PROJECT REPORT ON

ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

SUBMITTED IN PARTIAL REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE

SUBMITTED BY

S SOWMYAMITHRA	ANUSHA K	SHULOMITHI T S
17011A0533	17011A0504	17011A0532

UNDER THE SUPERVISION OF

Dr K P SUPREETHI
PROFESSOR OF CSE , JNTUHCEH

DEPARTMENT OF COMPUTER SCIENCE
JNTUH COLLEGE OF ENGINEERING
KUKATPALLY , HYDERABAD
TELANGANA
2020 - 2021

J.N.T.U.H. COLLEGE OF ENGINEERING
KUKATPALLY, HYDERABAD – 500 085

CERTIFICATE

This is to certify that the thesis entitled "***ONLINE ASSIGNMENT PLAGIARISM CHECKER FOR IMAGES***" submitted to JNTUH college of engineering , Hyderabad by *S SOWMYAMITHRA , ANUSHA K , SHULOMITHI TS* for the award of *BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE* is a record of bonafide research work carried by them. The results are verified and found to be satisfactory.

Supervisor

Dr K P SUPREETHI

PROFESSOR OF CSE

JNTUHCEH , HYDERABAD

Head of the Department

Dr M CHANDRA MOHAN

PROFESSOR OF CSE & HOD

JNTUHCEH , HYDERABAD

DECLARATION

We hereby declare that the project work entitled "***ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES***" submitted to JNTUH college of engineering , Hyderabad is a record of an original work carried out by *S SOWMYAMITHRA , ANUSHA K , SHULOMITHI TS* under the guidance of *Dr K P SUPREETHI, PROFESSOR OF CSE, JNTUHCEH*.This report is submitted in partial fulfilment of requirement for the award of BACHELOR OF TECHNOLOGY . We declare that the work reported in this report has not been submitted and will not be submitted to any university or institution for the award of degree.

SIGNATURE

SOWMYAMITHRA [17011A0533]

ANUSHA K [17011A0504]

SHULOMITHI T S [17011A0532]

ACKNOWLEDGEMENT

We would like to express my gratitude to our project supervisor *Dr K P SUPREETHI , PROFESSOR OF CSE, JNTUHCEH*. For her guidance and constant encouragement throughout the course of the dissertation work and I also thank *Dr M. CHANDRA MOHAN , Professor of CSE & Head of the Department , JNTUHCEH* for his valuable support and motivation regarding project work.I also thank all the faculty of the DEPARTMENT OF COMPUTER SCIENCE for their valuable suggestion at all our approaches.

DATE

PLACE KUKATPALLY ,
HYDERABAD

SIGNATURE

S SOWMYAMITHRA
[17011A0533]

ANUSHA K
[17011A0504]

SHULOMITHI T S
[17011A0532]

INDEX

ABSTRACT

Plagiarism is defined as a copy of someone's work and presenting it as one's own work . Plagiarism affects the education quality of the students and thereby reduces the economic status of the country.Plagiarism is done by Copying an image from a book or the internet without citing the original source.

Recently, the problem of plagiarism has become an important issue in the field of Education and Technology. The wide use and availability of electronic resources makes it easy for students, authors and even academic people to access and use any piece of information and embed it into his/ her own work without proper citation. The problem is raising in an exponential manner that puts the education process under threat.Preventing digital plagiarism requires an enormous amount of work from educators.The project we concentrate on implementation of a well known anti-plagiarism algorithm for local and global search for the original source of plagiarized images submitted in assignments.

Data mining is the field which can help in detecting plagiarism . Here we used an Reverse Image Processing approach. The approach consists of three main parts: assignment preprocessing, calculate and compare hash values , and visualization. Hash Values can be calculated using different hashing techniques and finally the result is represented in two ways: visual and numerical.In this project , it finds similarity between images by integrating not only a numerical estimation but also using visualization.

LIST OF FIGURES

INTRODUCTION

The proposed system helps in detecting image plagiarism in students assignments using Reverse Image Processing.Reverse Image Lookup is a Content-Based Image Retrieval (CBIR) technology that involves searching for visually similar images on the Database.

Similarly, for a reverse image search, we upload an image on the search engine and the search engine queries its database and matches the image color by color and pixel by pixel to return a list of exactly same or visually similar photos on the internet.It's an effective technique for checking image plagiarism.

PROBLEM DEFINITION -

In recent times , more often, assignments of students are submitted in electronic forms , but it leads towards the easy opportunity of plagiarism.

With the spread of information over the globe, it is very easy to copy the data from different sources and paste it in a single work without giving any acknowledgement to the sources. These actions lead towards lack of learning in students.So there is a need for detecting plagiarism to increase and improve the learning of students . Manual detection of plagiarism is difficult and time consuming.This arises the need to design an automated system to detect plagiarism and also to improve the quality of learning of the student

OBJECTIVES -

1. To develop an interface where students can submit their assignments
2. Collect the submitted assignments , extract images from them
3. Compute hash values and compare them with the images in database to detect image plagiarism in students assignments
4. Display the result in the form of bar chart and save the results in an excel sheet

SCOPE **-**

- ➢ Helps in detecting plagiarism of images in submitted students assignments
- ➢ Also works for images which are different in size , brightness , background , grey scaling but have similar content

SYSTEM REQUIREMENTS **-**

1. PYTHON
2. INBUILT PYTHON LIBRARIES
    a. ImageHash
    b. PyMuPDF Pillow
    c. Opencv
    d. Os-sys
    e. Matplotlib
    f. Xlsxwriter

g. Python-docx

CONCEPT **-**

1. The proposed system helps to detect plagiarism in students assignments using Reverse Image Processing
2. The technique calculates the hash value of images that are extracted from the submitted assignments
3. And compare then with the hash values of the already existing images in the database
4. Finally displays all the images that are plagiarised

LITERATURE REVIEW

Image Plagiarism means act of copying of an image in different ways like copying pictures online, copying images from your Instagram, Facebook, your blog or any online platform and then reuse them somewhere else without properly citing those pictures, often in a way that disregards you as the copyright owner and shows as if the photos were their own. Previously this issue was handled using the method called 'Reverse Image Processing'.

***HOW DOES REVERSE IMAGE PROCESSING WORK*?**

- ➔ Reverse Image Lookup is a Content-Based Image Retrieval (CBIR) technology that involves searching for visually similar images on the internet.
- ➔ When we search for something online, we usually enter keywords/key- phrases in a search engine like Google.
- ➔ Similarly, for a reverse image search, we upload an image on the search engine and the search engine queries its database and matches the image color by color and pixel by pixel to return a list of exactly same or visually similar photos on the internet.
- ➔ It's an effective technique for checking image plagiarism.

IMPORTANT USES OF REVERSE IMAGE SEARCH -

- ➔ Number one, it lets you find out who is using your copyrighted images on the internet
- ➔ Helps you find out the original source of the image – in case you are curious to find out.

BENEFITS OF USING REVERSE IMAGE SEARCH -

- ➔ It helps you find out where your pictures are getting misused or where your copyrighted images are getting used on the internet
- ➔ It also helps you in knowing the source of the picture – in case you want to find out and also helps in finding out the high-resolution images on the internet that can be used for your projects.

LIMITATIONS OF REVERSE IMAGE SEARCH IN CHECKING IMAGE PLAGIARISM -

- ➔ The tools and techniques mentioned above largely rely on web crawlers' ability to index the image files and web pages. If someone does not publish the stolen images online, or in case web crawlers are blocked from crawling a certain web page, it becomes almost impossible to check image plagiarism.

PROPOSED METHOD

The below represented is the block diagram of the proposed method of online assignment plagiarism checker for images.

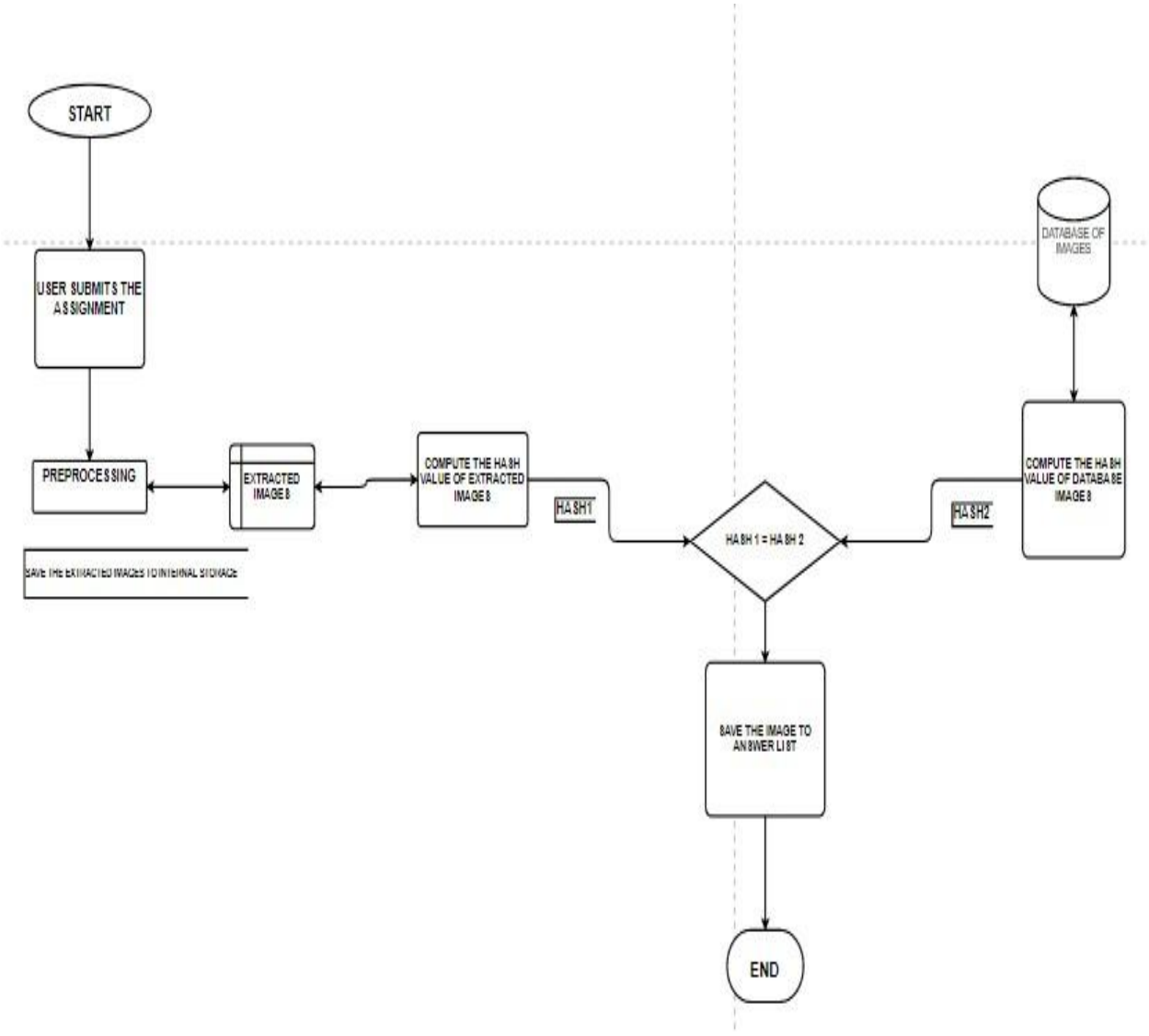

FIG 3.1 BLOCK DIAGRAM OF ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

HOW DOES THE PROPOSED SYSTEM WORK?

STEP 01

- ➢ The user (i.e., the student) submits the assignment with the provided link.
- ➢ The assignment that is being submitted is extracted and saved on local computer

STEP 02

- ➢ Preprocessing is done on the submitted document in which the data is segregated into text and images. All these images get stored in the userdata folder.
- ➢ And also we will be storing some original images in a database folder with which the comparison is done.

STEP 03

- ➢ For every image in the database, a hash value is being generated to it. The generated hash values are used in detecting the plagiarism of images.
- ➢ Now, hash values are also generated for the images that are being submitted in the form of a document by the user.

***HOW HASH VALUES ARE COMPUTED*?**

AHASH -

This algorithm is quite fast but not sensitive to such transformations like scaling of initial image, compressing and stretching, brightness and contrast.It is based on the average value, and, as a result, is sensitive to the operations that change this average value (for instance, change of levels or color balance).

To build aHash one should perform the following steps -

1. Decreasing the image size. The initial image is compressed to some reasonable size (usually it is 8x8 or 16x16 points which will show a 64 or 256 bytes respectively).
2. Image grey-scaling. This move helps to decrease the hash size in three times as it describes the number of components from 3 values of RGB to one level of grey.
3. Computing the average value. Then the average on all the image points is calculated.
4. Simplifying the image. Every pixel gives a value of 0 if it is less than the average value and it gives a value of 1 when its value is greater than average. Thus, the image is converted into the set of the bits. It's read line by line, and the set of values becomes the hash.

STEP 04

- Once the hash values are being calculated, the submitted images hash value is compared with each and every hash value of images in the database.

STEP 05

- Once all the comparisons are done, the pair of images whose hash values are equal are considered as plagiarised
- The obtained data is saved in a result excel sheet and the plagiarised images are saved in a word document.

DESIGN

The following diagrams represent the USE CASE and Activity diagrams -

USE CASE DIAGRAM **-**

- ➢ The first and foremost step in this process is the user has to submit the assignment i.e., the document online.
- ➢ When the assignment file is submitted by the user, the file is being extracted and checked against plagiarism.
- ➢ First, preprocess the file by isolating the text and images which it contains.
- ➢ Once the images are extracted, compute hash values for each of those images.
- ➢ Hash values of images which are in the database are also computed beforehand.
- ➢ Now these hash values of individual images of submitted documents are compared against the values of the images stored in the database.
- ➢ Finally, these comparisons give us the plagiarised percent and the images that are being copied.

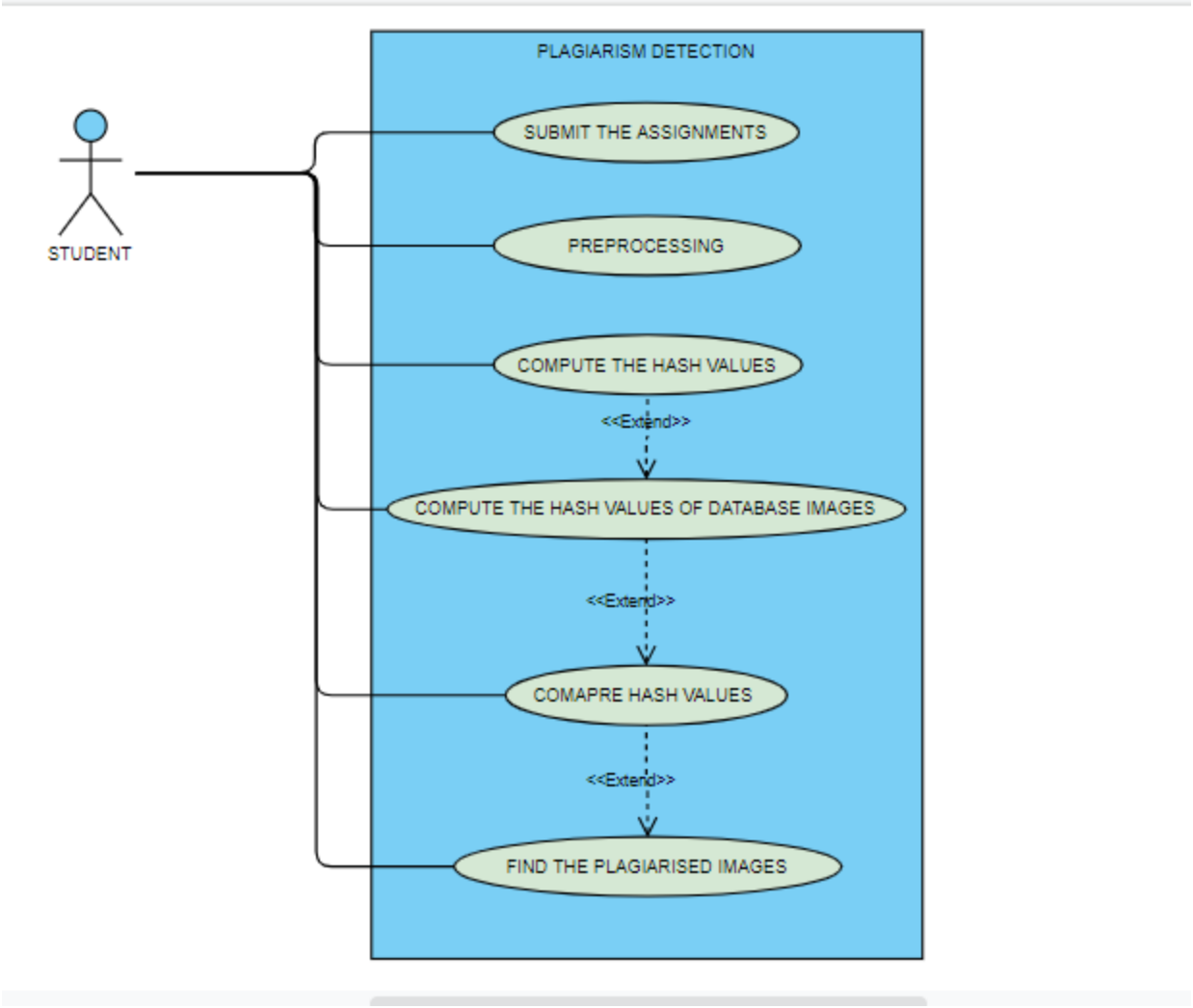

FIG 4.1 USE CASE DIAGRAM OF ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

ACTIVITY DIAGRAM **-**

The above activity diagram describes flow from one activity to another activity. The activity can be described as an operation of the system.

The following is the brief description of the flow of how system works:

1. User Interface: where the user is able to submit his/her assignment online.
2. Preprocessing: The step in which the file undergoes preprocessing. The result we would be getting here is, the content of the file and images get segregated.
3. Hash Value Computation: Hash value is something that it is important in detecting plagiarism.
    - For every image that is found, we compute hash value for it and store it for further purpose.
    - Even for those images that are being stored in the database, hash values are computed.
4. Result: Hash values that are generated from both user images and the database are compared in detecting plagiarism. The final list with the comparisons is generated. In which if i.e.,hash value(user image) - hash value(database image) == 0 it says that the image is copied and the result is printed and also the plagiarised images are being displayed.

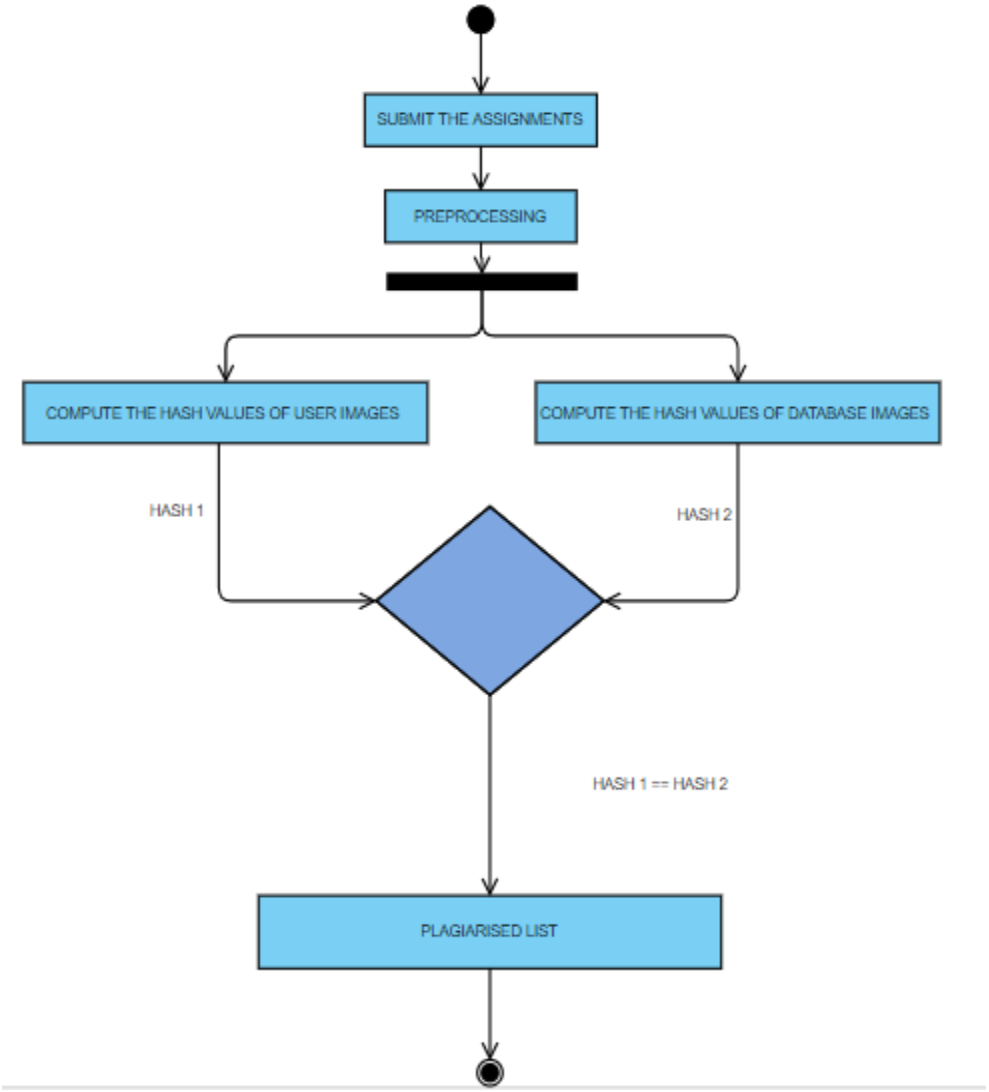

FIG 4.2 ACTIVITY DIAGRAM OF ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

IMPLEMENTATION

This section provides the pseudo code of the functions used in the project -

USER INTERFACE -

➢ A user interface is created that takes file(pdf , docs.) as input and has client side and server side validation

```
<%@ page language="java" %>
<html>
        <head>
                <title> Login form </title>
        </head>
        <body>
                <div align ="center">
                <form ENCTYPE="multipart/form-data" ACTION="Download.jsp" METHOD=POST>
                <table>
                        <tr>
                                <td>Name:</td>
                                <td><input type = "text" name = "name" required></td>
                        </tr>
                        <tr>
                                <td>Roll No: </td>
                                <td><input type = "number" name = "num" required></td>
                        </tr>
                        <tr>
                                <td>Choose a file:</td>
                                <td><input type="file" name="file" id = "file"><br></td>
                        </tr>
                </table>
                <input type = 'submit' value = 'submit'>
                </form>
                </div>
        </body>
</html>
```

PSEUDO CODE FOR PREPROCESSING -

- For the submitted doc. do preprocessing where images are separated from text and saved on local computer

Pseudo code -

```
1. OPEN THE FILE PREPROCESSED USING fitz.open
2. FOR EVERY PAGE IN OPENED PDF :
      a. EXTRACT THE LIST OF IMAGES USING getImageList()
      b. FOR EVERY IMAGE IN THE EXTRACTED LIST :
          i.   GET THE IMAGE BYTES USING extractImage()
         ii.   GET THE IMAGE EXTENSION AND LOAD IT TO PIL
        iii.   SAVE THE IMAGE TO LOCAL DISK
```

PSEUDO CODE FOR HASH VALUE COMPUTATION -

- Now for the extracted images , calculate hash values also compute hash values for all images in the database.Now compare the hash value of user input images to that of images already in the database.The images whose hash value is same as that of user image are displayed as output

Pseudo code -

```
1. FUNCTION LOAD_IMAGES(FOLDER) - TO EXTRACT THE LIST OF IMAGES IN THE
   GIVEN FOLDER
2. ANSWER LIST = []
      a. FOR EVERY FILE IN THE FOLDER :
           i.   IF FILE HAS EXTENSION FROM (.JPEG , .JPG , .PNG)
                   1. ADD TO THE ANSWER LIST
      b. RETURN ANSWER LIST

1. EXTRACT THE IMAGES IN THE DATABASE AND USER INPUT USING LOAD_IMAGES
2. ANSWER LIST = []
3. FOR EVERY IMAGE IN USER INPUT :
      a. CALCULATE THE HASH VALUE USING imagehash.average_hash() AS HASH1
      b. FOR EVERY IMAGE IN DATABASE :
           i.   CALCULATE THE HASH VALUE USING imagehash.average_hash() AS
                HASH2
          ii.   IF HASH1 == HASH2
                   1. ADD THE IMAGE IN USER INPUT TO THE ANSWER LIST
```

Output **–**

- Display output in a document plot a bar graph of image vs hash value difference for each and every user input image also save the results in an excel sheet

RESULTS

The below are the screenshots of the results/output that are generated from the above proposed system.

1. ***Excel sheet*** **-**

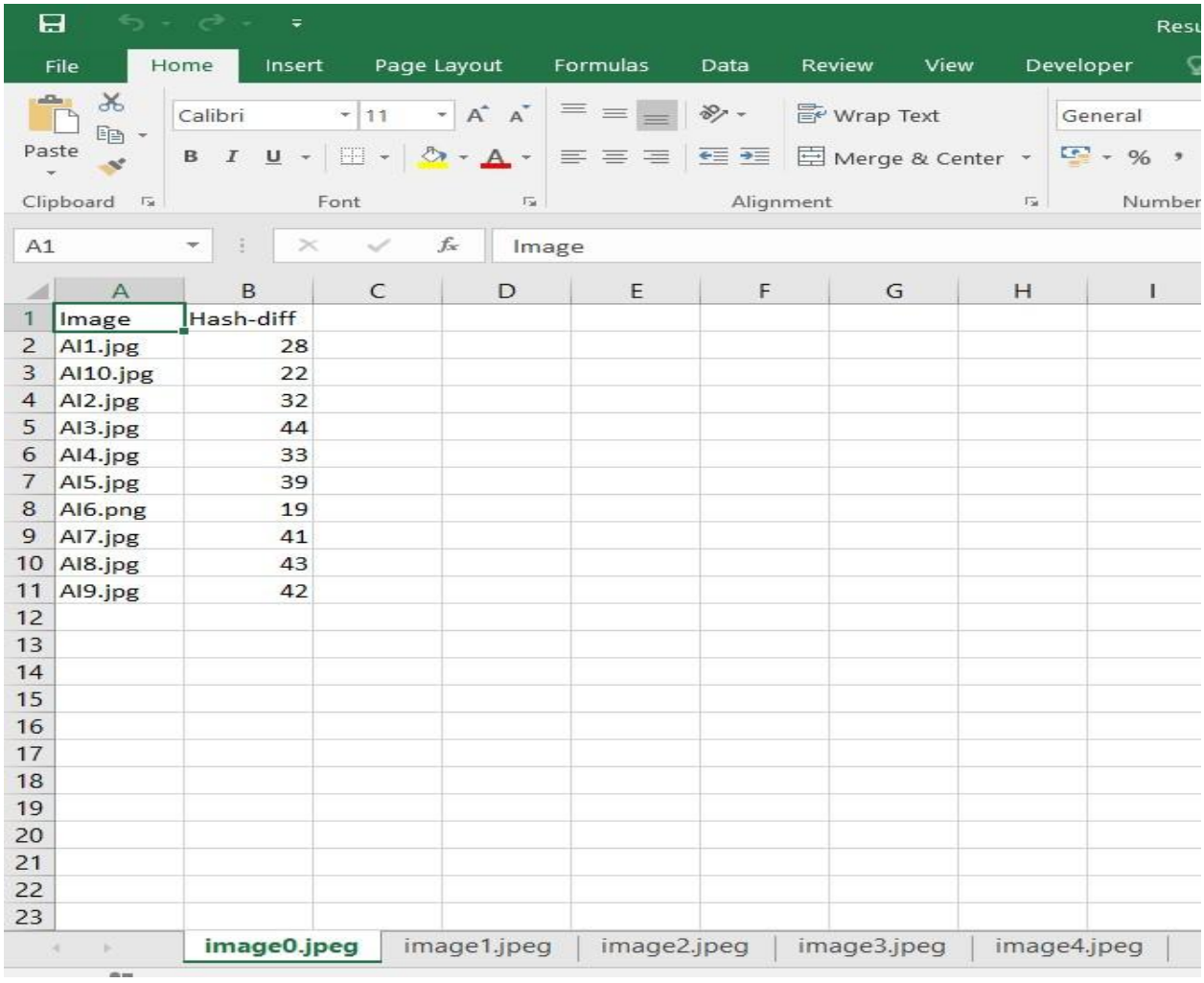

FIG 5.1 OUTPUT OF ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

The above figure shows or represents the data about the hash differences between the images that are submitted and the images that are already stored in the database.

Each and every image has its comparisons with all the images in the database in respective excel sheets.

2. ***Bar Graph*** -

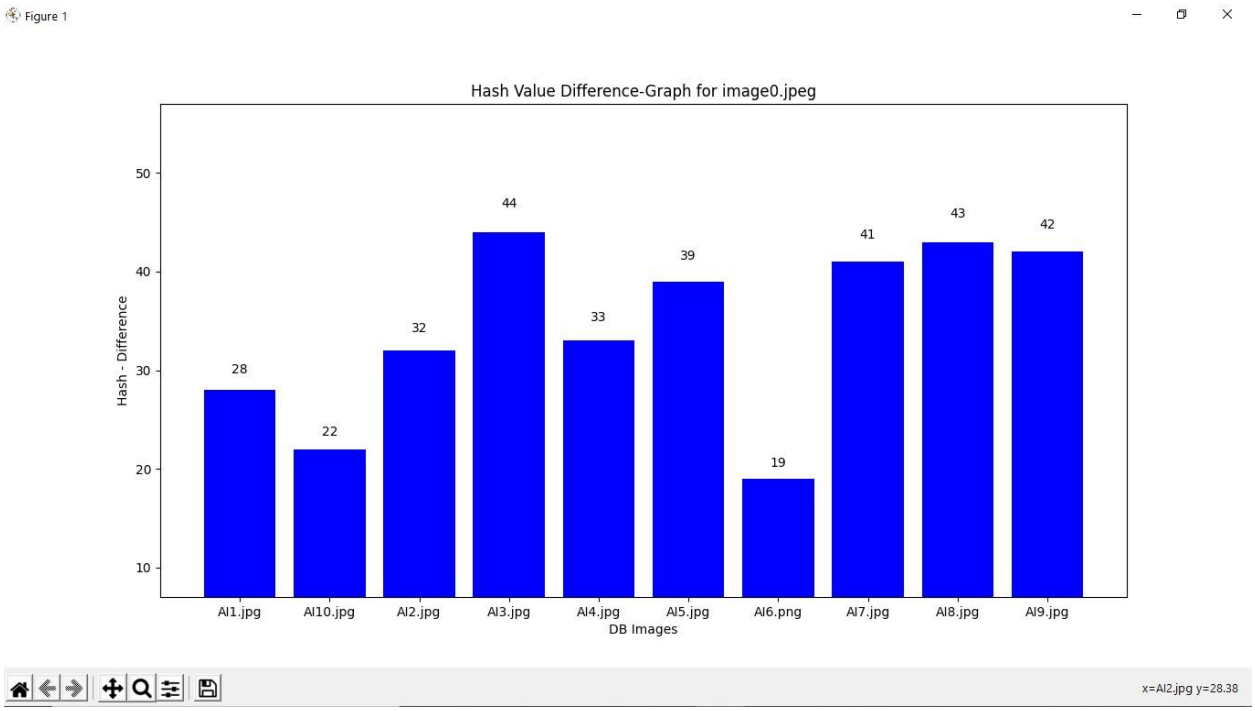

FIG 5.2 BAR GRAPH OF ONLINE ASSIGNMENT PLAGIARISM DETECTION FOR IMAGES

➢ The above graph illustrates the relationship between the hash value difference and the database images. Hash value difference on the y-axis and every image of the database on x-axis.

3. ***Result Document*** -

- These are the plagiarised images when performed plagiarism on the submitted assignment document.i.e., the images that are copied and are similar to the images (for which the hash value difference is "zero") that are already existing in a database.
- This is the final output of the system.

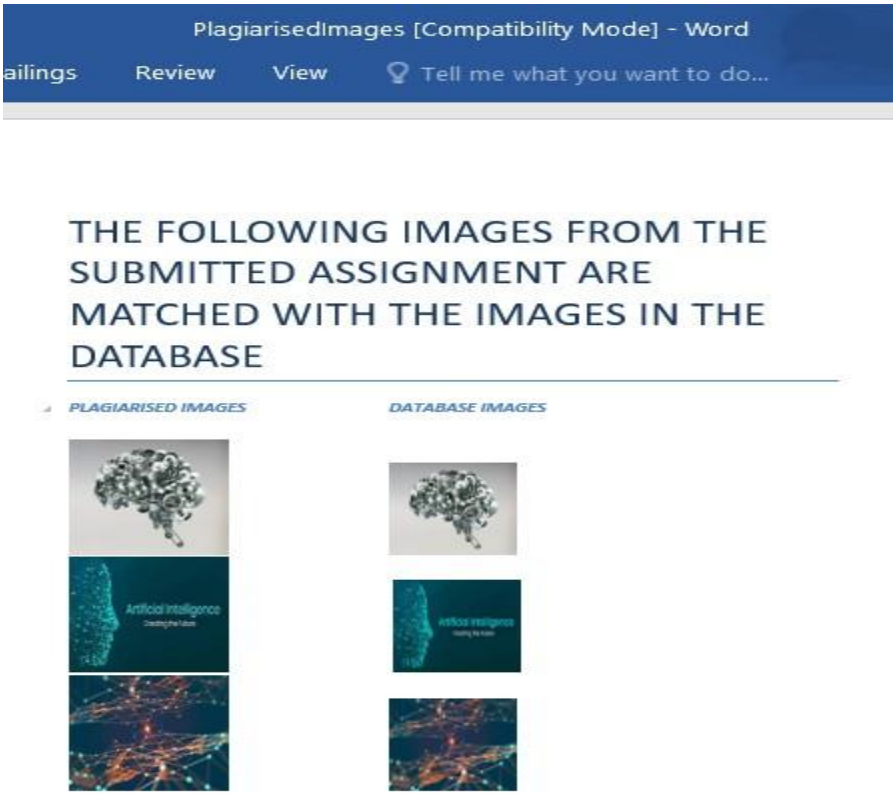

FIG 5.3 LIST OF PLAGIARISED IMAGES

CONCLUSIONS

- The proposed system helps in detecting plagiarism in students assignments using Reverse Image Processing
- Hash values of images are computed and compared with the hash values of suspected images.
- Hash values are chosen as the best measure because they ignore grey scaling , brightness , background colour of images if they have same content
- Finally the obtained result is displayed in the form of bar graphs for better understanding and also results are saved in excel sheet for future use

REFERENCES

1. https://github.com/piyush-kansal/reverse-image-search
2. https://github.com/ag-gipp/imageplag
3. https://towardsdatascience.com/image-similarity-detection-in-action-with-tensorflow-2-0-b8d9a78b2509
4. https://github.com/eisbilen/ImageSimilarityDetection
5. https://7webpages.com/blog/image-duplicates-detection-python/
6. https://www.pyimagesearch.com/2017/11/27/image-hashing-opencv-python/
7. https://ourcodeworld.com/articles/read/1006/how-to-determine-whether-2-images-are-equal-or-not-with-the-perceptual-hash-in-python
8. https://stackoverflow.com/questions/37220055/pip-fatal-error-in-launcher-unable-to-create-process-using
9. https://en.wikipedia.org/wiki/Hash_function
10. https://pypi.org/project/ImageHash/
11. https://www.thepythoncode.com/article/extract-pdf-images-in-python
12. https://www.w3schools.com/python/python_mysql_getstarted.asp
13. https://www.tutorialspoint.com/servlets/servlets-file-uploading.htm
14. https://www.programmersought.com/article/2796755717/
15. http://commons.apache.org/proper/commons-io/
16. http://commons.apache.org/proper/commons-fileupload/
17. https://www.w3schools.com/php/php_file_upload.asp
18. https://www.webucator.com/how-to/how-run-jsp-program-apache-tomcat-windows.cfm

19. https://pypi.org/project/opencv-python/
20. https://pypi.org/project/os-sys/
21. https://www.codegrepper.com/code-examples/delphi/how+to+read+all+images+from+a+folder+in+python+using+opencv
22. https://stackoverflow.com/questions/17358722/python-3-how-to-delete-images-in-a-folder
23. https://www.w3schools.com/python/default.asp
24. https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/
25. https://stackoverflow.com/questions/22276066/how-to-plot-multiple-functions-on-the-same-figure-in-matplotlib
26. https://www.geeksforgeeks.org/python-create-and-write-on-excel-file-using-xlsxwriter-module/
27. https://www.reddit.com/r/learnpython/comments/97boo6/dictionary_to_excel_file/
28. https://www.geeksforgeeks.org/python-pil-image-show-method/
29. https://python-docx.readthedocs.io/en/latest/
30. https://www.dummies.com/programming/use-labels-annotations-legends-matplotlib/
31. https://stackoverflow.com/questions/30228069/how-to-display-the-value-of-the-bar-on-each-bar-with-pyplot-barh
32. https://stackoverflow.com/questions/11216319/automatically-setting-y-axis-limits-for-bar-graph-using-matplotlib
33. https://python-docx.readthedocs.io/en/latest/dev/analysis/features/text/breaks.html
34. https://stackoverflow.com/questions/57361943/add-two-images-in-same-line-in-python-docx
35. https://www.slideshare.net/kmkmanoj1991/review4cs15