical character recognition of handwritten Devanagari script.” 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). Vol. 1. IEEE, 2019.
16. Deore, Shalaka Prasad, and Albert Pravin. “Devanagari Handwritten Character Recognition using fine-tuned Deep Convolutional Neural Network on trivial data.”cSadhana 45.1 (2020): 1-13.
17. Memon, Jamshed, et al. “Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR).” IEEE Access 8 (2020): 142642-142668.
18. Narang, Sonika Rani, Munish Kumar, and Manish Kumar Jindal. “DeepNetDevanagari: a deep learning model for Devanagari ancient character recognition.” Multimedia Tools and Applications 80.13 (2021): 20671-20686.
19. Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. “Thamizhimorph: A morphological parser for the Tamil language.” Machine Translation 35.1 (2021): 37-70.