Comparative Analysis of Machine Learning Models on Fashion MNIST Dataset

Aliyu Bello, Ahmad Garouda

Supervisor: Dr. Ahmed Elsheikh

Faculty of Computer Science

University of Prince Edward Island, Cairo Campus

December 2023

Abstract

The Fashion MNIST dataset, a standard in image classification, is used in this study's broad comparison of four well-known machine learning algorithms: Support Vector Machine (SVM), XGBoost, Random Forest, and Convolutional Neural Network (CNN). The main goal was to assess and contrast these models' performance regarding generalizability, computational efficiency, and accuracy. Our findings reveal that while CNN achieved the highest accuracy (92%) and F1 score (92%), it required the most considerable training time (approximately 35 minutes). In contrast, XGBoost demonstrated remarkable efficiency, with a training time of just 1 minute and a testing time of 0.25 seconds, while achieving an impressive accuracy of 88% and an F1 score of 89%. SVM and Random Forest, although proficient in training, showed signs of overfitting, as evidenced by their lower testing accuracies. The study underscores the importance of balancing computational demands with model performance and highlights the suitability of different algorithms based on specific application requirements. These insights offer valuable guidance for future research in machine learning, particularly in image classification, emphasizing the need for tailored algorithm selection and the potential of exploring advanced model tuning, hybrid approaches, and ethical considerations in AI applications.

Abbreviations

RF - Random Forest

ML - Machine Learning

SVM - Support Vector Machine

CNN - Convolutional Neural Network

XGBOOST - eXtreme Gradient Boosting

AI - Artificial Intelligence

TP - True Positives

FP - False Positives

TN - True Negatives

FN - False Negatives

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The project's primary motivation is to effectively interpret and categorize a wide variety of fashion items represented in grayscale images within the Fashion MNIST dataset. Renowned for its diversity in fashion products, this dataset presents unique challenges in image classification, establishing it as an ideal platform for evaluating the efficacy of different machine learning approaches. The algorithms selected for this purpose – XGBoost, Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Random Forest – have been chosen for their distinctive strengths and potential limitations in this specific context.

XGBoost, known for its superior performance in structured data classification, offers high accuracy and speed, making it a strong contender for the Fashion MNIST dataset Chen (2016). However, its linear approach may not be as effective in capturing the complex spatial patterns in image data compared to neural network-based methods. SVM, noted for its efficiency in high-dimensional spaces, is versatile and potent in image classification tasks due to its kernel functions Cortes (1995). Its downside lies in its computational intensity, particularly with large datasets, and its dependency on the choice of the kernel.

The inclusion of CNN is driven by its unparalleled performance in image recognition tasks Krizhevsky (2012). With its ability to learn spatial hierarchies, CNN is particularly suited for

the Fashion MNIST dataset, although it requires significant computational resources and may be prone to overfitting. Random Forest, selected for its simplicity and efficiency with non-linear data, is robust against overfitting and versatile in handling classification and regression tasks Breiman (2001). However, its performance might not be as robust with high dimensional data, such as images, compared to CNN.

The rationale for selecting these algorithms is to utilize their respective strengths in addressing the complexities of fashion image classification while recognizing their potential limitations. This project aims to conduct a comprehensive comparison of these techniques, illuminating their practical applicability and performance in a real-world dataset. By understanding the pros and cons of each algorithm in this context, the project seeks to contribute meaningful insights to the field of image classification and machine learning.

1.2 Outline

The paper outline follows: Chapter 2 provides basic definitions and Background describing each model's theory. Chapter 3 shows overall performance results, analysis, and discussions. Chapter 4 includes conclusions and future work.

Chapter 2

Background

2.1 Basic Definitions & Terminologies

2.1.1 Support Vector Machines

Support vector machine (SVM) is a robust supervised learning technique utilized in both classification and regression tasks.Gammermann (2000) SVM excels in handling problems with extremely high dimensions that were previously unmanageable for other learning machines.

2.1.2 Random Forest

In the Fashion MNIST dataset, a Random Forest is like a group of fashion judges (decision trees), where each judge specializes in identifying different clothing types. They individually analyze fashion items like shirts, shoes, and pants, and then collectively decide the category of each item. The model combines multiple decision trees to improve accuracy and handles complex relationships Ángel Hernández García. (2018). This method improves accuracy in classifying fashion products, as it combines the expertise of multiple judges, reducing the chance of misidentifying an item.

2.1.3 XGboost

XGBOOST (eXtreme Gradient Boosting) is an advanced decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It implements machine learning algorithms under the Gradient Boosting framework and provides a parallel tree boosting that solves many data science problems in a fast and accurate way Chen (2016).

2.1.4 Convolutional Neural Network CNN

Image processing and recognition have been shown to depend heavily on a particular kind of neural network known as a convolutional neural network (CNN). Originating from the biologically-inspired models of the human visual cortex, the foundations of CNNs were laid in the early works of Kunihiko Fukushima's "neocognitron" and later refined by Yann LeCun's team in the 1990s LeCun (1998). The unique architecture of CNNs allows them to excel in tasks that involve the recognition and classification of images, making them a pivotal tool in the advancement of deep learning technologies. Convolutional, pooling, and fully linked layers are the many layers that make up a CNN's architecture.

The convolutional layers utilize learnable filters to detect spatial patterns such as edges and textures in images Goodfellow (2016). Pooling layers, typically following the convolutional layers, help in reducing the dimensionality of the data, thus aiding in reducing computational load and mitigating overfitting. The network concludes with fully connected layers, akin to a traditional neural network, where each neuron is connected to all activations in the preceding layer. The discipline of computer vision has been greatly impacted by their efficiency in tasks including object identification, face recognition, and image categorization.

2.2 Data Exploration

2.2.1 Dataset Overview:

The Fashion MNIST dataset, created by Zalando Research, serves as an advanced alternative to the classic MNIST dataset for benchmarking machine learning models. Zalando (2017) It's specifically designed to represent a more challenging classification task with a focus on fashion items.

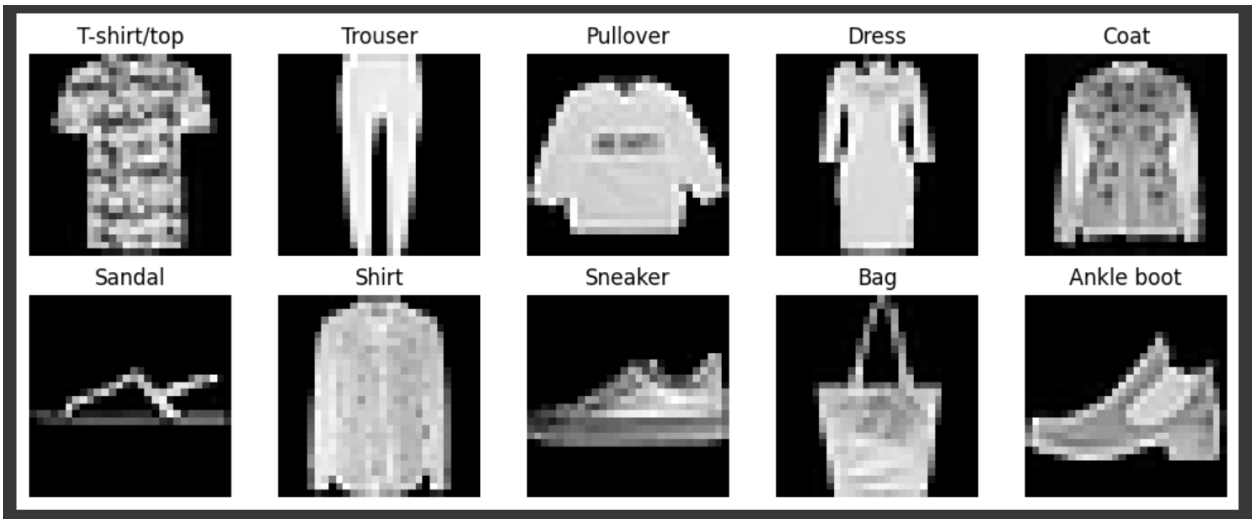
2.2.2 Data Composition and Structure:

This dataset comprises 60,000 training and 10,000 test images, each 28x28 pixels in grayscale. It encompasses 10 different classes, including a variety of apparel and accessory items like trousers, shirts, dresses, and bags. The dataset is balanced, with each class having an equal number of images, which is crucial for avoiding bias in model training.

2.2.3 Sample Visualizations:

Visual inspection of the dataset reveals diverse fashion items with varying shapes and styles. Sample images from each category provide a glimpse into the dataset's variety, highlighting the distinct features that models need to learn for accurate classification.

Figure 2.1: Sample classes for the Fashion MNIST



2.2.4 Data Quality Assessment:

The dataset is high-quality, with no missing values or label inconsistencies. The uniformity in image size and the absence of color simplify the preprocessing steps but also limit the complexity of features available for learning.

2.3 Data Preprocessing

2.3.1 Extraction of Features and Labels:

- *Xtrain* and *Xtest* are created by dropping the 'label' column from the *fashiontrain* and *fashiontest* datasets, respectively. This step isolates the image data (features) from the labels.
- *ytrain* and *ytest* are arrays containing the labels from the 'label' column of *fashiontrain* and *fashiontest*. These labels represent the classes of the fashion items.

2.3.2 Data Type Conversion:

- The image data in *Xtrain*and *Xtest* is converted to float32. This conversion is crucial for the normalization step that follows and ensures computational efficiency.

2.3.3 Normalization

- The pixel values in *Xtrain* and *Xtest* are divided by 255. Since pixel values range from 0 to 255, this step normalizes these values to a range of 0 to 1. Normalization is important for machine learning models as it leads to faster convergence during training.

2.3.4 Reshaping the Data:

- The data in *Xtrain* and *Xtest* is reshaped into a 4D array with dimensions (-1, 28, 28, 1).
- The '-1' infers the number of samples automatically, 28x28 is the dimension of each image, and '1' signifies that the images are in grayscale (single channel).
- This reshaping is necessary for CNN models, which expect input data in this specific format.

Chapter 3

Results and Discussion

The results of our comprehensive machine learning testing are presented in this chapter. The testing set is used to assess the model's performance and capacity for generalisation on untested data, whereas the training set is used to train the model. The F1-score, accuracy, and training/testing speeds are these parameters, as displayed below.

$$\textit{Accuracy} = \frac{TP + TrueNegatives(TN)}{TP + TN + FP + FN} \tag{3.1}$$

$$\textit{Time} = Train - TestSpeed \tag{3.2}$$

$$\textit{F1-Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.3}$$

Table 3.1: Table for average value for the datasets.

Best Hyperparameter	Most frequent values
n_estimators	100
max_features	auto
max_depth	20
min_sample_split	2
bootstrap	True

CNN key hyperparameters

Number of Epochs, Dropout rate, Network architecture (Convolutional Layers, Pooling Layers, Dense Layer)

3.0.1 Random forest

An analysis of the Fashion mnist dataset using the Random Forest classifier. The evaluation looks at several performance measures and tracks how different hyperparameters affect the model's efficiency, and uses the above evaluation metrics to see the results.

Table 3.2: Random Forest evaluation metrics

Evaluation Metrics	Training Result	Testing Result
Accuracy	100%	86%
F1-score	100%	87%
Time	2m	0.36

3.0.2 Support Vector Machines SVM

A Support Vector Machines classifier analysis of the Fashion mnist dataset. The evaluation uses the evaluation metrics mentioned above to examine a number of performance metrics and monitor the impact of various hyperparameters on the model's efficiency.

Table 3.3: Support Vector Machines evaluation metrics

Evaluation Metrics	Training Result	Testing Result
Accuracy	99%	87%
F1-score	100%	88%
Time	21m	42s

3.0.3 XGBOOST

A Gradient-boosting classifier analysis was applied to the Fashion MNIST dataset using XGBoost. The assessment employs a comprehensive set of evaluation metrics to see various performance aspects and assess the impact of diverse hyperparameters on the model's overall efficiency.

Table 3.4: XGBOOST evaluation metrics

Evaluation Metrics	Training Result	Testing Result
Accuracy	99%	88%
F1-score	100%	89%
Time	1m	0.25

3.0.4 Convolutional Neural network CNN

An exploration of Convolutional Neural Networks (CNNs) applied to the Fashion MNIST dataset using TensorFlow and Keras. This analysis delves into the performance evaluation of the CNN model, leveraging a diverse set of metrics to assess accuracy, precision, recall, and F1-score. Furthermore, the study investigates the impact of varying hyperparameters, such as dropout rates and filter sizes, on the overall efficiency of the CNN. By conducting this in-depth examination, our goal is to clarify CNN's effectiveness in accurately classifying fashion-related images within the Fashion MNIST dataset.

Table 3.5: Classification Report for CNN

	Precision	Recall	F1-Score	Support
0	0.89	0.84	0.87	1000
1	0.99	0.99	0.99	1000
2	0.91	0.85	0.88	1000
3	0.93	0.93	0.93	1000
4	0.85	0.90	0.87	1000
5	0.99	0.98	0.99	1000
6	0.75	0.80	0.77	1000
7	0.96	0.96	0.96	1000
8	0.98	0.99	0.99	1000
9	0.97	0.98	0.97	1000
Accuracy			0.92	10000
Macro Avg	0.92	0.92	0.92	10000
Weighted Avg	0.92	0.92	0.92	10000

Table 3.6: CNN evaluation metrics

Evaluation Metrics	Testing Result
Accuracy	92%
F1-score	92%
Time	35m

3.1 Discussion

Random Forest

- **Training and Testing Time:** Random Forest had a reasonable training time of 2 minutes and a testing time of 0.36 seconds, indicating good computational efficiency.
- **Accuracy and F1 Score:** It achieved the highest training accuracy (100%) but a slightly lower testing accuracy of 86% and an F1 score of 87%. The perfect training accuracy indicates an overfitting to the training set, which is reflected in the lower testing accuracy.

Support Vector Machine (SVM)

- **Training and Testing Time:** SVM showed the longest training time at 21 minutes and a moderate testing time of 42 seconds. This indicates that SVM may be computationally intensive compared to the other models.
- **Accuracy and F1 Score:** While SVM achieved a high training accuracy of 99%, its testing accuracy dropped to 87%, with an F1 score of 88%. This suggests a potential overfitting of the training data or a lack of generalization to the test dataset.

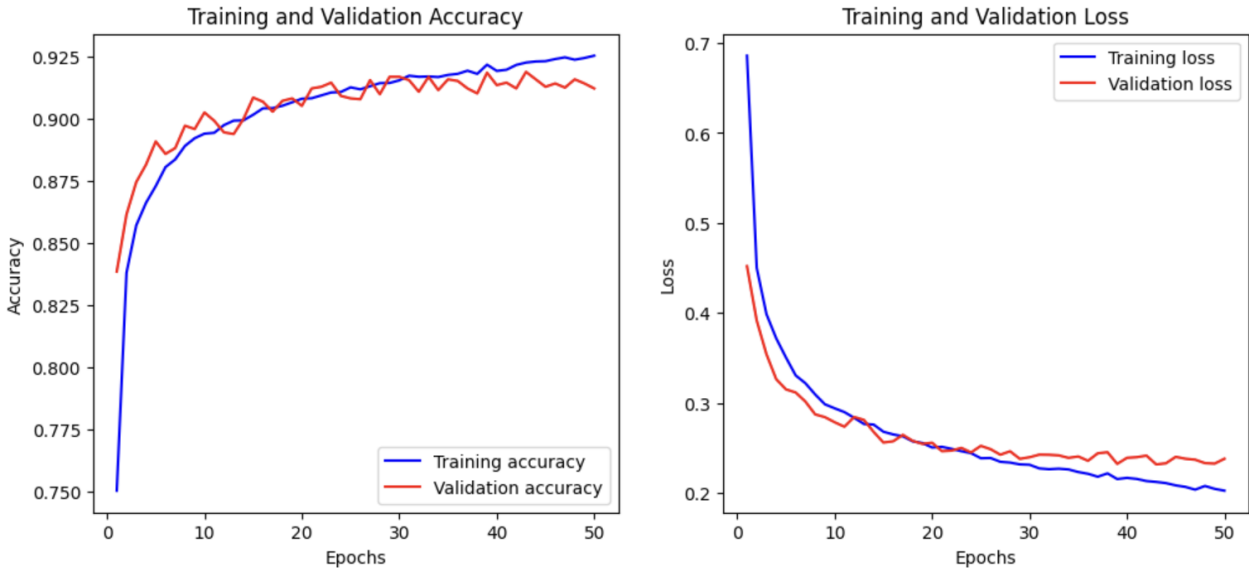
XGBOOST

- **Training and Testing Time:** XGBoost was remarkably efficient in training time, taking only 1 minute, and had a speedy testing time of 0.25 seconds. This efficiency makes XGBoost a suitable option for limited computational resources or time.
- **Accuracy and F1 Score:** XGBoost also achieved a high training accuracy (99%) and the highest testing accuracy among the three models (88%), with an F1 score of 89%. This suggests a good balance between learning from the training data and generalizing to unseen data.

Convolutional Neural Network (CNN):

- **Training Time:** CNN had the longest training time at approximately 35 minutes, which is significantly higher than the other models. This reflects the generally higher computational demands of deep learning models, especially for image data.
- **Testing Performance:** The CNN achieved a test accuracy and F1 score of 92%, both higher than those achieved by SVM, XGBoost, and Random Forest. This superior performance is likely due to the CNN's ability to extract hierarchical spatial features from images, making it highly effective for image classification tasks.

Figure 3.1: Learning process of CNN



The left graph is labeled "Training and Validation Accuracy" and the right graph is labeled "Training and Validation Loss". Both graphs have the x-axis labeled as "Epochs", which indicates the number of complete passes through the entire training dataset.

- **Training and Validation Accuracy:** The graph displays the model's accuracy on training (blue) and validation (red) data across epochs. Training accuracy is slightly higher, a common pattern indicating good learning without significant overfitting, as shown by the close proximity of both accuracies.

- **Training and Validation Loss:** This graph shows the model's loss on the training set (in blue) and the validation set (in red) over the same number of epochs. Loss is a metric that reflects how well the model's predictions match the actual data; the lower the loss, the better. Typically, loss decreases sharply at the beginning and then more slowly, which is what we see here. The training loss decreases more and ends up lower than the validation loss, which is also common. As long as the validation loss is not increasing, the model is generally considered to be learning effectively."

3.1.1 General insights

The comparative analysis of SVM, XGBoost, Random Forest, and CNN on the Fashion MNIST dataset offers valuable insights into the trade-offs between computational efficiency and classification accuracy. While CNN stands out with the highest accuracy and F1 score, indicative of its superior capability in image classification, it requires the longest training time, highlighting the computational demands of deep learning models. In contrast, XGBoost strikes a balance, offering high efficiency, relatively short training and testing times, and commendable accuracy and generalization capabilities. The performance of SVM and Random Forest, though proficient in training, suggests potential issues with overfitting, as reflected in their lower testing accuracies. This comparison underscores the importance of considering the nature of the dataset and the specific application context when selecting a machine learning model. The choice between these models thus hinges on prioritizing either computational resource constraints and response time (favoring models like XGBoost) or maximizing classification accuracy (leaning towards CNN), depending on the specific needs and constraints of the intended application.

Chapter 4

Conclusion, Recommendations, and Future Work

4.1 Conclusion

In this analysis of the Fashion MNIST dataset, we delved into the performance of various machine learning models in image classification, including SVM, XGBoost, Random Forest, and CNN. The results were quite revealing. The CNN, in particular, stood out with its exceptional accuracy and F1 scores, highlighting its strength in handling image-based tasks. However, this came at the expense of increased computational demand and longer training times. On the other hand, XGBoost proved to be a strong competitor, offering a good balance between accuracy and efficiency. It emerged as an ideal choice for situations where quick processing and resource conservation are key. We also observed interesting patterns with SVM and Random Forest. While they demonstrated high training accuracy, their testing performance was somewhat lower, hinting at possible overfitting issues. This calls for careful interpretation of their effectiveness. This investigation reveals the unique attributes and appropriateness of each algorithm for fashion image classification and highlights the complex trade-offs between computational efficiency, model complexity, and accuracy. These insights are invaluable for guiding future research and application development in this field. They emphasize the importance of choosing the right model based on a project's

specific requirements and limitations. Overall, this study serves as an important reference for practitioners and researchers, offering a clearer understanding of the capabilities of machine learning models in image classification and setting the stage for further innovative explorations in this area.

4.2 Future Work

In light of the findings from this study on machine learning applications in fashion image classification, future researchers are encouraged to pursue a series of enhanced strategies. Precision in model selection is paramount; it is crucial to weigh the trade-offs between computational efficiency, model complexity, and accuracy to suit the specific requirements of the application at hand. Delving deeper into model tuning and advanced training methodologies can refine performance and address potential overfitting. Exploring hybrid models or ensemble techniques could prove fruitful, potentially combining the strengths of algorithms like XGBoost and CNN for superior outcomes. Expanding the application of these models to a broader range of datasets will enrich our understanding of their adaptability and robustness across different contexts. Furthermore, AI principles and inclusion must be incorporated in every model creation phase to guarantee various groups' impartiality, equity, and representation. These suggestions are meant to direct future research towards more creative, ethical, and significant applications of machine learning—not just for picture categorization but other purposes.

4.3 APPENDIX

Figure 4.1: CNN confusion matrix

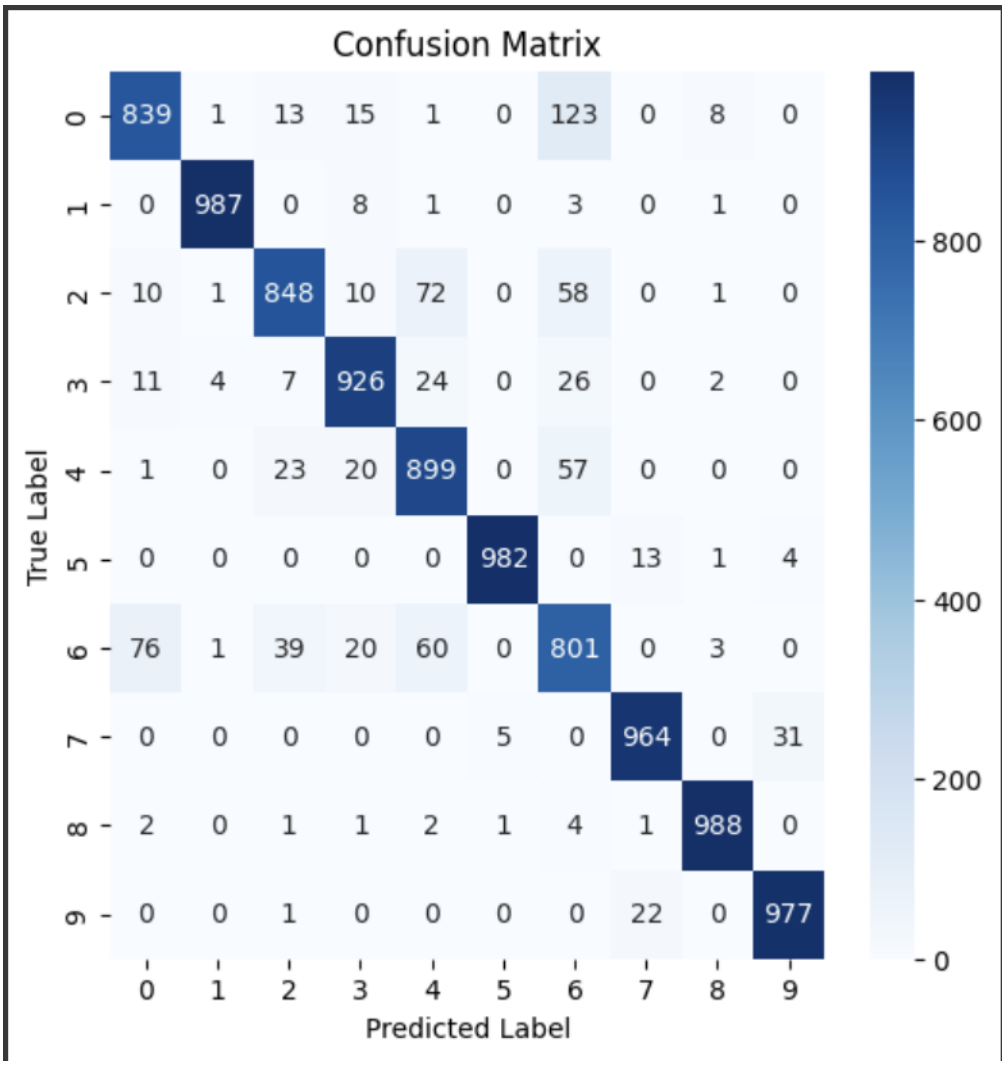


Figure 4.2: CNN correct predicted classes

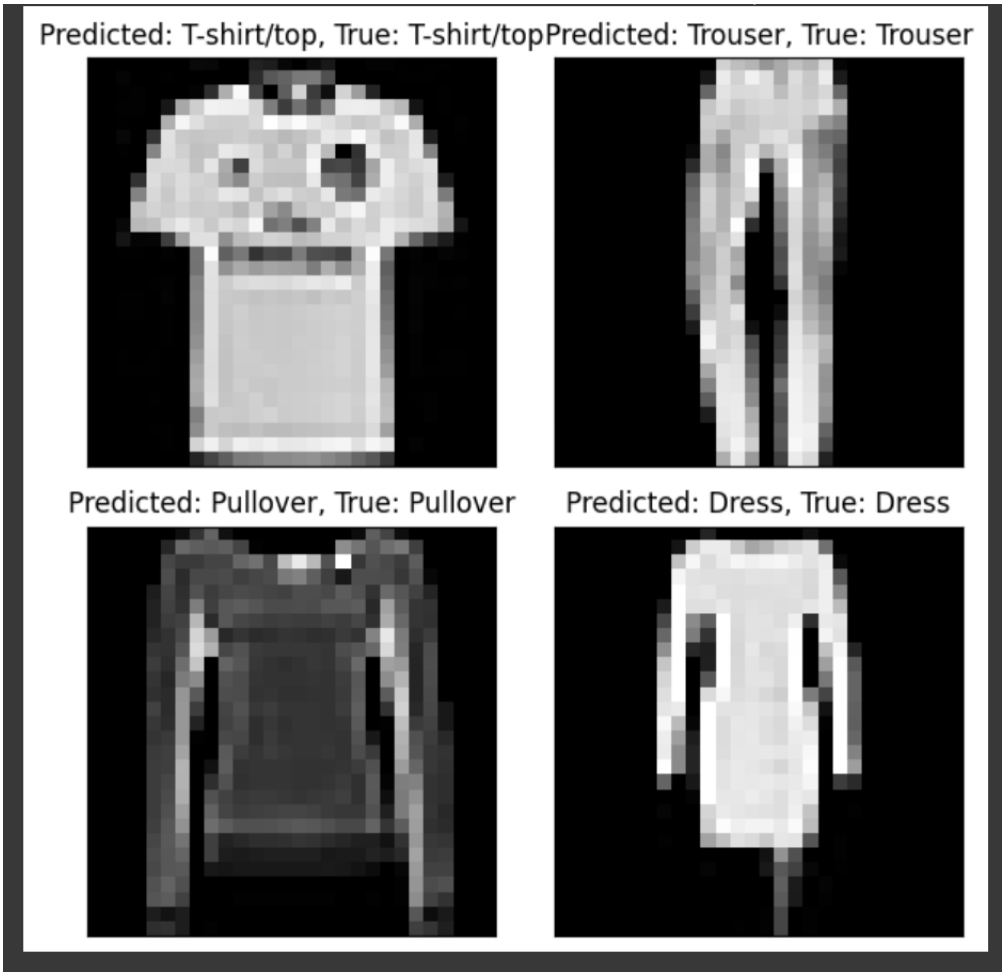


Figure 4.3: XGBOOST and Random forest accuracies comparison

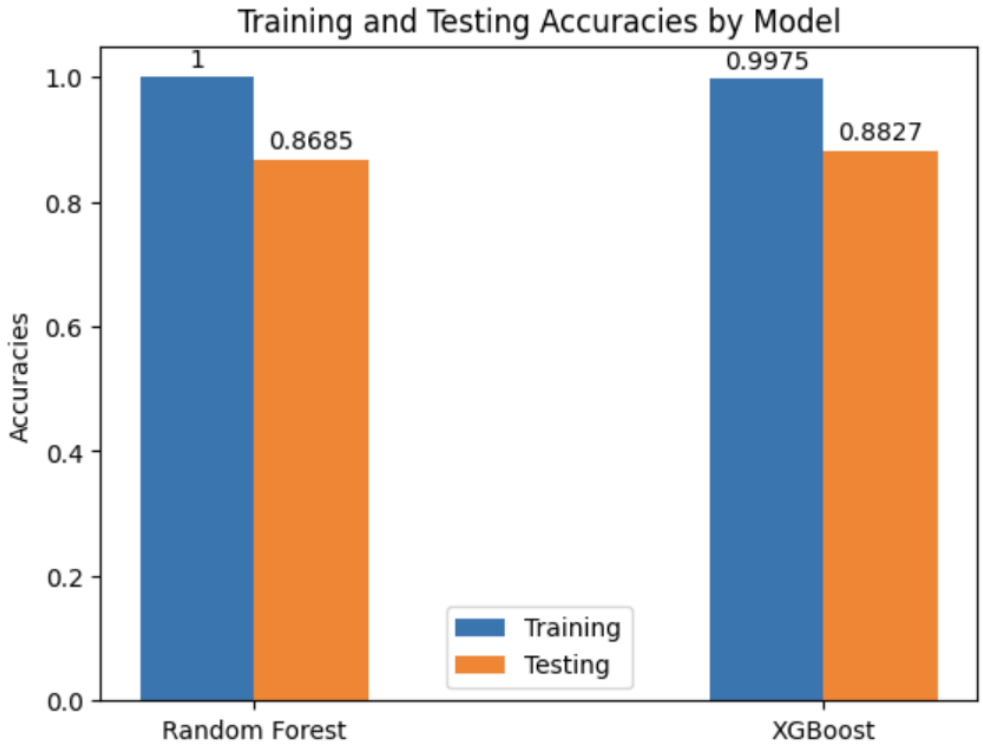


Figure 4.4: CNN incorrect predicted classes

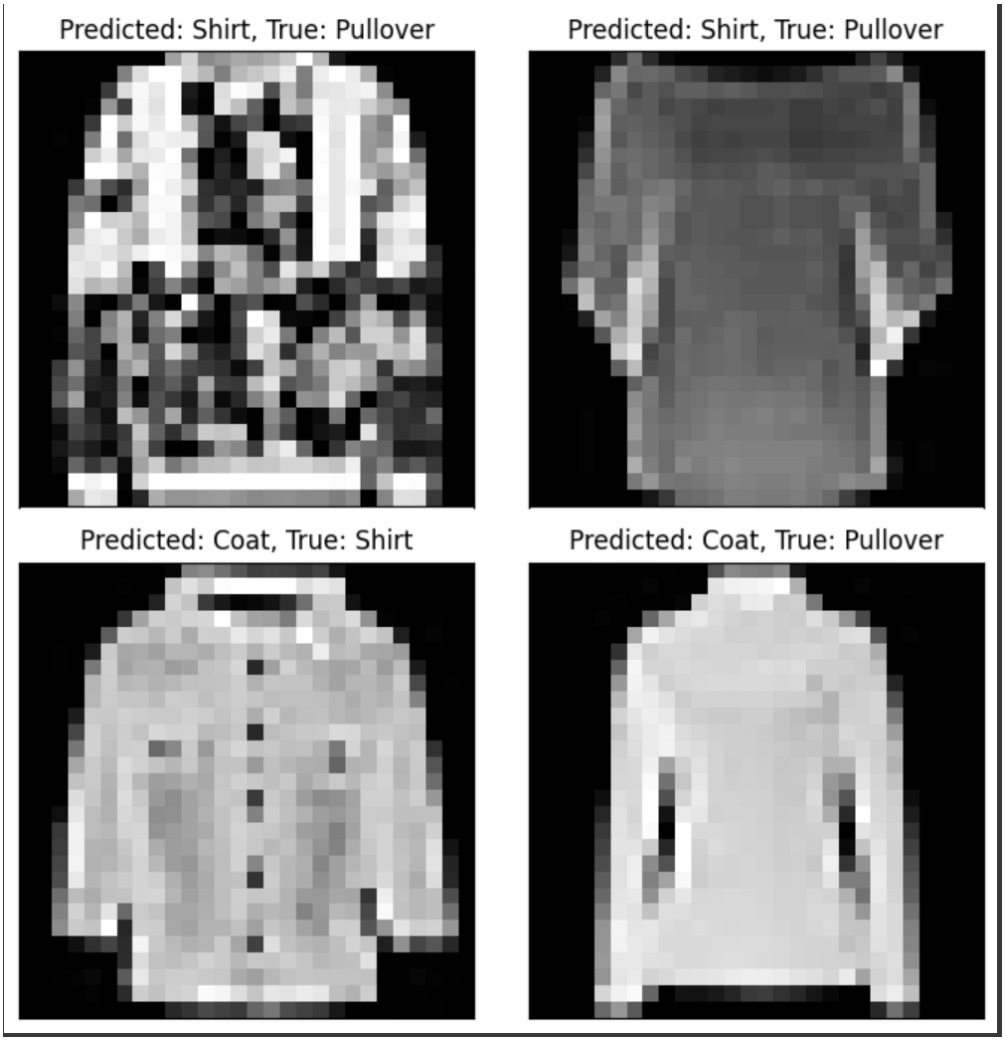


Figure 4.5: RF confusion matrix

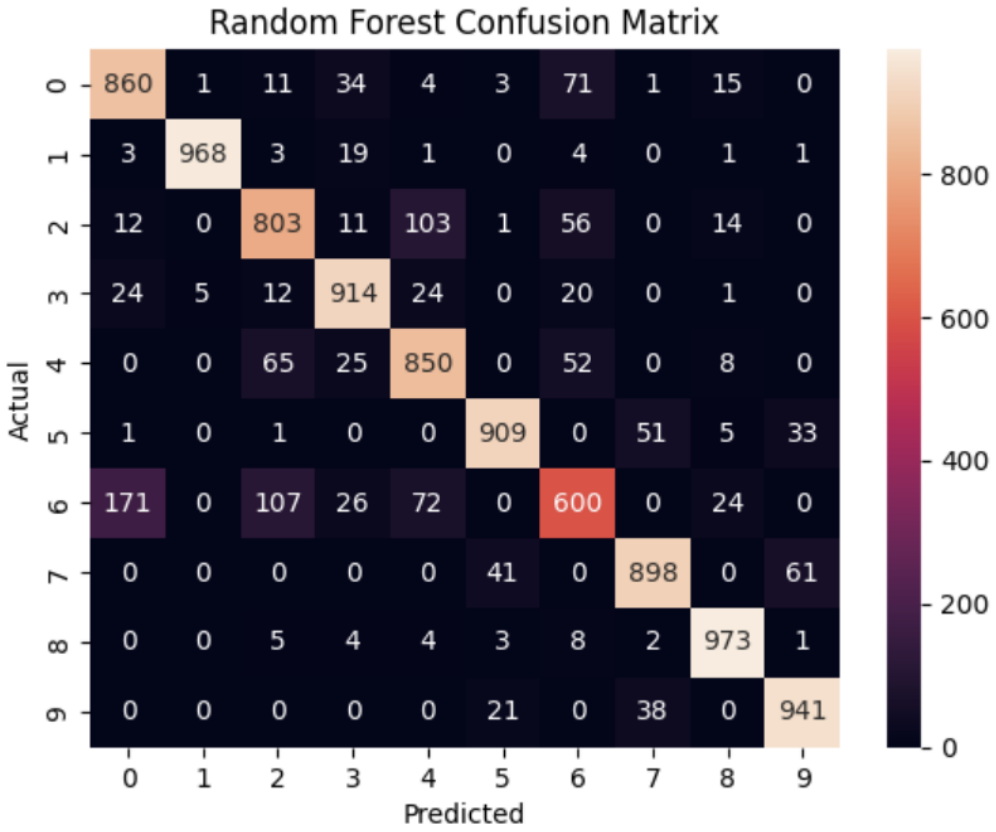


Figure 4.6: XGBOOST confusion matrix

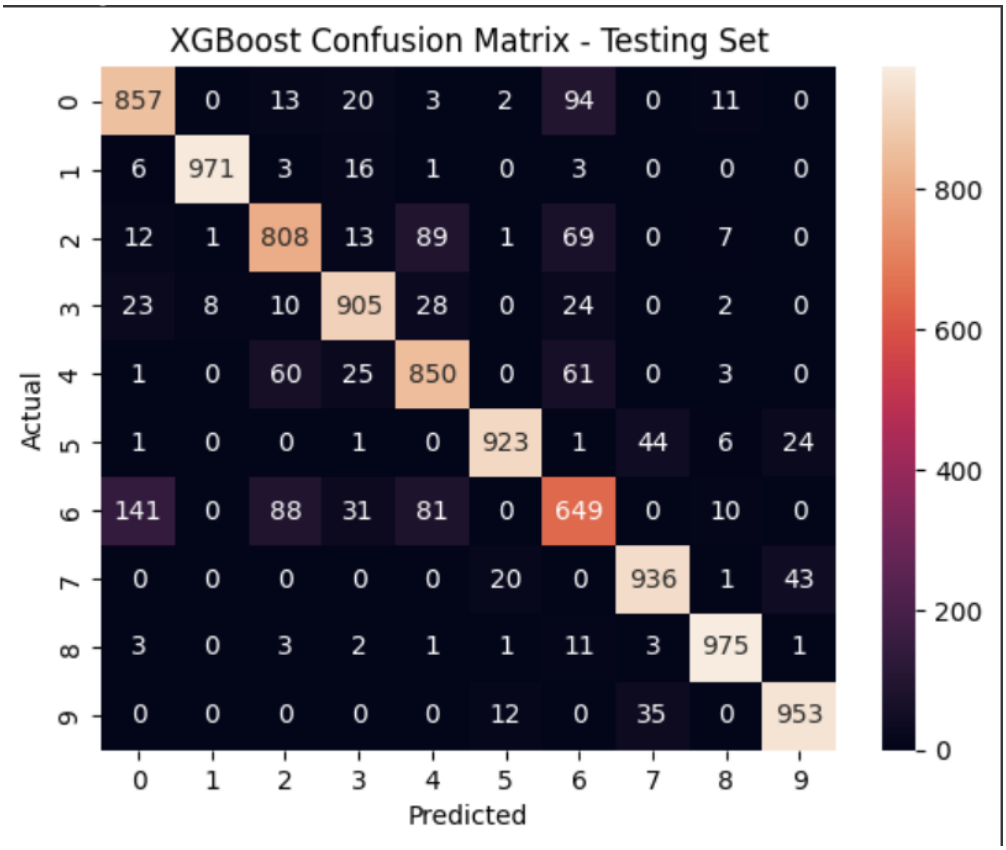
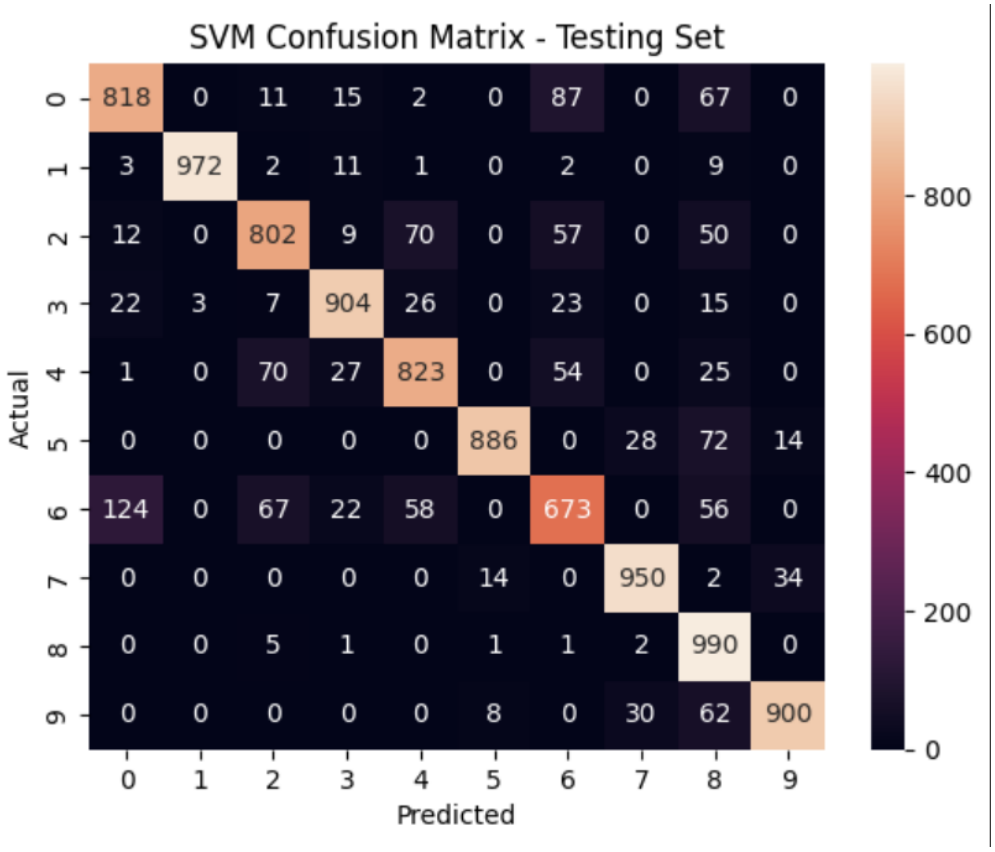


Figure 4.7: SVM confusion matrix


SVM Confusion Matrix - Testing Set
Actual
Predicted
0 1 2 3 4 5 6 7 8 9
818 0 11 15 2 0 87 0 67 0
3 972 2 11 1 0 2 0 9 0
12 0 802 9 70 0 57 0 50 0
22 3 7 904 26 0 23 0 15 0
1 0 70 27 823 0 54 0 25 0
0 0 0 0 0 886 0 28 72 14
124 0 67 22 58 0 673 0 56 0
0 0 0 0 0 14 0 950 2 34
0 0 5 1 0 1 1 2 990 0
0 0 0 0 0 8 0 30 62 900
800
600
400
200
0

Bibliography

Breiman, L. (2001). Random forests.

Chen, T., . G. C. (2016). Xgboost: A scalable tree boosting system.

Cortes, C., . V. V. (1995). Support-vector networks.

Gammermann (2000). Support vector machine learning algorithm and transduction.

Goodfellow, I., B. Y. . C. A. (2016). Deep learning.

Krizhevsky, A., S. I. . H. G. E. (2012). Imagenet classification with deep convolutional neural networks.

LeCun, Y., B. L. B. Y. . H. P. (1998). Gradient-based learning applied to document recognition.

Zalando (2017). Fashion mnist.

Ángel Hernández García., J. C.-P. (2018). Predicting teamwork group assessment using log data-based learning analytics.