Performance Estimation of the State of the Art Convolution Neural Networks (CNN) for Thermal Images-Based Gender Classification System

Muhammad Ali Farooq,[a,*] Hossein Javidnia, [a,b] Peter Corcoran,[a]
[a] National University of Ireland Galway (NUIG), College of Engineering and Informatics, University Road Newcastle, Galway, Ireland, H91TK33
[b] ADAPT Centre, Trinity College, Dublin 2, Dublin, Ireland, D02PN40

Abstract. Gender classification has found many useful applications in the broader domain of computer vision systems including in-cabin driver monitoring systems, human-computer interaction, video surveillance systems, crowd monitoring, data collection systems for the retail sector and, psychological analysis. In previous studies, researchers have established gender classification system using visible spectrum images of the human face. However, there are many factors affecting the performance of these system which includes illumination conditions, shadow, occlusions and, time of day. This study is focused on evaluating the use of thermal imaging to overcome these challenges by providing a reliable means of gender classification. As thermal images lack some of the facial definition of other imaging modalities, a range of state of art deep neural networks are trained to perform the classification task. For this study, the Tufts University thermal facial image dataset was used for training. This features thermal facial images from more than 100 subjects gathered in multiple poses and multiple modalities and provides a good gender balance to support the classification task. These facial samples of both male and female subjects are used to fine-tune a number of selected state-of-art Convolution Neural Networks (CNN) using transfer learning. The robustness of these networks is evaluated through cross-validation on the Carl thermal dataset along with an additional set of test samples acquired in a controlled lab environment using prototype uncooled thermal cameras. Finally, a new CNN architecture, optimized for the gender classification task, GENNet, is designed and evaluated with the pre-trained networks. EfficientNet-B4 outperformed the state of art CNNs with improved performance, achieving an accuracy of 93%.

Keywords: Deep Convolution Neural Networks (DCNN), Thermal imaging, Gender Classification, Long Wave Infrared (LWIR), Transfer Learning

***Muhammad Ali Farooq, E-mail: m.farooq3@nuigalway.ie

1 Introduction

Uncooled thermal imaging is approaching a level of maturity where it can be considered as an alternative to, or as a complimentary sensing modality to that of visible or NIR imaging. Thermal imaging offers some advantages as it does not require external illumination and provides a very different perspective on an imaged scene than a conventional CMOS based image sensor. The proposed research work is carried under HELIAUS[36] project which is focused on in-cabin driver monitoring systems using thermal imaging modality. The driver gender classification in a vehicle can help to improve personalization of various features (e.g. user interfaces, presentation of data

to the driver). It can also be used to better predict driver cognitive response[49], driver behavior and intent, and finally knowledge of gender can be useful for safety systems such as airbag deployment which may adapt to driver physiology. In summary, automotive manufacturers are interested to have the knowledge of driver gender within the vehicular environment for designing smarter and safer vehicles. Alongside this, there are many other applications of thermal human gender classification systems. In security systems thermal imaging can easily detect people and animals even in total darkness. In human-computer interaction systems, thermal imaging can provide complimentary information, determining subtle fluctuations in facial temperatures that can inform on the emotional status of a subject. In other human-computer interaction systems, the systems may need to classify the individual person and/or their facial expressions in order to effectively interact with them thus gender information serves as a source of soft biometrics[64]. In medical applications human thermography provides an imaging method to display heat emitted from a human body surface thus helping us to understand unique facial thermal patterns in both male and female gender[60]. Human thermography helps us to better understand that central and peripheral thermoreceptors are distributed all over the body including human face and are responsible for both sensory and thermoregulatory responses to maintain thermal equilibrium. Studies have shown that heat emission from the surface of the body is symmetrical. All these studies measured differences between the left and right side of different areas of the head[24, 46, 61].

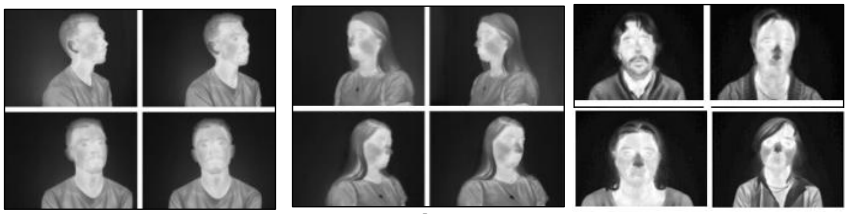
The literature reports that in healthy subjects the difference in skin temperature from side to side of the human body is as small as 0.2 degree centigrade[61]. Heat emission from the human body is related to cutaneious vascular activity, yielding enhanced heat output on vasodilation and reduced heat amount on vasoconstriction[62]. The medical literature reports that a significant difference has been observed between the absolute facial skin temperature of men and women during the clinical

studies of facial skin temperature[62]. Men were found to have higher temperature compared to women over all 25 anatomic areas measured on the face which includes upper lips, lower lips, chin, orbit, the cheek etc. According to another study, the basal metabolic rate of healthy 30 year old male with the height of 5 ft, 7 in weight of 64 kg, and who has surface area of about 1.6 m^2, dissipates about 50 W/m^2 of heat whereas on the other hand the basal metabolic rate of healthy 30 year old female with the height of 5 ft, 3 in, weight of 54 kg and who has surface area of 1.4 m^2, dissipates about 41 W/m^2 of heat. In addition to that women skin is expected to be cooler since less amount of heat is lost/unit (per area of body surface)[62]. However, thermal patterns whether in case of male or female also depends on many other factors such as age, human body intrinsic and extrinsic characteristics, outdoor environmental conditions and technical factors such as camera calibration, field of view (FoV). Moreover, it also depends on factors such as drinking, smoking, various diseases, using medications etc.

The preliminary focus of this study is on binary human gender classification, however the same system can be retrained for third or multi-class (non-binary) gender classification task if such datasets are available.

In this study the Tufts thermal faces[20–22] and Carl thermal faces datasets[23–24] are used to train and test a selection of state-of-art neural networks to perform the gender classification task. Figure. 1 shows some examples of thermal facial images with varying poses from the Tufts dataset and frontal facial poses from the Carl dataset. The complete workflow pipeline is detailed in Sec. 3 of this paper. In addition to using pre-trained neural networks, a new CNN architecture, GENNet, is provided. This is designed and trained specifically for the gender classification task and is evaluated against the pre-trained CNN networks. In addition, a new validation set of thermal images is acquired in controlled laboratory conditions using a new prototype uncooled thermal

camera and is used as a second means of cross-validating all the pre-trained models along with GENNet architecture. The evaluation results are presented in Sec. 4.



a b c

Fig. 1 Sample Images from Tufts and Carl thermal face database: (a) male subject with 4 different face poses from tufts dataset, (b) female subject with 4 different face poses from tufts dataset and (c) male and female subjects (frontal face pose) from Carl database

2 Background/ Related Work

This section focuses on background research and previous studies on gender classification using CNNs

2.1 Gender classification using Conventional Machine Learning Methods

E Makinen et al. [4] and Reid, D. A., et al.[5] provide a detailed survey of the gender classifications method in their studies. One of the early techniques for gender recognition reported in Ref 6 utilized a neural system trained on a small arrangement of close frontal face pictures. In[7] the consolidated 3D structure of the head (captured by a laser scanner) and picture intensities were utilized for characterizing genders. Support Vector Machine (SVM) classifiers were employed by[8] where the authors evaluated the performance of SVM with an overall error rate of 3.4% when compared with other traditional classifiers which include linear, fisher linear discriminant, nearest neighbor, and Radial Basis Functions (RBF). Instead of using SVM, Ref. 9 referred to AdaBoost for gender classification tasks using a set of low-resolution grayscale images. Perspective invariant age and gender recognition was performed by[10] using arbitrary viewpoints. Recently Ihsan et al. [11], utilized the Webers Local surface Descriptor[12] for the gender recognition system,

showing near-perfect execution on the FERET benchmark[13]. In Ref. 14, shape, texture, and color features were extracted from frontal faces, thus obtaining robust outcomes on the FERET benchmark. In an attempt by Arun et al. [15], unique mark pictures are used, and the input images were are represented by a feature vector consisting of ridge thickness to valley thickness ratio (RTVTR) and the ridge density. Further, they used SVM to categorize subjects into male and female classes accordingly. In addition to the gender classification system using the visible spectrum, the possibility of deducing gender information from thermal and NIR spectrum is also gaining much interest. Chen C, Ross et al. [53] claims to be the first proposing human faces-based gender classification system using thermal and NIR data. The authors have selected three different conventional feature extraction methods for gender representation which include Linear Binary Patterns (LBP), Principle Component Analysis (PCA), and pixels from low-resolution facial images. For gender recognition, they have used SVM, LDA, Adaboost, Random Forest, Gaussian Mixture Model (GMM), and Multi-Layer Perceptron (MLP) classifiers. Their experimental results conclude that SVM for histogram-based gender classification results in much better performance on NIR and thermal spectra. In Ref. 43, the authors proposed a gender classification system using joint visible and thermal spectrum data of the human body. The classification accuracies in Ref. 43 are measured by employing different feature extractors including HoG and MLBP [48]. Their experimental results demonstrated an improvement in classification accuracy using the joint data from visible and thermal image spectrums. Similarly, another study reported in Ref. 44, the author's utilized multimodal datasets consisting of audiovisual, thermal, and physiological recordings of male and female subjects. The authors extracted feature values from these datasets which were later used for automatic gender classification purposes. In both studies, authors used

conventional machine learning algorithms for feature extraction rather than using advanced deep learning methodologies.

2.2 Gender Classification using Deep Learning-based Methods

Due to the fact that much potential is laid in deep CNN structures, they are widely used for diversified applications especially where more precise and robust accuracy levels are required such as medical image analysis, surveillance systems, object detection, and autonomous classification systems. Canziani et al. [16] in his study listed many pre-trained models that can be used for various practical applications. He analyzed the overall performance of these pre-trained models by computing the accuracy levels and the inference time needed for each model. Dwivedi, N., & Singh, D. K in Ref. 54 provides a comprehensive review of deep learning methodologies for robust gender classification using GENDER-FERET[25] face dataset. In their study, they have compared the performance of various CNN architectures. Moreover, they have selected one of the architectures as a baseline model and by changing different parameters like the number of fully connected layers and the number of filters they have created different models. The authors achieved the best accuracy of 90.33% with the base model architecture of CNN. In Ref. 55 authors have investigated two different deep learning strategies which include fine-tuning and SVM classification using CNN features. They were applied on different networks which include their proposed task-specific GilNet model and pre-trained domain-specific VGG[32] and Generic AlexNet[33]-like CNN model for building robust age and gender classification system using Adience[58] visible spectrum dataset. The experimental results from their study show that transferred models outperform the GilNet model for both age and gender classification tasks by 7% and 4.5% respectively. In a more recent study by Anirudh, et al. [45] authors have investigated the overall performance of two CNN based methods for gender classification using Near-Infrared (NIR)

images. In the first method, a pre-trained VGG-Face[59] was used for extracting features for gender classification from a convolutional layer in the network, whereas the second method used a CNN model obtained by fine-tuning VGG-Face to perform gender classification from periocular images. The authors had achieved the classification accuracy of 81% on an in-house dataset which was gathered locally.

Further in a more recent study by Baek, Na Rae, et al. [56] authors have used the combined data of both visible and NIR spectrum for performing robust gender classification using full human body images in surveillance environment. The system works by deploying two CNN architecture to remove the noise of visible-light images and enhance the existing image quality to improve gender recognition accuracy. The overall system performance was evaluated on desktop pc as well as on Jetson TX2 embedded system.

3 Research Methodology

The goal of this research work is to evaluate the potential of thermal image facial data as a means of gender classification. The thermal image data is analyzed with a selected set of nine state-of-the-art neural networks. These pre-existing convolution neural networks are adapted for the thermal data using transfer learning. In addition, a new CNN model is proposed, and its performance is compared against nine state-of-art pre-trained networks.

Initially, all the pre-trained networks are first trained on the Casia Face dataset[34] since Tufts thermal training dataset[20–22] does not contain sufficient number of images, an important requirement for optimal training of deep neural networks. This face dataset is used to extract low-level features for building the baseline architecture. In the second stage the tufts thermal face database[20–22] is used for transfer learning. This dataset consists of 113 different subjects and comprises images from six different image modalities that include visible, near-infrared, thermal,

computerized sketch, a recorded video, and 3D images of both male and female classes. The thermal face dataset was acquired in a controlled indoor environment using constant lighting that was maintained using diffused lights. Thermal images were captured using FLIR Vue Pro Camera[40] which was mounted at a fixed distance and height.

Figure. 2 represents the complete workflow diagram of the overall gender classification system.

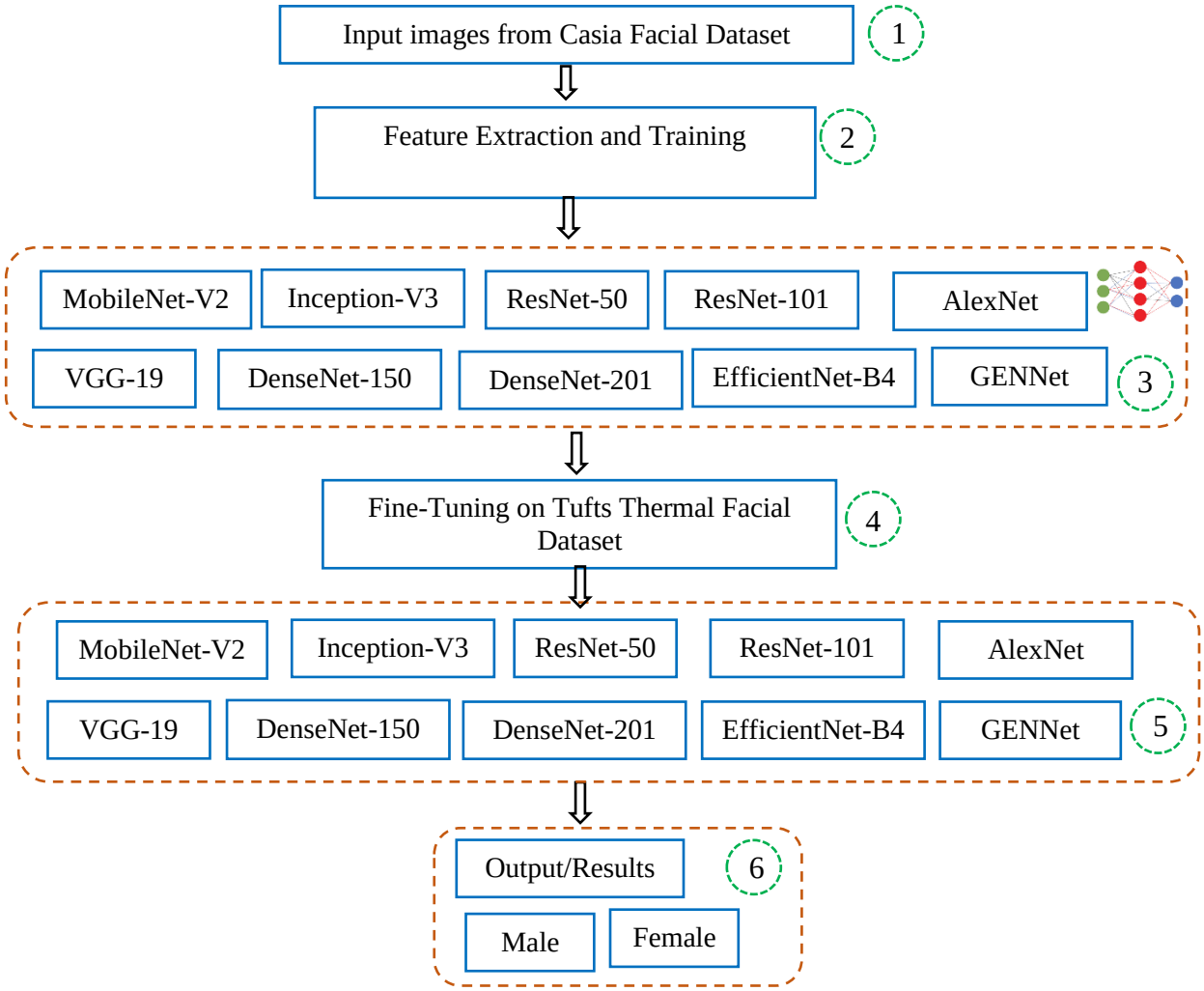


Fig. 2 Workflow diagram for autonomous gender classification system using thermal images

3.1 Initial Training and Transfer learning of pre-trained networks

This research takes advantage of the pre-trained networks by freezing and unfreezing all the layers and adding customized final layers to generalize the model for the target autonomous gender classification task from thermal image datasets. The main reason for using these pre-trained networks is they already learned low-level feature values such as edges and textures by training the networks on very large and varied datasets. This process helps in obtaining useful results even with a relatively small training dataset since the basic image features have already been learned by the pre-trained model using larger datasets like ImageNet[19]. Further, the classifier is trained to learn the higher-level features in the proposed thermal dataset images.

A typical CNN system comprises certain layers which include convolution layers, pooling layers, dense layers, and fully connected layers. There are various pre-trained networks available that can be efficiently used for different types of visual recognition, object detection, and segmentation tasks. For the proposed study the following pre-trained neural networks are utilised: ResNet-50[17], ResNet-101[17], Inception-V3[31], MobileNet-V2[30], VGG-19[32], AlexNet[33], DenseNet-121[29], DenseNet-20[29] and EfficientNet-B4[63] networks. These models are chosen as they are commonly trained using ImageNet[19] dataset each model has a different architectural styles, they provide a good trade-off between accuracy and inference time[28] and in addition are they are the state-of-the-art for image classification tasks. Thus, an impartial performance comparison of these networks can be made for the thermal gender classification task.

ResNet[17] architecture mainly relies on the residual learning process. The network is designed to solve complex visual tasks using more deeper layers stacked together. ResNet-50 is a 50-layer Residual Network. The other variants from ResNet family include ResNet-101[17] and ResNet-152[17]. Resnet-50 network was initially trained on ImageNet[19] which consists of a total of 1.28

million images from 1000 different classes. The Inception-v3 is made up of 48 layers stacked on top of each other[31]. The Inception-v3 model was initially trained on Imagenet[19] as well. These pre-trained layers have a strong generalization power as they are able to find and summarize information that will help to classify various classes from the real-world environment.

MobileNet-V2 is considered as one of the finest deep learning architectures proposed by Howard et al. [30] specifically designed for mobile and embedded vision applications. It is a lightweight deep learning architecture with the working principle of using depth-wise separable convolutions meaning that it performs single convolution operation on each color channel rather than combining all three and flattening them. This has the advantage of filtering the input channels.

DenseNet[29] architecture also referred to as Dense Convolutional Neural Network is a state of art variable-depth deep convolutional neural architecture. It was designed to improve the architecture of ResNet[17]. The principle design feature of this architecture is channel-wise concatenation, with every convolution layer which has access to the activations of every layer preceding it. DenseNet family has different variants which include DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264.

VGGNet[32] was developed by Visual Geometry Group from the University of Oxford. Like ResNet[17] and Inception-V3[31], this network was also originally trained on ImageNet[19]. The network was designed with significant improvement compared to AlexNet architecture[33] which was more focused on smaller window sizes and strides in the first convolutional layer. VGG architecture can be trained using images with (224×224) pixel resolution. The main attribute of VGG architecture is that it uses very small receptive fields (3×3 with a stride of 1) compared to AlexNet[33] (11×11 with a stride of 4). In addition to this, VGG incorporates 1×1 convolutional layers to make the

decision function more non-linear without changing the receptive fields. The architectures come in different variants which include VGG-11, VGG-16, and VGG-19.

EfficientNet[63] is recently published and is designed using compound scaling method. As the name suggests the network proved to be computationally efficient by achieving state of art result on the ImageNet dataset. Table 1[3] provides a more comprehensive comparison of these architectures highlighting their attributes, number of parameters, overall error rate on benchmark datasets and their respective depth.

Table 1 Performance comparison of state-of-the-art CNN

CNN	Number of Parameters	Top 5 Error Rate	Depth	Main Attributes
AlexNet	62 M	ImageNet: 16.4	8	Uses Relu, dropout and overlap Pooling
VGGNet	138 M	ImageNet: 7.3	19	Homogenous topology, uses small size kernals
Inception-V3	24 M	ImageNet: 3.5	159	Replace large size filters with small filters
MobileNet	2.2 M	ImageNet: 10.5	17	The width multiplier uniformly reduces the number of channels at each layer, fast inference
ResNet-50 ResNet-101	26 M 43 M	ImageNet: 3.6	152	Residual learning, Identity mapping-based skip connection
DenseNet-121 DenseNet-201	7.2 M 18.6 M	CIFAR-10+: 3.46	190	Cross-layer information flow
EfficientNet-B4	19M	ImageNet: 2.9		Compound coefficient scaling method, 8.4x smaller and 6.1x faster than other convnets

As discussed in the previous section, all the pre-trained networks are initially trained on the Casia Face database[34] since the Tufts thermal training dataset[20–22] does not contain a sufficient number of images. Casia facial dataset[34] consists of facial images of different celebrities (38,423 distinct subjects) in the visible spectrum. This facial dataset has been used to extract low-level feature values for building a baseline architecture. The networks are trained using a total of 30,887 frontal facial images of different celebrities from both genders. The data was split in the ratio of 90% for training and 10% for validation. To better generalize and regularize the base model for final fine-tuning on the thermal dataset, certain data transformations are performed on the Casia[34] training

data which includes random resizing of 0.8, random rotation of 15 degrees and flipping. The logic for performing these transformations is that it will bring supplementary data variations for optimal training of the baseline architectures keeping in view the final fine-tuning process on thermal images. Figure. 3 displays the Casia data samples along with training data transformation results. The initial training is done by adding a small number of additional final layers to enable generalization and regularization of all the pre-trained models. In the case of ResNet-50 and ResNet- 101 networks, the last fully connected layer is connected to a linear layer having 256 outputs. It is further fed into the Rectified Linear Unit (ReLU) [41] and dropout layers with the dropout ratio of 0.4 followed by a final fully connected layer which has binary output corresponding to the 2 classes in the Casia dataset. A similar formation of final layers is inserted by transforming the number of features to the number of classes in all the pre-trained networks. Each of these networks is further fine-tuned using a training dataset comprising of thermal facial image samples. The fine-tuning is achieved using transfer learning techniques.

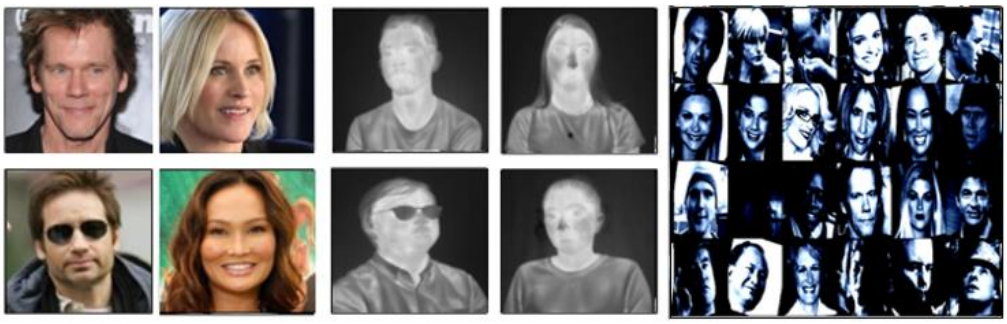


Fig. 3 Facial samples from two different datasets: (a) male and female data samples from Casia[34] database, (b) male and female samples from tufts thermal images[20–22] and (c) PyTorch data transformations on Casia dataset

The models were trained using the PyTorch framework[26]. Binary Cross-Entropy is used as the loss function during training along with a Stochastic Gradient Descent (SGD)[50] optimizer. The final training data includes male and female thermal images as shown in Fig. 4.

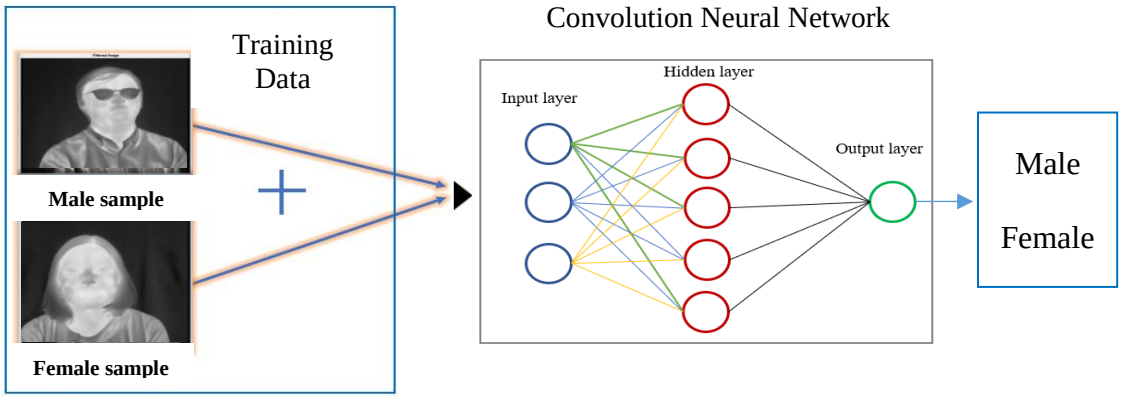


Fig. 4 Training data comprising of male and female samples for network training

In order to better fine-tune the networks, the thermal training data is augmented by introducing a selection of image variations. These are achieved using the transformation operations shown in Table 2.

Table 2 Training data transformation

Transformation Type	Data Variation
Resized cropping	Size=256, scale= (0.8, 1.0)
Rotation	15 degree
Flipping	Horizontal
Center cropping	Size: 224
Tensor Conversion	---
Mean and standard deviation normalization	[0.485, 0.456, 0.406], [0.229, 0.224, 0.225]

During the fine tuning phase the Stochastic Gradient Descent[50] and the Adam[51] optimizers are used to compare their respective performance. This is discussed in the experimental results (Section 4) of this paper. As compared to Gradient descent **(**GD) where the weights are updated in batch as the whole, in SGD[50] the weights are updated based on each training iteration thus the weights are updated element-wise. Moreover, SGD[50] is computationally less expensive and minimizes losses faster than GD since it doesn't cycle through the entire training dataset to update the weights. The Adam[51] optimizer is an adaptive learning rate optimizer and is considered to be one of the best optimizer for training convolution neural networks. As compared to SGD the Adam optimizer

implements an adaptive learning rate and can determine an individual learning rate for each parameter. Figure. 5 shows the generalized training structure for all the pre-trained networks. The training data is split in the ratio of 80% and 20% for training and validation purposes respectively. To achieve a fair evaluation baseline, all the pre-trained networks are fine-tuned using the same hyper-parameters on the one train dataset. These parameters are provided in Table 3.

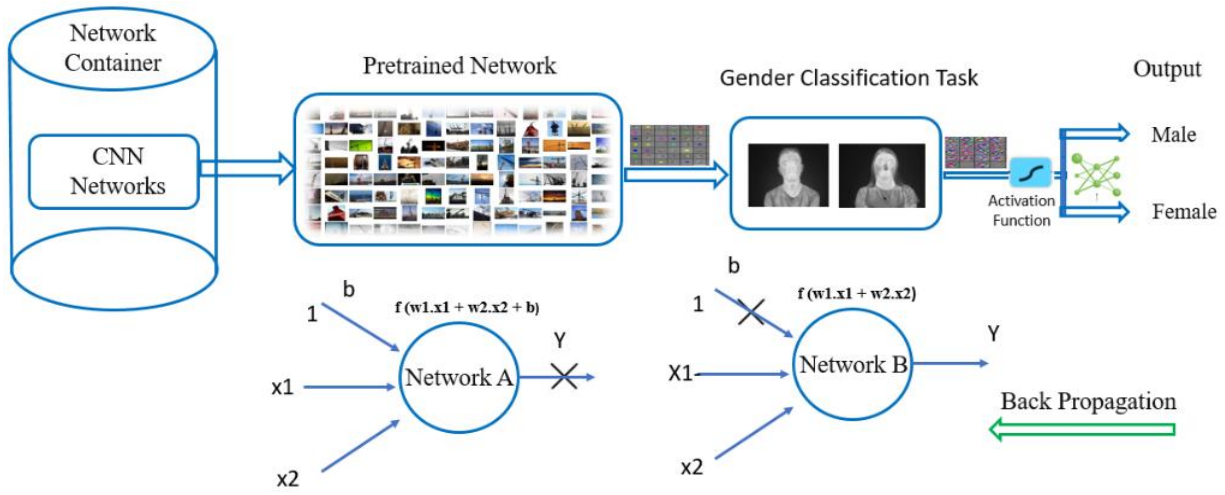


Fig. 5 CNN training structure, Network A indicates pre-trained networks with initial weights, Network B indicates transfer learning process with new weights for thermal gender classification

Table 3 Pretrained Networks Hyperparameters

Network Hyperparameters	
Batch Size	32
Epochs	100
Learning Rate	0.001
Momentum	0.9
Loss Function	Cross-Entropy
Optimizer	Stochastic Gradient Descent (SGD) and Adam

3.2 New CNN model GENNet

To analyze the validity of the existing thermal images, a novel CNN network is designed which is referred to as GENNet and its performance is compared against the pre-trained state of the art

architectures. The structural block diagram representation of the proposed network is shown in Fig. 6. The overall network structure is consisting of four main blocks. The first three blocks contain sequential layers in the form of 2D convolutions each followed by the Rectified Linear

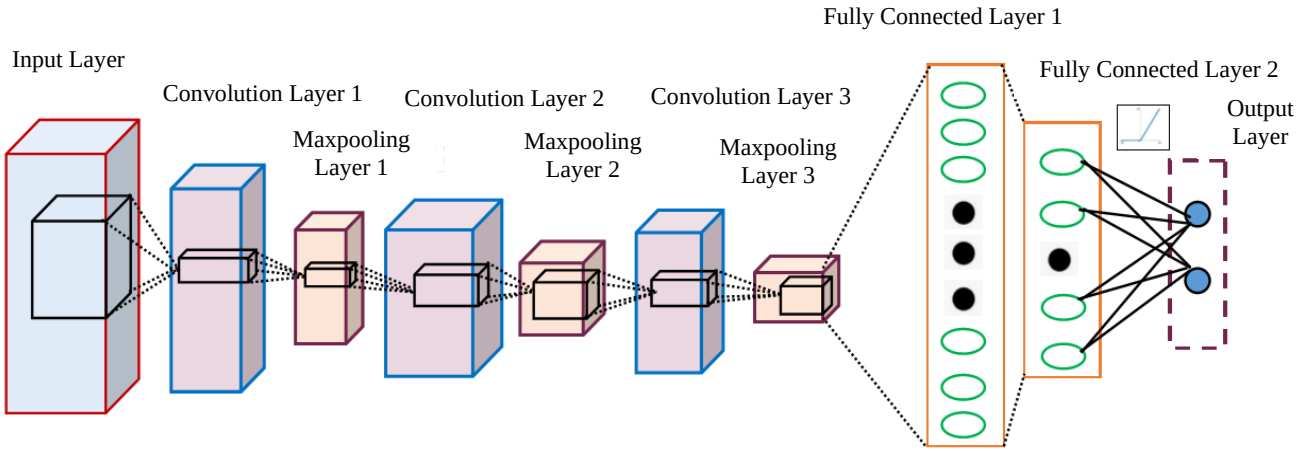


Fig. 6 Structural representation of GENNet CNN model for thermal images-based gender classification

Like all other pretrained networks, GENNet is initially trained on the Casia Facial database[34] and later fine-tuned on tufts thermal dataset[20–22]. The same division of thermal training data is used along with the same hyperparameters as it was utilized for other pre-trained models. Once the network is fine-tuned, it is tested on the combination of two new datasets as discussed in Sec. 4.3.

4 Experimental Results

PyTorch[26] deep learning platform is used to fine-tune and train all the pre-trained models as well as the proposed GENNet model. These experiments are performed on a machine equipped with NVIDIA TITAN X Graphical Processing Unit (GPU) with 12GB of dedicated graphic memory.

4.1 Training and validation results of CNN's architecture using SGD optimizer

The first part of the experiments includes analyzing the results of the nine state of the art deep learning networks by freezing the network layers as discussed in Sec. 3.1. Figure. 7 presents the overall performance of all the pre-trained architectures initially trained on Casia dataset[34] and fine-

tuned on thermal facial images from tufts dataset[20–22], using SGD optimizer. It is important to mention that during the training phase the data is divided subject-wise and all the eight poses of each particular subject are used for training and validation purposes respectively. This is done to avoid bias and doing optimal inductive learning. Figure. 7 presents the training and validation accuracy and loss chart of all the pre-trained models.

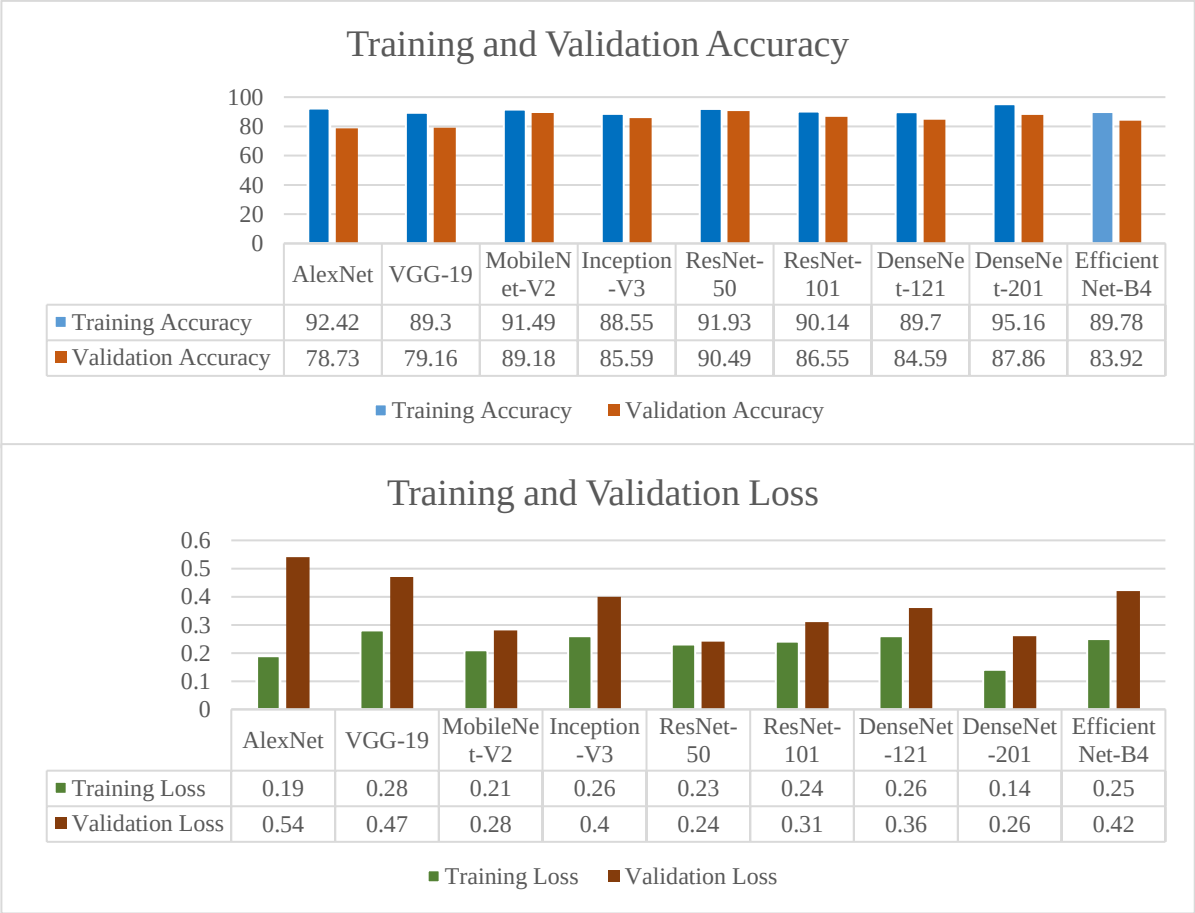
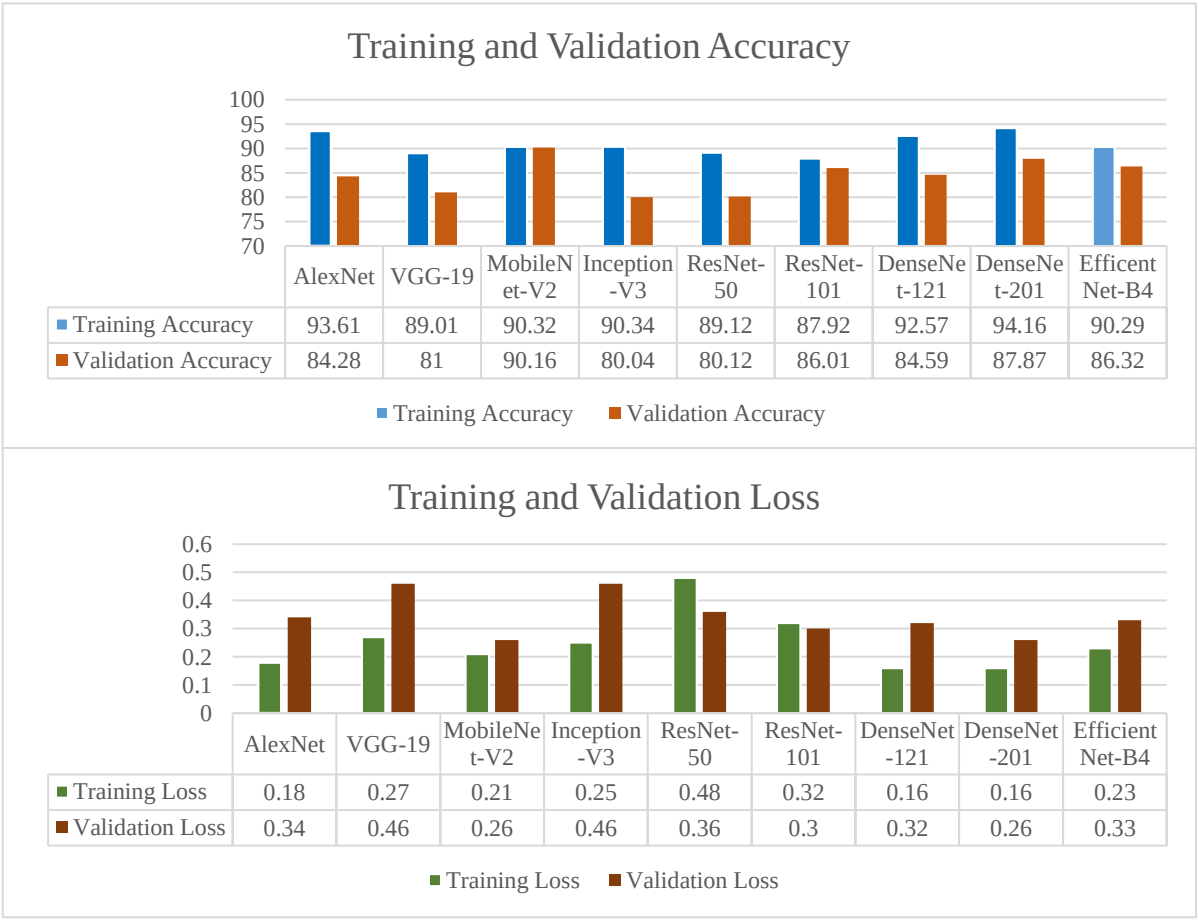


Fig. 7 Accuracy and loss charts of all the networks trained using SGD optimizer

Among all the models ResNet-50 architecture scores highest with the validation accuracy of 90.49% followed by MobileNet-V2 with validation accuracy of 89.18%. However, AlexNet and VGG architectures do not perform well as compared to other models thus getting the lower validation accuracy and higher loss values.

4.2 Training and validation results of CNN's using Adam optimizer

The second part of the experiments includes analyzing the results of the nine state of the art deep learning networks by freezing the network layers and using Adam optimizer. The same training procedure is followed as discussed in Sec. 4.1. Figure. 8 shows the accuracy and loss chart of all the pre-trained models.



	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	Efficent Net-B4
Training Accuracy	93.61	89.01	90.32	90.34	89.12	87.92	92.57	94.16	90.29
Validation Accuracy	84.28	81	90.16	80.04	80.12	86.01	84.59	87.87	86.32

	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	Efficient Net-B4
Training Loss	0.18	0.27	0.21	0.25	0.48	0.32	0.16	0.16	0.23
Validation Loss	0.34	0.46	0.26	0.46	0.36	0.3	0.32	0.26	0.33

Fig. 8 Accuracy and loss charts of all the networks using ADAM optimizer

The accuracy and loss chart in Fig. 8 reflect an increase in training and validation accuracy and lower loss values for most of the models compared to the models trained using SGD optimizer. However, it was not possible to achieve an optimal training outcome as most of the models have accuracy levels below 95%. By analyzing the accuracy and loss charts in Fig. 8 it is clear that during the finetuning process of all the pre-trained models DenseNet-201[29] and AlexNet achieves

the highest training accuracies of 94.16% and 93.61% with lowest training losses of 0.16 and 0.18 respectively. Whereas MobileNet-V2[30] architecture achieved the best validation accuracy of 90.16% with validation loss of 0.26. DenseNet-201 model scored best with validation accuracy of nearly 88%. The inception-V3 architecture was unable to achieve good accuracy scores compared to the other pre-trained models with overall validation accuracy of only 80.04% and the highest validation loss of 0.46.

4.3 Training and validation results of CNN's architecture by unfreezing the layers

As explained above, it is clear that transfer learning while freezing the network layers and using both SGD and ADAM optimizer we cannot achieve an optimal training and validation accuracy in case of most of the models as shown in Fig. 7 and Fig. 8 respectively. In the next stage of this experimental study all the networks are re-trained by unfreezing all the original network layers to improve the feature learning process on thermal data. During this fine-tuning process both Adam and SGD optimizers were employed and the best results in the case of each model were selected. Most of the models performed well, achieving better training and validation accuracy as shown in Fig. 9. AlexNET is specifically trained using a fixed learning rate and it utilizes a one cycle learning policy to achieve a better convergence. The initial learning rate of the network is set to 0.001 and momentum to 0.9. Using a smaller learning rate makes a model converge more efficiently but at the expense of the speed, while using a higher learning rate can lead to model divergence. Thus, in order to overcome this issue, the learning rate needs to be adjusted automatically. One cycle LR is useful for finding an optimal learning rate during the complete training process of a CNN. The main goal of performing these techniques is to optimize all the models as well as that of the newly proposed GENNET architecture. Figure. 9 shows the accuracy and loss chart of all the re-trained networks along with newly proposed GENNet architecture.

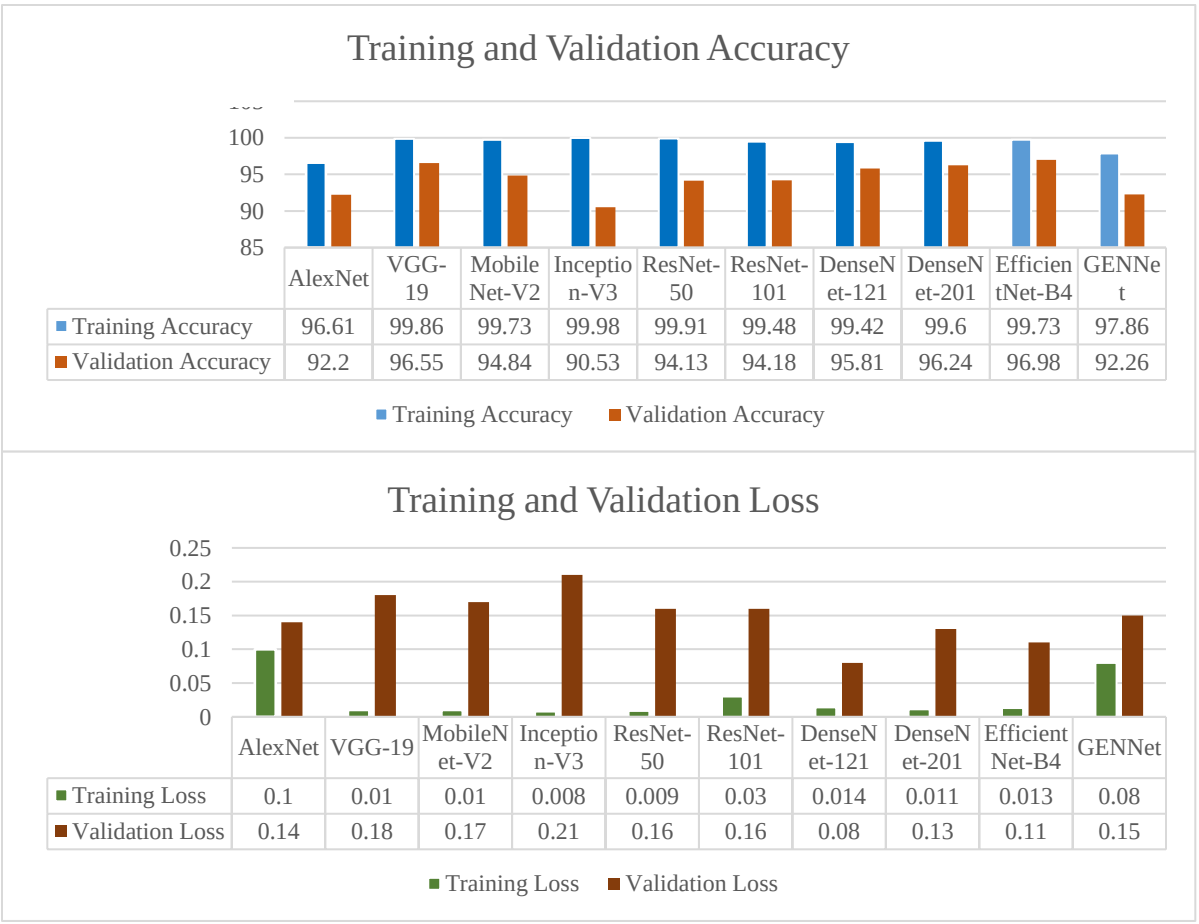


Fig. 9 Accuracy and loss charts of all the networks by unfreezing the network layers

It can be observed that most of models performed significantly well by getting training accuracy above 96% and validation accuracy greater than 90%. The trained models are further used for cross-validating their performance on a new test data as discussed and shown in next sections of experimental results.

4.4 Local thermal data acquisition

To further validate the effectiveness of all the pre-trained models and provide an additional mode of comparison with the new proposed CNN GENNet model, a live thermal facial dataset was gathered using a new prototype thermal camera. The data is acquired in an indoor lab environment using a camera-based on a prototype uncooled micro-bolometer thermal camera array that embeds

a Lynred[35] Long Wave Infrared (LWIR) sensor developed under the Heliaus EU project[36]. Figure. 10 displays the prototype thermal camera model being used for the proposed research work to gather this live dataset and whereas Table 4 provides the technical specifications of the camera.

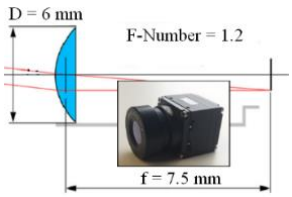
Fig. 10 Prototype thermal VGA camera model for acquiring local facial data

Table 4 Technical specifications

Prototype Thermal Camera Specifications	
Quality and Type	VGA, Long Wave Infrared (LWIR)
Resolution	640 x 480 pixels
Focal length (f)	7.5 mm
F-Number	1.2
Pixel Pitch	17 um
HFOV	90-degree, 890 mm

To take comprehensive facial information during the data acquisition process, we have calculated other important parameters including the lens aperture, angular Field of View (AFOV), height and width of the sensor, and working distance as shown below[37].

$$F - Number = \frac{Focal\ Length\ (f)}{Diameter\ (D)} \quad (1)$$



$$Diameter\ (D) = \frac{Focal\ Length\ (f)}{F\ Number} = \frac{7.5}{1.2} = 6.25\ \approx\ 6\ mm \quad (2)$$

$$Height\ of\ Sensor\ (h) = Horizontal\ Pixels * Pixels\ Pitch = 640 * 17 = 10.88\ mm \quad (3)$$

$$Width\ of\ Sensor\ (w) = Vetricle\ Pixels * Pixels\ Pitch = 480 * 17um = 8.16\ mm \quad (4)$$

$$AFOV = 2 * tan^{-1}\frac{h}{2f} = 2 * tan^{-1}\frac{10.88\,mm}{2 * 7.5\,mm} = 71.9 \approx 72\,Deg \quad (5)$$

$$Working\ Distance\ (WD) = \frac{Focal\ Length\ (f) * \ HFOV}{height\ of\ Sensor\ (h)} = \frac{7.5 * 890}{10.88} \approx 60\ cm \quad (6)$$

The data is collected by mounting a camera on a tripod at a fixed distance of 60-65 cm. The height of the camera is adjusted manually to align the subject's face centrally in the field of view. Shutterless[57] camera calibration at 30 FPS is used to acquire the data. The data acquisition setup is shown in Fig. 11. A total of five subjects consensually agreed to take part in this study. The data was gathered by recording videos stream of each subject covering different facial poses and then generating image sequences from the acquired videos.

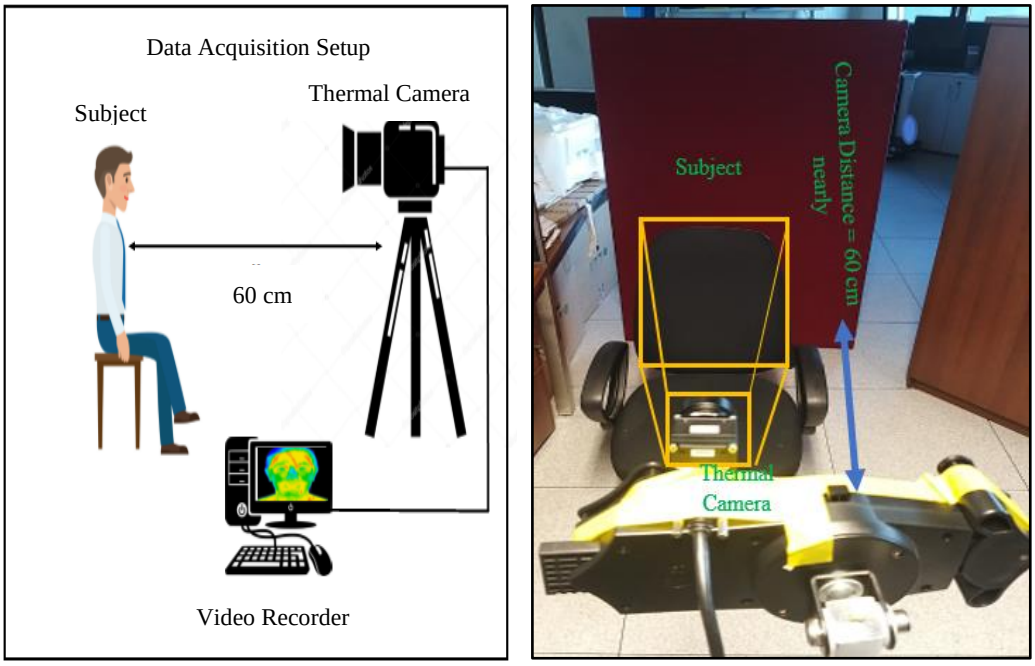


Fig. 11 Indoor lab environment data acquisition setup

Figure. 12 illustrates a few samples of the captured data including both male and female subjects.

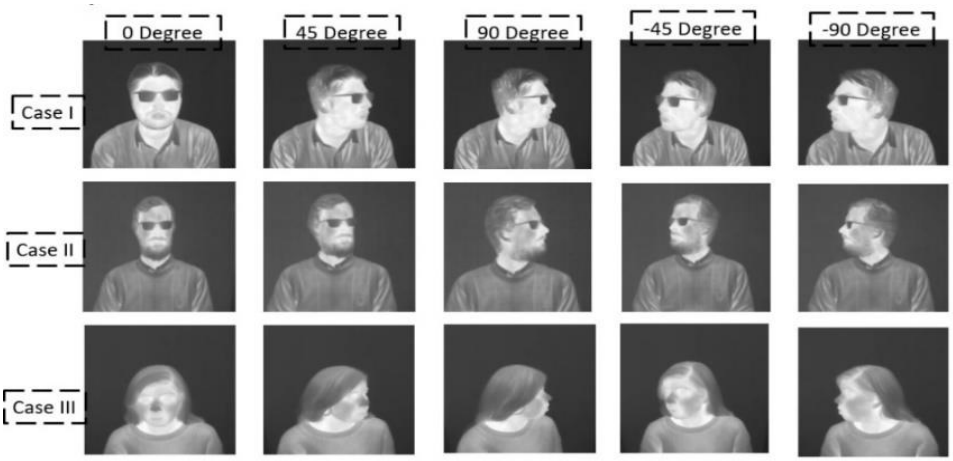


Fig. 12 Test cases of three different subjects acquired in the lab environment with varying face poses, first two-row show the varying facial angles of male subjects, the last row shows the different facial angles of a female subject

4.5 Testing results of state of the art CNN

All the trained models are tested on the combination of the two different datasets including Carl[23–24] and the locally gathered indoor thermal dataset. This is done in order to cross-validate the effectiveness of all the trained classifiers, as discussed in Sec. 1. The best models achieving the highest training and validation accuracy from Sec. 4.3 are selected for cross-validation experiment. The test data contains a total of ninety samples. The overall performance of all the networks on test data is measured using the accuracy metric as shown in Eq. (7)[27].

$$Accuracy\ (ACC) = \frac{tp + tn}{tp + tn + fp + fn} \times 100 \quad (7)$$

Where $tp, fp, fn, and\ tn$ refer to true positive, false positive, false negative, and true negative. ACC in (7) means overall testing accuracy.

Figure. 13 illustrates the calculated test accuracy chart of all the models trained on both original data and processed (fused) data. A confusion matrix for five of the best models is presented in Fig. 14 to better elaborate on the performance of each model on different genders.

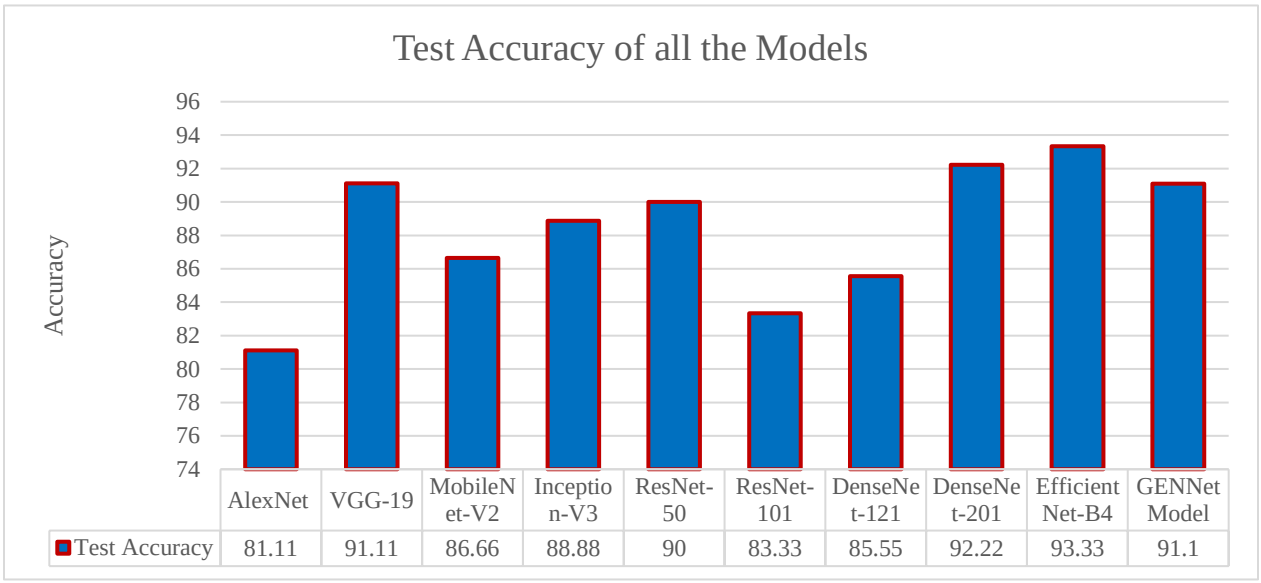


Fig. 13 Test accuracy chart of all the CNN architectures

VGG-19 N = 90	Class: Males True Positives	Class: Females True Negatives
Predicted Positives	51	4
Predicted Negatives	4	31

a

ResNet-50 N = 90	Class: Males True Positives	Class: Females True Negatives
Predicted Positives	51	5
Predicted Negatives	4	30

b

DenseNet-201 N = 90	Class: Males True Positives	Class: Females True Negatives
Predicted Positives	54	3
Predicted Negatives	1	32

c

EfficientNet-B4	Class: Males True Positives	Class: Females True Negatives
Predicted Positives	50	1
Predicted Negatives	5	34

d

GENNet	Class: Males True Positives	Class: Females True Negatives
Predicted Positives	54	7
Predicted Negatives	1	28

e

Fig. 14 Confusion Matrix depicting the performance of (a) VGG-19, (b) ResNet-50, (c) DenseNet-201, (d) EfficientNet-B4 and (e) GENNet models

Figure 15 shows a number of failed predictions by the studied state of the art models. The results display the model name along with predicted output class.

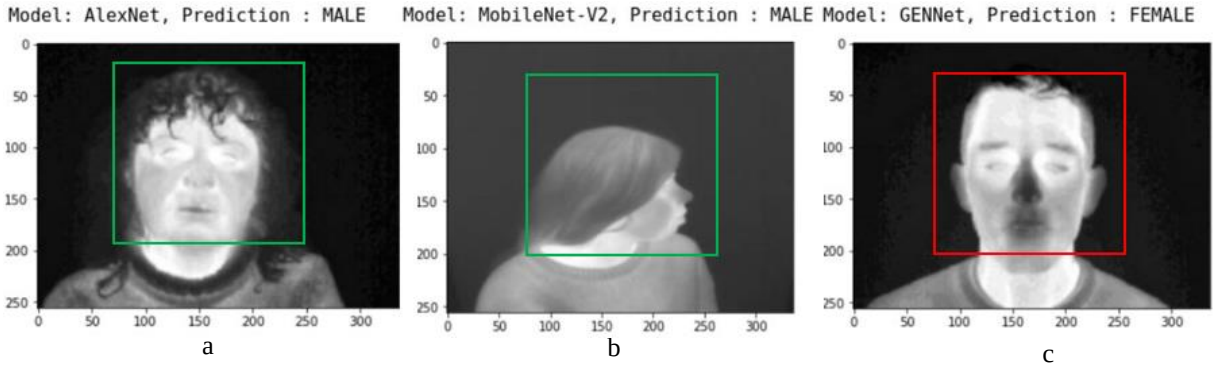


Fig. 15 Individual false prediction test case results (a) AlexNet model: female gender mis-classified as male gender, (b) MobileNet: female gender mis-classified as male gender, (c) GENNet: male gender mis-classified as female gender

In order to understand how effective, the models are for custom classification task, eight different quantitative metrics are employed in addition to the accuracy metrics thus providing a detailed performance comparison of all the trained models. The additional metrics include sensitivity, specificity, precision, negative predictive value, False Positive Rate (FPR), False Negative Rate (FNR), Matthews Correlation Coefficient (MCC) and F1-score. Sensitivity, specificity and precision are the conditional probabilities where sensitivity also termed as recall is defined as probability of given positive example results in positive test, specificity is the probability of given negative example results in negative test whereas precision provides what proportion of positive identifications was actually correct. The False Positive Rate (FPR) is the proportion of negative cases incorrectly identified as positive cases in the data whereas False Negative Rate (FNR) also known as miss rate is the proportion of positive cases incorrectly identified as negative cases. F1-score describes the preciseness (such that how many instances it predicts correctly) and robustness (such that it does not miss a significant number of instances) of the classifier. Matthews Correlation Coefficient (MCC) produces a more informative and reliable statistical score in evaluating binary classifications in addition to accuracy and F1-score. It produces a high score only if the trained classifier obtained good results in all the four confusion matrix categories which includes true

positives, false negatives, true negatives, and false positives. The numerical results are presented in Table 5. The best and worst value per metric is highlighted in green and light orange.

Table 5 Different Quantitative Metrics. The best value per metric is highlighted in green and emboldened, the worst value per metric is highlighted in light orange.

Quantitative Metrics Comparison of all the Models								
Models	Sensitivity	Specificity	Precision	Negative Predictive Value	False Positive Rate	False Negative Rate	F1-Score	Matthews Correlation Coefficient
AlexNet	**0.98**	0.54	0.77	0.95	0.45	**0.02**	0.86	0.61
VGG-19	0.93	0.88	0.93	0.88	0.11	0.07	0.92	0.81
MobileNet-V2	0.87	0.86	0.90	0.81	0.14	0.12	0.89	0.72
Inception-V3	0.96	0.77	0.87	0.93	0.23	0.04	0.91	0.77
ResNet-50	0.93	0.85	0.91	0.88	0.14	0.07	0.92	0.78
ResNet-101	**0.98**	0.60	0.79	0.95	0.40	**0.02**	0.87	0.66
DenseNet-121	0.93	0.74	0.85	0.87	0.25	0.07	0.88	0.69
DenseNet-201	0.93	0.91	0.94	0.88	0.09	0.07	0.93	0.83
EfficientNet-B4	0.90	**0.97**	**0.98**	0.87	**0.03**	0.09	**0.94**	**0.86**
GENNet Model	**0.98**	0.80	0.89	**0.96**	0.20	**0.02**	0.93	0.82

5. Discussions

This section will discuss the overall performance of each model along with its individual training and inference time required compared to other models and individual parameters of each model. Table 6 presents the numerical values of this comparison.

Table 6 Comparison of total training and testing time required by all the models and individual Model Parameters

Models	AlexNet	VGG-19	MobileNet-V2	Inception-V3	ResNet-50	ResNet-101	DenseNet-121	DenseNet-201	EfficientNet	GENNet
Average training time required for each epoch (seconds)	2.66	12.19	4.55	6.2	6.4	10.3	8.3	11.33	15.13	3.1

Overall Training time required (seconds)	266	1220	455	620	640	1030	830	1130	1513	310
Inference time required for complete test data (seconds)	3.6	13.2	4.1	8.3	7.2	11.2	7.4	9.3	7.2	3.6
Parameters (Million)	62.3	138	2.2	24	26	43	7.2	18.6	19M	16.8

- AlexNet model achieved the best inference time and sensitivity compared to the other models, but it has a low specificity and precision scores.
- EfficientNet-B4[63], DenseNet-201 and GENNet model has achieved an optimal F1-score followed by VGG-19 and ResNet-50 architectures. Also, EfficientNet-B4[63] achieved the highest testing accuracy of 93% and best Matthews Correlation Coefficient[42] scores however, efficientNet-B4 requires the highest training time.
- DenseNet-201 also proved to be one of the best models achieving second best specificity and second lowest false positive rate. The total test accuracy of model is 91%, however it requires highest inference time and relatively higher training time as compared to other models thus making it computationally expensive model.
- The bigger architectures like ResNet, DenseNet and EfficientNet have good sensitivity and less False Negative Rate (FNR), however, the inference time required by these architectures is relatively high compared to other models.
- Although the proposed model GENNet, has a high false-positive rate, but as a trade-off, it achieved the optimal test accuracy of 91% along with good sensitivity, F1 score, negative predictive value and lowest false negative rate when compared to other low or nearly equivalent parameter models. In addition to this, the model requires the least inference time like AlexNet.

- By analyzing the low specificity value of all the models except EfficientNet-B4 compared to the sensitivity metric as shown in Table 7, it can conclude that low specificity can be overcome by using a significant amount of thermal training data to better generalize the capabilities of DNN.
- Moreover, currently the main focus is on gender classification for in-cabin driver monitoring systems using thermal facial features. The current technique can be expanded to face recognition and obtaining other biometrics information in random outdoor environmental conditions. For instance, in law enforcement application this system can be made more effective by capturing data through CCTV recordings and doing more specific trainings and thus performing multi frame classification such that, with hat or without hat, with mask or without mask, then subsequently identifying person's gender. This can be achieved by training advanced deep learning algorithms such as human body instance segmentation and recognition.

6. Conclusions and Future Work

In the proposed study we have proposed a new CNN architecture GENNet for autonomous gender classification using thermal images. Initially, all the models which includes pre-trained models as well as newly proposed GENNet models are trained on large scale human facial structures which eventually helps us to fine-tune the model on smaller thermal facial data more robustly. In order to achieve optimal training accuracy and less error rate all the networks are trained using two different state of art optimizers which includes Stochastic Gradient Decent (SGD) and Adam optimizers and picked the best results in case of each model. The trained models are cross validated using two new thermal datasets which includes public as well as locally gathered dataset. EfficientNet-B4 model achieved the highest training accuracy of 93% followed by the DenseNet-

201 and the proposed network which has achieved an overall testing accuracy of 92 and 91% however, GENNet architecture is good for a compute-constrained thermal gender classification use-case as it performs significantly better than other low-parameter models.

For future work, we can work on the grouping of different datasets and fusions of features that can eventually push towards the horizon for the advancement of deep learning. In the same way, we can use techniques to generate new data from the existing data such as smart augmentation techniques, GANs and last but not least generating synthetic data that can aid us in increasing the accuracy levels and reducing the overfitting of a target network. Moreover, multi-scale convolutional neural networks can be designed for performing more than one human biometrics task such as face recognition, age estimation, and emotion recognition using thermal data. For example, face recognition using thermal imaging can be performed using blood perfusion data by extracting blood vessels patterns which is unique in all human beings. Similarly, emotion recognition can be performed by learning specific thermal patterns in human face while recording different emotions.

Appendices A

Table 1 Layer wise Architecture of GENNet. Output shape is shown in brackets along with kernel size, no of stride, padding and number of network parameters

Block-1	Block-2	Block-3	Block-4
Conv 2D - 1 [16, 16, 250, 250] Kernal size = 3 Stride = 1 Padding = 1 No of param = 448	Conv 2D - 5 [16, 32, 125, 125] Kernal size = 3 Stride = 1 Padding = 1 No of param = 4,640	Conv 2D - 9 [32, 64, 62, 62] Kernal size = 3 Stride = 1 Padding = 1 No of param = 18,496	FC-1/ Linear - 13 [65536, 256] No of param = 16,777,472
ReLU - 2 [16, 16, 250, 250]	ReLU - 6 [16, 32, 125, 125]	ReLU - 10 [32, 64, 62, 62]	ReLU - 14
MaxPool 2D - 3 [16, 16, 125, 125] Kernal size = 2 Stride = 2	MaxPool 2D - 7 [16, 32, 62, 62] Kernal size = 2 Stride = 2	MaxPool 2D - 11 [32, 64, 32, 32] Kernal size = 2 Stride = 2 Padding = 1	Dropout (0.5) - 15
Dropout (0.5) - 4 [16, 16, 125, 125]	Dropout (0.5) - 8 [16, 32, 62, 62]	Dropout (0.3) - 12 [32, 64, 32, 32]	FC-2/ Linear [256, 1] Total No of param = 16,801,570

Appendices B

During the experimental work, when training the GENNet model from scratch using only thermal dataset we were unable to achieve precise training and validation accuracy with greater loss values which eventually results in low testing accuracy. The experiments were carried using different optimizers which includes adaptive learning rate optimization Adam[51] as well as Stochastic

Gradient Descent (SGD)[50] but the same results were observed. The experimental results are demonstrated in below Figure.

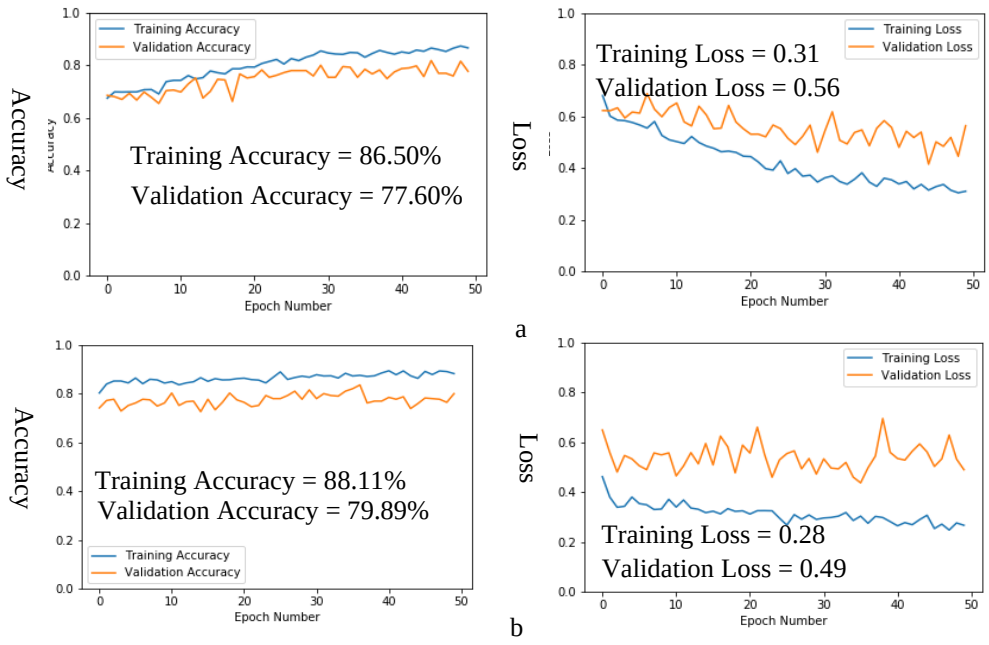


Figure. Training GENNet accuracies and loss graph using only thermal data: (a) training and validation accuracy and loss graph using Adam optimizer, and (b) training and validation accuracy and loss using SGD optimizer

Disclosures

Authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose.

Acknowledgments

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation program and National funding from France, Germany, Ireland (Enterprise Ireland International Research Fund), and Italy. The authors would like to acknowledge Joesph Lamley for providing his support on how to regularize and generalize the new DNN architecture with smaller datasets, Xperi Ireland team and Quentin Noir from Lynred France for giving their

feedback. Moreover, authors would like to acknowledge tufts university the contributors of the tufts dataset and carl dataset for providing the image resources to carry out this research work.

Author Biographies and Photographs

First Author is a PhD Researcher at the National University of Ireland Galway (NUIG). He received his BE in Electronics Engineering from IQRA University in 2012 and MS degrees in Electrical Control Engineering from the National University of Sciences and Technology (NUST) in 2017. His current research interests include machine learning, computer vision, medical image analysis, and sensor fusion. He has won the prestigious H2020 European Union (EU) scholarship and currently working as one of the consortium partners in HELIAUS EU project.

Hossein Javidnia is a Research Fellow at ADAPT Centre, Trinity College Dublin. A committee member at the National Standards Authority of Ireland working on the development of a national AI strategy in Ireland. He received his Ph.D. in Electronic Engineering from the National University of Ireland Galway focused on depth perception and 3D reconstruction. He is currently researching offline augmented reality and generative models.

Peter Corcoran is the Editor-in-Chief of the IEEE Consumer Electronics Magazine and a Professor with a Personal Chair at the College of Engineering & Informatics at NUI Galway. In addition to his academic career, he is also an Occasional Entrepreneur, Industry Consultant, and Compulsive Inventor. His research interests include biometrics, cryptography, computational imaging, and consumer electronics.

Ethical Conduct

For the proposed study informed consent was obtained from all the five subjects to publish their thermal facial data.

References

1. Raahul, A., et al. "Voice based gender classification using machine learning." *Materials Science and Engineering Conference Series. Vol. 263. No. 4*. (2017). [doi:10.1088/1757-899X/263/4/042083].
2. Smith, Philip, and Cuixian Chen. "Transfer Learning with Deep CNNs for Gender Recognition and Age Estimation." *2018 IEEE International Conference on Big Data (Big Data). IEEE*, (2018). [doi:10.1109/BigData.2018.8621891].
3. Khan, Asifullah, et al. "A survey of the recent architectures of deep convolutional neural networks." *Artificial Intelligence Review: 1-62*, (2019).
4. Makinen, Erno, and Roope Raisamo. "Evaluation of gender classification methods with automatically detected and aligned faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence 30.3*, (2008): 541-547.[doi:10.1109/TPAMI.2007.70800].
5. Reid, D. A., et al. "Soft biometrics for surveillance: an overview." Handbook of statistics. Vol. 31. Elsevier, (2013). 327-352.[doi: 10.1016/B978-0-444-53859-8.00013-8].
6. G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image and Vision Computing, vol. 32, no. 10, pp. 761 – 770, 2014*, best of Automatic Face and Gesture Recognition (2013). .[doi:10.1016/j.imavis.2014.04.011].
7. A. J. O'toole, T. Vetter, N. F. Troje, H. H. Bulthoff, et al. Sex ¨ classification is better with three-dimensional head structure than with image intensity information. Perception, 26:75–84, (1997). [doi:10.1068/p260075].
8. B. Moghaddam and M.-H. Yang. Learning gender with support faces. Trans. Pattern Anal. Mach. Intell., 24(5):707– 711, (2002). [doi:10.1109/34.1000244].
9. S. Baluja and H. A. Rowley. Boosting sex identification performance. Int. J. Comput. Vision, 71(1):111–119, (2007).[doi:10.1007/s11263-006-8910-9].
10. M. Toews and T. Arbel. "Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion". *Trans. Pattern Anal. Mach. Intell., 31(9):1567–1581*, (2009). [doi: 10.1109/TPAMI.2008.233].

11. Ullah, Ihsan, et al. "Gender recognition from face images with local wld descriptor." 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, (2012).
12. J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. Wld: A robust local image descriptor. Trans. Pattern Anal. Mach. Intell., 32(9):1705–1720, (2010). [doi:10.1109/TPAMI.2009.155].
13. P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. "The feret database and evaluation procedure for face-recognition algorithms". *Image and vision computing, 16(5):295–306,* (1998). [doi:10.1016/S0262-8856(97)00070-X].
14. C. Perez, J. Tapia, P. Estevez, and C. Held. "Gender classification from face images using mutual information and feature fusion". *International Journal of Optomechatronics, 6(1):92– 119,* (2012). [doi:10.1080/15599612.2012.663463].
15. Arun K.S., Rarath K.S. A, "Machine Learning Approach for Fingerprint Based Gender Identification"; *Proceedings of IEEE Conference on Recent Advances in Intelligent Computational Systems; Trivandrum, India. 22–24 September,* (2011); pp. 163–16. [doi:10.1109/RAICS.2011.6069294].
16. Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." arXiv preprint arXiv:1605.07678 (2016).
17. Resnet: He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2016). [doi:10.1109/CVPR.2016.90]
18. He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.
19. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition. IEEE,* (2009). [doi:10.1109/CVPR.2009.5206848].

20. Panetta, Karen, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Rao et al, The Tufts Face Database, <http://tdface.ece.Tufts.edu/> (Last accessed on 29 October 2019).
21. Panetta, Karen, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Rao et al. "A comprehensive database for benchmarking imaging systems." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2018). [doi: 10.1109/TPAMI.2018.2884458].
22. Shreyas Kamath K. M., Rahul Rajendran, Qianwen Wan, Karen Panetta, and Sos S. Agaian "TERNet: A deep learning approach for thermal face emotion recognition", *Proc. SPIE 10993, Mobile Multimedia/Image Processing, Security, and Applications 2019*, 1099309 (13 May 2019). [doi: 10.1117/12.2518708].
23. V. Espinosa-Duró, M. Faundez-Zanuy and J. Mekyska, "A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums*", Cognitive Computation, vol. 5, no. 1, pp. 119-135*, (2013). [doi:10.1007/s12559-012-9163-2].
24. V. Espinosa-Duró, M. Faundez-Zanuy, J. Mekyska and E. Monte-Moreno, "A Criterion for Analysis of Different Sensor Combinations with an Application to Face Biometrics", *Cognitive Computation, vol. 2, no. 3, pp. 135-141*, (2010).[doi:10.1007/s12559-010-9060-5]
25. Mivia Lab University of Salerno. Gender-FERET dataset, Weblink: http://mivia.unisa.it/database/gender-feret.zip, Last accessed on 30th June 2020
26. Pytorch deep learning framework, Weblink: https://pytorch.org/, Last accessed on 14th October 2019
27. M. Stojanovi et.al., "Understanding sensitivity, specificity, and predictive values", Vojnosanit Pregl, vol. 71, no11, pp. 1062–1065, (2014).[doi:10.2298/VSP1411062S].
28. Satya Mallick, Image Classification using Transfer Learning in Pytorch, <https://www.learnopencv.com/image-classification-using-transfer-learning-in-pytorch/> (Last accessed on 10th January 2020).
29. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017). [doi:10.1109/CVPR.2017.243]

30. Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018). [doi:10.1109/CVPR.2018.00474]
31. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016). [doi:10.1109/CVPR.2016.308]
32. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
33. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, (2012). [doi:10.1145/3065386]
34. Yi, Dong, et al. "Learning face representation from scratch." arXiv preprint arXiv:1411.7923 (2014).
35. Lynred France, <https://www.lynred.com/> (Last accessed on 27th January 2020).
36. Heliaus European Union Project, <https://www.heliaus.eu/> (Last accessed on 20th January 2020).
37. Camera optics measurements, <https://www.edmundoptics.eu/knowledge-center/application-notes/imaging/understanding-focal-length-and-field-of-view/> (Last accessed on 15th February 2020).
38. Zabłocki, Michał, et al. "Intelligent video surveillance systems for public spaces–a survey." Journal of Theoretical and Applied Computer Science 8.4 (2014): 13-27.[doi:10.1.1.958.799]
39. Malik, Hasmat, et al. "Applications of Artificial Intelligence Techniques in Engineering." SIGMA 1 (2018). [doi:10.1007/978-981-13-1819-1].
40. FLIR, FLIR Vuo Pro Thermal Camera, <https://www.flir.com/products/vue-pro/> (Last accessed on 14th October 2019).
41. Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10). (2010).

42. Matthews, Brian W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2 (1975): 442-451. [doi: 10.1016/0005-2795(75)90109-9].
43. Nguyen, Dat Tien, and Kang Ryoung Park. "Body-based gender recognition using images from visible and thermal cameras." Sensors 16.2 (2016): 156. [doi:10.3390/s16020156].
44. Abouelenien, Mohamed, et al. "Multimodal gender detection." Proceedings of the 19th ACM International Conference on Multimodal Interaction. (2017).[doi:10.1145/3136755.3136770].
45. Manyala, Anirudh, et al. "CNN-based gender classification in near-infrared periocular images." *Pattern Analysis and Applications 22.4*, (2019): 1493-1504. [doi:10.1007/s10044-018-0722-3].
46. Lewis, Debra A., Eliezer Kamon, and James L. Hodgson. "Physiological differences between genders implications for sports conditioning." *Sports medicine* 3.5: 357-369, (1986).
47. Elmir, Youssef, Zakaria Elberrichi, and Reda Adjoudj. "Support vector machine based fingerprint identification." *ctci 2012 conference*. (2012).
48. Xiao, Ling, et al. "Combining HWEBING and HOG-MLBP features for pedestrian detection." The Journal of Engineering *2018.16, 1421-1426*. (2018). [doi: 10.1049/joe.2018.8308].
49. Abdelrahman, Yomna, et al. "Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3: 1-20 (2017).
50. Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." *Proceedings of COMPSTAT'2010*. Physica-Verlag HD, 177-186 (2010).
51. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
52. Arjovsky, Martin, and Léon Bottou. "Towards principled methods for training generative adversarial networks." *arXiv preprint arXiv:1701.04862* (2017).

53. Chen, Cunjian, and Arun Ross. "Evaluation of gender classification methods on thermal and near-infrared face images." *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011. [doi: 10.1109/IJCB.2011.6117544].

54. Dwivedi, N., & Singh, D. K. “Review of Deep Learning Techniques for Gender Classification in Images”. In *Harmony Search and Nature Inspired Optimization Algorithms* (pp. 1089-1099) Springer, Singapore, (2019). [doi: 10.1007/978-981-13-0761-4_102].

55. Ozbulak, Gokhan, Yusuf Aytar, and Hazim Kemal Ekenel. "How transferable are CNN-based features for age and gender classification?." *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016. [doi: 10.1109/BIOSIG.2016.7736925]

56. Baek, Na Rae, et al. "Multimodal Camera-Based Gender Recognition Using Human-Body Image With Two-Step Reconstruction Network." *IEEE Access 7*,): 104025-104044, (2019). [doi: 10.1109/ACCESS.2019.2932146]

57. Tempelhahn, A., et al. "Shutter-less calibration of uncooled infrared cameras." *Journal of Sensors and Sensor Systems* 5.1, 9. (2016). [doi: 10.5194/jsss-5-9-2016].

58. Eran Eidinger, Roee Enbar, and Tal Hassner, Age and Gender Estimation of Unfiltered Faces, Transactions on Information Forensics and Security (IEEE-TIFS), *special issue on Facial Biometrics in the Wild*, Volume 9, Issue 12, pages 2170 - 2179, (Dec. 2014).

59. Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015).

60. Karjalainen, Sami. "Thermal comfort and gender: a literature review." *Indoor air 22.2*: 96-109. (2012). [doi: 10.1111/j.1600-0668.2011.00747.x.].

61. Christensen, J., Michael Væth, and Ann Wenzel. "Thermographic imaging of facial skin—gender differences and temperature changes over time in healthy subjects." *Dentomaxillofacial Radiology* 41.8, 662-667. (2012).

62. Bronzino, Joseph D., and Donald R. Peterson, eds. *Biomedical signals, imaging, and informatics*. CRC Press, 2014.

63. Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *arXiv preprint arXiv:1905.11946*, 2019.
64. Abdelwhab, Abdelgader, and Serestina Viriri. "A survey on soft biometrics for human identification." Machine Learning and Biometrics: 37, (2018).

List of Figures