National College of Ireland

Signature Forgery Detection

MSc Research Project
M.Sc. in Cyber Security

Anaz Bin Ashraf
Student ID: 20230443

School of Computing
National College of Ireland

Supervisor: Mr. Michael Pantridge

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. ANAZ BIN ASHRAF…….

Student ID: ……20230443……

Programme: …… M.Sc. in CYBER SECURITY…… **Year:** SEPT 2021-SEPT 2022

Module: …… M.Sc. RESEARCH PROJECT ……

Supervisor: …… Mr. MICHAEL PANTRIDGE……

Submission Due Date: …… 15th AUGUST 2022 ……

Project Title: …… SIGNATURE FORGERY DETECTION ……

Word Count: ……7517…… **Page Count**……23……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …… ANAZ BIN ASHRAF ……

Date: …… 15th AUGUST 2022 ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Digital Signature Forgery

Anaz Bin Ashraf
Student ID 20230443

Abstract

Digital signatures are widely used in recent times by organisations public and private alike. Similar to fingerprints the signatures are legally binding. Another reason why they are being used is that they are easy to handle and store. The digital signatures are utilised mainly in e-commerce websites for delivery authentication to the customers, bank customer procedures and verification, government organisations and various other businesses big and small. Nowadays, the government uses digital signatures for contracts and verifying documents. When there is an advancement in IT, it has its advantages and disadvantages. Signature is one of the important biometric techniques that may be used for manipulating the signature data and using them for malicious purposes. Two efficient machine learning algorithms VGG16 and random forest are implemented in the research attempt to identify a way to mitigate the risk caused by signature forgery. The machine language techniques are being trained and tested to check whether the signatures given are original or fake using a data set.

Introduction

Signatures were being used in the world for nearly a thousand years. It is written in a variety of patterns; some may be the letters in different languages. It is considered as evidence or consent of someone's agreement with someone else. This can be a contract agreement, sales agreement or anything relating to someone's identity that can be described in simple letters. As trade and market expanded, the use of signature has also expanded. After the digitalization of the industry, various conventional methods were converted into digital formats including signatures. This has helped in improving security standards but as the cybercriminals are getting smarter day by day cyber-crimes have also increased. Now it's been a great challenge for organizations to safeguard the digital information of the customers also identifying the authenticity of this digitally stored information has become a great challenge for organizations. Therefore, identifying the authenticity of digital signatures has been an important fact in the current scenario.

A signature can be easily categorised as one of the most important biometric techniques that dominate the life of an average man as it is used in every walk of a person's life say more than other biometric techniques such as eye scanning or fingerprint scanning. The documents that we use digital signature ranges from bank letters, property documents, contracts, bank checks, corporate documents, currencies, bonds, etc. The digital signatures are stored in a convenient location which can be easy to manage and retrieve whenever needed. Lack of knowledge on increasing threats of digital signature forgery where the attackers can access the systems stored

in a remote backup location or cloud, steal the data and do malicious activities on the signature data.

1.1 Background

A signature is one of the most popular writing styles used by people to formally bind a document or to indicate their preference for certain information and, if necessary, to convey obligation. Financial documents that are generally acknowledged by the government, business world, and other legal activities include bank letters, property documents, bank checks, corporate documents, currencies, bonds, etc. People used to duplicate forged signature patterns in the past largely for financial benefit. These methods changed throughout time, just like previous methods did. The fast growth of technology and computers in recent years has had a big impact on signatures. There are several reports of signature forgeries, which is a major cybercrime.

One of the most crucial biometric methods that run the risk of being tampered with and used maliciously is signature forgery. It is a topic that needs to be researched in order to maintain organizations' cyberspaces operational. During the COVID-19 pandemic, it saw exponential growth in these problems. Numerous factors are carefully examined, including size, letter shape, line quality, and orientation. Due to the frequent need for various forms of paperwork and signatures, it has an impact on the private sector. Therefore, much care must be taken in the matter of how to handle the security and integrity of the details as these will be increasing as the business expands. The study sheds light on various recent experiments. It gives details of the security patch which needs to be fixed to provide a way to recognize forged signatures. This can be accomplished by intensifying the forgery detection techniques using powerful and better deep learning techniques like Random Forest and VGG16.

1.2 Motivation

After realising the danger of how forged signatures can create severe privacy and vulnerability issues for the customers, it certainly compromises the security of the data. The data which may be collected through appropriate devices are usually stored in a database. Some organisations use the technology of storing them in the cloud as a way to eliminate paperwork and manage space over physical devices as they may slow down the process. Many organizations advise considering using a cloud-based electronic signing platform that can manage and facilitate to provide a sense of security by adding an additional layer of protection where there must be several steps and audit trails to take into consideration when there is a change in signature data or if there is a desire by the user to modify his existing signature. According to recent studies, the banking sector accounts for roughly 22% of all signature forging incidents, with counterfeit checks costing an estimated $900 million each year.
The research proposal’s goal is to find a state-of-the-art machine learning algorithm that will be efficient in detecting discrepancies involving signature forgery.

1.3 Research Question

How can we identify signature forgery using modern image recognition techniques such as Random Forest and VGG16?

The main goal of the research proposal was to search for and finalize an algorithm that should be both fast and accurate and used for detecting the discrepancies involving signature forgery in the various organisation data sets. The deep learning convolutional neural network VGG16 and Random Forest is used for the classification of the data and identifying whether the signatures are fake or not which are in the form of images. Both the algorithms are very efficient and are used for the analysis of such datasets. An application is being developed to compare the images of the signatures and verify them written in Python programming language. The application is a crucial part of the research project which increases its feasibility in the module. Here the output of the implementation is justified as we develop a simple application for this. There the application tries to compare the authenticity of the image which are uploaded. After verification, the output would be produced saying whether the signature is fake or original. The digital signatures are significant as they are valid and legally binding evidence in many instances. In the case of a criminal investigation, many forensic investigators try to analyse evidence that will include signatures to verify the crime is linked to the person having the same signature.

1.4 Research Objectives

The objectives that were pointed out to be achieved by the time the research works are completed are given as follows:

Feasibility: Before settling on Random Forest and VGG16, numerous strategies and algorithms were examined to see which would provide the best results for detecting forged signatures. The most recent studies on signature forgery have used Random Forest and VGG16 because they are more effective than the already used techniques and algorithms. This will ultimately help classify the entire set of classes in mode or calculate the mean of the class by prediction or regression technique. VGG16 is a convolutional neural network that aids in understanding how the layers of the digital image are connected, how well they are connected, and how to retrace them. They are versatile and offer much greater accuracy and outcomes than other techniques. To extract the fundamental features contained in the signature vectors, it will fine-tune them. To carry out the aforementioned process, VGG16 has a number of processes. The features will be extracted for further analysis at the fully connected layer, which follows after the convolution layer in the process.

Clarity: The clarity of the research is guaranteed as the research is carried out by collecting evidence and information from various sources; thus, how the research is carried out is simply maturely and practically. An application built with Python is used to carry out the research,

giving it a quantitative framework. Finding out which dataset is the real or fake signature will be the web application's primary goal. As a result, the outcome is merely confirmation of whether or not the signature is genuine.

Significance: Cybercrime is now more prevalent than ever. People attempt to hack various technologies and databases of systems in order to obtain personal information about others and use it for nefarious purposes. Obtaining a person's biometric information carries a larger risk because the attacker can commit their terrible act undetected and would be able to track it back to the person whose signature is being utilized. Organizations must aim to increase their efforts to avert such incidents from happening because this threat is connected to people's private life. If not, people would be more hesitant to provide their information in a digital form regarding the consequences. By transferring money, manipulating other people's work, gaining access to places that aren't regulated, and other methods, people attempt to gain economically. As a result, this study and its relevance are important for improving the review process that will aid in spotting such frauds.

• Ethicality: The objective of the research is to be ethical, meaning that a person's handwritten signature should be maintained securely because it may be exploited by criminals to commit crimes in that person's name. The privacy of a person's signature should be protected, just as the privacy of a password for a device, as it serves as a social network's primary means of identifying that person. As a result, our study solely attempts to be a part of society by contributing to social well-being and is never involved in any infringement of people's or organizations' rights in any way. Any data discovered during the study will be kept private and will not be made public. Furthermore, data collected from a person, or a group of people will be anonymous, keeping the identity under wraps for the kept safe and secure. All of this is done to keep the research ethical.

2. Related Work

2.1 Background of Digital Signature Forgery

A novel idea for identifying digital signature forgery was suggested by U.V Marti (2002) using a deep learning algorithm called k-NN or K-nearest neighbours. The technique is well defined and reaches an accuracy of 92% but with a considerably small dataset of 40 signatures. While checking for other approaches and techniques regarding digital signature forgery some papers did use cryptographic techniques to find the forged digital signatures.

A paper by Rashmi Kasodhan (2019) uses several cryptographic techniques including hash code, DNA encryption/decryption technique, SHA algorithm and Bio Gamal algorithm. The time complexity of the module is reduced but the final output has just 30-40% accuracy.

Another paper written by Longge Wang (2016) provides another cryptographic technique which uses elliptic curves. It is formally introduced to ensure the security of customer data and

ensure confidentiality. It checks authentication efficiency and security needs of signature data stored in a cloud platform. But the major limitation here includes the non-effectiveness that is surrounded by the elliptic curve discrete logarithm problem. Another limitation is the failure of attaining high security in the program. The only thing that improves within the program is the modular inverse operation that proportionally improves the operational efficiency of the cloud platform. Thus, from this paper, it is understood that ECDCP is not applicable for finding forged signatures. Even though the cryptographic techniques are used it is not recommended as they are obsolete compared to the high efficiency of machine learning and deep learning algorithms. The technique is becoming easier to crack and break in through the system by other means. It even can't protect the data from unencrypted documents. Another paper written by Zuraidasahana Zulkarnain (2015) uses a mean to standard deviation method to find the forged signature from offline signature data. This technique is good but using machine learning is far better as it has evolved from the above paper a lot.

2.2 Digital Signature using machine learning algorithms

Deep learning and machine learning are considered the breakthroughs that have happened in this century. It helps in achieving better accuracy and efficiency among the models by extending the depth and complexity of the existing network layers. The major important point to be noted is in finding the genuine datasets which are used to train the algorithms so that high output is obtained. Dr Santosh Kumar Bharti (2020) proposed the use of machine learning algorithms for digital signature forgery. The paper uses the CNN method along with the SURF algorithm & Harris corner detection algorithm. It proves that using machine and deep learning algorithms can be used for getting higher accuracy and precision and thus ensure them getting the results. Here the associated accuracy hits near 85%. During signature forgery detection it has an accuracy of 85-89% while during signature detection it has 90-94%.

Lu'lu'il Ayunin Fakhiroh (2021) talks about a methodology for finding the accuracy of offline signatures based on mobile devices. This paper propped a new approach based on CNN in mobiles. It trains two different networks separately to test the same dataset. It provides the highest accuracy for training with 99.33%. During testing, this has provided a 100% accuracy rate with a loss value of 0.032. But the drawback that was found was it had a loss value during verification cases greater than the normal cases. Another drawback that was present was the testing is more optimal and more suitable for mobile applications. The obstacles on these mainly occurred due to the number of datasets and processing them during data augmentation. The training process is also flawed due to the loss value feature of information during pre-processing.

Manikantha K (2021) has a paper that gives more insight into how deep learning methods are analysed in signature forgery detection. It uses the Siamese network, a network which is originally a twin network architecture which makes use of 2 identical networks. They have the same configuration in the system where similar observations could be used for future references. In the results, VGG16 attains the highest accuracy among the algorithms which are used together. Thus, it gives strength to the fact that deep learning algorithms can achieve

maximum accuracy than normal machine learning algorithms. Different algorithms are producing highly accurate results for different datasets. No one algorithm has an upper hand on the number 1 algorithm among them although VGG16 comes close to it. The epoch value is low as they only use 15. Testing them using fewer epoch values is a safe way to identify the accuracy of the signature and find the forged ones with high accuracy results. But the percentage of accuracy may differ if the epoch value is greater. So, even though the accuracy of the algorithms used is high this was considered a drawback.

The main target was to train the two algorithms here to understand how efficient both algorithms are as one is deep learning and the other one is a conventional neural network. It also needs to be understood how the testing and training speed differs in both as deep learning takes more time for training the system and testing them for the results which we are aiming for. While doing the initial research for the dataset, both real-time and already established datasets. But it was decided to choose a predefined and existing dataset which has all the information we needed, and which is publicly accessible as well.

The properties of various algorithms were studied before finalising the used algorithms in the research project. Various indicators were factored in while finalising the algorithm including the efficiency of the algorithm, speed, error rate, deep layered analysis, etc. While using two algorithms two different methods were able to implement in the program which benefitted positively.

Machine learning models

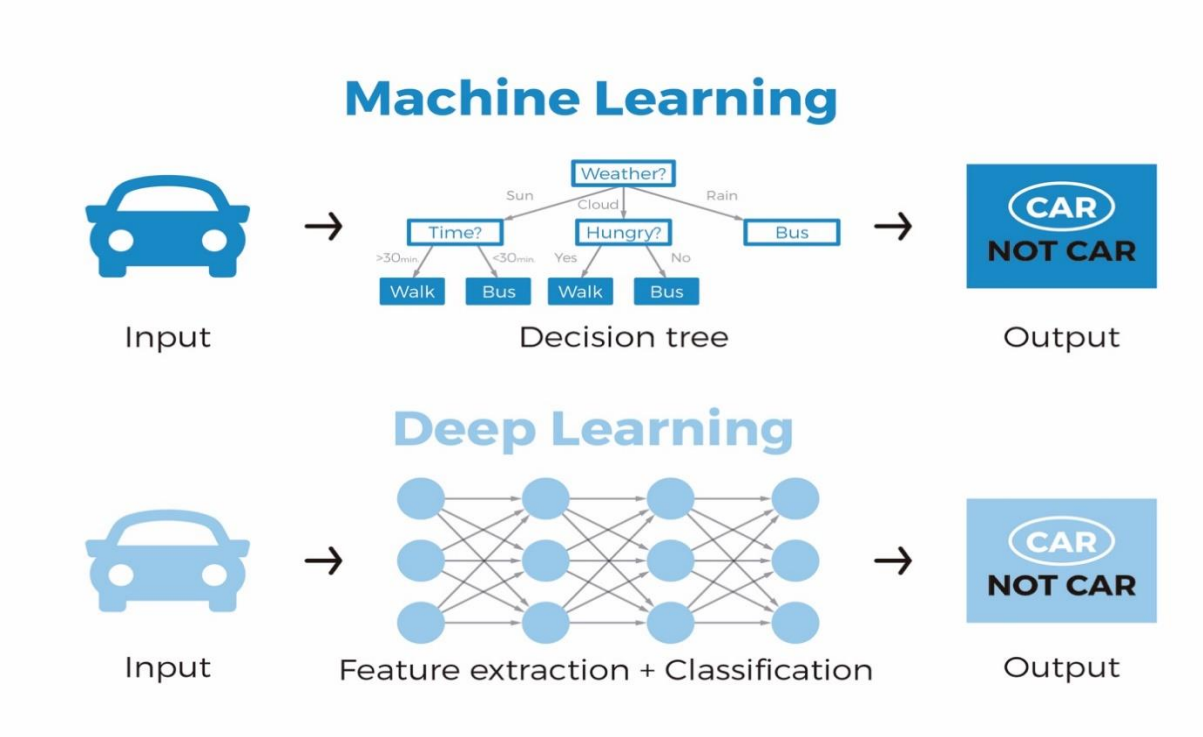
Atefeh Foroozandeh (2020) mentions various machine learning and deep learning algorithms which can be used for finding the original and forged image from the dataset. The machine learning algorithm was run with various matching functions to understand it's working. There are several databases which are chosen from various resources and organisations. These datasets were grouped, and the tests were done on each of these datasets by all the algorithms. The results are compared, and the best algorithm and technique are found. We have made use of neural networks, decision trees, and deep learning techniques to do our research method.
The algorithms which we used in the research project are both superb. We trained and tested VGG16 and checked the output by comparing it. When checking the other algorithms VGG16 performs far better as it's faster as well as accurate.

Justification for the need for the research question.

Comparing all the research papers some methods use a significantly small dataset whereas some have a very limited number of individuals' signature samples as some have just one authentic individual signature and two or three fake signatures. Thus, it was decided to perform the research question of digital signature with a comparatively large dataset even though it's not that large enough. Using a large dataset has its difficulties as it may encounter the

limitations of the previous work as well as new problems. By solving them we are trying to create a new model for digital signature forgery.

3. Research Methodology



[1] **Figure 1: Difference between machine learning and deep learning**

The research methodology is clearly stated below. Handwritten signatures alter for a person over time. Thus, the verification and authentication of these signatures take a lot of time and various processes to be followed to carry out these tasks. So, it becomes more difficult when a person tries to replicate the signature of another person and try to forge it. Thus, it's very important to train the data to the machine learning algorithms to minimise the errors found in some cases according to Yazan M. Al-Omari (2011). The following steps are followed for carrying out the methodology in the research project:

Dataset loading:- The dataset is downloaded to the machine and loaded into our machine for running in the program for training and testing the data. The dataset is a public dataset which consists of nearly 1600 images of signature data. The signature data consists of images which are both fake and original images. Each of the images in the dataset consists of a directory number which is the name of the particular image. If the image is an original image it is denoted

[1] https://blog.bismart.com/en/difference-between-machine-learning-deep-learning

by the user number. The fake or forged images are denoted by the user number along with adding "_forg" i.e. 001 for the original image and 001_forg for the forged image.

Data pre-processing:- The data for the research method is taken from a dataset which consists of a large number of signature images. The images are so large that no two images may be in the same orientation and have similar features. For this purpose, we need to implement pre-processing for the images so that we can treat each image equally which will allow for testing and training to be done efficiently. The pre-processing process includes removing the colours from RGB to greyscale, resizing the image, removal of blurs on the image, etc. according to Douglas J. Kennard (2012). The images will also be cropped and rotated in order to fit in the uniform style. We try to remove the noise from the images using the Gaussian denoising method. Canny edge detectors are also implemented to get the edges of the images. Binary processing is also carried out in order to convert the image into s binary format.

Visualisation:- Visualisation is the step where we are planning on how the graphs can be plotted in the program and based on this information the data for testing is chosen. The data testing follows this basic step thus making the visualisation an important step in the flow diagram.

Dataset splitting:- In this step, we are ready to split the data for training them to the respective algorithms we are going to use. From the training data, we are going to get a smaller set which will be sent to the testing data from which we compare the images and get the output.

Define algorithm model:- After finalising the data for splitting we have to define the algorithms which we are going to use in the machine. There are two algorithms which we are going to use here VGG16 and random forest. Each of the algorithms is not used simultaneously. Firstly, we run a particular algorithm, either VGG16 or random forest, do the training and testing and find the accuracy of the image which indicates the authenticity of the signature image. Next, we run the second algorithm in the same way and check the output. Thus, the process for both algorithms is the same but is run in different steps.

Training the data:- Training is important as we have to give information to our algorithms on what we are going to do with the data. We need to familiarise the data with algorithms then only it will be able to efficiently perform the task it was aimed for.

Cross-validation:- After the training and testing, the data are cross-validated in the next process to compare and check for the comparison between the images and to find out which one is fake and which one is original. This result will go through as the final output that will be displayed and thus we get the metrics.

4. Design Specification

Based on the research done on the research topic, it is found that random forest and VGG16 are the most efficient machine learning models available for checking whether the signature

provided is fake or not. They are trained and tested with the dataset to provide the desired results. Each of the machine learning algorithms is provided with their detailed architecture as given below. There are different types of machine learning algorithms. They can be categorised as supervised and unsupervised. Both are used in different scenarios in varying datasets. So, it must be needed to know which algorithm is which type. During supervised learning, we train the machine learning using the data we have from our dataset. It is used for two types of problems i.e., classification and regression.
Unsupervised learning is the type of machine learning algorithms which are trained without labelled data as input. They are mainly used to find the structure, patterns, and models from the given data. They are not needed to be supervised as a teacher a student as they learn data on their own. In unsupervised, the major usage is based on finding two problems. They are clustering and association.

4.1 Random Forest:-

As discussed by Pooja Gaikwad (2021) it is an algorithm which is used for doing both classification and regression tasks. It enables hyper-parameter tuning in nearly 90% of the cases in which they are used. It is a combination of a number of tree vectors that are dependent on the value of the random vector which was sampled independently. Random forests are a combination of classification for a range of tasks, regression techniques, feature selection and classification of the dataset of signature combined by constructing a large number of decision trees. Each of the trees will have the same distribution in this. It tries to understand the best among the various features already there in the algorithm instead of choosing the most important feature in the nodes according to Taraggy M Ghanim (2018). Another main advantage of this architecture is that it is never exhausted with overstuffing of samples as it employs random subsets for all the features and all are accommodated through building smaller subsets. The final output of the random forest is calculated and generated by taking the average of the outputs of each tree through a voting system (Leo Breiman, 2001).

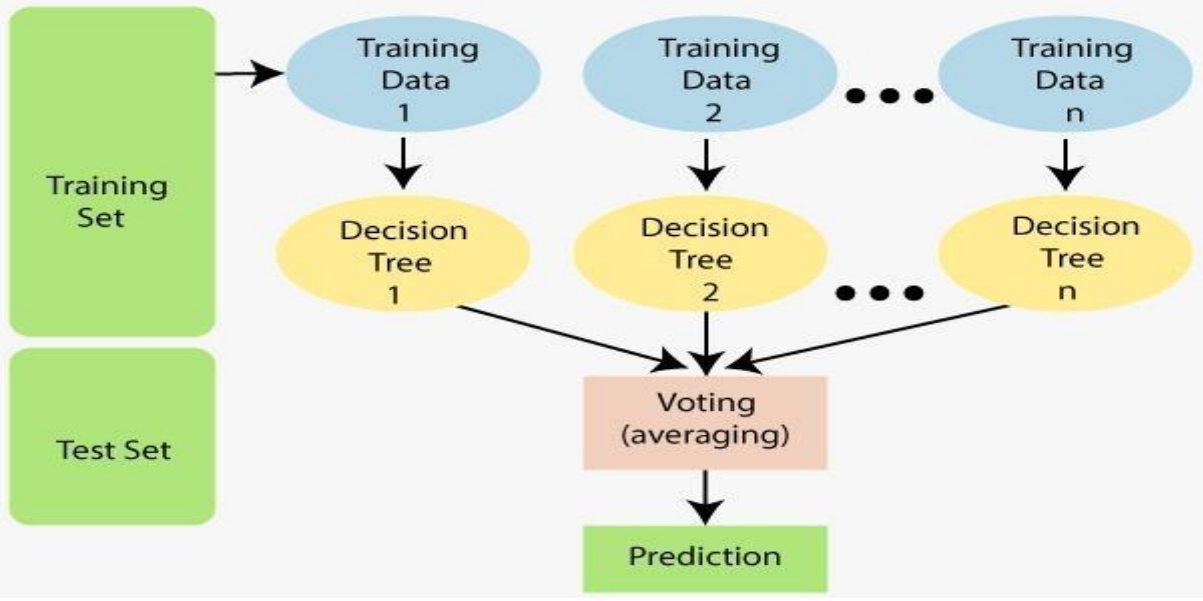


Figure 2: Random forest architecture

The workflow for the Random forest algorithm is given as follows:

-The main aim of the research question is to find how digital signature forgery can be identified and what can be done to train the models to find them.
- Divide the dataset into training and testing data.
-Select k value and calculate the distance.
-Sample space is selected from the test data and distance is calculated for every n number of training data sample.
-We obtain a set of distance and firstly they are sorted and then the k-nearest data is picked.
-The class that has the largest number of k neighbours is selected for the test class.

4.2 VGG16:-

As discussed in M. Muzaffar Hameed (2021), it is a convolutional neural network which is deep and uses 16 layers. It is one of the best model architectures for vision. It promotes high accuracy. Instead of providing a large number of hyper-parameter, they instead focus on convolutional layers of 3x3 filter and a max pool layer 2x2 filter. The first one has a stride 1 whereas the second one has stride 2. It has 16 layers that have weight. It has got training from at least a million images and thus the network has a representation that covers rich features for the wide range of images. The image input size of this ranges in the size of 224x224. It tries to normalise the pixel of the input image's gradient, successfully providing a smooth gradient ascent process that skilfully avoids large and small gradients.

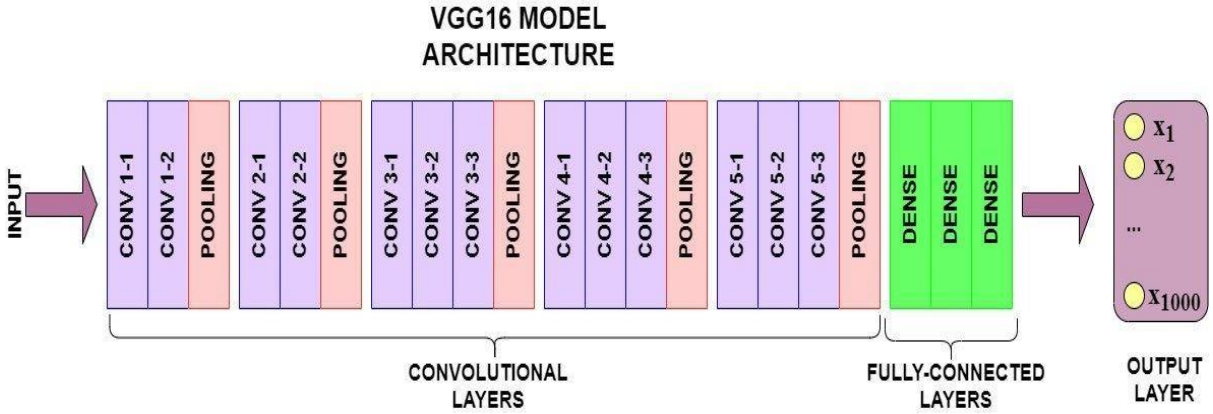


Figure 3: VGG16 algorithm architecture

4.3 The algorithm steps implemented for VGG16:-

Step 1: Start
Step 2: Loading the data as an image array.
Step 3: Convert images from RGB to greyscale images.
Step 4: Resize the image size into 256*256.

Step 5: Create an array of images and array labels.
Step 6: Split the dataset for training and testing.
Step 7: Define the VGG16 model.
Step 8: Send the training part to the model.
Step 9: Evaluate the model.
Step 10: Create a VGG16 model as a classifier.
Step 11: Train this model with the record pairs.
Step 12: Check the validation matrix for performance
Step 13: Tune the model parameters to get better performance.
Step 14: Get the new data after tuning.
Step 15: Confirm whether there is any duplicate data that is there or not.
Step 16: If yes, reject the record.
Step 17: Else if no, accept and store the record.
Step 18: End

5 Implementation

5.1 Digital Signature Forgery Implementation

The digital signature forgery is run with modern machine learning and deep learning tools rather than using cryptographic techniques for finding the output from the values between mean and standard deviation that is used earlier. This approach provides much faster and high accuracy than the traditional methods. The major implementation technique that is required to run the program is to train and test the data, search for similar records and combine these results to know whether the selected signature image is fake or not (M. Muzaffar Hameed,2021).

Pre-processing: - Before doing any testing of the data, it is important to do the pre-processing of the data to make the process easier. The dataset will be going through a lot of processes to make every one of the images in the dataset uniform such as resizing, RGB to greyscale, normalisation, cropping, orientation, etc. Normalisation is done to change the pixel value of the image from a higher value to a lower value or vice versa. We try to change the pixel value to 256 * 256 as 128 bits may not be that clear and may be unable to get the edges of images or the image is blurred. Blurry images are sharpened during this process as well.

Noise removal: - We deal with noisy images by using the Gaussian denoising method. This removes noises from images especially if we try to use salt and pepper noise and blurring effect. These would remove the noises. In case, if there is an instance where more noise, we use auto encoded to remove noise from the image.

Defining the model and training data: - After cleaning the data and making it uniform, it is time to define the model to be used to train the data and further train them. The similarities and differences between the images must be trained for the model to understand to get accurate

results and make fewer errors. The data for training and testing was selected in a 70:30 ratio. The optimizer used for our proposed research is the R-Adam optimizer. We have used this because it is one of the best optimizer techniques that are available right now and way better than the optimizer tool that has been used in the earlier research proposals. It will allow us to reach the global minima during the time of training. It helps to get out of the local minima and try to help reach the global minima. It adapts to new learning environments and datasets much better than other optimizers available and thus is much more accurate and reliable compared to the other optimizers

Testing and cross-validation: - Testing is carried out to check the efficiency attained by the algorithms during the training. The aim of testing is fulfilled if we are receiving the desired output from the input we provided. The success rate of finding the forged and real image provides accuracy in the entire research process. The testing results are processed, and we carry out a five-fold verification process as well to check the other properties of the output and we display the metrics. The models are run separately each time and not done together. Firstly, we run the random forest and it provides a set of results and plotted graphs showing the level of accuracy it has achieved while doing the project. Similarly, we run the VGG16 algorithm in the same manner in 100 epochs which give another level of accuracy and speed as it is a deep learning algorithm with a lot of layers beneath it. Both algorithms are compared together to understand which algorithm fared better in finding the forged and original image and which one is having the least errors and loss rate.

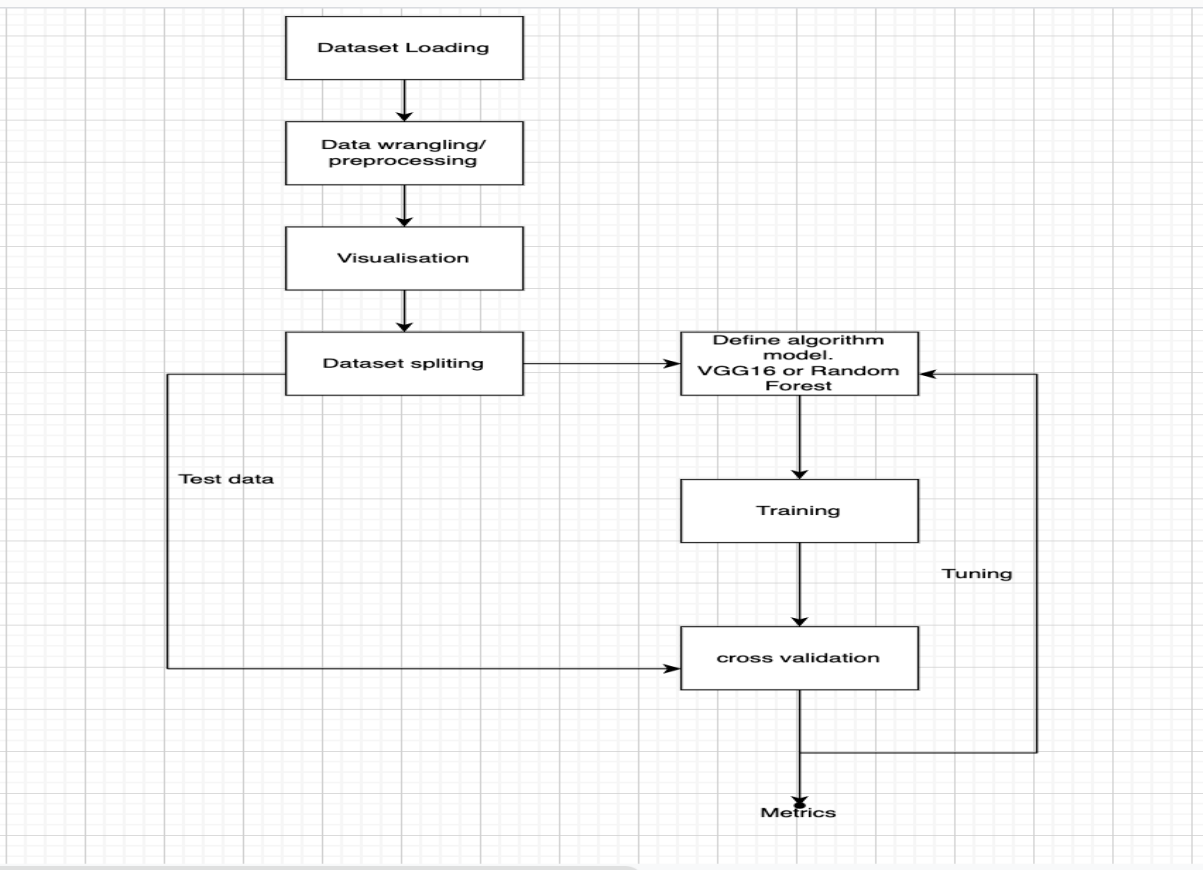


Figure 4: Implementation steps using the models VGG16 and Random Forest

5.2 Tools/ Language used for implementation:

-Jupyter Notebook 3: It is a browser-based independent application. The major uses of the Jupyter notebook are as follows:
-It helps in creating documents.
-It helps in displaying live codes, equations and other snippets in codes.
-The formatted or edited document can share among other tools in the machine.
-It also has many other features including cleaning of the data, statistical modelling, visualization of data, etc.

-Anaconda 3: It is an open source python package that allows the data scientists and M.L enthusiasts to ensure that the same packages and dependencies are installed even if there is two different OS used for running. It helps to launch applications without having to use the help of command-line commands.

-Python 3.7: It is the latest version of python used for doing the research project. It is an object-oriented programming language that is interpreted. It is very easy to use due to its reliability, clarity and usage. It is preferred for running algorithms and comparing datasets for most of the programs. Python is comparatively simple than the other programming language which also makes them more popular. The syntax for python is less and easy to code making them more efficient and code-friendly.

-NumPy: It is a library function that is used in Python. They are imported to the program for doing mathematical functions in arrays like multiplication, algebra, matrices, etc.
-Scikit learn:- It is another library which is used in python. Mainly they are developed to create and assist new models. It uses several supervised and unsupervised techniques in it. This library helps mainly in the cross-validation and conducting multiple hyperparameter searches across the models. It also helps in finding the best model from the program.

-Pandas:- It is one of the open source packages used for python. They are very fast and efficient and help in manipulating the data, help find missing data, cleaning the data, etc. It also helps in reading or loading data in a wide range of formats such as CSV, Excel, etc.

-TensorFlow library:- IT is another open source library which is used to provide data flow graphs during running and testing the model. Without the help of TensorFlow, we can't train the models that are using deep learning algorithms.

-Matplotlib:- It is introduced as a plotting library that is run on Python. This along with NumPy is used as an alternative to MATLAB in Python. Thus, for the machine learning model applications, this is a very important package for plots and graphs.

6 Evaluation

Evaluation of a model is carried out when we provide experimental evidence by carrying out the implementation of the research under a limited environment. This is carried out to evaluate whether the results are matching with our predefined requirements including matching the properties such as they are feasible and reliable, etc. Such implementation under controlled conditions will allow the researcher to notice whether the accuracy matches with the desired accuracy the researcher has in mind to deem it a success. There are mainly two methods which are used to evaluate the model in machine learning. They are hold-out and cross-evaluation techniques. Both these techniques employ a test set to avoid problems that arise due to overfitting. The hold-out method is used for generally large datasets and is divided into three sections namely, training set, validation set and test set.

Whereas the cross-evaluation or k-fold cross-evaluation method, is mainly focused on datasets which are comparatively smaller in size. Since our dataset has a limited amount of data we employ the k-fold cross-evaluation method. In this, we divide the dataset into k number of subsets each having an equal size.

For this research proposed, we are using a five-fold cross-validation method. The implementation is carried out through various methods. Firstly, the dataset is divided into 5 parts and each subset has an equal number of data present. Out of 5, one is selected as the test set whereas the remaining is processed as the training model set. While the training is evaluated several calculations are being performed based on various parameters such as true-positive, false-positive, true-negative and false-negative. After this, the records of both the test set and the training set are calculated and the average of the results is obtained. One portion of this dataset from the test is chosen for the training models. This would be needed for the final output as the results for the final test.

The accuracy of the data is obtained, so we calculate the four distinct parameters. These are mentioned below as follows:

Precision: It is used to display the actual positive values that are found from the estimated positives that are displayed.

Precision = TP / TP + FN

Recall: This is the expected positive that is deduced from the total positive values.

Recall = TP / TP + FN

F1-Score: It is calculated as the average of Precision and Recall values.

F1 score = 2*(Precision*Recall) / (Precision + Recall)

6.1 Experiment 1

The dataset is trained and tested using the algorithm model Random forest is used and from this testing, we get a confusion matrix and two bar graphs. The figure below depicts the following:

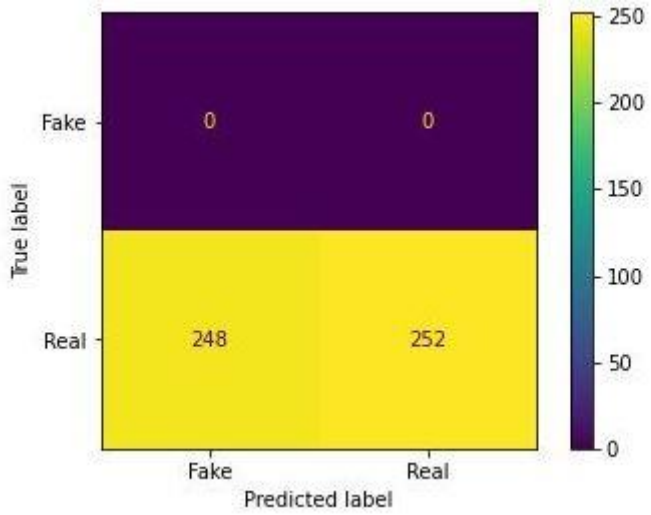
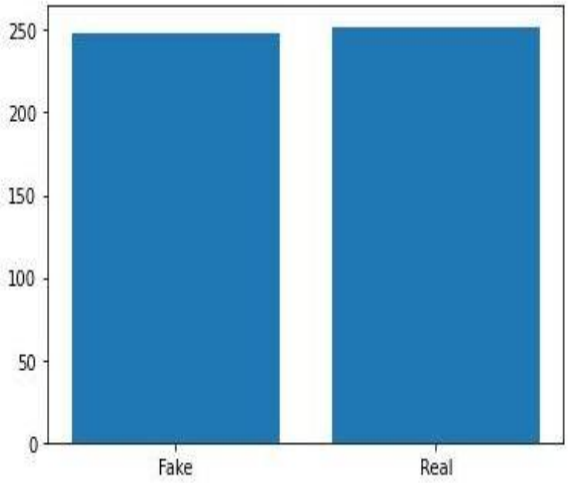


Figure 5: Confusion matrix for Random forest algorithm



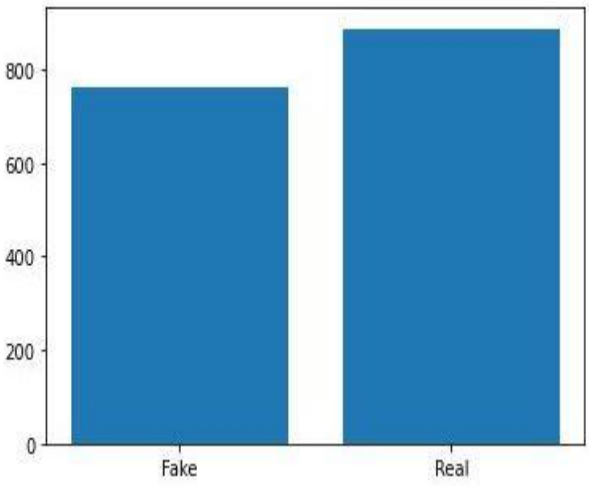


Figure 6: Bar graph for Random forest algorithm representing real and fake images

6.2 Experiment 2

While training and testing the model using VGG16, a Convolutional neural network, we get the following results during implementation. The sample image of the training data is given below.

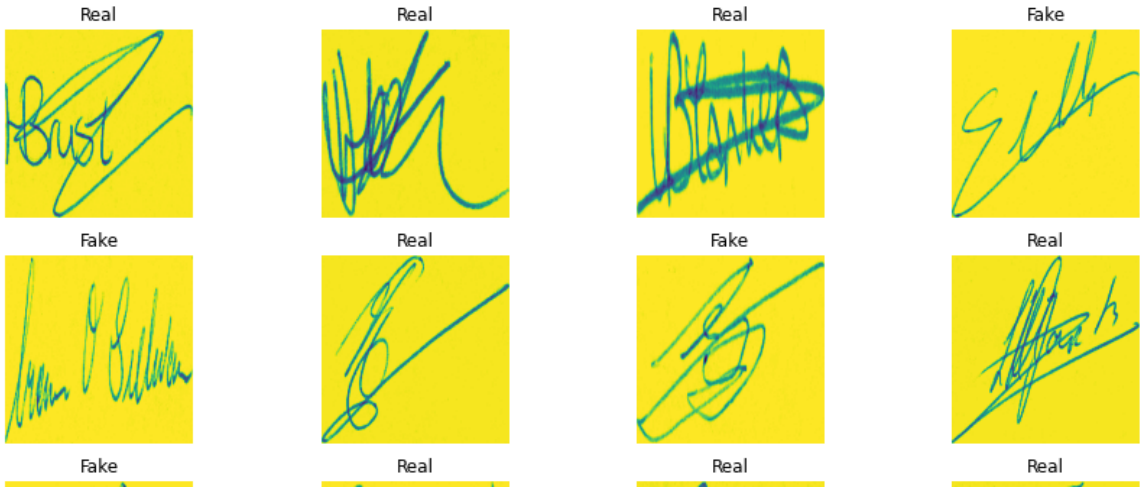


Figure 7: Sample image for training data for VGG16 model

The accuracy plot and confusion matrix of the VGG16 algorithm is given below.

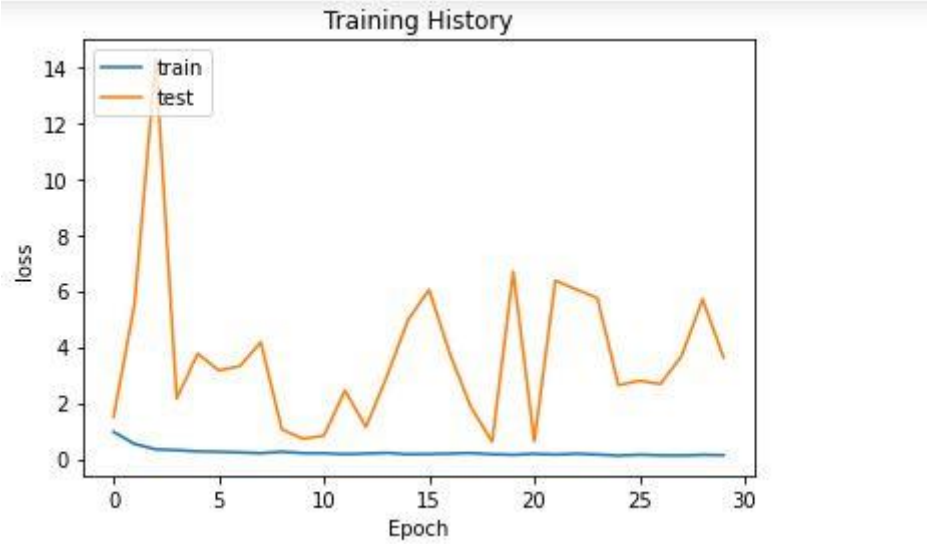


Figure 8: Line graph for training history

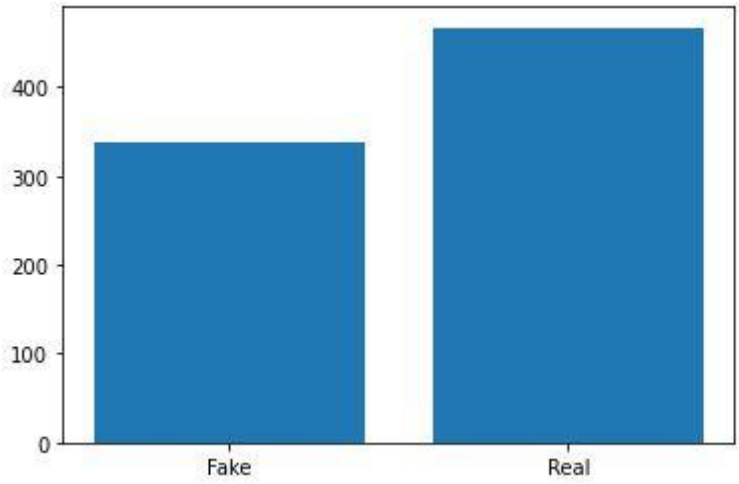


Figure 9: Bar graph for VGG16 representing real and fake image

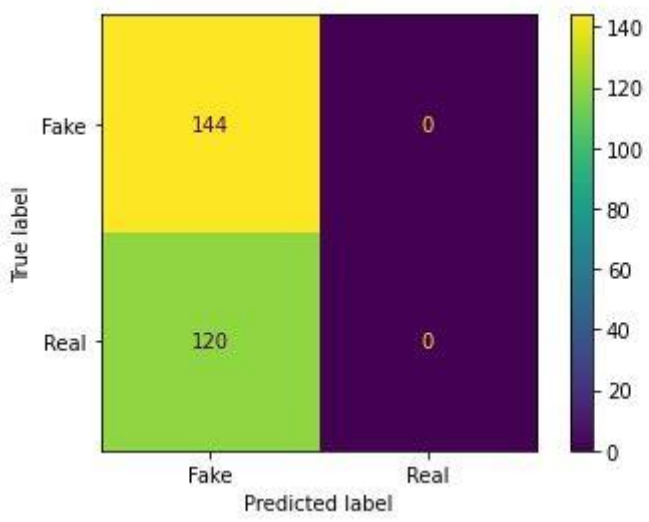


Figure 10: Confusion matrix for VGG16

6.3 Discussion

From running this project, we understand that VGG16 has better efficiency in finding output than random forest. This is supported by the accuracy graphs and confusion matrix plots of both of these algorithms obtained while running the program. The accuracy of VGG16 is 54.54% Whereas the accuracy for the random forest is found to be 50.4% From this we can conclude that signature data forgery can be better tested when we have VGG16. This is possible because VGG16 is a deep learning algorithm. Deep learning algorithms have more layers and use a ReLU which is efficient in reducing training time. It also uses a technique called LRN that tries to improve the memory capability and consumption of the data to the algorithm. Because of all the features involved and techniques used by VGG16, the collective output produces high accuracy. The objectives that were mentioned during the initial stage of the research process were all met, and a satisfactory result was achieved.

6.3.1 Comparison of Machine Learning Models implemented

Model Implemented	Accuracy	Precision	Recall	F1-Score
VGG16	54.54	1	0.54	0.70
Random Forest	50.4	0.50	1	0.66

We have used two algorithms, VGG16 and Random Forest. They have provided good accuracy and were perfectly working throughout the program. When comparing the accuracy, precision, and F-1 score of both the algorithms, the VGG16 algorithm is found to be better than the

Random. Forest algorithm with 54.54, 1 and 0.70 respectively. VGG16 makes use of the advantages of deep learning such as it adds weight to its layers which are effective in speeding up the training process.

This is better compared to the value of random forest whose accuracy comes to be 50.4% whereas the recall is higher than the VGG16 as it has 1 whereas VGG16 is lagging behind by 0.54.

7 Conclusion and Future Work

7.1 Conclusion

The output obtained while evaluating the evaluation metrics on both algorithms was satisfactory. Among them, VGG16 proved to be much better in detecting signature image forgery. This fuels the fact that implementing the machine learning with the minimum required number of data would give the expected output. Thus, the research was satisfying as it ticked all the objectives that were initially pointed out. Therefore, the research emphasises the fact that implementing more deep learning algorithms can provide high accuracy results compared to the convolutional machine learning algorithms.

7.2 Limitations

The major limitations that were involved in the research project were as follows:

-Limited time constraints on finishing the research project. While running the VGG16 model initially the number of epochs that were planned was 100. But it took nearly 76+ hours to run 100 epochs in the host computer. Due to the limitation of time constraints and the hardware and software limitation of host computer, running 100 epochs deemed a difficult task. The system crashed during various instances. Due to this, the epochs were reduced to 30 in which we found the accuracy of the matrix.

-Restricted network access on network devices.

-Non-availability of a greater dataset as it was difficult to get a dataset which has all the criteria matched along with licences and public access.

7.3 Future Work

The research conducted was able to implement VGG16 and random forest algorithms to find the forged signature data from the data set containing signature data both original and fake. The data was trained and tested on them without any difficulties. In future, we can try to analyse what can be done to increase the accuracy of the random forest algorithm as its accuracy was low when compared to VGG16. Due to the limitation of time the proposed 100 epochs could not be run which would have provided much better accuracy than the current result. In the

future, I am proposing to use the same technique with much better host machine with the required hardware and software and to have an additional period of time so that everything can be done much more efficiently. The use of larger datasets or multiple datasets can be implemented given more time in the future. With the help of more than one dataset, we can compare and contrast the difference in accuracy ranging among different datasets. Additionally, we can also try to update the project by including more algorithms. The number can be increased from two to five or six, which will help in testing and comparing the data and understanding the precision and accuracy of each algorithm. More output will mean more insight into how to tackle the issue of signature forgery efficiently. Thus, a better outlook on the subject can be attained in this way.

References

[1] Breiman, L., 2001. Random forests - machine learning. SpringerLink. Available at: https://link.springer.com/article/10.1023/A:1010933404324 [Accessed August 1, 2022].

[2] Marti, U.-V. & Bunke, H., 2002. The IAM-database: An English sentence database for offline handwriting recognition - International Journal on Document Analysis And Recognition (IJDAR). *SpringerLink*. Available at: https://link.springer.com/article/10.1007/s100320200071 [Accessed August 5, 2022]

[3] Kasodhan , R. &; Gupta, N., 2019. A new approach of digital signature verification based on BioGamal Algorithm Available at: https://www.researchgate.net/publication/335498239_A_New_Approach_of_Digital_Signature_Verification_based_on_BioGamal_Algorithm [Accessed August 7, 2022].

[4] Wang, L. &; Song, T., 2016. An improved digital signature algorithm and authentication protocols in cloud platform. IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/7796194 [Accessed August 7, 2022].

[5] Zulkarnain, Z., Rahim, M.S.M. &; Othman, N.Z.S., 2015. Feature selection method for offline signature verification. CORE. Available at: https://core.ac.uk/display/78378430?source=2 [Accessed August 7, 2022].

[6] Poddara, J., Parikha, V. &; Bhartia, S.K., 2020. Offline signature recognition and forgery detection using Deep Learning. Available at: https://www.researchgate.net/publication/340636385_Offline_Signature_Recognition_and_Forgery_Detection_using_Deep_Learning [Accessed August 8, 2022].

[7] L. A. Fakhiroh, A. Fariza and A. Basofi, "Mobile Based Offline Handwritten Signature Forgery Identification using Convolutional Neural Network," Available at: https://ieeexplore.ieee.org/abstract/document/9594019 [Accessed August 8, 2022].

[8] K , M. et al., 2021. A comparative study of transfer learning models for offline signature verification and forgery detection. Journal of University of Shanghai for Science and

Technology. Available at: https://jusst.org/a-comparative-study-of-transfer-learning-models-for-offline-signature-verification-and-forgery-detection/ [Accessed August 10, 2022].

[9] Foroozandeh, A. &; Hemmat, A.A., 2020. Offline handwritten signature verification and recognition based on Deep Transfer Learning. IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/9187481 [Accessed August 10, 2022].

[10] Kennard, D.J., Barrett, William A. & Sederberg, T.W., 2012. Offline signature verification and forgery detection using a 2-D geometric warping approach. IEEE Xplore. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6460976 [Accessed August 10, 2022].

[11] Hameed, M.M. et al., 2021. Machine learning-based offline signature verification systems: A systematic review. Signal Processing: Image Communication. Available at: https://www.sciencedirect.com/science/article/pii/S0923596521000047 [Accessed August 10, 2022].

[12] Al-Omari, Y.M., Huda Sheikh Abdullah, S.N. &; Omar, K., 2011. State-of-the-art in offline signature verification system. IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/5976912 [Accessed August 11, 2022].

[14] Ghanim, T.M. &; Nabil, A.M., 2018. Offline signature verification and forgery detection approach. IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/8639420 [Accessed August 12, 2022].

[15] Gaikwad, P. et al., 2021. Handwritten signature verification system using machine ... - IJARIIE. Available at: https://ijariie.com/AdminUploadPdf/Handwritten_Signature_Verification_System_using_machine_Learning_Approach__A_Review_of_Literature_ijariie14036.pdf [Accessed August 11, 2022].

[16] Hameed, M.M. et al., 2021. Machine learning-based offline signature verification systems: A systematic review Available at: https://www.researchgate.net/publication/348475222_Machine_learning-based_offline_signature_verification_systems_A_systematic_review [Accessed August 12, 2022].

[17] M. Arathi & A. Govardhan., 2014. An efficient offline signature verification system - ijmlc.org. Available at: http://www.ijmlc.org/papers/468-A1001.pdf [Accessed August 13, 2022].