*Journal of Universal Computer Science, vol. 30, no. 2 (2024), 153-178*
*submitted: 3/4/2023, accepted: 9/10/2023, appeared: 28/2/2024 CC BY-ND 4.0*

# Transfer Learning with EfficientNetV2S for Automatic Face Shape Classification

**Petra Grd**
(University of Zagreb Faculty of Organization and Informatics, Varazdin, Croatia
https://orcid.org/0000-0003-3530-5851, petra.grd@foi.unizg.hr)

**Igor Tomičić**
(University of Zagreb Faculty of Organization and Informatics, Varazdin, Croatia
https://orcid.org/0000-0002-8626-9507, igor.tomicic@foi.unizg.hr)

**Ena Barčić**
(University of Zagreb Faculty of Organization and Informatics, Varazdin, Croatia
https://orcid.org/0000-0002-4194-0512, ena.barcic@foi.unizg.hr)

**Abstract:** The classification of human face shapes, a pivotal aspect of one's appearance, plays a crucial role in diverse fields like beauty, cosmetics, healthcare, and security. In this paper, we present a multi-step methodology for face shape classification, harnessing the potential of transfer learning and a pretrained EfficientNetV2S neural network. Our approach comprises key phases, including preprocessing, augmentation, training, and testing, ensuring a comprehensive and reliable solution. The preprocessing step involves precise face detection, cropping, and image scaling, laying a solid foundation for accurate feature extraction. Our methodology utilizes a publicly available dataset of female celebrities, comprising five face shape classes: heart, oblong, oval, round, and square. By augmenting this dataset during training, we magnify its diversity, enabling better generalization and enhancing the model's robustness. With the EfficientNetV2S neural network, we employ transfer learning, leveraging pretrained weights to optimize accuracy, training speed, and parameter size. The result is a highly efficient and effective model, which outperforms state-of-the-art approaches on the same dataset, boasting an outstanding overall accuracy of 96.32%. Our findings demonstrate the efficiency of our approach, proving its potential in the field of face shape classification. The success of our methodology holds promise for various applications, offering valuable insights into beauty analysis, cosmetic recommendations, and personalized healthcare.

**Keywords:** EfficientNetV2, CNN, neural networks, face shape classification, transfer learning
**Categories:** I.2.0, I.4.0, I.4.9
**DOI:** 10.3897/jucs.104490

## 1 Introduction

The shape of a person's face is an important aspect of their physical appearance, and based on the proportions and contours of their facial features such as jawline, cheekbones, and forehead, can be classified into various types such as heart, oblong, oval, round, and square [Tio 2019]. Each of these face shape types has its own unique characteristics, which can be used to classify individuals into different categories. This can be viewed as a multiclass classification problem, which involves classifying an input image into one of several possible classes. In this case, the classes correspond to different face shape types.

Face shape classification has numerous applications in fields such as beauty industry, plastic surgery, security, entertainment industry, and others. For example, in the beauty industry, knowledge of a person's face shape can be used to recommend makeup products and application techniques that will complement their features, to recommend hairstyles, glasses frames or hats. In the plastic surgery industry, face shape classification can be used to help patients visualize the potential results of a particular procedure. In the entertainment industry, face shape classification can be used to select actors and actresses for different roles based on their facial features [Sukumaran et al. 2021].

In recent years, neural networks have become the most often used method for image classification. Neural networks are a type of machine learning algorithm that are modeled after the structure and function of the human brain. They consist of layers of interconnected nodes, each of which performs a simple computation on the input data [Tan and Le 2019]. Neural networks can be used for a variety of machine learning tasks, including image classification, natural language processing and speech recognition. In recent years, neural networks have been widely used for face shape classification, and transfer learning - a technique used to improve the performance of neural networks by leveraging the knowledge learned from one task to another related task - has emerged as an effective technique for improving the performance of these networks. In the context of face shape classification, transfer learning involves using a neural network that has been pretrained on a large dataset such as ImageNet, and then fine-tuning it on a smaller dataset of face shape images. Prior to deciding which neural network will be used in this paper, preliminary research was conducted with multiple neural networks; EfficientNetV2 demonstrated the greatest accuracy-to-time ratio. This may vary based on the specific dataset, task, and experimental conditions, and the most appropriate model was determined through experimentation and evaluation of the problem at hand.

This paper proposes a multi-step methodology for face shape classification using a pretrained EfficientNetV2S neural network, which involves preprocessing, augmentation, training, and testing. The EfficientNetV2 family of networks has been specifically designed to address the challenges of training large neural networks, and has been shown to achieve state-of-the-art performance on a variety of image classification tasks [Tan and Le 2021]. In addition, transfer learning is used to fine-tune the pretrained EfficientNetV2S network on the face shape dataset, which further improves the performance of the network. The proposed methodology involves preprocessing, augmentation, training, and testing of a neural network. The preprocessing step includes face detection, cropping, and image scaling. The dataset used for training and testing is a publicly available dataset [Lama 2023] of female celebrities categorized into five face shape classes: heart, oblong, oval, round, and square. The neural network was trained using the augmented dataset and tested using a holdout method.

The proposed method for face shape classification presents several novel contributions and advantages, which sets it apart from existing approaches in the field. These contributions stem from the use of transfer learning with the EfficientNetV2S neural network and the incorporation of specific techniques for image preprocessing and data augmentation: (1) Transfer learning with EfficientNetV2S: The key novelty of this method lies in leveraging transfer learning with the EfficientNetV2S neural network. Transfer learning allows the model to benefit from the knowledge and representations learned from a large-scale dataset (e.g., ImageNet), and then fine-tune it for the face shape classification task. By utilizing EfficientNetV2S, which is specifically designed to address training challenges in large neural networks, the proposed method can achieve better performance in face shape classification compared to traditional neural network architectures. The discovered effectiveness of EfficientNetV2S in this context is a significant contribution to the field. (2) Multi-step methodology: The proposed method adopts a comprehensive multi-step approach, encompassing crucial stages like image preprocessing, data augmentation, training, and testing. The incorporation of these steps demonstrates a systematic and well-considered strategy to tackle the face shape classification problem. Each step is carefully designed to ensure optimal feature extraction, model generalization, and performance enhancement. Such a holistic approach contributes to improved accuracy and robustness of the classification results. (3) Image preprocessing and Data Augmentation: Preprocessing is a critical factor that significantly impacts the performance of image classification models. The proposed method incorporates sophisticated techniques for face detection, cropping, and image scaling. These preprocessing steps are tailored to ensure that the facial features are properly aligned and represented, leading to better discriminative features during the training process. Additionally, data augmentation techniques are employed to augment the training dataset, thereby enriching the model's exposure to variations in facial shapes, poses, and expressions. This improves the model's ability to generalize to unseen faces. (4) State-of-the-art results: The experiments conducted in this research demonstrate the superiority of the proposed method in achieving state-of-the-art results for face shape classification. By outperforming existing methods, the proposed approach showcases its efficacy and potential to be a leading solution in the field. The obtained results provide relevant evidence of the significance and practical applicability of the method in real-world scenarios. (5) Broad applicability: Face shape classification has a wide range of applications, from beauty recommendations to plastic surgery simulations and casting decisions in the entertainment industry. The proposed method's accuracy and efficiency make it well-suited for various real-world use cases, where accurate and rapid face shape classification is essential.

A literature overview and state of the art is covered within Section 2. A proposed approach, describing elements such as preprocessing, data augmentation, EfficientNetV2S and other model implementation specifics are presented within Section 3. Section 4 describes the environment and the experiments, including the model training, validation and summary of results. Section 5 discusses the achieved results with regard to other research results within the field and positions our work within them. Section 6 concludes this work with a short overview of used methods and achieved results, and provides an insight into further research.

## 2 Related Work

With several practical applications, such as facial identification, virtual try-on, and planning facial surgery, face shape classification is an interesting area of research in computer vision and soft biometrics. Due to the creation of powerful deep learning models and the accessibility of a new dataset, this field has made considerable strides recently. In this section, we review the existing literature on face shape classification with an emphasis on the most recent and significant research.

The largest number of analyzed papers [Duan et al. 2022, Mehta and Mahmoud 2022, Abdullah et al. 2022, Nabil et al. 2021, Marinescu 2021, Weerasinghe and Vidanagama 2020, Alzahrani 2019, Pasupa et al. 2019, Rahmat et al. 2018, Rajapaksha and Kumara 2018] focused on hairstyle and eyeware recommendation and how it correlates to face shape. In addition, most of the papers use deep learning based approaches [Duan et al. 2022, Abdullah et al. 2022, Nabil et al. 2021, Weerasinghe and Vidanagama 2020, Tio 2019, Alzahrani 2019, Pasupa et al. 2019, Rahmat et al. 2018] and were able to achieve higher accuracy results then machine learning approaches. The approaches presented in [Duan et al. 2022, Nabil et al. 2021, Tio 2019] achieved the highest accuracy scores of all analyzed papers, with booth [Tio 2019] and [Nabil et al. 2021] using a modified Inception V3 model.

The authors in [Duan et al. 2022] propose a new face shape classification algorithm. They present the M-RetinaFace network for the purpose of aligning face images and combine the attention mechanism with the EfficientNet bilinear network. Following that the AB-CNN network is proposed for feature extraction. Finally, in the last step, a bilinear pooling layer is used for face shape classification. The authors use a female-only face dataset for their classification. In addition, by replacing mobilenetv3 with resnet50 for face alignment they were able to achieve a more precise alignment, and using MRetinaFace improved the classification accuracy for female faces. Their proposed method heavily relies on the selection of face images and they lack comparison to other state-of-the-art approaches.

Work by authors [Abdullah et al. 2022] presents a machine learning and engineering-based method that combine facial landmarks, geometrical measures, and pre-train model features. Their method consists data pre-processing, hand-crafted facial features are extracted. In this step, 68 facial points are used and 35 features are selected to train the model. Next, eye feature extraction based on a pre-train VGG19 model is performed. Finally, the pre-train model and the handcrafted features are combined into a single feature vector. Hand-pick features and machine self-learned features when merged present effective results for face shape classification. The authors identify the biggest shortcoming of their research as the lack of a proper dataset and note that both the quality and quantity of training data needs to be improved.

The authors in [Nabil et al. 2021] employ the Inception v3 model for classifying different face shapes. This paper uses a publicly available dataset, CelebA, and enriches it with additional data collected from the web. The authors apply image pre-processing before using the images to retrain the models, mainly image straightening, image cropping and resizing, and visual normalization. The training was unsupervised, the learning rate was set to 0.12, and the Cross-Entropy Loss Function was utilized. The authors were able to achieve the highest training accuracy of 94.5% and the highest validation accuracy of 93.9%. The highest reported testing accuracy of 100% was able to be achieved using a size of 500 and 1000 images. Like in the beforementioned papers, a shortcoming of this research is its dataset.

In [Weerasinghe and Vidanagama 2020] the authors present a framework for selecting the most suitable hairstyle or haircut by classifying face shape. The approach also considers beauty expert's knowledge. Their framework consists of four steps, the first one is data collection, next step is face detection using HaarCascade classifier in OpenCV, the next step is landmark detection using a pre-trained facial landmark detector inside the DLib library. The final step is haircut/hairstyle recommendation which is done by first training the system on the pre-processed dataset using the leave-one-out method and Naïve Bayes calcification algorithm. The system was evaluated wuth the Evuni academy of beauty and the accuracy of the hair recommendation module was 83%, making it slightly lower than the face shape classification accuracy. The authors presented their own approach but not a comparison with state-of-the-art approaches.

Work by authors [Tio 2019] presents experimental results that were obtained by retraining the last layer of the Inception v3 model. They create their own dataset and extract 19 features to train a non-CNN based classifiers and used the OpenCV library to detect the face and use it as the input for the DLib's 68-point facial landmark detector. They train the Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), MLP, and K-Nearest Neighbor (KNN) classifiers and retrain the last layer of the Inception v3 model with the default parameters with the learning rate being 0.01 and the number of iterations being 4000. Finally, the authors measured the overall accuracy and concluded that the retrained Inception v3 model achieves the highest accuracy results. The biggest drawback of this paper is the small amount of training and testing images used.

In [Alzahrani 2019] the authors collect their own face dataset and present their three-step approach. The first step is face detection and cropping using a face detection model trained on Histogram of Oriented Gradients (HOG) features. Those features are then used to train a Linear SVM classifier to detect the region of the face. The next step is landmark detection and face alignment and the final step is classification by Inception v3. Training images are fed to Inception v3 along with HOG features and landmarks. The authors conclude that combining hand-crafted features with automatically extracted features can improve the overall performance of face shape classification. However, they note that more labeled data as well as more handcrafted features and more advanced CNN architecture needs to be implemented in the future, to achieve higher accuracy.

The authors in [Pasupa et al. 2019] present a novel framework for a hairstyle recommender system that is based on face shape classification. Their model is based on SVM, and is evaluated on hand-crafted features, deep-learned features and a VGG-face fine-tuned version. They employ two combination techniques: Vector Concatenation and Multiple Kernel Learning (MKL) techniques. They discovered that it is difficult to classify round-against-square and oval-against-heart shaped faces in all cases, and use a Relief algorithm to achieve a more precise ranking of attributes. They found that the model with hand-crafted features yielded better performance. This paper however has several drawbacks, most notably the small dataset size and lack of descriptions, and no comparison with other state-of-the-art models.

In [Rahmat et al. 2018] the authors focus on face shape classification for the purpose of men's grooming. This is the only paper we came across that deals exclusively with men's faces. The authors proposed method consists of image acquisition and data grouping, pre-processing, and conversion to grayscale images. Then contrast is added and segmentation by converting the image into a binary image and bring up facial features using Gabor filters is performed. Next facial features are extracted using Invariant Moment algorithm and the face classification process is performed using a probabilistic neural network (PNN). The authors found that square-shaped faces tended to be the most difficult to classify correctly, and diamond-shape and round-shape tended to be most easy to classify. Because this paper uses only male faces comparison with other works is difficult.

The authors in [Mehta and Mahmoud 2022, Marinescu 2021, Rajapaksha and Kumara 2018] focus on face shape classification based on machine learning methods.

In [Mehta and Mahmoud 2022] the authors look at the possible application of face shape classification in dental reconstruction, as well as in the beauty industry. Their new method uses 13 facial landmarks and calculates 14 features. Their workflow consists of image annotation with facial landmarks using Dlib's 68-point facial landmark detector. Next, the location of landmarks is visually inspected, and images whit correctly placed landmarks are selected. After that distances, ratios, and angles are selected, and different classifiers are used to determine the most important features. The final step is model selection. Their results show that the proposed model can predict round and square face shapes with decent accuracy, and the best results were achieved using the GBM. This paper focuses mostly on face shape classification and leaves out clustering similar shaped images and landmark imputation for future research, implementing those two elements could result in higher accuracy.

Work by authors [Marinescu 2021] extracted facial landmark measurements in combination with a naive Bayes classifier unit. They used an off-the-shelf facial landmark detector, and a U-Net architecture for the shape segmentation of the hair area, and estimate the lower face region by combining the output of the facial landmark detector with the created segmentation mask. They then trained a naive Bayes classifier using the Chicago face database which was enriched with their own dataset. They compared the naive Bayes classifier with SVMs and conclude that the Bayes classifier achieves better overall results. The proposed approach presents two shortcomings, firstly they assume that images are in a near frontal pose, and secondly, the method of extracting depth information and measurements still needs to be perfected.

The authors in [Rajapaksha and Kumara 2018] create a hairstyle recommendation system based on face shapes and suitable hairstyles that is combined with data from experts in the beauty industry. Their approach consists of five steps, firstly, data collection where the authors created their own dataset, next they use face detection. with the Haar Cascade classifier in OpenCV. Third, they perform landmark detection using Dlib machine learning library and the OpenCV/C++ library. Next several calculations are performed, mainly: face length, forehead length, jawline length, chin width, cheekbone width, and angles. The final choice was based on a combination of beauty experts' suggestions and machine learning. A big shortcoming of this paper is the fact that the dataset consists of images that were uploaded by users presenting a large variation in noise, lighting, etc.

While other presented papers focus on the choice of hairstyle and eye-ware [Zhao et al. 2020] focuses on determining face shape for the purpose of facial attractiveness using a deep learning approach.

In [Zhao et al. 2020] the authors present a novel facial attractiveness evaluation system based on face shape structural features. They use a unified 81 landmark template for landmark detection and retrain the Inception v3 image classifier. The authors divide facial structure features into geometric features and triangular area features, and analyze skin texture features using an LBP algorithm. They have selected different machine learning algorithms to achieve the best results for classifying the five face shape subsets. KNN for oval faces, SVM-LIN for oblong and heart shaped faces, and the SVM-RBF for round and square faces. One drawback of this paper is that the authors only consider a limited set of facial features for attractiveness evaluation, namely face shape, facial structure, and skin.

In [Moussa et al. 2022] authors use MobileNet v1 architecture and focus on real-time application. They chose MobileNet v1 because it decreases the computational and space complexities with classification precision loss by utilizing depthwise separable convolutions. Their model was trained for 25 epochs and achieved the best results at the 21-st epoch. The authors were able to achieve an F1-score of 0.781, recall of 0.782, precision of 0.78.

The authors in [Zeng et al. 2023] present a unique approach to face shape classification as it is connected to traditional chinese medicine (TCM). They establish a new TCM-based face shape dataset and propose a new region symmetry mask (RSM) based approach for face shape classification. In addition they introduce five specific TCM face shapes named wood, fire, earth, gold and water. They tested their approach on the publically available SCUT-FBP dataset as well as their newly constructed TCM dataset. The experimental results on the new TCM-based face shape dataset show that the accuracy of the method with RSM has increased by 3

In [Hossam et al. 2021] the authors present a comparative study of different face shape classification techniques. They test different classification algorithms that use landmark distance ratios and angles as features, mainly, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), AdaBoost, and Naive Bayes. They discovered that the SVM classifier with radial basis function kernel achieved the highest overall accuracy of 82

The most notable drawback that can be seen across all analyzed papers is the need for a more diverse, larger, and properly labeled face shape dataset. With most of the papers either creating their own dataset [Duan et al. 2022, Abdullah et al. 2022, Zhao et al. 2020, Tio 2019, Alzahrani 2019, Pasupa et al. 2019, Rahmat et al. 2018, Rajapaksha and Kumara 2018] or combining an existing dataset with their own newly constructed dataset [Marinescu 2021]. It has also been noted that most research focuses on female faces with only [Rahmat et al. 2018] working exclusively with male faces. In addition, all papers noted that image quality such as noise, background, and resolution play a large part in the final face shape classification accuracy. One subject that could be further explored is the inclusion of more racially diverse datasets. Some papers note that they used pre-existing datasets [Mehta and Mahmoud 2022, Nabil et al. 2021, Marinescu 2021, Weerasinghe and Vidanagama 2020, Moussa et al. 2022, Zeng et al. 2023] which do contain some level of diversity, but a large number of papers [Duan et al. 2022, Abdullah et al. 2022, Zhao et al. 2020, Tio 2019, Alzahrani 2019, Pasupa et al. 2019, Rahmat et al. 2018, Rajapaksha and Kumara 2018] states that they created their own dataset ”from the internet”. This can lead to misclassifications or underrepresentation of certain groups. In addition, as some papers like [Tio 2019] point out labeling face shapes is a hard and time consuming process and often involves several experts classifying the same face to ensure proper classification. That is one explanation as to why celebrity faces are most commonly used as datasets because several beauty experts agree on their face shapes. Authors in [Pasupa et al. 2019, Rajapaksha and Kumara 2018] rely on human beauty experts and volunteers in the first step of their classification. A good example is shown in [Pasupa et al. 2019, Rajapaksha and Kumara 2018] where no clearly labeled datasets were available so volunteers were asked to determine face shapes. As a result of that human error and subjectivity can find their way into the dataset. An important finding was also that there was no consistent number of face shapes between the analyzed papers. In our overview we have found between 4 [Abdullah et al. 2022] and 7 [Marinescu 2021] face shape classes being used for shape classification. A good step for unifying the field of face shape recognition would be to clearly define and outline the number and necessary criteria for face shapes. A problem that is evident in all papers is also the segmentation of the hairline. With some papers using color-based skin segmentation [Zhao et al. 2020, Pasupa et al. 2019, Rahmat et al. 2018] and others relying on deep learning approaches [Marinescu 2021]. The problem lies in determining the boundary between the skin and hairline as it is a region that is often occluded by hair, accessories, or other occlusions. An additional problem can be seen in papers that use male faces [Rahmat et al. 2018, Rajapaksha and Kumara 2018] where additional occlusions like beards and/or mustaches being present. In addition, a large focus has been placed on the kind of features that are used with deep learning features [Duan et al. 2022, Nabil et al. 2021, Marinescu 2021, Tio 2019, Alzahrani 2019] being able to achieve higher accuracy ratings then hand-crafted [Pasupa et al. 2019] or geometric [Zhao et al. 2020] features. The highest accuracy was achieved in [Nabil et al. 2021, Tio 2019] where the authors used deep features in combination with the Inception v3 model. One additional notable finding was that most of the analyzed papers still rely on facial landmarks with only [Duan et al. 2022, Rahmat et al. 2018] not using them in their research. The number of landmarks also varies significantly ranging from 60 [Weerasinghe and Vidanagama 2020] to 81 [Mehta and Mahmoud 2022, Nabil et al. 2021, Zhao et al. 2020]. Most of the approaches rely on the Dlib’s 68 point facial landmark detector [Mehta and Mahmoud 2022, Marinescu 2021, Weerasinghe and Vidanagama 2020, Tio 2019, Rajapaksha and

Kumara 2018] with some adding additional landmark points [Mehta and Mahmoud 2022]. In addition, only [Nabil et al. 2021, Tio 2019, Pasupa et al. 2019] rely on pre-determined features or feature extraction methods while most other analyzed papers implement their own feature extraction approaches.

| **Paper** | **Pre-processing** | **Feature extraction** | **Classification** |
|---|---|---|---|
| [Zeng et al. 2023] | - | RSM | VGG16/ResNet +RSM |
| [Moussa et al. 2022] | - | - | MobileNet |
| [Duan et al. 2022] | M-RetinaFace | AM and EfficientNet | AM and EfficientNet |
| [Mehta and Mahmoud 2022] | - | PCA | Gradient Boosted Tree |
| [Abdullah et al. 2022] | HaarCascade | Hand-crafted, VGG19 | Ensemble (DT, RF, GBM, XGB, MLP) |
| [Hossam et al. 2021] | - | FRL | SVM, KNN, BN, MLP, RF, AdaBoost |
| [Nabil et al. 2021] | HaarCascade, Convex Hull | Inception v3 | Inception v3 |
| [Marinescu 2021] | - | U-Net | Naive Bayes |
| [Weerasinghe and Vidanagama 2020] | - | HaarCascade | Naive Bayes |
| [Zhao et al. 2020] | - | Multi-feature fusion | Inception v3 |
| [Tio 2019] | Manual, Inception v3 | Inception v3 | Inception v3 |
| [Alzahrani 2019] | HOG, SVM, ERT | Inception v3, HOG, Landmarks | Inception v3 |
| [Pasupa et al. 2019] | AAM, skin segmentation | AAM, VGG-Face | SVM |
| [Rajapaksha and Kumara 2018] | HaarCascade | - | - |
| [Rahmat et al. 2018] | Grayscaling, CLAHE, Gabor Filter | Invariant Moments | Probabilistic NN |

*Table 1: Overview of face shape classification state-of-the-art*

Upon analyzing the selected research papers, we observed that the current state-of-the-art accuracy for face shape classification stands at 91% on a publicly available dataset. While this accuracy is commendable, it indicates that there is still room for improvement. Despite the significant success of deep learning in various computer vision tasks, we found that it has been relatively underutilized in face shape classification. This underutilization may be attributed to the small number of training images available for this specific task. By devising novel data augmentation techniques and exploring transfer learning from related tasks, we leverage existing knowledge and address the data scarcity issue.

## 3 Proposed Approach

This research proposes a multi-step methodology for face shape classification using transfer learning (Figure 1). To begin, as is discussed in Section 3.1 in more detail, all of the training images are preprocessed. Following the completion of the preprocessing step, the dataset with resized and cropped images is formed which is then augmented. Upon the completion of image augmentation, a new dataset including original and augmented images is created. Following that, the neural network is trained using the augmented dataset. Following training and validation of the network comes the testing phase. First, an input image is preprocessed in the same manner as training images. When the image has been preprocessed, it is then input into a trained neural network, which assigns the image to one of several face shape classes (heart, oblong, oval, round and square) based on its appearance. Heart-shaped faces have broad cheekbones and foreheads that descend to a small chin. Oblong faces are typically longer than they are wide. Oval-shaped faces have a curved form with a height-to-width ratio of around 3:1. Faces with a round form are spherical and roughly as long as they are wide, while square-shaped faces have strong, angular jaws and a forehead that is nearly as wide as the cheekbones and jaw line [Tio 2019].

### 3.1 Preprocessing

The preprocessing method includes three stages (Figure 2): face detection, face cropping and image scaling. Initially, face detection is performed using HOG and Linear SVM on all images in the dataset. Following the detection of a face in an image, the face area is cropped into a square face image. The next step is to resize all the cropped images to 150px x 150px, which is the input size for the neural network.

### 3.2 EfficientNetV2S

Neural network used in this research is a model from EfficientNetV2 [Tan and Le 2021] family of networks. The authors identified several bottlenecks in EfficientNet models: (1) slow training when using large image sizes, (2) slowness when using depthwise convolutions in early layers, (3) scaling up every stage equally is not optimal. In order to mitigate those problems, they designed EfficientNetV2 family of networks where they utilize Fused-MBConv along with MBConv (Figure 3), and apply training-aware NAS and scaling to optimize accuracy, training speed, and parameter size. The authors also state that the resulting networks train 4 times faster than EfficientNetV1 models and have 6.8 times less parameters.

EfficientNetV2 family of networks has three basic models: EfficientNetV2S, EfficientNetV2M and EfficientNetV2L. Models M and L are based on the S model and scaled using compound scaling [Tan and Le 2021].

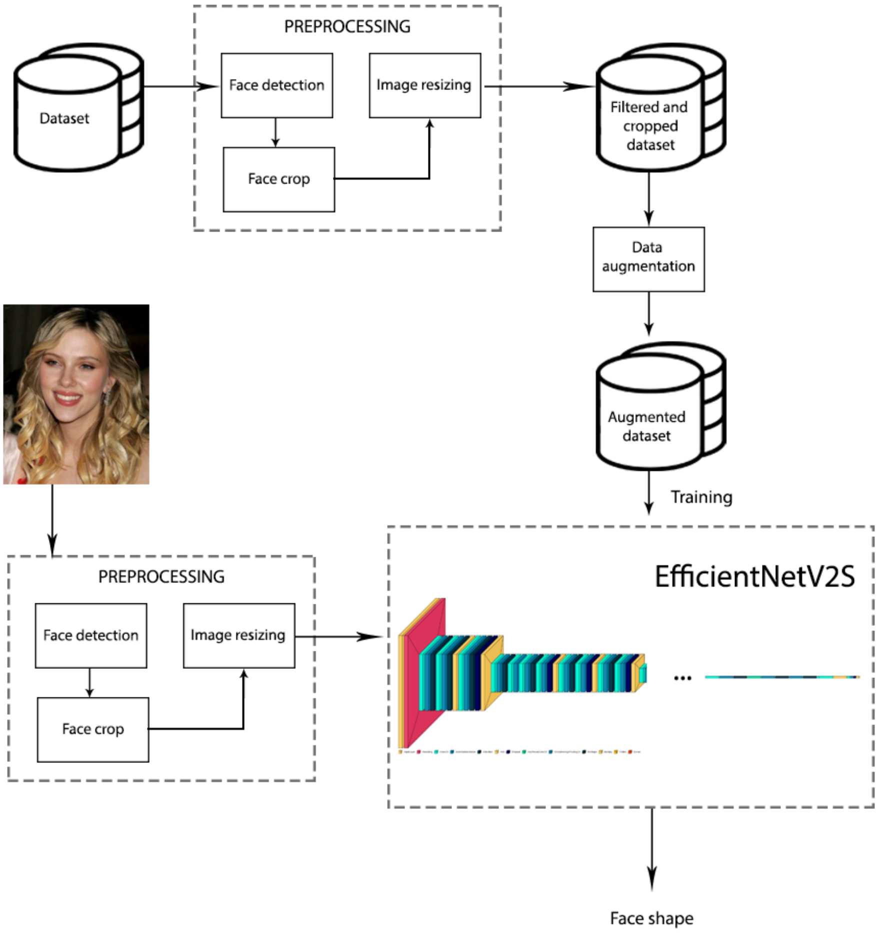

*Figure 1: Architecture of the proposed face shape classification system*

In order to select the base model for this research, we tested all three models which were pretrained on the ImageNet dataset after which finetuning on the face shape dataset (Section 4.1.) was performed. All images were preprocessed (Section 3.1.) and resized to 150px x 150px. Adam Optimizer was used, Rectified Linear Unit (ReLU) activation function in hidden layers and Softmax in the output layer. Loss was calculated using Sparse Categorical Crossentropy. The network was trained for 100 epochs with Batch size 32, Learning rate 0.0001 and Dropout of 0.2 in the dense layer. With the baseline settings, the best validation accuracy (0.6902) was achieved with EfficientNetV2S model (Table 2) which we will be further experimenting with in this research. The selected model also has the smallest number of parameters (53 105 509) and training time (84s per epoch).

The EfficientNetV2S model consists of 42 layers in total, distributed through eight stages (Table 3) and combines Fused-MBConv in stages 1-3 with MBConv in stages

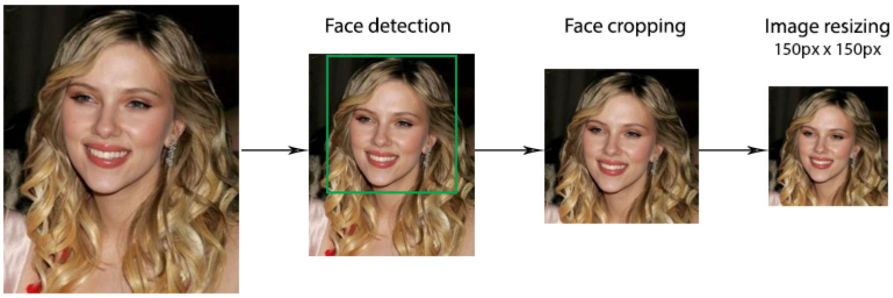

*Figure 2: Image preprocessing diagram*

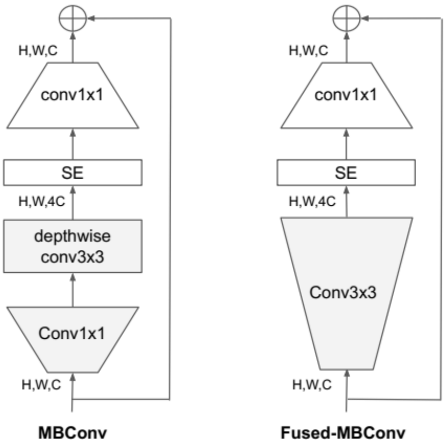

*Figure 3: Difference between MBConv and Fused-MBConv [Tan and Le 2021]*

4-6 with smaller kernel sizes (3x3). Also, MBConv uses squeeze-and-excitation (SE) optimization of 0.25 [Tan and Le 2021]. Since we use transfer learning and use weights from pretrained EfficientNetV2S model, we only fine tune the top part of the model.

**Activation function.** One of the most important parts of the neural network is the activation function. In hidden layers, it will determine how successfully the network model acquires knowledge from the training set. The most often used activation functions in hidden layers are: (1) Rectified Linear Activation (ReLU), (2) Logistic (Sigmoid) and (3) Hyperbolic Tangent (Tanh). In output layer, the choice of activation function will determine the sort of predictions that can be made by a model. Authors in [Goodfellow et al. 2016] recommend the researchers to use ReLU in hidden layers of modern neural networks. The idea of ReLU is to threshold values at 0. When $x < 0$ it outputs 0, and when $x \geq 0$ it outputs a linear function. ReLU can be defined as in Eq. 1 [Agararp 2018].

$$f(x) = max(0, x), \tag{1}$$

For output layer, the most common activation functions are: (1) Linear, (2) Logistic

| Model | Training accuracy | Validation accuracy |
|---|---|---|
| EfficientNetV2S | 0.7681 | **0.6902** |
| EfficientNetV2M | 0.5268 | 0.5328 |
| EfficientNetV2L | 0.6269 | 0.5700 |

*Table 2: Training and validation accuracy of EfficientNetV2 models for face shape classification*

| Stage | Operator | Stride | Channels | Layers |
|---|---|---|---|---|
| 0 | Conv3x3 | 2 | 24 | 1 |
| 1 | Fused-MBConv1, k3x3 | 1 | 24 | 2 |
| 2 | Fused-MBConv4, k3x3 | 2 | 48 | 4 |
| 3 | Fused-MBConv4, k3x3 | 2 | 64 | 4 |
| 4 | MBConv4, k3x3, SE0.25 | 2 | 128 | 6 |
| 5 | MBConv6, k3x3, SE0.25 | 1 | 160 | 9 |
| 6 | MBConv6, k3x3, SE0.25 | 2 | 256 | 15 |
| 7 | Conv1x1 & Pooling & FC | - | 1280 | 1 |

*Table 3: EfficientNetV2S architecture [Tan and Le 2021]*

(Sigmoid) and (3) Softmax. Their usage depends on the prediction problem. Since this research problem is multi-class classification problem, a Softmax activation function is used (Eq. 2). The input into Softmax is a vector of real numbers and Softmax function normalizes it into a probability distribution consisting of probabilities proportional to the exponentials of the input numbers. The result is a vector where each component is in the (0,1) interval [Wang et al. 2018].

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} (i = 1, 2, ..., N) \tag{2}$$

**Batch normalization.** Batch normalization (BN) is implemented by EfficientNetV2 in order to accelerate and standardize the training process. BN is implemented in the design of EfficientNetV2 after each convolutional layer. For the duration of the mini-batch, the BN operation will normalize the activations of the previous layer, which will result in a distribution of activations that is more consistent and centered around a mean of zero and a unit variance. This helps to minimize internal covariate shift, which can slow down the training process and lead to subpar performance if it is not addressed. In addition, BN performs the function of a regularizer and assists in preventing overfitting [Brownlee 2019, Tan and Le 2021].

**Pooling.** In EfficientNetV2 architecture, a global average pooling layer is employed in each block to minimize the spatial dimensions of the feature maps in order to reduce computational cost and control overfitting while maintaining high accuracy [Lin et al. 2013, Tan and Le 2019]. Global average pooling is a pooling procedure used in CNNs to provide the ability to learn invariant features and to minimize the spatial dimensions of the feature maps to a fixed size while preserving crucial feature information. The operation computes the average activation of each feature map across its whole spatial domain, yielding a single value for each feature map [Sharma and Mehra].

# 4 Experiments

The following machine characteristics were used for every experiment that was carried out: the central processing unit is an AMD Ryzen 7 3800X, the graphics processing unit used for EfficientNet training is an Nvidia GTX 1660 6 GB, and there are 32 gigabytes of RAM in total.

## 4.1 Dataset

Since this research uses transfer learning for face shape classification, pretrained EfficientNetV2S weights were used. The network was pretrained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) image classification and localization dataset [Russakovsky 2015] which is the most utilized subset of ImageNet dataset [Deng et al. 2009]. ImageNet is a database of images based on WordNet's hierarchical structure. Each relevant concept in WordNet, defined by numerous words or phrases, is referred to as a "synonym set" or "synset" and an average of 1000 images is provided to illustrate each synset. Currently, there are 5247 categories in ImageNet. The ILSVRC contains 1,281,167 training images, 50,000 validation images, and 100,000 test images over 1000 object classes.

While selecting a dataset for neural network finetuning, it is necessary to take a number of factors into account. A larger number of images in a dataset contributes to an increase in overall accuracy, whereas a balanced distribution of images per class results in a more balanced network training performance. The finetuning of the network for face shape classification proved to be a challenge, mostly because of the lack of available datasets with labeled face shape classes. Most of the papers work with small, private datasets of face shape images (Table 4).

In order to finetune the network, the largest available dataset of face shapes from Kaggle [Lama 2023] was used. The dataset (Original) consists of 4998 images of female celebrities categorized according to their face shape. There are five face shape classes: Heart (999 images), Oblong (999 images), Oval (1000 images), Round (1000 images) and Square (1000 images). The training part of the dataset consists of 800 images from each class, while the test set consists of the rest of the images of each class. From this dataset, Cropped and Augmented datasets were created. The goal of the Cropped dataset is to see if elimination of background and other noise in an image results in better classification performance, while the goal of the Augmented dataset is to create a larger dataset and analyze the impact of dataset size on the performance.

Cropped dataset was created by using HOG and SVM to detect face in an image and saving only the face as a new image. The Cropped dataset consists of the same classes, but with slightly less images: Heart (845 images), Oblong (791 images), Oval (808 images), Round (777 images) and Square (813 images). The difference between the number of images in these datasets comes from inability of the face detector to detect faces in some of the images. Next, the Augmented dataset was created by using data augmentation which takes one image and generates several variants of that image. The augmentation method increases the number of images in the dataset while making it more difficult for the network to learn, because none of the images are completely standard [Belcar et al. 2022]. In this research, the image variants include image rotation for a value between -50°and 30°, adding Gaussian noise to an image, horizontal image mirroring and changing the contrast of an image for gamma value of 2 (Figure 4). The distribution of images per class can be seen in Table 5.

| Paper | Dataset | No. of images | Face shape | Gender |
|---|---|---|---|---|
| [Zeng et al. 2023] | SCUT-FBP + FBP5500 | 4490 | H, Ob, Ov, R, S | M/F |
| [Moussa et al. 2022] | Face shape | 5000 | H, Ob, Ov, R, S | F |
| [Duan et al. 2022] | Private | 5500 | H, Ob, Ov, R, S | F |
| [Mehta and Mahmoud 2022] | Face shape | 5000 | H, Ob, Ov, R, S | F |
| [Abdullah et al. 2022] | Private | 400 | Ob, Ov, R, S | M/F |
| [Hossam et al. 2021] | Private | 500 | H, Ov, R, S, L | F |
| [Nabil et al. 2021] | CelebA | 2500 | H, Ob, Ov, R, S | F |
| [Marinescu 2021] | Chicago face + Private | 719 | Re, R, S, H, D, T, Ov | / |
| [Weerasinghe and Vidanagama 2020] | Face shape | 5000 | H, Ob, Ov, R, S | F |
| [Zhao et al. 2020] | Private | 600 | H, Ob, Ov, R, S | M/F |
| [Tio 2019] | Private | 500 | H, Ob, Ov, R, S | F |
| [Alzahrani 2019] | Private | 500 | H, Ob, Ov, R, S | F |
| [Pasupa et al. 2019] | Private | 500 | H, Ob, Ov, R, S | F |
| [Rahmat et al. 2018] | Private | 120 | Ov, R, D, R, T, S | M |
| [Rajapaksha and Kumara 2018] | Private | 5000+ | H, Ob, Ov, R, S | M/F |

*Table 4: Face shape datasets*

### 4.2 Training and Validation

In order to evaluate the performance of the proposed approach, a holdout method was used. The dataset was split into a train (80%) and test (20%) set. The train set was further split into train (80%) and validation (20%) set. The train set is the collection of images used to train the model, while the validation set is used to update the model's parameters. The test set is used to evaluate the model's performance on unseen data.

For training the network, we monitor the loss. The goal of a loss function is to compare the target and predicted output values and to measure how effectively the neural network predicts the training data. The goal is to minimize loss between the predicted and desired outputs during training. There are different loss functions and the selection depends on the problem that is being solved. Since this research is a multi-class classification

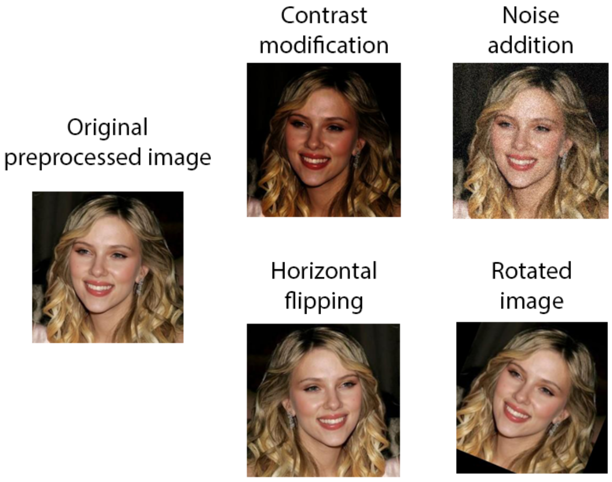

*Figure 4: Image augmentation example*

| Class | Original | Cropped | Augmented |
|---|---|---|---|
| Heart | 999 | 845 | 4995 |
| Oblong | 999 | 791 | 4995 |
| Oval | 1000 | 808 | 5000 |
| Round | 1000 | 777 | 5000 |
| Square | 1000 | 813 | 5000 |
| **Total** | **4998** | **4034** | **24990** |

*Table 5: Number of images per class in the datasets*

problem, we use Categorical Cross-Entropy Loss (Eq. 3) [Goodfellow et al. 2016].

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{3}$$

To validate the network, we monitor the network accuracy. The accuracy metric represents the performance of the model across all classes. It is measured as a proportion of the total number of accurate predictions (Eq. 4) [Grd 2021].

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}} \tag{4}$$

In order to select the best hyperparameters for this research, we conducted different experiments. We wanted to analyse the impact of different parameters on the validation accuracy. We trained the network on four datasets, with various dropout values in fully connected layer, with different optimizers, batch sizes and learning rates. The training in all of the experiments was conducted for 100 epochs and the other variables differ based on what we are analysing.

**Dataset.** As we mentioned in Section 4.1, we use three face shape datasets. Original dataset with 4998 face images, Cropped dataset with 4034 images and Augmented dataset with 24 990 images. To analyse the impact of the dataset size and image content on validation accuracy we conducted experiments on each of the datasets. The three experiments were conducted for 100 epochs, with batch size 32, learning rate of 0.0001, Adam Optimizer and Dropout rate of 0.2 in fully connected layers. Table 6 shows the train and validation accuracies for the three datasets. It is interesting to see that the network trained on the Cropped dataset achieves better validation accuracy than the network trained on the Original dataset, while the network trained on the Augmented dataset vastly outperforms both Original and Cropped. It is our opinion that the increased accuracy of the network on the Cropped dataset probably comes from two facts: (1) the most problematic images to classify are those where faces cannot be detected and in the Cropped dataset those images have been eliminated because no faces to be cropped were found and (2) in the Cropped dataset, only closely cropped face images are present and the background noise has been eliminated which in turn does not confuse the classifier.

| Dataset | Train accuracy | Validation accuracy |
|---|---|---|
| Original | 0.8389 | 0.5760 |
| Cropped | 0.7681 | 0.6902 |
| Augmented | 0.8667 | **0.9236** |

*Table 6: Overview of training results for different datasets*

**Optimizer.** The next parameter analysed is the optimizer. An optimizer is an algorithm that modifies the network's parameters during training in order to minimize the loss function and achieve better classification results. There are a number of different optimizers [Kingma and Ba 2014], such as Stochastic Gradient Descent (SGD), RMSProp and Adaptive Moment Estimation Optimizer (Adam) but in recent years, Adam has emerged as most commonly used in CNNs. To choose the best optimizer for this research we conducted experiments and trained the model with Adam, SGD and RMSProp (Table 7). The experiments were conducted on Cropped dataset for 100 epochs, with batch size 32, learning rate of 0.0001 and Dropout rate of 0.2 in fully connected layers. Cropped dataset was used in all training and validation experiments to decrease the training time and resources needed. The best validation accuracy was achieved by using Adam Optimizer which aligns with the suggestions in literature.

| Optimizer | Training accuracy | Validation accuracy |
|---|---|---|
| Adam | 0.7681 | **0.6902** |
| RMSprop | 0.7447 | 0.6568 |
| SGD | 0.4487 | 0.4808 |

*Table 7: Training and validation accuracy of EfficientNetV2S model for different Optimizers*

**Dropout.** One of the main problems of CNNs is overfitting. In order to mitigate this problem, Dropout is often used. Most commonly, Dropout is applied to fully connected layers. In EfficientNetV2 Dropout, RandAugment, and Mixup are used for regularization in convolutional layers [Tan and Le 2021]. In fully connected layers, we experimented with different Dropout values (0.1, 0.2, 0.4 and 0.5) and also without using Dropout to analyse the impact of Dropout values on the validation accuracy (Table 8). We trained EfficientNetV2S with Adam Optimizer, on Cropped dataset for 100 epochs with Batch size 32, and Learning rate 0.0001. The best validation accuracy was achieved with Dropout value of 0.1, but the difference between accuracy values for Dropout of 0.1, 0.2 and without Dropout are small, so we further experimented with different Dropout values on various batch sizes on Augmented dataset.

| Dropout | Training accuracy | Validation accuracy |
|---|---|---|
| 0.0 | 0.8072 | 0.6890 |
| 0.1 | 0.7950 | **0.7237** |
| 0.2 | 0.7681 | 0.6902 |
| 0.4 | 0.6970 | 0.6766 |
| 0.5 | 0.6556 | 0.6791 |

*Table 8: Training and validation accuracy of EfficientNetV2S model for different Dropout values*

**Batch size.** Training a CNN requires selecting a suitable batch size, which can have a substantial impact on the network's performance. Batch size is the number of training instances transmitted through the network in a single pass [Brownlee 2019]. It specifies the amount of samples to be processed before changing the model's internal parameters. During training, networks are often trained on data batches as opposed to individual data points. The appropriate batch size is determined by the size of the dataset, the network's complexity, and the available computer resources. Using smaller batch size leads to [Keskar et al.2016]: (1) more frequent parameter updates during each epoch which leads to faster convergence as the model learns from smaller and more frequent updates, (2) higher memory efficiency because small batch sizes require less memory to store intermediate activations during training, making them suitable for situations where memory resources are limited, (3) increased noise in gradient estimation which can have a regularizing effect and help generalize better, particularly when dealing with limited training data which is the case in this research. For this research, we tested batch sizes from 8 to 512, and varied Dropout rates (Table 9). The best results were achieved when Batch size was set to 16 with 0.2 Dropout rate in fully connected layers. If we take a look at these results in more detail, the connection between Dropout value and Batch size can be observed. For smaller Batch sizes (8, 16, 32 and 64), better results are achieved with a lower Dropout rate of 0.2, while for the larger Batch sizes (128 and 256), better validation accuracy is achieved for larger Dropout value of 0.5. Since the accuracy results for batch sizes 8, 16 and 32 are close, we further experimented with those three batch sizes on Augmented dataset.

In total, we performed 56 experiments with different hyperparameter configurations, and the best validation accuracy was achieved with the Model 50 whose hyperparameter configuration is presented in Table 10. The model was trained on the Augmented dataset for 500 epochs, with batch size 16, Adam Optimizer, learning rate of 0.0001 and without Dropout. The Accuracy and Loss curves for training with this configuration can be seen

| Batch size | Dropout | Training accuracy | Validation accuracy |
|---|---|---|---|
| 8 | 0.2 | 0.7369 | 0.6902 |
| 8 | 0.5 | 0.6312 | 0.6605 |
| 16 | 0.2 | 0.7450 | **0.7286** |
| 16 | 0.5 | 0.6675 | 0.6617 |
| 32 | 0.2 | 0.7681 | 0.6902 |
| 32 | 0.5 | 0.6556 | 0.6791 |
| 64 | 0.2 | 0.7721 | 0.6791 |
| 64 | 0.5 | 0.6718 | 0.6642 |
| 128 | 0.2 | 0.7551 | 0.6357 |
| 128 | 0.5 | 0.6702 | 0.6568 |
| 256 | 0.2 | 0.7476 | 0.6418 |
| 256 | 0.5 | 0.6876 | 0.6642 |
| 512 | 0.2 | 0.7329 | 0.6766 |
| 512 | 0.5 | 0.6641 | 0.6245 |

*Table 9: Training and validation accuracy of EfficientNetV2S model for different Batch sizes and Dropout values*

in Figure .

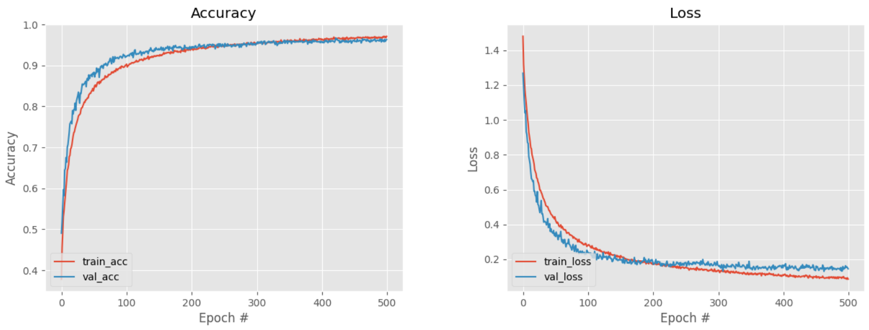

*Figure 5: Accuracy and Loss curves for the training of the model with the highest validation accuracy*

### 4.3 Results

As we mentioned in Section 4.2, a holdout method was used to evaluate the model performance. The dataset was split into a train (80%) and test (20%) set. The images in the test set were not used in model training and validation. Accuracy is only one of the performance measures that can be used to evaluate the model. During testing, other than accuracy, we also created a confusion matrix and calculated precision, recall, specificity and F1-score.

| **Hyperparameter** | **Value** |
|---|---|
| Activation function (hidden layers) | ReLu |
| Activation function (output layer) | SoftMax |
| Loss function | Categorical Crossentropy |
| Pooling | Global Average Pooling |
| Optimizer | Adam |
| Dropout | - |
| Batch size | 16 |
| Learning rate | 0.0001 |
| Trainable parameters | 32 774 149 |

*Table 10: Hyperparameters of the EfficientNetV2S model with the highest validation accuracy*

Confusion matrix is a two-dimensional matrix, with one dimension indexed by the true class of an item and the other by the class assigned by the classifier. It is a summary of the classification performance of a classifier relative to given test data [Ting 2010]. In confusion matrix, TP indicates the number of accurate positive predictions, TN indicates the number of accurate negative predictions, FP indicates the number of inaccurate positive predictions, and FN indicates the number of inaccurate negative predictions.

From the confusion matrix, precision, recall, specificity and F1-score are calculated for each class. Precision is the proportion of positive predictions that are accurate out of the total positive predictions [Ting 2010] and is calculated as in Eq. 5.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall (Sensitivity) is the proportion of positive predictions that are accurate out of the total number of actual positives. Recall is calculated as in Eq. 6.

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{6}$$

Specificity is the proportion of correctly predicted negative instances out of all actual negative instances and is calculated as in Eq. 7

$$Specificity = \frac{TN}{FP + TN} \tag{7}$$

F1-score can be defined as a harmonic mean of Precision and Recall and is calculated as in Eq.8.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{8}$$

During training and validation we experimented with different hyperparameters and trained and validated 56 different models. We tested three models with the best validation accuracy and the results of the overall accuracy can be seen in Table 11. Model 52 was finetuned on the Augmented dataset with image size of 150px x 150px, with learning rate of 0.0001, Batch size 32, Dropout of 0.1, used an Adam Optimizer and was trained for 500 epochs. Model 51 was similar to Model 52, but used Batch size 16. And the highest testing results were achieved with the Model 50 whose hyperparameters can be seen in Table 10.

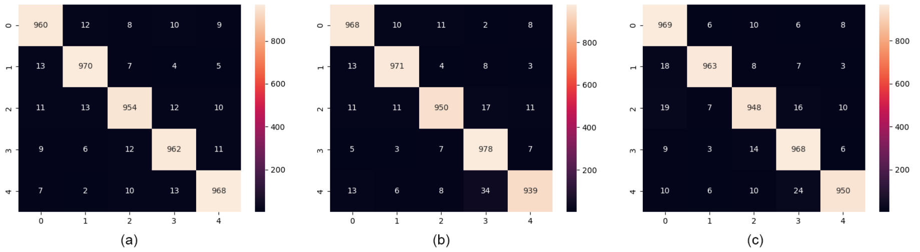

*Figure 6: Heatmap of the confusion matrix for (a) Model 50, (b) Model 51 and (c) Model 52*

The confusion matrices from which we calculated the performances can be seen in heatmaps in Figure 6. We analysed the Precision, Recall (Sensitivity), Specificity and F1-score values per class for the three models which achieved the best testing accuracy and the results can be seen in Table 11.

## 5 Discussion

Table 11 presents detailed testing results for three different models (Model 50, Model 51, and Model 52) in a multi class face shape classification problem. In Model 50, for the face shape category Heart, the model achieves a precision of 0.9600, indicating that 96% of the predicted positive instances are correct. The recall (or sensitivity) is 0.9610, meaning 96.1% of the actual positive instances are correctly identified. The specificity, measuring the true negative rate, is 0.9900, implying that 99% of actual negative instances are accurately recognized. The F1-score, a balanced metric of precision and recall, is 0.9605. Similar evaluation metrics are observed for the other face shape categories (Oblong, Oval, Round, and Square) with high precision, recall, specificity, and F1-scores. The overall accuracy for Model 50 is 0.9632, representing the proportion of correctly classified instances out of the total instances in the test set. The model achieves an exceptional overall specificity of 0.9908, indicating its ability to accurately identify negative instances across all face shape categories. Model 51 shows similar high performance in precision, recall, specificity, and F1-score for all face shape categories (Heart, Oblong, Oval, Round, and Square). The overall accuracy for Model 51 is 0.9616, indicating its effectiveness in predicting face shapes across all classes. The model achieves an overall specificity of 0.9904, further highlighting its ability to correctly identify negative instances. Model 52 demonstrates high precision, recall, specificity, and F1-scores for all face shape categories (Heart, Oblong, Oval, Round, and Square). The overall accuracy for Model 52 is 0.9600, showcasing its ability to predict face shapes effectively. The model achieves an overall specificity of 0.9900. Overall, all three models show strong performance in the face shape classification task, with consistently high precision, recall, specificity, and F1-scores for all face shape categories. These results indicate the models' effectiveness in accurately distinguishing different face shapes. The high accuracy and specificity values further

| | Model 50 | | | |
|---|---|---|---|---|
| **Shape** | **Precision** | **Recall** | **Specificity** | **F1-score** |
| **Heart** | 0.9600 | 0.9610 | 0.9900 | 0.9605 |
| **Oblong** | 0.9671 | 0.9710 | 0.9917 | 0.9690 |
| **Oval** | 0.9627 | 0.9540 | 0.9907 | 0.9583 |
| **Round** | 0.9610 | 0.9620 | 0.9902 | 0.9615 |
| **Square** | 0.9651 | 0.9680 | 0.9912 | 0.9666 |
| | | | | |
| **Accuracy** | | | | **0.9632** |
| | | | | |
| | **Model 51** | | | |
| **Shape** | **Precision** | **Recall** | **Specificity** | **F1-score** |
| **Heart** | 0.9584 | 0.9690 | 0.9895 | 0.9637 |
| **Oblong** | 0.9700 | 0.9720 | 0.9925 | 0.9710 |
| **Oval** | 0.9694 | 0.9500 | 0.9925 | 0.9596 |
| **Round** | 0.9413 | 0.9780 | 0.9847 | 0.9593 |
| **Square** | 0.9700 | 0.9390 | 0.9927 | 0.9543 |
| | | | | |
| **Accuracy** | | | | **0.9616** |
| | | | | |
| | **Model 52** | | | |
| | **Precision** | **Recall** | **Specificity** | **F1-score** |
| **Heart** | 0.9454 | 0.9700 | 0.9860 | 0.9575 |
| **Oblong** | 0.9777 | 0.9640 | 0.9945 | 0.9708 |
| **Oval** | 0.9576 | 0.9480 | 0.9895 | 0.9528 |
| **Round** | 0.9481 | 0.9680 | 0.9867 | 0.9579 |
| **Square** | 0.9724 | 0.9500 | 0.9932 | 0.9611 |
| | | | | |
| **Accuracy** | | | | **0.9600** |

*Table 11: Testing results for each class for Model 50, Model 51 and Model 52*

validate their overall performance, demonstrating their ability to avoid misclassifications of negative instances. The presented table demonstrates that the proposed models (Model 50, Model 51, and Model 52) are effective in face shape classification, achieving high accuracy and specificity, and providing a reliable and robust solution for the given task and outperform the current state of the art algorithms in the field, with Model 50 achieving the best performance results.

Comparison of the results between different papers and our approach is a difficult task, partly because of different face shape classes used, but mainly because no benchmark dataset for face shape classification exists. Most papers use small private datasets to test the performance of their approaches, and only three papers [Mehta and Mahmoud 2022, Weerasinghe and Vidanagama 2020, Moussa et al. 2022] use publicly available Face shape dataset.

There are fifteen papers that proposed and tested algorithms for face shape classification. Most of those papers [Duan et al. 2022, Mehta and Mahmoud 2022, Nabil et al. 2021, Weerasinghe and Vidanagama 2020, Zhao et al. 2020, Tio 2019, Alzahrani 2019, Pasupa et al. 2019, Rajapaksha and Kumara 2018] use the same five face shape classes as this research (Heart, Oblong, Oval, Round and Square) and only three of those papers [Mehta and Mahmoud 2022, Weerasinghe and Vidanagama 2020, Moussa et al. 2022] use a publicly available dataset. All of the papers report Accuracy as a perfomance measure, and if we compare the Accuracy achieved in this paper (96.32%) with the Accuracy achieved in the mentioned papers (89.8%, 91% and 79.2% respectively) we can see that our approach with using image augmentation and EfficientNetV2S neural network outperforms the state of the art approaches on the same dataset. If we disregard the used dataset, we can see that the best results were obtained in [Tio 2019] where author employed pretrained Inception v3 neural network to perform face shape classification. The paper mentioned has several drawbacks. Firstly, a small private dataset used for testing consisting of 500 images. Secondly, it is not clear which testing method has been used. Author uses a dataset of 500 images and trains and tests the network on the same images. And thirdly, testing accuracy values reported in the abstract, conclusion and results tables are inconsistent. In order to try and compare the results, we trained, validated and tested the Inception v3 on the same augmented dataset and the testing accuracy was 64% which shows that our approach with image augmentation and transfer learning with a pretrained EfficientNetV2S network achieves the highest accuracy to date.

## 6 Conclusion

In conclusion, the field of face shape classification has witnessed a wide array of methodologies, from traditional machine learning to cutting-edge deep learning techniques. While the current state-of-the-art accuracy at 91% on a publicly available dataset was commendable, there was ample room for improvement. Despite the remarkable success of deep learning in various computer vision tasks, it has been relatively underutilized in face shape classification. This underutilization can be primarily attributed to the scarcity of training images specific to this task. Deep learning models thrive on large-scale datasets, but the process of acquiring and annotating comprehensive facial shape datasets is laborious and time-consuming. To overcome the challenge of limited training data, we proposed data augmentation techniques and leveraged transfer learning from related tasks. By doing so, we capitalized on existing knowledge and made strides in addressing the data scarcity issue.

In this paper, we have presented a novel and effective multi-step methodology for face shape classification using the EfficientNetV2S neural network and transfer learning. Our approach outperforms state-of-the-art methods on the same dataset, achieving an significant accuracy of 96.32% (as presented in Section 5). The proposed methodology encompasses essential phases, including preprocessing, data augmentation, training, validation, and testing, ensuring a robust and reliable solution for face shape classification.

One of the contributions of our research lies in identifying research gaps and analyzing the state of the art in face shape classification and by addressing these gaps, we devised a multi-step methodology that leverages image preprocessing techniques and data augmentation to enhance the model's performance and adaptability. The inclusion of transfer learning using pretrained EfficientNetV2S weights further optimized the neural network's accuracy, training speed, and parameter size, significantly reducing the need for extensive computational resources. Our approach's versatility enables its

| Paper | Shape classes | Dataset | Accuracy |
|---|---|---|---|
| [Mehta and Mahmoud 2022] | H, Ob, Ov, R, S | Face shape dataset | 70% |
| [Pasupa et al. 2019] | H, Ob, Ov, R, S | Private - 500 images | 70.33% |
| [Moussa et al. 2022] | H, Ob, Ov, R, S | Face shape dataset - 5000 images | 79.2% |
| [Rahmat et al. 2018] | Ov, R, D, R, T, S | Private - 120 images | 80% |
| [Alzahrani 2019] | H, Ob, Ov, R, S | Private - 500 images | 81.1% |
| [Hossam et al. 2021] | H, Ov, R, S, L | Private - 500 images | 82% |
| [Marinescu 2021] | Re, R, S, H, D, T, Ov | Chicago face<br>Private - 290 images | 85% |
| [Zeng et al. 2023] | H, Ob, Ov, R, S | TCM face shape dataset - 6542 images | 85.52% |
| [Rajapaksha and Kumara 2018] | H, Ob, Ov, R, S | Private | 86% |
| [Abdullah et al. 2022] | Ob, Ov, R, S | Private | 86.5% |
| [Duan et al. 2022] | H, Ob, Ov, R, S | Private - 5500 images | 89.8% |
| [Weerasinghe and Vidanagama 2020] | H, Ob, Ov, R, S | Face shape dataset | 91% |
| [Zhao et al. 2020] | H, Ob, Ov, R, S | Private - 600 images | 93.0% |
| [Zeng et al. 2023] | H, Ob, Ov, R, S | SCUT-FBP + FBP5500 - 4490 images | 94.15% |
| [Nabil et al. 2021] | H, Ob, Ov, R, S | CelebA - 2500 images<br>Private - 300 images | 94.9% |
| [Tio 2019] | H, Ob, Ov, R, S | Private - 500 images | 97.8% |
| **Ours** | **H, Ob, Ov, R, S** | **Face shape dataset** | **96.32%** |

*Table 12: Comparison of the results with state of the art*

application in diverse areas such as beauty and cosmetics, healthcare, and security, providing valuable insights and applications in these fields. By achieving high accuracy in a multiclass classification problem, our methodology showcases its relevance and potential impact across various domains where face shape classification is essential. Moreover, our research contributes to the field by improving the state-of-the-art results for face shape classification. The implementation of transfer learning with EfficientNetV2S enables us to build on the existing knowledge base and establish more accurate and efficient classification models. This contributes to the advancement of face shape classification techniques and paves the way for better solutions in related research.

Despite our method's success, we acknowledge some potential limitations and areas for future exploration. One of the main restrictions is the reliance on the availability and quality of the dataset. While our approach performs remarkably well on the current dataset, we recognize the significance of using larger and more diverse datasets to further enhance its accuracy and generalizability. Future research should focus on creating and utilizing such datasets to ensure the broader applicability of our methodology across different populations and demographics. Furthermore, as the field of deep learning rapidly evolves, future studies could investigate alternative neural network architectures to compare their performance against EfficientNetV2S. Additionally, exploring various preprocessing and data augmentation techniques may provide insights into optimizing the model further.

In conclusion, our paper presents a highly efficient and accurate methodology for face shape classification, surpassing the state-of-the-art approaches on the same dataset. We highlight the originality of our research by addressing research gaps and proposing a comprehensive methodology with transfer learning. While acknowledging the potential restrictions and areas for improvement, our work contributes to the field and lays the groundwork for further advancements in face shape classification research.

### Acknowledgements

This research was funded by the project "Development of CSTI platform for retrieval and analysis of structured and unstructured data". The project received funding from the European Regional Development Fund through OP Competitiveness and Cohesion 2014–2020 within the Call for Proposals "Development of the products and services arising from research and development activities—Phase II", under grant number KK.01.2.1.02.0310.

## References

[Abdullah et al. 2022] Abdullah, A. Hussain, S. Ali, H.-C. Kim, M. Sain, and S. Aich, "Hybrid Based Model Face Shape Classification Using Ensemble Method for Hairstyle Recommender System," in Proceedings of 2nd International Conference on Smart Computing and Cyber Security, ser. Lecture Notes in Networks and Systems, P. K. Pattnaik, M. Sain, and A. A. Al-Absi, Eds. Singapore: Springer Nature, 2022, pp. 61–68.

[Agararp 2018] A. F. Agarap, "Deep learning using rectified linear units (relu)," CoRR, vol. abs/1803.08375, 2018. [Online]. Available: http://arxiv.org/abs/1803.08375

[Alzahrani 2019] T. Alzahrani, W. Al-Nuaimy, and B. Al-Bander, "Hybrid Feature Learning and Engineering Based Approach for Face Shape Classification," in 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Dec. 2019, pp. 1–4.

[Belcar et al. 2022] D. Belcar, P. Grd, and I. Tomičić, "Automatic ethnicity classification from middle part of the face using convolutional neural networks," in Informatics, vol. 9, no. 1. MDPI, 2022, p. 18.

[Brownlee 2019] J. Brownlee, Deep Learning for Computer Vision: Image Classification, Object Detection and Face Recognition in Python. Machine Learning Mastery, 2019.

[Deng et al. 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in CVPR09, 2009.

[Duan et al. 2022] J. Duan, X. Su, J. Ren, and L. Xie, "Face Shape Classification Based on Bilinear Network with Attention Mechanism," Journal of Physics: Conference Series, vol. 2278, no. 1, p. 012041, May 2022.

[Goodfellow et al. 2016] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[Grd 2021] P. Grd, "A survey on neural networks for face age estimation," in Central European Conference on Information and Intelligent Systems. Faculty of Organization and Informatics Varazdin, 2021, pp. 219–227.

[Hossam et al. 2021] M. Hossam, A.A. Afify, M. Rady, M. Nabil, K. Moussa, R. Yousri, and M.S. Darweesh, “A Comparative Study of Different Face Shape Classification Techniques on Face Shape and Facial Structure Features, ”International Conference on Electronic Engineering (ICEEM), pp. 1–6, Jul. 2021.

[Keskar et al.2016] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, “On large-batch training for deep learning: Generalization gap and sharp minima,” arXiv preprint arXiv:1609.04836, 2016.

[Kingma and Ba 2014] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[Lama 2023] N. Lama. (2023) Face shape dataset. [Online]. Available: https://www.kaggle.com/datasets/niten19/face-shape-dataset

[Lin et al. 2013] M. Lin, Q. Chen, and S. Yan, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[Marinescu 2021] A.-I. Marinescu, “Automatic Face Shape Classification Via Facial Landmark Measurements,” Studia Universitatis Babe□-Bolyai Informatica, vol. 66, p. 69, Dec. 2021.

[Mehta and Mahmoud 2022] A. Mehta and T. Mahmoud, Human Face Shape Classification with Machine Learning, Aug. 2022.

[Moussa et al. 2022] K. Moussa, M. Wessam, R. Yousri, and M.J. Darweesh, “ Light-Weight Face Shape Classifier for Real-Time Applications on Face Shape and Facial Structure Features, ”International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), pp. 291–295, May. 2022.

[Nabil et al. 2021] M. Nabil, M. Rady, K. Moussa, M. Wessam, M. Hossam, R. Yousri, and M. S. Darweesh, “A Preprocessing Approach to Improve the Performance of Inception v3-based Face Shape Classification,” in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Oct. 2021, pp. 205–209.

[Pasupa et al. 2019] K. Pasupa, W. Sunhem, and C. K. Loo, “A Hybrid Approach to Building Face Shape Classifier for Hairstyle Recommender System,” Expert Systems with Applications, vol. 120, pp. 14–32, Apr. 2019.

[Rahmat et al. 2018] R. F. Rahmat, M. D. Syahputra, U. Andayani, and T. Z. Lini, “Probabilistic Neural Network and Invariant Moments for Men Face Shape Classification,” IOP Conference Series: Materials Science and Engineering, vol. 420, no. 1, p. 012095, Sep. 2018.

[Rajapaksha and Kumara 2018] S. V. Rajapaksha and B. Kumara, “Hairstyle Recommendation Based on Face Shape Using Image Processing,” 2018.

[Russakovsky 2015] 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,” International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.

[Sharma and Mehra] S. Sharma and R. Mehra, “Implications of pooling strategies in convolutional neural networks: A deep insight,” Foundations of Computing and Decision Sciences, vol. 44, no. 3, pp. 303–330, 2019.

[Sukumaran et al. 2021] Sukumaran, A., Brindha, T. Optimal feature selection with hybrid classification for automatic face shape classification using fitness sorted Grey wolf update. Multimed Tools Appl 80, 25689–25710 (2021). https://doi.org/10.1007/s11042-021-10710-9

[Tan and Le 2021] M. Tan and Q. Le, “Efficientnetv2: Smaller models and faster training,” in International conference on machine learning. PMLR, 2021, pp. 10 096–10 106.

[Tan and Le 2019] Mingxing Tan, Quoc V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International conference on machine learning.PMLR, 2019, pp. 6105–6114.

[Ting 2010] K. M. Ting, Encyclopedia of Machine Learning, Confusion Matrix Boston, MA: Springer US, 2010, pp. 209–209. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_157

[Ting 2010] K. M. Ting, Encyclopedia of Machine Learning, Precision. Boston, MA: Springer US, 2010, pp. 780–780. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_651

[Tio 2019] A. E. Tio, "Face Shape Classification Using Inception V3," Nov. 2019.

[Wang et al. 2018] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, "A high-speed and low-complexity architecture for softmax function in deep learning," in 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 2018, pp. 223–226.

[Weerasinghe and Vidanagama 2020] H. Weerasinghe and D. Vidanagama, "Machine Learning Approach for Hairstyle Recommendation," in 2020 5th International Conference on Information Technology Research (ICITR), Dec. 2020, pp. 1–4.

[Zeng et al. 2023] X. Zeng, Z. Yang, and C. Wen, " Region Symmetry Mask for TCM-based Face Shape Classification on Face Shape and Facial Structure Features, "International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 315–319, May. 2023.

[Zhao et al. 2020] J. Zhao, M. Zhang, C. He, X. Xie, and J. Li, "A Novel Facial Attractiveness Evaluation System Based on Face Shape, Facial Structure Features and Skin,"Cognitive Neurodynamics, vol. 14, no. 5, pp. 643–656, Oct. 2020.