Computers & Security 125 (2023) 103058

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study

Muhammad Atif Butt[a], Adnan Qayyum[a], Hassan Ali[a], Ala Al-Fuqaha[b], Junaid Qadir[c,*]

[a] Information Technology University (ITU), Punjab, Lahore, Pakistan
[b] Information and Computing Technology (ITC) Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[c] Qatar University, Doha, Qatar

ARTICLE INFO

Article history:
Received 29 May 2022
Revised 16 October 2022
Accepted 8 December 2022
Available online 12 December 2022

Keywords:
Embedded machine learning
Human-Centered artificial intelligence
Adversarial machine learning
Privacy-awareness
Trustworthiness
Security
Robustness
Tiny machine learning

ABSTRACT

The use of artificial intelligence (AI) at the edge is transforming every aspect of the lives of human beings from scheduling daily activities to personalized shopping recommendations. Since the success of AI is to be measured ultimately in terms of how it benefits human beings, and that the data driving the deep learning-based edge AI algorithms are intricately and intimately tied to humans, it is important to look at these AI technologies through a human-centric lens. However, despite the significant impact of AI design on human interests, the security and trustworthiness of edge AI applications are not foolproof and ethicalneither foolproof nor ethical; Moreover, social norms are often ignored duringin the design, implementation, and deployment of edge AI systems. In this paper, we make the following two contributions: *Firstly*, we analyze the application of edge AI through a human-centric perspective. More specifically, we present a pipeline to develop human-centric embedded machine learning (HC-EML) applications leveraging a generic human-centric AI (HCAI) framework. Alongside, we also analyzediscuss the privacy, trustworthiness, robustness, and security aspects of HC-EML applications with an insider look at their challenges and possible solutions along the way. *Secondly*, to illustrate the gravity of these issues, we present a case study on the task of human facial emotion recognition (FER) based on AffectNet dataset, where we analyze the effects of widely used input quantization on the security, robustness, fairness, and trustworthiness of an EML model. We find that input quantization partially degrades the efficacy of adversarial and backdoor attacks at the cost of a slight decrease in accuracy over clean inputs. By analyzing the explanations generated by SHAP, we identify that the decision of a FER model is largely influenced by features such as eyes, alar crease, lips, and jaws. Additionally, we note that input quantization is notably biased against the dark skin faces, and hypothesize that low-contrast features of dark skin faces may be responsible for the observed trends. We conclude with precautionary remarks and guidelines for future researchers.

1. Introduction

Artificial Intelligence (AI) is rapidly transforming the living standards of people around the world. In particular, deep learning (DL) methods have demonstrated superior performance as compared to conventional machine learning (ML) based approaches in several domains including medical imaging (Litjens et al., 2017), scene understanding (Butt and Riaz, 2022; Rasib et al., 2021b), natural language processing (NLP) (Otter et al., 2020), intelligent transportation (Sumalee and Ho, 2018), and surveillance systems (Shidik et al., 2019). Despite their proliferation, these DL models are computation-intensive as they require data-driven training using large-scale data and high-performance resources such as graphical processing units (GPUs). In addition, GPUs are also required to make inferences in real-time applications. Consequently, these models are not efficiently executable on resource-constrained mobile and edge devices.

With the emergence of cloud computing (Tange et al., 2020), cloud-hosted AI services have demonstrated the potential to leverage AI-driven intelligence on edge devices. For instance, voice-controlled personal assistants such as Alexa, Siri, and Astro are capable of processing human voice commands and responding accordingly (McLean and Osei-Frimpong, 2019). These applications utilize centralized cloud-hosted NLP models to analyze user commands and natural language generation models to produce con-

* Corresponding author.
E-mail address: jqadir@qu.edu.qa (J. Qadir).

https://doi.org/10.1016/j.cose.2022.103058

tent in a textual or audible format, understandable to common users. However, such applications raise data privacy and security concerns due to the continuous transmission of users' sensitive information to cloud servers (Liao et al., 2019). Recently, embedded-ML (EML) has emerged as a promising solution to perform ML inferences on ultra-low powered edge devices (David et al., 2021). Rather than relying on cloud-hosted AI services, EML allows the shrinking of complex DL models into a more compact form that can be deployed on resource-constrained edge devices. Subsequently, EML enables greater data security, privacy, and responsiveness while also mitigating latency and communication issues (Shafique et al., 2021). In the last couple of years, a booming growth is observed in the adaptation of EML-based applications in diverse domains including, but not limited to, smart healthcare, biometric systems, industrial services, and personal assistants (Banbury et al., 2020). However, most of human patrons remain wary of EML-based applications due to longstanding privacy and security concerns. Whereas, EML-based applications must be aware of their social responsibilities such as social norms, ethical values, transparency, and safety while interacting with the human stakeholders (Holzinger, 2021; Shneiderman, 2020b).

Our Contributions: To the best of our knowledge, this study represents the first attempt towards analyzing the human-centric challenges (i.e., privacy, trustworthiness, security, and robustness) associated with the development and deployment of EML applications. It is important to mention here that the concepts of key terms covered in this study i.e., privacy, trustworthiness, security, and robustness often overlap in literature. Therefore, before moving further, we elaborate on these four terms in the context of HC-EML applications.

Shneiderman (2020b) defines privacy as an assurance that the confidentiality of the data generated from the edge devices is protected. On the other hand, according to Qayyum et al. (2020a), security is referred to the possibility of an attack or threat which can be realized on ML/DL models to achieve the desired output. In addition, Qayyum et al. (2020a) defines robustness as the survivability and resistance of ML/DL models against adversarial attacks. We discuss the security of EML models in terms of attacks attempting to get control of the human-centric applications executing on edge devices to get intended outcomes. Although adversarial robustness is generally coined under the security of DL models in current literature, to differentiate it from other security issues, we refer to it as a *robustness issue* in this paper and evaluate it as the model's resilience to standard adversarial perturbations. Lastly, referring to Rasheed et al. (2022), trustworthiness is a mapping approach that transforms abstract ideas into a human-understandable domain and explains the influential factors that contributed to producing the decision of DL models. Keeping this definition in view, we analyzed the explainable methods that can be used in developing human-centric EML applications to identify the key features contributing to a decision.

In Table 1, we present a comparison between the existing research articles which are focusing on analyzing the aforementioned key challenges. The following are the key contributions of this paper.

1. Keeping the generic human-centric AI (HCAI) framework in view, we present a novel pipeline for developing human-centric EML (HC-EML) applications.
2. We analyze human-centric aspects of EML applications along four major dimensions: (1) privacy; (2) trustworthiness; (3) security; and (4) robustness. We also note that human-centric implications are not limited to these four components. However, as the focus of this paper is on analyzing the EML applications from a human-centric perspective, therefore, we specifically focused on these four yet very important dimensions of the HCAI framework, as shown in Fig. 1. These HCAI guidelines can be leveraged in developing EML models which must (i) understand social norms, i.e., preserve privacy, and human values (Shneiderman, 2020b), (ii) be explainable and understandable to ensure a high level of human control (Holzinger, 2021), and (iii) provide a reliable, secure, and safe experience to end-users (Shneiderman, 2020b).
3. To illustrate the gravity of the aforementioned issues, we present a case study on the human facial emotion recognition (FER) task based on the AffectNet dataset, where we analyze the effects of widely used input quantization on the security, robustness, fairness, and trustworthiness of an EML model. We re-validate that input quantization partially degrades the efficacy of adversarial and backdoor attacks at the cost of a slight decrease in accuracy over clean inputs (Ali et al., 2019; Khalid et al., 2019). By analyzing the explanations generated by SHAP, we identify that the decision of a FER model is largely influenced by features such as eyes, alar crease, lips, and jaws. Additionally, we note that input quantization is notably biased against the dark skin faces, and hypothesize that low-contrast features of dark skin faces may be responsible for the observed trends.

The rest of the paper is organized as follows. Section 2 presents the related work of embedded ML, HCAI, and adversarial ML. Section 3 presents a comprehensive analysis of the challenges in HCAI inspired EML pipeline. While, Section 4 presents a taxonomy that can be employed to ensure privacy awareness, security, trustworthiness, and efficacy in HC-EML applications. In Section 5, a case study is presented to evaluate the security vulnerabilities and bias analysis of EML-based human emotion-aware face recognition models against state-of-the-art adversarial attacks and model quantization. Lastly, the learned lessons are discussed in Section 6, and concluding remarks are summarized in Section 7.

2. Related work

Over the past couple of decades, ML has grown tremendously to help the consumers, enterprises, and other organizations in improving their business processes and optimizing their decision-making in various domains including smart healthcare (Chen et al., 2021), intelligent transportation (Butt et al., 2021a; Lv et al., 2021), robotics (Rasib et al., 2021a; Vrontis et al., 2022), and surveillance systems (Jordan and Mitchell, 2015). However, conventional ML-based applications require enormous power along with heavy computing equipment like graphical processing units to maintain their effective execution.

2.1. Embedded machine learning and human-centered AI

Recently, EML has opened new opportunities to develop tiny yet efficient ML models which can be deployed on resource-constrained embedded devices (Giri et al., 2020). In general practices, ML models are trained over cloud data centers, or on computing clusters, then the resultant model is converted into the tiny model using various model compression techniques such as model pruning, quantization, low-rank factorization, and knowledge distillation (Cheng et al., 2018). After the successful conversion, the embedded model (of size in kilobytes) is deployed on resource-constrained edge devices such as ultra-low powered micro-controllers, and small-scale embedded kits to perform real-time inference tasks. To date, EML-based applications are widely deployed in commercial products. For instance, wake-word detection (Chen et al., 2014; Gruenstein et al., 2017; Zhang et al., 2017) is a well-known EML application that is deployed in Apples Siri, Amazons Alexa, and many other virtual assistants to get user input.

Table 1
Comparison of our paper with existing related papers focusing on analyzing privacy, trustworthiness, robustness, and security of HC-EML applications. (Legend: ✓ → Covered, × → Not covered, ≈ → Partially covered).

Year	Reference	Focused Domain	Generic EML Pipeline	HC-EML Framework	Challenges and Research Directions in HC-EML				Case Study	Open Issues
					Privacy	Trustworthy	Security	Efficacy		
2020	Sanchez-Iborra and Skarmeta (2020)	TinyML-Enabled Frugal Smart Objects	✓	×	×	×	×	✓	✓	×
2020	Banbury et al. (2020)	Benchmarking Tiny-ML Systems	≈	×	×	×	×	✓	×	×
2021	Dutta and Bharali (2021)	TinyML-as-a-service	✓	×	≈	×	≈	✓	×	≈
2021	Shafique et al. (2021)	Evolution of DL towards Tiny-ML	×	×	×	×	×	✓	×	×
2021	Ray (2021)	Prospects of Tiny-ML Applications	≈	×	✓	×	✓	✓	×	✓
2021	Tsoukas et al. (2021)	TinyML for Healthcare	×	×	≈	×	≈	≈	×	×
2021	Doyu et al. (2021)	TinyMLaaS Ecosystem for ML in IoT	✓	×	≈	×	≈	✓	×	✓
2022	Rajapakse et al. (2022)	Reformable Tiny-ML	≈	×	✓	×	✓	✓	×	✓
2022	Giordano et al. (2022)	Milliwatts Micro Controllers for Tiny-ML	×	×	×	×	×	✓	✓	×
2022	Our Work	Development of Private and Secure HC-EML	✓	✓	✓	✓	✓	✓	✓	✓

Fig. 1. HCAI framework: Key social and design requirements, stakeholders and possible threats.

Also, there are various other EML applications like predictive maintenance (Susto et al., 2014), object detection (Chowdhery et al., 2019), anomaly detection (Koizumi et al., 2019), and human activity recognition (Chavarriaga et al., 2013; Zhang and Sawchuk, 2012) which are being used in consumer and industrial applications.

However, the installation of such an immense amount of aforementioned EML applications in close proximity to human society while ignoring the user experience factor raises the trust and confidence related concerns in human patrons, which are among the main barriers to the proliferation of smart technology in social and industrial sectors (Haney et al., 2020). In order to address these compelling prospects, there is a need for human-centered EML devices and support services, enriched with human-like capabilities such as understanding human language, behavior (Butt et al., 2021b), and emotions to provide an efficient and safe experience (Dafoe et al., 2021; Zhu et al., 2022). Alongside, Human-centered EML (HC-EML) devices must ensure awareness regarding social responsibilities in mind, such as catering to the privacy, interpretability, explainability, and transparency issues while interacting with the human stakeholders in close proximity (Angerschmid et al., 2022; Siregar, 2021). In this way, this paradigm shift may lead to a more efficient, understandable, and safe future while mitigating the prospects such as threats of privacy, intelligent systems-driven unemployment, and fear of uninterpretable or out-of-control technology (Peeters et al., 2021).

2.2. Adversarial attacks and embedded-ML

The discovery of adversarial attacks[1] by Szegedy et al. (2013) enabled a plethora of works analyzing the adversarial vulnerabilities of DL models (Ali et al., 2021; Latif et al., 2018; Qayyum et al., 2020b; Usama et al., 2018; 2019). Numerous attacks have been proposed with different assumptions regarding an attacker's knowledge of the model. White-box attacks (Carlini and Wagner, 2017; Goodfellow et al., 2014; Madry et al., 2017) assume an

[1] Adversarial attacks are small imperceptible perturbations that are added to a given input in order to significantly change the model's behavior on the input.

attacker knowledgeable of the model architecture and weights. On the contrary, black-box attackers can only read either the prediction probabilities (Chen et al., 2017a) or the output class (Brendel et al., 2017; Khalid et al., 2020). Similarly, several defense methods have been proposed (Ali et al., 2019; Dhillon et al., 2018; Goodfellow et al., 2014; Khalid et al., 2019; Papernot et al., 2016) to robustify DL models, most of which are rendered completely ineffective against the stronger adaptive attacks (Athalye et al., 2018). To date, certified defenses (Huang et al., 2021b) and adversarial training (Madry et al., 2017) are the only effective defenses that provide reliable robustness to DL models (Lee et al., 2021).

Although many research works have focused on improving the performance, latency, and efficiency of EML models, the security and reliability of EML models have received little to no attention (Huang and Chen, 2022; Huang et al., 2021a). This is because the Tiny-ML models, in general, are extremely difficult to reverse, and hence, these models cannot be directly evaluated against a number of standard adversarial attacks (Huang and Chen, 2022). Therefore, previous works on the adversarial evaluation of EML models have mainly resorted to the black-box or transfer attack strategies (Sun et al., 2021), which do not directly consider the model architecture and weights, thus, significantly degrading the attack performance. We, on the contrary, first convert an EML model (typically.tflite extension) into a Keras model (.h5 extension) by manually reading, sorting, and processing the EML model weights, and transferring these weights to the Keras model. As the Keras model supports gradients, it can be directly evaluated against state-of-the-art white-box adversarial attacks. In this study, we use three standard adversarial attacks (formalized below) based on their frequent use in literature. For all three attacks, we use the standard implementation provided by the state-of-the-art adversarial research library, Adversarial Robustness Toolbox (ART).

Let an input $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the valid input feature space, be correctly classified by a classifier $\mathcal{F}_\theta$ as $\mathcal{F}_\theta(x) : \mathbb{R}^C$, where θ denotes the learnable parameters of $\mathcal{F}$ and C is the total number of output classes. The goal of an attack is to compute an adversarial input x^*, such that $argmax\mathcal{F}_\theta(x) \neq argmax\mathcal{F}_\theta(x^*)$ (untargeted attack Kurakin et al., 2016), and $\max|x^* - x| \leq \epsilon$, where ϵ is the maximum perturbation budget in x.

Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014): FGSM attack computes the adversarial input, x^*, in a single step as,

$$x^* = x - \epsilon \times \text{sign}\left(\frac{\partial \mathcal{F}_\theta(x)}{\partial x}\right) \tag{1}$$

Projected Gradient Descent (PGD) Madry et al. (2017): Let $\mathcal{B}_\mathcal{X}(x, \epsilon)$, be a bounded feature space in $\mathcal{X}$, such that $\forall x_b \in \mathcal{B}_\mathcal{X}(x, \epsilon)$, $\max|x_b - x| \leq \epsilon$. PGD attack formulates adversarial example, x^*, as,

$$x^* = argmax_{x_b \in \mathcal{B}_\mathcal{X}(x,\epsilon)} \mathcal{L}(\mathcal{F}_\theta(x), \mathcal{F}_\theta(x_b)) \tag{2}$$

where $\mathcal{L}$ is the cross-entropy loss. The optimization in Eq. (2) is achieved by repeating the steps in Eq. (3) for $[0, N-1]$ steps,

$$x^*_{i+1} = x^*_i + \alpha \times sign\left(\frac{\partial \mathcal{L}(\mathcal{F}_\theta(x), \mathcal{F}_\theta(x_b))}{\partial x^*_i}\right)$$
$$x^*_{i+1} = \text{clip}(x^*_{i+1}, x - \epsilon, x + \epsilon) \tag{3}$$

where $x^*_0 = x$, $x^* = x^*_N$, and α denotes the PGD-step. Note that α is the tunable hyper-parameter of the PGD attack.

Auto Attack (Croce and Hein, 2020b): The Auto attack first modifies the PGD attack to make it parameter-free (independent of α), and then combines it with two other attacks—FAB (Croce and Hein, 2020a), and Square attack (Andriushchenko et al., 2020)—to make a stronger ensemble of parameter-free attack.

2.3. Security of embedded-ML

The backdoor attack aims to create a victim model that links a certain target label with a specified backdoor trigger (Saha et al., 2020). The attacks can ensure that model produces appropriate predictions for clean input samples in order to maintain its utility. However, the backdoor is activated when the trigger is present in the input image. In this way, the backdoor induces the model to predict the input as the target label and the behavior of the infected model is manipulated based on the preferences of the attackers. In literature, two types of backdoor approaches are presented: (i) dirty-label attacks (Chen et al., 2017b; Liu et al., 2017; 2020; Tran et al., 2018) —that manipulate the training samples and assign the corresponding labels as a target, (ii) clean-label attacks (Luo et al., 2022; Zhao et al., 2020) —that do not substitute the actual labels. From the literature, it is observed that the aforementioned backdoor attacks have significantly influenced the performance of state-of-the-art learning models (Li et al., 2022; 2021; Liu et al., 2020; Zhao et al., 2020). However, to the best of our knowledge, the performance of these attacks has not been analyzed against EML models.

2.4. Explainability of embedded-ML

With the rapid growth of ML in both— academia and the industry, its influence along with the potential side effects can no longer be taken for granted. In human-centric applications, failure is not an option: Even a slight dysfunction in healthcare or related applications can lead to fatality (Tjoa and Guan, 2020). Therefore, the explainability of ML methods has become a critical issue: Can it be explained which factors are influencing the learning models to reach some decision? Angerschmid et al. (2022). If yes, then what is the correctness of these factors? Holzinger (2021) Several papers have recommended different criteria (including causality Khanal et al., 2022; Kuang et al., 2020; Schölkopf, 2022, reliability Tjoa and Guan, 2020 and usability Kenny and Keane, 2021) and frameworks (such as Captum Kokhlikyan et al., 2020 and tf-explain Sicara) to capture explainability of ML methods. However, these methods are not explored with EML models.

3. Developing HC-EML applications: Pipeline and challenges

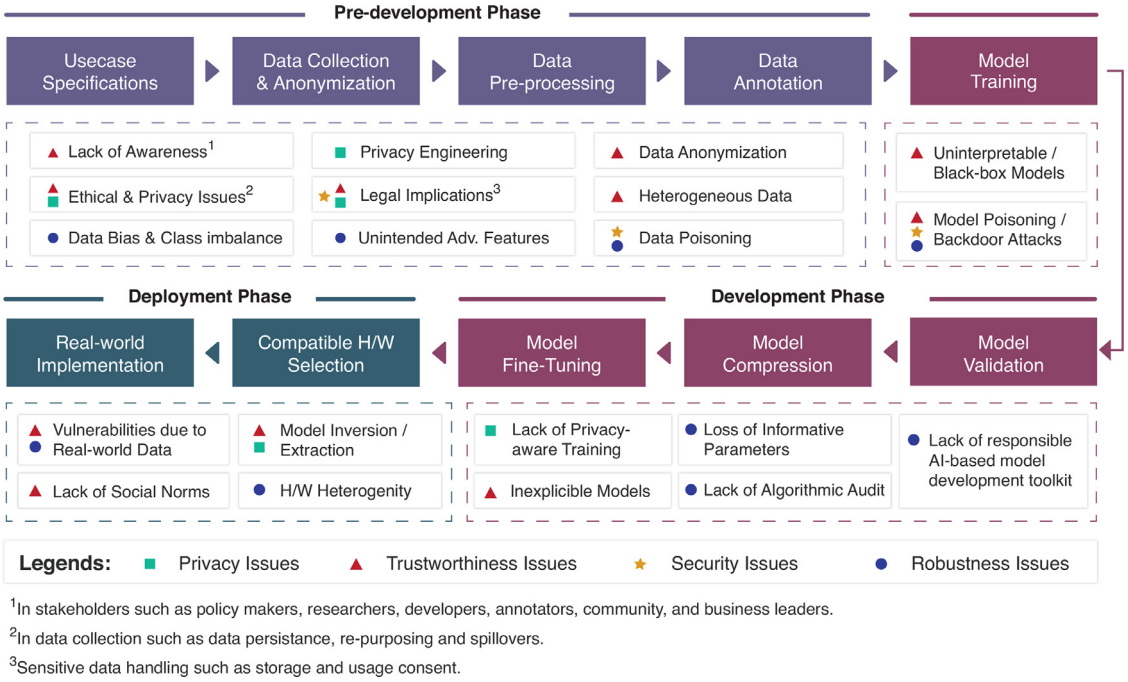
Considering the HCAI framework, we present a pipeline for developing HC-EML applications along with highlighting various challenges involved at each stage (as shown in Fig. 2). The proposed pipeline is mainly categorized into three major phases of *pre-development*, *development*, and *deployment*, which are described next.

3.1. HCAI Pipeline development phases

3.1.1. Pre-development phase

Before initiating the development of HC-EML applications, user specifications must be defined to clarify the stakeholders (such as policymakers, developers, and the community), business goals, application design, and success criteria. Then, application-specific data can be collected through various heterogeneous sources such as motion, acoustic, and biometric sensors, attached to human-centric edge devices. These sensors capture continuous users' sensitive data that should be stored over local/cloud servers after performing anonymization to ensure data privacy.

The data captured through the above-mentioned sensors often contain distortions due to operability in varying conditions that can significantly affect the performance of the underlying ML/DL model (Mitev et al., 2019). For instance, voice-controlled personal assistants can encounter a scenario, where a user is providing a



[1]In stakeholders such as policy makers, researchers, developers, annotators, community, and business leaders.

[2]In data collection such as data persistance, re-purposing and spillovers.

[3]Sensitive data handling such as storage and usage consent.

Fig. 2. The **HC-EML applications pipeline** involving various privacy, trustworthiness, security, and robustness issues at each stage.

voice command, while some music playing in the background. In this case, music gets intermixed with user input which can result in undesirable output. Such data in the training set can also influence the performance of the developed HC-EML model. In this regard, data pre-processing is performed for data cleaning and quality assurance.

In general, supervised learning methods are widely adopted for developing HC-EML applications that require labeled training data. However, data annotation may be incorrect and coarse-grained due to the perceptive subjectivity of annotators (Najafabadi et al., 2015). Therefore, this step requires careful monitoring to avoid ground truth errors.

3.1.2. Development phase

Once the data is ready, ML/DL models are developed either on the cloud or local servers. It also involves certain important decisions such as model and training framework selection, suitable hyperparameters selection and tuning, and model evaluation. Generally, the data captured from edge sensors are pushed over central servers/clouds, and a centralized model is trained over global data. However, it leads to various challenges including communication and data privacy issues (Mohanta et al., 2020). Recently, Federated Learning (FL) has emerged as a promising solution to train a model over multiple decentralized edge devices/servers without exchanging local data (Konečnỳ et al., 2016). This technique can be employed to ensure privacy and security while training the HC-EML model. After learning the DL model in FL settings, different compression techniques (such as model pruning, quantization, and knowledge distillation (Cheng et al., 2017; Stickel et al., 2009)) can be employed to make it deployable on low-powered edge devices. These methods compact the DL models through connections/neurons elimination, knowledge transfer, or mapping floating values into the finite set. However, a continuous model fine-tuning process is performed to obtain the best-performing EML model.

3.1.3. Deployment phase

The resultant compressed EML model is deployed on human-centric edge devices to perform inferences in a real-world environment. This phase involves certain compatible decisions such as suitable operating systems and hardware selection to ensure the smooth execution of the EML model.

3.2. Challenges in HC-EML pipeline

In this section, we discuss various human-centric challenges along four dimensions, i.e., privacy, trustworthiness, security, and robustness issues, which are described below.

3.2.1. Privacy issues in HC-EML pipeline

EML applications frequently collect data containing personal information such as daily routine activities, biometric data, and personal preferences (such as shopping behavior, TV shows, and preferred food choices). Such data is very precisely aligned with the user commands which can be a great aid in improving EML models. However, HC-EML devices (such as smart wearables and personal assistants) contain ultra-low powered architectures with small storage units which lead to the tradeoff between data generation and storage at the edge level.

To address these shortcomings, EML applications transmit their raw data to local or global data centers, which raise privacy concerns regarding access permissions and data protection of the consumers (Zhang et al., 2018). In addition, these practices raise legal questions regarding the consent of consumers (e.g., which type of data is being collected and shared over the servers). Manufacturers claim that they utilize collected data for improving their ML models, which also elevates privacy concerns over data collection, disclosure, and sharing to unknown parties such as data annotators.

3.2.2. Trust issues in HC-EML pipeline

In the perspective of automated systems, trust can be defined as "*an attitude that an agent helps in achieving an individual's goals in a situation characterized by uncertainty and vulnerability.*" (Shneiderman, 2020a). Data scientists are developing data-driven HC-EML applications to assist human beings in various domains that can include highly sensitive applications such as wearable drug delivery systems, symptoms detection, and prescription generation in healthcare bots; as well as low-sensitive applications

such as music and shopping recommendations, and daily routine task management through virtual assistants.

Although the discussed HC-EML applications have demonstrated better applicability in aiding human beings, they also raise questions such as: which parameters have influenced an EML model in producing a certain decision? and what is the probability of correctness and trustworthiness of those parameters? To date, ML/DL models deployed on edge devices operate as black-box, which impacts the trust and confidence that human patrons place in EML applications (Ali et al., 2022). These challenges are mainly categorized into (i) *Need for Truthful AI*—which ensures that AI-based text or speech analyses produced by EML models are not affected by misinformation or falsehood; (ii) *Need for Transparency*—which ensures the transparency of EML models; and (iii) *Need for Explainability*—which provides explainable outputs against the corresponding input (Evans et al., 2021).

3.2.3. Security issues in HC-EML pipeline

HC-EML pipeline encounters various security issues at each stage, as depicted in Fig. 2. For instance, data generated from the HC-EML applications is stored over a local/cloud server which is later used for data pre-processing. However, the collected data is often exposed to third parties, e.g., for annotation purposes that compromise data confidentiality and integrity.

Similarly, the training of the EML models is also a crucial phase as the training process is performed online on global servers either using central data or distributed data. This process is highly vulnerable to attacks like model poisoning that tampers training data to influence the model's outcomes. Also, backdoor attacks can influence the training data by embedding hidden patterns in DL models Kurakin et al. (2016). These patterns activate on inputs containing the same patterns as embedded in the trained model, which ultimately results in prediction errors. Moreover, the compressed models (the shallow version of a complex DL model) can be more prone to evasion attacks. Such attacks occur when EML models are fed with carefully crafted imperceptible adversarial examples to get the intended behavior or outcomes.

3.2.4. Robustness issues in HC-EML pipeline

Due to the heterogeneous architecture of EML devices and being operating in varying environments (such as parks, shopping sites, homes, and offices), data collected from HC-EML applications may contain unintended adversarial features (Ilyas et al., 2019)—data biases, inconsistencies, and imbalanced representations. Therefore, a DL model trained over such data produces skewed outcomes with low accuracy and analytical anomalies.

In addition to the adversarial vulnerabilities exhibited by DL models in general (Khalid et al., 2020; Szegedy et al., 2013), these models are compressed to compact them for resource-constrained edge devices using different compression techniques, which may add to the adversarial vulnerabilities of a DL model. Similarly, DL models are compressed to compact them for resource-constrained edge devices using different compression techniques. For instance, it is quite challenging to determine what to prune or eliminate without compromising the performance of the model. Therefore, loss of informative/influential parameters and inefficient fine-tuning can highly influence the robustness of resultant models. Moreover, the performance of the EML models becomes more vulnerable to unseen data due to comparatively less informative parameters in the compact model as compared to the actual model.

4. Towards developing private, secure, and trustworthy HC-EML applications

Here we present various methods to develop private, secure, trustworthy, and robust HC-EML applications to overcome the above-discussed challenges (an illustration is presented in Fig. 3) and are discussed next.

4.1. Solutions for pre-development phase

4.1.1. Educating stakeholders

HC-EML applications involve several stakeholders notably researchers, developers, data engineers, community, and business leaders. It is often reported that most of the stakeholders such as policymakers, community, and business leaders do not trust EML applications due to privacy issues. For instance, the majority of stakeholders argue that personal assistants keep track of user-sensitive data that can expose undesirable (privacy-related) insights in response to any malicious attack. Therefore, developers must consider this issue while developing HC-EML applications to make this technology secure and trustworthy. To this end, we suggest performing ML inference on edge devices through EML models and to transmit aggregated output to cloud servers for storage/backup purposes only. Also, there is a demanding need to create more awareness by organizing public educational seminars while also engaging different stakeholders.

4.1.2. Handling noise effect and uncertainty in data

EML models deployed on human-centric edge devices are trained over the data collected from multi-modal sensors, attached to billions of edge devices, in-use by the community. The majority of EML applications operate in varying environments under different atmospheric conditions. The data collected under such conditions contain multiple distortions which can influence the performance of an EML model. Moreover, due to the heterogeneity of edge devices, the attached sensors capture data with variable sensitivity and sample rates. Therefore, the data collected from the same sources can vary on different sensors, which may be challenging for EML-based sensing applications.

These shortcomings can be addressed by neutralizing the variation and noise effects from the data. There are two possibilities to cater to such data, (i) *data augmentation*—which enriches the data by adding variation while training the model to make it more efficient to noise; and (ii) *representation learning*—which focuses on extracting the most influential features while training the model to make it more robust to inconsistency/variation in data.

4.1.3. Data retention and disposal policy

Data anonymization techniques have demonstrated better applicability in preserving confidential information. However, data anonymization techniques cannot be applied to some HC-EML applications like biometric systems which need to store their data and logs (such as user identities, fingerprints, and routine activities) for reference points. However, storing such sensitive information on a server for an unspecified amount of time raises privacy and security concerns. In this regard, we suggest that a relevant data retention and disposal policy must be devised for the particular HC-EML use case (that should clearly specify the usage, minimum storage time, and data disposal) to mitigate such security risks.

4.1.4. Paid data acquisition

Existing voice-controlled personal assistants require cloud services to remotely analyze user input through different NLP techniques. Also, their developers need to collect regular operational data from these applications to analyze user experience which helps them with improving their products. Subsequently, these applications push a huge amount of data over servers. Though the companies claim that this data is required to fine-tune their models to improve user experience, however, it also compromises

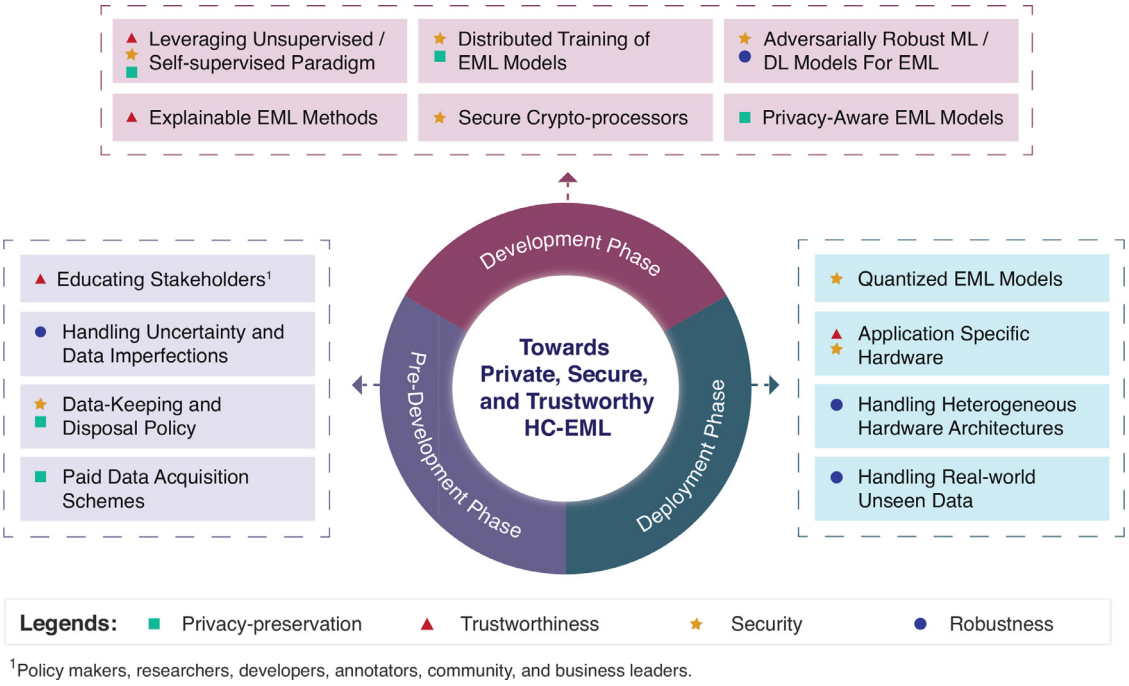


Fig. 3. Potential solutions for developing private, secure, and trustworthy HC-EML models.

users' privacy. To ensure the wider adaptability and trust of stakeholders, we recommend that the manufacturers of such products should initiate paid data acquisition schemes to ensure data collection in legal manners with the clear consent of data subjects.

4.2. Solutions for development phase

4.2.1. Moving beyond supervised learning paradigm

Personal assistants do not have a strong data protection mechanism against unauthorized access and are therefore vulnerable to malicious actors. This issue can be addressed by developing transient HC-EML-based NLP applications that capture user input, perform ML inference locally, and return the output to the user while disposing of user input data. Data security and privacy concerns can be effectively addressed using this technique, however, it will affect the traditional model improvement mechanism in which user data is transmitted over servers, which is pre-processed and labeled for re-training of the model. To handle this issue, we suggest the development of application-specific self-supervised learning techniques that do not require human input in data labeling.

4.2.2. Using differential privacy

Various differential privacy-based methods have been presented in the literature such as deferentially-private stochastic gradient descent, private-aggregation of teacher ensembles for private-ML, and exponential noise-based differential that can be used to hide users' sensitive information (Abadi et al., 2016). These methods can demonstrate better applicability in various HC-EML applications and can greatly augment trust and privacy-related issues.

4.2.3. Explainable models

EML-based edge devices can provide numerous benefits to human beings in diverse applications. However, their wide-scale adoption and acceptability in society need EML applications to be trustworthy. As a solution, the use of explainable methods is proposed in the literature that aims to explain how different model parameters and components contribute to the results being produced by DL models. In simple words, a model is said to be explainable, if its decision against some input is logically understandable (e.g., it is transparent that which factors influenced the model to reach its decision). Several approaches have been proposed to embed explainability in DL models such as generalized linear models, extreme linear models, tree SHAP, Local Interpretable Model-agnostic Explanations (LIME), and smooth gradient (Gunning et al., 2019). These methods can be leveraged to develop explainable ML/DL models for HC-EML applications.

4.2.4. Distributed training of the EML models

EML applications still require connectivity with the cloud to share data for model training and storage purpose, which ultimately raises privacy and security concerns. These limitations can be addressed by employing FL, which can perform distributed training on the edge level without sharing local data with servers. Intuitively, basic FL-based methods consist of a collaborative training framework where each participant such as an edge device can independently train a model using its local data (Li et al., 2020). These edge devices collaboratively share their model's information, i.e., model parameters without sharing actual data. Subsequently, this sharing mechanism allows EML-based applications to learn from diverse large-scale data and form a global function in a privacy-aware decentralized fashion.

FL is mainly categorized into two types; namely, data-centric FL, and model-centric FL. In model-centric FL, non-identical data distributed over remote end-user devices is used to train a central model through averaging and FederatedSGD. Whereas, in data-centric FL, end-users allow external organizations/companies to access their data to train and improve their models without sharing local data. In this approach, data is iteratively improved to achieve the best-performing model. Such methods can demonstrate better applicability in terms of dealing with sensitive data in various HC-EML applications.

4.2.5. Secure crypto-processors

Sensors attached to HC-EML devices generate an enormous amount of continuous data. Due to the resource-constrained nature of these devices, ensuring the privacy and security of the data

stream is challenging. For instance, an unauthorized person can gain physical possession of an edge device and can extract sensitive information through intrusive techniques. To address these limitations, crypto-processors can be used to secure the data on HC-EML edge devices. These are dedicated computer-on-a-chip-based microprocessors which carry out data encryption and decryption to avoid data tampering.

4.2.6. Adversarially robust HC-EML models

EML models, due to model pruning and quantization, become more prone to adversarial attacks. These attacks can potentially cause great harm to humans in various HC-EML applications (Zhao et al., 2021). In this regard, different adversarially robust techniques can be employed that have been broadly categorized into two groups, i.e., model modification, and data modification (Ross and Doshi-Velez, 2018).

Model modification refers to those methods that modify the parameters/features of trained ML/DL models to defuse the effect of adversarial perturbations. The most commonly used model modification methods include gradient regularization, defensive distillation, network verification, and model masking. Gradient regularization allows DL networks to bring a partial surge in training computational complexity to improve the network performance regardless of any prior knowledge about adversarial attacks. Defensive distillation is a knowledge transfer-based approach in which a DL model is trained on hard labels to predict the probabilities of output, produced by another model that is trained over baseline standards. Whereas, network verification refers to a verification method that verifies certain properties of the EML models over an adversarial example.

Data modification refers to defense methods in which training data or its features are modified to mitigate the effect of adversarial perturbations. These methods include adversarial re-training, feature squeezing, and input reconstruction. Adversarial re-training is a basic defense method in which ML models are trained over augmented training data comprising normal and adversarial examples. Similarly, input reconstruction is a transformation-based defense method that attempts to enhance the robustness of DL models through data pre-processing techniques such as input cleaning and denoising. Feature squeezing-based defense methods squeeze input feature space which can be exploited by an adversary in constructing adversarial input. We suggest using the aforementioned defense techniques in developing HC-EML models to mitigate the effect of adversarial attacks on real-world HCAI applications.

4.2.7. Privacy-aware EML models

Preserving the privacy of end-users in HC-EML applications is fundamentally important. Also, the underlying ML/DL models should not learn any privacy-related information during the data-driven training process. The literature suggests that these models are capable of learning sensitive information patterns even in the presence of appropriate privacy-preserving methods and different model extraction attacks can be realized. Therefore, to cater to such attacks, the development of such models is highly required that are privacy-aware by design, however, this problem is still an open question demanding significant research attention.

4.3. Solutions for deployment phase

4.3.1. Application specific hardware

The existing HC-EML applications contain general-purpose microphones, cameras, and other sensors to capture the data. The developers indirectly leverage sensitive user data generated from these applications to train and optimize their models, which raises serious privacy and data security concerns. To address this issue, we suggest using customized low-resolution task-specific sensors. For instance, the smart band should contain a $0.2 - 0.5$ MP camera that can easily capture the human face in close proximity for face recognition, however, struggle to capture surrounding objects. In this way, user privacy can be preserved while interacting with edge devices in the home, office, and other private places.

4.3.2. Quantized model

The EML models deployed on edge devices can be easily accessed due to the lack of appropriate security mechanisms. It is often reported that EML models are stolen by competitors which are then examined for understanding and replication. To cater to such issues, we suggest deploying quantized models which do not reveal the actual source weights and parameters. In this way, the unauthorized actors would not be able to understand and modify the trained models for any other use case.

4.3.3. Handling heterogeneous hardware architectures

HC-EML applications encounter heterogeneity challenges among connected devices such as smart wristbands, personal assistants, and home appliances. All of these devices are configured with different communication protocols, due to which, these devices generate different types of data. Therefore, it becomes a critical task to enable communication across such devices. To overcome this shortcoming, we suggest using data normalization techniques before providing the input to HC-EML models that translate the data onto the unit sphere to maintain standard distribution of source input data. In this regard, several data normalization methods including linear scaling, clipping, log scaling, and zero-score can be employed to translate heterogeneous data to compact it for HC-EML models.

4.3.4. Handling real-world unseen data

In real-world settings, HC-EML applications may encounter data, which is very different from the training data. This phenomenon is known as the distribution shift/drift problem. This difference is one of the major reasons behind the adversarial vulnerability of ML/DL models. In HC-EML applications, such shifts/drifts are highly expected due to the temporal and dynamic data collection and can have adverse effects on the overall performance of the system. To address this issue, we suggest the use of different domain adaptation techniques to efficiently handle unseen real-world data. Moreover, task-specific domain adaptation techniques can be developed for HC-EML applications.

5. Analyzing the security, robustness, and trustworthiness of HC-EML: Emotion-aware facial recognition case study

In this section, we present a case study to highlight the practical implications of adversarial threats on HC-EML applications. For this purpose, we have selected one of the widely used EML application nowadays, i.e., automatic face recognition, which have become one of the fundamental tools in intelligent surveillance and monitoring systems. In addition, emotion-aware face recognition systems are now widely used in key applications across different domains including educational institutions (such as schools and colleges) to comprehend students' states and help them identify and cope with stressful conditions. Generally, a camera is installed in a classroom to get a visual frame and the DL models are employed at the back-end to detect emotions from the faces in the frame. Despite their success, such human-centric systems encounter various kinds of issues that raise serious concerns about the privacy, security, and ethical values of end-users during data transmission and storage over back-end servers. EML can be leveraged to perform inference at the edge, instead of at the back-end server, to avoid the aforementioned issues.

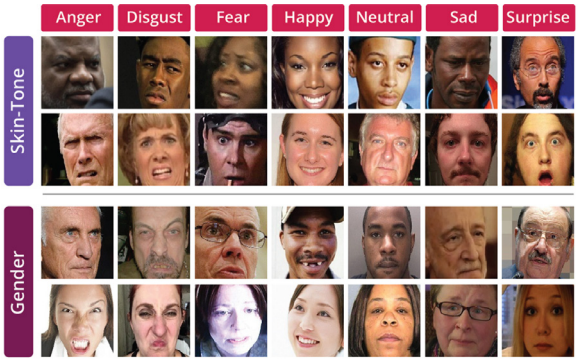


Fig. 4. Subsets of **AffectNet Dataset** for skin-tone and gender-based evaluation—First two rows demonstrate our skin-tone-based samples which are categorized into dark and fair skin-tone, respectively. Whereas, the third and forth rows show the representation of our gender-based selected subset of males and females..

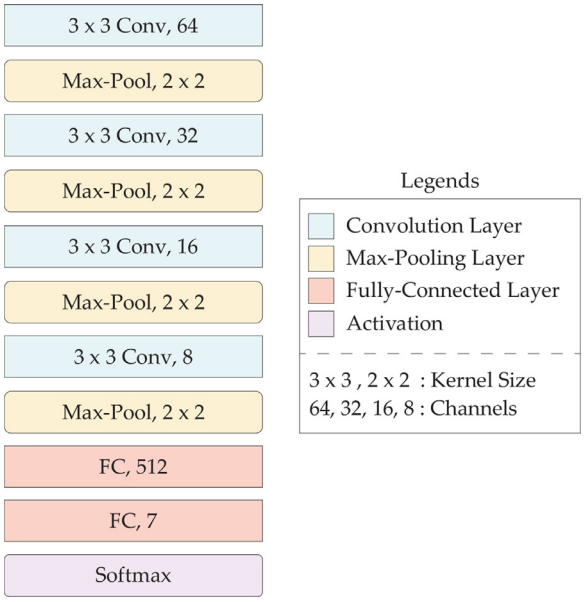


Fig. 5. Architecture of our customized HC-EML model for human facial emotion recognition tasks. The proposed model comprises four convolution layers followed by two dense layers. ReLU activation and max-pooling layer is employed after each convolution to downscale the feature map while retaining the useful information. Lastly, softmax activation function is used to get multinomial probability distribution.

Leveraging our proposed HC-EML pipeline, we have developed an EML model for human facial emotion recognition—which is essentially a human-centric application and analyzes its security vulnerabilities and fairness for human facial emotion recognition tasks. Firstly, we analyzed the performance of our HC-EML model against state-of-the-art adversarial attacks that include Projected Gradient Descent (PGD) (Madry et al., 2017), Auto attack (Croce and Hein, 2020b), and Fast Gradient-Sign Method (FGSM) (Goodfellow et al., 2014). The outcomes demonstrate that adversarial ML attacks have been quite successful in evading the outcomes of our developed HC-EML model, even though using very slight adversarial perturbations. Therefore, we also explore a potential solution to withstand such attacks (as shown in Fig. 3) in HC-EML applications for human facial recognition tasks. Secondly, we analyzed the performance of our HC-EML model on gender-specific (male/female) and skin tone-based (dark/white) facial emotions data to demonstrate the bias factor in HC-EML applications.

5.1. Experimental setup

5.1.1. Dataset

In this study, we have used one of the widely used datasets for training emotion-aware face recognition models, i.e., AffectNet (Mollahosseini et al., 2017) that contains seven distinct classes: Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. The actual AffectNet dataset contains around one million images with approximately 440k labels. We prepared three datasets to analyze the performance of the HC-EML-based facial recognition model against two aspects i.e., adversarial attacks and fairness. The first one is the actual AffectNet dataset that is used to train the HC-EML facial emotion recognition model. Whereas, the second and third datasets i.e., skin-tone and gender-specific datasets are manually constructed from the test set of the actual AffectNet dataset.

We conduct appropriate quality control to ensure the correctness of sample distribution. Firstly, we constructed the skin-tone and gender-based subsets by manually categorizing the whole test. For the skin-tone subset, we carefully split the fair and dark skin-tone samples from each class distribution. Similarly, we split male and female samples based on their appearance to form a gender subset. Some of the examples from both subsets are demonstrated in Fig. 4. After a successful data split, we designated a third party to verify the data distribution for each class and evaluate the quality of images as well; the low-quality images that contain reflection, or are unclear are discarded. Some of the main characteristics of each dataset are discussed below.

(Actual) AffectNet dataset. To train the HC-EML model for facial emotion recognition tasks, we choose the AffectNet-37k set which contains approximately 37k labeled images having seven aforementioned facial emotions classes. The dataset images are normalized to 180x180 dimensions and randomly distributed into train and set with 85 : 15 ratio (i.e., training set: 31,450, and test set: 5,550).

Skin-tone AffectNet dataset. To analyze the skin tone-based possible bias in the emotion-aware facial emotion recognition model, we formed a subset from the test set of the AffectNet dataset, comprising 5,249 images that are equally categorized into dark and white skin-based face samples of the seven emotion classes i.e., Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

Gender AffectNet dataset. Following the aforementioned practice, we manually constructed a subset of male and female samples from the test set of the AffectNet dataset, comprising 5,120 images—distributed into seven classes to perform gender-based analysis. To avoid wrong selection, we carefully cross-checked the selection against each class of both, the male and female groups.

5.1.2. Model architecture

In general HC applications, EML models are deployed on resource-constrained devices such as smart-wrist bands, smartwatches, and tiny personal assistants. These devices contain ultra-low powered tiny processors with small random-access memory, operational power, and memory space. Therefore, existing state-of-the-art large-scale ML models (He et al., 2016; Krizhevsky et al., 2017; Xie et al., 2017) cannot be deployed on such small-scale hardware architectures. Recently, Giordano et al. (2020), Mohan et al. (2021) have presented tiny CNN-based classification networks to perform face recognition tasks. However, to the best of our knowledge, there is no standard benchmark available for the training and evaluation of EML models for face recognition or facial emotion recognition tasks. In this regard, considering the memory and processing limitations of embedded microcontrollers, we developed (Giordano et al., 2020; Mohan et al., 2021) inspired a customized tiny CNN-based EML model—comprising of four convolutional layers followed by two dense layers, demonstrated in Fig. 5 to perform facial emotion recognition tasks.

After successfully training, it is required to compress the model in order to efficiently fit the model in EML devices. Till recently, four types of model compression techniques are employed to fit it into low-powered architectures (Cheng et al., 2017): (i) model pruning and quantization, (ii) low-rank factorization, (iii) transferred convolutional filters and (iv) knowledge distillation. Based on the vast adaptation and better performance (Chmiel et al., 2020), we utilized model quantization to compress our model. Leveraging our pipeline in Fig. 2, we compressed our trained model using three standard model quantization techniques, i.e., dynamic quantization, float16 quantization, and integer quantization. The ultimate aim of employing multiple quantization techniques is to analyze the intrinsic robustness of the model while considering the trade-off between performance and size of the model for facial emotion recognition tasks.

5.1.3. Adversarial evaluation

We have evaluated the security of both the uncompressed and the compressed models against three standard adversarial ML attacks that include PGD attack (Madry et al., 2017), Auto attack (Croce and Hein, 2020b), and Fast Gradient-Sign Method (FGSM) attack (Goodfellow et al., 2014). For each attack, we set the upper bound to the perturbation introduced by an adversary to each pixel to 0.01.

Note that the discovery of adversarial examples triggered a plethora of papers aiming to defend against adversarial perturbations. However, most of these defense papers exploited some preprocessing or distillation mechanism as a filter to remove the adversarial perturbations. However, as several later works would find out, these defenses generally exhibited the so-called phenomenon of gradient obfuscation that fools an attacker into computing incorrect gradients. Ultimately, the defenses appeared robust to the adversarial perturbations, not because they were actually robust, but because the adversarial perturbations were incorrectly computed. To fool such defenses, several adaptive attack algorithms were proposed (generally customized against a set of similar defenses). To the best of our knowledge, these customized adaptive adversarial attacks now constitute state-of-the-art complex adversarial attacks. However, for vanilla models, the effectiveness of standard attacks such as PGD and i-FGSM is comparable to the state-of-the-art complex attacks. Additionally, following the guidelines laid out by Athalye et al. (2018), while attacking the quantized model, we adapt these attacks such that the quantization layer is assumed to work as a linear layer to the attacker, which lets the gradients of the subsequently deployed model backpropagate.

Many recent works have exploited the robustification properties of quantized inputs to counter adversarial attacks on DNNs (Ali et al., 2019; Khalid et al., 2019). Therefore, we analyze how these models can be partially robustified if we only allow quantized inputs to be processed by the model. More specifically, we quantize an input image to either 1-bit, 2-bit, 3-bit, or 4-bit values, and refer to the aforementioned quantization configurations as Q-1, Q-2, Q-3, and Q-4 quantization, respectively. Figure 6a illustrates the effect of different quantization configurations on a random image from the dataset. Note that Q-1 achieves the strongest quantization effects followed by Q-2, Q-3, and Q-4, respectively.

5.1.4. Fairness evaluation

Recent research studies (Stickel et al., 2009) have triggered a controversial debate—the performance of DL-based facial/emotions recognition systems is greatly influenced by the gender or skin tone in input data. These reports raise imperative concerns regarding the adaptation of facial/emotion recognition-based HC-EML applications in human society. Therefore, we analyze two major aspects of our HC-EML model: (i) gender-based (male/female) analysis, and (ii) skin tone-based (dark/white) analysis. For this purpose, we evaluate our aforementioned best-performing facial emotion recognition HC-EML model against the two custom-developed datasets—Skin-tone AffectNet, and Gender AffectNet dataset. Also, it is a common practice to employ input quantization to reduce the size of data before passing it to EML models. Therefore, we evaluate our model against variable quantized inputs i.e., 1-bit, 2-bits, 3-bits, and 4-bits referred to as Q-1, Q-2, Q-3, and Q4 in Figs. 7, and 8, respectively.

5.1.5. Security evaluation

In step with the recent advancements in data poisoning and manipulation approaches, backdoor attacks have emerged as a serious security threat to DL models. This experiment particularly focuses on exploring the effectiveness of backdoor attacks against Tiny-ML models with different quantization configurations for facial emotion recognition tasks. More specifically, following recent practices (Ali et al., 2020; Gao et al., 2019), we poison the training data by randomly selecting 3% of the training data samples, injecting a 5×5 red square patch in the top-left corner of each of the selected samples, and assigning an attacker's chosen target class—in our case, *Happy*—to the perturbed image. The model trained on the poisoned dataset is referred to as the poisoned model in the future.

To evaluate the effectiveness of the backdoor attack, we first create poisoned test data by poisoning all the test data inputs by pasting a 5×5 trigger/patch in the top-left corner of the image. We then use two commonly used metrics, the Attack Success Rate (ASR)—the ratio of the poisoned test samples that are misclassified into the target class—and poisoned accuracy—the ratio of the poisoned test samples correctly classified into the original class. Both of these metrics are two of the widely used metrics in backdoor attacks literature (Ali et al., 2020; Gao et al., 2019).

5.1.6. Explainability evaluation

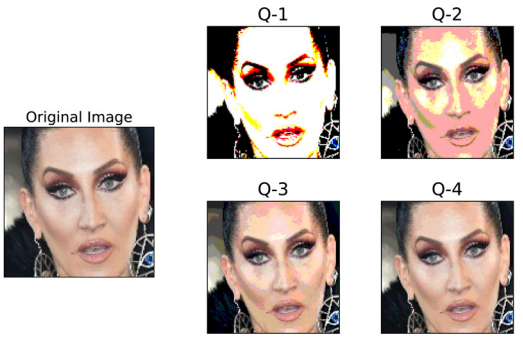
In this section, we analyze the trustworthiness of EML models in human-centric applications focusing on the specific setting of facial emotion recognition. More specifically, we generate explanations of our tiny model for the randomly selected images from the test set using a well-known black-box XAI method, SHAP (Lundberg and Lee, 2017), and qualitatively analyze the generated explanations to get important insights.

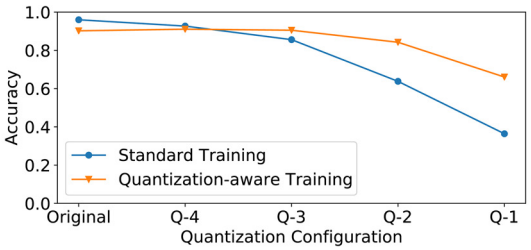
5.2. Evaluating the accuracy of trained models

Our uncompressed model achieves an accuracy of $\sim 96\%$ over the original test set under standard training settings (training over non-quantized original images[2]) as shown in Fig. 6c. When compressed, the accuracy slightly drops to 95% irrespective of the compression technique. However, this slight drop in the accuracy is also coupled with a significant reduction in the model size. More specifically, our uncompressed model occupies a space of 6.4MB in the hard drive, while the sizes of the compressed models are 0.55MB, 1.1MB, and 0.55MB for dynamic quantization, float16 quantization, and integer quantization, respectively.

From Fig. 6b, we note that as the quantization strength increases, the accuracy of the model trained under the standard settings significantly decreases. We highlight two main reasons for this behavior. Firstly, a strongly quantized image contains lesser information as compared to the original input image. Secondly, as the model is trained over non-quantized inputs, the quantized inputs are out of the distribution of the model's training set. This is also illustrated in Fig. 6a, where we note a significant difference between the original image and the Q-1 image.

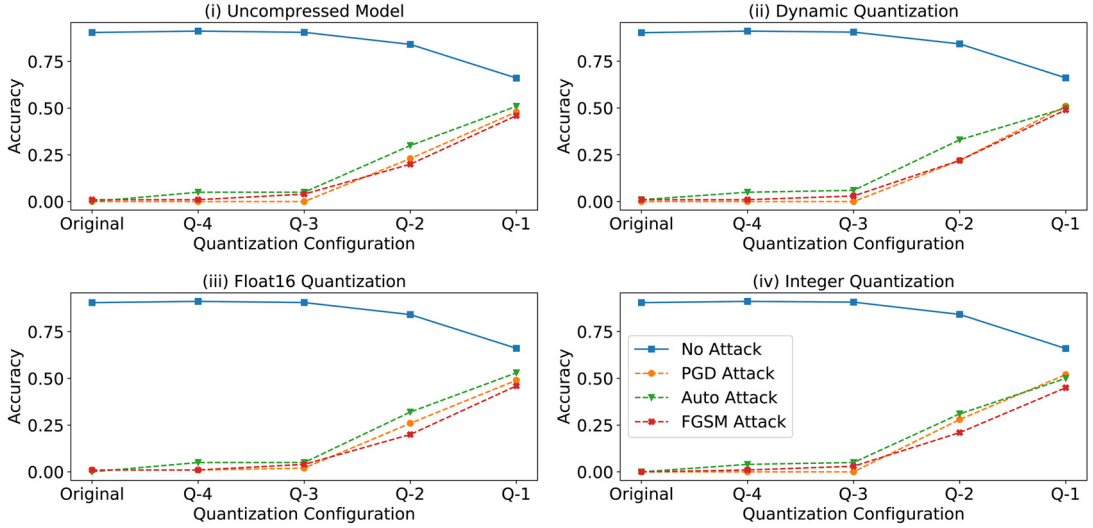
[2] In stricter terms, the standard settings generally assume Q-32 quantized images.





(a) **Comparison of different quantization configurations** with the original image (left) and the quantized images (right). *A more strongly quantized image contains considerably less information than the original image.*

(b) **Effect of quantization on model accuracy over clean inputs for the standard training and the quantization-aware training**. *Increasing the quantization strength of the input images, decreases the accuracy of a model.*



(c) **Accuracy of quantized HC-EML models over adversarially attacked images** generated by three standard attacks for the respective models for different quantization configurations assumed in our case study. *Quantizing the input images before processing by the models notably increases the accuracy of respective models over the adversarially attacked images.*

Fig. 6. Experimental results of our HC-EML Emotion-Aware Facial Recognition models under adversarial attacks.

To increase the performance of models on the quantized images, we perform the quantization-aware training by augmenting our training set with the quantized images and then fine-tuning the original model with augmented data. The results depicted in Fig. 6b show that a quantization-aware model exhibits a significantly greater performance over strongly quantized inputs—for example, the accuracy increases from 37% to 66% and from 64% to 84%, respectively over Q-1 and Q-2 quantized images. However, the cost of this superior performance over the quantized images is a slight decrease ($\sim$ 5%) in the accuracy over the original images as shown in Fig. 6b.

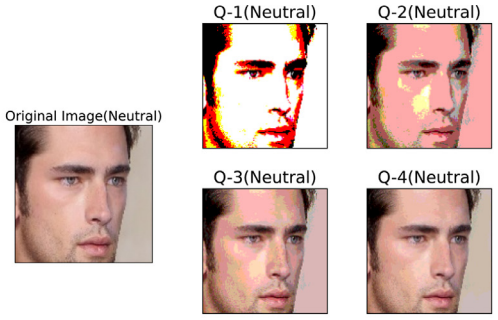
5.3. Evaluating the robustness of emotion-aware facial recognition models

Figure 6 c presents the results of the three aforementioned adversarial attacks on (i) the uncompressed and (ii)-(iv) the compressed models trained under the quantization-aware setting. We also provide the accuracy of the quantization-aware models on the unperturbed (No Attack) images for the sake of comparison. Both the uncompressed and compressed models achieve similar accuracy with 0.02% to no decrease in the overall accuracy over the unperturbed images. For the non-quantized (original) images, all three adversarial attacks reduce the accuracy of the models from $\sim$ 91% down to 0%. Therefore, we conclude that the adversarial vulnerabilities of the uncompressed DL models fully exhibit themselves in the compressed variants of DL models.

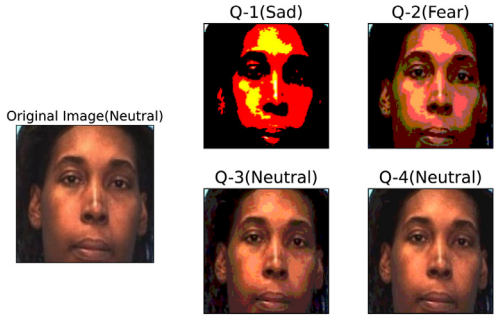
For the quantized images, we note a considerable improvement in the accuracy of quantization-aware models over the adversarially attacked images, notably for stronger quantization. For example, for the PGD attack, the accuracy of the quantization-aware models on average increases from 0% to $\sim$ 50%. We specifically attribute this to the robustification property of the quantization function. For example, note that Q-1 quantization maps all the values less than 0.5 to 0, and the values greater than 0.5 to 1. Consider an image pixel of value 0.18, which is mapped to 0 when Q-1 quantization is applied. While attacking, if an adversary were to perturb the pixel value by 0.01, the new pixel value would either become 0.19 (if the perturbation is positive) or 0.17 (if the perturbation is negative). However, Q-1 quantizing the perturbed image would still map the perturbed pixel value to 0. This robustification property of the quantization function is what notably increases the accuracy of the quantization-aware models over the adversarial examples.

5.4. Evaluating security of emotion-aware facial recognition models

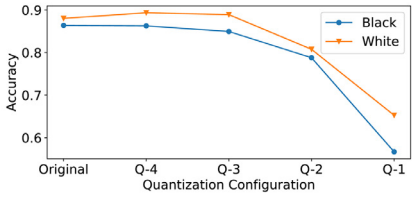
Figure 9 illustrates the effect of the backdoor attack on the poisoned model by comparing its predictions on three clean test images (randomly chosen from the test set) and the poi-



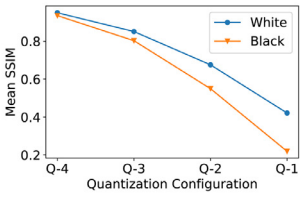
(a) Demonstrating the **effects of different quantization configurations on white skin tone samples**. *White face images, when quantized, retain greater feature information of the face than black faces, specifically notable for Q-1 configuration.*



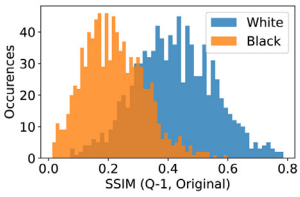
(b) Demonstrating the **effects of different quantization configurations on dark skin tone samples**. *Quantized dark face images typically retain much less emotional features of the face than white faces, specifically notable for Q-1 configuration.*



(c) Comparing the accuracy of HC-EML model on quantized white and dark skin images for different quantization configurations considered in the paper. *Degradation of accuracy due to quantization is more severe on dark skin faces, as compared to the white skin faces.*

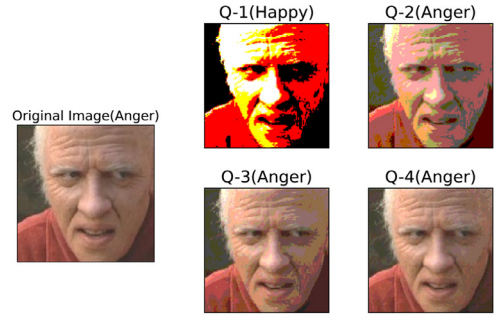


(d) Mean SSIM of original and quantized images (different configurations), for dark and white skin faces. *Dark skin face images are degraded more severely by quantization than the white skin faces.*



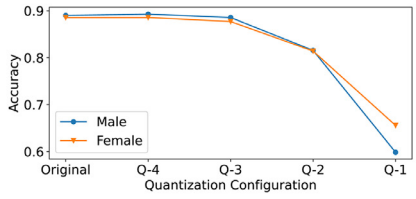
(e) Histograms of SSIM of original and quantized (Q-1) images for dark and white skin faces. *Quantized dark skin face images consistently show poor SSIM values as compared to the white skin faces.*

Fig. 7. Evaluating Skin Tone-based Bias Factor of our HC-EML Facial Emotion Recognition Model.

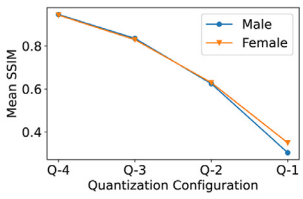


(a) Demonstrating the **effects of different quantization configurations on male face samples**. *Male face images are comparatively more severely degraded than female faces, specifically notable for Q-1 configuration.*

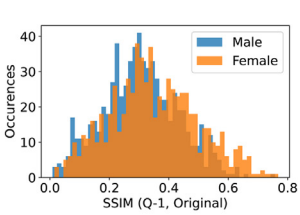
(b) Demonstrating the **effects of different quantization configurations on female face samples**. *Female face images, when quantized, retain comparatively more information than male faces, specifically notable for Q-1 configuration.*



(c) Comparing the accuracy of HC-EML model on quantized male and female face images for different quantization configurations considered in the paper. *Degradation of accuracy due to Q-1 quantization is slightly more severe on male faces, as compared to the female faces.*



(d) Mean SSIM of original and quantized images (different configurations) of male and female faces. *Male face images are degraded slightly more severely by Q-1 quantization than the female face images.*



(e) Histograms of SSIM of original and quantized (Q-1) images of male and female faces. *Quantized male face images show comparatively poor SSIM values, though slightly, than female face images.*

Fig. 8. Evaluating Gender-based Bias Factor of our HC-EML Emotion-Aware Facial Recognition Model.

soned counterparts of the same clean images characterized by 5×5 red patch trigger in the top-left corner of the image. As shown in the figure, the backdoor attack significantly influences the performance of EML models in facial emotion recognition tasks on the poisoned test inputs. Notably, our EML model produces accurate predictions on clean examples with good confidence scores. However, the same EML model malfunctions with comparatively higher confidence when inferring the poisoned inputs.

To study the effect of quantization on the efficacy of backdoor attacks, we train a quantization-aware model on the poisoned data, and report the clean accuracy and poisoned accuracy of the quantization-aware EML model and the ASR of the attack in Fig. 10. We note that a stronger quantization slightly decreases the efficacy of backdoor attacks, as illustrated by a decreasing ASR and increasing poisoned accuracy in Fig. 10. One of the key reasons for the observed decrease in the ASR is that the quantization significantly decreases the precision of different colors in an image. Ultimately,

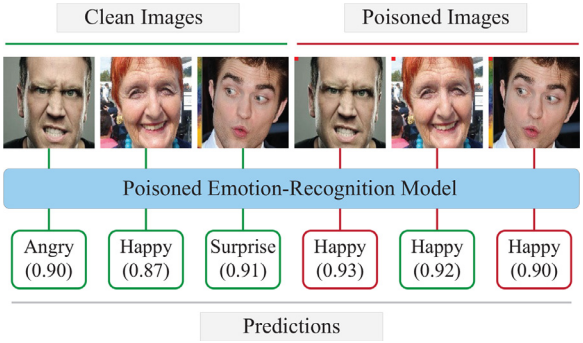


Fig. 9. An comparison of the decisions made by the model for a set of clean images and the respective poisoned images (defined by a 5×5 red square patch in the top-left corner) are input to the model. *The red square patch changes the model output to happy, indicating a successful backdoor attack.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

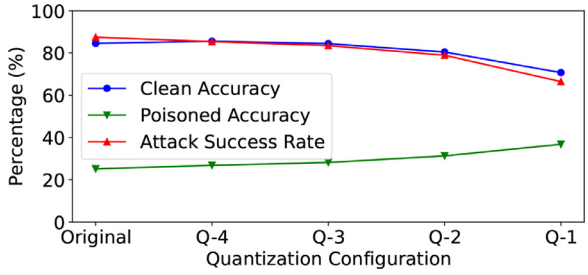


Fig. 10. Analyzing the performance of backdoor attacks on quantized face images for different quantization configurations considered in the paper. *Stronger quantization slightly decreases the efficacy of backdoor attacks.*

while training a quantization-aware model, quantization increases the probability of a red patch of random size naturally appearing in the top-left part of an image. However, because the labels of these images from the poisoned train set have been possibly left unperturbed (following Gao et al., 2019, we only poisoned 3% of the images), such images act as a natural defense against the perturbed images inserting a backdoor in the model during training. Ultimately, the efficacy of the backdoor attack is decreased, for example, from ~87% for the non-quantized images to ~60% for the Q-1 quantized images. We conjecture that this natural backdoor-resisting capability of the quantization-aware training may be used in the future to increase the effectiveness of a backdoor defense in an efficient (in terms of power and memory consumption on edge devices) manner.

5.5. Evaluating explainability of emotion-aware facial recognition models

We generated and carefully analyzed the explanations produced by SHAP for tens of images from each class. Figure 11 reports the explanations for three randomly selected images of three different classes—*Disgust, Happy* and *Angry*. From the explanations, we observe that the output of the facial emotion recognition model is mainly influenced by the facial areas such as the region around the eyes, the alar crease and nostrils, the lips, and the jaws. For example, the top-row image in Fig. 11 is classified into the *disgust* class by the model. The explanations identify the eyes, lips, and eyebrows to be the most influential features of the face when deciding the class *disgust*. Similar observations are made for the bottom-row image that is classified as *angry*. Similarly, for the image in the middle, eyes, nose, and lips are the key features in deciding the class *happy*, with the area around the lips being the most contributing factor. By analyzing several images, we have noted that the lips are generally the most contributing features of the class *happy*.

Interestingly, the explanations indicate that the background region is also slightly influencing the final outcome for all three images. This also partially explains why the facial emotion recognition model is vulnerable to adversarial attacks. One implication of this is that adding slight perturbations to tens of pixels in the background region can significantly affect the model output while also being inconspicuous.

5.6. Evaluating fairness of emotion-aware facial recognition models

5.6.1. Skin-tone analysis

The results (in Fig. 7) demonstrate that the performance of our employed model decreases with the increase in the strength of quantization. But interestingly, the model drops the accuracy significantly on the Q-1 image set of dark skin tone-based facial images, as compared to the white skin tone-based images. We note that one of the main reasons for this significant difference in Q-1 images is that dark skin tone facial images lose a huge amount of information due to the strong contrast effect of quantization, as shown in Fig. 7b. Whereas, Q-1 quantization exposes the white skin tone samples, however, retain greater emotional features of the face than black faces, specifically notable for Q-1 configuration (in Fig. 7a) the facial expressions remain visible with more available features than dark skin tone Q-1 images. Resultantly, the prediction of the HC-EML model is negatively influenced on the dark skin tone sample, while it generated an accurate prediction against white skin tone facial images. To further strengthen this assumption, we analyzed the difference between the original and the quantized images for the white and the dark skin tone-based facial images by calculating the Mean SSIM under different configurations. Resultantly, it is noted that the dark skin face images are degraded more severely by quantization than the white skin faces, as depicted in Fig. 7d. Specifically, we note that the quantized dark skin face images consistently show poor SSIM values as compared to the white skin faces on Q-1 images, as shown in Fig. 7e.

However, we would like to emphasize that several unknown factors—the lighting conditions of an image or the background—may be implicitly influencing the trends observed previously. For example, insufficient or poor lightning can possibly conceal key facial features determining the emotion of the face. To more comprehensively understand the reasons for the observed trends, we recall that a quantized image typically remains unaffected by smaller perturbations in the original image due to the natural robustness of the quantization mechanism to small changes in the input. Therefore, we conjecture that the images mostly composed of the low contrast regions are more significantly degraded when quantized. To understand this, consider two neighboring pixel values, p_1 and p_2, mapped to $q_1 = Q(p_1)$ and $q_2 = Q(p_2)$ by the quantization mechanism, $Q(\cdot)$, such that the difference (contrast) of the pixel values, $\epsilon = p_2 - p_1$, represents a relevant feature for the emotion recognition. If ϵ is sufficiently small (characterizing a lower contrast), then $\mathbb{E}[q_1] = \mathbb{E}[q_2]$, where $\mathbb{E}[q_1]$ and $\mathbb{E}[q_2]$ denote the expected value of q_1 and q_2 respectively, which removes the relevant feature, as $\mathbb{E}[q_2 - q_1] = 0$ in the quantized image. This can result in significant degradation of low contrast features, which validates our conjecture.

The aforementioned formalization highlights one of the possible reasons for the observed bias against dark skin-tone faces. Figure 12 compares the contrast images composed of region-wise contrast of randomly chosen dark and fair skin-tone faces for different filter sizes—2×2, 3×3, 4×4—where the filters define the region over which the contrast is computed. Interestingly, we note that even if the dark skin-tone face image is captured in natural light (top row of Fig. 12), it exhibits significantly lower contrast

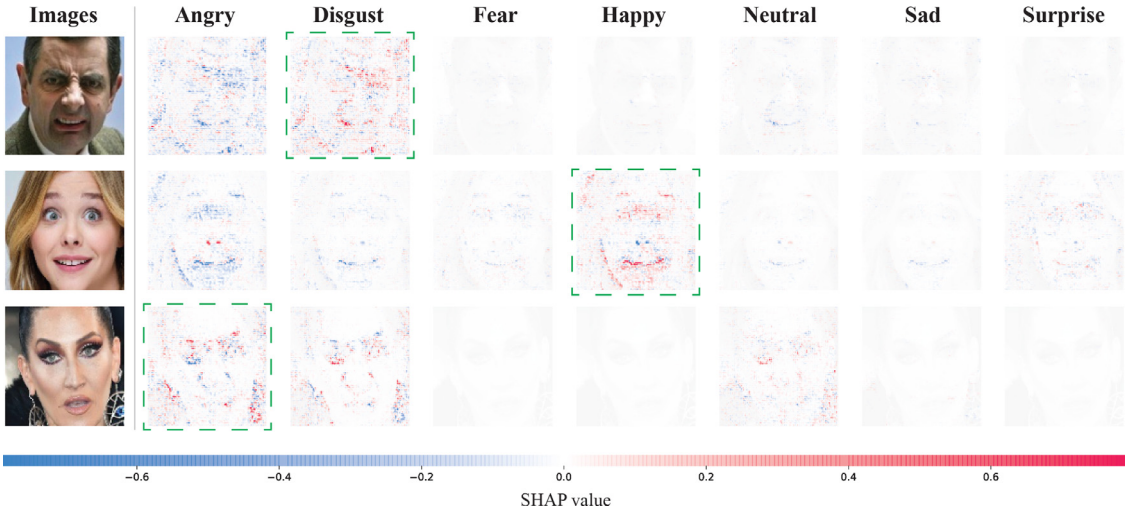


Fig. 11. Explanations generated by SHAP for three randomly selected images of three different classes—*Disgust, Happy* and *Angry. The model output is mainly influenced by the region around eyes, the alar crease and nostrils, the lips and the jaws.*

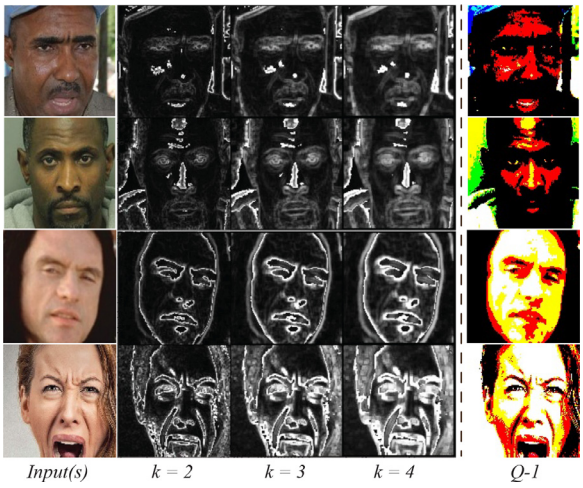


Fig. 12. Visualizing and comparing the contrast images of different skin-tones in multiple lighting conditions. The contrast image is computed by respectively combining the contrast value of a subset of image pixels as a pre-defined filter of different sizes strides through the image in a convolutional manner. ($k = 2$ denotes that a filter of size 2×2 has been used to compute the contrast image). *Dark skin-tone faces have typically lower visual contrast in facial features which may significantly degrade the image quality when quantization is applied.*

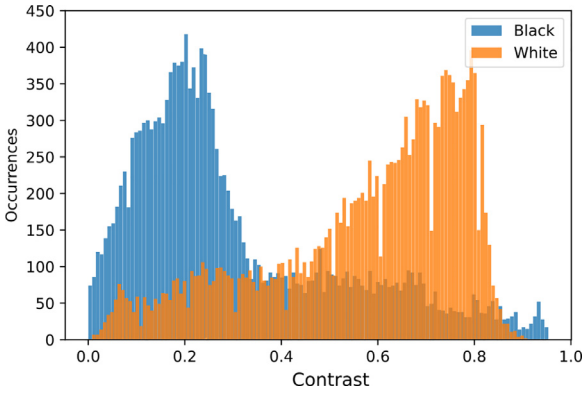


Fig. 13. Histograms of the contrast images for the skin-tone dataset. *The contrast values for the faces of darker skin-tone are significantly smaller than those for the fairer skin-tone, which may lead to significant image degradation when quantization is applied.*

values around key contributing regions—such as eyes, alar creases, and lips, as identified previously—as compared to the fair skin-tone faces causing a significant degradation in quantized images illustrated in the last column (Q-1) in Fig. 12. For a more comprehensive analysis, we compute the contrast images of all the dark and white skin-tone samples from the custom skin-tone dataset and report the values in the form of a histogram in Fig. 13. As evident from the figure, the dark skin-tone faces generally exhibit significantly lower contrast values as compared to the white skin-tone faces, which may be one of the reasons for a relatively poorer performance of the quantization-aware facial emotion recognition EML model.

5.6.2. Gender-based analysis

It can be seen from Fig. 8 that the performance of our employed model decreases with the increase in the strength of quantization, similar to the aforementioned white and dark skin-tone analysis. However, it is worth noting that male face images are comparatively more severely degraded than female faces, specifically notable for the Q-1 configuration. One of the reasons for this difference is that the male set loses more informative features due to the contrastive effect of strong quantization on Q-1 images, as shown in Fig. 8a. Whereas, a similar contrast effect caused by quantization only exposes the images, while retaining comparatively more information than male faces, especially for Q-1 configuration, as shown in Fig. 8b. Resultantly, the prediction of the HC-EML model prejudices against the male set, while it generated an accurate prediction against the female set, as depicted in Fig. 8c. We further analyzed the difference between the data distribution of male and female sets against Q-1, Q-2, Q-3, and Q-4 quantization strengths by calculating the SSIM to build up our assumption. It can be noted from Fig. 8d that male face images are degraded slightly more severely by Q-1 quantization than female face images. Specifically, quantized male face images show comparatively poor SSIM values, though slightly, than female face images, as shown in Fig. 8e which is an influential factor in reducing the performance of our model on male test samples.

6. Lessons learned and future work directions

Our case study presents two major findings. First, quantization-aware training can significantly increase the robustness of the EML

model against adversarial attacks over quantized inputs. Although more strongly quantized inputs reduce the accuracy of the model, they also more strongly resist adversarial attacks due to the inherent resilience of the quantization mechanism.

Secondly, although quantized images are hardware-friendly and require significantly less processing time and computational resources, we identify major stereotypical issues associated with quantizing the input images. Specifically, we have discovered that quantized inputs significantly favor white skin faces over dark skin faces. Ultimately, the model performs significantly poorly on the quantized inputs for dark skin faces as compared to white skin faces. We have also highlighted similar stereotypical discrimination between male and female faces, where quantization notably favors the female faces over the male faces, specifically the Q-1 quantization configuration.

Future works should focus on evaluating HC-EML models against privacy attacks—for example, the membership inference attacks (Choquette-Choo et al., 2021) and the model inversion attacks (Kahla et al., 2022)—and proposing novel defense methods to mitigate and counter such privacy attacks. Also, our analysis indicates a dire need for provably robust and private HC-EML models with verifiable guarantees to increase the proliferation of HC-EML-based applications in human patrons. Lastly, we strongly recommend that future researchers while developing HC-EML models utilize state-of-the-art bias mitigation techniques such as differential privacy and auto encoder-based latent structure learning methods (Amini et al., 2019) to improve the fairness and trustworthiness of EML models across different segments of society.

7. Conclusions

In this paper, we discuss the need for human-centric embedded machine learning (HC-EML), in particular, such applications that are enriched with social norms and human values to provide an efficient and safe experience. Keeping the human-centric artificial intelligence (HCAI) framework in view, we presented a pipeline for the development of HC-EML applications while highlighting major challenges and vulnerabilities encountered at each stage. Specifically, we categorized these challenges into four major groups, i.e., privacy, security, trustworthiness, and robustness. Based on the analysis, we discuss potential solutions to address the aforementioned challenges. To demonstrate the effectiveness of the proposed HC-EML pipeline, we presented a case study to investigate the security vulnerabilities and fairness of our human emotion-aware EML model. Our case study shows that although more strongly quantized inputs reduce the accuracy of the model, they also provide greater resistance to adversarial attacks. Further, we discovered major racial and gender stereotypical issues associated with quantizing the input images. Although quantized images have been widely used because of their hardware-friendliness, care must be taken in the future while deploying such systems in real-world applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Muhammad Atif Butt: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Adnan Qayyum:** Conceptualization, Methodology, Writing – review & editing. **Hassan Ali:** Conceptualization, Methodology, Writing – review & editing. **Ala Al-Fuqaha:** Conceptualization, Methodology, Funding acquisition, Writing – review & editing. **Junaid Qadir:** Conceptualization, Methodology, Funding acquisition, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgment

This publication was made possible by NPRP grant # [13S-0206-200273] from the Qatar National Research Fund (a member of Qatar Foundation). Open Access funding provided by the Qatar National Library The statements made herein are solely the responsibility of the authors.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318.

Ali, H., Khalid, F., Tariq, H.A., Hanif, M.A., Ahmed, R., Rehman, S., 2019. SSCNets: robustifying DNNs using secure selective convolutional filters. IEEE Des. Test 37 (2), 58–65.

Ali, H., Khan, M.S., Al-Fuqaha, A., Qadir, J., 2022. Tamp-X: attacking explainable natural language classifiers through tampered activations. Comput. Secur. 102791.

Ali, H., Khan, M.S., AlGhadhban, A., Alazmi, M., Alzamil, A., Al-Utaibi, K., Qadir, J., 2021. All your fake detector are belong to us: evaluating adversarial robustness of fake-news detectors under black-box settings. IEEE Access 9, 81678–81692.

Ali, H., Nepal, S., Kanhere, S. S., Jha, S., 2020. HaS-Nets: a heal and select mechanism to defend DNNs against backdoor attacks for data collection scenarios. arXiv preprint arXiv:2012.07474.

Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D., 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 289–295.

Andriushchenko, M., Croce, F., Flammarion, N., Hein, M., 2020. Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision. Springer, pp. 484–501.

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., Holzinger, A., 2022. Fairness and explanation in ai-informed decision making. Mach. Learn. Knowl. Extraction 4 (2), 556–579.

Athalye, A., Carlini, N., Wagner, D., 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International conference on machine learning. PMLR, pp. 274–283.

Banbury, C. R., Reddi, V. J., Lam, M., Fu, W., Fazel, A., Holleman, J., Huang, X., Hurtado, R., Kanter, D., Lokhmotov, A., et al., 2020. Benchmarking TinyML systems: challenges and direction. arXiv preprint arXiv:2003.04821.

Brendel, W., Rauber, J., Bethge, M., 2017. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248.

Butt, M.A., Khattak, A.M., Shafique, S., Hayat, B., Abid, S., Kim, K.-I., Ayub, M.W., Sajid, A., Adnan, A., 2021. Convolutional neural network based vehicle classification in adverse illuminous conditions for intelligent transportation systems. Complexity 2021.

Butt, M.A., Riaz, F., 2022. CARL-D: a vision benchmark suite and large scale dataset for vehicle detection and scene segmentation. Signal Process. Image Commun. 104, 116667.

Butt, M.A., Riaz, F., Mehmood, Y., Akram, S., 2021. REEEC-AGENT: human driver cognition and emotions-inspired rear-end collision avoidance method for autonomous vehicles. Simulation 97 (9), 601–617.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 39–57.

Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Tröster, G., Millán, J.d.R., Roggen, D., 2013. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recognit. Lett. 34 (15), 2033–2042.

Chen, G., Parada, C., Heigold, G., 2014. Small-footprint keyword spotting using deep neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4087–4091.

Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M., 2021. Ethical machine learning in healthcare. Annu. Rev. Biomed. Data Sci. 4, 123–144.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.-J., 2017. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26.

Chen, X., Liu, C., Li, B., Lu, K., Song, D., 2017b. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.

Cheng, Y., Wang, D., Zhou, P., Zhang, T., 2017. A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282.

Cheng, Y., Wang, D., Zhou, P., Zhang, T., 2018. Model compression and acceleration for deep neural networks: the principles, progress, and challenges. IEEE Signal Process. Mag. 35 (1), 126–136.

Chmiel, B., Banner, R., Shomron, G., Nahshan, Y., Bronstein, A., Weiser, U., et al., 2020. Robust quantization: one model to rule them all. Adv. Neural Inf. Process. Syst. 33, 5308–5317.

Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N., 2021. Label-only membership inference attacks. In: International Conference on Machine Learning. PMLR, pp. 1964–1974.

Chowdhery, A., Warden, P., Shlens, J., Howard, A., Rhodes, R., 2019. Visual wake words dataset. arXiv preprint arXiv:1906.05721.

Croce, F., Hein, M., 2020. Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International Conference on Machine Learning. PMLR, pp. 2196–2205.

Croce, F., Hein, M., 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning. PMLR, pp. 2206–2216.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T., 2021. Cooperative AI: machines must learn to find common ground.

David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Wang, T., et al., 2021. TensorFlow lite micro: embedded machine learning for TinyML systems. Proc. Mach. Learn. Syst. 3, 800–811.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., Anandkumar, A., 2018. Stochastic activation pruning for robust adversarial defense. arXiv preprint arXiv:1803.01442.

Doyu, H., Morabito, R., Brachmann, M., 2021. A TinyMLaas ecosystem for machine learning in IoT: overview and research challenges. In: 2021 International Symposium on VLSI Design, Automation and Test (VLSI-DAT). IEEE, pp. 1–5.

Dutta, L., Bharali, S., 2021. TinyML meets IoT: a comprehensive survey. Internet Things 16, 100461.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., Saunders, W., 2021. Truthful AI: developing and governing AI that does not lie. arXiv preprint arXiv:2110.06674.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S., 2019. Strip: A defence against trojan attacks on deep neural networks. In: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 113–125.

Giordano, M., Mayer, P., Magno, M., 2020. A battery-free long-range wireless smart camera for face detection. In: Proceedings of the 8th International Workshop on Energy Harvesting and Energy-Neutral Sensing Systems, pp. 29–35.

Giordano, M., Piccinelli, L., Magno, M., 2022. Survey and comparison of milliwatts micro controllers for tiny machine learning at the edge. In: 2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS). IEEE, pp. 94–97.

Giri, D., Chiu, K.-L., Di Guglielmo, G., Mantovani, P., Carloni, L.P., 2020. ESP4ML: platform-based design of systems-on-chip for embedded machine learning. In: 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, pp. 1049–1054.

Goodfellow, I. J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Gruenstein, A., Alvarez, R., Thornton, C., Ghodrat, M., 2017. A cascade architecture for keyword spotting on mobile devices. arXiv preprint arXiv:1712.03603.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI–explainable artificial intelligence. Sci. Rob. 4 (37), eaay7120.

Haney, J.M., Furman, S.M., Acar, Y., 2020. Smart home security and privacy mitigations: consumer perceptions, practices, and challenges. In: International Conference on Human-Computer Interaction. Springer, pp. 393–411.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Holzinger, A., 2021. The next frontier: AI we can really trust. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 427–440.

Huang, Y., Chen, C., 2022. Smart app attack: hacking deep learning models in android apps. IEEE Trans. Inf. Forensics Secur..

Huang, Y., Hu, H., Chen, C., 2021. Robustness of on-device models: adversarial attack to deep learning models on android apps. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, pp. 101–110.

Huang, Y., Zhang, H., Shi, Y., Kolter, J.Z., Anandkumar, A., 2021. Training certifiably robust neural networks with efficient local lipschitz bounds. Adv. Neural Inf. Process. Syst. 34.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A., 2019. Adversarial examples are not bugs, they are features. Adv. Neural Inf. Process. Syst. 32.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.

Kahla, M., Chen, S., Just, H. A., Jia, R., 2022. Label-only model inversion attacks via boundary repulsion. arXiv preprint arXiv:2203.01925.

Kenny, E.M., Keane, M.T., 2021. Explaining deep learning using examples: optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. Knowl. Based Syst. 233, 107530.

Khalid, F., Ali, H., Hanif, M.A., Rehman, S., Ahmed, R., Shafique, M., 2020. FaDec: a fast decision-based attack for adversarial machine learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Khalid, F., Ali, H., Tariq, H., Hanif, M.A., Rehman, S., Ahmed, R., Shafique, M., 2019. QuSecNets: quantization-based defense mechanism for securing deep neural network against adversarial attacks. In: 2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS). IEEE, pp. 182–187.

Khanal, S., Brodie, B., Xing, X., Lin, A.-L., Jacobs, N., 2022. Causality for inherently explainable transformers: CAT-XPLAIN. arXiv preprint arXiv:2206.14841.

Koizumi, Y., Saito, S., Uematsu, H., Harada, N., Imoto, K., 2019. ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp. 313–317.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al., 2020. Captum: a unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., Bacon, D., 2016. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z., 2020. Causal inference. Engineering 6 (3), 253–263.

Kurakin, A., Goodfellow, I., Bengio, S., 2016. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Latif, S., Rana, R., Qadir, J., 2018. Adversarial machine learning and speech emotion recognition: utilizing generative adversarial networks for robustness. arXiv preprint arXiv:1811.11402.

Lee, S., Lee, W., Park, J., Lee, J., 2021. Towards better understanding of training certifiably robust models against adversarial examples. Adv. Neural Inf. Process. Syst. 34.

Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 37 (3), 50–60.

Li, Y., Jiang, Y., Li, Z., Xia, S.-T., 2022. Backdoor learning: a survey. IEEE Trans. Neural Netw. Learn. Syst..

Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16463–16472.

Liao, Y., Vitak, J., Kumar, P., Zimmer, M., Kritikos, K., 2019. Understanding the role of privacy and trust in intelligent personal assistant adoption. In: International Conference on Information. Springer, pp. 102–113.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., Zhang, X., 2017. Trojaning attack on neural networks.

Liu, Y., Ma, X., Bailey, J., Lu, F., 2020. Reflection backdoor: a natural backdoor attack on deep neural networks. In: European Conference on Computer Vision. Springer, pp. 182–199.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Luo, N., Li, Y., Wang, Y., Wu, S., Tan, Y.-a., Zhang, Q., 2022. Enhancing clean label backdoor attack with two-phase specific triggers. arXiv preprint arXiv:2206.04881.

Lv, Z., Li, Y., Feng, H., Lv, H., 2021. Deep learning for security in digital twins of cooperative intelligent transportation systems. IEEE Trans. Intell. Transp. Syst..

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

McLean, G., Osei-Frimpong, K., 2019. Hey alexaQ examine the variables influencing the use of artificial intelligent in-home voice assistants. Comput. Human Behav. 99, 28–37.

Mitev, R., Miettinen, M., Sadeghi, A.-R., 2019. Alexa lied to me: skill-based man-in-the-middle attacks on virtual assistants. In: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp. 465–478.

Mohan, P., Paul, A.J., Chirania, A., 2021. A tiny CNN architecture for medical face mask detection for resource-constrained endpoints. In: Innovations in Electrical and Electronic Engineering. Springer, pp. 657–670.

Mohanta, B.K., Jena, D., Satapathy, U., Patnaik, S., 2020. Survey on IoT security: challenges and solution using machine learning, artificial intelligence and blockchain technology. Internet Things 11, 100227.

Mollahosseini, A., Hasani, B., Mahoor, M.H., 2017. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. 10 (1), 18–31.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. J. Big Data 2 (1), 1–21.

Otter, D.W., Medina, J.R., Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 604–624.

Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A., 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 582–597.

Peeters, M.M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M.A., Schraagen, J.M., Raaijmakers, S., 2021. Hybrid collective intelligence in a human–ai society. AI Soc. 36 (1), 217–238.

Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A., 2020. Secure and robust machine learning for healthcare: asurvey. IEEE Rev. Biomed. Eng. 14, 156–180.

Qayyum, A., Usama, M., Qadir, J., Al-Fuqaha, A., 2020. Securing connected & autonomous vehicles: challenges posed by adversarial machine learning and the way forward. IEEE Commun. Surv. Tutor. 22 (2), 998–1026.

Rajapakse, V., Karunanayake, I., Ahmed, N., 2022. Intelligence at the extreme edge: a survey on reformable TinyML. arXiv preprint arXiv:2204.00827.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J., 2022. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput. Biol. Med. 106043.

Rasib, M., Butt, M.A., Khalid, S., Abid, S., Raiz, F., Jabbar, S., Han, K., 2021. Are self-driving vehicles ready to launch? An insight into steering control in autonomous self-driving vehicles. Math. Probl. Eng. 2021.
Rasib, M., Butt, M.A., Riaz, F., Sulaiman, A., Akram, M., 2021. Pixel level segmentation based drivable road region detection and steering angle estimation method for autonomous driving on unstructured roads. IEEE Access 9, 167855–167867.
Ray, P.P., 2021. A review on TinyML: state-of-the-art and prospects. J. King Saud Univ.-Comput.Inform. Sci..
Ross, A., Doshi-Velez, F., 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.
Saha, A., Subramanya, A., Pirsiavash, H., 2020. Hidden trigger backdoor attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 11957–11965.
Sanchez-Iborra, R., Skarmeta, A.F., 2020. TinyML-enabled frugal smart objects: challenges and opportunities. IEEE Circuits Syst. Mag. 20 (3), 4–18.
Schölkopf, B., 2022. Causality for machine learning. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 765–804.
Shafique, M., Theocharides, T., Reddy, V.J., Murmann, B., 2021. TinyML: current progress, research challenges, and future roadmap. In: 2021 58th ACM/IEEE Design Automation Conference (DAC). IEEE, pp. 1303–1306.
Shidik, G.F., Noersasongko, E., Nugraha, A., Andono, P.N., Jumanto, J., Kusuma, E.J., 2019. A systematic review of intelligence video surveillance: trends, techniques, frameworks, and datasets. IEEE Access 7, 170457–170473.
Shneiderman, B., 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. Interact. Intell. Syst.(TiiS) 10 (4), 1–31.
Shneiderman, B., 2020. Human-centered artificial intelligence: reliable, safe & trustworthy. Int. J. Hum.–Comput.Interact. 36 (6), 495–504.
Sicara. Sicara/tf-explain: interpretability methods for tf.keras models with tensorflow 2.x. https://github.com/sicara/tf-explain.
Siregar, E., 2021. Learning human insight by cooperative AI: Shannon-Neumann measure. IOP SciNotes 2 (2), 025001.
Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., Holzinger, A., 2009. Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. In: International Conference on Universal Access in Human-Computer Interaction. Springer, pp. 615–624.
Sumalee, A., Ho, H.W., 2018. Smarter and more connected: future intelligent transportation system. Iatss Res. 42 (2), 67–71.
Sun, Z., Sun, R., Lu, L., Mislove, A., 2021. Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 1955–1972.
Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A., 2014. Machine learning for predictive maintenance: amultiple classifier approach. IEEE Trans. Ind. Inf. 11 (3), 812–820.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
Tange, K., De Donno, M., Fafoutis, X., Dragoni, N., 2020. A systematic survey of industrial internet of things security: requirements and fog computing opportunities. IEEE Commun. Surv. Tutor. 22 (4), 2489–2520.
Tjoa, E., Guan, C., 2020. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. 32 (11), 4793–4813.
Tran, B., Li, J., Madry, A., 2018. Spectral signatures in backdoor attacks. Adv. Neural Inf. Process. Syst. 31.
Tsoukas, V., Boumpa, E., Giannakas, G., Kakarountas, A., 2021. A review of machine learning and TinyML in healthcare. In: 25th Pan-Hellenic Conference on Informatics, pp. 69–73.
Usama, M., Qadir, J., Al-Fuqaha, A., 2018. Adversarial attacks on cognitive self-organizing networks: the challenge and the way forward. In: 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops). IEEE, pp. 90–97.
Usama, M., Qayyum, A., Qadir, J., Al-Fuqaha, A., 2019. Black-box adversarial machine learning attack on network traffic classification. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, pp. 84–89.
Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., Trichina, E., 2022. Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. Int. J. Hum. Resour.Manage. 33 (6), 1237–1266.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500.
Zhang, J., Chen, B., Zhao, Y., Cheng, X., Hu, F., 2018. Data security and privacy-preserving in edge computing paradigm: survey and open issues. IEEE Access 6, 18209–18237.
Zhang, M., Sawchuk, A.A., 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1036–1043.
Zhang, Y., Suda, N., Lai, L., Chandra, V., 2017. Hello edge: keyword spotting on microcontrollers. arXiv preprint arXiv:1711.07128.
Zhao, S., Ma, X., Zheng, X., Bailey, J., Chen, J., Jiang, Y.-G., 2020. Clean-label backdoor attacks on video recognition models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14443–14452.
Zhao, Y., Xu, K., Wang, H., Li, B., Jia, R., 2021. Stability-based analysis and defense against backdoor attacks on edge computing services. IEEE Netw. 35 (1), 163–169.
Zhu, L., Xu, X., Lu, Q., Governatori, G., Whittle, J., 2022. Ai and ethics–operationalizing responsible AI. In: Humanity Driven AI. Springer, pp. 15–33.

Muhammad Atif Butt is currently pursuing master's degree in computer science at the School of Electrical Engineering and Computer Science (SEECS), National University of Science and Technology (NUST), Islamabad, Pakistan. Meanwhile, he is working as a Research Assistant at IHSAN Lab, Information Technology University (ITU), Lahore, Pakistan. Before joining ITU in August 2021, he worked as a Research Associate with the Control, Automotive, and Robotics Laboratory, which is an affiliated lab of the National Centre of Robotics and Automation (NCRA), Pakistan. His research interests are in the areas of computer systems and intelligent information systems; applied artificial intelligence (AI) and adversarial machine learning (AML).

Adnan Qayyum is a PhD candidate at the Information Technology University (ITU) of Punjab, Lahore, Pakistan. His research interests include inverse medical imaging problems, healthcare, and secure and robust ML. He received the Bachelor's degree in Electrical Engineering from COMSATS, Pakistan, in 2014 and MS degree in Computer Engineeringfrom UET, Taxila, Pakistan, in 2016.

Hassan Ali is currently with IHSAN Lab, Information Technology University (ITU). He got his MS from the School of Electrical Engineering and Computer Sciences, NUST, Pakistan, with the President's gold medal. His research interests include embedded systems, machine learning, artificial intelligence, and security.

Ala Al-Fuqaha [S'00-M'04-SM'09] received PhD degree in Computer Engineering and Networking from the University of Missouri-Kansas City, Kansas City, MO, USA. He is currently a professor at the Information and Computing Technology (ICT) division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU). His research interests include use of ML in general and DL in particular in support of data-driven management of large-scale deployments of IoT and smart-city infrastructure and services, VANETs, and management and planning of SDNs. He is a senior member of IEEE and ABET Program Evaluator (PEV). He serves on editorial boards of multiple journals including IEEE Communications Letter and Network Magazine. He also served as chair, co-chair, and technical program committee member of multiple international conferences including IEEE VTC, Globecom, ICC, and IWCMC.

Junaid Qadir [SM'14] is a Professor of Computer Engineering at the Qatar University in Doha, Qatar, where he directs the IHSAN Research Lab. His primary research interests are in the areas of computer systems and networking, applied machine learning, using ICT for development (ICT4D); human-beneficial artificial intelligence; ethics of technology, artificial intelligence, and data science; and engineering education. He has published more than 150 peer-reviewed articles at various high-quality research venues including publications at top international research journals including IEEE Communication Magazine, IEEE Journal on Selected Areas in Communication (JSAC), IEEE Communications Surveys and Tutorials (CST), and IEEE Transactions on Mobile Computing (TMC). He was awarded the highest national teaching award in Pakistan the higher education commission's (HEC) best university teacher award for the year 2012–2013. He has obtained research grants from Facebook Research, Qatar National Research Fund, and the HEC, Pakistan.He has been appointed as ACM Distinguished Speaker for a three-year term starting from 2020.He is a senior member of IEEE and ACM.