******

***Dissertation on***

**“Detection of Deepfake Videos”**

*Submitted in partial fulfilment of the requirements for the award of degree of*

**Bachelor of Technology**

**in**

**Computer Science & Engineering**

**UE20CS390A – Capstone Project Phase - 1**

***Submitted by:***

|  |  |
| --- | --- |
| **Thejas N U**  **Mohammed Anas Danish**  **Abbu Bucker Siddique**  **Hithesh Dinesh Nayak** | **PES1UG20CS606**  **PES1UG20CS614**  **PES1UG20CS617**  **PES1UG20CS646** |

*Under the guidance of*

|  |
| --- |
| **Dr. Mamatha H. R.**  Professor |

**January - May 2023**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

FACULTY OF ENGINEERING

**PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)

100ft Ring Road, Bengaluru – 560 085, Karnataka, India



**PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)

100 Feet Ring Road, Bengaluru – 560 085, Karnataka, India

**FACULTY OF ENGINEERING**

**CERTIFICATE**

*This is to certify that the dissertation entitled*

**‘Detection of Deepfake Videos’**

*is a bonafide work carried out by*

|  |  |
| --- | --- |
| **Thejas N U**  **Mohammed Anas Danish**  **Abbu Bucker Siddique**  **Hithesh Dinesh Nayak** | **PES1UG20CS606**  **PES1UG20CS614**  **PES1UG20CS617**  **PES1UG20CS646** |

in partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE19CS390A) in the Program of Study - **Bachelor of Technology in Computer Science and Engineering** under rules and regulations of PES University, Bengaluru during the period Jan. 2022 – May. 2022. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

|  |  |  |
| --- | --- | --- |
| Dr. Mamatha H R  Professor | Dr. Shylaja S S  Chairperson | Dr. B K Keshavan  Dean of Faculty |

**External Viva**

|  |  |
| --- | --- |
| **Name of the Examiners**  **1.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Signature with Date**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_    **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

**DECLARATION**

We hereby declare that the Capstone Project Phase - 1 entitled **“Detection of Deepfake Videos”** has been carried out by us under the guidance of Dr.Mamatha H R, Professor and submitted in partial fulfilment of the completion of sixth semester of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2022. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

|  |  |  |
| --- | --- | --- |
| **PES1UG20CS606** | **Thejas N U** |  |
| **PES1UG20CS614** | **Mohammed Anas Danish** |  |
| **PES1UG20CS617** | **Abbu Bucker Siddique** |  |
| **PES1UG20CS646** | **Hithesh D N** |  |

**ACKNOWLEDGEMENT**

We would like to express our gratitude to Dr.Mamatha H R, Department of Computer Science and Engineering, PES University, for her/his continuous guidance, assistance, and encouragement throughout the development of this UE19CS390A - Capstone Project Phase – 1.

We are grateful to the project coordinator, Dr. Priyanka H., all the panel members & the supporting staff for organizing, managing, and helping the entire process.

We take this opportunity to thank Dr. Shylaja S S, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support we have received from her.

We are grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, Dr. B.K. Keshavan, Dean of Faculty, PES University for providing us various opportunities and enlightenment during every step of the way.

Finally, this project could not have been completed without the continual support and encouragement we have received from our family and friends.

**TABLE OF CONTENT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chapter No.** | **Title** | | **Page No.** | |
|  | | **ABSTRACT** | | 8 |
|  | | **PROBLEM STATEMENT** | | 9 |
|  | | **LITERATURE REVIEW** | | 10 |
|  | | **DATA** | | 18 |
|  | | 4.1 Deepfake Detection Challenge | | 18 |
|  | | 4.2 FaceForensics++ | | 19 |
|  | | 4.3 CelebDF | | 19 |
|  | | 4.4 DeepFake Image-high-quality | | 20 |
|  | | 4.5 Summary of datasets | | 20 |
|  | | **SYSTEM REQUIREMENTS SPECIFICATION**  5.1 Introduction  5.1.1 Project Scope  5.2 Product Perspective  5.2.1 Product Features  5.2.2 Operating Environment  5.2.3 General Constraints, Assumptions and Dependencies  5.2.4 Risks  5.3 Functional Requirements  5.4 External Interface Requirements  5.4.1 User Interfaces  5.4.2 Hardware Requirements  5.4.3 Software Requirements  5.4.4 Communication Interfaces  5.5 Non-Functional Requirements  5.5.1 Performance Requirements  5.5.2 Safety Requirements  5.5.3 Security Requirements  5.6 Other Requirements | | 22  22  22  22  22  22  22  23  23  24  24  24  24  24  24  24  24  25  25 |
|  | | **SYSTEM DESIGN**  6.1 Introduction  6.2 Current System  6.3 Design Considerations  6.3.1 Design Goals  6.3.2 Architecture Choices  6.3.3 Constraints, Assumptions and Dependencies  6.4 High Level System Design  6.4.1 Steps  6.4.2 Block Diagram  6.4.3 High Level Design Diagram  6.5 Design Details  6.5.1 Novelty  6.5.2 Innovativeness  6.5.3 Interoperability  6.5.4 Performance  6.5.5 Security  6.5.6 Reliability  6.5.7 Maintainability  6.5.8 Portability  6.5.9 Legacy to Modernization  6.5.10 Reusability  6.5.11 Application Compatibility | | 26  26  26  26  26  26  27  28  28  28  29  29  29  29  29  29  30  30  30  30  30  30  31 |
|  | | **IMPLEMENTATION AND PSEUDOCODE**  7.1 Code  7.2 Output | | 32  32  33 |
|  | | **CONCLUSION OF CAPSTONE PROJECT PHASE-1** | | 34 |
|  | **PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2** | | | 35 |
| **REFERENCE/ BIBLIOGRAPHY**  **APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS** | | | | 36  38 |
|  | | | |  |
|  | | | |  |

**LIST OF FIGURES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Figure No.** | | **Title** | **Page No.** | |
| Figure 1 | Statistics of Datasets | | | 20 |
| Figure 2 | Comprehensive overview of existing datasets | | | 21 |
| Figure 3 | Block Diagram | | | 28 |
| Figure 4 | High Level Design Diagram | | | 29 |
| Figure 5  Figure 6 | Output original image  Output cropped image | | | 33  33 |

|  |  |  |
| --- | --- | --- |
|  |  |  |

**1. ABSTRACT**

Deepfake videos have emerged as a major threat to cybersecurity and digital forensics in recent years, with the potential to cause significant harm by manipulating public opinion, spreading fake news, and damaging the reputation of individuals and organizations. To combat this threat, researchers and practitioners have developed various techniques and algorithms for deep fake detection.

Through a thorough literature review, this report finds that deep fake detection is a challenging problem that requires a multidisciplinary approach combining “computer vision, machine learning, and forensic analysis”. While current deepfake detection techniques have shown promising results, they are still limited in their ability to detect highly realistic and sophisticated deep fakes. Moreover, the effectiveness of these techniques may depend on the “specific characteristics of the deep fake and the context in which it is deployed”.

We aim to develop a DeepFake Detection model that can differentiate artificially created videos using deep learning from the real videos. This model uses a combination of deep learning techniques to process the input video and based on the data extracted, it will classify the video either as deep fake or real. This model can be useful in various fields such as “social media, news platforms, news channels'', etc., where they can verify if the video is real before it reaches the people and starts spreading virally.

Our model will have 2 components, one for extracting features and data from the frames and another one for processing the extracted data using deep learning techniques. The deep learning component will classify the video as either deepfake or not based on the data given by the data extraction layer. Our model will be robust against “image orientation”, “noise in the data” and “quality of the image”.

**2. PROBLEM STATEMENT**

With the rapid advancements in deep learning and computer vision technologies, the creation and proliferation of deepfake videos have become more accessible than ever before. Deepfakes are “synthetic videos or images that are created using deep learning algorithms”, and they have the potential to cause significant harm by “manipulating public opinion, spreading fake news, and damaging the reputation of individuals and organizations”. As a result, the detection and prevention of deepfakes have become critical challenges in the field of cybersecurity and digital forensics.

Our aim is to develop a model that can effectively detect deepfake videos(only the Visual Part), even when they are realistic and sophisticated, while minimizing the false positives and false negatives, which are also robust against different deepfake evasion methods and is scalable to detect different types of deepfakes like “face swapping”, “facial manipulations”, “identity swap”, “face reenactment”, “attribute manipulation”, and “entire face synthesis”, etc.

**3. LITERATURE REVIEW**

[1] “An explainable deepfake detection framework on a novel unconstrained dataset” research paper proposes a novel deepfake detection framework that uses “deep learning techniques” to detect deep fake videos. The proposed framework is evaluated on a new unconstrained dataset. Dataset created and used here is “DeepFake Image-high-quality (DFIM-HQ)” dataset which contains “70000 real images and 70000 fake images” which have been created without any race,gender and age bias. It also includes various scenarios such as image orientation changes, low quality images and illumination degradations. The features learnt from the new high quality dataset include visual artifacts, such as unnatural skin tones and inconsistent head poses, while the features are learnt using Grad-CAM technique and the features will be trained on deep learning based model Meso-Net or MesoInception-Net. However, Mesonet and Meso Inception Net require a large amount of labeled data to be trained effectively. Acquiring and labeling data can be time-consuming and costly and they are not robust against small changes in the input data, they are prone to adversarial attacks. The accuracy achieved here was 94.52% for Meso-Net and 99.87% for MesoInception-Net.

[2] “Deep fake Detection using Biological Features: A Survey” presents a comprehensive survey of deepfake detection using biological features. The survey covers various techniques that utilize biological features such as “facial expressions, eye movements, heart rate, and brain waves to detect deep fake videos”. This paper mainly discusses various biological features that have been used for deep fake detection, including “facial expressions, eye movements, heart rate, and brain waves”. These features are extracted using sensors and devices such as “cameras, eye trackers, electrocardiogram (ECG), and electroencephalogram (EEG)”. Models used in the papers mentioned in the survey are “Support Vector Machines (SVMs)”, “Xception net”, “LSTM” and “CNN”.

[3] “Multi-attentional Deepfake Detection” research paper mentions a method that leverages attention mechanisms to detect deep fake videos. The proposed method consists of two stages: “feature extraction and classification”. The feature extraction stage uses a “deep convolutional neural network” to extract features from the input video frames, while the classification stage uses a “multi-attentional mechanism” to weigh the importance of different regions in the video frames for the final classification decision. The proposed method uses a multi-attentional mechanism to weigh the importance of different regions in the video frames for the final classification decision.

The datasets used here are “FaceForensics++, CelebDF and DFDC”. Drawback of the proposed model is there is no detailed analysis of the complexity of the proposed model architecture and also it is unclear how the model performs when there are changes in the lighting conditions or camera angles in the input videos. The accuracy achieved here is 86.95% on low quality images and 96.37% on high quality images when Xception-net was used and it became 88.69% for low quality images, 97.60% for high quality images when Efficient-B4 was used.

[4] “Deepfake video detection using convolutional vision transformers” proposes a deep fake video detection method that uses a “Convolutional Vision Transformer (CVT)”. This model was trained on the “DFDC” dataset. The proposed method aims to detect deepfake videos by analyzing the visual features of the input frames and capturing the “temporal dynamics” between them. This model consists of a “convolutional layer that extracts spatial features” and a “transformer layer that captures temporal dynamics' ' between the frames. The method also incorporates a “self-attention mechanism” to capture the most relevant features for deep fake detection. However the model is sensitive to resolution of images due to the use of “Vision Transformer '' and it has a loss value of 0.32. Accuracy on the validation set is 87.25%.

[5] “Unsupervised learning-based framework for deep fake video detection” proposes a method which is different from the previously mentioned methods. They tried “unsupervised learning” which nobody has done till now. Authors have used “FaceForensics++” dataset to train this model. The proposed model has two components: “ Photo-Response Non-Uniformity(PRNU)” and “Noise Print”. First the PRNU runs on multiple video frames for “multi classification”, later using the source ,which is obtained from the previous component there will be a “binary classification” using noiseprint mechanism. The AUC value for this model is 0.925 for original images and 0.887 for cropped images.

[6] The authors of “A Performance Enhancement of Deepfake Video Detection through the use of a Hybrid CNN Deep Learning Model” propose a “hybrid convolutional neural network” model for deep fake video detection, which combines the strengths of both “traditional CNNs and residual networks (ResNets)” to enhance performance. Authors here have selected the “DFDC” dataset to train and test. The proposed model consists of a feature extraction module, a ResNet module, and a classification module. The model used here is a combination of “InceptionResnet v2” and

“Xception-net” to form a Hybrid CNN. But, both techniques used in hybrid models don't work well on low quality images. It is mentioned in the paper that the F1-Score obtained is 0.98.

[7] The authors of “Sharp multiple instance learning for deep fake video detection” research paper propose a deep fake video detection method called “Sharp Multiple Instance Learning (SMIL)” that is based on the “Multiple Instance Learning (MIL)” framework. Here authors have created the dataset called “FaceForensics Plus with Mixing samples (FFPMS)” by mixing their own samples with already existing and widely used FaceForensics dataset. The proposed method aims to learn a “sharp decision boundary” between real and fake video instances. The proposed SMIL method is based on the MIL framework and utilizes a combination of different classifiers to learn the decision boundary between real and fake video instances. The method first extracts the visual features from the video and then uses a “bag-level pooling” operation to obtain a single feature representation for each video. The final decision is made using a combination of “SVM, Random Forest, and AdaBoost'' classifiers. AUC obtained here on the test set is 0.9714.

[8] “Improving Deep Fake Detection Using Dynamic Face Augmentation” proposes a pre-processing method, in the form of dynamic face cutouts using the face landmark information in order to dynamically cutout regions from the face and use it for training in order to avoid overfitting of models.The datasets used were “DFDC, FaceForensics++, Celeb-DF”. The paper mainly focuses on preprocessing training data by using face landmark information and also deep fake locations to dynamically select the best cutout regions. The model performs 2 types of cutouts: “Sensory group removal” and “ Convex hull removal”. This data augmentation technique can be introduced to any existing deepfake detection pipeline without need for any significant modifications. The models implemented by the author included a combination of dynamic face cutout along with “EfficientNet-B4 and XceptionNet”. The proposed model was able to significantly improve the model accuracies giving an accuracy of 99.05% for EfficientNet-B4 and 95.6% for the same dataset using XceptionNet. The proposed model did have a few drawbacks which include inability to properly map face landmark information when image orientations are distorted/ noisy. The model also adds extra computation to already compute-heavy CNN based models.

[9] The paper “Fighting deepfake by exposing the convolutional traces on images'' proposes an Expectation-Maximization algorithm which can be trained to detect and extract fingerprints left behind by “Generative Adversarial Networks(GANs)” called “Convolutional Traces(CT)”. These Convolutional Traces can be used to not only to classify images as deep fake or real but also help identify which GAN model might have produced it. One of the key features of the proposed model is that, since it is not a deep learning model, it can easily be run on computers with limited cpu power. Using “Expectation-Maximization” steps, the model extracts the convolutional traces from the given image which can then be passed through a simple classifier like random forest. The model displayed a good accuracy of 87.1% on average for datasets including “STYLEGAN, STYLEGAN2, FACE FORENSICS ++”, also data generated from “STARGAN, ATTGAN, GDWCT, CYCLEGAN, PROGAN, IMLE”. The model displayed the highest accuracy of 99.32% on STYLEGAN2 dataset using a 7x7 kernel. One of the major drawbacks of the model is that since EM is an iterative process, it requires a lot of time in order to extract CTs from an image making it very difficult to implement in a video deep fake identification task as the number of frames will be very high.

[10] “Transformer-based feature compensation and aggregation for deepfake detection” suggests two main methods to compensate for the two common drawbacks of using a transformer: “Lack of local information” and “Failure to capture hierarchical data”. The first method is to make use of a “Local Compensation Block(LCA)” in order to capture local information effectively and then combine it with global features using “Global-Local Cross Attention(GLCA)”. And in order to aggregate hierarchical data, the author suggests using Multi-head Clustering Projection and Frequency-guided Fusion Module(FFM). The model archives a high accuracy of 97.88% on FaceForensics++ dataset which is very close to SOTA models. The major drawback of the given model is that it suffers heavily whenever it has to work with noisy or low quality images.

[11] “Combining efficientnet and vision transformers for video deep fake detection” suggests using vision transformers along with a feature extract such as “EfficientNet” to carry out classifications of deep fake images. Inorder to determine if faces are manipulated, the model first extracts the faces from the image using MTCNN which is a SOTA face detector. the convolutional cross ViT has two branches S-branch and L-branch which handle local and global perception respectively. After extraction the results from both S-branch and L-branch are combined to produce the final output. The Convolutional Cross Vit produced an average accuracy of 94.7% on FaceForensics++ and DFDC datasets. The author’s papers suggests an alternate method which makes use of vision transformers which are considered good for image data, but fails to produce good accuracy compared to SOTA models.

[12] “Generalization of Forgery Detection With Meta Deepfake Detection Model” focuses on detecting deep fakes using a meta deep fake detection model.The three main data sets used were “FaceForensics++,CelebDF, DFDC”. The model can be evaluated on unseen domains" without requiring any updates, following its training on a set of source domains. To enable this, the researchers employ data preprocessing techniques and apply block shuffling transformations. Experiments on various datasets show that the proposed method outperforms existing forgery detection approaches, particularly on previously unseen forgery types. The model achieves an average precision of 0.94 on the FaceForensics++ dataset and 0.89 on the Celeb-DF dataset, outperforming the other methods by a significant margin. Further experiments on a new dataset of unseen forgery types demonstrate the model's ability to generalize to previously unseen scenarios.Overall, the proposed method provides a promising solution to the challenge of forgery detection in the age of deepfakes, demonstrating the potential of meta-learning techniques for improving generalization in this domain.

[13] “Detection Of Deep Fake Videos Using Long-Distance Attention” describes a new approach for detecting deep fake videos using “long-distance attention mechanisms”. It focuses on analyzing the generated video frames for signs of manipulation using the long distance attention mechanism.It focuses on specific regions of video frames that are most likely to contain manipulations.The two main Data sets used are “FaceForensics++ and CelebDF”. The proposed spatial-temporal model has two essential components,first one is the “spatial attention component” and the second one is the “temporal attention component”. It allows to focus on specific regions of the generated video frames that are most likely to have manipulations using the attention maps which is in contrast to the traditional attention mechanisms that only consider local regions of video frames. They have also used a combination of supervised and unsupervised learning techniques to train the built algorithm. However this model fails to work with “abnormal inputs” such as faces, sideways faces. It also generates abnormal attention maps for few inputs.The overall performance is extremely good as it has claimed to achieve 99.13 ACC and 99.87 AUC.

[14] “DeepFake Detection Based on Discrepancies Between Faces and Their Context '' proposes a model to detect deepfakes using two main aspects namely “face region " and"context ". The three main

datasets used are “FaceForensics++, CelebDF,DFDC, VGGFace2 ''. The researchers introduced a unique method to detect fake images by comparing the inner face area, which is directly manipulated, with the outer context that remains unaltered by all known face manipulation methods. The approach involves two networks: “face identification network” and “context recognition network” that considers the face context. Experiments on several publicly available datasets show that the proposed method is robust against different types of manipulations. Even when the manipulated face is in motion and the context changes dynamically this model performs well.

[15] “Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection”.The models used are “Xception Net for encoding” and then “Bi directional LSTM for deep fake detection”. This paper uses a “3-Phase architecture” where the first phase would be to use the “Dlib library” to take only the region of interest that is the face or body. The second phase would be the neutral network which will use the weights and other parameters to represent the Image input to a latent representation.This includes the Xception Net which would be responsible for filtering only the spatial information and masking the temporal Information. then the output from that would be passed onto the 2-layered bi directional LSTM Model which would then generate the output which will be passed to the next phase.The Third Phase would contain the outputLayer of the class layer. The 3rd phase will also contain the probability distributions of real and fake detections that have happened and then we have two models in the last phase so we have two loss functions which will be used one would be cross entropy between the real and fake output from neurons and other will be “KL divergence loss” for the probability distribution. The datasets used are Face Forensics ++ and Celeb DF. In the accuracy part we have 50 to 98% from the class layer output and 97 to 99% from the probability distribution of the two real and fake.

[16] “Learning Features of Intra-consistency and Inter-diversity: Keys towards Generalizable Deep Fake Detection ”.In this paper the models used are “Deep CNN”. The architecture is divided into two phases, those would be first “self supervised pre-training with masking” which is a type of unsupervised learning technique where initially all the input images are passed to a masking layer which masks redundant or unneeded parts of the image, later this masked image is passed onto a “swin-L encoder” , this is ViT bases encoder. This encoder would generate a latent representation of the image in concern which the CNN model would take to predict. The second phase of the architecture contains the fine tuning of the already generated CNN model with labeled data . This is done to boost its accuracy on predictions.The

datasets used are “Celeb-DF and DFDC”. This model fails to enhance the forges that are generated due to the entire face or head synthesis methods. Celeb-DF :91% AUC ,SBI: 87.32%, ICT: 83.05%.

[17] “Detecting Real-Time Deep-Fake Videos Using Active Illumination”, this paper deals with real time videos which come from the feed of for example the webcam of a client sitting in for an online exam. The method that they use is to convert the video into frames using “DLib” and then convert the image to “HSI components” and take the hue part. The illumination that they use is from a periodic light source, then the model uses this data and the predicted hue from the formula that they gave to draw a correlation between those two and then using that data they can tell whether the real time video is real or artificially generated. The accuracy that they got was 94.8%.

[18] “Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection”, this is a paper which compares the generally used deepfakes detection methods and the “semi-supervised-GAN” architecture in detecting deep fakes.The feature of the body that they consider in a person or object of interest would be the general physiological signs which could be the “periodic eye blinks” and the “smile” that the subject gives facial region is identified by the face recognition libraries such as Dlib,etc. The general model that they used is a “CNN with LSTM” combo. Some of the researchers use spatial information with the already used deep pixel wise information that is the temporal information. Coming to the GAN model used, the training is done using two types of data majority of unsupervised data methods and minority of supervised data which is perfect of the GAN as the GAN has discriminator part of the code which is the part which does unsupervised learning and the presence of the supervised data will help it to become more perfect during validation phase.The accuracy that they have got is following direct relation to the size of dataset used for training for 40,000 images the model gives 92.3 % accuracy and for 10,000 images it gives 86.7% accuracy. The paper authors say that the same high accuracy could be achieved by a smaller dataset size by using some encoders and improving on the architecture.

[19] “Short And Low Resolution Deep Fake Video Detection Using CNN”, this paper explains various methods of deepfake detection ,first they talk about “Pure CNN” model which is only applicable to deep fake images and then we can see the CNN used with a particular type of RNN model which is called the LSTM which takes into account the past information , this model can be used to detect deep fake videos. Next they use the “CNN + LSTM + inception V3” where there is a encoder being used which converts the data into a latent representation which is a transformation of the actual image into another form which can be interpreted easily by the model.Next they use the pretrained weights from “ResNet-50 or VGG-16”, they are doing transfer learning here. Different variants of the pretrained models are three which the paper talks about which are as follows: “InceptionResNetV2 , MobileNet , and DenseNet121”. The overall accuracy that they got is 93.7% with InceptionResNetV2, 94.93% with MobileNet and 93.86% DenseNet121.

[20] Author of the research paper “Deepfakes Creation and Detection Using Deep Learning ” talks about the “MesoNet” and how it plays an important role in DeepFake Detection. In some cases where the image has been compressed due to storage issue,etc the information that the image is deepfake is lost and a normal network cannot identify those “micro alterations” ,This is where MesoNet is useful, due to its “small number of layers” and “many neurons in the hidden layers” this model is efficient in finding such minute manipulations. The model they used has “4 hidden layers” with “consecutive 2D-conv and maxpool”. The Accuracy that their model gave was “over 80% for 5000 images of training”. The Gaps that the paper has is that the MesoNet created cannot detect the face swaps when the person sees directly towards the camera.

**4. DATA**

Deepfake datasets are collections of “images, videos, and audio recordings that are used to train algorithms to create convincing fake media”. These datasets typically consist of real media that has been manipulated using deep learning techniques to produce fake content. Deepfake datasets have become extremely popular in recent years due to the growing concern about the misuse of this technology.

**4.1 DeepFake Detection Challenge**

The “DeepFake Detection Challenge (DFDC)” dataset is a large-scale dataset created by Facebook to aid in the development of deepfake detection technologies. The dataset consists of “over 100,000 videos, both real and manipulated”, and is divided into training, test and validation sets. The manipulated videos were created using various deepfake creation methods, including GANs and autoencoders, to generate realistic fake media.

The dataset was created by collecting videos from publicly available sources, such as YouTube, and then manipulating them to create deepfakes. The manipulation process involved altering the faces of individuals in the video to create fake media. The manipulated videos were then labeled as "fake" and included in the dataset, along with the corresponding real videos.

The DFDC dataset includes a diverse range of subjects, lighting conditions, and camera angles to challenge deepfake detection algorithms. The dataset also includes a track for both image and video-based deepfakes, allowing researchers and developers to evaluate the performance of their deepfake detection algorithms on both types of media.

To check the performance of different deepfake detection algorithms on the dataset, Facebook organized a competition in which participants developed and submitted their algorithms for evaluation. The competition included a leaderboard that ranked the performance of each algorithm on the test set of the dataset. The DFDC dataset has spurred the development of new deepfake detection algorithms and helped researchers and developers to better understand the complexities of detecting manipulated videos.

**4.2 FaceForensics++**

The “FaceForensics++” dataset is a large-scale dataset created to aid in the development of deepfake detection technologies. It contains “363 real and 3,068 deepfake” videos. The manipulated videos were created using four different deepfake creation techniques: “DeepFake”, “Face2Face”, “FaceSwap”, and “NeuralTextures”.

The DeepFake method uses GANs to generate manipulated images, while the Face2Face method uses facial tracking to animate the expressions of one face onto another. The FaceSwap method uses deep learning algorithms to swap the faces of two individuals, and the NeuralTextures method uses GANs to synthesize facial features onto a target image.

The dataset includes multiple variations of each video, such as different compression levels and resolutions, to simulate real-world scenarios where deepfakes may be found. Each video is labeled as either real or manipulated, and the manipulated videos are further labeled with the creation method used.

In addition to the videos, the FaceForensics++ dataset also includes precomputed image features for each video, allowing for faster training and validation of deepfake detection models. This dataset has been widely used by researchers and developers to develop and test new deepfake detection methods and technologies.

**4.3 CelebDF**

The “Celeb-DF(v2)” dataset is a large-scale dataset created for the purpose of training and testing deepfake detection algorithms. It includes “590 videos” which are collected from “Youtube” and “5639 synthesized deepfake videos”. The manipulated videos were created using several deepfake creation methods, including “DeepFake”, “FaceSwap”, and “Face2Face”.

The DeepFake method uses GANs to generate manipulated images, while the FaceSwap method uses deep learning algorithms to swap the faces of two individuals, and the Face2Face method uses facial tracking to animate the expressions of one face onto another.

The Celeb-DF(v2) dataset was created by collecting videos of celebrities from publicly available sources, such as YouTube, and then manipulating them to create deepfakes. The dataset also includes various metadata for each video, such as the video's resolution and frame rate, to provide additional information for deepfake detection algorithms.

**4.4 DeepFake Image-high-quality (DFIM-HQ)**

“DeepFake Image-high-quality (DFIM-HQ)” has been created with the intention of making deepfakes dataset free from any kind of bias or constraints. Most of the datasets show bias towards particular gender,age and race. To overcome this, DFIM-HQ has been created created without any race,gender and age bias, where it includes pictures of people from young age to old age, all races and all genders. It also takes care of various scenarios such as image orientation changes, low quality images and illumination degradations. The dataset contains “70000 real images and 70000 fake images”, which have been divided further as train,validation and test datasets. Train in each contains 52500 images, Validation contains 10500 images and Test contains 7000 images.

**4.5 Summary of datasets**

|  |
| --- |
| **Fig1. Comparison of datasets** |
| Fig1. Statistics of Datasets |

|  |
| --- |
|  |
| Fig2. Comprehensive overview of existing datasets |

**5. SYSTEM REQUIREMENTS SPECIFICATION**

**5.1 Introduction**

**5.1.1 Project Scope**

* Our project aims to develop a system that can detect deepfakes in videos with high accuracy, which is robust against image orientation changes, image quality and occlusion.
* The project scope includes features such as video analysis and classification of video.
* The target audience for the system is media companies, law enforcement agencies, and other organizations that need to detect deepfakes in videos.

**5.2 Product Perspective**

**5.2.1 Product Features**.

* The system will analyze video content to identify signs of manipulation, such as inconsistent facial expressions or lighting, and use machine learning algorithms to detect deepfakes with high accuracy.
* When a deepfake is detected, the system will trigger an alert and generate a report with details about the video and the detection results.

**5.2.2 Operating Environment**

* The deepfake detection system will require high-performance hardware, including a powerful CPU and GPU, and a large amount of memory and storage.
* The software requirements include operating systems such as Windows and Linux, and deep learning frameworks such as TensorFlow and PyTorch.

**5.2.3 General Constraints, Assumptions and Dependencies**

* The project depends on the availability of diverse and reliable data sources
* The project’s accuracy depends on the quality of the training data
* The project assumes that there will be a large number of deepfakes in the wild, and that the system will need to be updated regularly to keep up with new types of deepfakes.
* The project assumes that deepfake techniques used in the dataset are similar to those used in the real world
* The project assumes that the entire length of the video will either be real or deepfake
* The project may be constrained by limited resources, such as time and funding, and dependencies on third-party software libraries and frameworks.

**5.2.4 Risks**

* The system may produce false positives, which may lead to innocent users being falsely accused of creating or sharing deepfake
* The system may infringe on user privacy by analyzing and storing user data, leading to privacy concerns and potential legal issues
* The system may be vulnerable to adversarial attacks, where attackers may attempt to deceive the system by creating deepfakes that can evade detection
* Other risks include the potential for malicious actors to develop new types of deepfakes that the system cannot detect, and the risk of false confidence in the system's detection results.

**5.3 Functional Requirements**

* The deepfake detection system must be able to analyze videos for signs of manipulation, such as inconsistent facial expressions or lighting, and use machine learning algorithms to detect deepfakes with high accuracy.
* Data Preprocessing: The system should be able to preprocess the dataset to extract relevant features from the images or videos to train the model
* Model Training: The system should be capable of training deep learning models to detect deepfakes
* Model Testing: The system should be able to test the model on a separate dataset to evaluate its performance
* Real-Time Detection: The system should be able to perform real-time deepfake detection on images and videos as they are uploaded
* Accuracy: The system should have a high accuracy rate in detecting deepfakes and minimizing false positives

**5.4 External Interface Requirements**

**5.4.1 User Interfaces**

The deepfake detection system will have a web-based user interface that allows users to upload videos for analysis and view detection results.

**5.4.2 Hardware Requirements**

The deepfake detection system will require high-performance hardware, including a powerful CPU and GPU, and a large amount of memory and storage.

**5.4.3 Software Requirements**

The software requirements include operating systems such as Windows and Linux, and deep learning frameworks such as TensorFlow,PyTorch,DlibNumpy,Pickle and OpenCV.

**5.4.4 Communication Interfaces**

The deepfake detection system will require communication interfaces, such as APIs or network connections, to integrate with other systems or applications.

**5.5 Non-Functional Requirements**

**5.5.1 Performance Requirements**

* The deepfake detection system must be able to analyze videos quickly and accurately, with a “high detection rate” and “low false positive rate”.
* The system must also be scalable, able to handle large volumes of video content and multiple users simultaneously.

**5.5.2 Safety Requirements**

* The deepfake detection system must not cause harm or endanger users in any way.
* It must not leak user data or use any user data without their consent

**5.5.3 Security Requirements**

The deepfake detection system must ensure the confidentiality, integrity, and availability of user data, and must be designed to prevent cyber attacks.

**5.6 Other Requirements**

The deepfake detection system must comply with legal and regulatory requirements, such as data privacy laws and intellectual property rights.

**6. SYSTEM DESIGN**

**6.1 Introduction**

This section provides an overview of the proposed system design for deepfake detection in videos, with a focus on facial visualization part .

**6.2 Current System**

Existing deepfake detection systems have limitations in detecting deepfakes that primarily focus on facial visualization. These techniques are not that effective in detecting sophisticated deepfakes. They very much rely on detecting artifacts, such as inconsistent lightning or blurriness that are introduced during the manipulation process.However, with the use of advanced machine learning algorithms, it is becoming increasingly difficult to detect these artifacts and deepfakes are becoming more realistic and sophisticated.

There is a need for a deepfake detection system that specifically focuses on the facial visualization part of the videos and uses advanced machine learning techniques to detect sophisticated deepfakes in real-time and at scale.The current system is also proven to fail when there is a change in orientation.The proposed system aims to address these limitations and provide a more effective and reliable solution for deepfake detection.

**6.3 Design Considerations**

**6.3.1 Design Goals**

* Higher accuracy even for the low quality images
* Real-time performance for deepfake videos
* Ability to detect sophisticated deepfakes
* Good model performance even when the the face orientation is bad

**6.3.2 Architecture Choices**

* Convolutional Neural Networks (CNNs): CNNs have been shown to be effective at detecting deepfakes in videos. One approach is to use a pre-trained CNN on image recognition tasks to extract features from individual frames of a video, and then use a

recurrent neural network (RNN) to classify the video as real or fake based on the temporal sequence of these features.

* Two-stream CNNs: Two-stream CNNs use separate networks to analyze spatial and temporal information in videos. One stream processes individual frames, while the other analyzes motion between frames. These streams are then combined to classify the video as real or fake.
* Siamese Networks: Siamese networks are a type of neural network architecture that compares two inputs and determines whether they are similar or dissimilar. For deepfake detection, a Siamese network can be trained to compare pairs of frames from a video, and classify the video as real or fake based on the similarity of the frames.
* Gradient-weighted Class Activation Mapping(Grad-CAM): is a visualization technique used to understand which regions of an image a convolutional neural network (CNN) is focusing on when making a particular classification decision. It works by taking the gradient of the class score with respect to the feature maps of a CNN layer and then weighting these feature maps by their gradient values. This produces a heat map that highlights the regions of the input image that are most relevant to CNN's decision. Grad-CAM has been widely used in computer vision research and has applications in fields such as object detection, image segmentation, and visual question answering.
* Multi-Attentional Maps (MAM): It is a technique used to visualize the attention mechanisms in neural networks. MAM works by identifying the most relevant words or phrases in a given sentence or document, based on their relative importance to the task at hand. This is done by training a neural network to generate attention weights for each word in the input sequence, and then using these weights to create a heat map that highlights the words with the highest attention scores. MAM can be used to gain insights into how neural networks process.

**6.3.3 Constraints, Assumptions and Dependencies**

* The model assumes a uniform video data format.
* The model assumes that the entire length of the video will either be real or deepfake.
* The model’s accuracy depends on the quality of training data.
* The model cannot have a fixed size for the feature extraction kernel.
* The project depends on the availability of diverse and reliable data sources
* The project’s accuracy depends on the quality of the training data
* The project assumes that there will be a large number of deepfakes in the wild, and that the system will need to be updated regularly to keep up with new types of deepfakes.
* The project assumes that deepfake techniques used in the dataset are similar to those used in the real world
* The project assumes that the entire length of the video will either be real or deepfake
* The project may be constrained by limited resources, such as time and funding, and dependencies on third-party software libraries and frameworks.

**6.4 High Level System Design**

**6.4.1 Steps**

* Obtain datasets from references research
* Merge all the datasets,extract frames from videos and pre process it
* Pass the processed data through feature extraction model
* Pass the above obtained data to deep learning model
* Predict the output

**6.4.2 Block Diagram**

|  |
| --- |
|  |
| Fig3. Block Diagram |

**6.4.3 High Level Design Diagram**

|  |
| --- |
|  |
| Fig4. High Level Design Diagram |

**6.5 Design Details**

**6.5.1 Novelty**

* Effective on low quality images also
* Robust against bad orientation of face
* Works even when occlusion occurs

**6.5.2 Innovativeness**

The project can be considered innovative in its approach to detecting deepfakes in real-time or in its use of novel deep learning architectures.

**6.5.3 Interoperability**

The deepfake detection project can work with different forms of video files, such as mp4,mov,mkv,etc,. It will also be compatible with different platforms, such as mobile devices or cloud-based services. Ensuring interoperability may involve designing an appropriate data schema, creating adapters for different media formats, or using standardized APIs.

**6.5.4 Performance**

The performance of our model will be the same on low quality images or bad face orientations. The model will trained and optimized properly to get higher accuracies and faster detection speeds

**6.5.5 Security**

The deepfake detection project incorporates advanced security measures, such as using secure communication protocols and encryption, to prevent data leakage and protect against attacks. It also uses adversarial training and other techniques to make the model more robust to attacks and manipulations.

**6.5.6 Reliability**

The deepfake detection project is highly reliable, achieving consistent levels of accuracy across different types of deepfakes and media file formats. It is also designed to be easily scalable and adaptable to different contexts, ensuring reliability in production environments.

**6.5.7 Maintainability**

The deepfake detection project is designed with maintainability in mind, using modularized code and automated testing to make it easy to manage code changes and updates. It also incorporates version control tools and other software development best practices to ensure long-term maintainability.

**6.5.8 Portability**

The deepfake detection project is designed to be portable across different hardware and software environments.

**6.5.9 Legacy to Modernization**

The deepfake detection project is designed to transition from older, rule-based methods of detecting manipulated media to newer, machine learning-based methods. It uses algorithms to automatically generate labeled data and collects a large dataset of manipulated and unmanipulated media files to improve detection accuracy.

**6.5.10 Reusability**

The deepfake detection project is designed to be reusable across different applications and contexts, using standard APIs to make it easy to integrate the system with other software tools and platforms. It also incorporates custom adapters to handle different media formats and ensure compatibility with different applications.

**6.5.11 Application Compatibility**

The deepfake detection project is designed to be compatible with different types of applications, such as social media platforms and video editing software. It incorporates custom adapters to handle different media formats.

**7. IMPLEMENTATION AND PSEUDOCODE**

The below code shows the preprocessing part of the code, in which the data from the dataset is extracted and stored in different directories. The videos are converted to frames and they are dynamically selected and stored based on the number of frames in the video. Later the frames are cropped to get only the face part and they are stored in a separate directory

**7.1 Code**

****

**7.2 Output**

Original image:

|  |
| --- |
|  |
| Fig5. Output original image |

Cropped image:

|  |
| --- |
|  |
| Fig6. Output cropped image |

**8. CONCLUSION OF CAPSTONE PROJECT PHASE-1**

Phase-1 of Capstone Project named “DeepFake Detection For Videos” ensures the completion of below mentioned milestones:

* Completed research survey.
* We have finalised gaps to be worked upon.
* Collecting datasets.
* Basic preprocessing of the data.

**9. PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2**

In the phase-2 of the Capstone Project we are planning to work on the below mentioned points:

* Do more preprocessing on the datasets and combine the datasets.
* Test existing architectures by implementing them on the new combined dataset.
* Implementation of our new model.
* Testing, Improving the accuracy and Parameters fine tuning.
* Work on a research paper for publishing it in journals or conferences.

**REFERENCE/ BIBLIOGRAPHY**

[1] Mathews, Sherin, et al. "An explainable deepfake detection framework on a novel unconstrained dataset." Complex & Intelligent Systems (2023): 1-13.

[2] Patil, Kundan, et al. "Deepfake Detection using Biological Features: A Survey." arXiv preprint arXiv:2301.05819 (2023).

[3] Zhao, Hanqing, et al. "Multi-attentional deepfake detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[4] Wodajo, Deressa, and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer." arXiv preprint arXiv:2102.11126 (2021).

[5] Zhang, Li, et al. "Unsupervised learning-based framework for deepfake video detection." IEEE Transactions on Multimedia (2022).

[6] Ikram, Sumaiya Thaseen, Shourya Chambial, and Dhruv Sood. "A Performance Enhancement of Deepfake Video Detection through the use of a Hybrid CNN Deep Learning Model." International journal of electrical and computer engineering systems 14.2 (2023): 169-178.

[7] Li, Xiaodan, et al. "Sharp multiple instance learning for deepfake video detection." Proceedings of the 28th ACM international conference on multimedia. 2020.

[8] Das, S., Seferbekov, S., Datta, A., Islam, M. and Amin, M., 2021. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3776-3785).

[9] Guarnera, L., Giudice, O. and Battiato, S., 2020. Fighting deepfake by exposing the convolutional traces on images. IEEE Access, 8, pp.165085-165098.

[10] Tan, Z., Yang, Z., Miao, C. and Guo, G., 2022. Transformer-based feature compensation and aggregation for deepfake detection. IEEE Signal Processing Letters, 29, pp.2183-2187.

[11] Coccomini, D.A., Messina, N., Gennaro, C. and Falchi, F., 2022, May. Combining efficientnet and vision transformers for video deepfake detection. In Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III (pp. 219-229). Cham: Springer International Publishing.

[12] V. -N. Tran, S. -G. Kwon, S. -H. Lee, H. -S. Le and K. -R. Kwon, "Generalization of Forgery Detection With Meta Deepfake Detection Model," in IEEE Access, vol. 11, pp. 535-546, 2023, doi: 10.1109/ACCESS.2022.3232290.

[13] W. Lu et al., "Detection of Deepfake Videos Using Long-Distance Attention," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3233063.

[14] Y. Nirkin, L. Wolf, Y. Keller and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6111-6121, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3093446.

[15] H. A. Khalil and S. A. Maged, "Deepfakes Creation and Detection Using Deep Learning," 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 2021, pp. 1-4, doi: 10.1109/MIUCC52538.2021.9447642.

[16] H. Chen, Y. Lin, B. Li and S. Tan, "Learning Features of Intra-consistency and Inter-diversity: Keys towards Generalizable Deepfake Detection," in IEEE Transactions on Circuits and Systems for Video Technology, 2022, doi: 10.1109/TCSVT.2022.3209336.

[17] C. R. Gerstner and H. Farid, "Detecting Real-Time Deep-Fake Videos Using Active Illumination," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 53-60, doi: 10.1109/CVPRW56347.2022.00015.

[18] J. John and B. V. Sherif, "Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection," 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Dharan, Nepal, 2022, pp. 516-521, doi: 10.1109/I-SMAC55078.2022.9987265.

[19] A. Chintha et al., "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 1024-1037, Aug. 2020, doi: 10.1109/JSTSP.2020.2999185.

[20] A. Rahman et al., "Short And Low Resolution Deepfake Video Detection Using CNN," 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), Hyderabad, India, 2022, pp. 259-264, doi: 10.1109/R10-HTC54060.2022.9929719.

**APPENDIX A: DEFINITIONS, ACRONYMS AND ABBREVIATIONS**

|  |  |
| --- | --- |
| CNN | Convolutional Neural Networks |
| RNN | Recurrent Neural Networks |
| DFDC | DeepFake Detection Challenge |
| FF++ | FaceForensics++ |
| GAN | Generative Adversarial Networks |
| EM | Expectation Maximisation |
| CT | Convolutional Traces |
| SOTA | State Of The Art |
| ViT | Vision Transformers |
| MTCNN | Multi Task Cascaded Convolutional Neural Networks |
| MDD | Meta Deepfake Detection |
| AUC | Area Under Curve |
| ACC | Accuracy |
| LSTM | Long Short Term Memory |