

**İSTANBUL TECHNICAL UNIVERSITY  
FACULTY OF COMPUTER AND  
INFORMATICS**

**PERFORMANCE ANALYSIS OF DEEPPFAKE  
DETECTION USING DIFFERENT FACIAL  
REGIONS**

**Graduation Project Final Report**

**Egehan Orta  
150160124**

**Department: Computer Engineering  
Division: Computer Engineering**

**Advisor: Prof. Dr. Hazım Kemal EKENEL**

**June 2021**

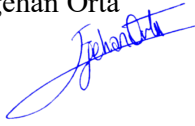
## Statement of Authenticity

I/we hereby declare that in this study

1. all the content influenced from external references are cited clearly and in detail,
2. and all the remaining sections, especially the theoretical studies and implemented software/hardware that constitute the fundamental essence of this study is originated by my/our individual authenticity

İstanbul, June 2021

Egehan Orta



## Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would like to express my deepest appreciation to my advisor Prof. Dr. Hazım Kemal Ekenel, whose expertise is invaluable in computer vision to give me an opportunity to work such attractive and beautiful topic. I would also like to thank Prof. Dr. Ekenel again for his support throughout the year and for contributing greatly to the shaping of my academic career by inviting speakers to Smart Interaction and Machine Intelligence Technologies Lab (SiMiT Lab) and organizing events.

I would also like to extend my deepest gratitude to Prof. Dr. Ekenel's assistant MSc student Alperen Kantarcı for his support and valuable help. During this whole process, he did not spare me any kind of technical assistance and information.

I would like to extend my sincere thanks to members of SiMiT Lab for helping me gain different perspectives in realizing the project.

I would also like to extend my gratitude to Istanbul Technical University and SiMiT Lab for giving an opportunity to work with such successful people and allowing me to use powerful servers which helped a lot to finalize this project.

In addition, thanks also to my colleague Doruk Çanga from my internship at Allianz Insurance for his advices.

# **PERFORMANCE ANALYSIS OF DEEPPAKE DETECTION USING DIFFERENT FACIAL REGIONS**

## **(SUMMARY)**

Computer Vision is one of today's the most popular fields of study of Computer Science. Thanks to the opportunities by developing technology and widespread use of technology, having more processing power and visual data have led to positive impact on development speed of Computer Vision in recent years. Among the subjects of Computer Vision, there are subjects such as face recognition systems, object tracking systems and visual manipulations, whose usage has increased in recent years. Deepfake, one of the subheadings of visual manipulation, is undoubtedly in the top ranks of the most popular topics thanks to social media. Since deepfake is so popular, face manipulations are often referred to as deepfakes lately.

Deepfake is visual data that created using the deep learning to replace human faces inside the video or image with another face within the video or image. Deepfake, whose name comes from the combination of "deep learning" and "fake", is still an active field of study of Computer Vision. Successful models developed by people working in the field of Computer Vision, have an ability to create good deepfakes such that people cannot understand whether the created visual data is real or fake.

Deepfake process, which is easily performed even by using mobile phone applications within the scope of the possibilities offered by developing technology, has been recognized by the whole world by being included in the camera filters of social media applications. Deepfake images, which have a common usage as entertainment, pose a great threat in terms of spreading false information. Since it is difficult to detect whether data is deepfake or not, and there exist a threat of using deepfake to harm people or communities, people working in field of Computer Vision have started to building models to detect deepfake. Although great success has been achieved in deepfake detection, the increase in detection difficulty is inevitable with the developments in the deepfake field. For this reason, studies are continuing to increase performance of deepfake detectors by using the weaknesses of deepfake technology. This situation has also been the focus of attention in this study, and considering that it will contribute to the development of deepfake detection models, it has been tried to determine which facial region has got more importance to detect deepfake.

# FARKLI YÜZ BÖLGELERİ KULLANILARAK YAPILAN DEEPFAKE TESPİTİNİN PERFORMANS ANALİZİ

## (ÖZET)

Bilgisayarla Görü günümüzün popüler Bilgisayar Bilimi çalışma alanlarından bir tanesidir. Gelişen teknolojinin sunduğu fırsatlar ve teknolojinin yaygın kullanımı sayesinde daha fazla işlem gücüne ve daha fazla görsel veriye sahip olmak, son yıllarda Bilgisayarla Görü'nün gelişim hızına olumlu katkılar sağlamıştır. Son yıllarda kullanımında artış yaşanan Bilgisayarla Görü konularının başında yüz tanıma sistemleri, obje takip sistemleri ve görsel manipülasyonlar gibi konular yer almaktadır. Görsel manipülasyonun alt başlıklarından olan deepfake konusu da şüphesiz sosyal medya sayesinde en popüler konuların baş sıralarında yer almaktadır. Deepfake konusunun bu denli popüler olması, yüz manipülasyonlarının genel olarak deepfake olarak adlandırılmasına da sebep olmaktadır.

Derin öğrenme kullanılarak bir videodaki veya fotoğraftaki insan yüzlerinin farklı video veya fotoğraflardaki insan yüzleri ile değiştirilmesiyle oluşan görsel veriye deepfake adı verilmektedir. İsmi İngilizce “deep learning” ve “fake” kelimelerinin birleşiminden gelen deepfake, Bilgisayarla Görü alanının halen aktif bir çalışma konusudur. Bilgisayarla Görü alanında çalışan kişilerin geliştirdikleri başarılı modeller, yaratılan görsel verilerin gerçek olup olmadığı insanlar tarafından anlaşılamayacak derecede iyi deepfake verisi oluşturma kabiliyetine sahiptir.

Gelişen teknolojinin sunduğu fırsatlar çerçevesinde cep telefonlarındaki uygulamaların kullanımıyla bile kolaylıkla gerçekleştirilen deepfake işlemi sosyal medya uygulamalarının kamera filtrelerinde de yer alarak tüm dünya tarafından tanınmıştır. Eğlence alanında kullanımı yaygın olan deepfake görüntüleri yalan bilgi yayılımı konusunda çok büyük bir tehdit oluşturmaktadır. Yaratılan deepfake görüntülerinin gerçekliğinin kolaylıkla tespit edilememesi ve sahte görüntüler oluşturularak kişilere veya topluluklara zarar verilebilme potansiyelinin yüksek oluşu Bilgisayarla Görü alanında çalışma yapan kişilerin dikkatini çekmiş ve deepfake tespiti yapabilen modellerin geliştirilmesine sebep olmuştur. Her ne kadar deepfake tespitinde büyük başarılar sağlansa da deepfake alanındaki gelişmelerle tespit zorluğundaki artış kaçınılmazdır. Bu sebeple deepfake teknolojisinin zayıflıkları kullanılarak deepfake tespitindeki performansın yükseltilmesi için çalışmalar sürdürülmektedir. Bu durum bu çalışmanın da ilgi odağı olmuş ve deepfake tespit modellerinin gelişmesine katkı sağlayacağı düşünülerek, yüzdeki farklı bölgeler kullanılarak deepfake tespitinde hangi bölgenin ne derecede önem taşıdığı tespit edilmeye çalışılmıştır.

# Contents

|   |           |
|---|-----------|
| 1 Introduction and Problem Definition ..... | 1         |
| 2 Literature Survey.....                    | 2         |
| 3 Developed Approach and System Model ..... | 3         |
| 3.1. Dataset.....                           | 3         |
| 3.2. Models.....                            | 6         |
| 4 Binary Detection Models.....              | 9         |
| 4.1 Preprocessing Images .....              | 9         |
| 4.2 Training Models .....                   | 9         |
| 5 Results .....                             | 10        |
| 6 Conclusion and Future Work.....           | 12        |
| <b>7 References.....</b>                    | <b>13</b> |

# 1 Introduction and Problem Definition

Image processing techniques have been the focus of attention of people dealing with computer science for years. After discovering how to process visual data thanks to the technology we have, various studies conducted to increase the use of this data have contributed greatly to the development of computer vision. After we realized we could process and manipulate visual data by using computers, with the help of mathematics, the way to find elements such as edges, corners, objects and even movement directions in images was made possible. The world of science was not limited to these inventions, it continued the development of the field of computer vision by using the opportunities provided by the developing and renewed technology. Developments in the field of machine learning and artificial intelligence have taken the field of computer vision to a different dimension, allowing visual data to be learned, processed and manipulated by machines. Today, with the technological power we have, the widespread use of artificial intelligence and computer vision has increased, and thanks to mobile phones and computers, these technologies have turned into technologies that are easily accessible by everyone.

The development and combination of artificial intelligence and computer vision undoubtedly led to very important discoveries. Today, using these technologies, many applications such as face recognition, motion analysis and tumor detection have been developed, as well as visual syntheses such as the development of 3D objects from 2D objects, creation of virtual cities and image transformations called deepfake. Deepfake, one of these issues, is at the heart of this project and studies on deepfake detection have been carried out in this project.

The deepfake process, which is performed by synthesizing fake images using deep learning, takes as the main goal the replacement of a person's face in a video or photo with a face video or photograph of another person. Scientists conducting studies in this field have developed successful models that create deepfakes so well that people cannot distinguish between real and fake. Deepfake is still an active field of computer vision.

Although Deepfake is an academic research area, it has become a technology that has attracted the attention of the whole world with the effect of the widespread use of mobile phones and social media. Nowadays, it is possible to easily create deepfakes even with popular social media applications. Although deepfake seems like a successful and amusing technology, there exist serious problems it may create. For example, today, thanks to the use of deepfake, while people who are not alive can be included in movies, by manipulating the videos of important statesmen, the way has been enabled for the creation of dangerous discourses. For this reason, the subject of deepfake detection has entered the radar of the Computer Vision field and models that detect deepfake have started to be developed by people working in this field. Although among the developed detection models, there are models that have achieved great success, deepfake is still a developing field of study that tries to cover its weaknesses. As deepfakes develop, deepfake detection models will also be developed. This development of detection models will only be possible if weaknesses of deepfakes can be found. However, the presence of many different manipulation methods, making it difficult to form a comprehensive detection model. This study aims to contribute to the detection of weak spots in deepfakes by measuring the success of distinguish of different facial regions in deepfake detection using dataset consisting a mixture of images created by different methods.

## 2 Literature Survey

Sharing images or videos is one of the most popular activity in today's world. People like to share visual data on the internet, and recently, fake data called deepfake has also been included in this visual data. It is predicted that deepfake elements can cause many problems as well as having entertainment features. For this reason, people working in the Computer Vision field have tried to create deepfake detection models and have achieved successful results in this.

Rösslar et al. have made a great and important contribution to deepfake detection systems with their work, and has become the performance metric of many studies in this field. They created a dataset that has 1000 original youtube videos, 363 original videos of paid actors, 3068 manipulated videos called Deepfake Detection Dataset and manipulated video datasets each containing 1000 videos created by Deepfakes, Face2Face, FaceSwap and NeuralTextures methods. They served this dataset as in low quality, high quality and raw. They also created and tested detection models on this dataset with for every quality group. They prepared a report with 5 different models. As a result of the models they made, they were able to obtain 99.26 percent success as a binary detection. Furthermore, they created a benchmark dataset to make able to other researchers compare their models' performance. [1]

A research done by Li and Lyu showed that deepfakes have got serious weaknesses. In their research they trained the models without any deepfake data. They collected images as positive examples and applied basic image operations like resize and blur to create negative examples. Since images got degraded after deepfake operations, they tried to simulate these degradations with image operations. They trained 4 different methods: VGG16 [2], ResNet50, ResNet101 and ResNet152 [3]. They evaluated these methods with 2 different datasets called UADFV [4] and DeepfakeTIMIT [5]. Top results according to these datasets are 97.4 percent for UADFV [4], 99.3 percent and 93.2 percent for DeepfakeTIMIT's low-quality and high-quality dataset [5] respectively. Their methods showed that Some features of deepfakes offer great advantages in terms of detection. [6]

According to the research of Ciftci, Demir and Yin, deepfakes can be detected by using information gained from not whole face but some regions. They extracted photoplethysmography (PPG) signals from different ROIs and created their models by using these signals. They used FaceForensics++ [1] dataset to train and test models. They made their researches with 7 different models. Their results showed that their approach can detect deepfakes with 97.29% accuracy and the source of deepfake method with 93.39% accuracy. This research is showed that some regions in face has got more information for deepfake detection. [7]

The topic of analyzing the importance of different facial regions, is engaged attention of Tolosano et al. and they made a research on this topic. They masked eye, nose, mouth and rest face area and trained models separately. They used UADFV [4], FaceForensics++ [1], Celeb-DF [8] and DFDC [9]. They trained each facial region and each dataset separately. They used Xception [10] and Capsule Network [11]. Based on their results they managed to get 83.6% AUC score and some facial parts give more accuracy than whole face, also their approach gave more accurate results than some of previous researches. [12]



### 3 Developed Approach and System Model

This project is aimed to find important facial regions for deepfake detection. Unlike similar studies, in this project, it was decided to use all of the datasets together instead of using them separately. Also, in the study that Tolosano et al. made [12], Xception [10] and Capsule Network [11] are used as model, however this project will focus on different models. In this project Vgg19 [2], Xception [10] and ResNet50 [3] are used to detect deepfakes. Dataset is created by using FaceForensics++ [1], Celeb-DF [8] and DeepfakeTIMIT [5]. While measuring the performance of models and different facial regions accuracy and area under curve (AUC) are used as performance metrics.

#### 3.1. Dataset

Three different datasets are selected for this project. These datasets will be used combined during the training phase of model. The selected models are: FaceForensics++ high-quality (c23 compression) dataset [1], Celeb-DF [8] dataset and DeepfakeTIMIT high-quality dataset [5]. Information about datasets can be seen in Table 3.1.

**Table 3.1:** A Sample table

|                            | <i>Original Video Count</i> | <i>Manipulated Video Count</i> |
|----------------------------|-----------------------------|--------------------------------|
| <b>FaceForensics++ [1]</b> | 1363                        | 8068                           |
| <b>Celeb-DF [8]</b>        | 890                         | 5639                           |
| <b>DeepfakeTIMIT [5]</b>   | 0                           | 320                            |
| <b>Total</b>               | <b>2253</b>                 | <b>14027</b>                   |

The distribution of data is not balanced. So, 20% of manipulated videos were selected to work while all of the original videos were selected. However, there was still imbalanced data problem. To overcome this problem, 7 frames planned to be extracted from each manipulated video and 10 frames planned to be extracted from each original video yet some faces could not detected by selected detection method and total number became slightly less than expected. At the end data was almost distributed balanced.

After the videos to be used in the project were determined, the dataset was divided into three as train, validation and test. Train dataset consists 64%, validation data consists 16% and test data consists 20% of whole dataset. Detailed information about the selected dataset is shown in table 3.2, while detailed information about the frames to be used is shown in tables 3.3 and 3.4.

**Table 3.2:** Detailed information about selected dataset

|               | Celeb-DF    |          |         | FaceForensics++    |           |             |             |          |                 |          |         | DeepfakeTIMIT | Total       |          |
|---------------|-------------|----------|---------|--------------------|-----------|-------------|-------------|----------|-----------------|----------|---------|---------------|-------------|----------|
|               | manipulated | original |         | Deepfake Detection | Deepfakes | manipulated |             |          |                 | original |         | manipulated   | manipulated | original |
|               | synthesis   | real     | youtube |                    |           | Face2Face   | FaceShifter | FaceSwap | Neural Textures | Actors   | Youtube |               |             |          |
| All Data      | 5639        | 590      | 300     | 3068               | 1000      | 1000        | 1000        | 1000     | 1000            | 363      | 1000    | 320           | 14027       | 2253     |
| Selected Data | 1127        | 590      | 300     | 613                | 200       | 200         | 200         | 200      | 200             | 363      | 1000    | 64            | 2804        | 2253     |
| Train         | 720         | 377      | 192     | 392                | 128       | 128         | 128         | 128      | 128             | 232      | 640     | 40            | 1792        | 1441     |
| Validation    | 181         | 95       | 48      | 98                 | 32        | 32          | 32          | 32       | 32              | 58       | 160     | 11            | 450         | 361      |
| Test          | 226         | 118      | 60      | 123                | 40        | 40          | 40          | 40       | 40              | 73       | 200     | 13            | 562         | 451      |

**Table 3.3:** Detailed information about splits

|            | Image Count |          |             |          |
|------------|-------------|----------|-------------|----------|
|            | planned     |          | actual      |          |
|            | manipulated | original | manipulated | original |
| Train      | 12544       | 14410    | 12048       | 13732    |
| Validation | 3150        | 3610     | 3061        | 3447     |
| Test       | 3934        | 4510     | 3746        | 4283     |

**Table 3.3:** Detailed information about test data

| Dataset             | Type            | Count |
|---------------------|-----------------|-------|
| FaceForensics++ [1] | Deepfake        | 767   |
|                     | Detection       |       |
|                     | Deepfakes       | 260   |
|                     | Face2Face       | 264   |
|                     | FaceShifter     | 263   |
|                     | FaceSwap        | 263   |
|                     | Neural Textures | 258   |
|                     | Original        | 2542  |
| Celeb-DF [8]        | Fake            | 1580  |
|                     | Original        | 1741  |
| DeepfakeTIMIT [5]   | Fake            | 91    |

### 3.1.1 Methods of Used Dataset Classes

The dataset used on this project created by using different manipulation methods and different original samples.

#### 3.1.1.1 FaceForensics++ [1]

**FaceSwap [1]:** FaceSwap is basically a method that matches source facial landmarks with the targets' and renders the source face on the target image.

**DeepFakes [1]:** Deepfake class is created by using deep learning to swap target face with source face.

**Face2Face [1]:** Face2Face method transfers the expression of source face to the target face by reconstructing target without destroying target identity.

**NeuralTextures [1]:** NeuralTextures method uses deep learning to transfer facial expression of source image to target by learning neural textures.

**DeepFakeDetection [13]:** DeepFakeDetection class created by using deep learning. This method transfers source face to the target image.

**FaceShifter [14]:** FaceShifter class created by using deep learning. This method transfers source face to the target image.

**Original [1]:** Original videos of FaceForensics++ dataset created by using 1000 Youtube videos and 363 paid actor videos.

### 3.1.1.2 Celeb-DF [8]

**Fake [8]:** Fake videos of Celeb-DF dataset created by deep learning methods by swapping target face with the source face.

**Original [8]:** Original videos created by using Youtube videos and celebrity videos.

### 3.1.1.3 DeepfakeTIMIT [5]

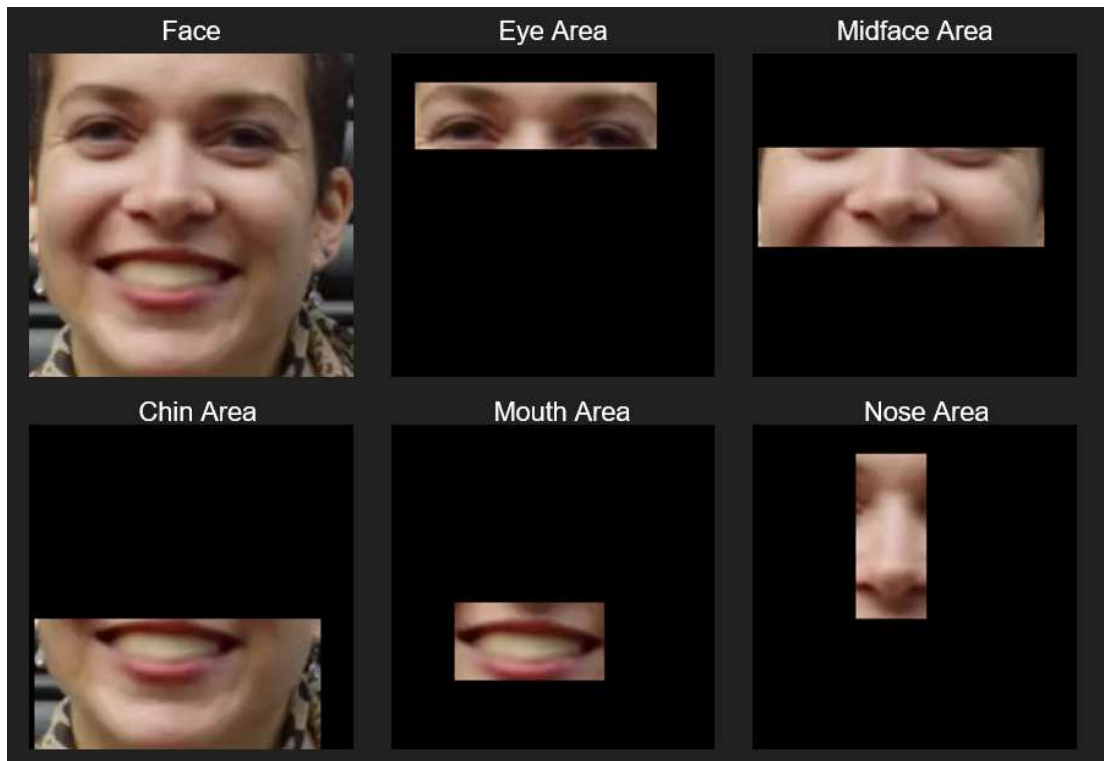
**Fake [5]:** Fake videos of DeepfakeTIMIT database are created by using deep learning to swap target face with source face.

## 3.1.2 Extracting Facial Regions

Videos of selected datasets have got average duration of 15 second. 7 frames at equal distance from each other from manipulated videos and 10 frames at equal distance from each other were extracted from original videos by using OpenCV [15]. Then, 68 facial spots were found by using DLIB [16] and ROIs were extracted. Five facial regions have been extracted from the selected dataset. While extracting faces from videos some faces could not detected and final frame count became fewer than the planned count. The selected facial regions were eye area, mouth area, midface area, nose area and chin area. While creating these areas below boundaries were used for regions as in order top, bottom, left and right edges.

- **Eye Area:** Top 18-27 / Mid-point of bottom 37-48 and spot 29 / Leftmost of 18 and 37 / Rightmost 27 and 46.
- **Mouth Area:** Bottom 32-36 / Bottom 55-6 / Leftmost of 49,50,60 and 61 / Rightmost 54,55,56 and 65
- **Midface Area:** Bottom 37-48 / Top 49-55 / Leftmost 1-17 / Rightmost 1-17
- **Nose Area:** Top 49-55 / Top 18-27 / Spot 32 / Spot 36
- **Chin Area:** Top 49-55 / Bottom 1-17 / Leftmost 1-17 / Rightmost 1-17

Example image can be seen in Figure 3.1.



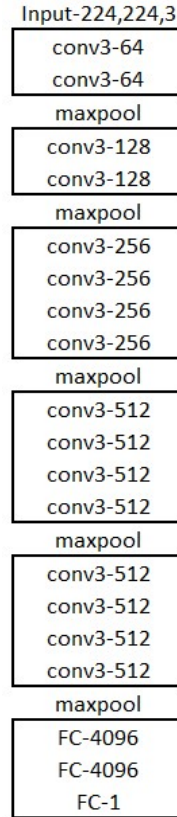
**Figure 3.1:** A sample figure from FaceForensics++ Dataset [1]

## 3.2. Models

Since this project aimed to find performances of detection of different facial regions, different models for each 6 classes must be created. Vgg19 [2], Xception [10] and ResNet-50 [3] were selected as models to be tested. Each CNN model trained for each 6 classes. During the training phase convolution part of each model used with the weights of Imagenet [17]. Only the ANN parts of each model were trained.

### 3.2.1 VGG19 [2]

In all classes same structured Vgg19 [2] model is used. Imagenet [17] weights were used for convolution layers. After convolution layers, 3 layered ANN was used which structured as 4096 fully connected nodes with ReLU activation, 4096 fully connected nodes with ReLU activation and 1 fully connected node with Sigmoid activation. Related graph can be seen in Figure 3.2.



**Figure 3.2:** Structure of VGG-19 [2]

### 3.2.2 Xception [10]

In all classes same structured Xception [10] model is used. Imagenet [17] weights were used for convolution layers. After convolution layers 1 fully connected node with Sigmoid activation is used. Related graph can be seen in Figure 3.3.

### 3.2.3 ResNet-50 [3]

In all classes same structured ResNet-50 [3] model is used. Imagenet [17] weights were used for convolution layers. After convolution layers 1 fully connected node with Sigmoid activation is used. Related graph can be seen in Figure 3.4.

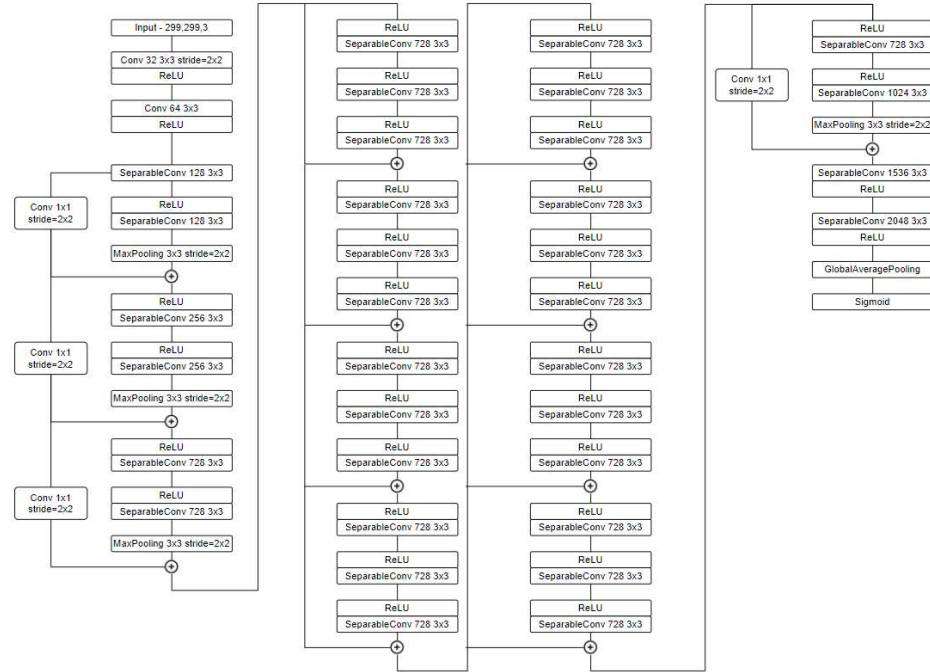


Figure 3.3: Structure of Xception [10]

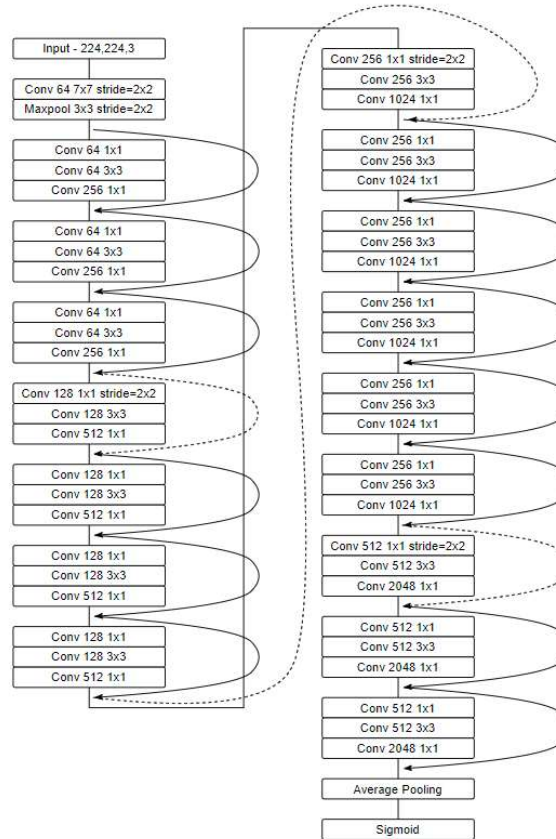


Figure 3.4: Structure of ResNet-50 [3]

## 4 Binary Detection Models

### 4.1 Preprocessing Images

Before feeding networks with images every image must be preprocessed as Imagenet [17] weights were used. Each image randomly was rotated in range of 90 degrees. The brightness of each image was changed to a randomly chosen value between 80 percent and 120 percent of the original brightness. Finally, images were randomly flipped horizontally and preprocessing function of related model was applied. To apply this preprocessing function, Tensorflow [18] was used.

### 4.2 Training Models

During the training phase 18 different models were trained. For each 6 selected facial region, 3 models trained as described before. While training these models learning rate of 0.0002 and a batch size 32 were used. Each model trained with Adam optimizer with the default moment parameters ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1e-8$ ) and binary cross-entropy loss. Each model was trained for 20 epochs.

While training these models, best training accuracy (Figure 4.1), training AUC score (Figure 4.2), training loss (Figure 4.3), validation accuracy (Figure 4.4) and validation AUC score (Figure 4.5) were obtained in whole face region by using VGG19 [2] model. However, the validation loss (Figure 4.6) of the VGG19 [2] model in the whole face area performed worse than the other VGG19 [2] and Resnet50 [3] models.

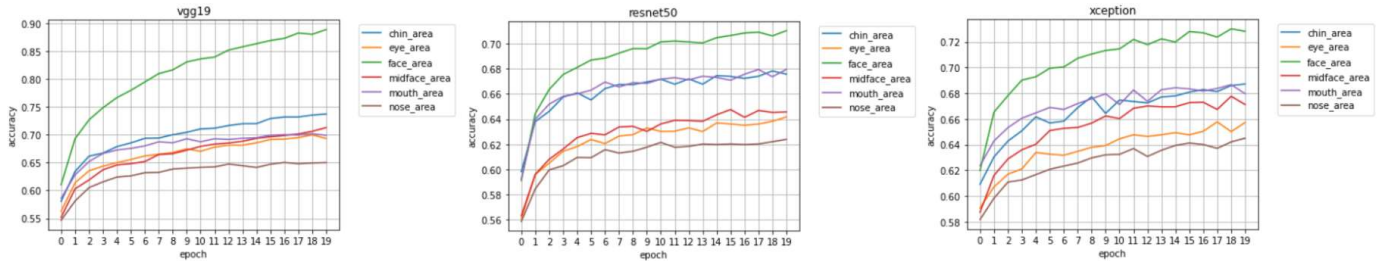


Figure 4.1: Training accuracies of each model in each region over epochs.

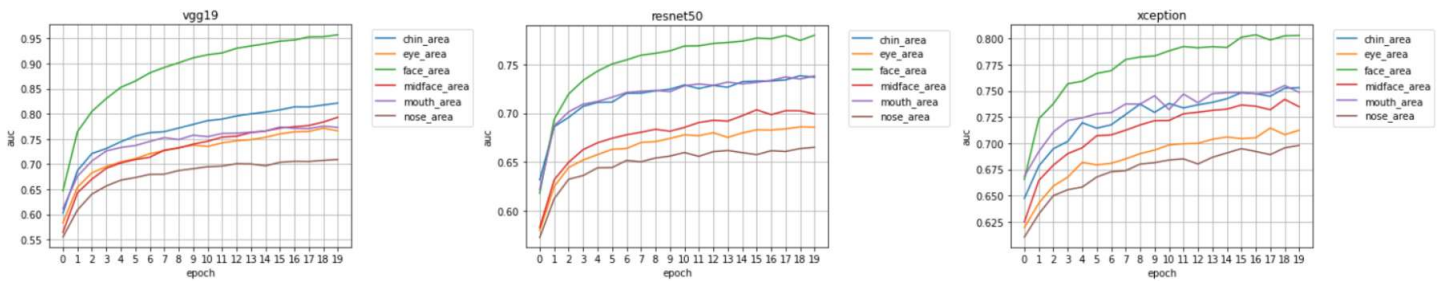


Figure 4.2: Training AUC scores of each model in each region over epochs.



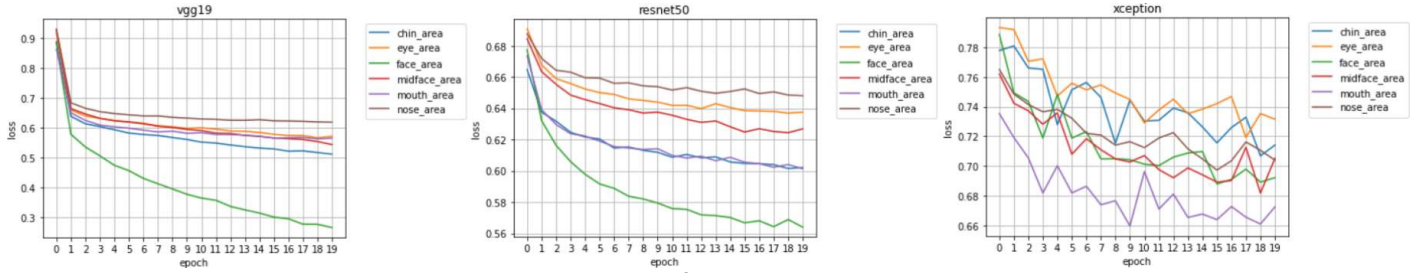


Figure 4.3: Training losses of each model in each region over epochs.

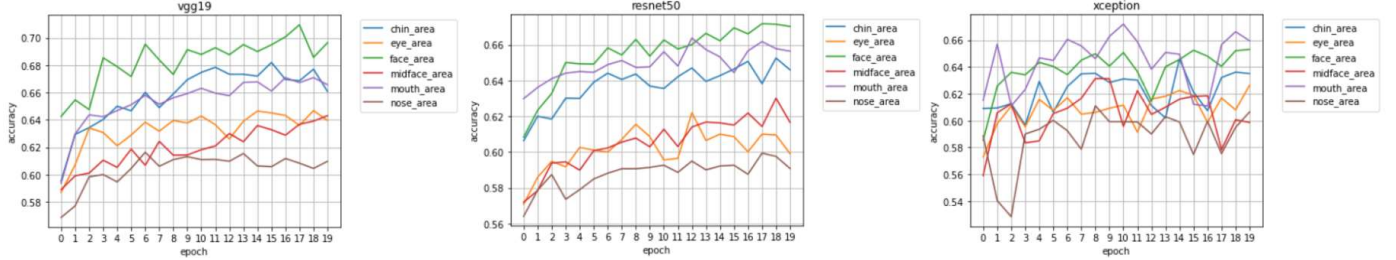


Figure 4.4: Validation accuracies of each model in each region over epochs.

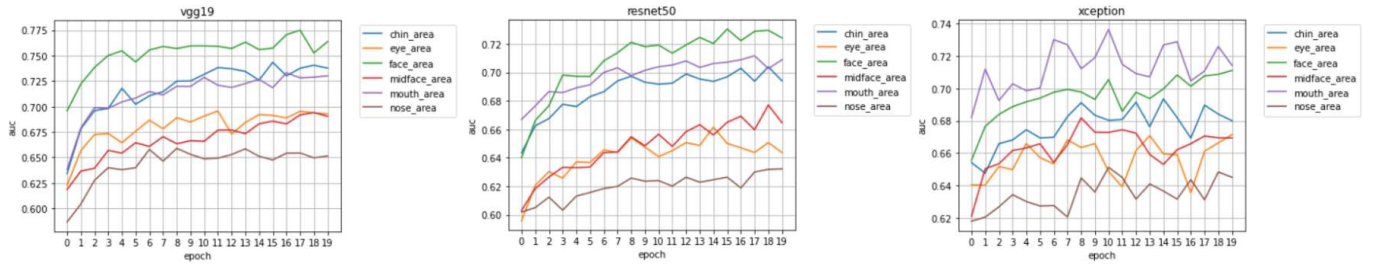


Figure 4.5: Validation AUC scores of each model in each region over epochs.

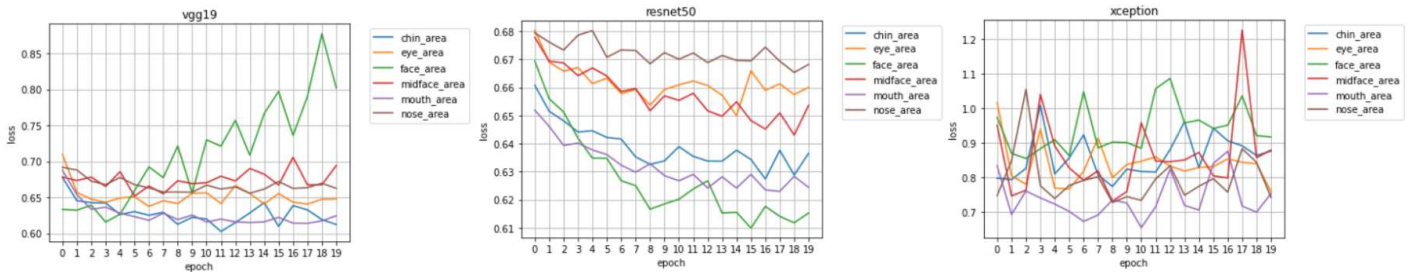


Figure 4.6: Validation losses of each model in each region over epochs.

## 5 Results

In the Table 5.1, accuracies of all data are listed. Considering that 53% of the dataset consists of original images, we can say that by looking accuracies the models are not very successful. Also, it was not a surprise since training and validation accuracies were also not high. To understand the reason for this low accuracy, the datasets were tested separately. Looking at Table 5.3, it is seen that where the accuracies of the original images are high, the fakes generally have low accuracy and vice versa. In addition, the imbalanced distribution of the fake class within itself shows that some models learn specific methods rather than all fake class. It can be seen that there is a tradeoff in some cases between fake classes of Faceforensics++ [1] dataset and other fakes. Also, the number of fake images created by



using Deepfakes approach greater than others and that causes low accuracies on other manipulation methods.

The listed AUC scores (Table 5.2) showed that, some models performed better than Celeb-DF [8] in terms of AUC score (Table 5.3). However, since used dataset of this project way different than Celeb-DF [8] study it will be inappropriate to compare the results directly.

**Table 5.1:** Accuracies of models and facial regions.

| region       | VGG19 [2]   | ResNet50 [3] | Xception [10] |
|--------------|-------------|--------------|---------------|
| chin_area    | 0.66185081  | 0.65250963   | 0.645534933   |
| eye_area     | 0.635446489 | 0.616639674  | 0.628347218   |
| face_area    | 0.679412127 | 0.674305618  | 0.613899589   |
| midface_area | 0.620874345 | 0.631211877  | 0.642172098   |
| mouth_area   | 0.676422954 | 0.663594484  | 0.666832745   |
| nose_area    | 0.610536814 | 0.607049465  | 0.602192044   |

**Table 5.2:** AUC scores of models and facial regions.

| region       | VGG19 [2]   | ResNet50 [3] | Xception [10] |
|--------------|-------------|--------------|---------------|
| chin_area    | 0.720196605 | 0.704772115  | 0.697542071   |
| eye_area     | 0.687221289 | 0.667445362  | 0.677461565   |
| face_area    | 0.74442941  | 0.736271441  | 0.642947078   |
| midface_area | 0.664794207 | 0.685279429  | 0.691203296   |
| mouth_area   | 0.731904447 | 0.722529709  | 0.728539944   |
| nose_area    | 0.649096489 | 0.642353654  | 0.639755368   |

**Table 5.3:** Scores of similar studies.

| Study                | Dataset                      | Method            | Metric   | Score |
|----------------------|------------------------------|-------------------|----------|-------|
| Rössler et al. [1]   | FaceForensics++ HQ [1]       | Xception [10]     | Accuracy | 0.957 |
| Li et al. [8]        | Celeb-DF [8]                 | Xception-c23 [10] | AUC      | 0.653 |
| Li et al. [8]        | Celeb-DF [8]                 | Xception-c40 [10] | AUC      | 0.655 |
| Tolosana et al. [12] | FaceForensics++ FaceSwap [1] | Xception [10]     | AUC      | 0.994 |
| Tolosana et al. [12] | Celeb-DF [8]                 | Xception [10]     | AUC      | 0.836 |
| Li et al. [6]        | DeepfakeTIMIT HQ [5]         | ResNet-50 [10]    | AUC      | 0.932 |

**Table 5.4:** Detailed accuracies of models, facial regions and methods. Green cells are the highest accuracy for method, red cells are the lowest accuracy for method.

| Dataset         | Category          | chin_area   |             |             | eye_area    |             |             | face_area   |             |             | midface_area |             |             | mouth_area  |             |             | nose_area   |             |          |
|-----------------|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------|
|                 |                   | Resnet50    | VGG19       | Xception    | Resnet50    | VGG19       | Xception    | Resnet50    | VGG19       | Xception    | Resnet50     | VGG19       | Xception    | Resnet50    | VGG19       | Xception    | Resnet50    | VGG19       | Xception |
| Celeb-DF        | Real              | 0.507397711 | 0.534377694 | 0.734551787 | 0.463881642 | 0.533507407 | 0.570931256 | 0.619669259 | 0.704090536 | 0.679721475 | 0.731070518  | 0.658833742 | 0.358572662 | 0.587467372 | 0.506527424 | 0.675369859 | 0.660574436 | 0.51697129  | 0.555265 |
|                 | Fake              | 0.793037951 | 0.794936717 | 0.533544302 | 0.793037951 | 0.81265825  | 0.694303811 | 0.745569646 | 0.78607595  | 0.681645572 | 0.558227837  | 0.637974679 | 0.845569611 | 0.772784829 | 0.831012666 | 0.643037975 | 0.543037951 | 0.737341762 | 0.701266 |
|                 | Youtube           | 0.653716207 | 0.697635114 | 0.817567587 | 0.535472989 | 0.581081092 | 0.660472989 | 0.748310804 | 0.847972989 | 0.756756783 | 0.75         | 0.731418908 | 0.417229742 | 0.66385138  | 0.603040516 | 0.716216207 | 0.733108103 | 0.630067587 | 0.603041 |
| DeepfakeTIMIT   | Fake              | 0.780219793 | 0.802197814 | 0.648351669 | 0.648351669 | 0.835164845 | 0.703296721 | 0.747252762 | 0.725274742 | 0.692307711 | 0.417582422  | 0.736263752 | 0.912087917 | 0.560439587 | 0.747252762 | 0.593406618 | 0.571428597 | 0.802197814 | 0.89011  |
| Faceforensics++ | DeepfakeDetection | 0.614080846 | 0.757496715 | 0.670143425 | 0.555410683 | 0.757496715 | 0.607561946 | 0.688396335 | 0.576271176 | 0.683181226 | 0.413298577  | 0.560625792 | 0.844850063 | 0.692307711 | 0.747066498 | 0.644067824 | 0.35853976  | 0.53455019  | 0.584094 |
|                 | Face2Face         | 0.534090936 | 0.450757563 | 0.545454562 | 0.405303031 | 0.363636374 | 0.38257575  | 0.465909094 | 0.375       | 0.488636374 | 0.454545468  | 0.38257575  | 0.791666687 | 0.428030312 | 0.560606062 | 0.435606062 | 0.265151501 | 0.352272719 | 0.613636 |
|                 | FaceShifter       | 0.281368822 | 0.376425862 | 0.433460087 | 0.486692011 | 0.433460087 | 0.444866925 | 0.334600747 | 0.231939167 | 0.342205316 | 0.30038023   | 0.281368822 | 0.646387815 | 0.269961983 | 0.330798477 | 0.266159683 | 0.285171092 | 0.444866925 | 0.509506 |
|                 | FaceSwap          | 0.20152092  | 0.269961983 | 0.3041825   | 0.220532313 | 0.410646379 | 0.338403046 | 0.178707227 | 0.334600747 | 0.315589368 | 0.250950575  | 0.262357414 | 0.646387815 | 0.121673003 | 0.235741451 | 0.20152092  | 0.159695819 | 0.231939167 | 0.410646 |
|                 | NeuralTextures    | 0.616279066 | 0.515503883 | 0.581395328 | 0.391472876 | 0.352713168 | 0.372093022 | 0.220930234 | 0.422480613 | 0.511627913 | 0.321705431  | 0.72868216  | 0.523255825 | 0.600775182 | 0.468992233 | 0.313953489 | 0.383720934 | 0.507752    |          |
|                 | Original          | 0.737608194 | 0.700236022 | 0.68055649  | 0.762391806 | 0.68646735  | 0.707317054 | 0.80133754  | 0.812745869 | 0.728166819 | 0.830055058  | 0.77694726  | 0.466168374 | 0.798583806 | 0.740775532 | 0.806845009 | 0.833202183 | 0.697088897 | 0.625885 |
|                 | Deepfakes         | 0.596153855 | 0.684615374 | 0.649999976 | 0.569230795 | 0.496153831 | 0.580769241 | 0.573076904 | 0.484615386 | 0.61153847  | 0.392307699  | 0.465384603 | 0.77692306  | 0.665384591 | 0.707692325 | 0.646153867 | 0.38846153  | 0.534615397 | 0.596154 |

## 6 Conclusion and Future Work

Although the results of the studies were not as successful as expected, it has created an impression that some facial regions may be more successful (such as mouth area) in detecting deepfakes than other regions. The effort to create a more general model by combining different datasets was the most challenging part of this project. In the future, with a more balanced distributed dataset and more complex models (especially for Xception [10] and ResNet-50 [3]), regions where deepfakes generally fail can be detected more successfully. In addition, when we look at similar studies, we can see that different face detectors are used. Instead of using DLIB's 68 facial landmark detector [16], using different detectors can also lead to more successful models.

As the technology behind deepfake technology develops and the use of deepfake technology becomes widespread, I will be continuing to research on this subject and improve my models. First, I will further increase my knowledge on this subject. And then, I will improve the structure of dataset and I will try different CNN models and face detection methods.

## 7 References

- [1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *ICCV*, 2019.
- [2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.
- [3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
- [4] X. Yang, Y. Li and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP*, 2019.
- [5] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," 2018.
- [6] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," 2019.
- [7] U. A. Ciftci, İ. Demir and L. Yin, "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals," 2020.
- [8] Y. Li, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, United States, 2020.
- [9] B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," 2019.
- [10] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern*, 2017.
- [11] H. H. Nguyen, J. Yamagishi and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," 2019.
- [12] R. Tolosana, S. Romero-Tapiador, J. Fierrez and R. Vera-Rodriguez, "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance," 2020.
- [13] N. Dufour and G. Andrew, "Google AI Blog," [Online]. Available: <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>. [Accessed 25 6 2021].
- [14] L. Li, J. Bao, H. Yang, D. Chen and W. Fang, "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping," 2020.
- [15] OpenCV, "Open Source Computer Vision Library," 2015.
- [16] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.
- [17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," 2015.
- [18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard and R. Jozefowicz, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015.