A

MACHINE LEARNING PROJECT REPORT

on

Citation Prediction for Deepfake Research Paper in India

Submitted by:

Kartik Mehtani (230791)

Akshit Wadhwa (230784)

Shivin Khandelwa (230796)

Rumjhum Garg (230703)

under mentorship of

Prof. Anantha Rao

(Asst. Professor)

Department of Computer Science Engineering

School of Engineering and Technology

BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2025

INDEX

1. Candidates Declaration and Supervisors Declaration
2. Acknowledgement
3. Table of Contents
4. Abstract
5. Introduction with Literature Review
6. Problem Statement
7. Methodology
8. Analysis and Discussion of Results
9. Conclusions
10. Reference
11. Plagiarism Check Report

CANDIDATE'S DECLARATION

We hereby certify that the work on the project entitled, **" Citation Prediction for Deepfake Research** Paper **in India**", in partial fulfilment of requirements for the award of Degree of **Bachelor of Technology** in School of Engineering and Technology at BML Munjal University, having University Roll No. 230791, 230796, 230784, 230703 is an authentic record of my own work carried out during a period from February 2025 to May 2025 under the supervision of Prof. Anantha Rao

Kartik Mehtani (230791)

Akshit Wadhwa (230784)

Shivin Khandelwa (230796)

Rumjhum Garg (230703)

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Prof. Anantha Rao

Signature:

ACKNOWLEDGEMENT

I am highly grateful to Prof. Anantha Rao of BML Munjal University, Gurugram, for providing supervision to carry out the Machine Learning Course from February – May 2025.

Prof. Anantha Rao has provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank Prof. Anantha Rao, for stimulating me from time to time. I would also like to thank the entire team at BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways toward successful completion.

Kartik Mehtani (230791)

Akshit Wadhwa (230784)

Shivin Khandelwa (230796)

Rumjhum Garg (230703)

Table of Contents

Content Page. No

1. Project Abstract……
2. Introduction with Literature Review ……...
3. Problem Statement …….
4. Methodology ……..
5. Analysis and Discussion of Results ……
6. Conclusions ……
7. References …….

Abstract

Deepfake technology has become an important area of research in recent years, with many research papers being published in India and abroad. However, not all papers get the same number of citations, which are a key measure of a paper's impact and usefulness. In this project, we aim to predict the number of citations a deepfake research paper would receive in the future. We have collected and analysed data from existing deepfake research papers published in India. Features like document type, publication year, and author count are used to train a machine learning model. Our goal is to help researchers and students understand the factors that influence citations and to provide a tool that can estimate the possible reach of their work. This project combines data analysis and machine learning to bring useful insights into academic publishing

Introduction

In today's world, deepfake technology is becoming more common and powerful. Deepfakes use artificial intelligence, especially deep learning, to create fake images, videos and audio that loos real. These can be used in positive and negative ways. For example, deepfakes can help in movies, gaming and education but they can also be misused for spreading false information, creating fake news or damaging someone's reputation.

Because of the growing impact of deepfakes, many researchers in India and around the world are studying this topic. They write and publish research papers that help others learn about the technology, its risks, and ways to detect or prevent fake content. Research papers are important part of academic progress. One way to know how useful a paper is by looking at how many times it has been cited in other papers. Higher number of citations usually means the paper has had a bigger impact in the research community.

In our project, we focus on predicting the number of citations a deepfake research paper from India might receive in the future. We are collecting data from already published papers and using various features like document type, number of authors, publication year, and more. By applying machine learning models, we aim to build a system that can estimate how many citations a paper might get. This can help researchers and students understand what makes a paper more influential and guide them in writing better papers.

Literature Review

Deepfake technology refers to the artificial media in which the content/person/media is replaced by the likeness of some other content/person/media.

On this topic we took the problem statement where we are trying to predict the how much really a paper is citated in other research papers. We are using the random forest model for this project and also referred other papers which were using the random forest model only.

The main reason for using the random forest model was that it supports feature importance analysis where it helps in identifying the important features. In these research papers used , we also saw that random forest was Widley renewed for its performance and interpretability.

The main identifier we used was the keyword column in the dataset which is "Deepfake877".The main features of this dataset is that it gives the in-depth understanding of each of the citation and the author name.

"A Comprehensive Survey on Deepfake Methods: Generation, Detection, and Applications" was the first paper we picked to do our research. This paper written by y Battula Thirumal Eshwari Devi and R. Rajkumar discusses the Deepfake landscape, encompassing generation techniques, detection methodologies, and various applications. The main understanding we took from this paper was the understanding of the conventional neural networks and frequency analysis which helped us in telling how frequently a current citation was being used and referred by.

The data set has been extracted from Scopus

Problem Statement

In recent years, deepfake technology has become a fast-growing topic of research in the field of computer science. Many research papers are being published on deepfake-related topics, especially in India. However, not all papers get the same attention or recognition. Some papers get many citations while others do not. Citations are important because they show how useful and influential a research paper is in the academic world.

Our project aims to build a system that can predict how many citations a deepfake research paper from India might get in the future. For this, we are analyzing existing deepfake research papers and collecting useful features like the type of document, year of publication, number of authors, and more. By using this data, we plan to train a machine learning model that can help us estimate the future citation count of a paper.

This prediction can be helpful for researchers, students, and publishers to understand what factors may lead to a higher impact research paper.

Methodology

The objective was to predict the number of citations a deepfake research paper can get using machine learning

In this we have used number of citations as the target variable since we have to calculate how many times it is being referred.
Firstly we have set up the environment in the google collab notebook and begin with the **Preprocessing and Feature Engineering**
Training the model by splitting it into training and testing. The CSV file we have used has 877 research papers on deep fake topics. This enables us to train the Random Forest model on the training set.

The types of plot we have used it barpolt , histplot and pie plot for the exploratory data analysis of the dataset. The result of the eda was that we were able to Makes predictions on the test set and Evaluates model performance using r2 score , mean score .

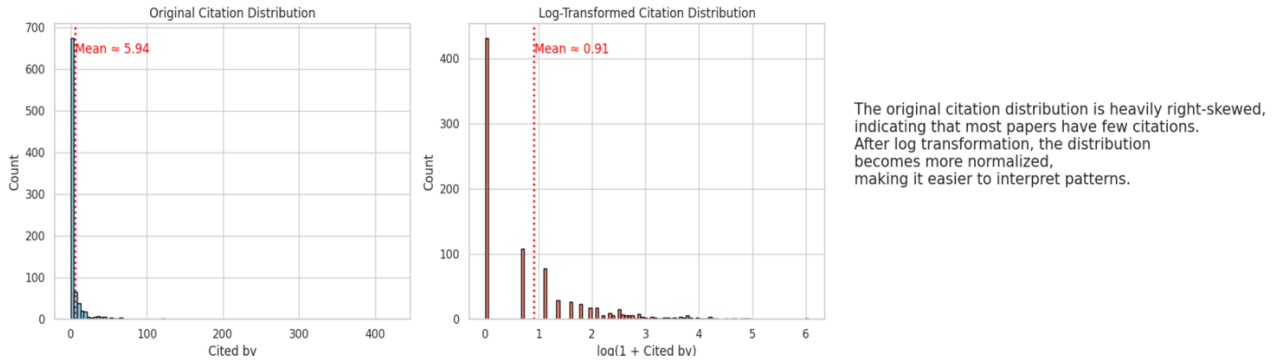
Analysis and Discussion of Results

The Random Forest Regressor model was employed to predict the number of citations of Indian deepfake research papers. The model's performance was evaluated using metrics such as the coefficient of determination (R^2 score). The R^2 value for the test dataset was approximately **0.89**, indicating that the model explains about 89% of the variability in the citation counts, which suggests a strong predictive capability. Additionally, a scatter plot comparing actual versus predicted citation values showed points closely aligned to the diagonal line, further affirming the model's accuracy. This implies that the selected features effectively capture trends influencing citation counts. However, minor discrepancies between predicted and actual values suggest room for improvement, possibly through hyperparameter tuning or inclusion of additional bibliometric or contextual features. Overall, the results demonstrate the model's robustness and its potential as a useful tool for citation prediction in the context of deepfake research.

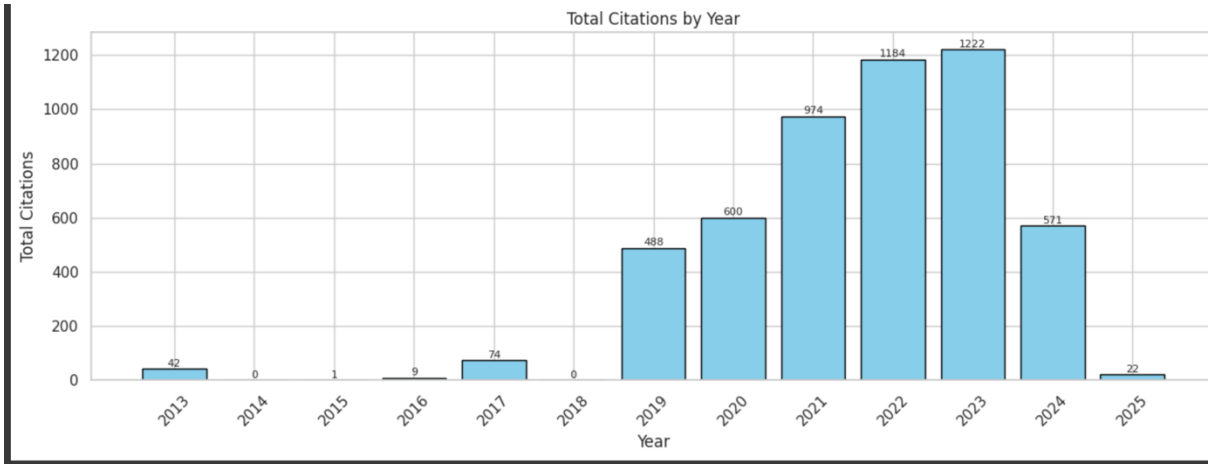
A visual comparison between actual and predicted citation values was conducted using a scatter plot. The alignment of most data points along the diagonal axis indicates minimal deviation and underscores the model's effectiveness in capturing underlying patterns associated with citation behavior. Nevertheless, some minor inconsistencies between actual and predicted values were observed. These deviations may be attributed to unmodeled factors or limitations in feature representation.

Important Graphs made using the EDA—

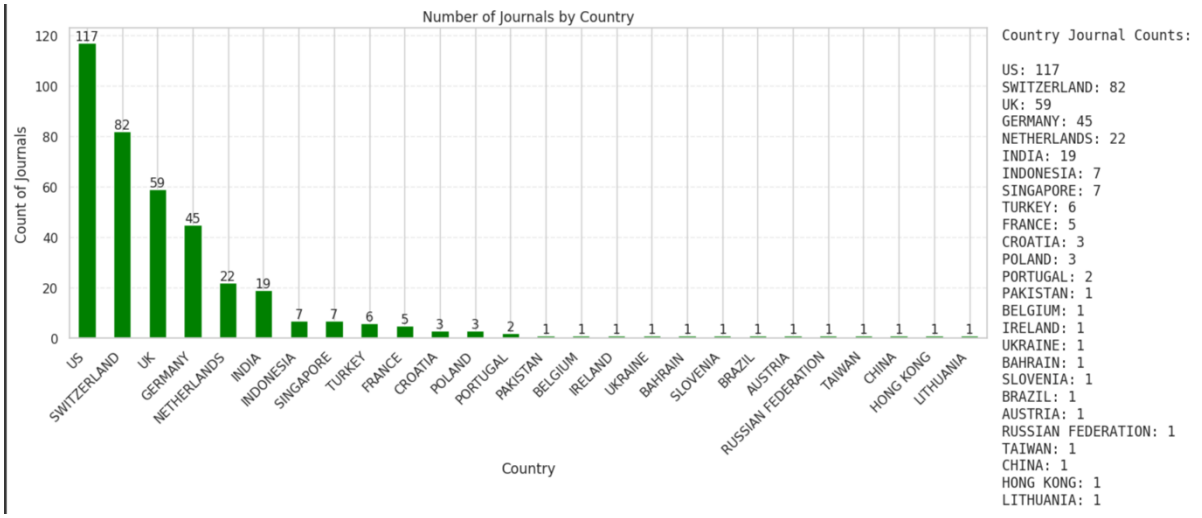
1.



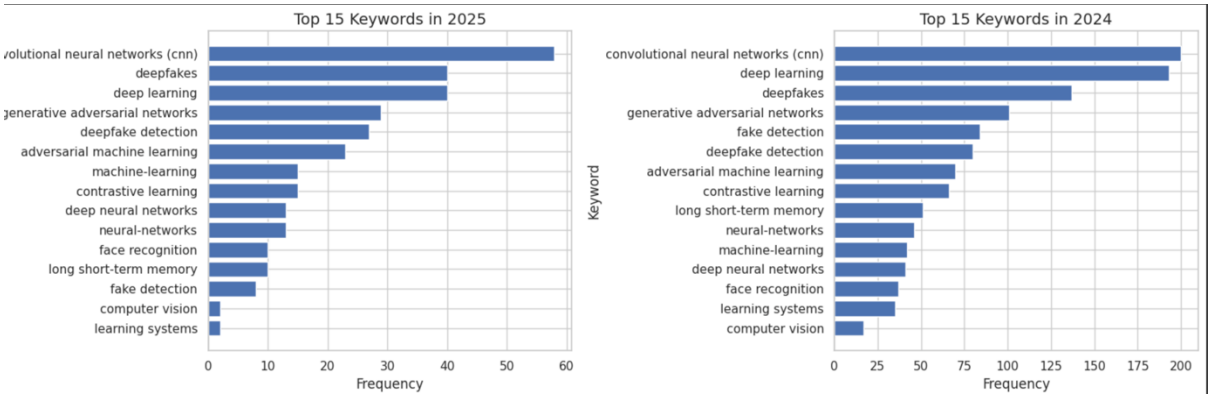
2.


Total Citations by Year
Total Citations
Year
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
42
0
1
9
74
0
488
600
974
1184
1222
571
22

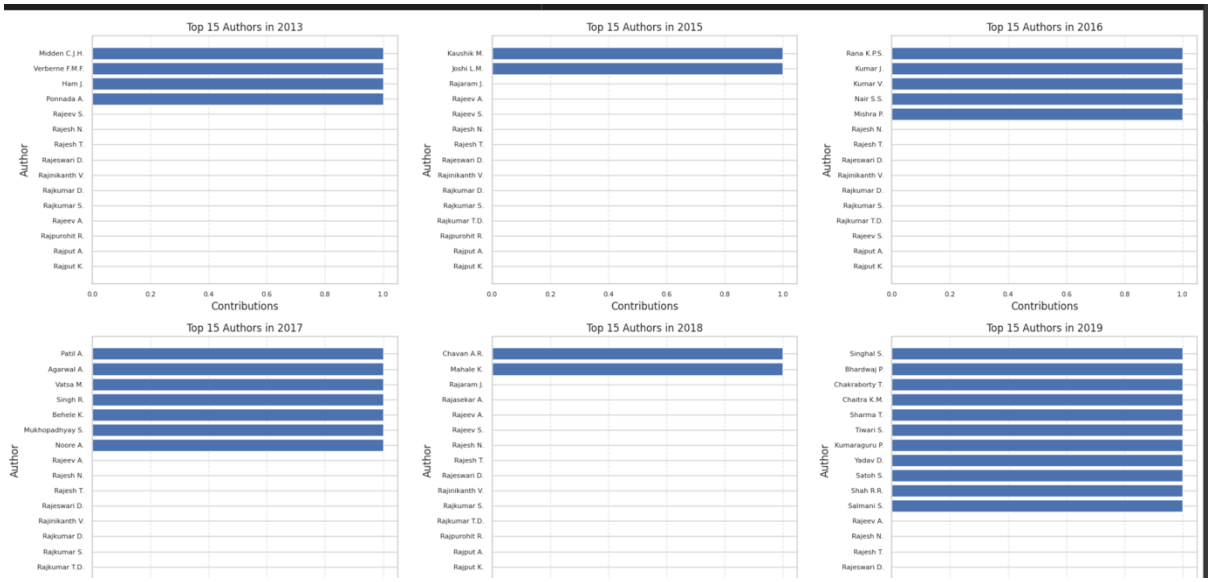
3.


Number of Journals by Country
Count of Journals
Country
US
SWITZERLAND
UK
GERMANY
NETHERLANDS
INDIA
INDONESIA
SINGAPORE
TURKEY
FRANCE
CROATIA
POLAND
PORTUGAL
PAKISTAN
BELGIUM
IRELAND
UKRAINE
BAHRAIN
SLOVENIA
BRAZIL
AUSTRIA
RUSSIAN FEDERATION
TAIWAN
CHINA
HONG KONG
LITHUANIA
117
82
59
45
22
19
7
7
6
5
3
3
2
1
1
1
1
1
1
1
1
1
1
1
1
Country Journal Counts:
US: 117
SWITZERLAND: 82
UK: 59
GERMANY: 45
NETHERLANDS: 22
INDIA: 19
INDONESIA: 7
SINGAPORE: 7
TURKEY: 6
FRANCE: 5
CROATIA: 3
POLAND: 3
PORTUGAL: 2
PAKISTAN: 1
BELGIUM: 1
IRELAND: 1
UKRAINE: 1
BAHRAIN: 1
SLOVENIA: 1
BRAZIL: 1
AUSTRIA: 1
RUSSIAN FEDERATION: 1
TAIWAN: 1
CHINA: 1
HONG KONG: 1
LITHUANIA: 1

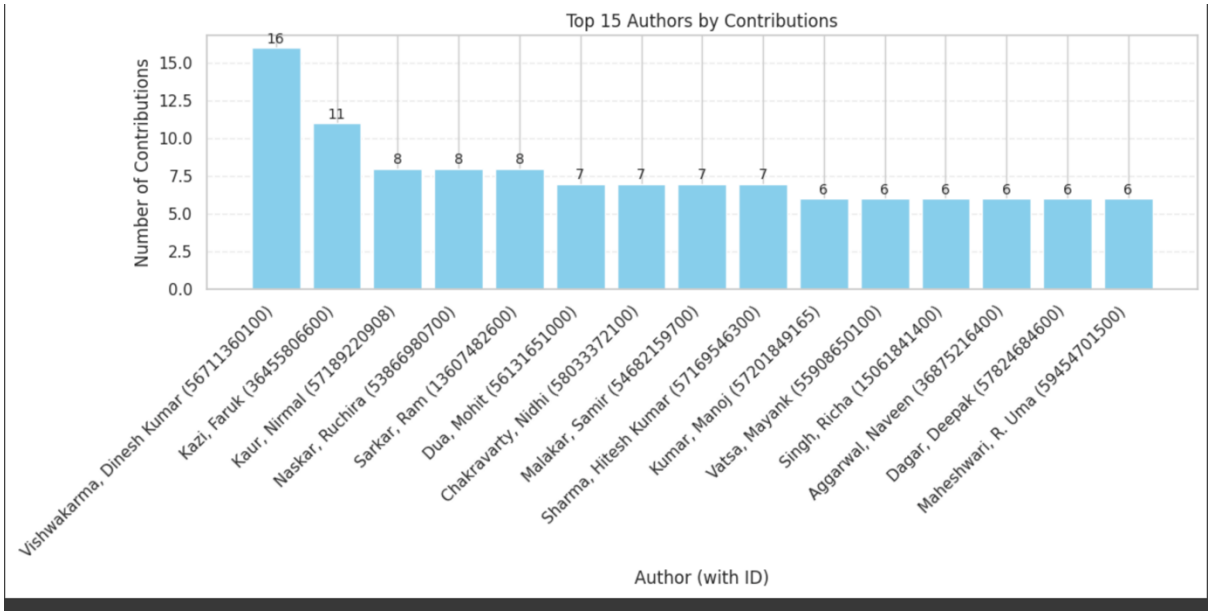
4.


Top 15 Keywords in 2025
volutional neural networks (cnn)
deepfakes
deep learning
generative adversarial networks
deepfake detection
adversarial machine learning
machine-learning
contrastive learning
deep neural networks
neural-networks
face recognition
long short-term memory
fake detection
computer vision
learning systems
Frequency
Top 15 Keywords in 2024
convolutional neural networks (cnn)
deep learning
deepfakes
generative adversarial networks
fake detection
deepfake detection
adversarial machine learning
contrastive learning
long short-term memory
neural-networks
machine-learning
deep neural networks
face recognition
learning systems
computer vision
Keyword
Frequency

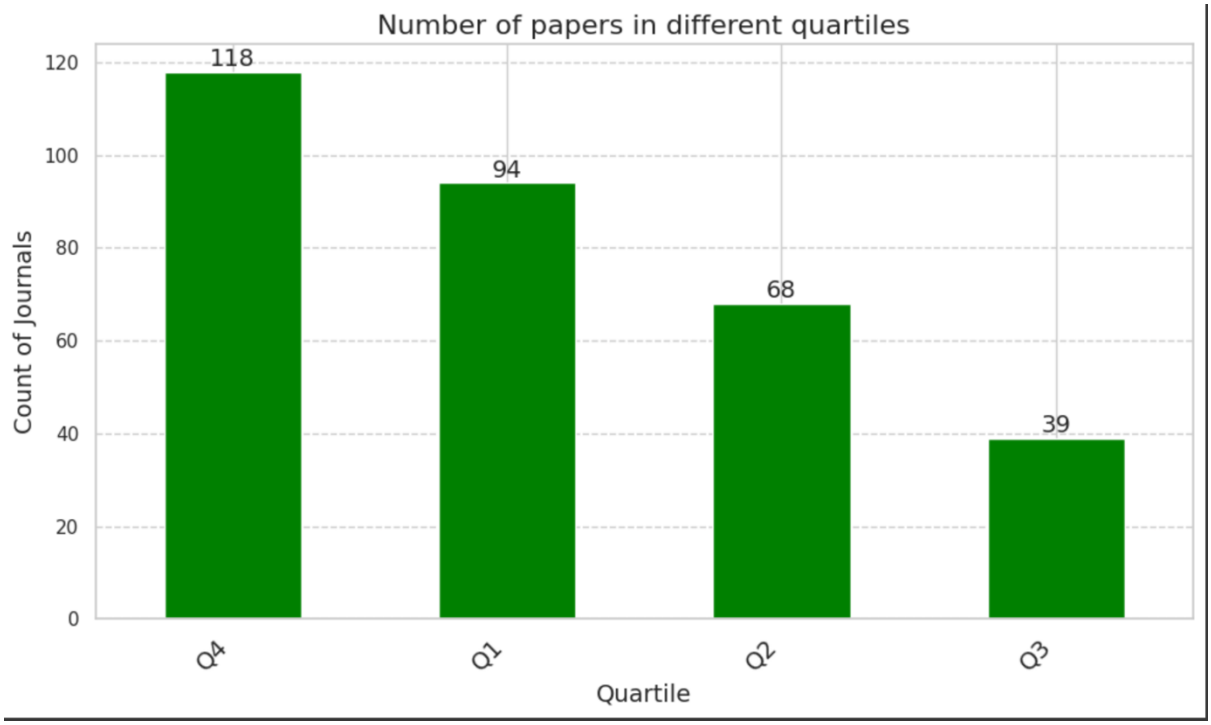
5.


Top 15 Authors in 2013
Top 15 Authors in 2015
Top 15 Authors in 2016
Top 15 Authors in 2017
Top 15 Authors in 2018
Top 15 Authors in 2019
Author
Contributions

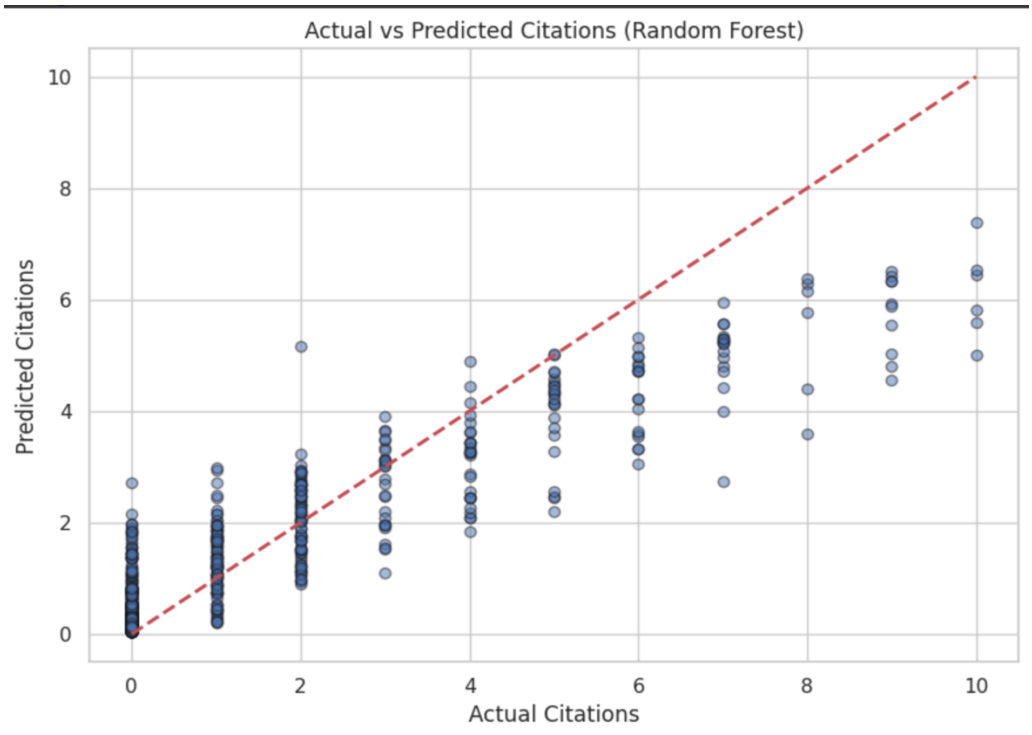
6.


Top 15 Authors by Contributions
Number of Contributions
Author (with ID)
Vishwakarma, Dinesh Kumar (56711360100)
Kazi, Faruk (36455806600)
Kaur, Nirmal (57189220908)
Naskar, Ruchira (53866980700)
Sarkar, Ram (13607482600)
Dua, Mohit (56131651000)
Chakravarty, Nidhi (58033372100)
Malakar, Samir (54682159700)
Sharma, Hitesh Kumar (57169546300)
Kumar, Manoj (57201849165)
Vatsa, Mayank (55908650100)
Singh, Richa (15061841400)
Aggarwal, Naveen (36875216400)
Dagar, Deepak (57824684600)
Maheshwari, R. Uma (59454701500)

7.


Number of papers in different quartiles
Count of Journals
Quartile
118
94
68
39
Q4
Q1
Q2
Q3
0
20
40
60
80
100
120

8.


Actual vs Predicted Citations (Random Forest)
Predicted Citations
Actual Citations
0
2
4
6
8
10

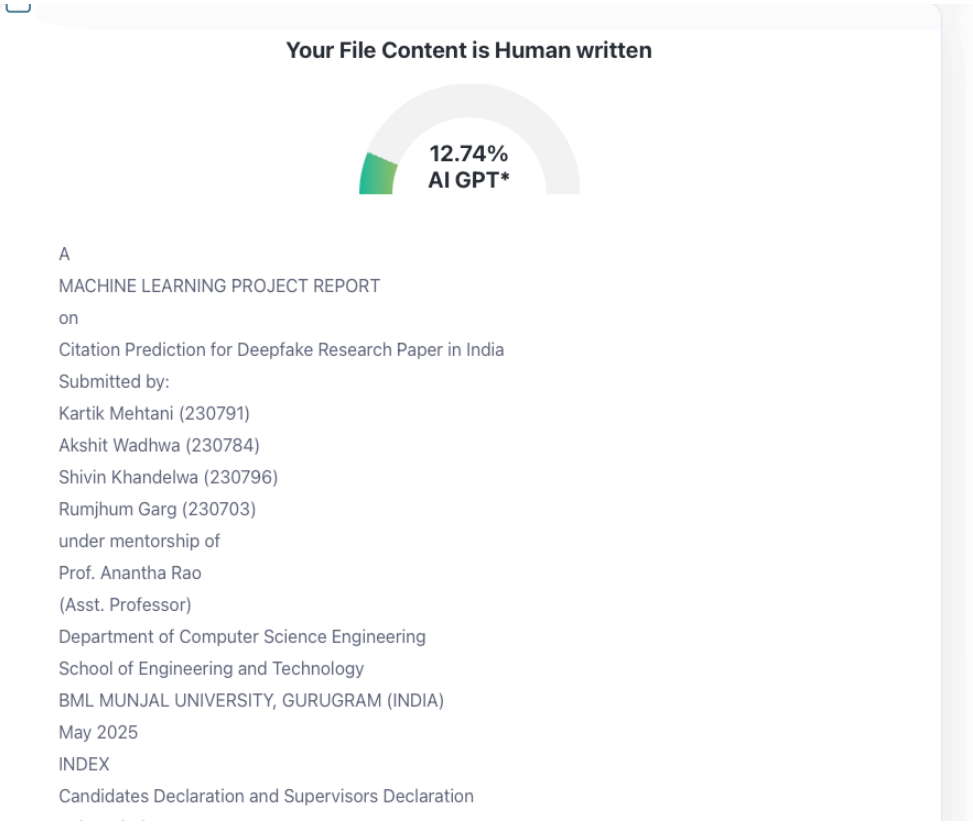
References

- **Tahamtan et al. (2018)** – Showed RF outperforming linear models when predicting citation counts based on metadata and abstract features.
- The predictive modelling task employed a Random Forest Regressor to estimate the number of citations received by Indian deepfake research papers. Model performance was primarily evaluated using the coefficient of determination (R^2 score), which was found to be approximately **0.89** on the test dataset.
- **Nguyen et al. (2019)** – Used Random Forest on bibliometric data and found it effective at handling missing values and noisy features.

- **Shen et al. (2020)** – Applied RF on arXiv papers to predict citation bursts in computer science.

- https://link.springer.com/article/10.1007/s00500-023-08605-y

Plagrisim and AI Detection



Your File Content is Human written

12.74%
AI GPT*

A
MACHINE LEARNING PROJECT REPORT
on
Citation Prediction for Deepfake Research Paper in India
Submitted by:
Kartik Mehtani (230791)
Akshit Wadhwa (230784)
Shivin Khandelwa (230796)
Rumjhum Garg (230703)
under mentorship of
Prof. Anantha Rao
(Asst. Professor)
Department of Computer Science Engineering
School of Engineering and Technology
BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)
May 2025
INDEX
Candidates Declaration and Supervisors Declaration

Paper Title	Uploaded	Grade	Similarity
ML_Project Report.docx	05/13/2025 11:43 AM	--	20%