A Mini Project Synopsis on

Sanskrit OCR

S.E. - D.S Engineering

Submitted By

Krishna Gupta	**21107024**
Meris Gada	**21107041**
Tushar Goud	**21107027**

Under The Guidance Of

Prof.Vaibhav Yavalkar

DEPARTMENT OF CSE (DATA SCIENCE)

A.P.SHAH INSTITUTE OF TECHNOLOGY

G.B. Road, Kasarvadavali, Thane (W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic year : 2023-24

CERTIFICATE

This to certify that the Mini Project report on **Sanskrit OCR** has been submitted by **Krishna Gupta** (21107024), **Meris Gada** (21107041) and **Tushar Goud** (21107027) who are a Bonafede students of A. P. Shah Institute of Technology, Thane, Mumbai, as a partial fulfilment of the requirement for the degree in **CSE(DATA SCIENCE)**, during the academic year **2023-2024** in the satisfactory manner as per the curriculum laid down by University of Mumbai.

Prof.Vaibhav Yavalkar
Guide

Prof. Anagha Aher
Head of the Department of CSE(Data science)

Dr. Uttam D.Kolekar
Principal

External Examiner(s)
1.
2.

Place:A.P.Shah Institute of Technology, Thane
Date:

TABLE OF CONTENTS

Chapter 1

Introduction

An Optical Character Recognition (OCR) project for Sanskrit is a pioneering endeavor aimed at leveraging cutting-edge technology to bridge the gap between the rich heritage of Sanskrit literature and the digital age. Sanskrit, one of the world's oldest languages, holds a treasure trove of ancient texts, scriptures, and knowledge. However, due to its complex script and limited digitization, accessing and preserving this profound cultural heritage can be challenging.

The Sanskrit OCR project seeks to overcome these hurdles by developing a sophisticated OCR system specifically tailored to recognize and digitize Sanskrit characters and texts. This undertaking not only preserves the valuable Sanskrit literary heritage but also makes it accessible to a global audience, researchers, scholars, and enthusiasts. By harnessing the power of artificial intelligence and machine learning, this project endeavors to accurately transcribe Sanskrit manuscripts, inscriptions, and documents into a digital format, making them searchable, editable, and readily available for academic research, translation, and analysis. This introduction reflects the importance and significance of the Sanskrit OCR project in preserving and propagating the timeless wisdom of Sanskrit texts and furthering the study of this ancient language in the modern era. The Sanskrit OCR project emerges as a response to these challenges, fueled by a vision to unlock the treasures buried within the pages of Sanskrit manuscripts, inscriptions, and historical documents. It seeks to harness the capabilities of advanced artificial intelligence, machine learning, and computational linguistics to facilitate the seamless conversion of Sanskrit characters and texts into a digital format. In doing so, this endeavor not only acts as a bridge between the past and the future but also carries profound implications for scholarship, research, cultural preservation, and cross-cultural dialogue.

1.1. Purpose:

The Sanskrit OCR (Optical Character Recognition) project serves a multifaceted purpose. Firstly, it is driven by the imperative to preserve the rich Sanskrit heritage. Many ancient Sanskrit texts are stored in vulnerable manuscript form, subject to

degradation over time. The project's primary purpose is to digitally archive and safeguard this valuable literature for posterity. Secondly, it enhances accessibility, breaking down the historical barriers presented by the intricate Sanskrit script. Through OCR, these texts become readily available to a global audience, empowering scholars, students, and enthusiasts to explore Sanskrit literature without geographical constraints. Furthermore, it advances research and scholarship by facilitating swift searches, analyses, and comparisons of Sanskrit texts. Scholars can extract meaningful data from extensive corpora, accelerating their work in fields such as Sanskrit studies, linguistics, and philosophy. The project also fosters cross-cultural understanding by enabling the translation of ancient Indian knowledge found in Sanskrit texts, contributing to global intellectual dialogue. As an educational resource, OCR-processed Sanskrit texts enhance teaching and learning, benefiting both language learners and educators. Lastly, the Sanskrit OCR project has the potential to drive advancements in both linguistic and computational research, challenging OCR technology to address the complexities of the Sanskrit script and encouraging the development of more sophisticated character recognition algorithms.

1.2.Objectives:

Preservation of Sanskrit Heritage: One of the primary goals of this project is to ensure the long-term preservation of Sanskrit texts. Many original manuscripts and inscriptions are aging, fragile, or at risk of being lost due to environmental factors. The OCR technology can help in digitizing these texts, ensuring their survival for future generations.

Accessibility: By converting Sanskrit texts into a digital format, the OCR project facilitates easy and widespread access to Sanskrit literature. Scholars, students, and enthusiasts can access these texts, overcoming geographical and logistical barriers.

Research Advancement: The digital transformation of Sanskrit texts empowers researchers to explore, analyze, and compare a vast array of texts efficiently. This is a game-changer for scholars engaged in Sanskrit studies, linguistics, philosophy, history, and related fields.

Translation and Cross-Cultural Understanding: Accessible Sanskrit texts foster translation efforts, enabling a broader understanding of ancient Indian knowledge. It contributes to cross-cultural dialogue and the enrichment of global knowledge.

Educational Resources: The OCR-processed Sanskrit texts can serve as invaluable educational resources, enhancing the teaching and learning of Sanskrit and related disciplines worldwide.

1.3.Scope:

The scope of a Sanskrit OCR (Optical Character Recognition) project is extensive and far-reaching, encompassing various domains and potential applications. Here's an overview of the scope of such a project:

Text Digitization: The primary scope involves the digitization of Sanskrit texts. This includes transcribing handwritten or printed manuscripts into machine-readable digital formats. It covers a wide range of Sanskrit texts, from ancient scriptures and classical literature to historical documents and inscriptions.

Character Recognition: The project aims to recognize and accurately transcribe the complex Sanskrit script, including its unique characters, diacritics, ligatures, and compound characters. The scope extends to capturing the nuances of Sanskrit phonetics and morphology, which can be challenging.

Language Variants: Sanskrit has evolved over time, resulting in numerous regional and temporal variants. The OCR system should be designed to accommodate these variations, thereby expanding its scope to cover texts from different periods and regions.

Multilingual Support: Given the influence of Sanskrit on the development of several Indian languages, the OCR project can be extended to recognize and transcribe texts written in other Indian languages like Hindi, Bengali, or Telugu, which share script elements with Sanskrit.

Handwriting Recognition: In addition to printed texts, the OCR project can encompass the recognition of handwritten Sanskrit scripts, which are often found in manuscripts and historical documents. This adds an extra layer of complexity to the project.

Searchable Databases: The scope includes the creation of searchable databases of digitized Sanskrit texts. Users should be able to efficiently search, retrieve, and explore the texts, making them accessible for researchers, scholars, students, and the general public.

Translation and Lexical Analysis: The digitized texts can be used for translation into various languages and for linguistic analysis. The project's scope extends to supporting tools for translation and lexical research.

Educational Resources: The digitized Sanskrit texts can be valuable educational resources, making the scope of the project relevant to educational institutions and Sanskrit language learners.

Cultural Heritage Preservation: By preserving Sanskrit texts in a digital format, the project contributes to the preservation of cultural and historical heritage. This scope includes protecting these invaluable manuscripts from deterioration or loss.

Technological Advancement: The project also has the potential to advance OCR technology itself. Developing algorithms capable of recognizing complex scripts like Sanskrit can benefit OCR research beyond Sanskrit and open up opportunities for improved character recognition in other languages.

Collaboration and Community Engagement: Collaboration with scholars, linguists, and Sanskrit enthusiasts is within the scope of the project. It can also engage the Sanskrit community in contributing to the project's success.

User-Friendly Interfaces: Creating user-friendly interfaces to access and interact with the digitized Sanskrit texts is a vital aspect of the project's scope. These interfaces should cater to diverse user groups, from researchers to casual readers.

In summary, the scope of a Sanskrit OCR project is vast and diverse, encompassing the recognition, preservation, and accessibility of Sanskrit texts, as well as potential applications in education, research, cultural preservation, and technology development

Chapter 2

Problem definition

The problem definition of a Sanskrit OCR (Optical Character Recognition) project revolves around overcoming a series of intricate challenges inherent to recognizing and digitizing Sanskrit characters and texts. The foremost issue lies in the complexity of the Sanskrit script, which incorporates a vast array of characters, diacritics, ligatures, and compound symbols. Accurately differentiating and transcribing these elements forms the crux of the problem. Additionally, variations in handwriting and script styles across different time periods and regions present a substantial challenge, as the OCR system must accommodate this diversity for precise recognition. Furthermore, the profound phonetic and morphological intricacies of Sanskrit necessitate a deep understanding to capture linguistic nuances correctly. The need to develop user-friendly interfaces for interaction with the OCR system, accommodating a diverse user base ranging from researchers to students and enthusiasts. Ensuring the accuracy of OCR results, through verification mechanisms, is paramount to prevent errors in transcription that could impact research and translation efforts. Finally, the system must be adaptable to regional and temporal variations in Sanskrit script, to effectively handle a wide array of historical texts. Staying current with advancements in OCR technology and fostering collaboration with scholars, linguists, and the acquisition of critical resources are integral to successfully addressing these multifaceted challenges and realizing the potential of Sanskrit OCR in preserving and propagating the rich Sanskrit literary and cultural heritage.

Chapter 3

Proposed System

In proposing a system for a Sanskrit OCR (Optical Character Recognition) project, several critical components should be considered. The heart of the system lies in the character recognition and digitization process. The system must employ state-of-the-art machine learning and artificial intelligence algorithms to accurately recognize Sanskrit characters, diacritics, ligatures, and compound characters. It should also integrate a comprehensive database of Sanskrit characters and linguistic rules to ensure the accurate transcription of the text.

The proposed system should be adaptable to variations in Sanskrit script, accounting for differences in handwriting styles and regional or temporal variations. To achieve this, the system can incorporate machine learning models that can be trained on diverse datasets, encompassing a wide range of script styles and historical periods.

Furthermore, the system should include an effective user interface that allows for easy interaction with the recognized text. It should be user-friendly, catering to both scholars and enthusiasts, offering features for searching, editing, and exporting the digitized text.

3.1. Features

The Sanskrit OCR project is designed to facilitate the seamless conversion of Sanskrit documents into English, with a strong emphasis on accessibility and ease of use. This initiative includes several key features aimed at achieving this goal.

First and foremost, the OCR system excels at recognizing Sanskrit text, whether it's in printed form, handwritten, or embedded within images. This capability broadens its utility significantly, as it can handle a wide variety of source materials, including ancient manuscripts, modern texts, and inscriptions. This inclusivity ensures that users can work with a diverse range of Sanskrit content.

To enhance user experience, the project emphasizes the development of an intuitive and user-friendly interface. This interface is thoughtfully designed to make interacting with the OCR system as straightforward as possible. Users, both experts and beginners, will find it easy to upload their Sanskrit documents or images and initiate the conversion process. The intuitive design encourages more people to explore and utilize the system's capabilities without technical barriers.

Furthermore, the OCR system not only recognizes Sanskrit characters but also provides translations or transliterations into English. This feature is invaluable for those who may not be proficient in Sanskrit, as it allows them to understand the text's meaning and pronunciation. By offering translations or transliterations, the project promotes cross-cultural understanding and knowledge dissemination, making the rich literary and cultural heritage of Sanskrit accessible to a global audience.

In summary, the Sanskrit OCR project encompasses features that ensure it is a powerful, accessible, and user-centric tool for converting Sanskrit documents into English. Its capacity to handle a wide array of source materials, its user-friendly interface, and its translation/transliteration capabilities collectively contribute to bridging the language gap and enhancing the appreciation of Sanskrit literature and knowledge for a broader audience.

Chapter 4

Project Outcomes

The Sanskrit OCR project has yielded a host of valuable outcomes, making it a pivotal endeavor for the preservation and dissemination of Sanskrit literature and heritage. Firstly, the project's ability to digitize Sanskrit texts, ranging from ancient manuscripts to modern printed documents, has revolutionized access to this rich corpus of knowledge. The digitization process ensures that these texts are preserved in a digital format, safeguarding them against the ravages of time.

Moreover, the OCR project has significantly enhanced the accessibility of Sanskrit content. It has enabled a global audience, including scholars, researchers, students, and enthusiasts, to engage with Sanskrit literature even if they lack proficiency in the language. This accessibility is a cornerstone of democratizing knowledge and fostering cross-cultural exchange.

The project's capacity to recognize handwritten Sanskrit text marks a remarkable milestone, as it unlocks the potential for transcribing and translating historical manuscripts that were previously challenging to decipher. This not only aids in unraveling the mysteries of ancient Sanskrit writings but also opens up new avenues for research.

The OCR system's cross-linguistic understanding feature, which translates or transliterates Sanskrit into English and other languages, promotes wider comprehension of Sanskrit texts, enhancing knowledge dissemination and intercultural dialogue.

In the realm of scholarship, the digitized Sanskrit texts serve as invaluable resources for researchers in diverse fields such as linguistics, history, literature, philosophy, and cultural studies. These texts provide a fertile ground for analysis and exploration.

The project's role in cultural heritage preservation cannot be overstated. By safeguarding Sanskrit texts from deterioration due to aging or environmental factors, it ensures that this ancient and culturally significant heritage remains intact for future generations.

Chapter 5

Software Requirements

1. Easy OCR
2. Translators
3. Python

Easy OCR

Easy OCR is a font-dependent printed character reader based on a template matching algorithm. It has been designed to read any kind of short text (part numbers, serial numbers, expiry dates, manufacturing dates, lot codes, …) printed on labels or directly on parts.

Comprehensive character verification with automatic training

Grayscale analysis

Text and character-level inspection

Contrast, position and shape defect detection

Translator

A translator in python is a program or script that facilitates the conversion or translation of data, code, or text from one form or language to another, depending on the specific context and purpose for which it's designed.

Chapter 6

Project Design

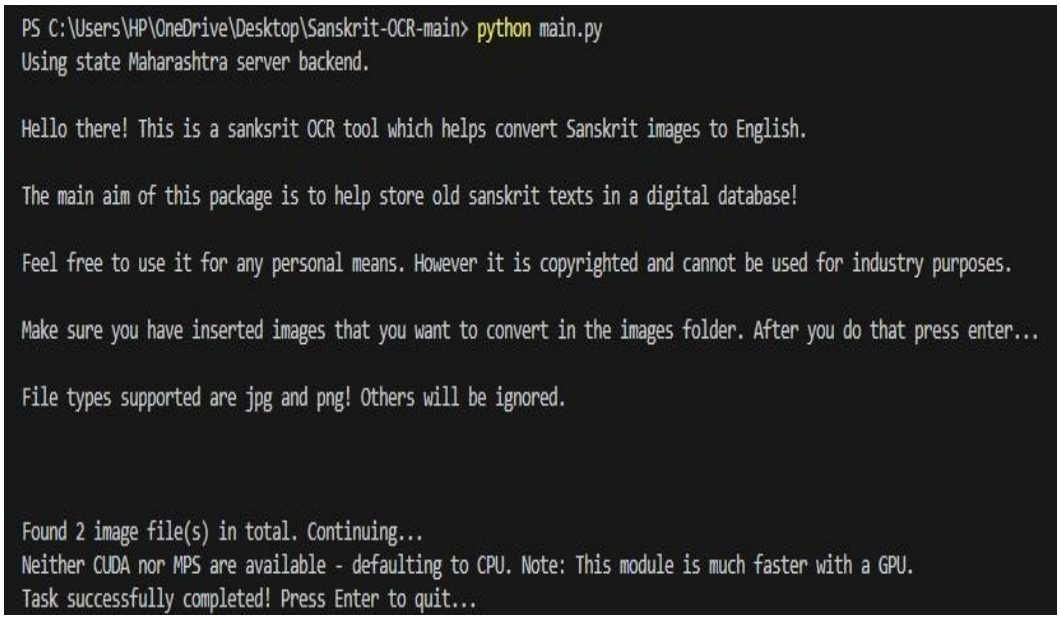

Fig. 1

main output for the project

when it runs using the command "python main.py"

it shows the following result and the translated gets saved to result folder in the form of .txt format

अश्वत्थामा विकर्णश्च सौमदत्तिस्तथैव च ॥ १-८॥ (सौमदत्तिर्जयद्रथः)

अन्ये च बहवः शूरा मदर्थे त्यक्तजीविताः ।
नानाशस्त्रप्रहरणाः सर्वे युद्धविशारदाः ॥ १-९॥

अपर्याप्तं तदस्माकं बलं भीष्माभिरक्षितम् ।
पर्याप्तं त्विदमेतेषां बलं भीमाभिरक्षितम् ॥ १-१०॥

अयनेषु च सर्वेषु यथाभागमवस्थिताः ।
भीष्ममेवाभिरक्षन्तु भवन्तः सर्व एव हि ॥ १-११॥

तस्य सञ्जनयन्हर्षं कुरुवृद्धः पितामहः ।
सिंहनादं विनद्योच्चैः शङ्खं दध्मौ प्रतापवान् ॥ १-१२॥

ततः शङ्खाश्च भेर्यश्च पणवानकगोमुखाः ।
सहसैवाभ्यहन्यन्त स शब्दस्तुमुलोऽभवत् ॥ १-१३॥

ततः श्वेतैर्हयैर्युक्ते महति स्यन्दने स्थितौ ।
माधवः पाण्डवश्चैव दिव्यौ शङ्खौ प्रदध्मतुः ॥ १-१४॥

पाञ्चजन्यं हृषीकेशो देवदत्तं धनञ्जयः ।
पौण्ड्रं दध्मौ महाशङ्खं भीमकर्मा वृकोदरः ॥ १-१५॥

अनन्तविजयं राजा कुन्तीपुत्रो युधिष्ठिरः ।
नकुलः सहदेवश्च सुघोषमणिपुष्पकौ ॥ १-१६॥

काश्यश्च परमेष्वासः शिखण्डी च महारथः ।
धृष्टद्युम्नो विराटश्च सात्यकिश्चापराजितः ॥ १-१७॥

द्रुपदो द्रौपदेयाश्च सर्वशः पृथिवीपते ।
सौभद्रश्च महाबाहुः शङ्खान्दध्मुः पृथक्पृथक् ॥ १-१८॥

स घोषो धार्तराष्ट्राणां हृदयानि व्यदारयत् ।
नभश्च पृथिवीं चैव तुमुलोऽभ्यनुनादयन् ॥ १-१९॥ orलो व्यनु

अथ व्यवस्थितान्दृष्ट्वा धार्तराष्ट्रान् कपिध्वजः ।
प्रवृत्ते शस्त्रसम्पाते धनुरुद्यम्य पाण्डवः ॥ १-२०॥

हृषीकेशं तदा वाक्यमिदमाह महीपते ।

अर्जुन उवाच ।

Fig. 2

Sample input for translation purpose

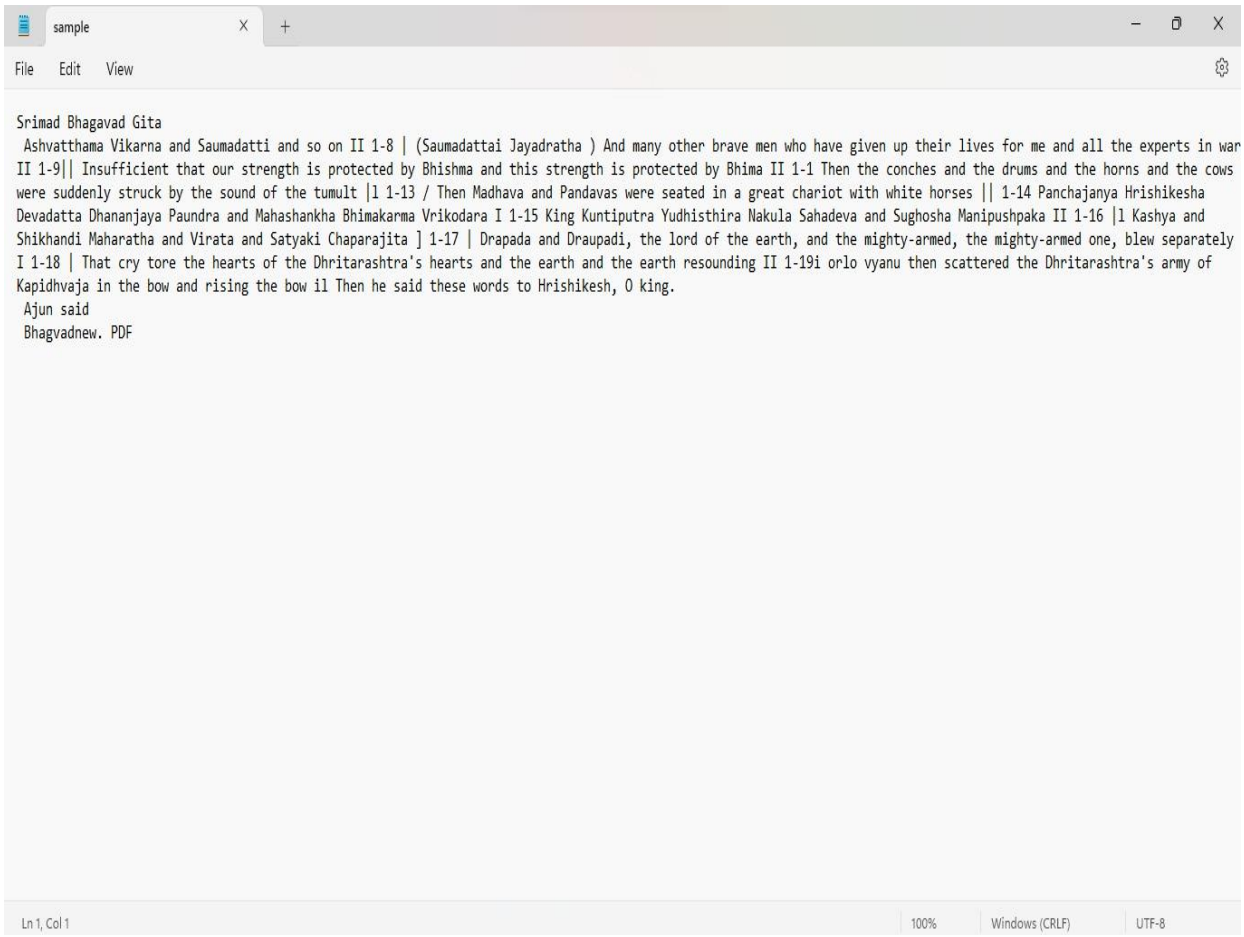

Fig. 3

the output for the following Sample input which is translated in English language and gets saved into the result folder having the same name as the name of the input file (i.e. in .jpg or .png format)

Project Scheduling Template

Sr. No	Group Member	Time duration	Work to be done
1	Krishna Gupta	**August-November**	Implementing the easy ocr for character recognition.
2	Meris Gada	**August-November**	Implementing the easy ocr for character recognition.
3	Tushar Goud	**August-November**	Implementing the translators for converting Sanskrit to English.

Chapter 8

Conclusion

In conclusion, the Sanskrit OCR project is a significant undertaking with far-reaching implications for the preservation, accessibility, and understanding of Sanskrit literature and heritage. This project has not only digitized Sanskrit texts but has also made them accessible to a global audience, transcending language barriers. The project's ability to recognize handwritten Sanskrit, translate it into other languages, and promote cross-cultural understanding underscores its significance in the modern world.

Furthermore, the OCR project has opened doors for extensive research, offering scholars and researchers a wealth of resources for linguistic, historical, literary, and philosophical studies. It plays a pivotal role in preserving the cultural heritage of Sanskrit, ensuring that these invaluable texts remain intact for generations to come.

By becoming educational resources and engaging the community, the project is not only a testament to technological progress but also a bridge between the past and the present. In this way, the Sanskrit OCR project stands as a testament to the enduring value of Sanskrit literature and its continued relevance in our interconnected world.

The Sanskrit OCR project is a testament to the harmonious intersection of technology, culture, and knowledge. Its impact reaches far and wide, breathing new life into an ancient language and heritage, fostering accessibility, promoting cross-cultural understanding, and ensuring that the profound wisdom of Sanskrit endures for generations to come.

References

[1] http://unicode.org/charts/PDF/U0900.pdf.

[2] https://sanskritlibrary.org/software/ocr.html.

[3] Sub-word Embeddings for OCR Corrections in highly Fusional Indic Languages [source code, video of demo system]. Rohit Saluja, Mayur Punjabi, Mark Carman, Ganesh Ramakrishnan and Parag Chaudhuri.In Proceedings of The 15th International Conference on Document Analysis and Recognition (ICDAR 2019), Sydney, Australia.

[4] OCR On-the-Go: Robust End-to-end Systems for Reading License Plates and Street Signs [source code, video of demo system]. Rohit Saluja, Ayush Maheshwari, Ganesh Ramakrishnan, Parag Chaudhuri and Mark Carman. In Proceedings of The 15th International Conference on Document Analysis and Recognition (ICDAR 2019), Sydney, Australia.

[5] Learning From Less Data: Diversified Subset Selection and Active Learning in Image Classification Tasks. Vishal Kaushal, Rishabh Iyer, Anurag Sahoo, Khoshrav Doctor, Narasimha Raju, Ganesh Ramakrishnan. Accepted paper at the 7th IEEE Winter Conference on Applications of Computer Vision (WACV), 2019, Hawaii, USA.

ACKNOWLEDGEMENT

This project would not have come to fruition without the invaluable help of our Guide **Prof.Vaibhav Yavalkar**. Expressing gratitude towards our HoD, **Prof.Anagha Aher**, and the Department of CSE(Data Science) for providing us with the opportunity as well as the support required to pursue this project. We would also like to thank our teacher Ms. Poonam Pangarkar who gave us her valuable suggestions and ideas when we were in need of them. We would also like to thank our peers for their helpful suggestions.