How May Deep Learning Testing Inform Model Generalizability? The Case of Image Classification

Giammaria Giordano, Valeria Pontillo, **Giusy Annunziata**, Antonio Cimino, Filomena Ferrucci, Fabio Palomba

University of Salerno (Italy)
Department of Computer Science
Software Engineering (SeSa) Lab

https://giusyann.github.io/

@Giusy_A_

gannunziata@unisa.it

Background

Artificial Intelligence (**AI**) systems are becoming increasingly popular due to their many uses, especially when it comes to ML-intensive IoT systems.

However, AI systems are extremely complex to implement. In response to this challenge, the software engineering research community has come into play through the definition of novel methods and instruments to engineering ML-intensive systems effectively and efficiently: this research field is known as **SE4AI**. [2]

(1) G. Giordano, F. Palomba, F. Ferrucci, On the use of artificial intelligence to deal with privacy in IoT systems: A systematic literature review, Journal of Systems and Software 193 (2022)
(2) E. Nascimento, A. Nguyen-Duc, I. Sundbø, T. Conte, Software engineering for artificial intelligence and machine learning software: A systematic literature review

Background

One of the most popular use cases of ML-intensive systems is represented by **Image Classification,**[1] which is instrumental for a large variety of real-world tasks, like video surveillance and facial recognition, just to name a few.

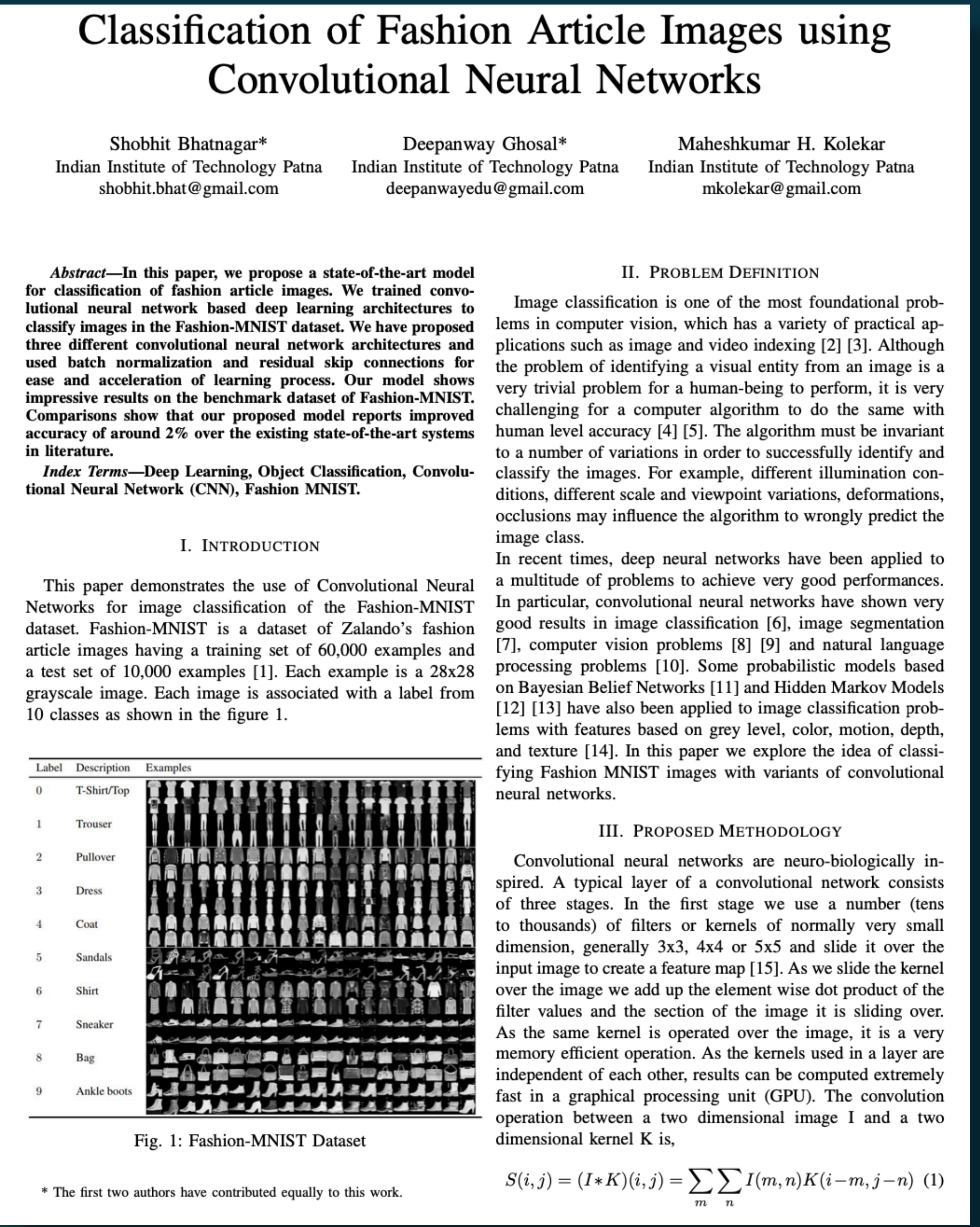
We argue that those systems are of interest for **SE4AI** as well, as they enclose critical SE properties, e.g., robustness, privacy, fairness, security, and performance.

(1) G. Giordano, F. Palomba, F. Ferrucci, On the use of artificial intelligence to deal with privacy in IoT systems: A systematic literature review, Journal of Systems and Software 193 (2022)
(2) E. Nascimento, A. Nguyen-Duc, I. Sundbø, T. Conte, Software engineering for artificial intelligence and machine learning software: A systematic literature review

Background

The work proposed three different Convolutional Neural Network Architectures using batch Normalization and Residual Skipped Connections.

Results: They achieved 92.54% accuracy using a two-layer Convolutional Neural Network with batch normalization and skipped connections.



Classification of Fashion Article Images using Convolutional Neural Networks

Shobhit Bhatnagar*
Indian Institute of Technology Patna
shobhit.bhat@gmail.com

Deepanway Ghosal*
Indian Institute of Technology Patna
deepanwayedu@gmail.com

Maheshkumar H. Kolekar
Indian Institute of Technology Patna
mkolekar@gmail.com

Abstract—**In this paper, we propose a state-of-the-art model for classification of fashion article images. We trained convolutional neural network based deep learning architectures to classify images in the Fashion-MNIST dataset. We have proposed three different convolutional neural network architectures and used batch normalization and residual skip connections for ease and acceleration of learning process. Our model shows impressive results on the benchmark dataset of Fashion-MNIST. Comparisons show that our proposed model reports improved accuracy of around 2% over the existing state-of-the-art systems in literature.**

Index Terms—**Deep Learning, Object Classification, Convolutional Neural Network (CNN), Fashion MNIST.**

I. Introduction

This paper demonstrates the use of Convolutional Neural Networks for image classification of the Fashion-MNIST dataset. Fashion-MNIST is a dataset of Zalando's fashion article images having a training set of 60,000 examples and a test set of 10,000 examples [1]. Each example is a 28x28 grayscale image. Each image is associated with a label from 10 classes as shown in the figure 1.

Label	Description	Examples
0	T-Shirt/Top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandals	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boots	

Fig. 1: Fashion-MNIST Dataset

* The first two authors have contributed equally to this work.

II. Problem Definition

Image classification is one of the most foundational problems in computer vision, which has a variety of practical applications such as image and video indexing [2] [3]. Although the problem of identifying a visual entity from an image is a very trivial problem for a human-being to perform, it is very challenging for a computer algorithm to do the same with human level accuracy [4] [5]. The algorithm must be invariant to a number of variations in order to successfully identify and classify the images. For example, different illumination conditions, different scale and viewpoint variations, deformations, occlusions may influence the algorithm to wrongly predict the image class.
In recent times, deep neural networks have been applied to a multitude of problems to achieve very good performances. In particular, convolutional neural networks have shown very good results in image classification [6], image segmentation [7], computer vision problems [8] [9] and natural language processing problems [10]. Some probabilistic models based on Bayesian Belief Networks [11] and Hidden Markov Models [12] [13] have also been applied to image classification problems with features based on grey level, color, motion, depth, and texture [14]. In this paper we explore the idea of classifying Fashion MNIST images with variants of convolutional neural networks.

III. Proposed Methodology

Convolutional neural networks are neuro-biologically inspired. A typical layer of a convolutional network consists of three stages. In the first stage we use a number (tens to thousands) of filters or kernels of normally very small dimension, generally 3x3, 4x4 or 5x5 and slide it over the input image to create a feature map [15]. As we slide the kernel over the image we add up the element wise dot product of the filter values and the section of the image it is sliding over. As the same kernel is operated over the image, it is a very memory efficient operation. As the kernels used in a layer are independent of each other, results can be computed extremely fast in a graphical processing unit (GPU). The convolution operation between a two dimensional image I and a two dimensional kernel K is,

$$S(i,j) = (I*K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (1)$$

Classification of Garments from Fashion MNIST Dataset Using CNN LeNet-5 Architecture

Mohammed Kayed
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-suef, Egypt, 62511
mskayed@gmail.com

Ahmed Anter
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-suef, Egypt, 62511
sw_anter@yahoo.com

Hadeer Mohamed
Faculty of Science, Beni-Suef University, Beni-suef, Egypt, 62511
hadeer.mohamed98@ymail.com

ABSTRACT

Recently, deep learning has been used extensively in a wide range of domains. A class of deep neural networks that give the most rigorous effects in solving real-world problems is a Convolutional Neural Network (CNN). Fashion businesses have used CNN on their e-commerce to solve many problems such as clothes recognition, clothes search and recommendation. A core step for all of these implementations is image classification. However, clothes classification is a challenge task as clothes have many properties, and the depth of clothes categorization is highly complicated. This complicated depth makes different classes to have very similar features, and so the classification problem becomes very hard. In this paper, CNN based LeNet-5 architecture is proposed to train parameters of the CNN on Fashion MNIST dataset. Experimental results show that LeNet-5 model achieved accuracy over 98%. Therefore, it outperforms both the classical CNN model and the other existing state-of-the-art models in literatures.

Keywords

Deep learning architectures; Fashion MNIST; Fashion Classification; Convolutional Neural Network (CNN); LeNet-5.

1. INTRODUCTION

Over past few years, with the assistance of various layers, deep learning [1] has been widely used and achieved very good results in different domains such as computer vision [2], big data [3], automatic speech recognition [4] and natural language processing [5]. A common architecture of deep neural networks is CNN. CNN is a multi-layer perceptron neural network that extracts properties from the input data and is trained with the neural network back-propagation algorithm. CNN can learn complex, high-dimensional, non-linear mappings from a very large number of data (images). Moreover, CNN gives an excellent classification average for images [6]. The main advantages of CNN are that it extracts the salient features that are never changed, and it is invariant to shifting, scaling and distortions of input data (images). CNN based LeNet-5 architecture has shown very good results in many domains such as image classification [7], pattern recognition [8], computer vision [9] and image segmentation.

One of the most challenging multi-classes classification problems is fashion classification in which labels that characterize the clothes type are assigned to the images. The difficulty of this multi-classes fashion classification problem is due to the richness of the clothes properties and the high depth of clothes categorization as well. This complicated depth makes different labels/classes to have similar features. This paper tries to enhance the performance of the fashion classification problem on the Fashion-MNIST Dataset [10], which contains 70,000 images (each image is labeled from the 10 categories shown in Figure 1: T-shirt/top, Trousers, Pullover, Dress, Coat, Sandals, Shirt, Sneaker, Bag and Ankle boot).

There are some issues to consider in classification of fashion [11]. First, garments can be easily distorted by lengthening pattern. Second, some garments might be considered as various according to the opinion, and various garments might be considered as same. Third, some garment items are robust to be recovered due to their small size. Fourth, photos can be taken in various cases such as the difference in the angle, light and noise backgrounds. Fifth, some garment classes have similar features and can be fuzzy, such as trouser and tights. Sixth, a garment image is different based on whether it is just a photo of a garment or a photo of the model's wearing garment. Therefore, an algorithm that could be used to get high multi-classes fashion classification performance is of great necessity. As well as this paper gives a brief review of the different CNN models for the classification of the Fashion-MNIST, the major contribution of this paper is that the multi-classes fashion classification problem will be solved by the CNN based LeNet-5 architecture. To the best of our knowledge, this model is not used before for this common MNIST dataset.

The rest of the paper is organized as follows: Section 2 gives a review of the related works. Section 3 describes the used dataset and methodology. Section 4 presents the proposed model. Section 5 presents the experiments and the classification results, while Section 6 concludes our work.

2. RELATED WORKS

Deep learning and CNN have been fully surveyed in [12]. Many CNN architectures have been used in image classification: LeNet [13], Alex Net [14], Google Net [15], VGGNet [16] and ResNet [17]. All of these architectures compete to correctly classifying and detecting images. Neural networks have also been applied to metrics learning with applications in image similarity estimation and visual search. Recently, two datasets have been published. MNIST [18] and Fashion-MNIST datasets for image classification [19] with 70,0000 annotated real-life images. In this section, we shall briefly review the works done on the Fashion-MNIST dataset as follows.

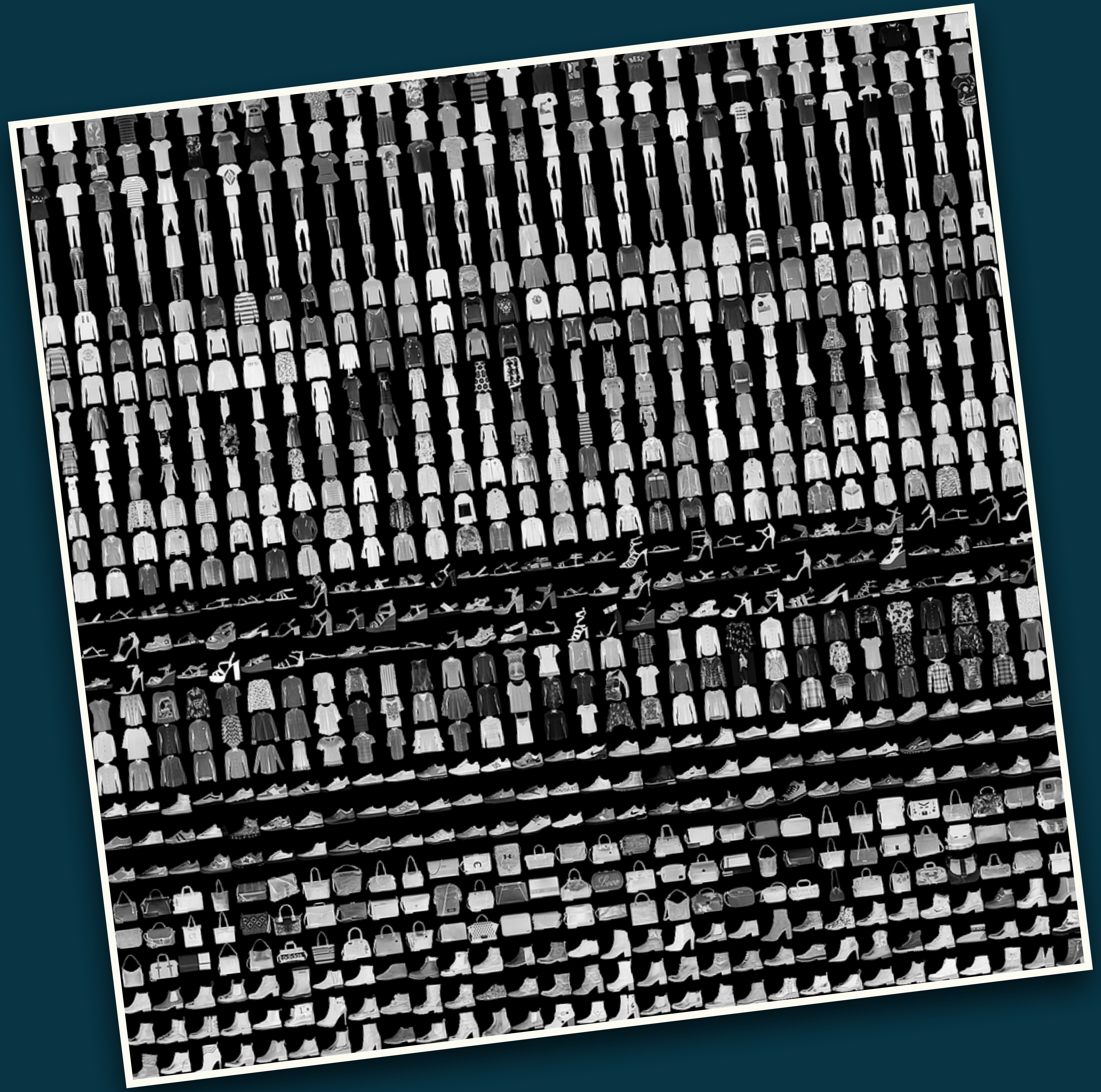
Shobhit et al. [20] proposed a model for classification of fashion article images. Convolutional neural network based deep learning architectures are trained to classify images of the Fashion-MNIST dataset. Also, three different CNN architectures used batch normalization and residual skip connections are suggested to accelerate the learning process. The results showed that the

The aim of the work is to improve Convolutional Neural Network's performance by leveraging a LeNet-5 architecture

Results: By enhancing the performance of Convolutional Neural Network by leveraging a LeNet-5 architecture, 98% accuracy can be achieved.

(3) S. Bhatnagar, D. Ghosal, M. H. Kolekar, Classification of fashion article images using convolutional neural networks, in: 2017 Fourth International Conference on Image Information Processing (ICIIP), IEEE, 2017
(4) M. Kayed, A. Anter, H. Mohamed, Classification of garments from fashion mnist dataset using cnn lenet-5 architecture, in: 2020 international conference on innovative trends in communication and computer engineering (ITCE), IEEE, 2020

Background



Most of the research conducted on image classification has been based on the use of the Fashion-MNIST dataset.[5] Why?

- Instances normalized in a dimension of 28×28 pixels
- Images converted into a gray scale
- Pixels composed of a different value (0-255) based on the color intensity
- 60,000 items divided into 10 classes of garments

As such, the dataset is widely considered as a valuable benchmark to experiment with image classification.

(5) The Fashion-MNIST dataset: https://github.com/zalandoresearch/fashion-mnist

Limitations

In-vitro experiments

Accuracy as a performance metric

Limitations

In-vitro experiments

- It is still unclear how robust ML-intensive systems trained on the Fashion-MNIST dataset are in a more realistic context, i.e., are the conclusions drawn generalizable and robust?
- Why does it matter? In an evolutionary context, data drift and data distorsions can occur and, as such, the performance of the model may significantly vary

Limitations

- Accuracy does not take into account the distribution of training and test sets and may be distorted due to the learning effect

- Why does it matter? It is not an appropriate measure for unbalanced data sets because it does not distinguish values such as false positive and false negative, possibly biasing the interpretation of the results

Accuracy as a performance metric

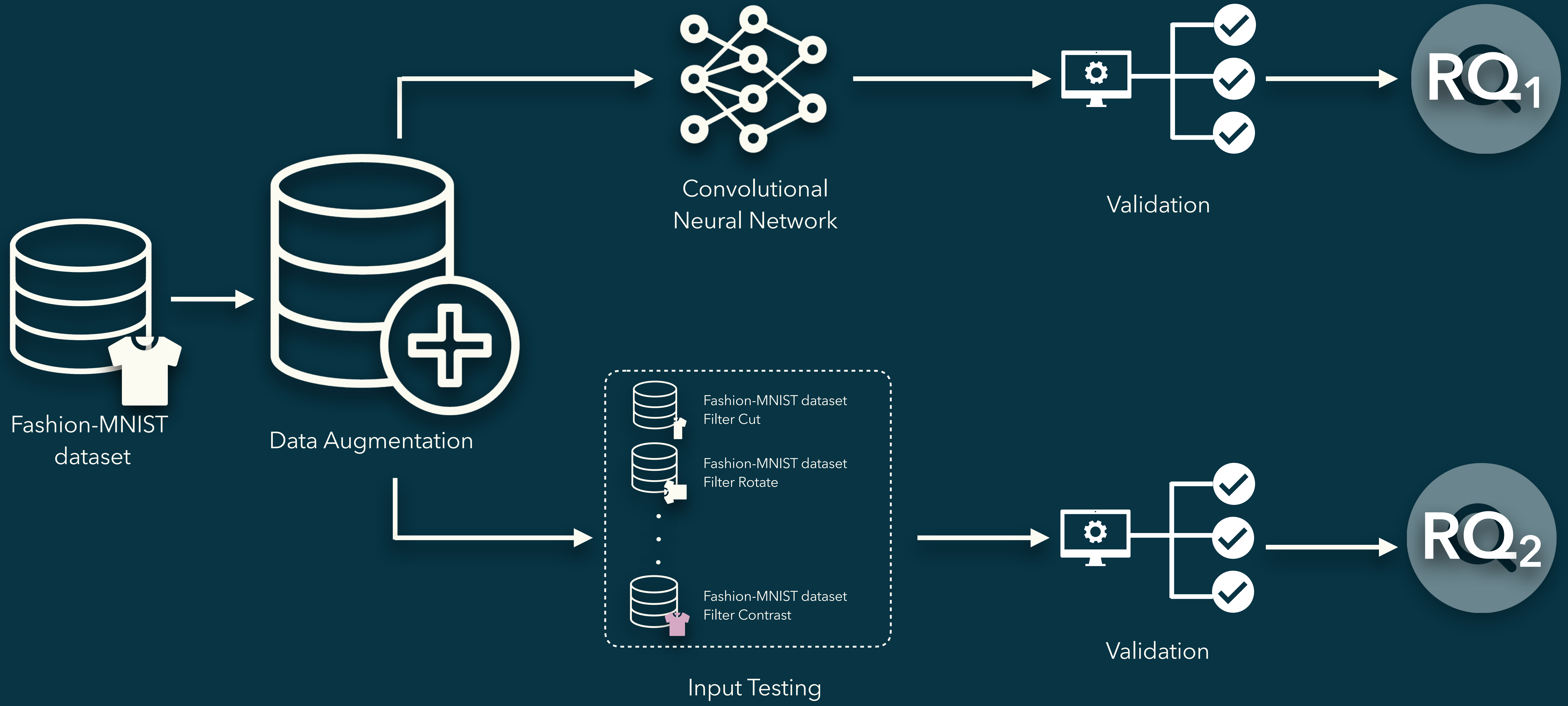
Assessing the robustness of systems simulating a more realistic evaluation scenario

Research Questions

RQ$_1$ ***Baseline.*** *What is the performance of an* ***engineered Convolution Neural Network*** *when applied for the task of image recognition?*

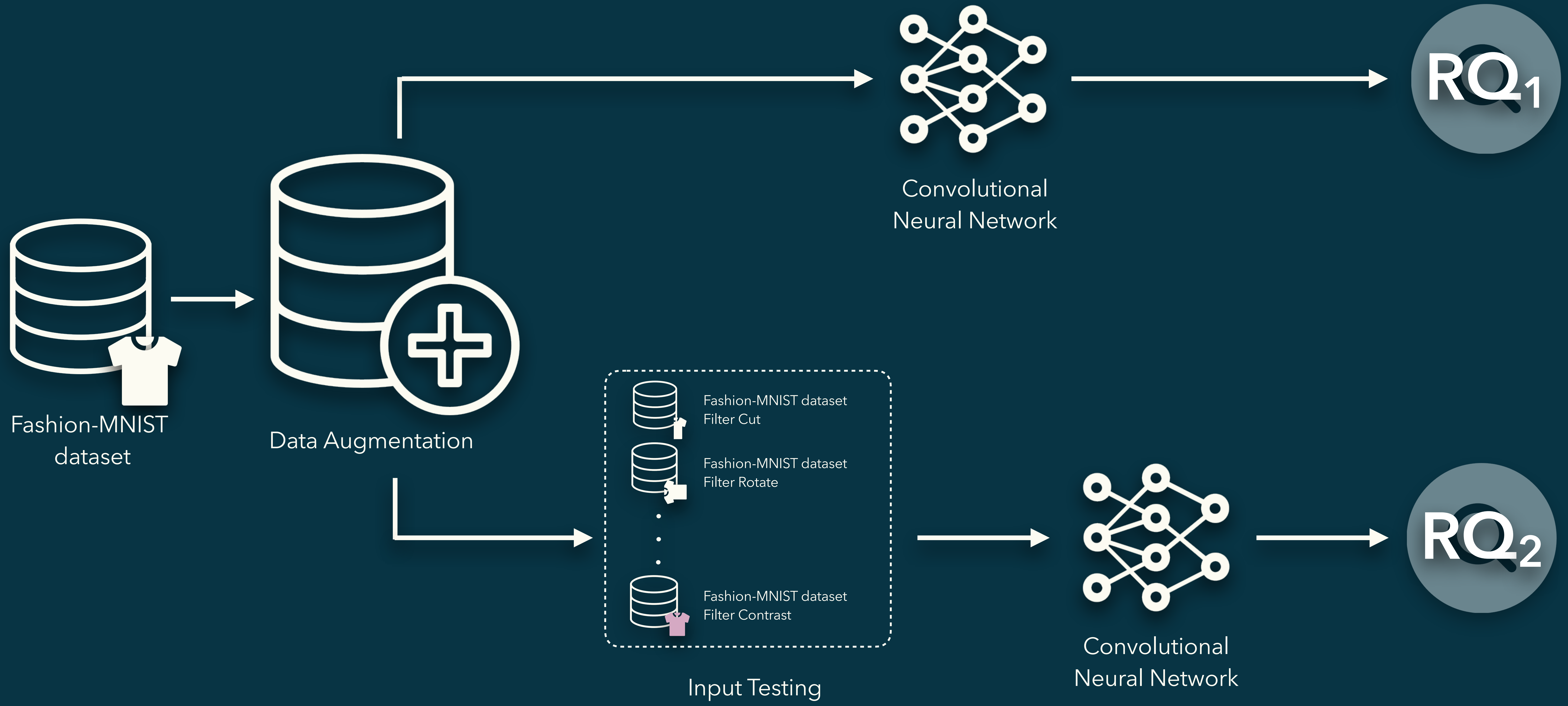
RQ$_2$ ***Goal.*** *To what extent the application of* ***input testing methods*** *impact the performance of an engineered Convolution Neural Network when applied for the task of image recognition?*

Research Method



Input Testing

Research Method



Results

Results of Existing Approach

91%

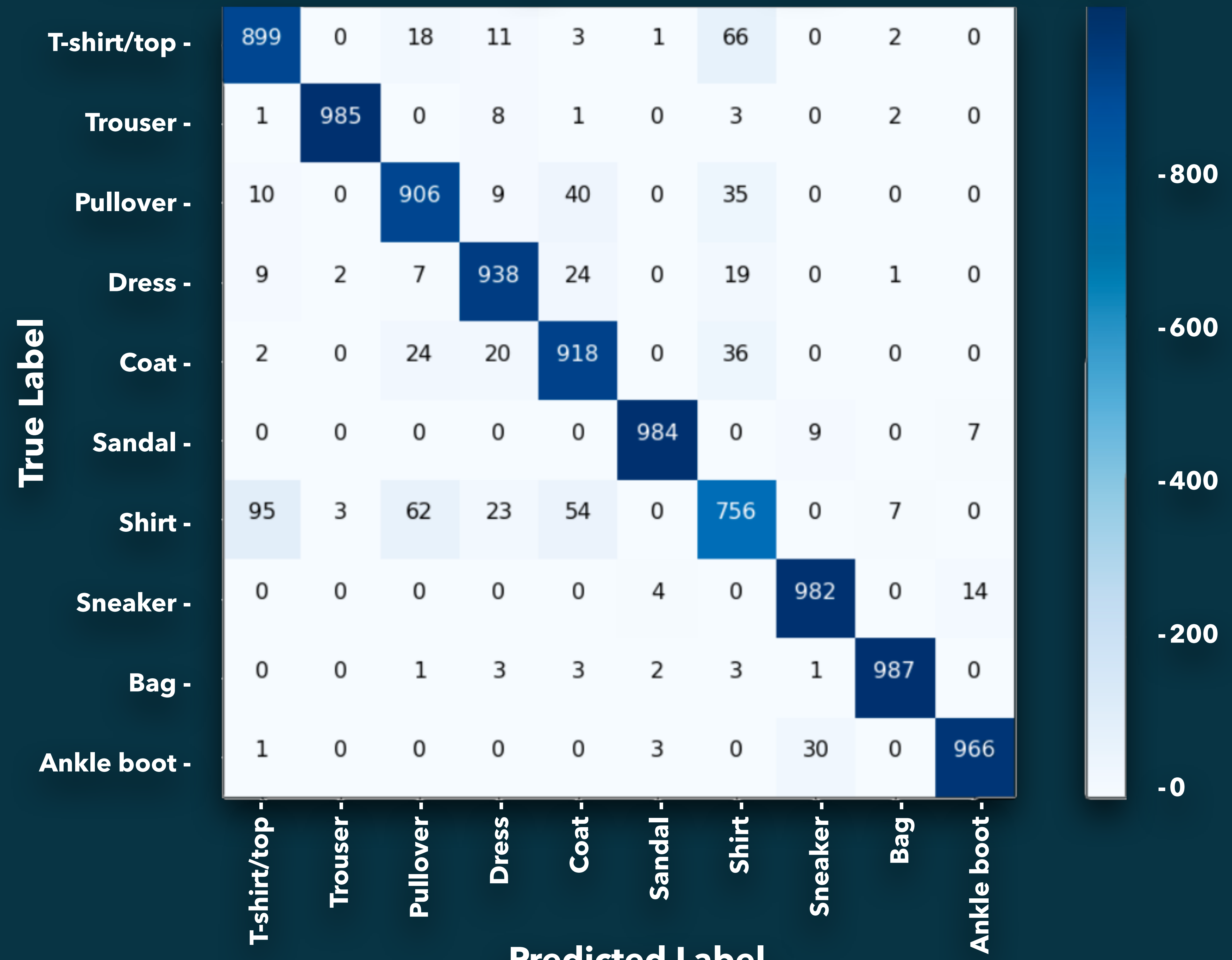
Precision

Recall

F-Measure

Accuracy

Results of the Engineered Approach – RQ_1



93%

Precision

Recall

F-Measure

Accuracy

Results

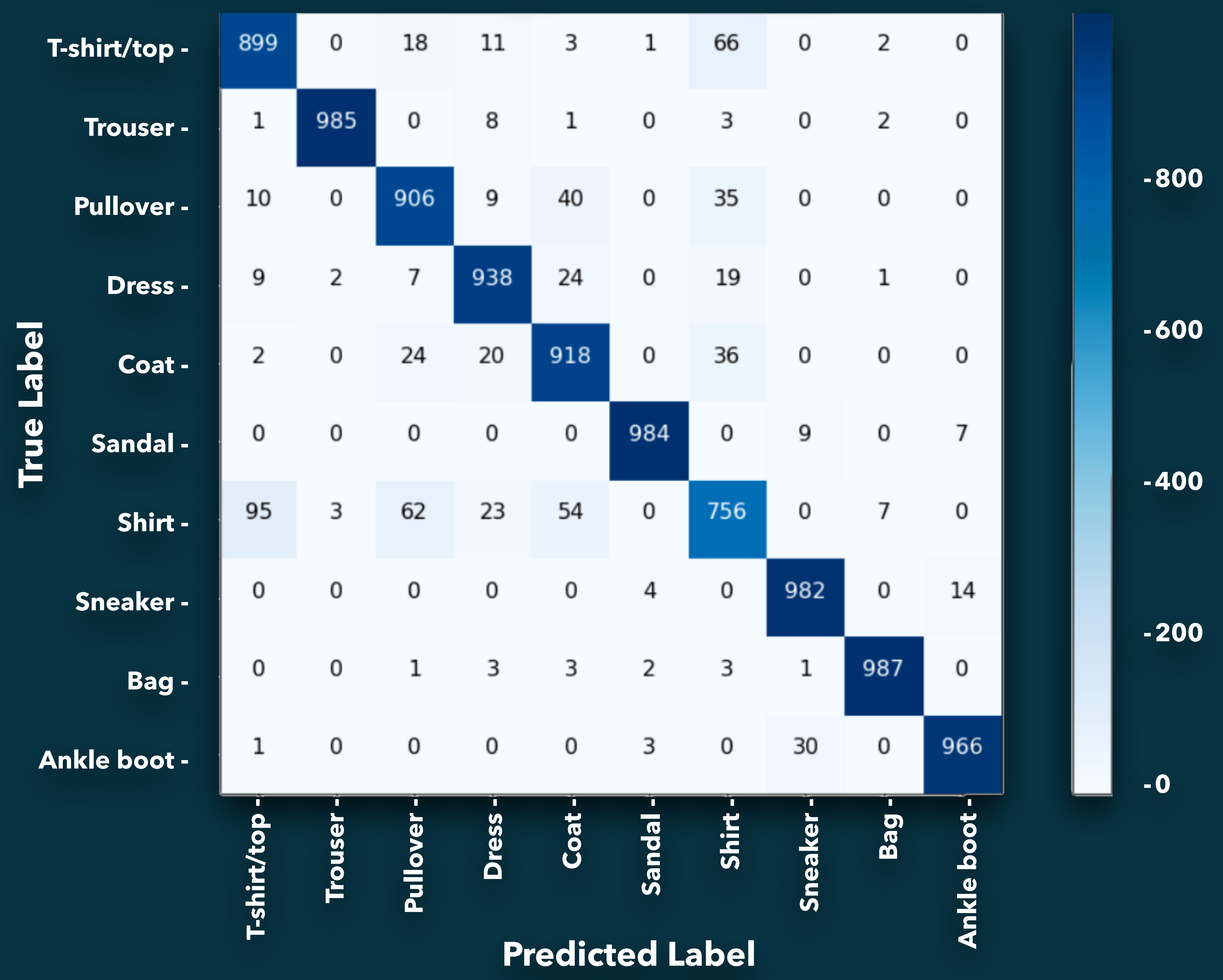
Results of Existing Approach

- The engineered approach achieved good levels of prediction for all garments, ranging from 89.9 percent to 98.7 percent.
- One out of four items is misclassified as a shirt, but is misclassified as a T-shirt or coat: this is probably because the three clothing classes are similar to each other.

91%

Precision
Recall
F-Measure
Accuracy

Results of the Engineered Approach – RQ_1

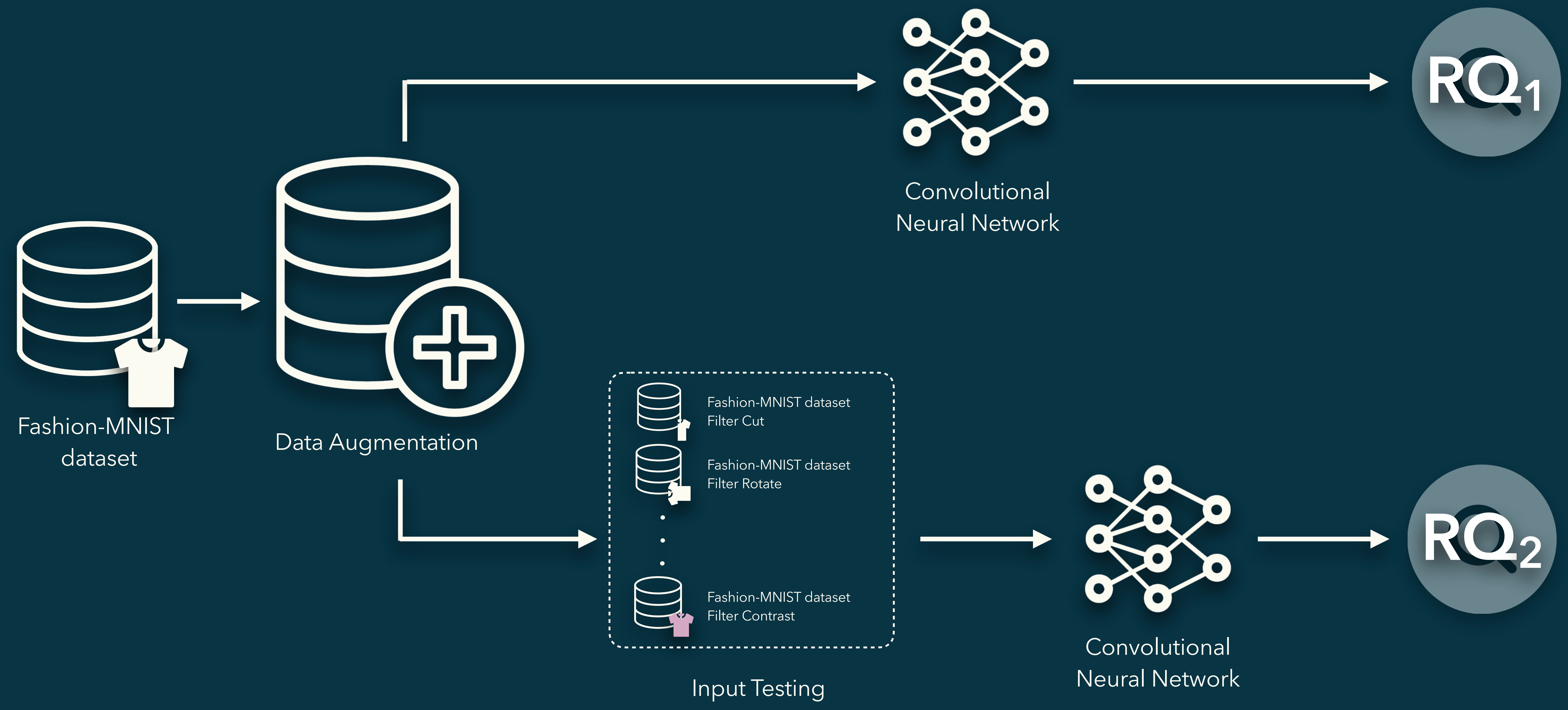


True Label \ Predicted Label	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
T-shirt/top	899	0	18	11	3	1	66	0	2	0
Trouser	1	985	0	8	1	0	3	0	2	0
Pullover	10	0	906	9	40	0	35	0	0	0
Dress	9	2	7	938	24	0	19	0	1	0
Coat	2	0	24	20	918	0	36	0	0	0
Sandal	0	0	0	0	0	984	0	9	0	7
Shirt	95	3	62	23	54	0	756	0	7	0
Sneaker	0	0	0	0	0	4	0	982	0	14
Bag	0	0	1	3	3	2	3	1	987	0
Ankle boot	1	0	0	0	0	3	0	30	0	966

93%

Precision
Recall
F-Measure
Accuracy

Research Method



Results

Result RQ_1

93%

Precision
Recall
F-Measure
Accuracy

Result RQ_2

Filter Cut	33% (- 60%)
Filter Rotate	14% (- 79%)
Filter Occlusion	60% (- 39%)
Filter Contrast	85% (- 8%)
Filter Blur	59% (- 34%)
Filter Rain	72% (- 21%)

Precision

Results

Result RQ_1

93%

Precision
Recall
F-Measure
Accuracy

Result RQ_2

Filter Cut	31% (- 62%)
Filter Rotate	9% (- 84%)
Filter Occlusion	50% (- 43%)
Filter Contrast	84% (- 9%)
Filter Blur	52% (- 41%)
Filter Rain	58% (- 35%)

Recall

Results

Result RQ_1

93%

Precision
Recall
F-Measure
Accuracy

Result RQ_2

Filter Cut	27% (- 66%)
Filter Rotate	7% (- 86%)
Filter Occlusion	49% (- 44%)
Filter Contrast	84% (- 9%)
Filter Blur	46% (- 47%)
Filter Rain	58% (- 35%)

F-Measure

Results

Result RQ_1

93%

Precision
Recall
F-Measure
Accuracy

Result RQ_2

Filter Cut	31% (- 62%)
Filter Rotate	9% (- 84%)
Filter Occlusion	50% (- 43%)
Filter Contrast	84% (- 9%)
Filter Blur	52% (- 41%)
Filter Rain	58% (- 35%)

Accuracy

When replicating in a real-world context, model performance drops dramatically

Considerations

1. Even engineered models fail with data that deviate from the training data

2. The only filter leading to similar results is *contrast*, because it does not deviate much from the training data

3. To test the robustness of the model, it would be good to conduct in-vivo experiments

4. Fashion-MNIST is a widely used dataset; it may be necessary to re-evaluate existing research and replicate it in a more realistic context

Future Works

Evaluate the performance of CNN-based models with other in-vivo scenarios

Investigate the main cause of the results to understand whether they depend on training/test data or validation procedures

Experiment with other dataset and use cases:

- self-driving cars
- safety evaluation

Software Engineering approaches can detect and succumb to issues related to AI contexts.

Leveraging SE4AI would be appropriate.

Thank You!

https://giusyann.github.io/

@Giusy_A_

gannunziata@unisa.it