



# Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks

Esma Mansouri-Benssassi<sup>1</sup> · Juan Ye<sup>1</sup>

Published online: 16 January 2021  
© The Author(s) 2021

## Abstract

Emotion recognition through facial expression and non-verbal speech represents an important area in affective computing. They have been extensively studied from classical feature extraction techniques to more recent deep learning approaches. However, most of these approaches face two major challenges: (1) robustness—in the face of degradation such as noise, can a model still make correct predictions? and (2) cross-dataset generalisation—when a model is trained on one dataset, can it be used to make inference on another dataset?. To directly address these challenges, we first propose the application of a spiking neural network (SNN) in predicting emotional states based on facial expression and speech data, then investigate, and compare their accuracy when facing data degradation or unseen new input. We evaluate our approach on third-party, publicly available datasets and compare to the state-of-the-art techniques. Our approach demonstrates robustness to noise, where it achieves an accuracy of 56.2% for facial expression recognition (FER) compared to 22.64% and 14.10% for CNN and SVM, respectively, when input images are degraded with the noise intensity of 0.5, and the highest accuracy of 74.3% for speech emotion recognition (SER) compared to 21.95% of CNN and 14.75% for SVM when audio white noise is applied. For generalisation, our approach achieves consistently high accuracy of 89% for FER and 70% for SER in cross-dataset evaluation and suggests that it can learn more effective feature representations, which lead to good generalisation of facial features and vocal characteristics across subjects.

**Keywords** Spiking neural network · Facial emotion recognition · Speech emotion recognition · Unsupervised learning

## 1 Introduction

Emotions recognition represents one of the most important aspects in affective computing with a wide range of applications areas from human–computer interaction, social robotics, and behavioural analytic (Hsu et al. 2013). Emotions are expressed through various means, such as verbal, non-verbal speech, or facial expression and body language. Emotion recognition from facial expression and speech is the most studied in affective computing, either as separate or joined modality (Vinola and Vimaladevi 2015).

Facial expressions represent non-verbal means of expressing emotions and mental states. They are defined by the deformation of multiple facial muscles, forming representations of different emotions. Emotion from speech defines all the non-verbal and verbal cues that represent different emotional states. Speech emotion recognition (SER) represents one of the most popular means for emotion recognition and has been extensively investigated.

Over years, FER and SER have developed significantly with advances in computer vision, speech signal processing, and deep learning. Convolutional neural networks (CNNs) have demonstrated promising results in both FER and SER, because of their ability to extract effective feature representations to distinguish different facial parts (Khorrami et al. 2015) and distinctive speech features from raw audio (Harár et al. 2017).

However, state-of-the-art techniques still face two major challenges, that is, *robustness* to input degradation with noise and *cross-dataset generalisation* capacity. When the quality of image data or speech samples is compromised, most of

---

Communicated by Miguel A. Vega-Rodríguez.

✉ Esma Mansouri-Benssassi  
emb24@st-andrews.ac.uk

Juan Ye  
jy31@st-andrews.ac.uk

<sup>1</sup> Computer Science, University of St Andrews, St Andrews, UK

the existing techniques will lead to significant decrease in recognition accuracy (Aghdam et al. 2016). Cross-datasets generalisation refers to the ability to generalise feature learning across datasets created using different subjects, ethnic groups, facial dimensions, and characteristics, for example, different shapes and sizes of key facial regions like eyes or mouth or even data acquisition conditions (Lopes et al. 2017). So far, existing approaches face difficulty achieving cross-dataset generalisation and they perform worse on unseen data. On top of these two challenges, most of the existing approaches rely on well-annotated training data in order to learn distinctive features to separate different types of emotions. However, the annotated training data are challenging to acquire.

This paper explores the use of biologically plausible models, spiking neural networks (SNNs), to directly address the above challenges. In contrast to artificial neural network (ANN), SNN has the advantage of capturing precise temporal pattern in spiking activity, which leads to crucial coding strategy in sensory information processing and the success in many pattern recognition tasks (Tavanaei et al. 2019). The main motivation of this paper is to investigate and explore neuromorphic algorithms with unsupervised learning for cognitive tasks such as facial expression recognition or speech emotion recognition. The key novelty of this paper is the adaptation of bio-inspired model with unsupervised learning for FER and SER to extract meaningful features that can be generalised across datasets and be robust to noise degradation. Presented as the third generation of neural networks (Maass 1997; Hodgkin and Huxley 1990), SNNs have been successfully applied to simulate brain processes for different tasks including pattern recognition and image processing (Jose et al. 2015). Our contributions are listed in the following.

1. We have successfully applied SNNs with unsupervised learning to FER and SER tasks on two types of data: static images in FER and time series data in SER.
2. We have achieved higher recognition accuracy than the state-of-the-art techniques through cross-dataset evaluation. It demonstrates the generalisation capability and subject independence of SNN for both SER and FER.
3. We have achieved higher accuracy than the state-of-the-art techniques in noise robustness experiments where we inject salt and pepper noise, Gaussian noise with different noise intensities to images and inject white, pink, or brown noise to speech data.

The rest of the paper is organised as follows. Section 2 presents the state of the art in FER, and Sect. 3.2 introduces spiking neural network (SNN) and describes how we apply SNN to support unsupervised learning in FER. Section 4 discusses the conducted experiments and results obtained on

overall accuracy, generalisation, and image degradation by noise tasks. We compare our results with some selected state-of-the-art approaches, namely HOG features extraction with SVM classifier and a CNN. Section 6 concludes the paper and points to future research direction.

## 2 Related work

Emotion recognition research from both images and speech data has developed significantly over the recent years with the advances in machine learning and the availability of more and larger datasets. This section will present an overview of the state of the art in emotion recognition through both audio and visual data.

### 2.1 Facial expression recognition

Extracting meaningful features from input images represents a crucial step in FER classification process. This can be achieved with the following three main approaches: appearance-based, model-based, and deep learning techniques.

Appearance features are a set of features based on the changes of the image texture (Mishra et al. 2015). One of the most used approaches is local binary pattern (LBP) for texture analysis. Liu et al. (2016) have used LBP, in combination of grey pixel values with the addition of principal component analysis (PCA) for dimensionality reduction of the obtained features. They have used active facial patches on region of interest (ROI) where major changes occur in facial expressions.

Histograms of ordered gradients (HOGs) is another popular approach (Dalal and Triggs 2005). HOG descriptors are based on constructing a histogram feature vector by computing the accumulation of gradient direction over each pixel of a small region. Carcagni et al. (2015) have conducted a comprehensive study on using HOG feature extraction for facial expression recognition. They have compared various parameters such as cell sizes and orientation bins.

Model-based techniques have been applied to track facial muscles deformation by constructing models of the face. Tie and Guan (2013) have proposed a 3D deformable facial expression model with 26 fiducial points that are tracked through video frames using multiple particle filters. They then use a discriminative Isomap-based classification to classify the tracked facial deformation into a facial expression of emotion. Gilani et al. (2017) have used a 3D face model to compute the correspondence between different constructed 3D models of different faces. This is achieved by morphing the model to new faces. They have achieved high accuracy for gender and facial recognition.

Recently, research has turned towards using deep learning for automatic facial expression recognition, which has achieved promising results for general pattern recognition. Kim et al. (2016) have used discriminative deep model to learn the mapping between aligned and non-aligned facial images. Lopes et al. (2017) have extended the CNN with specific data pre-processing and augmentation approaches in order to overcome small datasets training. They have added eye localisation, rotation correction, and intensity normalisation before feeding their training data to the CNN network. Mollahosseini et al. (2016) have proposed a novel architecture for a CNN with two convolutional layers where each is followed by a max pooling and four Inception layers. Using inception layers gives more depths and width to the network without affecting the computational cost. This is the result of using transfer learning on a pre-trained network. Mehendale (2020) has applied 2 CNNs: one for removing background and the other for extracting facial features. Rzaeva and Alasgarov (2019) have designed 5-layered CNN with multi-dimension of images and different pooling techniques to improve emotion recognition accuracy.

Unsupervised learning has also been researched. Majumder et al. (2016) have extracted geometric and LBP features from facial images and then fused them for unsupervised classification using Kohonen self-organizing map (SOM).

The majority of existing approaches have achieved high recognition accuracy with the ability in extracting and learning distinctive features in a supervised learning manner. However, the features are often subject to subtle changes in each facial area and thus sensitive to noise, thus lacking generalisation ability.

## 2.2 Speech emotion recognition

The most important step in speech emotion recognition (SER) tasks consists of extracting and learning features translating differences of various emotional states in speech (Akasay and Oauz 2020). Audio features can represent both speech and non-speech. We describe some of the most popular features in the following sections.

### 2.2.1 Mel frequency cepstral coefficients (MFCCs)

MFCCs are the most biologically plausible method and mimic how human processes sound (Gupta et al. 2018). They are computed as a linear cosine transformation of log power spectrum representing short-term power of signals.

### 2.2.2 Spectral centroid

Spectral centroid represents the centre mass of the spectrum magnitude indicating quick changes in the audio signal (Tavarez et al. 2017). They are computed with the centre

mass of the magnitude of spectrum. They have been successfully used along with convolution neural network (Cummins et al. 2017).

### 2.2.3 Pitch

Pitch is the quality of the signal, and it represents the nature of a tone being either low or high.

### 2.2.4 Energy

Energy is usually calculated from small time intervals and consists of finding the presence of a signal through a temporal interval.

### 2.2.5 Classifiers for SER

There are two types of learning in SER: static and dynamic. Static learning aims to recognise emotion through the whole utterance on auditory features (Chavhan et al. 2010; Yang et al. 2017; Papakostas et al. 2017; Tashev et al. 2017), while dynamic learning partitions an auditory signal into frames and focuses on learning temporal relationships between frames in emotion recognition (Anagnostopoulos et al. 2015). For example, Yang et al. have fed the above features in a support vector machine (SVM) to recognise different emotional states (Yang et al. 2017). Anagnostopoulos et al. have used hidden Markov model (HMM) in dynamic learning (Anagnostopoulos et al. 2015). Lech et al. (2020) have used a pre-trained image classification network AlexNet to enable real-time speech emotion recognition.

### 2.2.6 Deep learning

Using deep learning in SER becomes popular as deep learning has outperformed most classic machine learning techniques (LeCun et al. 2015; Papakostas et al. 2017). Methods using hand-crafted features focus on the whole signal rather than the temporal and dynamic relation, thus losing the temporal dimension. Lee and Tashev (2015) have used deep learning by investigating recurrent neural network (RNN) for learning different feature representations of an audio signal. Satt et al. (2017) first train a CNN with spectrogram features using various networks architectures. They then compare the results with the use of long short-term memory (LSTM) where LSTM proves to be useful for audio task. LSTM has an overall accuracy of 68% compared to 62% on CNN. Another type of deep learning technique is introduced by Niu et al. (2018). They propose the use of deep retinal convolution neural network (DRCNN). They first use data augmentation techniques on the spectrogram features extracted from audio signals. They then apply a deep CNN

on the extracted features and obtain an overall accuracy of 99%.

### 2.2.7 Bio-inspired approaches

Bio-inspired approaches have been rarely used for SER tasks. There have been some early attempts (Buscicchio et al. 2006), where Buscicchio et al. use SNN for extracting linguistic features by decomposing speech input in different parts for vowels occurrences.

Lotfidereshgi and Gournay (2017) use liquid state machine (LSM) for speech recognition using raw speech data. LSM includes a *reservoir* represented by a SNN (Gallicchio et al. 2017) that represents a form of reservoir computing. Their method goes first through pre-processing steps by applying linear prediction analysis. They have achieved accuracy of 82.35% which is comparable to the state-of-the-art techniques.

## 3 Proposed approach

In this section, we describe the application of SNNs with unsupervised learning in FER and SER tasks. We will start with a brief introduction to SNN and then describe the process in more details for both tasks.

### 3.1 Introduction to spiking neural networks

Information in the brain is transmitted between neurons using action potentials via synapses. When a membrane potential reaches a certain threshold, a spike is generated (Jose et al. 2015). There have been various attempts in the literature to create and simulate computational processes in the brain. Spiking neural networks represent the third generation of neural networks and are an attempt to model how the brain processes information (Maass 1997; Hodgkin and Huxley 1990). The main difference from artificial neural networks is that SNNs process information based on spikes, where neurons communicate through series of spikes by firing spikes when they reach a certain threshold (Filip and Andrzej 2011). The computation of SNNs is based on timing of spikes in that spikes that fire together get a stronger connection.

There exist various types of SNNs using different types of learning. Huxley–Hodgkin (Gavrilov and Panchenko 2016) represents an early attempt, which is based on modelling electrochemical information transmission between neurons using electrical circuits containing capacitors and resistors. It successfully models biologically realistic properties of membrane potentials, with realistic behaviours comparable to natural neurons. This is characterised by a sudden and large increase at firing time, which is followed by a refractory period where a neuron cannot spike again, followed by a

time interval where the membrane is depolarised. Although Hodgkin–Huxley model demonstrates to be very powerful to model neuronal behaviours realistically, its implementation is very complex for numerically solving the system of differential equation using SNNs.

Leaky integrate and fire model (LIF) is a simplification of Hodgkin–Huxley models by considering every spike as a uniform event defined solely by the time of spiking. Compared to Hodgkin–Huxley models, LIF models are less biologically plausible but less computationally costly. Therefore, we select LIF in our work.

Similarly to classical artificial neural networks (ANN), SNNs can be designed using different topologies:

- Feedforward: In this topology, information flows in one direction with no feedback connection. These kinds of topology are usually used in SNN to model low-level sensory systems, such as vision systems. They have also been used for binding tasks such as spatio-temporal spikes or spike synchronisation (Tapson et al. 2013; Sporea and Grüning 2012; Tavanaei and Maida 2017).
- Recurrent: Neurons interact through feedback connections, where a dynamic temporal behaviour represents the network. Although this topology is harder to compute, it can have higher computational power. Recurrent architectures are particularly useful for modelling or analysing dynamic objects. However, it is computationally more challenging to apply supervised learning on this type of architecture (Demin and Nekhaev 2018). Recurrent architectures can also be applied to investigate extensive population activities and analyse neuronal populations dynamics.

Feedforward topology is the most common topology for general pattern recognition as it mimics the hierarchical structure of visual cortex (Al-Yasari and Al-Jamali 2018). This topology represents the right candidate for tasks such as emotion recognition, which is therefore selected in our architecture.

Learning in SNNs also takes various forms:

- *Supervised learning* that is achieved through applying Hebbian learning. The supervision is done through a spike-based Hebbian process by reinforcing the post-synaptic neuron in order to fire at preset timing and not spike at other times. The reinforcement signal is transmitted through synaptic currents (Knudsen 1994).
- *Unsupervised learning* follows the basic Hebb's law, where neurons that fire together are connected (Hebb 1962). Automatic reorganisation of connection in the Hebbian learning permits the ability of unsupervised learning with various potential applications, such as clustering or pattern recognition. Unsupervised learning with

Hebbian formula enables learning of distinct patterns without using classes labels or having a specific learning goal (Hinton et al. 1999; Bohte et al. 2002; Grüning and Bohte 2014).

- *Reinforcement learning* that enables learning directly from the environment where SNN includes a rewarding signal spike (Farries and Fairhall 2007).

In this paper, we explore the use of unsupervised learning in speech and facial emotion recognition.

### 3.2 Application of SNN with unsupervised learning for FER

This section describes the application of SNN in FER tasks. The process follows different steps, including input encoding, choice of learning rules, and network topology of SNN.

#### 3.2.1 Image pre-processing

We apply pre-processing on input images by applying filters to defines contours of the input images. Filters such as difference of Gaussian (DoG) have been successfully applied to pre-process data and prepare it as input to SNN. For example, DoG has been applied on pre-process handwriting images (Kheradpisheh et al. 2017). In this work, we apply Laplacian of Gaussian (LoG) to extract contours and edges of facial expression on input images. Although LoG and DoG are quite similar, where the DoG represents an approximation of the LoG. LoG is selected for use as it achieves higher precision (Marr and Hildreth 1980) and is represented in Eq. 1.

$$\nabla^2 G_\sigma(x, y) = \frac{\partial G_\sigma(x, y)}{\partial x^2} + \frac{\partial G_\sigma(x, y)}{\partial y^2} \tag{1}$$

where  $\nabla^2$  is the Laplacian operator,  $\sigma$  is the smoothing value, and  $G_\sigma(x, y)$  is the Gaussian filter applied to the image, given by:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

Given a facial image, we first apply the Gaussian filter to smooth and remove noise and then apply the Laplacian filter to locate edges and corners of the image.

#### 3.2.2 Image encoding

From the contours defining faces and various facial features obtained from the pre-processing step using LoG, spike trains are created using Poisson distribution. The firing rates of the spikes trains are proportionate to the input pixels intensity.

The Poisson distribution  $P$  is given by the following equation:

$$P(n) = \frac{(rt)^n}{n!} \exp^{-rt} \tag{3}$$

where  $n$  is the number of spikes occurring in a time interval  $\Delta t$  and  $r$  is randomly generated in a small time interval where only one spike occurs. Each  $r$  has to be less than the firing rate in the  $\Delta t$  time interval.

#### 3.2.3 Network dynamics of SNN

There are several models translating neurons behaviour, including integrate-and-fire, leaky-integrate-and-fire, and Hodgkin–Huxley models (Kheradpisheh et al. 2017).

The leaky-integrate-and-fire is the most commonly used model as it is simple and computationally efficient. Its network dynamics are captured in the following equation. We have built from the work presented in Diehl and Cook (2015). Although the original work was for general pattern recognition task, we have identified a potential to use it for FER tasks:

$$\tau \frac{dV}{dt} = (E_{rest} - V) + g_e(E_{exc} - V) + g_i(E_{inh} - V) \tag{4}$$

$V$  is the membrane voltage and  $E_{rest}$  represents the resting membrane potential.  $E_{inh}$  and  $E_{exc}$  represent the equilibrium potential for the inhibitory and excitatory synapses, respectively.  $g_e$  and  $g_i$  represent the conductance of the synapses.  $\tau$  is a time constant representing the time a synapse reaches its potential, and it is longer for the excitatory neurons. When a membrane reaches a certain threshold, the neuron fires a spike and then enters into a resting phase  $E_{rest}$  for a certain time interval. This represents a refractory period where the neuron cannot spike.

#### 3.2.4 Unsupervised learning

We adopt spike timing-dependent plasticity (STDP) learning (Diehl and Cook 2015) to perform unsupervised learning of FER. STDP is a process based on the strength of connection between the neurons in the brain. The strength represents the conductance that is increased when a pre-synaptic spike arrives at a synapse. It will be adjusted based on the relative timing between the input represented as spikes and outputs.

The principle of STDP learning is based on the update of weights according to the temporal dependencies between pre-synaptic and post-synaptic spikes. The weights learn different features in the input images in an unsupervised manner without the provision of labels. Weights are updated when a pre-synaptic trace reaches a synapse. A trace represents the tracking of changes in each synapse. When a pre-synaptic

spike arrives at a synapse, the trace is increased by 1; otherwise, it decays exponentially. The learning rule, characterised in Eq. 5, defines how weights are updated in each synapse.

$$\Delta w = \eta(x_{\text{pre}} - x_{\text{tar}})(w_{\text{max}} - w)^\mu \quad (5)$$

where  $\Delta w$  represents the weight change,  $\eta$  represents the learning rate,  $\mu$  is a rate determining the dependence of the update on the previous weight,  $x_{\text{tar}}$  is the target value of the pre-synaptic trace, and  $w_{\text{max}}$  is the maximum weight. The target value  $x_{\text{tar}}$  ensures that pre-synaptic neurons that rarely lead to firing of the post-synaptic neuron will become more and more disconnected and is especially useful if the post-synaptic neuron is only rarely active.

There also exist other STDP learning rules, such as exponential weight dependence, and inclusion of a post-synaptic trace (Diehl and Cook 2015). We have chosen the learning rule that reports the best performance in the original paper.

### 3.2.5 SNN architecture

We first introduce the workflow of applying SNN to FER, from image pre-processing to classification in Fig. 1. Raw input goes through the first layer where an image filter is applied and the input is encoded into spike trains. It is then connected to a convolution layer where each input is divided into several features of the same size and a stride window that is moved throughout the whole input. Applying convolution layer proves beneficial for increasing the overall accuracy in Saunders et al. (2018). Each convolution window forms a feature, which represents an input to the excitatory layer. The number of neurons  $O$  in the convolutional layer is calculated through the formula:

$$O = \frac{(\text{in}_{\text{size}} - c_{\text{size}}) + 2P}{c_{\text{stride}}} + 1 \quad (6)$$

where  $\text{in}_{\text{size}}$  is the input image size in the input layer,  $c_{\text{size}}$  is the size of each feature in the convolutional layer,  $c_{\text{stride}}$  is the size of the stride in the convolutional layer, and  $P$  is the padding.  $O$  is the convolutional output size that represents the squared root of the number of neurons in the convolutional layer. The third layer represents an inhibitory layer where feature neurons are inhibited apart from the one that a neuron is connected to. The number of neurons in the inhibitory layer is proportional to the number of patches in the excitatory layer.

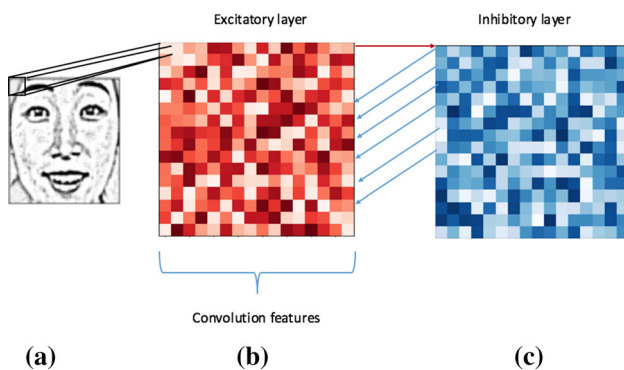
### 3.2.6 Classification

After training is completed by presenting training input images to the network and for each training interval, neurons for each features are assigned a class label based on their spiking pattern for each one.

## 3.3 Application of SNN with unsupervised learning for SER

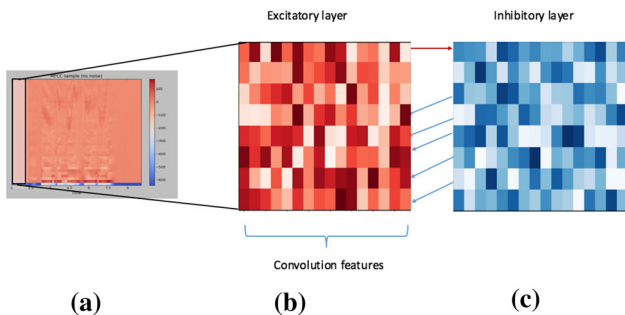
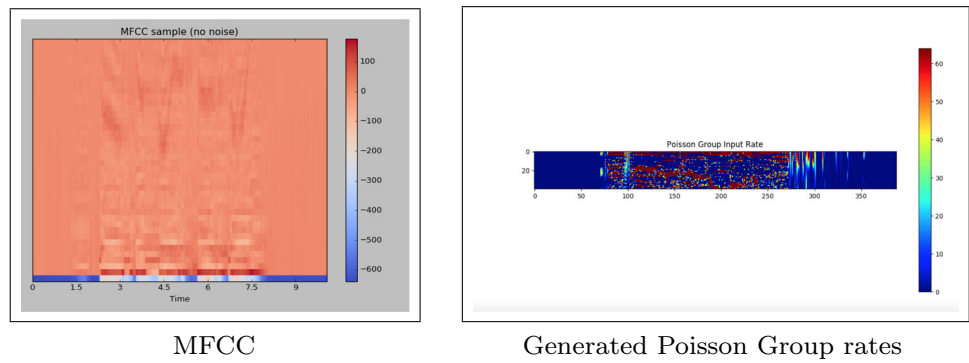
For SER task, we have used the same basic architecture as in Sect. 3.2. The main difference resides in the input layer and nature of the convolution layer configuration choice, where for SER tasks convolution window is moved through the temporal axis. We have initially experimented with two different network architectures. The first approach consists of dividing the input into different frames where each frame represents an input to the network with a 1D convolution layer. The second experimented approach takes extracted features such as MFCCs and inputs them to the convolution layer running across the temporal axis. Although we have experimented with raw audio data and spectrogram features, we have opted with the choice of MFCCs as input for SNN because they have achieved better performance. MFCC represents one of the most popular and mainly used feature extraction techniques for speech recognition tasks and speech emotion recognition tasks in particular (Gupta et al. 2018). They represent short-term power spectrum of an audio signal and are the closest to mimic the human hearing system.

We then encode MFCC features into a population of Poisson spike trains (Diehl and Cook 2015). Each extracted input represents a firing rate proportionate to its intensity, and each feature value over time is transformed into firing rate between 0 and 63.75 Hz. The input data are run through the network for 350 ms (Diehl and Cook 2015). After that, the network enters a resting phase for 150ms, in order to get back to its initial equilibrium before receiving the next input. Figure 2 presents the process of MFCC feature extraction from raw



**Fig. 1** SNN workflow for FER: **a** LoG filters are applied to raw input, and then, the input is processed to create Poisson spikes train. **b** Excitatory convolutional layer where a number of features, stride, and convolution window are chosen. **c** Inhibitory layer where each neuron inhibits all convolutional feature neurons apart from the one it receives input from

**Fig. 2** MFCC and corresponding Poisson group rates



**Fig. 3** SNN workflow for SER: **a** MFCC features are extracted, and Poisson spike train is created. **b** Excitatory convolution layer where a number of features, stride, and convolution window are chosen and convolution moves through temporal axis. **c** Inhibitory layer where each neuron inhibits all convolution features apart from the one it receives input from

audio data and the rates created on the input Poisson group for SNN.

## 4 Experiment set-up

The main objectives of the evaluation are (1) whether using more bio-inspired model such as SNNs with unsupervised learning achieves comparable accuracy to the state-of-the-art supervised learning techniques; (2) whether SNNs exhibit robustness to degradation such as noise; and (3) whether it has generalisation capacity. To assess the above questions, we design the following evaluation methodology.

### 4.1 Datasets

We have used various publicly available datasets for both FER and SER tasks.

#### 4.1.1 Datasets for FER tasks

For FER, we use two widely used datasets: JAFFE and CK+. The CK+ dataset consists of 3297 images of 7 basic emotions, including happy, surprised, sad, disgusted, fearful,

angry, and neutral. The emotions are recorded on 210 adults aged between 18 and 50 years with a higher proportion of females and different ethnic backgrounds. Each video starts with a neutral expression, progresses to an expression, and ends with the most intensified expression. As we focus on the main 6 expressions excluding the neutral one, we extract frames where the expression is more emphasised from the videos. JAFFE dataset consists of 221 images of the same 7 emotions. These emotions are acted by Japanese females in a controlled environment. We use only 6 emotions excluding the neutral one. We use OpenCV (Bradski 2000) to crop face area of each image, which will be then resized to a uniform size and converted to greyscale.

#### 4.1.2 Datasets for SER tasks

We have used two third-party datasets for SER tasks. Ryerson audio-visual database of emotional speech and song (RAVDESS) (Livingstone and Russo 2018) is a multi-modal dataset with all basic emotions where recordings include both songs and sentence reading. It comprises 24 participants where sentences are recorded as audio only, video only, and audio-visual.

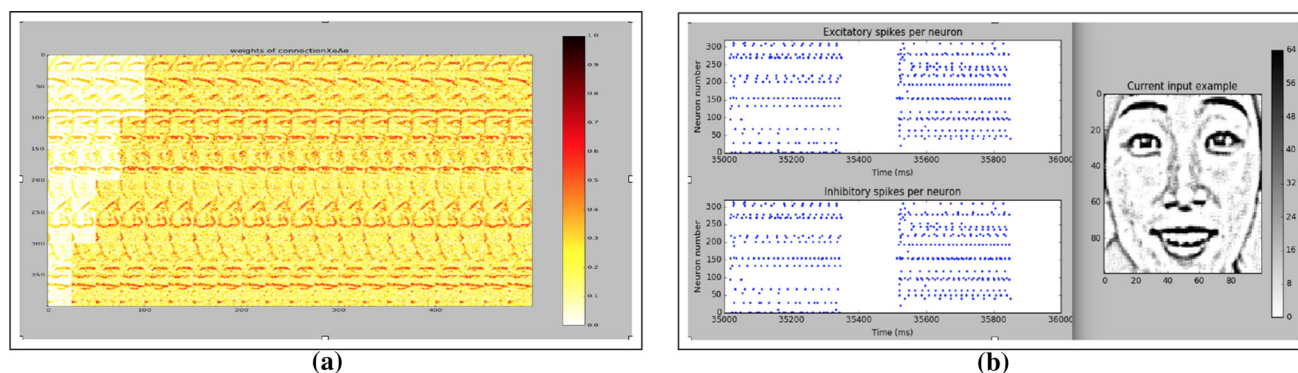
The other dataset is the eNTERFACE dataset (Pitas et al. 2006), which includes 42 subjects representing 14 nationalities where there are more male participants than female ones.

### 4.2 SNN configuration

This section describes the SNN configuration for the experiments conducted for both SER and FER. All experiments and settings are achieved with Brian open-source SNN simulator (Goodman and Brette 2008).

#### 4.2.1 FER

Here, we describe SNN configuration for FER task as described in (Mansouri-Benssassi and Ye 2018).



**Fig. 4** **a** Excitatory and inhibitory neuron spikes. **b** Learned weights for a convolution of size 25, stride 25, and feature size 20

*Image pre-processing and encoding for SNN* To prepare the input for the SNN, we apply LoG filter to extract the contours of different facial features. Each filtered image is then encoded into a Poisson spike train where the firing rate corresponds to the intensity of each pixel. The highest rate used is derived from the original paper (Diehl and Cook 2015) which is 63.73 Hz, and the lowest one is 0. They correspond to the highest and lowest pixel intensity (from 0 to 255).

*SNN configuration and learning* The chosen network configuration consists of a convolution layer containing 50 features, with a stride size of 15 and convolution size of 15. Various stride sizes, convolutions, and features are experimented. The higher the number of features, the smaller the stride size results in a better performance. This configuration is retained as it performs the best. The input data are all re-sized to  $100 \times 100$ . Thus, the number of neurons in the input layers is set to 10000. At the convolution layer, the number of neurons is calculated using the chosen number of strides and convolution size according to Eq. 6. We have used the same parameters for SNN as in the original paper (Diehl and Cook 2015), and the online STDP learning is applied. The weights are learned by either being increased when a post-synaptic neurons fire after a spike reaches a synapse, or decreased when the post-synaptic spike fires before a spike arrives at a synapse.

Figure 4a shows an example of learned weights for a configuration of 20 convolution features with size 25 and stride 25. In practice, this set-up is too coarse to capture fine-grained features so the actual configuration used for our experiment is the larger feature size 50 with the smaller convolution size 15 and the smaller stride size 15. When an input is presented for 350 ms, spikes are recorded for both excitatory and inhibitory layers as shown in Fig. 4b, where a group of neurons spike for different features. During learning, each group of neurons will learn a particular feature and neurons are assigned to a class label when the neurons have spiked the most.

#### 4.2.2 SER

As detailed in the approach section, SNN follows the same architecture for SER as in FER, and we use different feature extraction encodings for audio. We apply Librosa open-source library (McFee et al. 2015) to extract MFCCs features from audio inputs. The number of energies of filter banks is set at 40, which is the number of features per frame. All audio features are unified to have a temporal length of 388, which is the frame size. Audio signals which result in smaller size are padded to match the chosen setting.

## 5 Evaluation and results

This section presents the experiments results addressing the three evaluation objectives in Sect. 4.

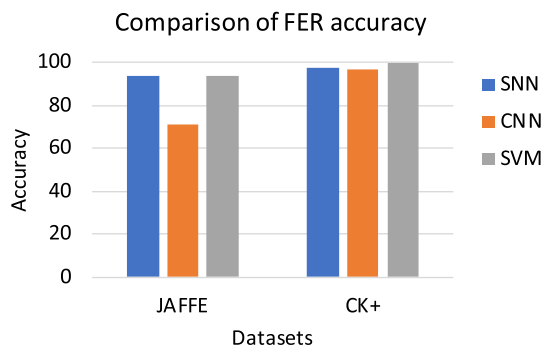
### 5.1 Performance of unsupervised learning of SNN

#### 5.1.1 FER

In order to demonstrate the advantage of SNN in FER, we compare its performance with the state-of-the-art techniques, including HOG, LBP, and geometrical/coordinates-based features applied with a SVM classifier (Dalal and Triggs 2005) and CNN. Among them, HOG with SVM classifier and CNN with data augmentation perform the best. They are selected as baselines for manual and automatic feature extraction techniques, respectively.

We use the scikit-image library (van der Walt et al. 2014) to extract HOG features for each image, resulting in a feature vector of 22500. The features are then fed into a linear SVM for classification, as SVM is one of the most popular classifiers for FER (Majumder et al. 2016). We first use a VGG16 pre-trained on ImageNet for general image classification task and then retrain the last layer with the features obtained. The small network is a one-layer dense network configured with 256 nodes with a softmax activation function. We have tried





**Fig. 5** Comparison of overall accuracy for FER tasks on 2 datasets between SNN, CNN, and HOG+SVM

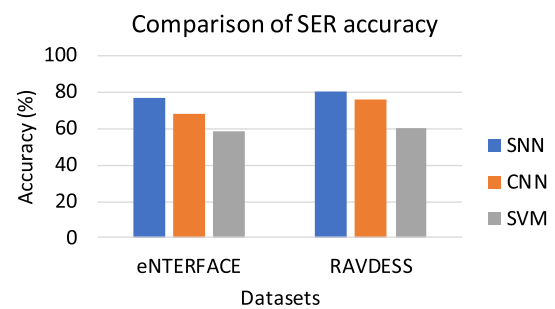
various configurations of the network with multiple layers and different numbers of nodes, but the performance does not vary much. We also added some commonly used data augmentation techniques such as cropping, rotation, and flip using Keras library (Chollet 2015).

Figure 5 compares FER accuracy of SNN, HOG with SVM, and CNN on CK+ and JAFFE datasets. On each dataset, we apply repeated holdout with 10 trials by splitting data into 60% for training, 20% for validation, and 20% for testing in CNN-based models. For SNN and SVM models, we split data into 80% for training and 20% for testing in SNN and SVM models. Data are shuffled randomly with a balanced distribution within classes on both training and testing data. We obtain the accuracy for the 10 trials and average the accuracy.

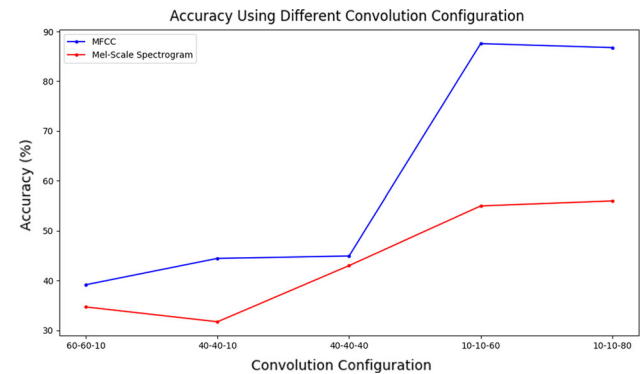
On CK+ dataset, SNN achieves an average accuracy of 97.7%, which outperforms CNN by 1% while lower than the HOG+SVM model by 2%. On JAFFE dataset, SNN achieves an average recognition accuracy of 94.0%, similar to HOG+SVM, and exceeds CNN by 23%. CNN model experiences the lowest performance which is mainly due to the small training size of JAFFE dataset compared to CK+ dataset, that is, not enough to train the network to generate effective feature representations without any data augmentation or pre-processing (Lopes et al. 2017).

### 5.1.2 SER

In order to evaluate SNN model for SER, we have implemented some classic methods for SER classification with SVM and CNN (Swain et al. 2018). Firstly, similar to SNN, we extract MFCC features from audio input using Librosa (McFee et al. 2015), with a total number of feature of 40 and temporal feature length of 388. MFCCs are used as an input for a simple SVM classifier. Many kernels have been experimented, such as linear, polynomial, or radial basis function (RBF). The linear kernel has been retained as it produces the best overall accuracy. CNN represents an effective way of extracting features for SER (Kim et al. 2017). Here,



**Fig. 6** Comparison of overall accuracy for SER tasks on 2 datasets between SNN, CNN, and SVM



**Fig. 7** Effect of convolution window configuration on overall accuracy

we choose a baseline CNN architecture that consists of three sets of convolution layers, each followed by a max-pooling and batch normalisation. It also consist of a fully connected layer at the end (Badshah et al. 2017).

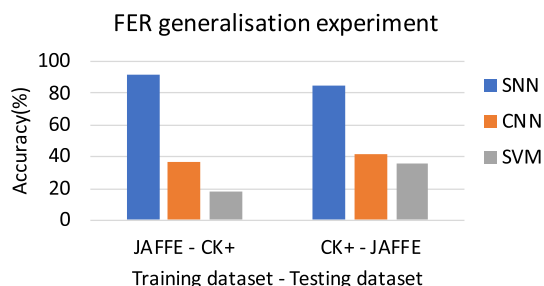
Figure 6 presents the overall accuracy for SER tasks on 2 datasets. On eINTERFACE, SNN outperforms CNN by 8.7% and SVM by 18.9%, and on RAVDESS, SNN outperforms CNN by 3.7% and SVM by 19.8%.

The accuracy of SNN for SER can be enhanced by choosing different parameters for number of features, window, and stride size for the convolution window. Results in Fig. 7 show that the overall accuracy increases when the convolutional size is smaller, and the number of features is higher.

Increasing the number of features leads to an increase in the number of excitatory neurons, leading to higher accuracy. The pattern is observed using both MFCCs and Mel-scale spectrogram features (Diehl and Cook 2015). Having more features and more excitatory neuron leads to learning more features. However, having more excitatory neurons is more computationally costly.

## 5.2 Computational cost

SNN is computationally more expensive compared to ANN. For example, the training time is about 6.9 h for FER on CK+ dataset and 1.05 h for SER on RAVDESS dataset. The size of



**Fig. 8** Comparison of FER accuracy on SNN, HOG+SVM, and CNN with models on cross-dataset

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	91.7	8.3	0	0	0	0
Disgusted	4.3	91.3	0	0	0	4.3
Fearful	4	0	80	0	0	16
Happy	4.2	0	4.2	91.7	0	0
Sad	0	8.3	12.5	4.2	70.8	4.2
Surprised	0	0	0	0	0	100

**Fig. 9** Confusion matrix of SNN when trained on CK+ and tested on JAFFE (accuracy in %)

the SNN models for both FER and SER is quite similar. The number of excitatory and inhibitory neurons is set as 5760 for SER and 6000 for FER. SNN computational cost can vary between different types of SNN simulator implementations. The simulator used in this paper is BRIAN 1.4 (Goodman and Brette 2008) which is a CPU-based implementation. Other implementation such as BINDSnet (Hazan et al. 2018) is GPU based and has a lower computational costs. A more comprehensive comparison on computation cost between SNN and ANN can be found (Deng et al. 2020).

### 5.3 Cross-dataset generalisation experiments

We have performed cross-dataset generalisation experiments by training models on one dataset and testing them using a different dataset with different distributions of data.

#### 5.3.1 FER

Figure 8 presents the accuracy of SNN, HOG+SVM, and CNN on generalisation capacity with cross-dataset validation. In both cases, SNN has achieved consistently high accuracy: 85%—trained on CK+ and tested on JAFFE, and 92%—trained on JAFFE and tested on CK+, which significantly exceed the HOG+SVM and CNN techniques.

Figures 9, 10, and 11 present the confusion matrices of SNN, CNN and HOG+SVM on cross-dataset validation. SNN has the best performance in all classes compared to CNN and HOG+SVM. The highest class accuracy for both

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	4.2	29.2	12.5	9.3	45.8	0
Disgusted	0	47.8	17.4	21.7	8.7	4.3
Fearful	0	0	48	12	20	20
Happy	0	0	41.7	54.2	0	4.2
Sad	4.2	8.3	37.5	16.7	33.3	0
Surprised	0	0	20.8	4.2	0	75

**Fig. 10** Confusion matrix of CNN when trained on CK+ and tested on JAFFE (accuracy in %)

	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Angry	1	44	44	11	0	0
Disgusted	0	12	88	0	0	0
Fearful	0	0	67	33	0	0
Happy	0	0	89	11	0	0
Sad	0	33	44	22		0
Surprised	0	0	100	0	0	

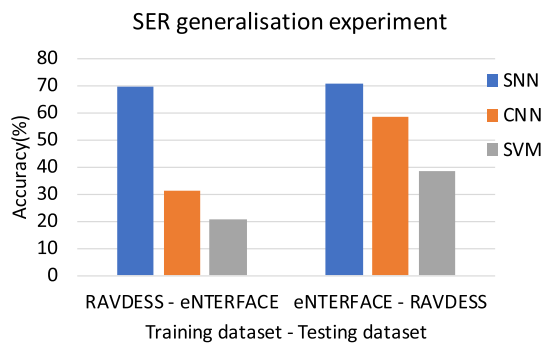
**Fig. 11** Confusion matrix of HOG+SVM when trained on CK+ and tested on JAFFE (accuracy in %)

methods is ‘surprise’, where SNN achieved 100% and CNN 75%, whereas the highest class accuracy using HOG features is ‘fearful’, and all classes are mainly classified as ‘fearful’.

The under-performance of CNN and SVM might be due to the following reasons. The supervised learning used in both CNN and SVM expects training and testing data to have the same distribution and is more biased by the dataset used for training. They also work better with larger datasets. Using JAFFE dataset with only ten subjects has a negative impact on the accuracy for CNN and SVM, due to limited variation in faces, facial expressions, and cultural differences. JAFFE dataset has exclusively Japanese females subjects, whereas the CK+ dataset includes more diverse subjects. Similar findings have also been reported in Lopes et al. (2017). SNN accuracy seems not affected much by this issue. The combination of applying LoG filter, unsupervised learning, and convolutional layer enables the model to generalise well without expecting the same distribution of the data, and the accuracy is dependent on the number of features/patches chosen. LoG filters help define contours and highlight key facial features.

#### 5.3.2 SER

Figure 12 compares the accuracy of SNN, CNN, and SVM on the generalisation experiment. Models trained with eNTERFACE exhibit the same patterns for results as the models trained with RAVDESS, with SNN outperforming SVM and CNN baselines for generalisation using RAVDESS as a test



**Fig. 12** Comparison of SER accuracy on SNN, HOG+SVM, and CNN with models on cross-dataset

dataset. That is, SNN achieves an overall accuracy of 70.8% compared to 68.5% and 58.3% for SVM and CNN, respectively. Exploiting the unsupervised learning using SNN and the feature learning using convolution layers, we obtain a more robust model that can learn features.

### 5.4 Robustness experiments

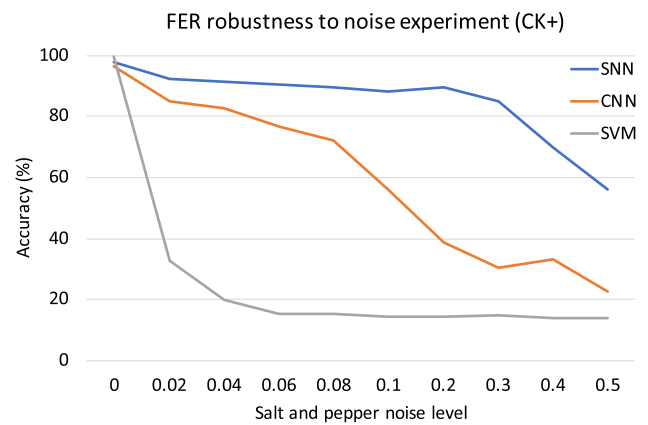
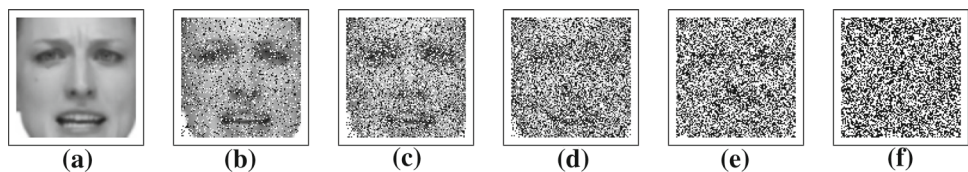
Here, we have experimented the sensitivity and robustness of SNN to noise for both audio and visual data, compared to the state-of-the-art models.

#### 5.4.1 FER

Various types of noise have been used in the literature to assess the sensitivity of models for image recognition tasks. There exist various ways of assessing model robustness to image degradation such as colours changing, noise such as salt and pepper or Gaussian noise (Karahan et al. 2016). Noise degradation is also used to assess the sensitivity of different CNN models (ALexNet, VGG, and GoogleNet) (Karahan et al. 2016).

We have experimented with different intensity parameters of salt and pepper noise degradation ranging from 0 to .5. Salt and pepper noise represent intensity and sparse disturbances to an image where original pixels are randomly replaced with black and white pixels. After .5 noise intensity, we have noticed that the image is completely covered, thus not any more useful to get insight of the performance. Figure 13 shows the samples of salt and pepper noise degradation of input images.

**Fig. 13** a Image no noise b .1 noise probability c .2 noise probability d .3 noise probability e .4 noise probability f .5 noise probability



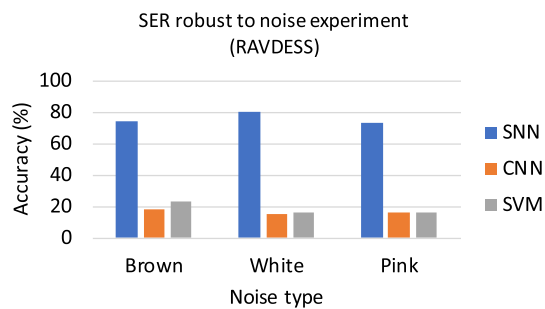
**Fig. 14** Models accuracy with different noise degradation intensities

Results for FER noise degradation tasks are summarised in Fig. 14. The initial results of the three models where no noise is applied are quite close. SVM model experiences the highest accuracy with 99.6% overall, followed by CNN and SNN with 97.6% and 97.4%, respectively.

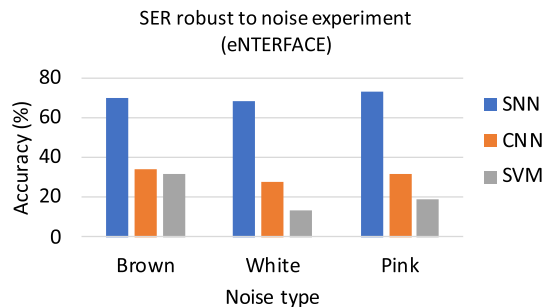
Starting from the lowest probability of noise degradation of 0.2%, we notice a drop in the overall accuracy for all three models. The drop for the SNN model down to 92.4% is not as significant as the drop in CNN to 85% or the significant drop for the SVM model to 32.6%. The higher noise intensity results in a lower overall accuracy for all three models. However, SNN performs best for all noise intensities. The lowest accuracy for SNN is using the 0.5 probability distribution of noise with only 56.2%. However, the lowest accuracy for CNN and SVM is significantly weaker: 22.6% and 14.1%, respectively. SVM is the most affected by the artificial noise degradation. The drop in accuracy pattern in all models does follow the results obtained in Roy et al. (2018) and Karahan et al. (2016), where the increase in noise affects feature identification. Although accuracy has dropped for SNN, it still maintains an accuracy over 65% up to the noise intensity of 0.4., whereas we notice a quicker drop on CNN and SVM starting from intensity .1 and .02, respectively. Figure 14 presents the trend of accuracy decreasing with the increase in noise ratio.

#### 5.4.2 SER

The results of SER noise degradation experiments on RAVDESS and eINTERFACE are summarised in Figs. 15



**Fig. 15** Comparison of SER accuracy for noise degradation tasks for RAVDESS



**Fig. 16** Comparison of SER accuracy for noise degradation tasks for eINTERFACE

and 16. The results on RAVDESS show an overall accuracy of 85% for SNN model, 76.3% for CNN, and 60.5% for SVM. Applying noise leads to a significant drop on CNN and SVM accuracy. However, a much less significant drop is noticed in SNN with the lowest accuracy experienced with pink noise at 73.3%. However, the accuracy of CNN drops significantly to lower than 20%.

Similar to the results on RAVDESS, we have observed a degradation in accuracy for all audio noise with SNN performing best for the three audio noise effects on eINTERFACE. Noise affects the overall accuracy of all tested models. However, the less affected model for both tasks is SNN, as with unsupervised learning, it can overcome various degrees of noise degradation for both images and audio inputs. Results are consistently in line with the generalisation tasks results, where SNN is better at learning intrinsic, more robust features.

## 6 Conclusions and future work

This paper investigates the application of bio-inspired spiking neural network with unsupervised learning for speech and facial emotion recognition tasks. It assesses the robustness of these networks in noise degradation and investigates generalisation capacity. We have set up various experiments for both SER and FER tasks by training the model with one dataset and testing it with a different one. SNN has

achieved consistently better accuracy compared to the state-of-the-art techniques such as SVM with HOG features and CNN networks. The evaluation results show that the SNN has better generalisation capability and more robust to noise degradation on different noise densities, compared to the state-of-the-art techniques SVM and CNN. We will extend this work in application for continuous multi-modal emotion recognition, by applying SNN to multi-sensory integration. In addition, we will further investigate the robustness of the architecture, for example, removing some of the synapses or deleting some of the neurons, and assess the impact of these behaviours to the accuracy of emotion recognition.

**Acknowledgements** We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro M5000 GPU used for this research.

## Compliance with ethical standards

**Conflict of interest** All authors declare that they have no conflict of interests.

**Ethical approval** This article does not contain any studies with human or animals performed by any of the authors.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aghdam HH, Heravi EJ, Puig D (2016) Analyzing the stability of convolutional neural networks against image degradation. In 'VISIGRAPP (4: VISAPP)'. pp 370–382
- Akasay MB, Ouz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun* 116:56–76
- Al-Yasari MMR, Al-Jamali NAS (2018) Modified training algorithm for spiking neural network and its application in wireless sensor network. *Energy* 5(10)
- Anagnostopoulos C-N, Iliou T, Giannoukos I (2015) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artif Intell Rev* 43(2):155–177. <https://doi.org/10.1007/s10462-012-9368-5>
- Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 international conference on platform technology and service (PlatCon). IEEE, pp 1–5

- Bohte SM, La Poutré H, Kok JN (2002) Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks. *IEEE Trans Neural Netw* 13(2):426–435
- Bradski G (2000) The opencv library. *Dr Dobb's J Softw Tools* 25:120–125
- Buscicchio CA, Górecki P, Caponetti L (2006) Speech emotion recognition using spiking neural networks. In: Esposito F, Raś ZW, Malerba D, Semeraro G (eds) *Foundations of intelligent systems*. Springer, Berlin, Heidelberg, pp 38–46
- Carcagni P, Coco MD, Leo M, Distanto C (2015) Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus* 4(1):645
- Chavhan Y, Dhore ML, Yesaware P (2010) Speech emotion recognition using support vector machine. *Int J Comput Appl* 1:8–11
- Chollet F et al (2015) Keras. <https://keras.io>
- Cummins N, Amiriparian S, Hagerer G, Batliner A, Steidl S, Schuller BW (2017) An image-based deep spectrum feature representation for the recognition of emotional speech. In: *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, pp 478–484
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol 1, pp 886–893
- Demin V, Nekhaev D (2018) Recurrent spiking neural network learning based on a competitive maximization of neuronal activity. *Front Neuroinform* 12:79
- Deng L, Wu Y, Hu X, Liang L, Ding Y, Li G, Zhao G, Li P, Xie Y (2020) Rethinking the performance comparison between SNNs and ANNs. *Neural Netw* 121:294–307
- Diehl P, Cook M (2015) Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front Comput Neurosci* 9:99
- Farries MA, Fairhall AL (2007) Reinforcement learning with modulated spike timing-dependent synaptic plasticity. *J Neurophysiol* 98(6):3648–3665
- Filip P, Andrzej K (2011) Introduction to spiking neural networks: information processing. *Learn Appl* 71:409–33
- Gallicchio C, Micheli A, Pedrelli L (2017) Deep reservoir computing: a critical experimental analysis. *Neurocomput* 268:87–99
- Gavrilov AV, Panchenko KO (2016) Methods of learning for spiking neural networks. a survey. In: *The 13th international scientific-technical conference on actual problems of electronics instrument engineering (APEIE)*, vol 2. IEEE, pp 455–460
- Gilani SZ, Mian A, Shafait F, Reid, I (2017) Dense 3d face correspondence. In: *IEEE transactions on pattern analysis and machine intelligence*, pp 1584–1598
- Goodman D, Brette R (2008) Brian: a simulator for spiking neural networks in python. *Front Neuroinform* 2:5
- Grüning A, Bohte S (2014) Spiking neural networks: principles and challenges. In: *Proceedings of the 22nd European symposium on artificial neural networks*. Computational intelligence and machine learning-ESANN
- Gupta D, Bansal P, Choudhary K (2018) The state of the art of feature extraction techniques in speech recognition. In: Agrawal SS, Devi A, Wason R, Bansal P (eds) *Speech and language processing for human-machine communications*. Springer, Singapore, pp 195–207
- Harár P, Burget R, Dutta MK (2017) Speech emotion recognition with deep learning. In: *The 4th international conference on signal processing and integrated networks (SPIN)*. IEEE, pp 137–140
- Hazan H, Saunders DJ, Khan H, Patel D, Sanghavi DT, Siegelmann HT, Kozma R (2018) Bindsnet: a machine learning-oriented spiking neural networks library in python. *Front Neuroinform* 12:89
- Hebb DO (1949) *The organization of behavior: a neuropsychological theory*. J Wiley, Chapman & Hall
- Hinton GE, Sejnowski TJ, Poggio TA (1999) *Unsupervised learning: foundations of neural computation*. MIT press, Cambridge
- Hodgkin AL, Huxley AF (1990) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Bull Math Biol* 52:25–71
- Hsu F, Lin W, Tsai T (2013) Automatic facial expression recognition for affective computing based on bag of distances. In: *Proceedings of 2013 Asia-Pacific signal and information processing association annual summit and conference*, pp 1–4
- Jose JT, Amudha J, Sanjay G (2015) A survey on spiking neural networks in image processing. In: El-Alfy E-SM, Thampi SM, Takagi H, Piramuthu S, Hanne T (eds) *Advances in intelligent informatics*. Springer, Cham, pp 107–115
- Karahan S, Kilinc Yildirim M, Kirtac K, Rende FS, Butun G, Ekenel HK (2016) How image degradations affect deep CNN-based face recognition? In: *2016 international conference of the biometrics special interest group (BIOSIG)*, pp 1–5
- Kheradpisheh S, Ganjtabesh M, Thorpe S, Masquelier T (2017) STDP-based spiking deep convolutional neural networks for object recognition. *Neural Netw* 99:56–67
- Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition? In: *CoRR*
- Kim B, Dong S, Roh J, Kim G, Lee S (2016) Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In: *IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. pp 1499–1508
- Kim J, Truong KP, Englebienne G, Evers V (2017) Learning spectro-temporal features with 3d CNNs for speech emotion recognition. In: *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, pp 383–388
- Knudsen EI (1994) Supervised learning in the brain. *J Neurosci* 14(7):3985–3997
- Lech M, Stolar M, Best C, Bolia R (2020) Real-time speech emotion recognition using a pre-trained image classification network: effects of bandwidth reduction and companding. *Front Comput Sci* 2:14
- LeCun Y, Bengio Y, Hinton GE (2015) Deep learning. *Nature* 521(7553):436–444. <https://doi.org/10.1038/nature14539>
- Lee J, Tashev I (2015) High-level feature representation using recurrent neural network for speech emotion recognition. In: *Proceedings of INTERSPEECH 2015*
- Liu Y, Cao Y, Li Y, Liu M, Song R, Wang Y, Xu Z, Ma X (2016) Facial expression recognition with PCA and LBP features extracting from active facial patches. In: *Proceedings of 2016 IEEE international conference on real-time computing and robotics (RCAR)*, pp 368–373
- Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American english. *PLOS ONE* 13(5):1–35
- Lopes AT, de Aguiar E, Souza AFD, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognit* 61:610–628
- Lotfifidreshgi R, Gournay P (2017) Biologically inspired speech emotion recognition. In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp 5135–5139
- Maass W (1997) Networks of spiking neurons: the third generation of neural network models. *Neural Netw* 10:1659–1671
- Majumder A, Behera L, Subramanian VK (2016) Automatic facial expression recognition system using deep network-based data fusion. *IEEE Trans Cybern* 99:1–12
- Mansouri-Bensassi E, Ye J (2018) Bio-inspired spiking neural networks for facial expression recognition: generalisation investigation. In: *International conference on theory and practice of natural computing*. Springer, pp 426–437

- Marr D, Hildreth E (1980) Theory of edge detection. *Proc R Soc Lond Ser B* 23:187–217
- McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O (2015) Librosa: audio and music signal analysis in python. In: *Proceedings of the 14th python in science conference*
- Mehendale N (2020) Facial emotion recognition using convolutional neural networks (FERC). *SN Appl Sci* 2(3):446
- Mishra B, Fernandes SL, Abhishek K, Alva A, Shetty C, Ajila CV, Shetty D, Rao H, Shetty P (2015) Facial expression recognition using feature based techniques and model based techniques: a survey. In: *2015 2nd international conference on electronics and communication systems (ICECS)*, pp 589–594
- Mollahosseini A, Chan D, Mahoor, MH (2016) Going deeper in facial expression recognition using deep neural networks. In: *2016 IEEE winter conference on applications of computer vision (WACV)*, pp 1–10
- Niu Y, Zou D, Niu Y, He Z, Tan H (2018) Improvement on speech emotion recognition based on deep convolutional neural networks. In: *Proceedings of the 2018 international conference on computing and artificial intelligence, ICCAI 2018*. ACM, New York, pp 13–18
- Papakostas M, Spyrou E, Giannakopoulos T, Siantikos G, Sgouropoulos D, Mylonas P, Makedon F (2017) Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation* 5(2):26
- Pitas I, Kotsia I, Martin O, Macq B (2006) The enterface05 audio-visual emotion database. In: *22nd international conference on data engineering workshops (ICDEW'06) (ICDEW)*, vol 00, p 8
- Roy P, Ghosh S, Bhattacharya S, Pal U (2018) Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*
- Rzayeva Z, Alasgarov E (2019) Facial emotion recognition using convolutional neural networks. In: *2019 IEEE 13th international conference on application of information and communication technologies (AICT)*, pp 1–5
- Satt A, Rozenberg S, Hoory R (2017) Efficient emotion recognition from speech using deep learning on spectrograms. *Proc Interspeech 2017*:1089–1093
- Saunders DJ, Siegelmann HT, Kozma R, Ruzinko M (2018) STDP learning of image patches with convolutional spiking neural networks. In: *2018 international joint conference on neural networks (IJCNN)*, pp 1–7
- Sporea I, Grüning A (2012) Classification of distorted patterns by feed-forward spiking neural networks. In: *International conference on artificial neural networks*. Springer, pp 264–271
- Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: a review. *Int J Speech Technol* 21(1):93–120
- Tapson JC, Cohen GK, Afshar S, Stiefel KM, Buskila Y, Hamilton TJ, van Schaik A (2013) Synthesis of neural networks for spatio-temporal spike pattern recognition and processing. *Front Neurosci* 7:153
- Tashev IJ, Wang Z.-Q, Godin K (2017) Speech emotion recognition based on gaussian mixture models and deep neural networks. In: *2017 information theory and applications workshop (ITA)*, pp 1–4
- Tavanaei A, Ghodrati M, Kheradpisheh SR, Masquelier T, Maida A (2019) Deep learning in spiking neural networks. *Neural Netw* 111:47–63
- Tavanaei A, Maida, AS (2017) Multi-layer unsupervised learning in a spiking convolutional neural network. In: *2017 international joint conference on neural networks (IJCNN)*, pp 2023–2030
- Tavarez D, Sarasola X, Alonso A, Sanchez J, Serrano L, Navas E, Hernández I (2017) Exploring fusion methods and feature space for the classification of paralinguistic information. *Proc Interspeech 2017*:3517–3521
- Tie Y, Guan L (2013) A deformable 3-D facial expression model for dynamic human emotional state recognition. *IEEE Trans Circuits and Syst Video Technol* 23:142–157
- van der Walt S, Schenberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T (2014) Scikit-image: image processing in python. *PeerJ* 2:e453
- Vinola C, Vimaladevi K (2015) A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electron Lett Comput Vision Image Anal* 14(2):24–44
- Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W, Sturge-Apple M (2017) Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. *Int J Speech Technol* 20(1):27–41

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.