Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system

Muhammad Ali Farooq[a,*] **Hossein Javidnia**[a,b] **and Peter Corcoran**[a]
[a]National University of Ireland Galway, College of Engineering and Informatics, Galway, Ireland
[b]ADAPT Centre, Trinity College, Dublin, Ireland

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 a gender classification system using visible spectrum images of the human face. However, there are many factors affecting the performance of these systems including illumination conditions, shadow, occlusions, and time of day. Our 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-the-art deep neural networks are trained to perform the classification task. For our 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 provided 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-the-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 pretrained networks. [DOI: 10.1117/1.JEI.29.6.063004]

Keywords: deep convolution neural networks; thermal imaging; gender classification; long-wave infrared; transfer learning.

Paper 200318 received May 1, 2020; accepted for publication Oct. 23, 2020; published online Nov. 18, 2020.

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[1] 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 the personalization of various features (e.g., user interfaces and presentation of data to the driver). It can also be used to better predict driver cognitive response,[2] driver behavior, and intent, and finally knowledge of gender can be useful for safety systems such as airbag deployment that 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

*Address all correspondence to Muhammad Ali Farooq, m.farooq3@nuigalway.ie

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 and voices[3] in order to effectively interact with them thus gender information serves as a source of soft biometrics.[4] 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.[5] Human thermography helps us to better understand that central and peripheral thermoreceptors are distributed all over the body including on the 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.[6,7,8]

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°C.[8] The heat emission from the human body is related to cutaneous vascular activity, yielding enhanced heat output on vasodilation, and reduced heat amount on vasoconstriction.[9] 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.[9] Men were found to have higher temperatures compared to women overall; 25 anatomic areas were measured on the face including upper lips, lower lips, chin, orbit, and the cheek. According to another study, the basal metabolic rate of a healthy 30-year-old male with a 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; on the other hand the basal metabolic rate of healthy 30-year-old female with the height of 5 ft, 3 in the weight of 54 kg, and who has surface area of 1.4 W/m^2 dissipates about 41 W/m^2 of heat. In addition, women's skin is expected to be cooler since less heat is lost per unit of body surface area.[9] However, thermal patterns whether in the case of male or female also depend on many other factors such as age, human body intrinsic and extrinsic characteristics, outdoor environmental conditions, and technical factors such as camera calibration, and the field of view (FoV). Moreover, it also depends on factors such as drinking, smoking, various diseases, and using medications.

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 tasks if such datasets are available.

In this study, the Tufts thermal faces[10–12] and Carl thermal faces datasets[13,6] are used to train and test a selection of state-of-the-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 pretrained 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 pretrained 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 pretrained models along with GENNet architecture. The evaluation results are presented in Sec. 4.

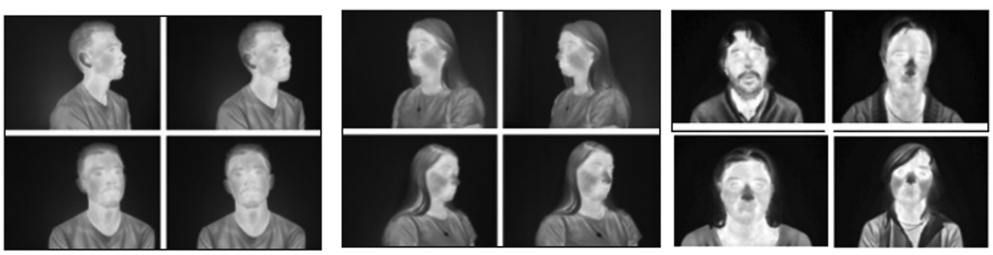


Fig. 1 Sample images from Tufts and Carl thermal face database: (a) male subject with four different face poses from the Tufts dataset; (b) female subject with four different face poses from the Tufts dataset; and (c) male and female subjects (frontal face pose) from Carl database.

2 Background/Related Work

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

2.1 *Gender Classification Using Conventional Machine Learning Methods*

Makinen and Raisamo[14] and Reid et al.[15] provided a detailed survey of the gender classifications method in their studies. One of the early techniques for gender recognition reported in Ref. 16 utilized a neural system trained on a small arrangement of close frontal face pictures. In Ref. 17, 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 Ref. 18 where the authors evaluated the performance of SVM with an overall error rate of 3.4% when compared with other traditional classifiers including linear, fisher linear discriminant, nearest neighbor, and radial basis functions. Instead of using SVM,[19] Baluja and Rowley[20] referred to AdaBoost for gender classification tasks using a set of low-resolution gray-scale images. Perspective invariant age and gender recognition was performed by Ref. 21 using arbitrary viewpoints. Recently, Ullah et al.[22] utilized the Webers local surface descriptor[23] for the gender recognition system, showing near-perfect execution on the facial recognition technology (FERET) benchmark.[24] In Ref. 25, shape, texture, and color features were extracted from frontal faces, thus obtaining robust outcomes on the FERET benchmark. In an attempt by Arun and Rarath,[26] unique mark pictures are used, and the input images are represented by a feature vector consisting of ridge thickness to valley thickness ratio and 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 and Ross[27] claimed 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 including linear binary patterns, principle component analysis, and pixels from low-resolution facial images. For gender recognition, they have used SVM, LDA, Adaboost, random forest, Gaussian mixture model, and multi-layer perceptron classifiers. Their experimental results conclude that SVM for histogram-based gender classification results in much better performance on NIR and thermal spectra. Nguyen and Park[28] proposed a gender classification system using joint visible and thermal spectrum data of the human body. The classification accuracies in Ref. 28 are measured by employing different feature extractors including HoG and MLBP.[29] Their experimental results demonstrated an improvement in classification accuracy using the joint data from visible and thermal image spectrums. Similarly, in another study reported in Ref. 30, 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.[31] Canziani et al.[32] listed many pretrained models that can be used for various practical applications in their study. They analyzed the overall performance of these pretrained models by computing the accuracy levels and the inference time needed for each model. Dwivedi and Singh[33] provided a comprehensive review of deep learning methodologies for robust gender classification using the GENDER-FERET[34] 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 (FC) 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. Ozbulak et al.[35]

have investigated two different deep learning strategies including fine-tuning and SVM classification using CNN features. They were applied on different networks including their proposed task-specific GilNet model and pretrained domain-specific VGG[36] and Generic AlexNet[37]-like CNN model for building robust age and gender classification system using the Adience[38] 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, Manyala et al.[39] investigated the overall performance of two CNN-based methods for gender classification using near-infrared (NIR) images. In the first method, a pretrained VGG-Face[40] 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, Baek et al.[41] 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 work is to evaluate the potential of thermal image facial data as a means of gender classification. The thermal image data are 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 pretrained networks.

Initially, all the pretrained networks are first trained on the Casia Face dataset[42] since Tufts thermal training dataset[10–12] does not contain enough 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[10–12] is used for transfer learning. This dataset consists of 113 different subjects and comprises images from six different image modalities that include visible, NIR, 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,[43] which was mounted at a fixed distance and height.

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

3.1 *Initial Training and Transfer Learning of Pretrained Networks*

This research takes advantage of the pretrained 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 pretrained 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 pretrained model using larger datasets like ImageNet.[44] Further, the classifier is trained to learn the higher-level features in the proposed thermal dataset images.

A typical CNN system comprises certain layers including convolution layers, pooling layers, dense layers, and FC layers. There are various pretrained networks available that can be efficiently used for different types of visual recognition, object detection, and segmentation tasks. For the proposed study, the following pretrained neural networks are utilized: ResNet-50,[45] ResNet-101,[45] Inception-V3,[46] MobileNet-V2,[47] VGG-19,[36] AlexNet,[37] DenseNet-121,[48] DenseNet-20,[48] and EfficientNet-B4[49] networks. These models are chosen as they are commonly

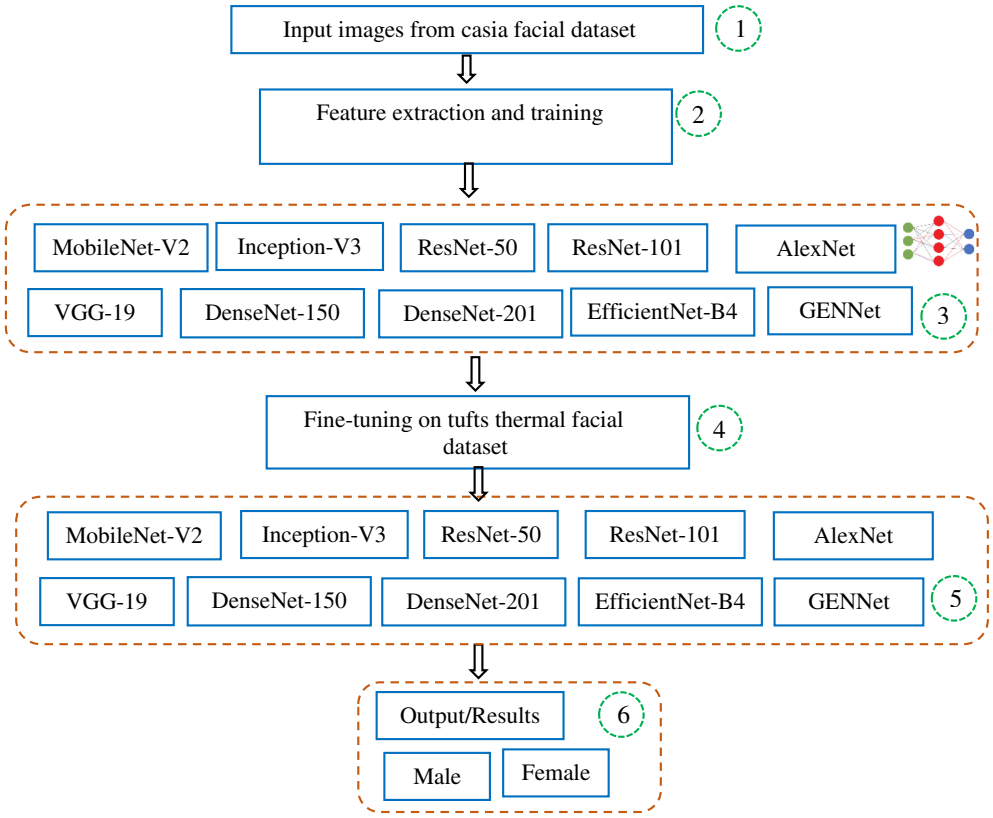


Fig. 2 Workflow diagram for autonomous gender classification system using thermal images.

trained using the ImageNet[44] dataset, each model has a different architectural style, they provide a good trade-off between accuracy and inference time,[50] and in addition, 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[45] 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 the ResNet family include ResNet-101[45] and ResNet-152.[45] Resnet-50 network was initially trained on ImageNet,[44] 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.[46] The Inception-v3 model was initially trained on Imagenet[44] as well. These pretrained 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 efficient deep learning architecture proposed by Sandler et al.[47] 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 a 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[48] architecture also referred to as dense convolutional neural network is a state-of-the-art variable-depth deep convolutional neural architecture. It was designed to improve the architecture of ResNet.[45] The principle design feature of this architecture is channel-wise concatenation, with every convolution layer that has access to the activations of every layer preceding it. DenseNet family has different variants including DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264.

VGGNet[36] was developed by the Visual Geometry Group from the University of Oxford. Like ResNet[45] and Inception-V3,[46] this network was also originally trained on ImageNet.[44] The network was designed with the significant improvement compared to AlexNet architecture,[37] which was more focused on smaller window sizes and strides in the first convolutional

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 kernels
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	26 M	ImageNet: 3.6	152	Residual learning, identity mapping-based skip connection
ResNet-101	43 M			
DenseNet-121	7.2 M	CIFAR-10+: 3.46	190	Cross-layer information flow
DenseNet-201	18.6 M			
EfficientNet-B4	19 M	ImageNet: 2.9		Compound coefficient scaling method, 8.4 × smaller and 6.1 × faster than other convnets

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[37] (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 including VGG-11, VGG-16, and VGG-19.

EfficientNet[49] was recently published and designed using a compound scaling method. As the name suggests the network proved to be a competent and optimum network by achieving state-of-the-art results on the ImageNet dataset. Table 1[51] provides a more comprehensive comparison of these architectures highlighting their attributes, number of parameters, the overall error rate on benchmark datasets, and their respective depth.

As discussed in the previous section, all the pretrained networks are initially trained on the Casia Face database[42] since the Tufts thermal training dataset[10–12] does not contain a sufficient number of images. Casia facial dataset[42] 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 were 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[42] training data including random resizing of 0.8, random rotation of 15 deg, 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 pretrained models. In the case of ResNet-50 and ResNet-101 networks, the last FC layer is connected to a linear layer having 256 outputs. It is further fed into the rectified linear unit (ReLU)[52] and dropout layers with the dropout ratio of 0.4 followed by a final FC layer, which has binary output corresponding to the two 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 pretrained 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.[53]

The models were trained using the PyTorch framework.[54] Binary cross-entropy is used as the loss function during training along with a stochastic gradient descent (SGD)[55] optimizer. The final training data include male and female thermal images as shown in Fig. 4.

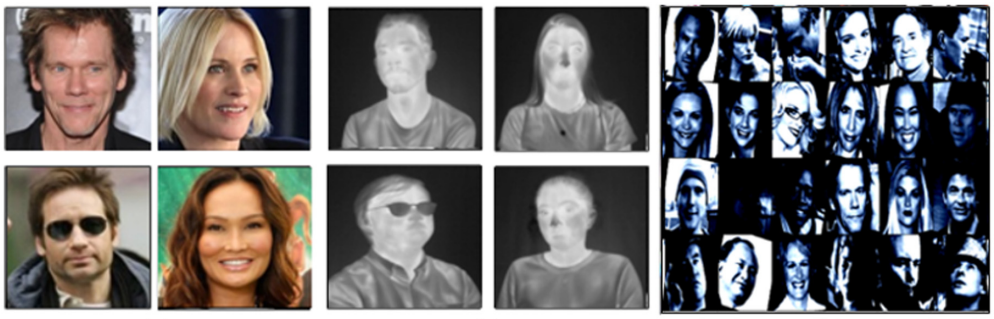


Fig. 3 Facial samples from two different datasets: (a) male and female data samples from Casia[42] database; (b) male and female samples from Tufts thermal images;[10–12] and (c) PyTorch data transformations on Casia dataset.

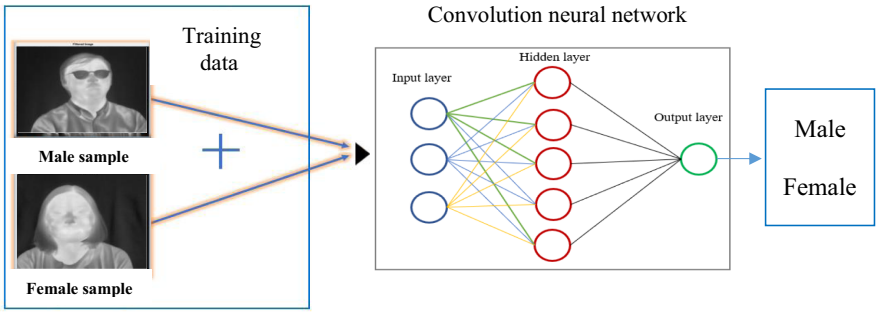


Fig. 4 Training data comprising of male and female samples for network training.

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

During the fine-tuning phase, the SGD[55] and the Adam[56] optimizers are used to compare their respective performance. This is discussed in Sec. 4. As compared to gradient descent (GD) where the full training set is used to update the weights in each iteration, in minibatch SGD,[55] the dataset is split into randomly samples minibatches, and the weights are updated in separate iterations for each minibatch (not element-wise unless minibatch size is 1). Moreover, minibatch SGD[55] is computationally less expensive and minimizes losses faster than GD as it cycles through the full training data, just in the form of chunks as opposed to all at once. The Adam[56] optimizer is an adaptive learning rate optimizer and is considered one of the best optimizers for training convolution neural networks. As compared to minibatch SGD, Adam optimizer also uses the SGD algorithm. However, it implements an adaptive learning rate and

Table 2 Training data transformation

Transformation type	Data variation
Resized cropping	Size = 256, scale = (0.8, 1.0)
Rotation	15 deg
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]

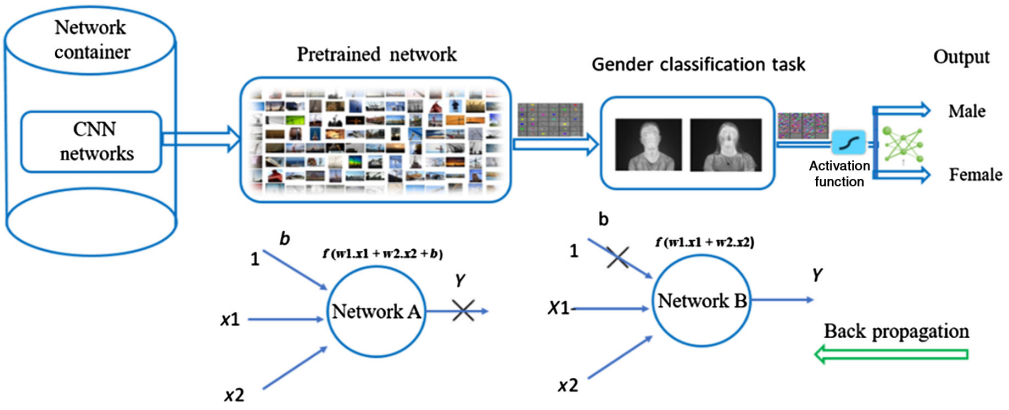


Fig. 5 CNN training structure: network A indicates pretrained networks with initial weights and network B indicates transfer learning process with new weights for thermal gender classification.

Table 3 Pretrained networks hyperparameters

Network hyperparameters	
Batch size	32
Epochs	100
Learning rate	0.001
Momentum	0.9
Loss function	Cross-entropy
Optimizer	SGD and Adam

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

3.2 *New CNN Model GENNet*

To analyze the validity of the existing thermal images, a novel CNN network is designed that is referred to as GENNet and its performance is compared against the pretrained 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

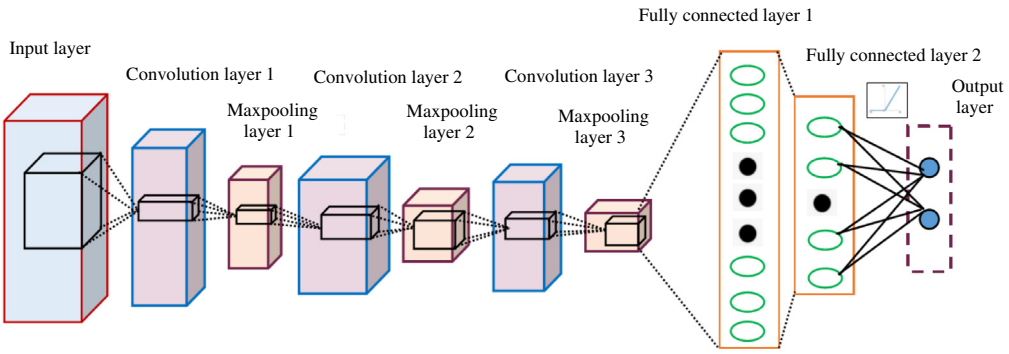


Fig. 6 Structural representation of GENNet CNN model for thermal images-based gender classification.

contain sequential layers in the form of 2D convolutions each followed by the ReLU[52] activation function, max-pooling, and dropout layers. The fourth block consists of two FC layers. The first FC layer is followed by the ReLU activation function[52] and dropout layer, whereas the second and last FC layer of the overall network converts the corresponding number of features to the number of outputs. The layer-wise detail of the GENNet model is provided in Appendix A (Table 7).

Like all other pretrained networks, GENNet is initially trained on the Casia facial database[42] and later fine-tuned on Tufts thermal dataset.[10–12] The same division of thermal training data is used along with the same hyperparameters as it was utilized for other pretrained 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[54] deep learning platform is used to fine-tune and train all the pretrained models as well as the proposed GENNet model. These experiments are performed on a machine equipped with NVIDIA TITAN X graphical processing unit with 12 GB of dedicated graphic memory.

4.1 *Training and Validation Results of CNN Architectures by Unfreezing the Layers*

In this part of the experimental study, all the networks are retrained by unfreezing all the original network layers to improve the feature learning process on thermal data. As described and shown in ablation study Sec. 6, transfer learning while freezing the network layers and using both SGD and ADAM optimizer we cannot achieve optimal training and validation accuracy in the case of most of the models. The experimental results using freezed network layer are depicted in Fig. 14. 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. 7. 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. The final learning rate of the network was 0.0003. Using a smaller learning rate makes a model converge more efficiently but at the expense of the speed, whereas using a higher learning rate can lead to model divergence. Thus to overcome this issue, the learning rate needs to be adjusted automatically. One cycle LR works by increasing and then decreasing the learning rate according to a fixed schedule 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 7 shows the training and validation accuracy chart of all the retrained networks along with the newly proposed GENNet architecture.

It can be observed that most of the models performed significantly well by getting training accuracy above 96% and validation accuracy greater than 90%. The inception-V3 achieved the highest training accuracy with the lowest training loss of 0.008. The Efficientnet-B4 network achieved the highest validation accuracy of 96.98% with a validation loss of 0.11. The newly proposed GENNet model for task-related thermal gender classification achieves the overall training and validation accuracy of 97.86% and 92.26% with loss of 0.08 and 0.15, respectively. The trained models are further used for cross-validating their performance on the new test data as discussed and shown in the subsections.

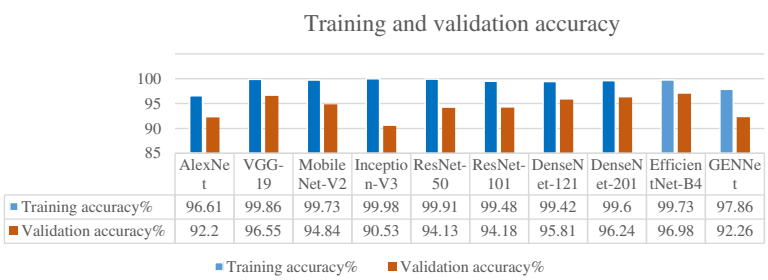


Fig. 7 Accuracy charts of all the networks by unfreezing the network layers.

4.2 *Local Thermal Data Acquisition*

To further validate the effectiveness of all the pretrained models and provide an additional mode of comparison with the newly proposed CNN GENNet model, a live thermal facial dataset was gathered using a new prototype thermal camera. The data are acquired in an indoor lab environment using a camera-based on a prototype uncooled microbolometer thermal camera array that embeds a Lynred[57] long-wave infrared (LWIR) sensor developed under the Heliaus EU project.[1] Figure 8 displays the prototype thermal camera model being used for the proposed research work to gather this live dataset, whereas Table 4 provides the technical specifications of the camera.

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 as follows:[58]

$$F-\text{number}=\frac{\text{focal length}(f)}{\text{diameter}(D)}, \tag{1}$$

$$\text{diameter}(D)=\frac{\text{focal length}(f)}{F\,\text{number}}=\frac{7.5}{1.2}=6.25\approx 6\ \text{mm}, \tag{2}$$

$$\text{height of sensor}(h)=\text{horizontal pixels}*\text{pixel spitch}=640*17=10.88\ \text{mm}, \tag{3}$$

$$\text{width of sensor}(w)=\text{vetricle pixels}*\text{pixel spitch}=480*17\ \mu\text{m}=8.16\ \text{mm}, \tag{4}$$

$$\text{AFOV}=2*\tan^{-1}\frac{h}{2f}=2*\tan^{-1}\frac{10.88\ \text{mm}}{2*7.5\ \text{mm}}=71.9\approx 72\ \ \text{deg}, \tag{5}$$

$$\text{working distance(WD)}=\frac{\text{focal length}(f)*\text{HFOV}}{\text{height of sensor}(h)}=\frac{7.5*890}{10.88}\approx 60\ \text{cm}. \tag{6}$$

The data are collected by mounting a camera on a tripod at a fixed distance of 60 to 65 cm. The height of the camera is adjusted manually to align the subject's face centrally in the FoV. Shutterless[59] camera calibration at 30 FPS is used to acquire the data. The data acquisition setup

Fig. 8 Prototype thermal VGA camera model for acquiring local facial data.

Table 4 Technical specifications

Prototype thermal camera specifications	
Quality and type	VGA and LWIR
Resolution	640 × 480 pixels
Focal length (f)	7.5 mm
F-number	1.2
Pixel pitch	17 μm
HFOV	90 deg, 890 mm

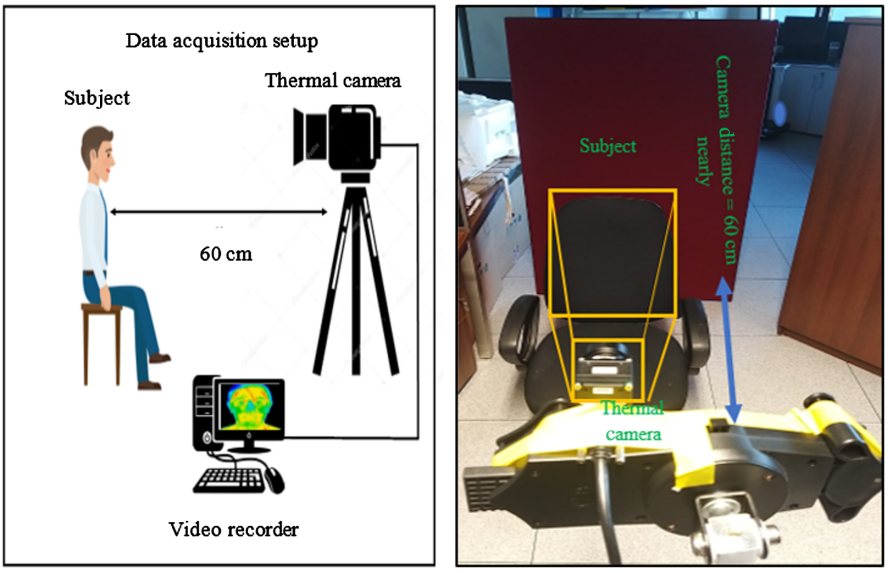


Fig. 9 Indoor lab environment data acquisition setup.

is shown in Fig. 9. A total of five subjects consensually agreed to take part in this study. The data were gathered by recording videos stream of each subject covering different facial poses and then generating image sequences from the acquired videos.

Figure 10 illustrates a few samples of the captured data including both male and female subjects.

4.3 *Testing Results of State-of-the-Art CNN*

All the trained models are tested on the combination of the two different datasets including Carl[13,6] and the locally gathered indoor thermal dataset. This is done 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 the cross-validation experiment. The test data contain a total of ninety samples. The overall performance of all the networks on test data is measured using the accuracy metric as shown in the following equation:[60]

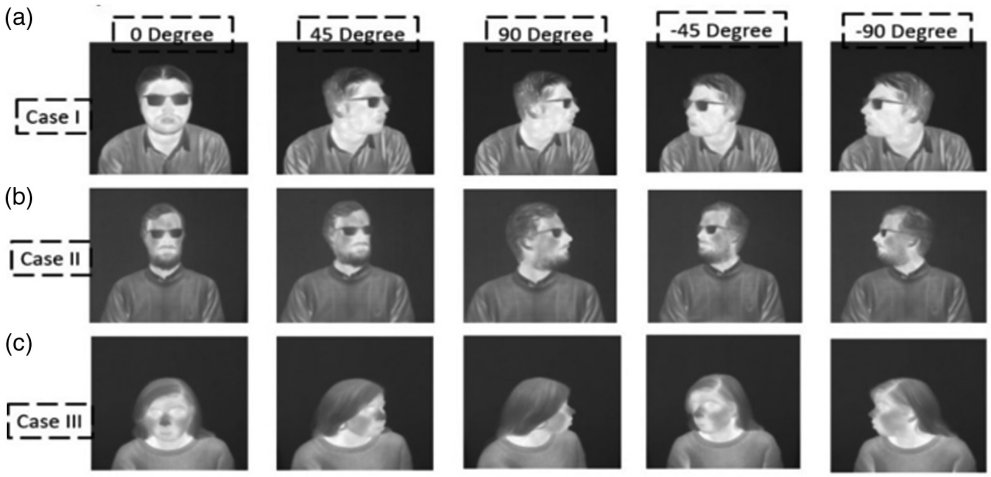
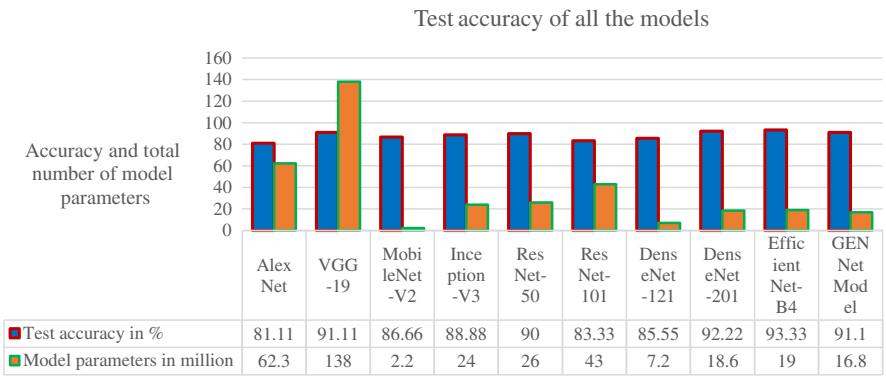


Fig. 10 Test cases of three different subjects acquired in the lab environment with varying face pose: (a), (b) the varying facial angles of male subjects and (c) the different facial angles of a female subject.



	Alex Net	VGG -19	Mobi leNet -V2	Ince ption -V3	Res Net-50	Res Net-101	Dens eNet -121	Dens eNet -201	Effic ient Net-B4	GEN Net Mod el
Test accuracy in %	81.11	91.11	86.66	88.88	90	83.33	85.55	92.22	93.33	91.1
Model parameters in million	62.3	138	2.2	24	26	43	7.2	18.6	19	16.8

Fig. 11 Test accuracy and model parameters chart of all the CNN architectures.

$$\text{accuracy}(\text{ACC}) = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}} \times 100, \tag{7}$$

where tp, fp, fn, and tn refer to true positive, false positive, false negative, and true negative, respectively. ACC in Eq. (7) means overall testing accuracy.

Figure 11 illustrates the calculated test accuracy along with total number of parameters chart of all the models. A confusion matrix for five of the best models is presented in Fig. 12 to better elaborate on the performance of each model on different genders.

By analyzing Fig. 11, we can observe that GENNet model performed significantly well among other low-parameter models by achieving total test accuracy of 91%, equal to the test accuracy of the VGG-19 model. However, VGG-19 has 138 million parameters, which is the highest number of parameters among all other models.

Figure 13 shows a number of failed predictions by the studied state-of-the-art models. The results display the model name along with the predicted output class.

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. 12 Confusion matrix depicting the performance of (a) VGG-19; (b) ResNet-50; (c) DenseNet-201; (d) EfficientNet-B4; and (e) GENNet models.

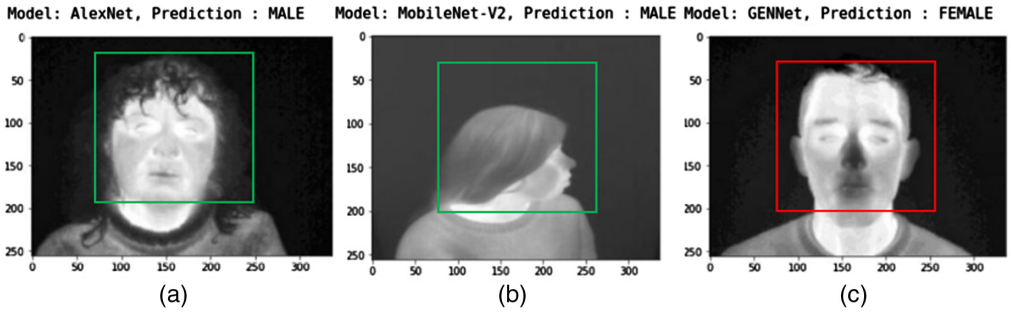


Fig. 13 Individual false prediction test case results: (a) AlexNet model: female gender misclassified as male gender; (b) MobileNet: female gender misclassified as male gender; and (c) GENNet: male gender misclassified as female gender.

In order to understand how effective, the models are for the 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 the 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 FPR is the proportion of negative cases incorrectly identified as positive cases in the data, whereas 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. 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 including 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 bold and italics.

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.

- 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,[49] DenseNet-201, and GENNet model has achieved an optimal $F1$-score followed by VGG-19 and ResNet-50 architectures. Also EfficientNet-B4[49] achieved the highest testing accuracy of 93% and best MCC[61] scores, however, EfficientNet-B4 requires the highest training time.
- DenseNet-201 also proved to be one of the best models achieving the second best specificity and second lowest FPR. The total test accuracy of the model is 91%, however, it requires the highest inference time and relatively higher training time as compared to other models thus making it a computationally expensive model.
- The bigger architectures such as ResNet, DenseNet, and EfficientNet have good sensitivity and less 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

Table 5 Different quantitative metrics. The best value per metric is highlighted in bold, and the worst value per metric is highlighted in italics.

Quantitative metrics comparison of all the models								
Models	Sensitivity	Specificity	Precision	Negative predictive value	FPR	FNR	*F*1-score	MCC
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

Table 6 Comparison of total training and testing time required by all the models and individual model parameters

Models	Alex Net	VGG-19	Mobile Net-V2	Inception-V3	Res Net-50	Res Net-101	Dense Net-121	Dense Net-201	Efficient Net	GEN Net
Average training time required for each epoch (s)	2.66	12.19	4.55	6.2	6.4	10.3	8.3	11.33	15.13	3.1
Overall training time required (s)	266	1220	455	620	640	1030	830	1130	1513	310
Inference time required for complete test data (s)	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	19 M	16.8

predictive value, and lowest FNR 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 be concluded that low 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 applications[62] this system can be made more effective by capturing data through CCTV recordings. The recorded data can be used for training and thus performing multi-frame detection and classification tasks such as hat and mask detection, and then subsequently classifying the person's gender. This can be achieved by training advanced deep learning algorithms[63,64] such as human body instance segmentation and recognition.

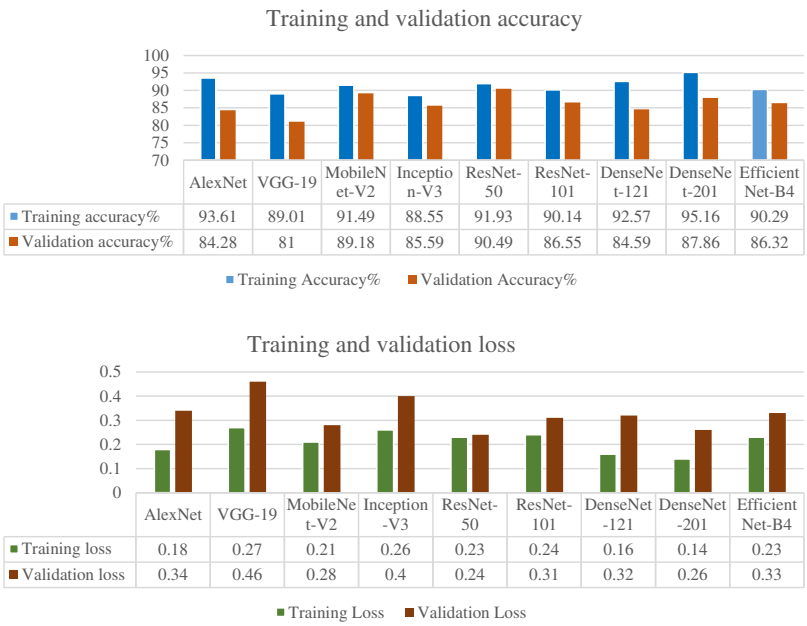


Fig. 14 Accuracy and loss charts of all the networks trained using freezed layer configuration.

6 Ablation Study

This section shows an ablation study by 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 14 presents the overall performance of all the pretrained architectures initially trained on Casia dataset[42] and fine-tuned on thermal facial images from Tufts dataset.[10–12] The networks were trained using both SGD and Adam optimizer, and the best training and validation results in the case of each model were selected. It is important to mention that during the training phase the data are 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 to do optimal inductive learning. Figure 14 presents the training and validation accuracy and loss chart of all the pretrained models.

Among all the models ResNet-50 architecture scores highest with the validation accuracy of 90.49% followed by MobileNet-V2 with a validation accuracy of 89.18% using the SGD optimizer. However, AlexNet, VGG, and EfficientNet architectures do not perform well as compared to other models thus getting the lower validation accuracy and higher loss values. However, it was not possible to achieve an optimal training outcome as most of the models have accuracy levels below 95% with freeze layer configuration. By analyzing the accuracy and loss charts in Fig. 14, it is clear that during the finetuning process of all the pretrained models DenseNet-201[48] and AlexNet achieves the highest training accuracies of 95.16% (using SGD optimizer) and 93.61% (using Adam optimizer) with the lowest training losses of 0.14 and 0.18, respectively. MobileNet-V2[47] architecture achieved the best validation accuracy of 89.18% with a validation loss of 0.28 (using SGD optimizer). However, it achieved a lower training accuracy of 90.32% with validation accuracy of 90.16% when the model was trained using Adam optimizer. The DenseNet-201 model scored second best with a validation accuracy of nearly 88% (using SGD optimizer). The VGG-19 architecture was unable to achieve good accuracy scores compared to the other pretrained models with overall validation accuracy of only 81% and the highest validation loss of 0.46.

7 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 including pretrained models

as well as newly proposed GENNet models are trained on a large-scale human facial structures, which eventually help 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-the-art optimizers including SGD and Adam optimizers and picked the best results in the case of each model. The trained models are cross-validated using two new thermal datasets including the public as well as the locally gathered dataset. The EfficientNet-B4 model achieved the highest training accuracy of 93% followed by the DenseNet-201, and the proposed network 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 toward 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 are unique in all human beings. Similarly, emotion recognition can be performed by learning specific thermal patterns in human faces while recording different emotions.

Appendix A

Table 7 shows the complete layer-wise architectural details of the newly proposed GENNet model for task-specific thermal gender classification.

Table 7 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]	Conv 2D-5 [16, 32, 125, 125]	Conv 2D-9 [32, 64, 62, 62]	FC-1/linear-13 [65536, 256]
Kernel size = 3	Kernel size = 3	Kernel size = 3	No of param = 16,777,472
Stride = 1	Stride = 1	Stride = 1	
Padding = 1	Padding = 1	Padding = 1	
No of param = 448	No of param = 4,640	No of param = 18,496	
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]	MaxPool 2D-7 [16, 32, 62, 62]	MaxPool 2D-11 [32, 64, 32, 32]	Dropout (0.5)-15
Kernel size = 2	Kernel size = 2	Kernel size = 2	
Stride = 2	Stride = 2	Stride = 2	
		Padding = 1	
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

Appendix 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 including adaptive learning rate optimization Adam[56] as well as SGD,[55] but the same results were observed. The experimental results are demonstrated in Fig. 15.

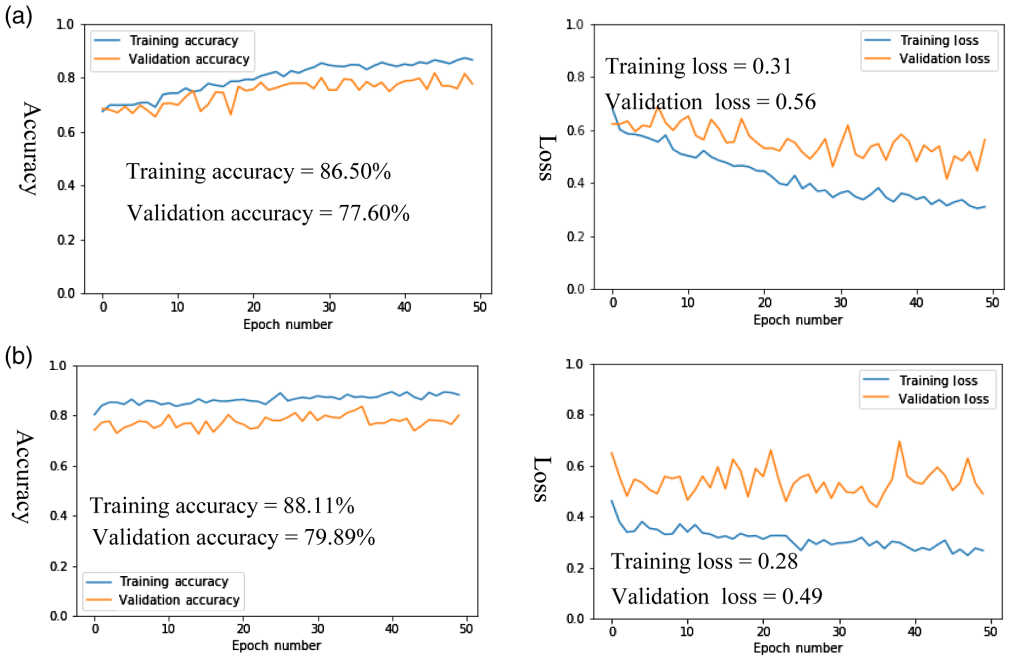


Fig. 15 Training GENNet accuracies and loss graph using only thermal data: (a) training and validation accuracy and loss graph using Adam optimizer and (b) training and validation accuracy and loss using SGD optimizer.

Acknowledgments

This thermal gender classification system using the public as well locally gathered dataset acquired using prototype thermal camera with measured accuracies of state-of-the-art models is part of the project that 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 the 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, Chris Dainty, and Quentin Noir from Lynred France for giving their feedback. Moreover, the authors would like to acknowledge Tufts University for the contributors of the Tufts dataset and Carl dataset for providing the image resources to carry out this research work. Authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose. For the proposed study informed consent was obtained from all the five subjects to publish their thermal facial data.

References

1. Heliaus European Union Project, https://www.heliaus.eu/ (accessed 20 January 2020).
2. Y. Abdelrahman et al., "Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **1**(3), 1–20 (2017).

3. A. Raahul et al., "Voice based gender classification using machine learning," *IOP Conf. Series: Mat. Sci. Eng.* **263**(4), 042083 (2017).
4. A. Abdelwhab and S. Viriri, "A survey on soft biometrics for human identification," in *Machine Learning and Biometrics*, J. Yang et al., Eds., p. 37 (2018).
5. S. Karjalainen, "Thermal comfort and gender: a literature review," *Indoor Air* **22**(2), 96–109 (2012).
6. V. Espinosa-Duró et al., "A criterion for analysis of different sensor combinations with an application to face biometrics," *Cognit. Comput.* **2**(3), 135–141 (2010).
7. D. A. Lewis, E. Kamon, and J. L. Hodgson, "Physiological differences between genders implications for sports conditioning," *Sports Med.* **3**(5), 357–369 (1986).
8. J. Christensen, M. Væth, and A. Wenzel, "Thermographic imaging of facial skin—gender differences and temperature changes over time in healthy subjects," *Dentomaxillofacial Radiol.* **41**(8), 662–667 (2012).
9. J. D. Bronzino and D. R. Peterson, *Biomedical Signals, Imaging, and Informatics*, CRC Press, Boca Raton, Florida (2014).
10. K. Panetta et al., "The tufts face database," http://tdface.ece.tufts.edu/ (accessed on 29 October 2019).
11. K. Panetta et al., "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 509–520 (2020).
12. K. M. S. Kamath et al., "TERNet: a deep learning approach for thermal face emotion recognition," *Proc. SPIE* **10993**, 1099309 (2019).
13. V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognit. Comput.* **5**(1), 119–135 (2013).
14. E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(3), 541–547 (2008).
15. D. A. Reid et al., "Soft biometrics for surveillance: an overview," in *Handbook of Statistics*, C. R. Rao and V. Govindaraju, Vol. **31**, pp. 327–352, Elsevier, North Holland (2013).
16. G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vision Comput.* **32**(10), 761–770 (2014).
17. A. J. O'Toole et al., "Sex classification is better with three-dimensional head structure than with image intensity information," *Perception* **26**, 75–84 (1997).
18. B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 707–711 (2002).
19. Y. Elmir, Z. Elberrichi, and R. Adjoudj, "Support vector machine based fingerprint identification," in *Conférence nationale sur l'informatique et les Technologies de l'Information et de la Communication*, Vol. **2012** (2012).
20. S. Baluja and H. A. Rowley, "Boosting sex identification performance," *Int. J. Comput. Vision* **71**(1), 111–119 (2007).
21. M. Toews and T. Arbel, "Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(9), 1567–1581 (2009).
22. I. Ullah et al., "Gender recognition from face images with local wld descriptor," in *19th Int. Conf. Syst., Signals and Image Process.*, IEEE (2012).
23. J. Chen et al., "WLD: a robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1705–1720 (2010).
24. P. J. Phillips et al., "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vision Comput.* **16**(5), 295–306 (1998).
25. C. Perez et al., "Gender classification from face images using mutual information and feature fusion," *Int. J. Optomechatron.* **6**(1), 92–119 (2012).
26. K. S. Arun and K. S. A. Rarath, "Machine learning approach for fingerprint based gender identification," in *Proc. IEEE Conf. Recent Adv. Intell. Comput. Syst.*, Trivandrum, India, pp. 163–16 (2011).
27. C. Chen and A. Ross, "Evaluation of gender classification methods on thermal and near-infrared face images," in *Int. Joint Conf. Biom.*, IEEE (2011).

28. D. T. Nguyen and K. R. Park, "Body-based gender recognition using images from visible and thermal cameras," *Sensors* **16**(2), 156 (2016).
29. L. Xiao et al., "Combining HWEBING and HOG-MLBP features for pedestrian detection," *J. Eng.* **2018**(16), 1421–1426 (2018).
30. M. Abouelenien et al., "Multimodal gender detection," in *Proc. 19th ACM Int. Conf. Multimodal Interaction* (2017).
31. H. Malik et al., "Applications of artificial intelligence techniques in engineering," in *SIGMA*, Vol. **698** (2018).
32. A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," arXiv:1605.07678 (2016).
33. N. Dwivedi and D. K. Singh, "Review of deep learning techniques for gender classification in images," in *Harmony Search and Nature Inspired Optimization Algorithms*, N. Yadav et al., Eds., Vol. **741**, pp. 327–352, Springer, Singapore (2019).
34. Mivia Lab University of Salerno, "Gender-FERET dataset," http://mivia.unisa.it/database/gender-feret.zip (accessed 30 June 2020).
35. G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are CNN-based features for age and gender classification?" in *Int. Conf. Biom. Special Interest Group*, IEEE (2016).
36. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. (ICLR)*, San Diego, California (2015).
37. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.* (2012).
38. E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Secur.* special issue on Facial Biometrics in the Wild **9**(12), 2170–2179 (2014).
39. A. Manyala et al., "CNN-based gender classification in near-infrared periocular images," *Pattern Anal. Appl.* **22**(4), 1493–1504 (2019).
40. O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proc. British Machine Vision Conf. (BMVC)*, pp. 1–12, BMVA Press (2015).
41. N. R. Baek et al., "Multimodal camera-based gender recognition using human-body image with two-step reconstruction network," *IEEE Access* **7**, 104025–104044 (2019).
42. D. Yi et al., "Learning face representation from scratch," arXiv:1411.7923 (2014).
43. FLIR, "FLIR Vuo Pro thermal camera," https://www.flir.com/products/vue-pro/ (accessed 14 October 2019).
44. J. Deng et al., "Imagenet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE (2009).
45. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
46. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
47. M. Sandler et al., "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2018).
48. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2017).
49. M. Tan and Q. V. Le. "Efficientnet: rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, Vol. **97**, pp. 6105–6114 (2019).
50. S. Mallick, "Image classification using transfer learning in Pytorch," https://www.learnopencv.com/image-classification-using-transfer-learning-in-pytorch/ (accessed 10 January 2020).
51. A. Khan et al., "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.* **53**, 5455–5516 (2020).
52. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.* (2010).
53. P. Smith and C. Chen, "Transfer learning with deep CNNs for gender recognition and age estimation," in *IEEE Int. Conf. Big Data*, IEEE, Seattle, Washington, pp. 2564–2571 (2018).
54. "Pytorch deep learning framework," https://pytorch.org/ (accessed 14 October 2019).

55. L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Physica-Verlag HD, pp. 177–186 (2010).
56. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
57. Lynred France, "Heliaus project coordinator and consortium partner," https://www.lynred.com/ (accessed 27 January 2020).
58. "Camera optics measurements," https://www.edmundoptics.eu/knowledge-center/application-notes/imaging/understanding-focal-length-and-field-of-view/ (accessed 15 February 2020).
59. A. Tempelhahn et al., "Shutter-less calibration of uncooled infrared cameras," *J. Sens. Sens. Syst.* **5**(1), 9 (2016).
60. M. Stojanovi et al., "Understanding sensitivity, specificity, and predictive values," *Vojnosanit Pregl* **71**(11), 1062–1065 (2014).
61. B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta* **405**(2), 442–451 (1975).
62. M. Zabłocki et al., "Intelligent video surveillance systems for public spaces—a survey," *J. Theor. Appl. Comput. Sci.* **8**(4), 13–27 (2014).
63. K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, pp. 2961–2969 (2017).
64. M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," arXiv:1701.04862 (2017).

Muhammad Ali Farooq received his BE degree in electronics engineering from IQRA University in 2012 and his MS degree in electrical control engineering from the National University of Sciences and Technology in 2017. He is a PhD researcher at the National University of Ireland Galway. His research interests include machine vision, computer vision, smart embedded systems, and sensor fusion. He has won the prestigious H2020 European Union (EU) scholarship and currently working on safe autonomous driving systems under the HELIAUS EU project.

Hossein Javidnia received his PhD in electronic engineering from the National University of Ireland Galway focused on depth perception and 3D reconstruction. He is a research fellow at ADAPT Centre, Trinity College, Dublin, Ireland, and a committee member at the National Standards Authority of Ireland working on the development of a national AI strategy in Ireland. 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 and Informatics of 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.