OCR Grading Assistant: Data field extraction from mark sheet using OCR

PCSE25-71

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY
Computer Science and Engineering

SUBMITTED BY

SHREYA SHARMA, SAMIKSHA GUPTA,

MOHAMMED FAAIZ, AKSSHAT GOVIND

October 2023

**KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)**

TABLE OF CONTENTS

Introduction

Imagine a world where the whispers of pen on paper aren't lost to the digital realm. Where handwritten answers on sheets dance seamlessly into databases, unlocking instant feedback and valuable insights. Where historical manuscripts reveal their secrets, their stories transformed into searchable digital symphonies. This is the vision that drives our innovative project: building a cutting-edge deep learning-powered Optical Character Recognition (OCR) system.

The current landscape of paper-based assessments and data capture is ripe for transformation. Manual data entry is tedious, prone to errors, and stifles the flow of valuable information. Existing OCR solutions often stumble with the inherent complexities of handwritten text, limiting their application and accuracy. We aim to rewrite the narrative, unlocking the vast potential of handwritten data by creating an OCR system that not only transcribes, but truly comprehends.

Our system promises a future where:

- Educators: Upload faculty-marked answer sheets effortlessly, with grades instantly digitized and analyzed for faster feedback and deeper understanding of student performance.
- Researchers: Delve into the whispers of historical documents, unlocking their stories and secrets through accurate text extraction and digital accessibility.
- Healthcare professionals: Access a symphony of medical records, their insights readily available for improved analysis and patient care.
- Everyone: Transforms hand-written forms, surveys, and documents into easily searchable and usable digital assets.

But how do we achieve this revolutionary leap? Through the power of deep learning, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) trained on vast datasets of diverse handwritten text. These algorithms waltz with every inkwell nuance, adapting to individualities and conquering complexities with exceptional accuracy.

Our system goes beyond mere transcription – it comprehends the essence of handwritten information, automatically identifying document types, classifying data, and even extracting structured information from complex forms.

RATIONALE

The laborious ballet of data entry waltzes with frustration and error in the world of paper-based assessments and document capture. Traditional Optical Character Recognition (OCR) systems attempt the pirouette, but often stumble on the complexities of handwritten text, their melodies marred by inaccuracies and limited scope. We propose a revolutionary fouetté – a deep learning-powered OCR system that not only transcribes, but truly comprehends, transforming handwritten expressions into a vibrant digital symphony.

This leap is driven by necessity. Manual data entry stifles the flow of vital information, its slow tempo frustrating educators, researchers, and professionals across industries. Existing OCR solutions, with their limited repertoire, cannot keep pace with the diverse rhythms of handwritten text, leaving valuable data trapped in the inkwell.

Our system pirouettes around these limitations with:

- Deep learning grace: Convolutional Neural Networks and Recurrent Neural Networks waltz with every nuance of handwritting, conquering complexities with breathtaking accuracy.
- Unleashing versatility: From answer sheets to historical manuscripts, our system adapts to a multitude of document types, each note of information readily transcribed.
- Beyond transcription: We don't merely listen – we comprehend. Automatic document identification, data classification, and structured information extraction make the dance of information retrieval smooth and effortless.

This is not just technology; it's a bridge between inkwell whispers and digital symphony. Join us as we rewrite the story of OCR, unlocking the vast potential of handwritten data and empowering a future where every stroke of pen contributes to a richer, more connected world.

OBJECTIVES

Our deep learning-powered OCR system aims to achieve the following objectives:

Accuracy and Versatility:

- Develop an OCR system that surpasses existing accuracy thresholds for recognizing diverse handwriting styles and document layouts.
- Ensure accurate character recognition even in challenging scenarios with noise, smudges, and variations in writing pressure.
- Make the system adaptable to a wide range of document types, including answer sheets, historical manuscripts, medical records, forms, and surveys.

Efficiency and Automation:

- Automate the process of extracting handwritten data from paper documents, eliminating the need for manual data entry and transcription.
- Significantly reduce data processing time compared to traditional methods, enabling faster feedback and information analysis.
- Streamline data capture processes for various applications, enhancing workflow efficiency and data accessibility.

Advanced Data Comprehension:

- Move beyond simple transcription by implementing intelligent algorithms that can identify document types, classify data, and extract structured information from complex forms.
- Develop contextual understanding of extracted text, enabling richer analysis and improved knowledge discovery.
- Integrate language processing capabilities to further enhance the semantic interpretation of handwritten information.

Impact and Open Access:

- Demonstrate the practical application of deep learning in solving real-world problems related to handwritten data extraction.
- Contribute to the advancement of OCR technology by making our system and training data open-source, fostering further research and development.
- Empower diverse stakeholders across education, research, healthcare, and various industries with a powerful tool for unlocking the potential of handwritten data.

LITERATURE REVIEW REPORT

PAPER 1

A Novel Approach – Automatic paper evaluation system

AUTHOR: Devaki Priya , Harini , Haripriyaa , Dharaniya

PUBLISHED DATE: 03 March 2019

The document discusses an automatic paper evaluation system that utilizes machine learning and natural language processing techniques. The system involves the conversion of handwritten answer sheets into text documents using optical character recognition (OCR) and then compares the extracted text with reference answers stored in a database to assign marks to students. The system consists of three main modules: text extraction, text comparison, and mark evaluation and updation.

The text extraction module captures answer sheet images and converts them into editable text using OCR. The text comparison module compares the extracted text with reference answers and provides a threshold value based on the similarity level. Marks are then allocated according to the similarity measures obtained using natural language processing (NLP) techniques. The mark evaluation and updation module stores the allocated marks in the database, accessible to both staff and students through separate login credentials.

The system's implementation is detailed, showcasing images of text extraction, text comparison, and mark evaluation and updation. The results and discussions section demonstrates the system's performance in evaluating student answers and allocating marks, with a focus on the importance of machine learning and NLP techniques in achieving a high level of accuracy, up to 85 percent.

The document also discusses various algorithms and techniques used in the system, such as the Cuckoo search algorithm for multi-document summarization, fuzzy rules for text summarization, and convolutional neural network (CNN) based Chinese text detection algorithms. Additionally, the document highlights the importance of automated data processing, spell checkers, and the use of synthetic data for text localization in natural images.

In summary, the document provides a comprehensive overview of the automatic paper evaluation system, including its architecture, performance, and the various algorithms and techniques used to enhance its efficiency.

PAPER 2

Text Extraction and Recognition from Image using Neural Network

AUTHOR: C. Misra, P.K Swain, J.K Mantri

PUBLISHED DATE: 02 February 2012

The document discusses a comprehensive approach for text extraction and recognition from images using a neural network. The primary objective is to develop an unconstrained image indexing and retrieval system. The authors propose a novel color reduction technique using the HSV color space and extract a set of features from each Region of Interest (ROI) for specific color planes. These features are used in a feature-based classifier to determine if the ROI contains text or non-text blocks. The text blocks identified are then processed by an Optical Character Recognition (OCR) system, and the output in the form of ASCII characters forming words is stored in a database for future retrieval.

The document outlines the method for text extraction and recognition, including the extraction of features such as size analysis, aspect ratio, contrast changes per unit length, inter-character gap, foreground pixel density, and ratio of foreground pixel to background pixel. The authors also discuss the evaluation of the performance of the feature set, including the determination of the contribution of each feature in the feature set using the F-ratio based method. The reduction of the feature vector dimension is also highlighted as a means to improve the classification result and computation efficiency.

Furthermore, the document presents the process of selecting the optimum feature set using Singular Value Decomposition (SVD) and the application of a multilayer perceptron (MLP) as a classifier to label the ROIs as text or non-text. The training and testing of the MLP using the back propagation algorithm are discussed, along with the method for text extraction and recognition, including color polarity detection, binarization of cropped images, removal of touching characters and noise, and OCR-based identification and storage of text in a database.

In conclusion, the document provides a comprehensive overview of the text extraction and recognition process, highlighting the need for further improvements, such as extending the system to handle non-horizontally oriented text, better tracking of text with complex motion, and improving the recognition accuracy for text with complex backgrounds. The document also references related work in the field of text extraction and recognition, providing a comprehensive overview of the current state of the art in this area.

PAPER 3

Data extraction from exam answer sheets using OCR with adaptive calibration of environmental threshold parameters

AUTHOR: Deepak Sharma, Himanshu Sharma, Avinav Sharan, Arpit Agarwal

PUBLISHED DATE: December 2013

The paper discusses a system for automating the extraction of data from exam answer sheets using Optical Character Recognition (OCR) with adaptive calibration of environmental threshold parameters. The system is designed to recognize enrollment numbers and corresponding marks from answer sheets, process the data, and store it in a computer. It consists of hardware that captures images of answer scripts and a software component that processes the images using an adaptive threshold for color detection (ATCD) algorithm. The algorithm extracts the region of interest (ROI), which contains the enrollment number and marks of the student, and passes it to the OCR system for character extraction. The hardware involves a conveyor belt system, a webcam for image capture, and a voice module for user instructions. The system aims to replace the manual process of data collection and entry, enabling automatic updating of student marks in the university's server.

The hardware processes papers one by one, capturing images of the front page of each answer sheet. The image processing algorithm involves preprocessing camera images to reduce noise, adaptive threshold for color detection to segment the box containing the characters, binarization, and image enhancement. Additionally, the original image is binarized using the Otsu method. The OCR process includes character segmentation, categorization, and template matching using a database of numerals.

The system demonstrated an average accuracy of 81% in various light and paper conditions, having been trained with 50 samples of numerals (0-9) for the OCR system.

The study concludes by emphasizing the potential of OCR technology in automating data entry processes, reducing labor, and contributing to a paperless world. The authors suggest that as OCR technology continues to advance, it will lead to 100% accuracy, making automated data entry through OCR a highly attractive feature for data entry systems.

In summary, the paper presents a comprehensive system for automating the extraction of data from exam answer sheets, combining hardware for image capture and software for image processing and OCR. The system aims to improve efficiency, accuracy, and automation in the process of recording and updating student marks. The study highlights the potential of OCR technology in revolutionizing data entry applications.

PAPER 4

InMAS: Deep Learning for Designing Intelligent Making System

AUTHOR: LEI SHAO, MAOYANG LI, LIANJUN YUAN, GUAN GUI

PUBLISHED DATE: April 29, 2019

The document presents a method called InMAS, based on the You Only Look Once (YOLOv3) algorithm, for developing intelligent making systems (InMASs) to automate the grading of students' test papers and assignments. The proposed method involves creating two datasets: one for localization and the other for recognition of arithmetic problems. The YOLOv3 network is used for identifying the location and extraction of each mathematical problem in images and recognizing the characters in each arithmetic problem.

The traditional OCR technology, particularly Baidu OCR, is shown to have low recognition accuracy for arithmetic problems, particularly with handwritten characters. To address this, the proposed method uses deep learning-based YOLOv3 algorithm for character recognition, achieving a recognition accuracy of 97.15%. The method also includes a template matching approach to mark the bounding box of incorrectly evaluated arithmetic problems in the original pictures.

Experimental results demonstrate that the proposed method has near-perfect accuracy for localization and outperforms the Baidu OCR in recognition accuracy for arithmetic problems. The proposed InMAS method is designed to reduce the workload of teachers and ensure the accuracy of scoring. The method can effectively help in correcting examination papers and reduce the workload of teachers and parents in education.

Overall, the proposed method offers a comprehensive solution for automating the grading of students' test papers and assignments, leveraging deep learning-based algorithms for superior localization and recognition accuracy compared to traditional OCR technology.

PAPER 5

Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)

AUTHOR: Jamshed Memon, Maira Sami , Rizwan Ahmed Khan

PUBLISHED DATE: 1 Jan 2020

The document is a comprehensive review of the research conducted on Optical Character Recognition (OCR) for handwritten documents. The review covers the period from 2000 to 2018 and focuses on six languages: English, Arabic, Indian, Chinese, Urdu, and Persian. The document aims to summarize the primary details, central ideas, and crucial conclusions of the review, which includes the review methodology, statistical results, classification methods, datasets, and languages studied.

The review describes the ubiquity of handwritten documents in human transactions and the invaluable practical worth of OCR in translating various types of documents or images into analyzable, editable, and searchable data. The document outlines the systematic literature review protocol, including the review background, search strategy, data extraction, research questions, and quality assessment criteria for the selection of study and data analysis. The review also presents the statistical results of the selected studies, including distribution of publication sources, citation count status, temporal view, and the type of languages and research methodologies studied.

Furthermore, the document details the classification methods of handwritten OCR, including artificial neural networks (ANN), kernel methods, statistical methods, template matching techniques, and structural pattern recognition. It discusses the datasets used for evaluation and benchmarking of OCR algorithms, such as CEDAR, MNIST, UCOM, IFN/ENIT, CENPARMI, HCL2000, and IAM. The document also provides information on the languages studied, with a focus on English, which has the highest number of publications in OCR research.

Overall, the review aims to highlight the state-of-the-art results and techniques in OCR for different languages, as well as to provide research directions and identify research gaps in the field of OCR for handwritten documents.

FEASIBILITY STUDY

1.Document Quality:

Evaluate the quality of the mark sheets. If the documents are well-printed, have a standardised format, and exhibit good readability, the feasibility of data extraction increases.

2. OCR Technology:

Select or develop an OCR technology that is capable of accurately recognizing characters from scanned or photographed images. Modern OCR tools can handle both printed and handwritten text to varying degrees.

3.Template Variability:

Assess the consistency of mark sheet templates. If mark sheets follow a standard format across different educational institutions or batches, it becomes easier to design an OCR solution that can identify and extract data fields consistently.

4.Data Field Complexity:

Consider the complexity of the data fields you need to extract. Some OCR systems may struggle with handwritten text or complex symbols. Ensure that the chosen OCR technology can handle the specific types of data present in mark sheets.

5. Accuracy Requirements:

Define the level of accuracy required for data extraction. Depending on the project's goals, you may need a high level of accuracy, especially if the extracted data will be used for critical processes or decision-making.

6. Preprocessing Techniques:

Implement preprocessing techniques to enhance the quality of the input images before OCR. Techniques such as image cleanup, noise reduction, and contrast adjustment can significantly improve OCR accuracy.

SIGNIFICANCE

Automation of data field extraction from mark sheets significantly reduces the time and effort required for manual data entry. This can lead to increased efficiency in educational institutions, especially during large-scale examination result processing.

Manual data entry is prone to errors, including typos and data entry mistakes. Implementing OCR for mark sheet data extraction can help minimize errors, ensuring more accurate and reliable data.

Automation of data extraction processes speeds up the overall data processing time. This is particularly crucial in educational institutions where timely release of examination results is essential.

The project contributes to better record-keeping and data management. Extracted data can be stored in digital formats, making it easier to organize, retrieve, and analyze over time.

Digitalized and extracted data can be easily accessed and shared across different departments or systems, facilitating better collaboration and communication within educational institutions.

Extracted data can be used for analytical purposes, such as generating reports and insights into student performance trends. This can aid educators, administrators, and policymakers in making data-driven decisions.

The project's scalability allows it to handle a large volume of mark sheets, making it suitable for educational institutions with varying sizes and examination scales.

METHODOLOGY / PLANNING OF WORK

Research Type: Applied research.

Unit of Study: Data extraction field using OCR.

1. Project Planning and Requirements Analysis:

1.1 Define Project Objectives:

Clearly articulate the goals and objectives of the project. Determine what specific data fields you need to extract from mark sheets.

1.2 Stakeholder Analysis:

Identify and engage with key stakeholders, including educators, administrators, and IT professionals, to understand their requirements and expectations.

1.3 Scope Definition:

Clearly define the scope of the project, specifying the types of mark sheets to be processed, the data fields to be extracted, and any specific constraints or limitations.

2. Literature Review and Technology Selection:

2.1 Review OCR Technologies:

Conduct a literature review to understand the latest advancements in OCR technologies. Identify OCR tools that are suitable for mark sheet data extraction.

2.2 Evaluate Template Recognition Tools:

Explore tools or techniques for template recognition, which can be crucial for identifying and extracting data fields in a structured manner.

3. Data Collection and Preprocessing:

3.1 Collect Sample Mark Sheets:

Gather a representative sample of mark sheets that will be used for testing and training the OCR system.

3.2 Data Preprocessing:

Implement preprocessing techniques to enhance the quality of the input images. This may include image cleanup, noise reduction, and contrast adjustment.

4. OCR Model Development:

4.1 Train OCR Model:

Use the collected data to train the OCR model. Consider training the model to recognize both printed and handwritten text, if applicable.

4.2 Template Recognition:
If needed, integrate template recognition algorithms to identify and locate specific data fields on mark sheets.

5. System Implementation:
5.1 Develop Data Extraction System:
Implement the data extraction system, integrating the trained OCR model and any template recognition tools. Ensure that the system is scalable and capable of handling a large volume of mark sheets.

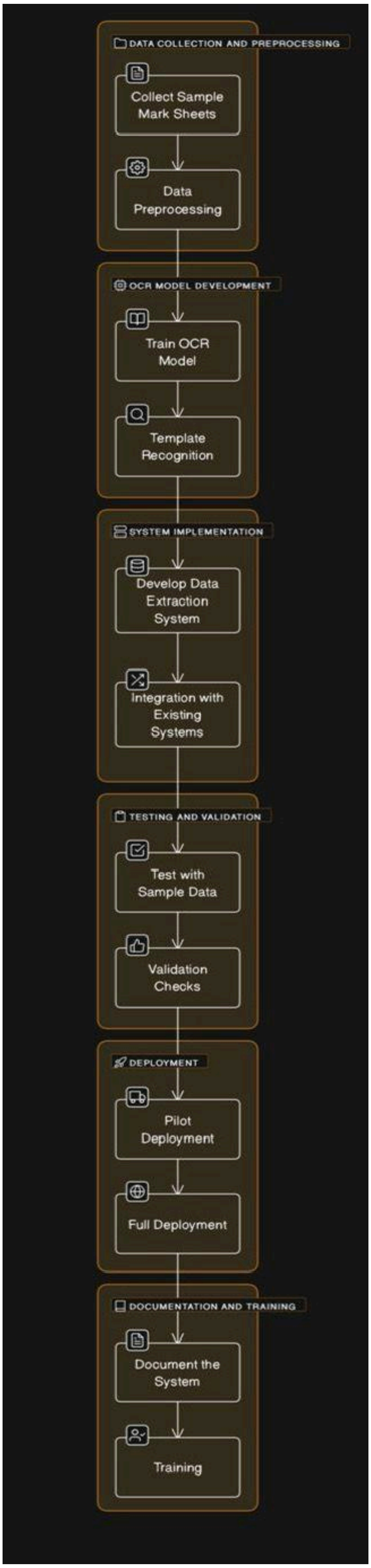
5.2 Integration with Existing Systems:
If applicable, integrate the data extraction system with existing educational management systems or databases.

6. Testing and Validation:
6.1 Test with Sample Data:
Conduct extensive testing using the sample mark sheets to validate the accuracy of data extraction. Identify and address any issues or inaccuracies.

6.2 Validation Checks:
Implement validation checks to ensure the accuracy and reliability of the extracted data. Include error handling mechanisms.

7. Deployment:
7.1 Pilot Deployment:
Deploy the system in a controlled environment for a pilot phase. Gather feedback from users and make any necessary adjustments.

7.2 Full Deployment:
Roll out the system for full-scale deployment, ensuring that it aligns with the operational needs of educational institutions.



8. Documentation and Training:

8.1 Document the System:

Create comprehensive documentation for the data extraction system, including user manuals and technical documentation.

8.2 Training:

Provide training sessions for end-users, administrators, and IT staff to ensure the proper use and maintenance of the system.

9. Monitoring and Maintenance:

9.1 Monitoring Tools:

Implement monitoring tools to track the performance of the data extraction system. Set up alerts for potential issues.

9.2 Ongoing Maintenance:

Establish a plan for ongoing maintenance, including regular updates to the OCR model and system enhancements based on user feedback.

10. Evaluation and Continuous Improvement:

10.1 Performance Evaluation:

- Periodically evaluate the performance of the data extraction system against defined metrics. Gather feedback from users to identify areas for improvement.

10.2 Continuous Improvement:

- Implement continuous improvement processes based on evaluation results. Consider updates to OCR models, system features, and user interfaces.

FACILITIES REQUIRED FOR PROPOSED WORK

Physical Facilities:

Classrooms or Evaluation Centers: Dedicated spaces with appropriate seating arrangements for manual evaluation by teachers.
Secure Storage: Facilities for securely storing physical answer sheets to prevent loss or damage.

Hardware Facilities:
Computers: For automated systems, a sufficient number of computers with the required specifications to run the answer sheet evaluation software.
Scanners: High-quality scanners capable of efficiently capturing text and images from physical answer sheets.
Printers: To generate hard copies of answer sheets if necessary.

Software Facilities:
Answer Sheet Evaluation Software: Whether commercial or custom-developed, software to facilitate the grading process, manage data, and generate reports.
OCR Software: If utilizing Optical Character Recognition, appropriate software for accurately extracting text from scanned answer sheets.

Database Management System (DBMS):
To store and manage evaluation data securely.

Network Infrastructure:
Local Area Network (LAN): For connecting computers, scanners, and printers within the facility.
Internet Connectivity: Required for online systems or for accessing updates and support services.

Training Facilities:

Training Rooms: Dedicated spaces for conducting training sessions for evaluators and other users.
Training Materials: Presentation tools, training manuals, and other materials for educating users on the answer sheet evaluation process and system usage.

Help Desk:

A centralized support system to address user queries and issues.
Maintenance Workshop: If applicable, a facility for repairing and maintaining hardware components.

EXPECTED OUTCOMES

Here are some anticipated outcomes:

1. **Efficient Data Processing:**

Outcome: Streamlined and accelerated data processing of mark sheets.

Impact: Reduction in the time and effort required for manual data entry, leading to more efficient and timely processing of examination results.

2.**Accuracy Improvement:**

Outcome: Increased accuracy in data extraction compared to manual entry.

Impact: Minimization of errors, typos, and data entry mistakes, contributing to more reliable and trustworthy educational records.

3. **Time Savings:**

Outcome: Faster turnaround time for the release of examination results.

Impact: Improved efficiency allows for quicker dissemination of student performance information, benefiting both students and educational institutions.

4. **Enhanced Data Accessibility:**

Outcome: Digitalized and easily accessible data.

Impact: Improved accessibility and availability of extracted data for various stakeholders, fostering better collaboration and informed decision-making.

5. **Improved Record-Keeping:**

Outcome: Well-organized digital records.

Impact: Enhanced record-keeping capabilities, making it easier to manage, retrieve, and analyze historical examination data.

6. **Data Analysis and Reporting:**

Outcome: Availability of structured data for analysis.

Impact: Empowerment of educators, administrators, and policymakers with data-driven insights into student performance trends, allowing for better decision-making.

7. **Consistency Across Institutions:**

Outcome: Standardization of data extraction processes.

Impact: Increased consistency in handling mark sheets across different educational institutions, facilitating comparisons and analyses on a broader scale.

REFERENCES

[1] *Hathaliya, Jigna J., et al. "Securing electronics health care records in healthcare 4.0: A biometric-based approach." Computers & Electrical Engineering 76 (2019): 398-410.*

[2] *Nwosu, Kingsley C. "Mobile Facial Recognition System for Patient Identification in Medical Emergencies for Developing Economies." (2016).*

[3] *Ali, S.; Abdullah; Armand, T.P.T.; Athar, A.; Hussain, A.; Ali, M.; Yaseen, M.; Joo M.-I.; Kim, H.-C. "Metaverse in Healthcare Integrated with Explainable AI and Blockchain: Enabling Immersiveness, Ensuring Trust, and Providing Patient Data Security."*

[4] *Orciuoli, Francesco, Francesco J. Orciuoli, and Angela Peduto. "A Mobile Clinical DSS based on Augmented Reality and Deep Learning for the home cares of patients afflicted by bedsores." Procedia Computer Science 175 (2020)*

[5] *Jain, Yash, et al. "Mental and physical health management system using ML, computer vision and IoT sensor network." (2020)*