UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Leonardo Antonio Pinheiro

Neuroevolutionary Architecture Search for Facial Expression Recognition

Florianópolis
22 de abril de 2022

Leonardo Antonio Pinheiro

Neuroevolutionary Architecture Search for Facial Expression Recognition

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de mestre em Ciência da Computação.
Orientador: Prof. Rafael de Santiago, Dr.

Florianópolis
22 de abril de 2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Pinheiro, Leonardo Antonio
Neuroevolutionary Architecture Search for Facial Expression Recognition / Leonardo Antonio Pinheiro ; orientador, Rafael de Santiago, 2022.
76 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2022.

Inclui referências.

1. Ciência da Computação. 2. Redes Neurais Convolucionais. 3. Algoritmos Genético. 4. Neuroevolução. 5. Expressões Faciais. I. de Santiago, Rafael. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Ciência da Computação. III. Título.

Leonardo Antonio Pinheiro

Neuroevolutionary Architecture Search for Facial Expression Recognition

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Diego Vrague Noble, Dr.
Pontifícia Universidade Católica do Rio Grande do Sul

Prof. Mauro Roisenberg, Dr.
Universidade Federal de Santa Catarina

Prof. Rudimar Luís Scaranto Dazzi, Dr.
Universidade do Vale do Itajaí

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciência da Computação.

GOV.BR
Documento assinado digitalmente
PATRICIA DELLA MEA PLENTZ
Data: 23/06/2022 14:13:58-0300
CPF: 959.243.700-97
Verifique as assinaturas em https://v.ufsc.br

Prof^{a}. Patricia Della Méa Plentz, Dr^{a}
Coordenadora do Programa

ICP-Edu
Documento assinado digitalmente
Rafael de Santiago
Data: 23/06/2022 14:30:01-0300
CPF: 041.277.019-94
Verifique as assinaturas em https://v.ufsc.br

Prof. Rafael de Santiago, Dr.
Orientador

Florianópolis, 22 de abril de 2022.

ACKNOWLEDGEMENTS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

You will never change your life until you change something you do daily
John C. Maxwell

RESUMO

O reconhecimento de expressão facial tem sido uma área ativa com aplicações para resolução de tarefas em diversos campos como medicina, psicologia, pedagogia e também para interação humano-computador. Para resolver esta tarefa, redes neurais convolucionais com técnicas de visão computacional para pré-processamento são as abordagens mais comuns. Além disso, nos últimos anos, a pesquisa reconhecimento de expressão facial passou de condições controladas em laboratório para condições desafiadoras na natureza. Recentemente, estudos utilizando algoritmos genéticos para atribuir automaticamente os melhores parâmetros à redes neurais convolucionais obtiveram melhores resultados do que aqueles feitos manualmente em tarefas de classificação de imagens utilizando as bases de dados CIFAR-10 e CIFAR-100. A abordagem utiliza parâmetros aleatórios para gerar arquiteturas redes neurais convolucionais, e alcança uma arquitetura competitiva em relação ao estado da arte em reconhecimento de expressões faciais. Nossa abordagem usa três dos principais conjuntos de dados para reconhecimento de expressões faciais, a saber: JAFFe, CK+ e AffectNet. O uso de um conjunto de dados não laboratorial como o AffectNet é relevante porque os ruídos da imagem se assemelham ao ambiente real. A metaheurística proposta nesta dissertação de mestrado superou os resultados do estado da arte para o conjunto de dados AffectNet.

Palavras-chave: Redes Neurais Convolucionais. Algoritmos Genético. Neuroevolução. Aprendizagem Profunda Evolucionária. Reconhecimento de Expressão Facial. Reconhecimento de Emoções.

RESUMO ESTENDIDO

Introdução

O reconhecimento de expressões faciais (FER) é utilizando em várias tarefas em diversas áreas como clínica, comercial, neurológica, pedagógica, psiquiátrica, psicológica e outras. As primeiras pesquisas em FER foram publicadas no ano de 1978. Devido os algoritmos de detecção facial demandarem uma demanda computacional, o progresso da área se desenvolvia lentamente. Pesquisadores abordam com aprendizagem profunda para lidar com FER, utilizando por exemplo, Redes Neurais Convolucionais (CNN), Redes Híbridas e *Long Short-Term Memory*. O potencial da aprendizagem profunda é automaticamente aprender os padrões mais discriminativos da expressão facial. Mas apesar disso o pre-processamento de imagens influenciam nessa performance. Neuro Evolução (NE) é uma forma de lidar com aprendizado de máquina que utiliza algorítimos evolucionários para designar os parâmetros, topologia e estrutura da rede neral. Considerando os resultados presentes das CNNs que são apresentados em pesquisas recentes, nós realizamos uma revisão sistemática de todos os estudos que utilizam NE para resolver o problema de reconhecimento expressão facial via imagens. O reconhecimento de expressões faciais são performados em três base de dados principais: JAFFE, CK e CK+. As três base de dados são bidimensionais, feitas em um ambiente controlado e são classificadas em 6 emoções básicas. Nessa dissertação foram utilizadas três base de dados a JAFFE, CK+ e AffectNet. Nessas bases foram utilizadas as 6 emoções básicas listadas por Ekman: felicidade, tristeza, surpresa, medo, raiva e desprezo combinado com desgosto. Finalmente, essa dissertação propõe uma estratégia evolucionária como algorítimo genético para resolver a tarefa de reconhecimento de expressões faciais. Sendo assim, é reportado os resultados e análises dessa estratégia, como também demonstrado que essa mesma estratégia pode ser utilizada para outros problemas de classificação.

Objetivos

Para o melhor de nosso conhecimento, pesquisando na literatura científica, identificamos que a abordagem NE não foi usada para resolver o problema de FER em condições não laboratoriais. A condição não laboratorial significa que o conjunto de imagens não é controlado por laboratório. O rosto, então, está mais próximo de situações da vida real, com diferentes ângulos e podendo sofrer algum tipo de oclusão. Portanto o objetivo principal dessa dissertação é uma abordagem Neuro Evolutiva para reconhecimento de expressões faciais em condições não laboratoriais. Para alcançar esse objetivo foi identificado o estado da arte na área de FER, foi desenvolvido uma revisão sistemática da literatura para estratégias evolutivas para FER, foi desenvolvido método Neuro Evolutivo utilizando dois dos principais conjuntos de dados e a maior base de dados não laboratorial. Por fim, foi avaliada a técnica Neuro Evolutiva.

Metodologia

Foi feita uma pesquisa bibliográfica sobre reconhecimento de emoções e expressões faciais, visão computacional e neuro evolução com intuito de identificar características necessárias para implementar a solução, fortalecer a importância do projeto e adquirir mais conhecimento sobre o assunto. Para isso foi seguido os seguintes procedimentos: (I) pesquisar o estado da arte das técnicas de reconhecimento facial, identificando pontos fortes e fracos, (II) avaliar os métodos mais utilizados na área de neuro evolução, (III) especificar os métodos e técnicas de algoritmos genéticos, (IV) desenvolver uma estratégia neuro evolutiva, (V) Conduzir experimentos comparando o método feito com abordagens existentes, (VI) analisar os resultados utilizando as principais bases de reconhecimento facial, (VII) analisar os resultados obtidos, (VIII) divulgar os resultados com a área de pesquisa.

Resultados e Discussão

Para os experimentos utilizando os conjuntos de dados de FER, primeiro utilizamos redes clássicas como VGGNet e ResNet e então a meta heurística proposta por essa dissertação. Para o preprocessamento das imagens. Primeiro, nós utilizamos o algoritmo Viola-Jones para fazer o corte da área de interessa, ou seja, o rosto das pessoas. Então, foi redimensionado as imagens para o tamanho de $64x64$, principalmente porque as imagens da base AffectNet não tem um padrão de tamanho para as imagens e por fim as imagens foram convertidas para padrões de cinza. Para a probabilidade de *crossover* e mutações, foi mantido os valores baseado em convenções, ou seja, 90% e 20% respectivamente. A estrategia para a separação entre dados de validação e treinamento são: (I) JAFFe: Aleatoriamente separado 70% para treinamento e 30% para validação. (II) CK+: Fora selecionado os três últimos quadros das sequencia de imagens e então separado aleatoriamente 70% para treinamento e 30% para validação.(III) AffectNet: Foi selecionado o conjunto de imagens anotado manualmente. E então feita a separação entre treinamento e teste utilizando o conjunto que o autor do conjunto provém. Para os testes utilizando as CNNs clássicas foi utilizado como base de dados a combinação entre JAFFe e CK+. As redes clássicas obtivera uma acurácia de 94% tanto a VGGNet quanto a ResNet. Nesse experimento, as redes foram mantidas na configuração mais tradicional e básica para utilizar esses valores como base comparativa. Para os testes com bases laboratoriais, a GA rapidamente converge a resultados superiores aos das redes clássicas. Em média, a acurácia obtida para JAFFe foi de 97,6% e 96,4% para a base CK+. Na matriz de confusão para o JAFFe, em que vemos apenas um erro de previsão para a emoção raiva. Este erro de previsão pode ser devido a muito músculo contração que essa emoção exerce. A emoção da tristeza é facilmente reconhecida pelos humanos ao notar microexpressões nos lábios, enquanto a raiva pode ocorrer contração muscular ao redor os lábios, mas tem sua diferenciação na contração muscular ao redor dos olhos. Foi realizado dois experimentos usando o conjunto de dados AffectNet. A primeira segue todas as instruções descritas no início da seção e usa 50 gerações máximas para cada execução. No segundo experimento, 4.000 imagens de cada emoção e 250 imagens foram separados aleatoriamente para treinamento e validação, então um quarto conjunto de dados foi criado, que é o junção dos três conjuntos de dados apresentados neste trabalho. Ao contrário dos experimentos feitos com JAFFe e CK+, a população não converge para um valor ótimo. Apesar disso, houve uma convergência para o valor ótimo. Na geração 50, o melhor indivíduo tem uma precisão de 73,3%. Como nos experimentos para os conjuntos de dados de laboratório, o GA foi executado um total de 5 vezes. Para o AffectNet, obteve-se uma precisão de 68,8% em média. Para o segundo experimento usando AffectNet, o AG executou apenas duas vezes. No entanto, 80 gerações máximas foram definidas para essas execuções. Em ambas as execuções, a solução converge rapidamente para valores acima de 80% e se mantém por gerações antes de alcançar uma pequena melhoria. Ao final desta execução, a precisão obtido para o melhor indivíduo foi de 87%. Para avaliar os resultados dos experimentos, 6 trabalhos de estado da arte que utilizam o mesma quantidade de emoções foram separados, sendo 2 trabalhos para cada um dos conjuntos de dados. Conforme demonstrado na comparação, a metaheurística apresentada nesta dissertação atinge o estado-de última geração para reconhecimento de expressão facial. O que demonstra a eficácia de um simples processo evolucionário. O algoritmo genético encontrou arquiteturas CNN que superaram o estado da arte curadoria para JAFFe e AffectNet. a GA encontrou uma arquitetura com desempenho de 98,8% para CK+, mas em média as execuções tiveram um desempenho de 96,4%. As arquiteturas CNN encontradas não eram muito profundas. Os melhores resultados variaram entre arquiteturas com profundidade entre 5 e 8 camadas.

E um máximo de 3 camadas na rede densa.

Considerações Finais

Nesta dissertação, investigamos o desempenho do algoritmo genético como um algoritmo de busca de arquitetura neural na área de Reconhecimento de Expressões Faciais. Embora trabalhos anteriores usando GA para encontrar a melhor arquitetura CNN decidimos remover a camada totalmente codificada. A metaheurística deste artigo a codifica, então a profundidade, parâmetros ou remoção desta camada é decidida pelo processo evolutivo. Em nossa pesquisa bibliográfica no campo da psicologia, levantamos dois pontos importantes. A primeira é que existe um consenso entre pelo menos 6 das emoções básicas, que são: felicidade, surpresa, medo, tristeza, raiva e desgosto. O segundo ponto é a técnica facial, FACS, onde cada emoção pode ser realizada de forma mais técnica. Estudos e desafios em computação que utilizam FACS estão principalmente relacionados à segmentação semântica, por exemplo, alguns estudos tratam FACS demonstrando a ocorrência de AUs e/ou sua intensidade. Continuamente, apresentamos algumas das principais bases de dados. As bases de dados que as imagens são fotografados em laboratórios têm uma quantidade muito menor quando comparados aos que não são. E todos eles apresentam as seis emoções básicas mencionadas a cima. Os bancos de dados mais usados são CK e JAFFe. Durante a análise de dados com o conjunto de dados Affectnet, notamos que no conjunto de dados é formado por várias imagens duplicadas e muitas delas foram classificadas incorretamente. Como pode haver divergências entre as opiniões dos pesquisadores ao classificar a emoção, precisa-se de um conjunto de dados em ambiente real produzido por especialistas usando técnicas como FACS para que exista menos divergência ou nenhuma. Esse problema persiste em outros conjuntos de dados em ambiente não controlado, como FER2013. Também foi possível observar a ausência de abordagens em ambiente real que utilizam grandes bancos de dados. Detectamos que ainda não existem abordagens para reconhecimento de expressões faciais que aplicam uma estratégia evolutiva para a criação ou alteração da rede neural artificial, incluindo adaptação de topologia. Na revisão sistemática da literatura dois estudos são os únicos que utilizam redes neurais convolucionais, e um é uma Rede Neural Inspirada no Cérebro e os demais utilizam redes neurais perceptron multicamadas, mostrando que ainda há muito espaço para pesquisas utilizando modelos de redes neurais para classificação de imagens. É importante notar que estudos usando estratégias evolutivas na extração de características etapa obtiveram grande sucesso na precisão da classificação. Retornando a pergunta da pesquisa: "Uma abordagem neuro evolucionária alcança melhor resultados para o reconhecimento de expressões faciais por meio de imagens quando comparado a outros métodos que usam redes neurais?". Com base em nossos resultados aplicando uma estratégia neuro evolucionária, a resposta é afirmativa. Uma abordagem NE alcança resultados semelhantes ou superiores quando comparados para outras Redes Neurais Artificiais. Assim, a hipótese alternativa: "A abordagem NE atinge resultados semelhantes ou superiores quando comparados a outras Redes Neurais Artificiais"é satisfeito por esta dissertação. Portanto, rejeitando a hipótese nula.

Palavras-chave: Redes Neurais Convolucionais. Algoritmos Genético. Neuroevolução. Aprendizagem Profunda Evolucionária. Reconhecimento de Expressão Facial. Reconhecimento de Emoções.

ABSTRACT

abstract. Facial Expression Recognition (FER) has been an active field with applications to solve tasks in several areas such as medicine, psychology, pedagogy and also for human-computer interaction. To solve this task, convolutional neural networks (CNN) with computer vision techniques for pre-processing are the most common approaches. In addition, in the last years, FER research has transitioned from laboratory-controlled to challenging in-the-wild conditions. Recently, studies using genetic algorithms (GA) to automatically assign the best parameters to CNN have obtained better results than those manually-designed in image classification tasks using the CIFAR-10 and CIFAR-100 databases. Using random parameters to generate CNN architectures, applying a competitive architecture compared to state of the art in facial expression recognition. Our approach uses three of the main datasets for facial expression recognition, namely: JAFFe, CK+, and AffectNet. The use of a non-laboratory dataset such as AffectNet is relevant because the image noises resemble the real environment. The metaheuristic proposed in this master's thesis surpassed the state-of-the-art results for the AffectNet dataset.

Keywords: Convolutional Neural Networks. Genetic Algorithms. Neuroevolution. Evolutionary Deep Learning. Facial Expression Recognition. Emotion Recognition.

LIST OF FIGURES

LIST OF TABLES

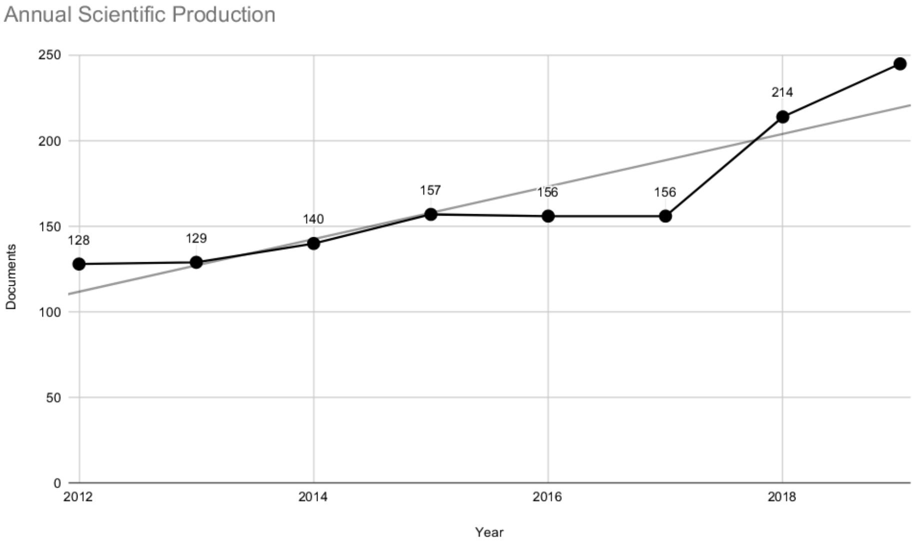
LIST OF ALGORITHMS

CONTENTS

1 INTRODUCTION

Facial Expression Recognition (FER) is used for several tasks in diverse areas such as clinic, e-commerce (LIANG et al., 2015; CORTIS et al., 2017; CHANG; WANG, 2018), neurology, pedagogy, psychiatry, psychology, pain assessment, and others (DAS, 2015; FARSI; MUNRO; AL-THOBAITI, 2015; YANG et al., 2018). In addition, the scientific production in the area of computer science has seen a positive increase in the last years. This growing demand can be seen in Figure 1

Figure 1 – Annual Scientific Production



The first research in FER was published by SUWA (1978). Due to the face detection algorithms that demand computational performance inaccessible at that time, the progress of the area developed at a slowly pace (YANG et al., 2020). Until today, the training databases still not large enough becoming it a challenge to FER methods, so the classification cannot be given precisely (JAIN et al., 2018).

Two methods are used for automatic FER: messages-based and movement of facial components based. The message-based method can be divided into two strategies, namely, discrete categorical and continuous dimensional (ZHANG et al., 2018).

In the discrete categorical strategy, facial expression is classified into one of the predefined states of emotion, including the six basic emotions: anger, aversion, fear, happiness, sadness, and surprise (EKMAN, 1992a), and non-basic emotions like depression and anguish. It is a convenient way to classify emotions by facial expressions; however, it doesn't contemplate the evaluation of more complex or mixed emotions (ZHANG et al., 2018).

The continuous dimensional method comes from the psychology field, which demonstrates emotions in a multidimensional way, using continuous axes in a multidimensional space and represents each expression as a point or a region in the space (RUSSELL, 1980). Spaces composed of two-dimensional and three-dimensional representations are the most commonly used (ZHANG et al., 2018).

The method based on the movement of facial components uses the movements of facial muscles to categorize a facial expression. One of the main examples of this method is the Emotional Facial Action Coding System (EMFACS) which is a derivation of FACS that uses FAC (Facial Action Coding) (EKMAN; FRIESEN, 1976). This method defines Actions Units(AU) that represents when a muscle contract or relax.

Another category of the method based on the movement of facial components, is focused on the recognition of Micro-Expressions (ME). This task is complex even for humans (ZHANG et al., 2018). To recognize these MEs, which last from $1/25$ to $1/3$ of a second, several tools have been created such as the recently highlighted ME analysis system (LI et al., 2018).

Several researchers use approaches with deep learning to address automatic FER, such as Convolutional Neural Networks (CNN) (HE; ZHANG, 2018; PITALOKA et al., 2017), hybrid networks (JAIN et al., 2018) and Long Short-Term Memory (LSTM) (Rodriguez et al., 2017). Deep neural networks are very popular due to their ability to reach the state-of-the-art (ZAFEIRIOU; ZHANG; ZHANG, 2015). The potential of deep learning techniques is to automatically learn the most discriminating patterns of facial expression (ZHANG et al., 2018). Despite this, the pre-processing of the images influences the performance of the deep neural networks applied in the context (PITALOKA et al., 2017).

The pre-processing is required in several strategies that use neural networks, such as NeuroEvolution. NeuroEvolution(NE) is a a form to deal with machine learning that uses evolutionary algorithms to designate parameters, topology and structure of a neural network. There are several methods of NE such as:

- NeuroEvolution of Augmenting Topologies (NEAT) (STANLEY; MIIKKULAINEN, 2002);
- Hypercube-based NeuroEvolution of Augmenting Topologies (HyperNEAT) (STANLEY; D'AMBROSIO; GAUCI, 2009);
- Evolvable Substrate Hypercube-based NeuroEvolution of Augmenting Topologies (ES-HyperNEAT) (RISI; STANLEY, 2012);
- Gradient evolution (GE) (KUO; ZULVIA, 2015);
- Evolutionary Acquisition of Neural Topologies (EANT) (KASSAHUN; SOMMER, 2005).

Traditionally neuroevolution is not used in Deep Learning architectures and is only used in shallow neural networks (SUGANUMA; SHIRAKAWA; NAGAO, 2017), although algorithms developed in the neuroevolution community can be tested with deep neural networks. Despite the high computational cost, this approach proved to be efficient in Deep Learning, *e.g.*, CNNs (SUCH et al., 2017).

Considering the current results of CNNs that are presented in surveys (ZHANG et al., 2018; LI; DENG, 2020; REVINA; EMMANUEL, 2018), we did a systematic review of all studies that use NE to solve facial expression classification problems in images, to evaluate the viability of this method.

About using Genetic Algorithm (GA) as an evolutionary strategy, to assess the fitness of each individual in the population, it is necessary to train the CNN that the individual represents so that it is possible to obtain the accuracy of this architecture, for that reason it is computationally expensive. For example, other works using GA to find the solution architecture, 17 (XIE; YUILLE, 2017) and 35 (SUN et al., 2020) GPU-days were processed in the CIFAR10 (KRIZHEVSKY et al., 2009) database. GPU-days means that the GA processed one day in a GPU when the GA finished processing, *e.g.*, if two GPUs are used and the GA processed for 10 days, the GA used 20 GPU-days.

FER usually is performed on three main databases to recognize emotions that are: The JApanese Female Facial Expression (JAFFE) (LYONS et al., 1998); The Cohn-Kanade (CK) (KANADE; COHN; TIAN, 2000) and its extended variation CK + (LUCEY et al., 2010). JAFFE has 213 images of posed emotions, while CK has 486 sequences of videos and CK + 593 sequences of posed expressions and 122 sequences of spontaneous smiles, all of which are encoded for FACS. The three databases are two-dimensional, made in a controlled environment and are classified by the six basic emotions. We also did some tests using the GA proposed by this work on the CIFAR10(KRIZHEVSKY et al., 2009) database, thus demonstrating that the algorithm can be used without changes in different data.

We used three, among the main, databases for FER: JAFFe(LYONS et al., 1998), CK+ (LUCEY et al., 2010) and AffectNet (MOLLAHOSSEINI; HASSANI; MAHOOR, 2017). On theses bases, we use the 6 basic emotions listed by Ekman (EKMAN; FRIESEN; ELLSWORTH, 2013): happiness, surprise, fear, sadness, anger, and disgust in combination with contempt. The combination of fear and surprise is also possible. Using the CK and CK+ datasets separately would be redundant as CK+ is an extension of the CK dataset.

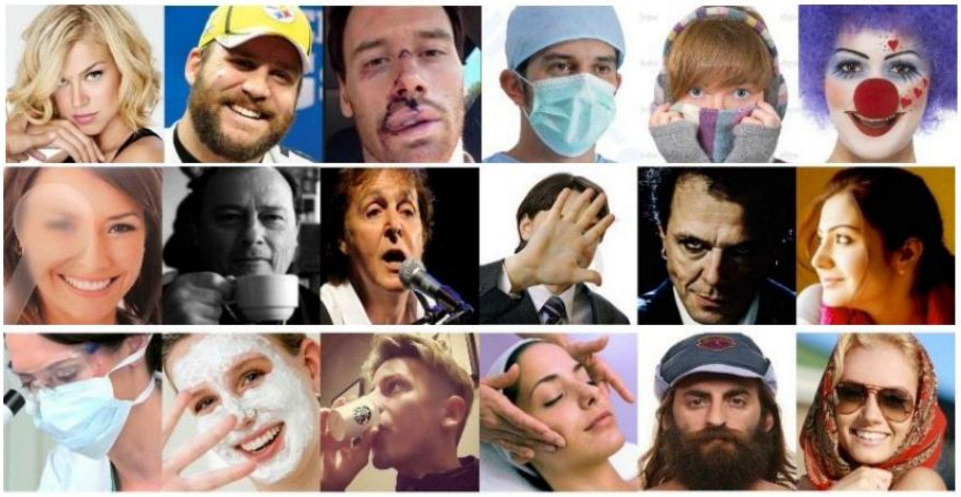
Finally, this research proposes an evolutionary strategy as genetic algorithm to solve the facial expression recognition task. Thus, we report the results and analysis of this strategy, as well as demonstrate that this same strategy can be used for other classification problems.

In the remainder of this chapter we present the research problem and question, the hypothesis, the objectives, the methodology and finally contributions. In the chapter 2, We rise the theoretical ground on facial expressions, Deep Learning and genetic algorithms. In chapter 3, the process of systematic literature review is carried out. In chapter 4 the metaheuristic operation is shown. In chapter 5, the experiments are displayed and the results are analyzed. Finally, chapter 6 concludes and points to future work.

1.1 RESEARCH PROBLEM

For the best of our knowledge, researching in the scientific literature, we have identified that the NE approach was not used to solve the problem of FER with in-the-wild conditions. The in-the-wild condition means that the set of images is not laboratory-controlled. The face is closest to real-life situations, with different angles, and may suffer some type of occlusion. Those conditions can be seen in Figure 2. Therefore, a research question and two hypotheses have been formulated.

Figure 2 – Example of In-The-Wild Conditions (ZHANG et al., 2018)



1.1.1 Research Question

We present the following research question:

Does a neuroevolutionary approach achieve better results for classifying facial expressions when compared to other state-of-the-art methods that use neural networks?

1.1.2 Hypothesis

The research considers the following hypotheses:

- Null: A NE approach achieves inferior results when compared to other Artificial Neural Networks.
- Alternative: A NE approach achieves similar or superior results when compared to other Artificial Neural Networks

1.2 OBJECTIVES

The objectives of this research are divided into main goal and specific objectives. The specific objectives consider results necessary to meet the main goal.

1.2.1 Main Goal

A NeuroEvolution approach for facial expression recognition in-the-wild.

1.2.2 Specific Objectives

To achieve the goal of this work, the following specific objectives was achieve:

- Identify the state-of-the-art in the area of facial expression recognition.
- Identify the best evolutionary strategies for pattern recognition in images.
- Develop a systematic reviews in literature for NE strategies for FER.
- Develop a NE method using facial expression recognition datasets as JAFFe, CK+ and AffectNet.
- Evaluate our NeuroEvolution technique in comparison to other state of art results for facial expression recognition.
- Evaluate our NeuroEvolution technique with in-the-wild condition.

1.3 METHODOLOGY

A bibliographic research (detailed in Chapter 2) was carried out regarding the recognition of emotions, computer vision and neuroevolution in order to identify the characteristics necessary to implement the proposed solution, strengthen the importance of the project and acquire greater knowledge on the subject.

In order for the research question to be answered, the following procedures will be followed:

1. Survey the state-of-the-art for facial recognition techniques, identifying strengths and weaknesses;
2. Evaluate the most used methods in the area of neuroevolution;
3. Specify methods and techniques of genetic algorithms;
4. Develop the neuroevolutionary strategy;

5. Conducting experiments comparing the proposed method with the existing approaches and analyzing the results, using the main databases for facial recognition in training;
6. Analyze the results obtained, generating documentation reporting them;
7. Share the research results to the field.

1.4 CONTRIBUTIONS

The contributions of this work resides in the use of random parameters to generate CNN architectures, allowing the generalization of the search space, *i.e.*, the search space is not influenced by the researcher. Our metaheuristic creates a competitive architecture compared to the state of the art in FER. The contributions of the metaheuristic are summarized as follows: (i) Diverse initial populations, allowing to demonstrate how each population behaves over the generations; (ii) As it has more parameters, it is possible to make small changes to verify which parameters are important in the solution architecture; (iii) A different approach to representing encoded CNNs; (iv) Our results have surpassed the state-of-the-art methods of tuning parameters for the AffectNet dataset; (v) Evaluation of the use of GA as automated machine learning in the area of facial expression recognition.

We also list the contributions of the systematic literature review presented in this thesis as well. The contributions can be divided into 3 lines. The first one is to offer a review addressing evolutionary techniques and specific strategies for Facial Expression Recognition. The second one is to present some gaps for future research, where other researchers can use it as a starting point for new research in the field of FER. In third, we compare these works and discuss the main contributions.

2 BACKGROUND

This chapter describes some fundamental works related to facial expressions and computational methods to recognize them. This requires a theoretical basis for facial expressions and emotions, and recognition methods.

In the section 2.1, we will describe about the concept of facial expression in the field of psychology. In this section, we also present how to recognize the six basic emotions from facial recognition foundations.

The next sections present the computational methods for FER. The section 2.2 we provide an overview of the classic methods for FER. In the section 2.3 we cover deep learning and how it is applied on FER. Finally, in the section 2.4 we conceptualize about NeuroEvolution.

2.1 FACIAL EXPRESSION

In the decade of the 1980s, the answer of psychology was that the face was the main way for understanding emotions and emotions the main one for understanding the face. Making the connection between emotions and facial expressions although it seems like common sense, but it is the most important concept in the psychology of emotions. Charles Darwin is the earliest writer whose work is still an important influence (RUSSELL; DOLS, 1997).

In both men and animals, Darwin (1872) noted that the expressions of emotions follow three principles:

1. The principle of serviceable associated habits.
2. The principle of antithesis.
3. The principle of actions due to the constitution of the nervous system, independently from the first of the will, and independently to a certain extend of habit.

The first principle states that certain complex actions are linked to certain feelings in order to alleviate or gratify any sensation. Due to habit certain feelings can be suppressed by will, so the muscles that are under control are more likely to act causing expressive movements. In other cases, habitual movements that require small movements are equally expressive (DARWIN, 1872). The second principle is explained when a directly opposite feeling is induced, there is an involuntary tendency to perform movements of the opposite nature and these movements, in some cases, are highly expressive. Finally, the third principle is directly linked to an action caused by the nervous system (DARWIN, 1872).

Given that Darwin's views were mainly observatory, much of his study was criticized in the most theoretical and technical areas (RUSSELL; DOLS, 1997). Several psychologists have found convincing evidence that six basic emotions are expressed in the same way in all cultures (CARLSON; HATFIELD, 1992).

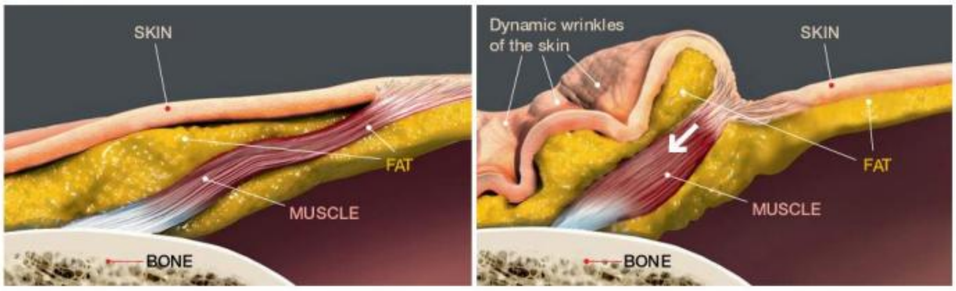
To allow a clearer and more prototypical view of what a facial expression is, Russell & Dols (1997) listed in a more heuristic way the premises and implications, which would be:

1. There are a small number of basic emotions.
2. Each basic emotion is genetically determined, universal and discrete.
3. The encoding and decoding of distinct facial expression constitute a signaling system.
4. Any state lacking its own facial signal is not a basic emotion.
5. All emotions other than basic ones are subcategories or combinations of basic emotions.
6. Voluntary facial expressions can simulate spontaneous one.
7. Any facial expression that deviates from the universal signals is a mixture of basic signals or stems from the operations of a culture display rules
8. Emotional state is revealed by facial measurement.
9. The subjective feelings associated with an emotion are due to proprioceptive feedback from facial movements.
10. Deliberately manipulation of the face into the appropriate configuration creates the neurological pattern of the corresponding emotion.
11. The basic facial expressions are easily recognized by all human beings regardless of their culture.
12. The ability to recognize the emotion of a facial expression is innate.
13. The mental categories by means of witch recognition occurs are genetically rather than culturally determined
14. The meaning of a facial expression is fixed by nature and invariant across changes in the context in which occurs.

Ekman, Friesen & Ellsworth (2013) found that all researchers obtained evidence for six basic emotions, namely: happiness, surprise, fear, sadness, anger, and disgust combined with contempt. They also noticed that in one preliterate culture, fear and surprise are not distinguished from each other (EKMAN, 1992b).

Facial muscles are of utmost importance. These muscles are responsible for nonverbal communication between humans. The muscles influence the appearance of the skin when contracting causing wrinkles and forming different expressions (ZARINS, 2017).

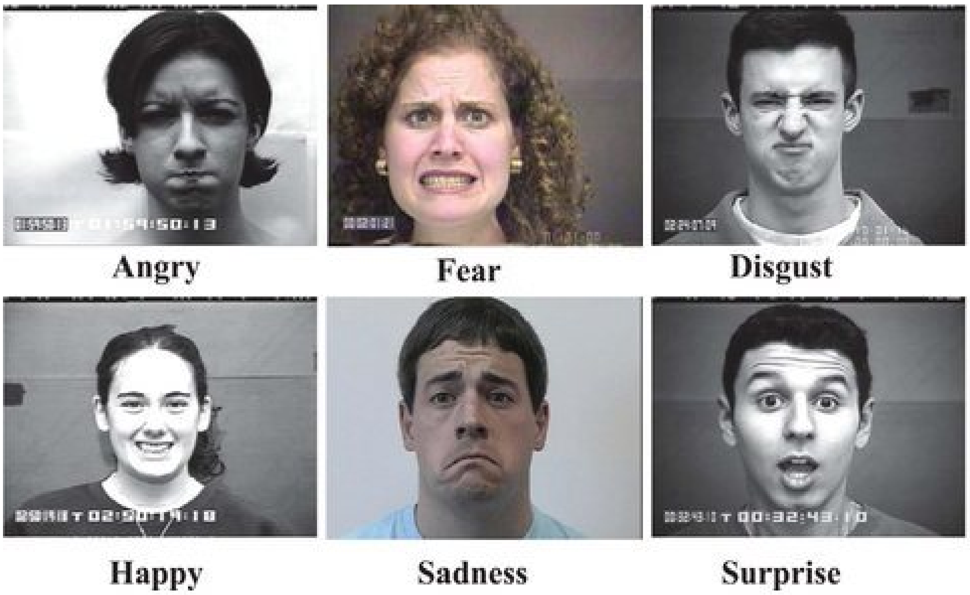
Figure 3 – How the facial muscle moves the skin (ZARINS, 2017, p. 49)



2.1.1 How to recognize basic emotions

The purpose of this subsection is to bring the study of basic emotions from a historical reading by the main researchers. At first, a view of how researchers describe each of the basic emotions by observation (Figure 4).

Figure 4 – Six basic emotions (KANADE; COHN; TIAN, 2000)



To recognize the emotion of happiness, the facial sign present in all variations of that emotion is the smile. Darwin (1872) also states that smile and laughter are the primary expressions of happiness. However, smiles can be confusing because they are also presented when people do not express happiness, e.g. in politeness. The happiness is also possible to notice due to the contraction of the *orbicularis oculi* muscle (EKMAN, 2004).

Surprise is the briefest emotion among all, its duration is a few seconds. Its main characteristics are the raised eyebrows, with the eyes wide open and the jaw drop open (EKMAN, 2004). As much as surprise is an emotion, startle is not, Ekman (2004) in his experiments shot a blank pistol at his research subjects, their reactions were that their eyes closed tightly, their lips stretched tensely and their eyebrows lowered, which differs from the expression of surprise.

The eyes are a crucial feature for surprise and fear, as well as to distinguish them from each other. The fear can be noticed in the lower eyelids. When tense lower eyelids accompany the raised upper eyelids and the rest the face is empty, it is almost always a sign of fear (EKMAN, 2004).

The eyebrows are very important, highly reliable signs of sadness. The inner corners of the eyebrows are pulled up only in the middle and not the entire eyebrow. The lip corners are pulled downward and the upper eyelids droop (EKMAN, 2004).

Under moderate anger, the eyes become bright, the wings of the nostrils are slightly raised, the mouth is usually compressed, and there is almost always a frown on the eyebrow (DARWIN, 1872). In anger the jaw is often thrust forward, the eyelids cover the lower edges of the iris and the lips narrow (EKMAN, 2004).

The most common method of expressing contempt is by movements about the nose, or around the mouth, but the last movements, when strongly articulated, indicate disgust. The nose may be slightly turned upward, together with a turn of the upper lip, or the movement may be a mere wrinkling of the nose (DARWIN, 1872). For Ekman (2004), the nostrils raised while wrinkles appear on the sides of the nose, the raising of the cheeks and lowering of the eyebrows are marks of extreme disgust.

2.1.2 FACS

The human being can perform more than 10,000 facial expressions (EKMAN, 2004). Then in 1978, a facial measurement tool, the Facial Action Coding System (FACS)(EKMAN; FRIESEN, 1978), was published and has been used by both psychologists and computer scientists to describe human facial expressions through muscle movements.

Hjortsjö (1969) was one of the biggest help for the creation of FACS, in his work, he describes the visible appearance changes for each muscle. Hjortsjö photographed his face in different ways and described them. However, he did not describe a set of rules necessary to distinguish facial muscle combinations.

The changes in facial expression described by the FACS use different action units (AU), each of which is anatomically related to the contraction or relaxation of any specific muscle or set of muscles (Pantic; Rothkrantz, 2004).

All AU have a corresponding number, it also has an annotation from A to E, with A being the weakest trace and E being the maximum intensity possible. There are also other modifiers present in FACS to represent which side of the action occurs, such as "R" on the right and "L" on the left (FRIESEN; EKMAN et al., 1983). The main AUs are listed in table 1, from this list, we can correlate the indexes with the facial muscle. We can also use this table to manually label any of the six basic emotions

More than 7,000 AU combinations are observed on a daily basis. When different AUs co-occur in different areas of the face, it is possible to notice changes and view different facial expressions. Therefore, when the AU affects the same facial area they are commonly non-additive (EKMAN; SCHERER, 1982). With FACS it is possible to describe each expression

Table 1 – Main Action Units in the Facial Action Code (EKMAN, 1977; EKMAN; FRIESEN, 1976)

AU	FAC name	*Muscular Basis*
0	Neutral face	
1	Inner Brow Raiser	*Frontalis; Pars Medialis*
2	Outer Brow Raiser	*Frontalis; Pars Lateralis*
4	Brow Lowerer	*Depressor Glabellae; Depressor Supercilli; Corrugator*
5	Upper Lid Raiser	*Levator Palpebrae Superioris*
6	Cheek Raiser	*Orbicularis Oculi; Pars Orbitalis*
7	Lid Tightener	*Orbicularis Oculi; Pars Palpebralis*
8	Lips toward each other	*Orbicularis Oris*
9	Nose Wrinkler	*Levator Labii Superioris; Alaeque Nasi*
10	Upper Lip Raiser	*Levator Labii Superioris; Caput Infraorbitalis*
11	Nasolabial Fold Deepener	*Zygomaticus Minor*
12	Lip Corner Puller	*Zygomaticus Major*
13	Cheek Puffer	*Caninus*
14	Dimpler	*Buccinnator*
15	Lip Corner Depressor	*Triangularis*
16	Lower Lip Depressor	*Depressor Labii*
17	Chin Raiser	*Mentalis*
18	Lip Puckerer	*Incisivii Labii Superioris; Incisive Labii Inferioris*
19	Tongue Show	
20	Lip Stretcher	*Risorius*
21	Neck tightener	*Platysma*
22	Lip Funneler	*Orbicularis Oris*
23	Lip Tightner	*Orbicularis Oris*
24	Lip Pressor	*Orbicularis Oris*
25	Lips Part	*Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris*
26	Jaw Drop	*Masetter; Temporal and Internal Pterygoid Relaxed*
27	Mouth Stretch	*Pterygoids; Digastric*
28	Lip Suck	*Orbicularis Oris*

objectively. Thus, relating the analysis of the AUs with their possible emotion as shown in table 2.

Table 2 – List of AUs Involved In Basic Emotion (MARTINEZ et al., 2017)

Basic Emotion	AU
Happiness	6, 12, 25
Sadness	1, 4, 6, 11, 15, 17
Surprise	1, 2, 5, 26, 27
Fear	1, 2, 4, 5, 20, 25, 26, 27
Anger	4, 5, 7, 10, 17, 22-26
Disgust	9, 10, 16, 17, 25, 26

The basic emotion of happiness is noticed by AU6, AU12 and AU25, but not necessarily by the combination of these three AUs, e.g., happiness is often expressed by the combination of AU12 and AU6 only (MARTINEZ et al., 2017).

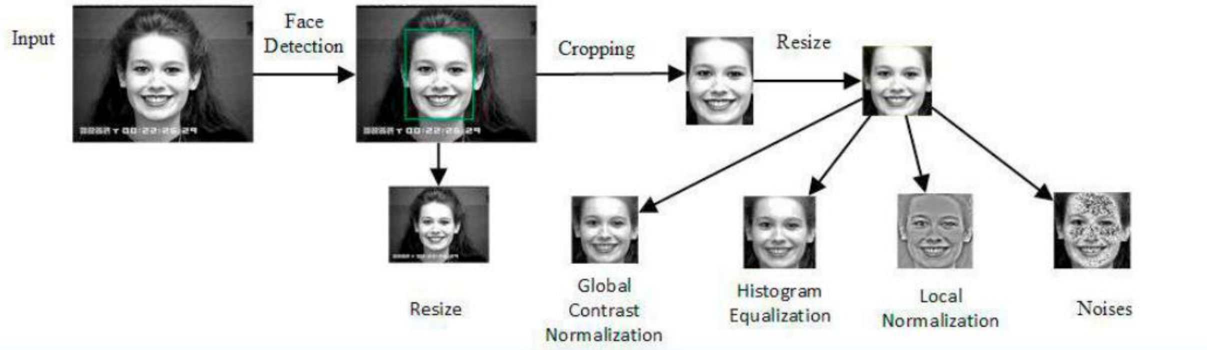
2.2 FACIAL EXPRESSION RECOGNITION

In this section, we present the overall performance of facial expression recognition in the area of computer vision. In addition, we present the main databases used to accomplish this task.

Facial feature detection was one of the first computer vision applications. Since then, research on facial detection and obtaining facial features has made significant progress, providing algorithms capable of detecting faces in the most diverse environments (ZAFEIRIOU; ZHANG; ZHANG, 2015).

The FER system consists of the main stages, such as face image pre-processing, feature extraction and classification (REVINA; EMMANUEL, 2018). Pre-processing methods on FER include face detection and cropping, resize, adding noise, image clarity and normalization. These methods allow better results both in feature extraction and in the classification (PITALOKA et al., 2017).

Figure 5 – Image pre-processing in FER (PITALOKA et al., 2017)



Feature extraction finds and describes features that concern image classification. In computer vision, this extraction is a significant stage for an image to obtain data, where this data will be used as input in the classification step. The feature extraction methods are categorized into five types (REVINA; EMMANUEL, 2018):

- Texture feature-based method;
- Edge feature-based method;
- Geometric feature-based method;
- Global feature-based method;
- Local feature-based method.

In the area of computer vision, the classification of facial expressions has several approaches such as the directed Line segment Hausdorff Distance (dLHD), Euclidean distance metric, Minimum Distance Classifier, KNN (k – Nearest Neighbors) and Support Vector Machine (SVM) (REVINA; EMMANUEL, 2018).

Various approaches are performed on FER by using multiples databases. However, because the annotation process is expensive, very few databases are FACS annotated and fewer are still public (MARTINEZ et al., 2017). In table 3, we list and describe 5 databases widely used: Conh-Kanade (KANADE; COHN; TIAN, 2000), Extended Cohn-Kanade (LUCEY et al., 2010), RAF-DB (LI; DENG; DU, 2017a), JAFFe (LYONS et al., 1998), MMI (Pantic et al., 2005) and AffectNet(MOLLAHOSSEINI; HASSANI; MAHOOR, 2017). In the column "samples", there is the number of images/videos of each database; the column "expression distribution" shows the emotions annotated in each database; and in the column "Condition‘‘ shows if the database is laboratory-controlled or not.

Table 3 – FER Database description

Database	Samples	Emotion	Condition	Year
Cohn-Kanade (CK)	486 image sequences	6 basic emotions	Controlled	2000
Extended Cohn-Kanade (CK+)	593 image sequences	6 basic emotions + contempt and neutral	Controlled	2010
JAFFE	213 images	6 basic emotions + neutral	Controlled	1998
MMI	740 images 2,900 videos	6 basic emotions + neutral	Controlled	2005
RAF-DB	29,672 images	6 basic emotions + neutral and 12 compound emotions	Wild	2017
AffectNet	1,000,000 images	6 basic emotions + contempt and neutral	Wild	2017

The CK, CK + and MMI databases are composed by sequenced images or videos, so they have a problem that the amount of images for training and classification ends up being much smaller since the sequences show from a neutral facial expression to a facial expression with a high intensity. RAF-DB, JAFFe and AffectNet have no AU annotations and CK+ is partially annotated. An important point to be emphasized is that every database uses labels of the 6 basic emotions pointed out by Ekman, Friesen & Ellsworth (2013).

In general, in most works in FER, the JAFFE and CK databases are the commonly used (REVINA; EMMANUEL, 2018). In the works that use FACS, the most used databases are MMI and CK, however in there are other larger and more recent databases for FACS (MARTINEZ et al., 2017).

2.3 DEEP LEARNING

Since 1950, researchers have been interested in imitating the human brain. In the same decade, one of these areas of study created a program capable of producing results outside the notion of programming. In this manner, there has since been a great interest in the field of machine learning (ML), due to its advantage of a single algorithm learning to solve several problems through its training process (SZE et al., 2017). Neural Networks (NN) and in its most recent sub-field, Deep Learning (DL), are of great prominence in this area of ML.

NN can be defined as a complex structure interconnected by simple processing elements (neurons), which have the ability to perform operations such as calculations in parallel, for data processing and knowledge representation (SCHMIDHUBER, 2015).

'Deep' refers to the number of layers in the network. DL uses a deep architecture of learning or a hierarchical approach to learning. Learning consists of a procedure where the model parameters are estimated so that the model learned performs a specific task (ALOM et al., 2018). Sze et al. (2017) points out three factors that made Deep Learning possible in the early 2010s:

1. The amount of data available to train networks;
2. The amount of computing capacity available;
3. The evolution of the algorithmic techniques.

The crucial difference between traditional machine learning methods when compared to DL is how features are extracted. Machine learning applies several algorithms for extracting features, meanwhile in DL this feature extraction process is part of the algorithm. Finally, classification algorithms are used and additionally other boosting approaches. In DL, features are learned automatically and are represented at the multiple levels of the neural network (ALOM et al., 2018).

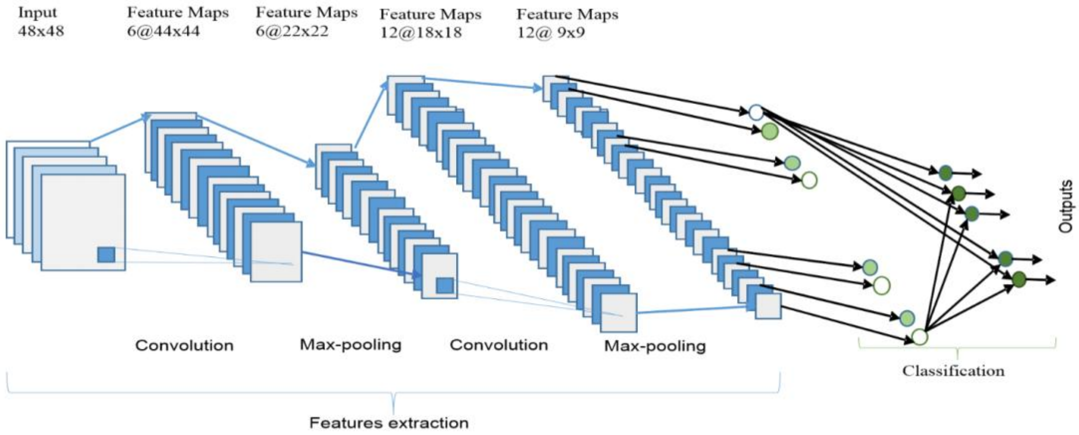
For the image classification based on the ImageNet dataset, CNNs show the state-of-art accuracy in this task (HE et al., 2016). The CNN architecture consists of a combination of three types of layers: convolution, pooling, and classification / regression (ALOM et al., 2018).

In the convolution layer, feature maps from previous layers are convoluted with a specific kernels. The output of the kernels go through an activation function to form the output feature maps. Each output feature maps can be combined with more than one input feature map (ALOM et al., 2018).

The kernel forms a filter that provides a measure of how closely a patch or input region resembles a feature. This filter traverses the image in strides. The size of the stride means how many pixels the filter will traverse through the image, that is, a $1x1$ stride will move the filter one pixel at a time, while a $2x2$ stride will move 2 pixels at a time.

The pooling layer performs the down sampled operation on the feature maps. In this layer, the number of input and output resource maps does not change. Due to reduced sampling

Figure 6 – CNN Example



the operation, the size of each dimension of the output maps will be reduced, depending on the size of the sample reduction mask (ALOM et al., 2018).

Finally, the classification layer is the fully connected layer that calculates the score of each class from the resources extracted from a convolution layer in the previous steps. The fully connected feed-forward neural layers are used to classify within a scalar vector (ALOM et al., 2018). An example of CNN structure can be seen in Figure 6

2.3.1 Deep Learning in FER

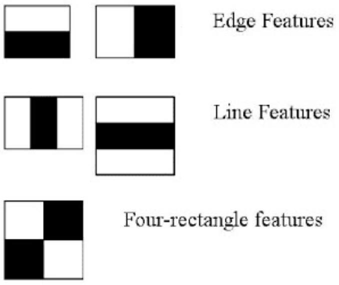
Variations in images that are irrelevant to the FER such as lighting, head position and backgrounds are commonly discarded before training of DL. In addition, preprocessing improves CNN to achieve a better accuracy (PITALOKA et al., 2017).

The first step in pre-processing is to detect the face and then remove unnecessary areas. The Viola-Jones (V&J) algorithm (VIOLA; JONES, 2004) is a classic method and is widely used in this stage of face detection. After this step, one algorithm is applied to detect the face alignment, such as Discriminative Response Map Fitting, Incremental, and Multi-task CNN (LI; DENG, 2020).

The OpenCV version of the V&J algorithm uses a combination of the original algorithm and machine learning. The algorithm scans the image computing 5 rectangular feature types, these rectangular feature types are shown in Figure. To obtain these characteristics, the sum of the pixels that correspond to the rectangles is made. So for face detection it is considered that: (i) the eye region tends to be darker than the cheek region, (ii) the nose region is lighter than the eye region. Finally, the AdaBoost algorithm is used in the entire set of features to select which ones correspond to the facial region.

DL neural networks require sufficient data to ensure that the training can generalize satisfactory for a given recognition task. However, most approaches use databases with an insufficient amount of data. Therefore, the data augmentation step is a crucial step for DL FER.

Figure 7 – The 5 different types of Haar-like features



Data augmentation strategies are divided into two groups: on-the-fly and offline (LI; DENG, 2020).

The On-the-fly data augmentation is used to decrease the necessity of adjustments. During training, the input samples are randomly cut, centered then horizontally inverted, resulting in a database up to ten times larger than the initial one. The operations most often used for offline data augmentation include: rotation, shifting, skew, scaling, noise, contrast and color jittering. Generative Adversarial Network (GAN) can also be applied at this stage to generate more images (LI; DENG, 2020). With the exception of using GANs, operations are also used on-the-fly.

DL techniques show a state-of-the-art performance in Facial Expression Recognition and Analysis Challenge. The challenge used a database having six intensity levels of seven AUs and nine head poses. Of these DL techniques, most were CNN (ZHANG et al., 2018). In addition to the use of CNN, some works also employ visualization techniques to apply qualitative analysis to CNN (ZEILER; FERGUS, 2014). The well-known models applied are AlexNetKrizhevsky, Sutskever & Hinton (2012), VGGNet(SIMONYAN; ZISSERMAN, 2014) and ResNet (HE et al., 2016).

Li & Deng (2020) present two very common problems in the application of DL in FER, inconsistent annotations and data bias. Researchers generally evaluate only on a specific set of data to achieve satisfactory performance. Another very common problem is the presence of unbalanced data in databases for facial expression recognition. They presents a solution that is to perform the balancing of the data in the pre-processing.

2.4 NEUROEVOLUTION

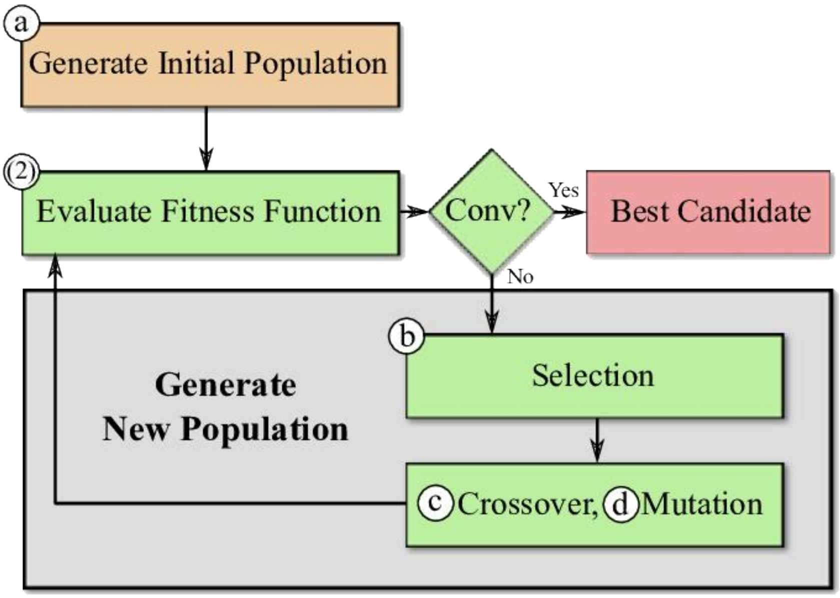
Neuroevolution (NE) is the use of artificial evolutionary strategies to “calibrate” or construct neural networks. At first, this technique was used as a form of reinforcement learning (RL) to find the best weights in a neural network. Until then, the traditional approaches of NE, the network topology was formed before the experiment begins. Stanley & Miikkulainen (2002) proposed the neuroevolution of augmenting topologies (NEAT), not only improving the weights of the neural network using an evolutionary method, but also the formation of the network topology.

2.4.1 Genetic Algorithm

Genetic algorithm is a population-based meta-heuristic inspired by biological evolution. This artificial evolution begins with a predefined or random population. Population adaptation takes place over generations, where individuals in the population are selected through a combination of crossover and mutation to form a new population. The new population undergoes the same process for each generation of the algorithm (HOLLAND, 1991).

A GA typically requires a genetic representation of the domain solution and a fitness function to evaluate the domain solution. Some stop criteria for GA can be taken, among them: (i) the solution to find a minimum that satisfies it, (ii) a maximum number of generations, (iii) the highest ranking solution's fitness is reaching, (iv) manually (WHITLEY, 1994). The traditional GA pipeline can be seen in Figure 8.

Figure 8 – Traditional GA Pipeline



2.4.2 Designing CNN Architectures

Considering the good results that CNNs showed for image classification, neuroevolutionary approaches to improve performance or create automatic CNNs were made as:

- Genetic CNN (XIE; YUILLE, 2017)
- Hierarchical Evolution (LIU et al., 2017)
- Cartesian genetic programming method (CGP-CNN) (SUGANUMA; SHIRAKAWA; NAGAO, 2017)
- CNN-GA (SUN et al., 2020)

The creation of CNNs on CGP-CNN and CNN-GA are fully automatic, while Genetic CNN and Hierarchical Evolution can be classified as “automatic + manually tuning” (SUN et al., 2020). All of these four approaches used the CIFAR-10 database to qualify the accuracy of the generated network. In Table 4, we compared the state of the art in this area. The GPU-day column means how many days the algorithm was run on a GPU, which represents the computational cost of each algorithm.

Table 4 – Comparison between manual CNN and state-of-the-art approaches to CNN neuroevolution

Approach	Accuracy	GPU-Days
Manual VGGNet	93.34%	-
Genetic CNN	92.90%	17
Hierarchical Evolution	96.37%	300
CGP-CNN	94.02%	27
CNN-GA	95.22%	35

Traditionally neuroevolution is not used in DL architectures and is only used in low-level of neuron connectivity (SUGANUMA; SHIRAKAWA; NAGAO, 2017). Despite the high computational cost, this approach proved to be efficient in DL, in particular, CNN. Algorithms developed in the neuroevolution community can be tested with deep neural networks (SUCH et al., 2017).

2.4.3 Automated Machine Learning

Automated machine learning (AutoML) is the process of automating the machine learning application process, which encompasses numerous parts of a pipeline, ranging from feature engineering to model training and serialization.

Among the main AutoML techniques are: Hyperparameter Optimization, Meta-Learning and Neural Architecture Search.

Neural Architecture Search (NAS) is a technique that aims to find the best architecture for a neural network, according to a data set that was passed to the network. Thus, the NAS comes to help reduce the empirical trial and error process for adjusting these networks, taking into account aspects such as (i) search space of these architectures, (ii) search strategy, and finally (iii) performance evaluation strategy for each of the architectures (ELSKEN; METZEN; HUTTER, 2019). This process is shown in Figure 9.

NAS became popular after work published by Zoph and Le (ZOPH; LE, 2016). In their work, they use a a recurrent neural network to compose neural network architectures. 800 GPUs for three to four weeks are used to achieve their results, *i.e.*, up to 22,400 GPU-Days.

A wide variety of methods have been published to reduce computing costs and get even more performance improvements. However, NAS is still an expensive computational task.

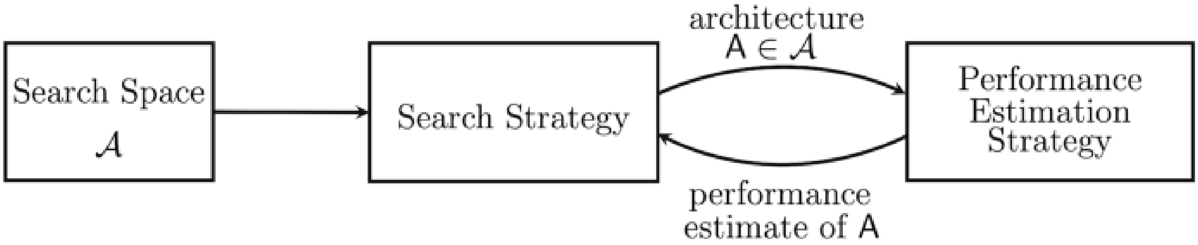


Figure 9 – Abstract illustration of NAS methods (ELSKEN; METZEN; HUTTER, 2019)

2.4.4 Genetic Approaches

In this subsection we will discuss the two approaches using GA to generate CNN architectures are GeNet (XIE; YUILLE, 2017) and CNN-GA (SUN et al., 2020).

One of the similarities between the two works is the CNN encoding representation. Both works use numeric strings to identify the layers of pooling and convolution. However, GeNet uses binary strings while CNN-GA uses numbers between 0 and 1 for the pooling layer and value pairs greater than one for convolution.

This binary representation for the GeNet structure is partitioned into several blocks. These blocks are convolution layers, neighboring blocks are connected via pooling operation. The fully connected layer is not encoded because the convolutional part of the network is transposed to another network.

The proposed CNN-GA strategy, by contrast, encodes the entire CNN into a single numerical sequence. The dense layer is also not used. The convolution layers are always added in pair. The numbers in the string are the value of a feature map, but numbers between 0 and 0.5 represent max-pooling layer and between 0.5 and 1 represent the mean-pooling layer.

About the initial population, CNN-GA initializes all convolution layers with filters at 3×3 and stride 1×1. No mutation changes these parameters, so all networks generated by this GA will always have these same values. The same is valid for pooling layers that have stride 2×2. In the GeNet strategy, the values are also fixed throughout the evolutionary operation.

3 LITERATURE REVIEW

The DL area is a recent field with much potential to be explored and the use of NE methods in DL is even more recent. Such et al. (2017) conclude in their research that the use of GA for real-life tasks is as competitive as the algorithms already proposed and consolidated. In this context, similar works were sought that applied evolutionary techniques together to neural networks for FER in the last 10 years. In Section 3.1, we present the research methodology. In Section 3.2, we describe the results of the literature review and we discuss the data obtained by answering the research questions.

3.1 METHODOLOGY

In this review, we follow the research methodology applied to the systematic literature review (SLR) developed by Kitchenham (KITCHENHAM, 2004). This methodology follows the strategy of defining research questions, declaring the search strategy, data synthesis and data analysis to answer each of the defined questions.

The main objective of this SLR is to identify which evolutionary strategies in union of neural networks were used in the period between 2009 and 2021 to solve the task of Facial Expression Recognition. So, we report how this data is validated and which research gaps are still present in the literature.

3.1.1 Research Question

In order to compare the results of the works that uses Neuroevolution and survey the current state of the area, our primary research questions are:

- RQ1: Does a Neuroevolutionary Image Classifier show significant results compared to CNNs Image Classifier to solve the problem of recognizing facial expressions?
- RQ2: Which Neuroevolution methods for FER have been proposed between 2009 to 2021?
- RQ3: What are the remaining problems to be solved in new proposed approaches?
- RQ4: Published studies present a systematic approach, step by step, for the construction of the model?
- RQ5: What evaluation methodology is used in the approaches?

The previous questions were asked using the following PICO (METHLEY et al., 2014) criteria:

- Population: Research that addresses Facial Expression Recognition.

- Intervention: NE models for FER approaches
- Comparison Intervention: State-of-art of CNNs for image classifier.
- Outcomes: The performance accuracy of a classification algorithm.

3.1.2 Search Strategy

The search process consisted of a search of articles and conference papers published by the period between 2009 and 2021 using titles, keywords, and abstracts as search fields. The strategy used to construct the search string was aided by the PICO criteria as follows:

- Population: “Facial Expression” OR Emotion OR Sentiment OR Affection;
- Intervention: Neuroevolution OR NEAT OR “Neuro evolution” OR (“Neural Network*”; AND (“Evolutionary Algorithms” OR “Genetic Algorithms” OR “Reinforcement Learning” OR “Genetic Computing”)).

Two collections of abstracts, bibliographical references and indexes will be used: Scopus and Web Of Science. We also did a broader search on the IEEE Xplore and ACM Digital Library databases. At IEEE Xplore and ACM Digital Library, we altered the intervention by removing the restriction by neural networks.

3.1.3 Inclusion and exclusion criteria

Articles were excluded from the search results following the criteria: Facial Expression research that does not use images or videos as data; approaches that don’t use a neural network; approaches that don’t apply an evolutionary method; studies that were not published between 2009 and 2021; studies that have a quality score of 5 or lower; works that are not written in English.

3.1.4 Data Extraction Strategy

The data to be extracted from each study are:

1. The title;
2. Name of the journal or conference of the article;
3. Abstract of the study;
4. Evolutionary method used;
5. Neural Network architecture;

6. The average accuracy of the experiment;
7. Database used;
8. Validation methodology used;
9. The conclusion of the study;
10. The future research proposals by the study (open problems).
11. The quality of the research;

Items 6, 7 and 9 are related to answering RQ1; Items 4 and 5 to RQ2; 10 to RQ3; 11 to RQ4 and 8 to RQ5. The remaining items are used for general analysis.

3.1.5 Quality assessment

To assess the quality of articles and papers, we follow rules similar to those adopted by Tealab (2018), a score metric from 0 to 8, with 0 being the worst possible score and 8 being the best one. The questions will be answered with Yes / No / Partially, being scored as 1/ 0 / 0.5 respectively. The 8 questions are:

1. Does the author specify the neural network and the evolutionary strategy?
2. Is the model building process well defined?
3. Does the study specify the criteria for selecting the evolutionary strategy?
4. Is the method applied validated?
5. Is the procedure for training the model specified, such as data transformation, stopping criteria or initial values?
6. Is the accuracy obtained with the model clearly specified?
7. Are comparisons made with other methods in the state-of-the-art?
8. Are suggestions for future work presented?

The highest score represents the study that most effectively answers research questions. Studies that scored less than 5 were discarded. Table 5 shows the selected articles and their respective scores.

3.2 DISCUSSION

In this section, initially we reviewed the literature raised by the methodology, then we answer the questions raised. Finally, we discuss about the larger In-The-Wild dataset.

Table 5 – Quality assessment

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Score
S1(CABADA et al., 2020)	1	1	1	0.5	0.5	1	0.5	1	6.5
S2(LI et al., 2019)	1	1	0.5	1	1	0.5	0	1	6
S3(CHENGETA, 2019)	1	1	1	0.5	1	0,5	0.5	0	5.5
S4(Mistry et al., 2017)	1	1	1	0.5	1	1	1	0.5	7
S5(MEI; TAN; LIU, 2017)	1	1	1	0.5	1	1	0.5	1	7
S6(WANG et al., 2017)	1	1	1	1	0.5	1	0	0	5.5
S7(NEOH et al., 2015)	1	1	1	1	0.5	0.5	1	0	6
S8(QIN; FANG; YANG, 2013)	1	1	1	1	1	1	0	1	7

3.2.1 Research

The results returned by the search strings were analyzed by title, abstract, and development in order to apply the inclusion and exclusion criteria. Finally, the quality of the remaining papers and conference papers is assessed.

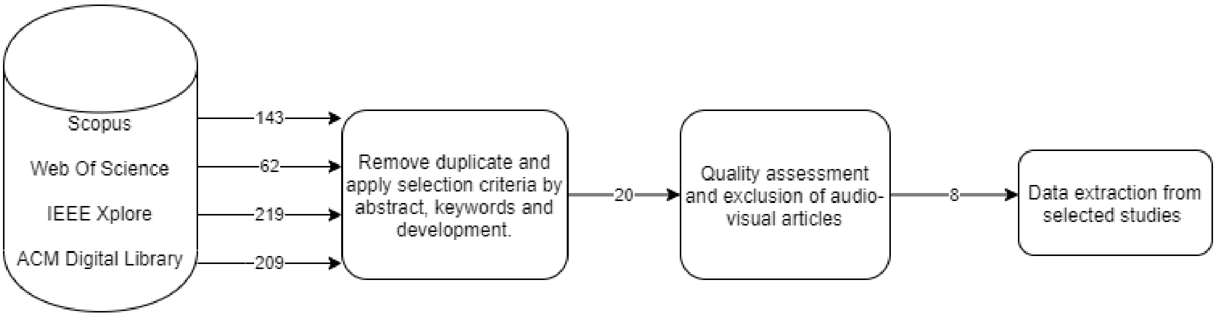

Scopus
Web Of Science
IEEE Xplore
ACM Digital Library
143
62
219
209
Remove duplicate and apply selection criteria by abstract, keywords and development.
20
Quality assessment and exclusion of audio-visual articles
8
Data extraction from selected studies

Figure 10 – Steps for data extraction

In this research, 143 articles and conference papers were returned in the Scopus database, 62 on Web of Science Core Collection, 219 on IEEE Xplore and 209 ACM Digital Library.Table 6 shows the studies evaluated in the systematic review and also presents the approaches of each study.

Table 6 – Selected studies. GA = Genetic Algorithm; RL = Reinforcement Learning; LCE = Layered Cascade Evolution; PSO&CSO = Swarm approaches.

ID	Year	Algorithm	Neural Network	Database	Accuracy	Evolves the NN
S1(CABADA et al., 2020)	2020	GA	CNN	Own	82%	Yes
S2(LI et al., 2019)	2019	RL	CNN	Own	-	Yes
S3(CHENGETA, 2019)	2019	GA	ANN	JAFFe	96.6%	Used for feature extraction
S4(Mistry et al., 2017)	2017	PSO + GA	ANN	CK+ MMI	100% 92.9%	Used for feature extraction
S5(MEI; TAN; LIU, 2017)	2017	GA	BEL	CK JAFFe	96.17% 95.57%	Yes
S6(WANG et al., 2017)	2017	CSO	ANN	Private	89.49%	Yes
S7(NEOH et al., 2015)	2015	LCE	ANN	CK+ + MMI	97.4%	Used for feature extraction
S8(QIN; FANG; YANG, 2013)	2013	GA	ANN	JAFFe	77.5%	Yes

Although CNNs have shown excellent results since 2012 with AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) for image classification, only the 2 most recent selected studies apply an approach using them. The other studies use traditional multi-layer or single-layer perceptrons neural networks and one uses a brain-inspired emotional learning (BEL).

Studies S3, S4 and S7 do not apply the evolution of neural networks. They apply an evolutionary strategy to extract features from the image so that they can then classify emotions using a neural network. However, we keep them in the systematic review of the literature and in the data extraction as they are approaches that show that an evolutionary strategy can be efficient for extracting characteristics in addition to improving the neural network used.

Table 7 – FER Database description

Database	Samples	Emotion
Cohn-Kanade(KANADE; COHN; TIAN, 2000)	486 image sequences	6 basic emotions
Extended Cohn-Kanade(LUCEY et al., 2010)	593 image sequences	6 basic emotions + contempt and neutral
JAFFe(LYONS et al., 1998)	213 images	6 basic emotions + neutral
MMI(PANTIC et al., 2005)	740 images 2,900 videos	6 basic emotions + neutral

In general, in most works in FER, the JAFFe and CK databases are the commonly used (REVINA; EMMANUEL, 2018). In the works that use Facial Action Code System (FACS)(EKMAN; FRIESEN, 1978), the most used databases are MMI and CK, however in there are other larger and more recent databases for FACS (MARTINEZ et al., 2017). In Table 7, we can notice that in the selected studies these databases are still the most used.

The 4 databases listed are laboratory-controlled images, that is, none represents a real-life case for facial expression recognition, where there may be some type of occlusion in the image. Even the study S1 and S2 where they apply their method directly in a real case, the images are in a controlled environment.

Studies S4 and S7 do not show how their validation set was built, it is possible that the neural network is overfitting, since a precision top-1 above 90% using the MMI as data is a very complex task, since this dataset, despite being made in laboratory, has a difference that are photos of people from different angles. These different angles impact because the neural network, in addition to analyzing the patterns on a frontal face, also needs to point out patterns on a face in profile.

3.2.2 Answers

RQ1. Neuroevolutionary Image Classifier shows significant results compared to CNNs Image Classifier to solve the problem of recognizing facial expressions?

We found 5 approaches that tried to obtain better results by improving the weights of the neural networks using an evolutionary strategy. The NE strategy was only used to improve the weights of a previously defined neural network. The following approaches are some of the state-of-the-art in the databases we cover(LI; DENG, 2020):

- JAFFe: 95.8% (HAMESTER; BARROS; WERMTER, 2015)
- CK+: 98.9%(ZHANG et al., 2016)
- MMI: 78.46% (LI; DENG; DU, 2017b)

The approaches that used the evolutionary strategies for feature extraction achieved slightly better results than those listed. Therefore, we found that when using evolutionary methods, studies such as S3 (CHENGETA, 2019); S4 (Mistry et al., 2017) and S5 (MEI; TAN; LIU, 2017) managed to obtain an accuracy in their neural networks with values above the state-of-the-art for CNN, however it is possible that the neural networks from the study S4 is overfitting for presenting 100% accuracy in the CK+ dataset.

In general, the answer to this research question is affirmative, the studies that applied evolutionary approaches obtained a better accuracy in their neural networks. However, there is insufficient research in the area with the same databases for it to be possible to apply a meta-analysis.

RQ2. Which Neuroevolution methods for FER have been proposed between 2009 to 2021?

The most common techniques found were genetic algorithms and particle swarm optimization variations. In the context of neural networks, it is frequent to use neural networks based on perceptrons. During the process of exclusion and inclusion criteria, it can be noted that Support Vector Machine is also widely used to classify emotions.

This demonstrates that there are many possibilities for research in the area of facial expression recognition, since there is little research using convolutional neural networks that are widely used for classification tasks. Only the two most recent studies apply this technique. Therefore, it is possible that in the coming years we will have more data and results with this technique.

RQ3. What are the remaining problems to be solved in new proposed approaches?

There is still a lot of space for research in this area, the main contributions to be made are:

- Construction of models of neural networks through evolutionary strategy.
- Use of In-The-Wild databases as AffectNet (MOLLAHOSSEINI; HASSANI; MAHOOR, 2017) and EmotionNet (BENITEZ-QUIROZ et al., 2017).
- Use of evolutionary methods in data preprocessing in more recent models of neural networks.

- Applications in a real environment with a larger sample size.
- Generalize to novel databases.

Although we point out how to use In-The-Wild bases as contributions to be made. These bases have problems in the image labels, they have duplicate images and their automated parts have a low accuracy. Therefore, the creation of a large set of images with people of different ages, cultures, genders, which refer to a real environment, is important.

We do not approach other emotion classification methods, during the methodological process we noticed that there are few approaches using valence or Facial Action Coding System. In Table 3, we list two databases that have Action Units annotated, MMI(PANTIC et al., 2005) and CK+(LUCEY et al., 2010).

RQ4. Published studies present a systematic approach, step by step, for the construction of the model?

Only 3 studies did not present a complete construction of the model, from the neural network to the construction of the evolutionary strategy, being the studies: S1, S6, S7. In the other studies, the construction of the neural network model was meticulously informed, as well as the information about the evolutionary strategy.

RQ5. What evaluation methodology is used in the approaches?

The validation method used is by benchmark, but some of the works do not compare their approach with others in state-of-the-art, only with other approaches similar to theirs. In studies referring to image classification it is the most commonly applied methodology, which was not different in the studies selected by our research.

However, the studies do not present how the validation sets were constructed, thus, unfair comparisons between the approaches may occur.

3.3 VALIDITY

To evaluate the validity of the study, we followed the checklist proposed by Maxwell (MAXWELL, 1992) and reviewed by Petersen and Gencel (PETERSEN; GENCEL, 2013), in which 4 different categories are used for validity evaluation: descriptive validity, theoretical validity interpretive validity and generalizability.

To reduce the threat for descriptive validity, i.e., the accuracy and objectivity of the information gathered. We have designed and followed a data collection form. The form was used in the data extraction process and could always be revisited. Hence, this threat is considered as being under control.

To reduce the threat of theoretical validity, we conducted a search for the keywords most used by the authors in the area, so that we could consider as many articles as possible for the area of emotion recognition in general.

In order for the interpretive validity to be achieved, and no bias from researchers, the discussion of the results, as well as the conclusions addressed, were made separately with a review by the co-author (of this research) so that there was no bias from the main author. Finally, the study follows the methodology of systematic literature reviews and mapping, for a specific scope, thus allowing generalization.

4 PROPOSED NEUROEVOLUTIONARY METAHEURISTIC

In this chapter, we present our proposed metaheuristic design, including its evolutionary operators.

4.1 METAHEURISTIC OVERVIEW

Algorithm 1 Metaheuristic Overview

```
Input: population size $N$, max generation number $g_{Max}$, image dataset $D$
Output: CNN model
$P_0 \leftarrow$ Initializes the population with N random individuals.
$g \leftarrow 0$
Evaluates the fitness of each individual in $P_0$ using dataset $D$
while $g < g_{Max}$ do
    $O_g \leftarrow$ Generate Offspring from $P_g$
    Evaluates the fitness of each individual in $O_g$ using dataset $D$
    $P_{g+1} \leftarrow$ tournament selection between $P_g$ and $O_g$
    $g \leftarrow g+1$
end while
return CNN with best fitness
```

Algorithm 1 shows the standard pipeline of a GA: Fitness evaluation, offspring generation and the environment selection. Crossover and mutation operations are performed during the offspring step. Note that we have used a fixed population size, so P and O have the same sizes.

In the next subsections, each operation of the metaheuristic is detailed, with a greater focus on the first step, initial population.

4.2 INITIAL POPULATION

Previous works claim that population is not important in the context of automatic CNNs generation. Although their results have reached this conclusion, we believe that their fixed-parameter strategy have frustrated broader analyses in the initial population and the use of the metaheuristic in different datasets.

Our initial population has its parameters completely random, including filter and stride. This randomness is also maintained in the mutation process. So it is possible to generate some networks that do not work properly, but the evolutionary process surrounds this problem. All follow the same base algorithm presented in Algorithm 2.

The approach in this work follows the traditional CNN model. Where there are multiple layers of convolution and *pooling* followed by a fully connected network. The existence of the layers and the fully connected network is decided by the evolutionary process. $D_c \bigcup D_f$

Algorithm 2 Initial Population

```
Input: population size $N$
Output: $P_0$
 1: $P_0 \leftarrow \emptyset$
 2: while $|P_0| < N$ do
 3:     $individual.fit \leftarrow 0$
 4:     $D_c \leftarrow$ Random integer between 2 and 30
 5:     $list_{conv} \leftarrow$ Linked list containing $D_c$ nodes.
 6:     foreach $node \in list_{conv}$ do
 7:         $node.dropout \leftarrow$ Random number between 0 and 0.5
 8:         $node.stride \leftarrow$ Random integer between 1 and 5
 9:         $r \leftarrow$ Random number between 0 and 1
10:         if $r < 0.5$ then
11:             $node.type \leftarrow 1$
12:             $node.map \leftarrow$ Random integer between 1 and 1024
13:             $node.filter \leftarrow$ Random integer between 1 and 10
14:             $node.depthwise \leftarrow$ Random integer between 0 and 5
15:         else
16:             $node.type \leftarrow 2$
17:             $node.pool \leftarrow$ Random integer between 1 and 3
18:         end if
19:     end foreach
20:     $D_f \leftarrow$ Random integer between 0 and 10
21:     $list_{full} \leftarrow$ Linked list containing $D_f$ nodes.
22:     foreach $node \in list_{full}$ do
23:         $node.type \leftarrow 3$
24:         $node.dropout \leftarrow$ Random number between 0 and 0.5
25:         $node.dense \leftarrow$ Random integer between 1 and 1024
26:     end foreach
27:     $individual.cnn \leftarrow list_{conv} \bigcup list_{full}$
28:     $P_0 \leftarrow P_0 \bigcup individual$
29: end while
30: return $P_0$
```

represent the depth of the network. The CNN is represented by a linked list where each node in the list is a layer of the CNN.

Each node can have a probability of dropout between 0 and 0.5. The values for the steps and the filter are maximized to 5 and 10, respectively (row 8 and row 13). Values are limited because extreme high values cause many non-functional neural networks, which in turn increases the computational cost of the GA.

The layer type is randomly chosen, being 1 for convolution and 2 for pooling. If the convolution operation is selected, a random value is generated for the map and filter. A number between 0 and 5 is also generated for the depthwise parameter, where the value dictates the number of channels for this type of convolution, if the value is 0, a traditional convolution operation is used.

If the pooling operation is selected, an integer between 1 and 3 is randomly selected, which represent, respectively: max-pooling, mean-pooling and min-pooling.

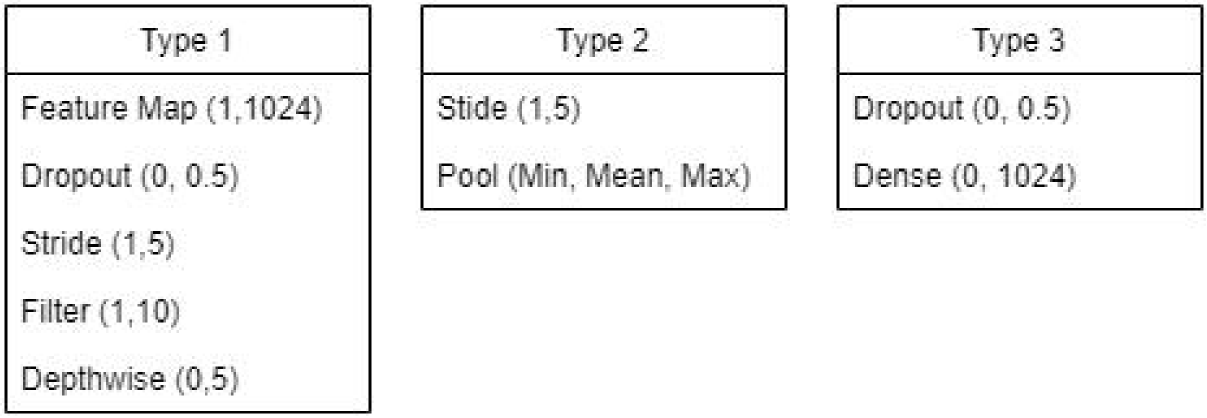


Figure 11 – Nodes attributes

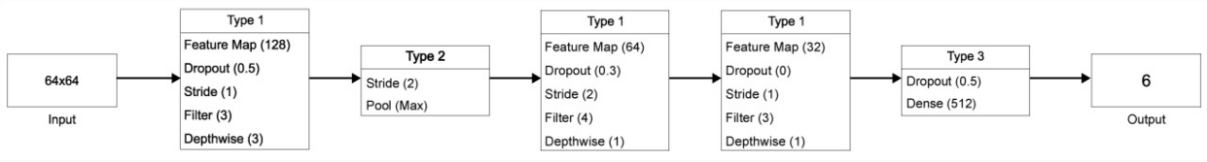


Figure 12 – Example of a CNN that can be created during the initial population using the nodes proposed in this work.

Then, a second Linked List is made for the fully connected network. All convolutional operations and the fully connected layer use ReLU activation. All these attributes are represented in Figure 11.

Finally, the algorithm returns a set of individuals with fitness equal to zero and a linked list representing the CNN, this set being the initial population. It is important to point out that both the CNN input and output are immutable as the input is the image size and the output the number of classes. An example of CNN that is generated by this algorithm is represented in Figure 12.

4.3 FITNESS EVALUATION

In order to assess the fitness of each individual, it is necessary to train all the neural networks. which makes the process computationally expensive. In this work, the algorithm for training the networks used is Adam (KINGMA; BA et al., 2015). This step can be optimized and accelerated if multiple GPUs are deployed, but as in this work only one GPU was available, the asynchronous training coding was discarded. Finally, the individual with the highest fitness is saved for use during tournament selection.

4.4 CROSSOVER, MUTATION, AND OFFSPRING

The offspring generation consists of two parts: crossover and mutation. In the crossover operation, binary tournament(MILLER; GOLDBERG et al., 1995) is used: two individuals are randomly selected from the population; the one with the greatest fitness is selected as a parent. Then the process is repeated to select the other parent.

Algorithm 3 Offspring Generating

```
Input: Population $P_t$; Crossover probability $p_c$; Mutation probability $p_m$
Output: Offspring population $O_t$
$O_t \leftarrow \emptyset$
while $|O_t| < |P_t|$ do
    $i_1, i_2 \leftarrow 0$
    while $i_1 = i_2$ do
        $i_1 \leftarrow$ Select the best fitness between 2 randomly selected individuals from $P_t$
        $i_2 \leftarrow$ Select the best fitness between 2 randomly selected individuals from $P_t$
    end while
    $r \leftarrow$ Random number between 0 and 1
    if $r < p_c$ then
        Select a random point in $i_1$ and $i_2$ CNN and divide it into two parts;
        Select a random point in $i_1$ and $i_2$ fully connected network and divide it into two parts;
        $o_1 \leftarrow$ Join the first part of $i_1$ and the second part of $i_2$;
        $o_2 \leftarrow$ Join the first part of $i_2$ and the second part of $i_1$;
        $O_t \bigcup o_1 \bigcup o_2$
    else
        $O_t \bigcup i_1 \bigcup i_2$
    end if
    foreach Individual i $\in O_t$ do
        $r \leftarrow$ Random number between 0 and 1
        if $r < p_m$ then
            Select one mutation operation $m$
            Do mutation $m$ on individual $i$
        end if
    end foreach
end while
return $O_t$
```

Then it is checked if the crossover will occur, if not, both parents are added to the set O_g. in the crossover operation, the CNN is separated into two parts, one with the convolutional part and pooling and the second part with the fully connected network. Then for each of the parents, a layer of the convolutional network is randomly selected, so the network is split in two from that point, thus uniting the first part of the parent-1 with the second part of the parent-2 and vice-versa. The same process is done with the fully connected network. Finally, jointing the CNN into a single linked list, the two new CNNs are then added to O_g. This iteration is illustrated in Figure 13.

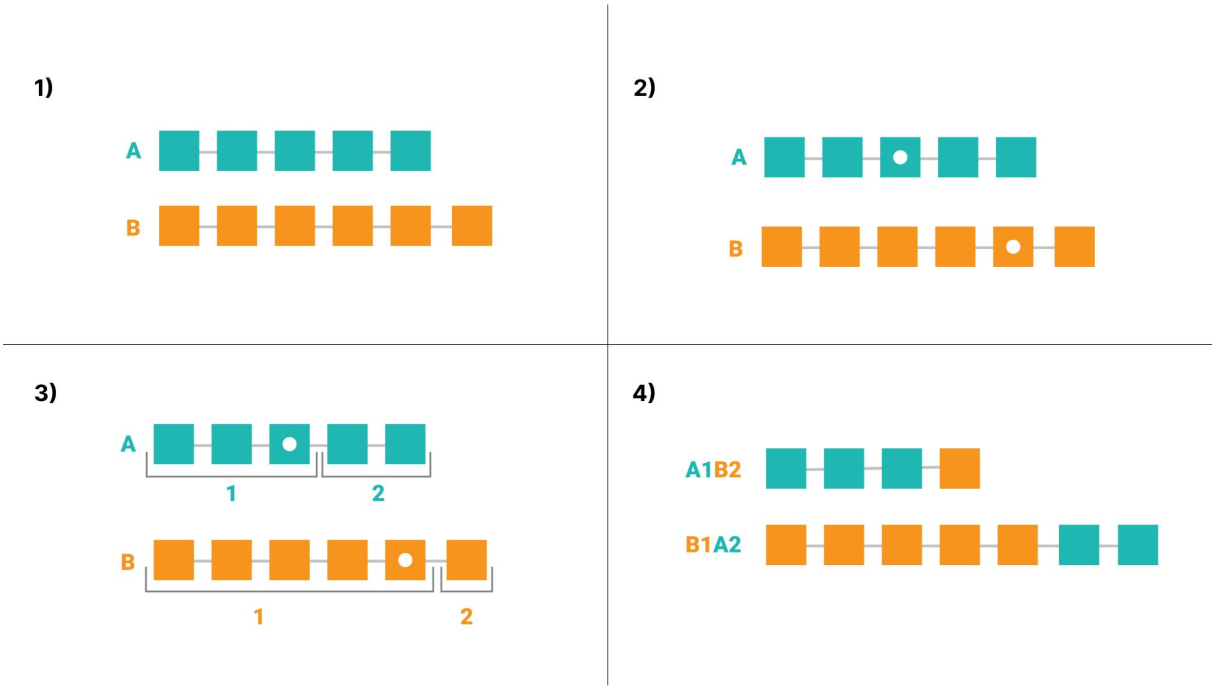


Figure 13 – Crossover steps

For each individual in O_g, it is checked if one of the 6 different mutations will occur: (i) add a convolution layer with random parameters; (ii) add a layer of pooling with random parameters; (iii) remove a random layer; (iv) change the parameters of a random layer; (v) remove a fully connected network layer; (vi) add a layer on the fully connected network.

The mutation process must preserve the properties of the surviving individuals while provides different individuals to the population, allowing new possibilities. Therefore, the mutation must cause minimal changes in the individual.

4.5 ENVIRONMENT SELECTION

The environmental selection by tournament takes place as follows: Two individuals are randomly selected from the union of population and offspring generation ($P_g \bigcup O_g$), then the one with the best fitness is added to the next generation population, following the principle of binary tournament selection.

Then, the algorithm check if the individual with the best fitness is part of the new population. If that individual does not belong the new population, the individual with the worst fitness is replaced by the one with best fitness. Although the metaheuristic converges to the best solutions naturally, it is important to maintain diversity in the population, thus composing new architectures without having to depend only on mutation (GOLDBERG; HOLLAND, 1988; MICHALEWICZ, 2013).

5 EXPERIMENTS AND ANALYSIS

First we present about the databases used in the experiments. Next, we will present the results using classical neural networks for laboratory data. Therefore, we present some uses of the metaheuristic. About these uses, we demonstrate that the metaheuristic can be used for any base without changing the algorithm, using a widely used base for image classification. Finally, the experiments for facial expression recognition are presented.

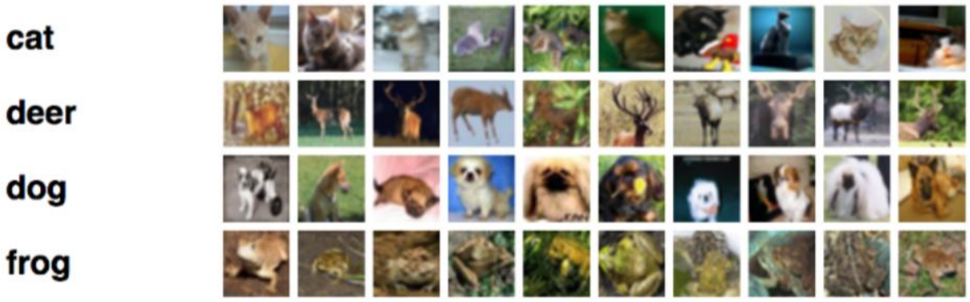
5.1 DATASETS USED

In this section we list and describe the datasets we used in the experiments.

5.1.1 CIFAR10 Dataset

The CIFAR10 dataset (KRIZHEVSKY et al., 2009) is widely used to measure the performance of deep learning algorithms. The CIFAR10 is an image classification benchmark for recognizing 10 classes of objects such as animals: cat, deer, dog, frog; but also objects such as airplane and automobile.

Figure 14 – CIFAR10 Images Example



The dataset consists of 60,000 RGB images in the 32x32 dimension, of which 50,000 are for the training set and 10,000 for the validation set. Each category has the same number of images, so there is no need to balance the data. We applied a data generator just to randomly shift the image by 10% and flip horizontally, the image augmentation not only expand the size of the dataset but also increase the variation in the dataset which allows a better generalization.

5.1.2 JAFFE and CK+

Both the JAFFE (LYONS et al., 1998) and CK+ (LUCEY et al., 2010) datasets are made up of images produced in the laboratory, *i.e.*, the photographs are placed with the expressions requested by the researchers.

The Japanese Female Facial Expression (JAFFE) dataset contains 213 samples of posed expressions from 10 Japanese females. Each image has averaged semantic ratings on 6 facial expressions by 60 Japanese viewers.

The Extended Cohn-Kanade (CK+) dataset contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age of different genders and ethnicity.

5.1.3 AffectNet

AffectNet (MOLLAHOSSEINI; HASSANI; MAHOOR, 2017) contains about 1 million facial images. 420,000 are manually annotated and 550,000 automatically annotated. The images are downloaded from the Internet by querying different search engines using 1,250 emotion related tags in 6 different languages (English, Spanish, Portuguese, German, Arabic, and Farsi).

Because the images are from the Internet, the images have noise, which does not happen in laboratory databases. These noises can be: beard, hand or objects covering the face, masks, makeup, tattoos, adornments, pose variations, low resolution. Thus, resembling the real environment.

Table 8 – Current state-of-the-art approaches using Affectnet Dataset

Approach	Accuracy %
PAENet (HUNG et al., 2019b)	65.29
CPG (HUNG et al., 2019a)	63.57
CAKE (KERVADEC et al., 2018)	61.70

In Table 8, we list 3 state-of-the-art approaches for the AffectNet Dataset. Both PAENet and CPG use similar learning approaches in combination with neural networks. The CAKE approach uses ResNet-18 for emotion classification, after which compress a k-dimensional representation of the characteristics. There is still a lot to be explored with In-The-Wild databases, the accuracy of studies that use AffectNet are low. One possibility of study is to improve the labels of the images that are automatically annotated.

5.2 CIFAR10 EXPERIMENTS

To evaluate the algorithm generalization and its impacts on the performance of the evolutionary algorithm, two experiments were conducted in an image classification task, specifically using the CIFAR10 (KRIZHEVSKY et al., 2009) dataset.

In the next subsections, the parameters that are common to the two experiments, and then each one is presented individually. Finally, a subsection to discuss the results.

5.2.1 CIFAR10 Parameter Settings

In this subsection, we present the parameters that are in common for the both experiments performed for the CIFAR10 dataset.

Since we only had only one GPU at our disposal, we ran the evolutionary algorithm for 10 generations, where the population of each generation is formed by 20 individuals. For the probability of crossover and mutation, we use values based on convention (BACK, 1996) being the probability of 90% and 20%, respectively.

During the fitness assessment of each individual. We use Adam optimizer (KINGMA; BA et al., 2015) and only 50 epochs, being $1/7$ of the 350 epochs used by CNN-GA (SUN et al., 2020) and less than half of the 120 epochs of GeNet (XIE; YUILLE, 2017).

Naturally, more epochs during the fitness assessment, more individuals and more generations, result in better performance up to a limit. At the end of 10 generations, we select the individual with the best fitness and train it for 350 epochs, Finally acquiring a competitive value of the architecture found. It is worth mentioning, we use the same seed value, so that the initial values of the neural network weights are the same, both in the first fitness assessment and in training after the end of the 10 generations,

Our work focusing on a GA that with less human impact, all mutations are equally likely to occur. The values of these mutations are the same as those presented in Alg. 2, any changes are presented in the following subsections. All experiments were performed on a single GPU card, Nvidia GeForce GTX 1080.

5.2.2 Totally random

The first CIFAR10 experiment, consists of a more random approach. The parameter values are those shown in Algorithm 2: Feature Map (1,512); Stride (1, 5); Filter (1, 10); etc.

Initially, some problems were presented in several networks generated entirely randomly, so in the first and second generations there are 6 and 4 individuals with fitness equal to zero, respectively. Note that with each generation, GA naturally eliminates individuals with less accuracy.

The 10 generations cost 40 GPU-Days, because from the 5th generation, each generation cost at least 5 GPU-Days to be executed, since GA converged to complex structures. Finally, the best individual achieved 84% accuracy on the validation set with 50 training epochs and 88.55% in 350 epochs.

5.2.3 Guided evolution

In our second experiment, the parameters differ in only two settings: the Filter Map of the convolutional layer; and the units of the densely-connected layer. In these two parameters, the values are random between the powers of 2, maximized at 2^9, *i.e.* $\{2,4,8,16,32,64,128,512\}$, thus, following values closer to the state-of-the-art found in the literature. Another change, the CNNs are less sensitive to mutation, *i.e.*, the changes caused by mutations are less impactful.

The computational cost of the second experiment is less than that of the two other experiments, since in 12 GPU-days it was possible to run the 10 generations. The populations have less sparse fitness values when compared to the other experiments and although no individual excels in the initial population, none obtained fitness equal to zero. At the end of the experiment, the best individual had 90% recognition accuracy in 50 epochs and when training for 350 epochs, the CNN achieves 98.65% in the training set and 93.05% in the validation set.

In the next subsection, we discuss the experiments and compare both experiments to similar works in the literature.

5.2.4 CIFAR10 Experiments Analysis

In general, the entire computational cost is in the network training, as our algorithm has more than 6,000 possible combinations per convolutional layer compared to CNN-GA, where each convolutional layer only has 3 options, being the feature map (64, 128 , 256) since the stride and kernel filter are fixed at 1×1 and 3×3, respectively. In both the GeNet and CNN-GA approaches, CNN's fully connected layer is discarded. We coded so that this choice to discard this layer or not is part of the evolutionary process of the evolutionary algorithm, it also helps to increase the computational cost. Therefore, we chose to reduce the number of epochs during training, knowing that it would reduce the performance of neural networks.

Despite running each experiment for only 10 generations and using fewer epochs for each training, our approach with more random values achieved competitive results, the second experiment achieved better results than GeNet. Our results have surpassed the state-of-the-art methods, considering 10 generations. These comparative data are presented in Table 9. All results shown are for validation set.

Table 9 – Comparison between our work and similar works.

	Accuracy %	GPU-Days	Total Generations
GeNet (XIE; YUILLE, 2017)	92.90	17	50
CNN-GA (SUN et al., 2020)	95.22	35	20
Totally Random (Our)	88.55	40	10
Guided Evolution (Our)	93.05	12	10

Our experiments showed that although the evolutionary algorithm for generating CNN architectures is a task with high computational cost, even higher when used in a totally random manner, it is a technique that brings excellent results. Due to this high computational cost, the best approach is to use parameters closest to the state-of-the-art, but allowing to find diversity in these values. So finally, apply it as an optimization technique. However, a random approach allows the metaheuristic to be used for different datasets with totally distinct objects.

5.3 FER EXPERIMENTS

In this section we present the experiments carried out with the databases for facial expression recognition. First, the experiments using the classical networks are described and then the metaheuristic proposed in this work.

For the image preprocessing: First, we use the V&J algorithm to cut the area of interest, *i.e.*, people's faces. Second, the images were resized to 64×64 size, mainly because the AffectNet database does not have a unique size for the images.

We apply an image generator to randomly rotate the images by 10% and also rotate them horizontally, thus helping to reduce overfitting. For the probability of crossover and mutation, values based on the conventions(BACK, 1996) were used, thus, 90% and 20% for crossover and mutation, respectively. The population was made up of 20 individuals and the training networks use 150 times to train.

This work focuses on the 6 basic emotions described by Ekman (EKMAN; FRIESEN; ELLSWORTH, 2013), so the class with "neutral" faces has been deleted. The strategies for separating data between training and validation are:

- JAFFe: We randomly separate 70% of the images for training and 30% for validation.
- CK+: We select the last three frames for the image sequence of each emotion, so we randomly separate 70% of the images for training and 30% for validation.
- JAFFE and CK+ cross-database: Combination of the items described above.
- AffectNet: We only use the set of images that are manually annotated, *i.e.* $450,000$ images. Finally, the separation between training and testing is the set provided by the author of AffectNet (510 images of each emotion for validation)

The other configurations or differences between the experiments are presented in the following subsections.

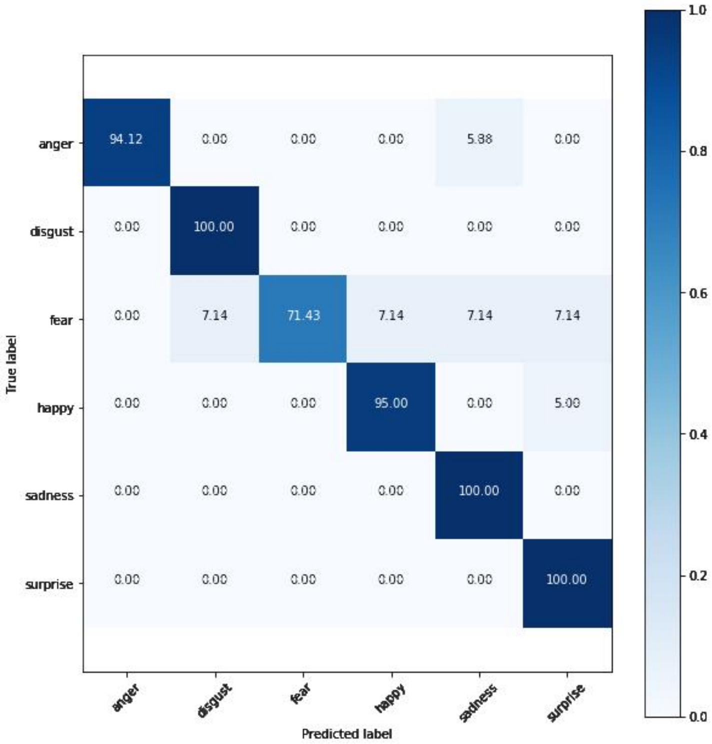
5.3.1 Classical CNNs

To assess the performance of CNNs already consolidated in the literature for image recognition, we have developed experiments using the VGG network and ResNet.

The first classical CNN experiment consisted of two stages, both using a cross-database of the JAFFE and CK+ dataset. So that we have an accuracy value in a controlled environment with CNN manually architected. The first stage consists of using the images without preprocessing and the second stage using V&J to cut the area of interest and also transform it to gray-scale.

We set up the CNN for 150 epochs, we achieved a validation accuracy of 89% with VGG and 88% with ResNet in the first stage. With the application of preprocessing the results reached an accuracy of 94% with both VGG and ResNet. Thus showing that with low computational power, using two of the consolidated CNNs in the literature, we were able to approach the accuracy obtained by the methods presented in the table 6.

Figure 15 – Confusion Matrix VGGNet



In these experiments, we have tried to keep the neural networks in a more traditional and basic configuration, so we can use the values as a comparative basis. The figure 15 shows the confusion matrix of the second stage of the experiment, in it we can see that the emotion of fear was the least accurate. The confusion of fear between surprise is already expected because there are only small differences such as eye muscle contraction.

5.3.2 Laboratory Datasets

A series of runs were made for both the JAFFe and CK+ datasets. The GA was run from scratch 5 times. In all executions, the accuracy of the networks quickly converged to optimal values in a few generations, as can be seen in Figure 16.

On average, the accuracy obtained was 97.6% for JAFFe dataset and 96.4% for CK+. Each execution, the GA does not complete a GPU day, as both datasets are small, and the execution time cannot be compared against the CIFAR10 (KRIZHEVSKY et al., 2009) dataset, which has 60000 images.

Note that the metaheuristic proposed by this thesis obtained more significant results than classical networks such as VGG and ResNet. When we train the neural network using JAFFe, if we train for many epochs we achieve 100% accuracy in different validation sets. In

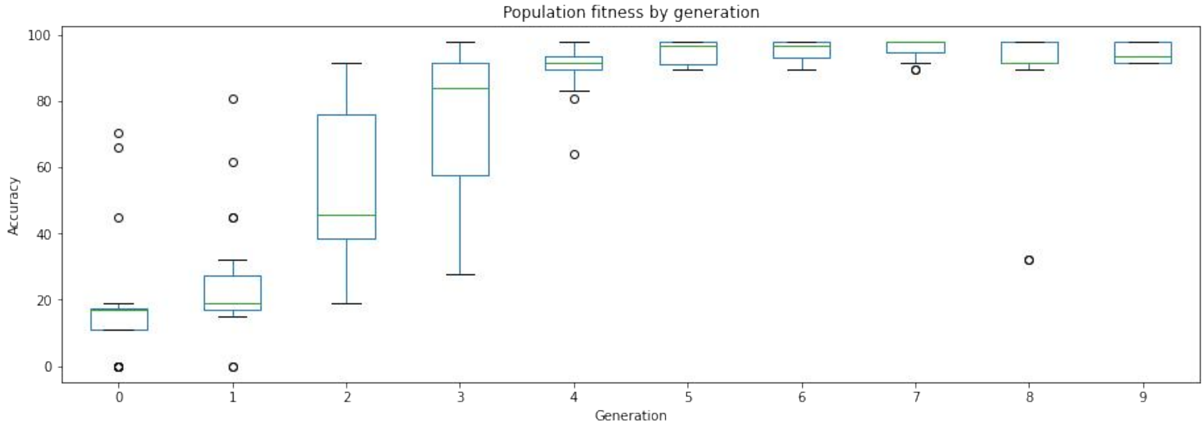


Figure 16 – Population fitness over generation for JAFFe dataset

Figure 17 – Confusion Matrix JAFFe

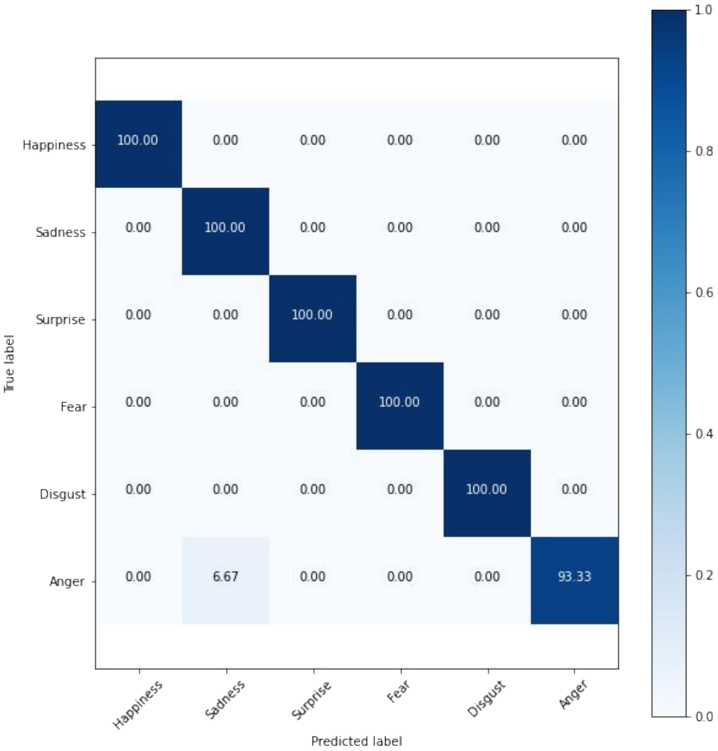


figure 17 we demonstrate the confusion matrix for the JAFFe dataset, in which we see only one prediction error for the anger emotion. This prediction error can be due to a lot of muscle contraction that this emotion exerts. The emotion of sadness is easily recognized by humans when noticing micro-expressions on the lips, while anger can occur muscle contraction around the lips, but it has its differentiation in the muscle contraction around the eyes.

In this experiment, it is already possible to say that using the metaheuristic just changing the input and output of the neural network, since the images have different sizes and the number of classes are different, works for different bases.

5.3.3 AffectNet

We performed two experiments using AffectNet dataset. The first one follows all the instructions described at the beginning of the section and uses 50 maximum generations for each execution. In the second experiment, 4000 images of each emotion and 250 images were randomly separated for training and validation, then a fourth dataset was created, which is the junction of the three datasets presented in this work.

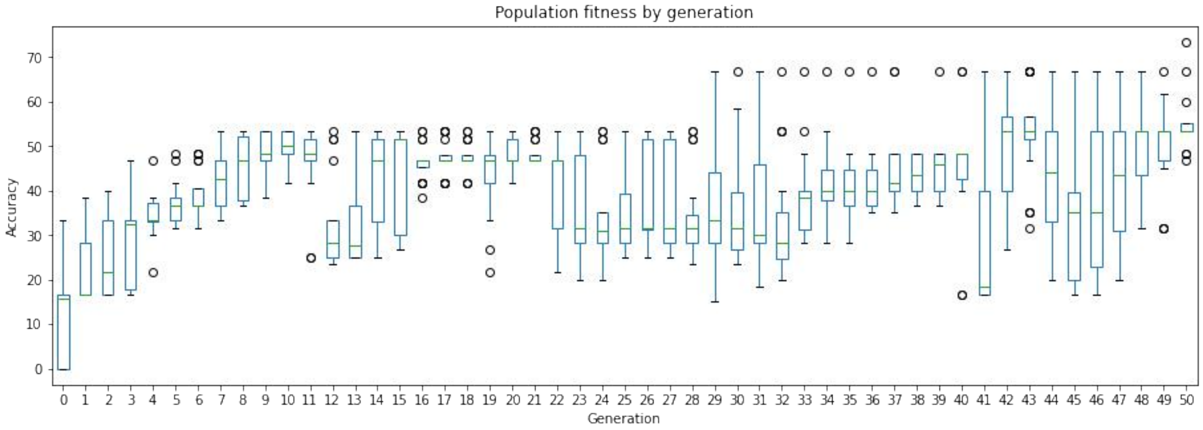


Figure 18 – Population fitness over generation for AffectNet dataset

Unlike the experiments done with JAFFe and CK+, the population did not quickly converge to an optimal value. Despite this, there was a convergence to the optimal value, as shown in the example in Figure 18. In this example, it is seen that in generation 50, the best individual has an accuracy of 73.3%. As in the experiments for the laboratory datasets, the GA was run a total of 5 times. For AffectNet, an accuracy of 68.8% on average was obtained.

In Figure 19, we present the confusion matrix of one of the results to AffectNet. As expected from the literature on facial expressions, the most common confusion was between surprise and fear and the least common between happiness and sadness.

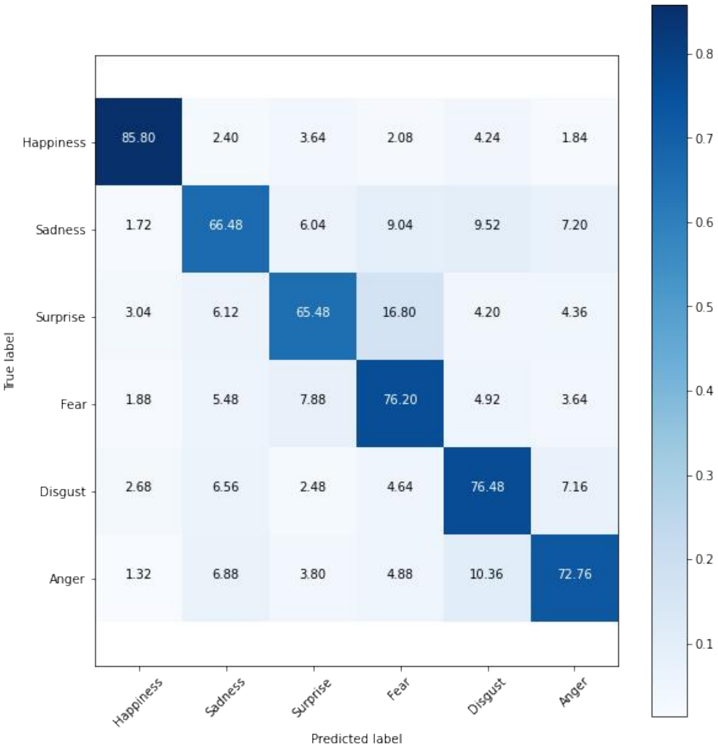
For the second experiment using AffectNet, the GA executed only twice. However, 80 maximum generations were set for these runs. In both runs, the solution quickly converges to values above 80% and holds for generations before achieving a small improvement. The example shown in Figure 20, demonstrates this behavior, at the end of this execution, the accuracy obtained for the best individual was 87%.

5.3.4 FER Benchmarking

In order to evaluate the results of the experiments, 6 state-of-the-art works that use the same amount of emotions were separated, being 2 works for each of the data sets.

As shown in Table 10, the metaheuristic presented in this paper achieves the state-of-the-art for facial expression recognition. Which demonstrates the effectiveness of a simple

Figure 19 – Confusion Matrix AffectNet



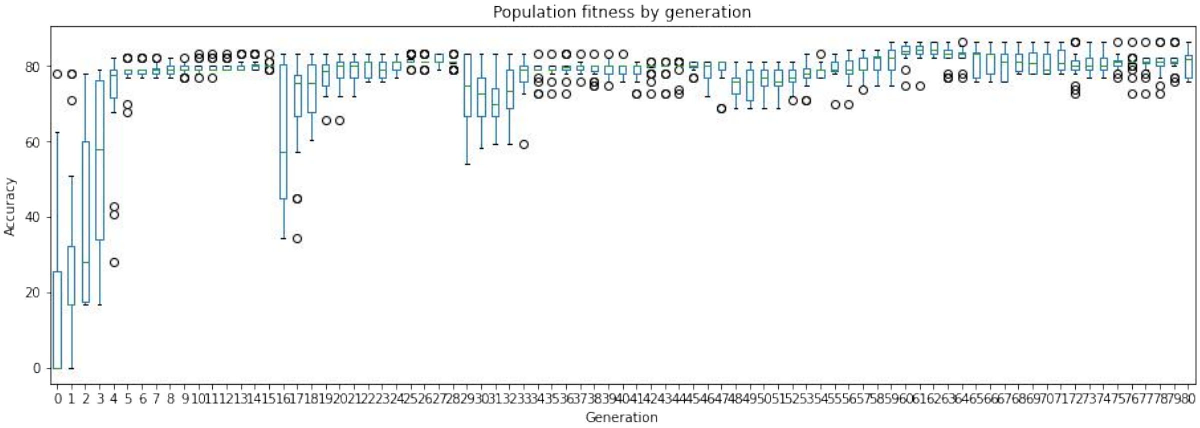


Figure 20 – Population fitness over generation for the combined dataset

evolutionary process. This evolutionary process allows obtaining efficient results for both laboratory and non-laboratory images.

The genetic algorithm found CNN architectures that surpassed the state-of-the-art accuracy for JAFFe and AffectNet. Our GA found an architecture with a performance of 98.8% for CK+, but on average the executions had a performance of 96.4%.

The CNN architectures found were not very deep. The best results varied between architectures with depth between 5 and 8 layers. And a maximum of 3 layers in the dense network.

Table 10 – State-of-the-Art accuracy with six prototypical facial expressions.

	JAFFe	CK+	AffectNet
MCCNN (HAMESTER; BARROS; WERMTER, 2015)	95.8%		
BDBN (LIU et al., 2014)	91.8%		
Zhang *et al* (ZHANG et al., 2016)		98.9%	
Zero-bias CNN (KHORRAMI; PAINE; HUANG, 2017)		95.7%	
PAENet(HUNG et al., 2019b)			65.2%
CAKE (KERVADEC et al., 2018)			61.7%
VGG / ResNet	94%	94%	59.5%
Genetic Approach (Our Approach)	97.6%	96.4%	68.8%

6 CONCLUSIONS

In this research, we investigate the performance of the genetic algorithm as a neural architectural search in Facial Expression Recognition area. While previous work using GA to find the best CNN architecture (XIE; YUILLE, 2017; SUN et al., 2020) decided to remove the fully encoded layer. The metaheuristic of this paper encodes it, so the depth, parameters or removal of this layer is decided by the evolutionary process.

In our bibliographic research in the field of psychology, we raised two important points. The first is that there is a consensus between at least 6 of the basic emotions, which are: happiness, surprise, fear, sadness, anger, and disgust. The second point is the facial technique, FACS, where each emotion can be performed in a more technical way. Studies and challenges in computing that use FACS are mostly related to semantic segmentation, e.g., the studies that Martinez et al. (2017) surveys treat FACS by demonstrating the occurrence of AUs and / or its intensity.

In continuity, we present some of the main databases. The databases that the images are photographed in laboratories have a much lower amount when compared to those that are not. And they all present the six basic emotions mentioned above. The most used databases are CK and JAFFe. During data analysis with the Affectnet dataset, we noticed that the dataset was formed by several duplicated images and many of them were misclassified. How can there be divergences between the opinions of researchers when classifying the emotion label, there needs to be an In-the-Wild dataset produced by experts using techniques such as FACS so that there are less divergences or none. This problem persists in other In-The-Wild datasets such as FER2013. It was also possible to observe the absence of In-The-Wild approaches that use large databases. The In-The-Wild condition means that the set of images is not laboratory-controlled. The face is closest to real-life situations, and only two studies considered experiments real-life benchmarks.

We have detected that there are still no approaches for facial expression recognition that apply an evolutionary strategy for the creation or alteration of the artificial neural network, including topology adaptation.

The S1(CABADA et al., 2020) and S2(LI et al., 2019) studies are the only ones that use convolutional neural networks, the S5(MEI; TAN; LIU, 2017) a Brain-Inspired Neural Networks and the others use multi-layer perceptron neural networks, showing that there is still a lot of space for research using more recent models of neural networks for classification of images.

It is important to note that studies using evolutionary strategies in the feature extraction stage have achieved great success in the accuracy of the classification (CHENGETA, 2019; Mistry et al., 2017; NEOH et al., 2015). And overall, with the exception of S8 (QIN; FANG; YANG, 2013), the accuracy of the studies was equal to or better than the state-of-the-art for CNN and was better than the traditional models of CNN as VGGNet and ResNet.

Returning the survey question:"Does a neuroevolutionary approach achieve better results for recognizing facial expressions through images when compared to other state-of-the-art methods that use neural networks?". Based on our results applying a neuroevolutionary strategy, the answer is affirmative. A NE approach achieves similar or superior results when compared to other Artificial Neural Networks. Thus,the alternative hypothesis: "NE approach achieves similar or superior results when compared to other Artificial Neural Networks" is satisfied by this thesis. Therefore, rejecting the null hypothesis.

The metaheuristic presented in this research was tested on three datasets that are among the main ones for facial expression recognition. Also, in a fourth dataset that is the combination of the three. A large amount of data with a large variety of faces is required for the metaheuristic to be non-exclusive.

For laboratory bases such as JAFFe and CK+, the GA quickly finds a CNN that achieves state-of-the-art, obtaining an average accuracy of 97.6% and 96.4%, respectively. By applying the GA to the AffectNet dataset, an average accuracy of 68.8% was achieved and a maximum accuracy of 73.3%, which surpasses state-of-the-art results. All networks are trained and built from scratch, without using transfer learning. For the experiments using the combination of the sets, the accuracy obtained was to 86.4%.

6.1 FUTURE RESEARCH

As future research, we suggest to expand the analysis, separating the parameters so that it is possible to verify the impact of each parameter on the solution. We also suggested to create a non-laboratory dataset that is balanced and carefully recorded on emotions, using techniques such as FACS (Facial Action Coding System) to reduce human error. Since these errors in annotations occur in several non-laboratory bases, such as FER2013 and AffectNet itself.

For the systematic literature review, we suggest two main further researches the creation of a non-laboratory database balanced and noted with emotions correctly, when analyzing the images using the FACS system, it is possible to verify that many of the facial images are in the wrong category. This occurs in many In-The-Wild datasets, *e.g.*, FER2013 and AffectNet. As well as the application of an evolutionary approach in the feature extraction stage using In-The-Wild datasets

BIBLIOGRAPHY

ALOM, M. Z. et al. The history began from alexnet: A comprehensive survey on deep learning approaches. **arXiv preprint arXiv:1803.01164**, 2018.

BACK, T. **Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms**. [S.l.]: Oxford university press, 1996.

BENITEZ-QUIROZ, C. F. et al. Emotionet challenge: Recognition of facial expressions of emotion in the wild. **CoRR**, abs/1703.01210, 2017.

CABADA, R. Z. et al. Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems. **SOFT COMPUTING**, v. 24, n. 10, p. 7593–7602, 2020.

CARLSON, J. G.; HATFIELD, E. **Psychology of emotion.** [S.l.]: Harcourt Brace Jovanovich, 1992.

CHANG, W.-L.; WANG, J.-Y. Mine is yours? using sentiment analysis to explore the degree of risk in the sharing economy. **Electronic Commerce Research and Applications**, v. 28, p. 141 – 158, 2018. ISSN 1567-4223.

CHENGETA, K. Enhanced Feature Selection for Facial Expression Recognition Systems with Genetic Algorithms. In: Macintyre, J and Iliadis, L and Maglogiannis, I and Jayne, C (Ed.). **Engineering Aplications of Neural Networks**. [S.l.: s.n.], 2019. (Communications in Computer and Information Science, v. 1000), p. 176–187. ISBN 978-3-030-20257-6; 978-3-030-20256-9. ISSN 1865-0929.

CORTIS, K. et al. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, p. 519–535, 2017. ISSN 0725-5136.

DARWIN, C. **The expression of the emotions in man and animals**. [S.l.]: JOHN MURRAY, London, 1872.

DAS, S. Virtual Classroom for Effective Learning in IT Industry. In: **2015 International Conference on Information Technology (ICIT)**. [S.l.]: IEEE, 2015. p. 221–226. ISBN 978-1-5090-0487-4.

EKMAN, P. Facial action coding system. Consultion Psychologists Press, 1977.

EKMAN, P. An argument for basic emotions. **Cognition and Emotion**, v. 6, n. 3-4, p. 169–200, may 1992. ISSN 0269-9931.

EKMAN, P. Are there basic emotions? American Psychological Association, 1992.

EKMAN, P. Emotions revealed. **BMJ**, British Medical Journal Publishing Group, v. 328, n. Suppl S5, p. 0405184, 2004.

EKMAN, P.; FRIESEN, W. Measuring facial movement. **Environmental Psychology and Nonverbal Behavior**, Kluwer Academic Publishers-Human Sciences Press, v. 1, n. 1, p. 56–75, 1976. ISSN 03613496.

EKMAN, P.; FRIESEN, W. Manual of the facial action coding system (facs). **Trans. ed. Vol. Consulting Psychologists Press, Palo Alto**, 1978.

EKMAN, P.; FRIESEN, W. V. Measuring facial movement. **Environmental psychology and nonverbal behavior**, Springer, v. 1, n. 1, p. 56–75, 1976.

EKMAN, P.; FRIESEN, W. V.; ELLSWORTH, P. **Emotion in the human face: Guidelines for research and an integration of findings**. [S.l.]: Elsevier, 2013. v. 11.

EKMAN, P.; SCHERER, K. R. **Handbook of methods in nonverbal behavior research**. [S.l.]: Cambridge University Press, 1982.

ELSKEN, T.; METZEN, J. H.; HUTTER, F. Neural architecture search. In: ____. **Automated Machine Learning: Methods, Systems, Challenges**. Cham: Springer International Publishing, 2019. p. 63–77. ISBN 978-3-030-05318-5.

FARSI, M.; MUNRO, M.; AL-THOBAITI, A. The effects of teaching primary school children the islamic prayer in a virtual environment. In: **2015 Science and Information Conference (SAI)**. [S.l.: s.n.], 2015. p. 765–769.

FRIESEN, W. V.; EKMAN, P. et al. Emfacs-7: Emotional facial action coding system. **Unpublished manuscript, University of California at San Francisco**, v. 2, n. 36, p. 1, 1983.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. Kluwer Academic Publishers-Plenum Publishers; Kluwer Academic Publishers ..., 1988.

HAMESTER, D.; BARROS, P.; WERMTER, S. Face expression recognition with a 2-channel convolutional neural network. In: **2015 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2015. p. 1–8.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HE, X.; ZHANG, W. Emotion recognition by assisted learning with convolutional neural networks. **Neurocomputing**, v. 291, p. 187 – 194, 2018. ISSN 0925-2312.

HJORTSJÖ, C.-H. **Man's face and mimic language**. [S.l.]: Studen litteratur, 1969.

HOLLAND, J. **Adaptation in natural and artificial systems, 1975**. [S.l.]: Univ. of Michigan Press. A. Kershenbaum, P. Kermani, GA Grover,"MENTOR: An ..., 1991.

HUNG, C.-Y. et al. Compacting, picking and growing for unforgetting continual learning. In: WALLACH, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2019. v. 32.

HUNG, S. C. Y. et al. Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. In: **Proceedings of the 2019 on International Conference on Multimedia Retrieval**. New York, NY, USA: Association for Computing Machinery, 2019. (ICMR '19), p. 339–343. ISBN 9781450367653.

JAIN, N. et al. Hybrid deep neural networks for face emotion recognition. **Pattern Recognition Letters**, Elsevier B.V., 2018. ISSN 01678655.

KANADE, T.; COHN, J. F.; TIAN, Y. Comprehensive database for facial expression analysis. In: **Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)**. [S.l.: s.n.], 2000. p. 46–53.

KASSAHUN, Y.; SOMMER, G. Efficient reinforcement learning through evolutionary acquisition of neural topologies. In: **ESANN**. [S.l.: s.n.], 2005.

KERVADEC, C. et al. CAKE: a compact and accurate k-dimensional representation of emotion. In: **British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018**. [S.l.]: BMVA Press, 2018. p. 316.

KHORRAMI, P.; PAINE, T. L.; HUANG, T. S. **Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?** 2017.

KINGMA, D.; BA, L. et al. Adam: A method for stochastic optimization. Ithaca, NYarXiv. org, 2015.

KITCHENHAM, B. Procedure for undertaking systematic reviews. **Computer Science Depart-ment, Keele University (TRISE-0401) and National ICT Australia Ltd (0400011T. 1), Joint Technical Report**, 2004.

KRIZHEVSKY et al. Learning multiple layers of features from tiny images. Citeseer, 2009.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). **Advances in Neural Information Processing Systems 25**. [S.l.]: Curran Associates, Inc., 2012. p. 1097–1105.

KUO, R.; ZULVIA, F. E. The gradient evolution algorithm: A new meta-heuristic. **Information Sciences**, v. 316, p. 246 – 265, 2015. ISSN 0020-0255. Nature-Inspired Algorithms for Large Scale Global Optimization. Disponível em: http://www.sciencedirect.com/science/article/pii/S0020025515002996.

LI, M. et al. Assisted therapeutic system based on reinforcement learning for children with autism. **COMPUTER ASSISTED SURGERY**, TAYLOR & FRANCIS LTD, 2-4 PARK SQUARE, MILTON PARK, ABINGDON OR14 4RN, OXON, ENGLAND, 2019.

LI, S.; DENG, W. Deep facial expression recognition: A survey. **IEEE Transactions on Affective Computing**, IEEE, 2020.

LI, S.; DENG, W.; DU, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 2852–2861.

LI, S.; DENG, W.; DU, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 2584–2593.

LI, X. et al. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. **IEEE Transactions on Affective Computing**, v. 9, n. 4, p. 563–577, 2018.

LIANG, T.-P. et al. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. **International Journal of Electronic Commerce**, Routledge, v. 20, n. 2, p. 236–260, 2015.

LIU, H. et al. Hierarchical representations for efficient architecture search. **CoRR**, abs/1711.00436, 2017. Disponível em: http://arxiv.org/abs/1711.00436.

LIU, P. et al. Facial expression recognition via a boosted deep belief network. In: **2014 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 1805–1812.

LUCEY, P. et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: . [S.l.: s.n.], 2010. p. 94 – 101.

LYONS, M. et al. Coding facial expressions with gabor wavelets. In: **Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition**. [S.l.: s.n.], 1998. p. 200–205.

MARTINEZ, B. et al. Automatic analysis of facial actions: A survey. **IEEE transactions on affective computing**, IEEE, 2017.

MAXWELL, J. Understanding and validity in qualitative research. **Harvard educational review**, Harvard Education Publishing Group, v. 62, n. 3, p. 279–301, 1992.

MEI, Y.; TAN, G.; LIU, Z. An improved brain-inspired emotional learning algorithm for fast classification. **Algorithms**, MDPI AG, v. 10, n. 2, 2017. ISSN 19994893.

METHLEY, A. M. et al. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. **BMC Health Serv Res**, v. 14, p. 579, Nov 2014.

MICHALEWICZ, Z. **Genetic algorithms+ data structures= evolution programs**. [S.l.]: Springer Science & Business Media, 2013.

MILLER, B. L.; GOLDBERG, D. E. et al. Genetic algorithms, tournament selection, and the effects of noise. **Complex systems**, [Champaign, IL, USA: Complex Systems Publications, Inc., c1987-, v. 9, n. 3, p. 193–212, 1995.

Mistry, K. et al. A micro-ga embedded pso feature selection approach to intelligent facial emotion recognition. **IEEE Transactions on Cybernetics**, v. 47, n. 6, p. 1496–1509, 2017.

MOLLAHOSSEINI, A.; HASSANI, B.; MAHOOR, M. H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. **CoRR**, abs/1708.03985, 2017. Disponível em: http://arxiv.org/abs/1708.03985.

NEOH, S. C. et al. Intelligent facial emotion recognition using a layered encoding cascade optimization model. **APPLIED SOFT COMPUTING**, ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS, v. 34, p. 72–93, 2015. ISSN 1568-4946.

Pantic, M.; Rothkrantz, L. J. M. Facial action recognition for facial expression analysis from static face images. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 34, n. 3, p. 1449–1461, 2004.

Pantic, M. et al. Web-based database for facial expression analysis. In: **2005 IEEE International Conference on Multimedia and Expo**. [S.l.: s.n.], 2005. p. 5 pp.–.

PANTIC, M. et al. Web-based database for facial expression analysis. In: **2005 IEEE International Conference on Multimedia and Expo**. [S.l.: s.n.], 2005. p. 5 pp.–. ISSN 1945-7871.

PETERSEN, K.; GENCEL, C. Worldviews, research methods, and their relationship to validity in empirical software engineering research. In: **2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement**. [S.l.: s.n.], 2013. p. 81–89.

PITALOKA, D. A. et al. Enhancing cnn with preprocessing stage in automatic emotion recognition. **Procedia Computer Science**, v. 116, p. 523 – 529, 2017. ISSN 1877-0509. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

QIN, W.; FANG, Q.; YANG, Y. A facial expression recognition method based on singular value features and improved BP neural network. **Communications in Computer and Information Science**, Springer Verlag, v. 363, p. 163–172, 2013. ISSN 18650929.

REVINA, I.; EMMANUEL, W. S. A survey on human face expression recognition techniques. **Journal of King Saud University - Computer and Information Sciences**, 2018. ISSN 1319-1578. Disponível em: http://www.sciencedirect.com/science/article/pii/S1319157818303379.

RISI, S.; STANLEY, K. O. An enhanced hypercube-based encoding for evolving the placement, density, and connectivity of neurons. **Artificial Life**, v. 18, n. 4, p. 331–363, 2012. PMID: 22938563. Disponível em: https://doi.org/10.1162/ARTL_a_00071.

Rodriguez, P. et al. Deep pain: Exploiting long short-term memory networks for facial expression classification. **IEEE Transactions on Cybernetics**, p. 1–11, 2017. ISSN 2168-2267.

RUSSELL, J. A circumplex model of affect. **Journal of Personality and Social Psychology**, v. 39, p. 1161–1178, 12 1980.

RUSSELL, J. A.; DOLS, J. M. F. **The psychology of facial expression**. [S.l.]: Cambridge university press Cambridge, 1997. v. 131.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural networks**, Elsevier, v. 61, p. 85–117, 2015.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

STANLEY, K. O.; D'AMBROSIO, D. B.; GAUCI, J. A hypercube-based encoding for evolving large-scale neural networks. **Artificial Life**, v. 15, n. 2, p. 185–212, 2009. PMID: 19199382. Disponível em: https://doi.org/10.1162/artl.2009.15.2.15202.

STANLEY, K. O.; MIIKKULAINEN, R. Evolving neural networks through augmenting topologies. **Evol. Comput.**, MIT Press, Cambridge, MA, USA, v. 10, n. 2, p. 99–127, jun. 2002. ISSN 1063-6560. Disponível em: http://dx.doi.org/10.1162/106365602320169811.

SUCH, F. P. et al. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. **CoRR**, abs/1712.06567, 2017. Disponível em: http://arxiv.org/abs/1712.06567.

SUGANUMA, M.; SHIRAKAWA, S.; NAGAO, T. A genetic programming approach to designing convolutional neural network architectures. In: **Proceedings of the Genetic and Evolutionary Computation Conference**. [S.l.: s.n.], 2017. p. 497–504.

SUN, Y. et al. Automatically designing cnn architectures using the genetic algorithm for image classification. **IEEE transactions on cybernetics**, IEEE, v. 50, n. 9, p. 3840–3854, 2020.

SUWA, M. A preliminary note on pattern recognition of human emotional expression. **Proc. of The 4th International Joint Conference on Pattern Recognition**, p. 408–410, 1978. Disponível em: https://ci.nii.ac.jp/naid/10006751528/en/.

SZE, V. et al. Efficient processing of deep neural networks: A tutorial and survey. **CoRR**, abs/1703.09039, 2017. Disponível em: http://arxiv.org/abs/1703.09039.

TEALAB, A. Time series forecasting using artificial neural networks methodologies: A systematic review. **Future Computing and Informatics Journal**, Elsevier, v. 3, n. 2, p. 334–340, 2018.

VIOLA, P.; JONES, M. J. Robust real-time face detection. **International journal of computer vision**, Springer, v. 57, n. 2, p. 137–154, 2004.

WANG, S.-H. et al. Facial emotion recognition via discrete wavelet transform, principal component analysis, and cat swarm optimization. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Springer Verlag, v. 10559 LNCS, p. 203–214, 2017. ISSN 03029743.

WHITLEY, D. A genetic algorithm tutorial. **Statistics and computing**, Springer, v. 4, n. 2, p. 65–85, 1994.

XIE, L.; YUILLE, A. L. Genetic CNN. **CoRR**, abs/1703.01513, 2017. Disponível em: http://arxiv.org/abs/1703.01513.

YANG, D. et al. An emotion recognition model based on facial recognition in virtual learning environment. **Procedia Computer Science**, v. 125, p. 2 – 10, 2018. ISSN 1877-0509. The 6th International Conference on Smart Computing and Communications.

YANG, H. et al. Real-time emotion recognition framework based on convolution neural network. **Smart Innovation, Systems and Technologies**, v. 157, p. 313–321, 2020.

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. **Computer Vision and Image Understanding**, v. 138, p. 1 – 24, 2015. ISSN 1077-3142. Disponível em: http://www.sciencedirect.com/science/article/pii/S1077314215000727.

ZARINS, U. **Anatomy of Facial Expression**. Exonicus Incorporated, 2017. ISBN 9780990341116. Disponível em: https://books.google.com.br/books?id=CqMyngAACAAJ.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. **European conference on computer vision**. [S.l.], 2014. p. 818–833.

ZHANG, L. et al. Facial expression analysis under partial occlusion: A survey. **CoRR**, abs/1802.08784, 2018. Disponível em: http://arxiv.org/abs/1802.08784.

ZHANG, Z. et al. From facial expression recognition to interpersonal relation prediction. **CoRR**, abs/1609.06426, 2016.

ZOPH, B.; LE, Q. V. Neural architecture search with reinforcement learning. **arXiv preprint arXiv:1611.01578**, 2016.