A REPORT
ON

AADHAR MASKING AND OCR BASED TEXT EXTRACTION

Submitted by,

Mr. ALLEN CONROY DSOUZA – 20211CSE0227

Under the guidance of,

Dr. M CHANDRA SEKHAR

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY

BENGALURU

MAY 2025

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Internship/Project report **"AADHAR MASKING AND OCR BASED TEXT EXTRACTION"** being submitted by "Mr. Allen Conroy Dsouza" bearing roll number "20211CSE0227" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

Dr. M Chandra Sekhar
Professor
PSCS
Presidency University

Dr. Asif Mohammed
HOD, CSE
PSCS
Presidency University

Dr. Mydhili Nair
Associate Dean
PSCS
Presidency University

Dr. Sameeruddin Khan
Pro-Vice Chancellor - Engineering
Dean –PSCS / PSIS
Presidency University

ii

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

I hereby declare that the work, which is being presented in the report entitled **"AADHAR MASKING AND OCR BASED EXTRACTION"** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of my own investigations carried under the guidance of **Dr. M Chandra Sekhar, Presidency School of Computer Science and Engineering, Presidency University, Bengaluru.**

I have not submitted the matter presented in this report anywhere for the award of any other Degree.

Name	Allen Gonroy D'souza
Roll No	20211CSE0227
Signature	

iii

ABSTRACT

This internship report presents the work conducted during my tenure at Abylle Solutions, focused on developing an automated Aadhar card number masking system using Optical Character Recognition (OCR) and image processing techniques. The core objective was to ensure privacy compliance by detecting and masking sensitive Aadhar digits in scanned or photographed documents. Leveraging Python-based tools such as Tesseract OCR and OpenCV, I engineered a pipeline to extract text from images, accurately identify Aadhar numbers using regular expressions, and apply masking algorithms to conceal the digits. The project involved a hands-on understanding of text extraction challenges from diverse image formats, handling noise, orientation corrections, and ensuring consistent masking across varying layouts. Through iterative testing and optimization, the model achieved robust accuracy in detecting and anonymizing Aadhar numbers, thereby contributing to secure data handling in line with data protection protocols. This experience enhanced my skills in image preprocessing, text recognition, and real-world deployment of data privacy solutions using computer vision.