Fashion MNIST Image Classification

Yamini Preethi Kamisetty
Texas A&M University
College Station, United States
yamini_preethi_k@tamu.edu

Bhavyasree Mohan
Texas A&M University
College Station, United States
bhavya_0108@tamu.edu

Sri Harsha Parimi Venkata
Texas A&M University
College Station, United States
parimi@tamu.edu

***Abstract*—Data mining, machine learning and deep learning are being extensively used in today's world across various sectors like e-commerce, fashion, automobile industry and others. With enormous data available in the fashion industry, there has been a rapid increase in the usage of the data science technologies in fashion e-commerce to address several problems like clothing classification, recognition, and recommendations. This paper presents various approaches to perform clothing classification using the Fashion MNIST dataset. The study aims to classify the images into ten categories. In this paper, classical machine learning algorithms like XGBoost and Random Forest have been used as baseline models to perform the classification. Deep learning algorithms like Convolutional Neural Networks (CNN) are utilized to perform this image classification task. The paper also aims at the interpretability of these models predictions by using various methods including LIME and SHAP.**

***Index Terms*—Machine Learning, Data Mining, Convolutional Neural Networks, Fashion MNIST, Interpretability**

I. Introduction

Image classification solves the problem of identifying specific entities from an image using computer algorithms. In the past few years, there has been a significant increase in the usage of deep learning models like Convolutional Neural Networks(CNN) for image classification [1] for processing of image data with improved performance. CNN represent the cutting-edge in neural network technology, learning to extract features from input images and map them to a predefined set of classes. CNN's have emerged as a powerful tool in image classification due to their ability to effectively capture spatial hierarchies and complex patterns within images.

The Fashion MNIST dataset [2], is a collection of 28x28 grayscale images of 10 different fashion categories, with a training set of 60,000 examples and a test set of 10,000 examples. Each image is associated with a label indicating the class of the clothing item it represents.

This study provides approaches to classify the Fashion MNIST images to one of the 10 pre-defined classes. The paper utilizes classical machine learning algorithms like XGBoost to create a baseline benchmark and then implement a Convolutional Neural Network with multiple hidden layers resulting in a better performance. The project also includes interpretability of the models which helps in understanding which pixels contributed to the predicted class using techniques like Local Interpretable Model-Agnostic Explanations(LIME) [3] and SHapley Additive exPlanations(SHAP).

II. Methods

This study is centered on employing two methodologies: the classical machine learning approach for establishing a baseline model and the deep learning approach. This paper also discusses the model performance in each of the approaches. XGBoost algorithm is a gradient boosting ensemble method that uses the power of decision tree classifiers where each tree rectifies the errors made by its predecessors. Convolutional Neural Networks' [4] have a unique architecture which enables local feature extraction and hierarchical representation. [5] [6] [7]. This study uses LeNet [8] architecture comprising of convolutional, subsampling and fully connected layers that are responsible for feature extraction, nonlinear transformations, and classification. After training the models, the model performance has been tested on out of sample images taken from google images, to validate the model for real world applicability. The results have been discussed below.

III. Experiments

A. Data Preparation and Visualization

The image data consists of 784 features where each feature is a pixel value. Fig. 1 visually represents the image data by plotting the pixels as a rendered pseudocolor image.

Fig. 1: Pixel data visualized as images

From Fig. 2, gives the descriptive statistics of the Fashion MNIST data and it can be inferred that all the 10 classes are equally balanced.

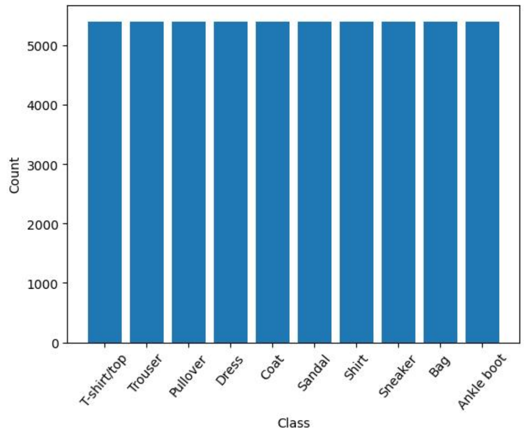

Fig. 2: Class Distribution of Fashion MNIST dataset

The data is split into train, test and validation sets where train set is used for model training, validation set is used for cross validating the trained model and test set for measuring model performance. The data has also been normalized using PyTorch which scales the input down to [0,1] for the CNN model.

Dimensionality reduction techniques Uniform Manifold Approximation and Projection(UMAP) and t-distributed Stochastic Neighbor Embedding(TSNE) have been used to visualize the data in 2 dimensions. It can be observed from Fig. 3 how different each class is from other classes. For example, "ankle boot", "sneakers" and "sandal" classes are grouped together since they belong to similar categories and are distant from classes like T-shirt or dress.

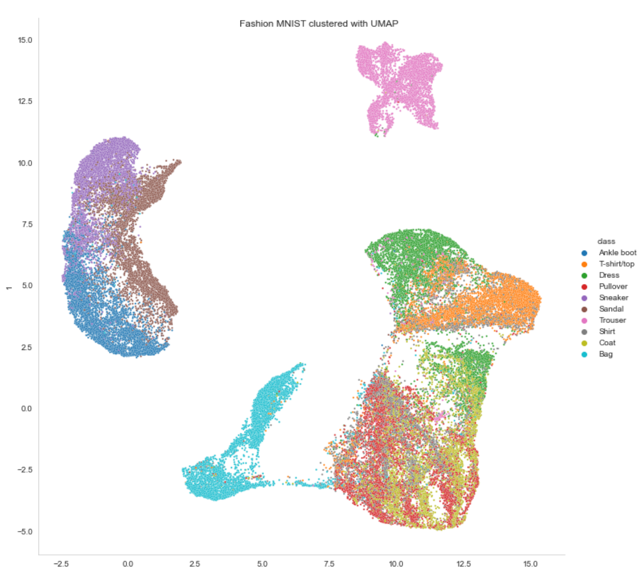

Fig. 3: UMAP dimensionality reduction plot

B. Machine Learning Approach

The XGBoost classifier is employed through the XGBoost library to establish a baseline model. Subsequently, hyperparameter tuning is conducted using grid search cross-validation to obtain optimal parameters for the XGBoost classifier. Essential parameters, such as colsample_bytree for column subsampling during tree construction and max_depth indicating the tree's maximum depth, are fine-tuned. The feature importance, depicted in Fig. 4, is derived from the trained XGBoost model, elucidating the pivotal pixel features contributing significantly to the model's predictive outcomes. This comprehensive approach ensures the robustness and efficacy of the baseline model in the context of the study.

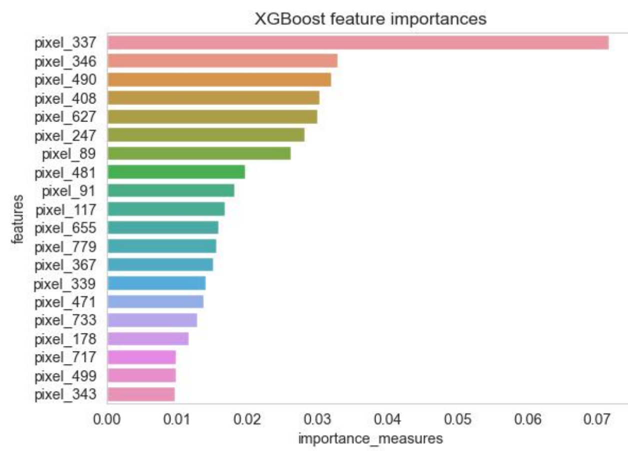

Fig. 4: XGBoost Feature Importance

C. Deep Learning Approach

The CNN model has been built based on the LeNet architecture. As depicted in Fig. 5, the model has 5 layers - 2 convolutional, 3 fully connected. The first convolutional layer is configured with a kernel size of 5x5 and 6 output channels, followed by a ReLU(rectified linear unit) activation [9]. Subsequently, an average pooling layer with a 2x2 kernel and stride 2 is employed to downsample the feature maps.

Building upon the feature hierarchy, a second convolutional layer is introduced with 16 output channels and a 5x5 kernel and passed through ReLU activation, followed by the same average pooling operation, further reducing spatial dimensions.

The convolutional outputs are flattened and passed through a series of fully connected layers. The first fully connected layer comprises 120 neurons, followed by ReLU activation. Subsequently, a second fully connected layer with 64 neurons and ReLU activation is incorporated, concluding with an output layer consisting of 10 neurons, representing class probabilities for the classification task.

This LenetCNN architecture showcases a sequential arrangement of convolutional and fully connected layers, adhering to the principles of feature extraction and hierarchical learning essential for effective image classification tasks and this model achieved an accuracy of 91.2%.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [100, 6, 28, 28]             156
              ReLU-2          [100, 6, 28, 28]               0
         AvgPool2d-3          [100, 6, 14, 14]               0
            Conv2d-4         [100, 16, 10, 10]           2,416
              ReLU-5         [100, 16, 10, 10]               0
         AvgPool2d-6           [100, 16, 5, 5]               0
            Linear-7                [100, 120]          48,120
              ReLU-8                [100, 120]               0
            Linear-9                 [100, 64]           7,744
             ReLU-10                 [100, 64]               0
           Linear-11                 [100, 10]             650
================================================================
Total params: 59,086
Trainable params: 59,086
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.30
Forward/backward pass size (MB): 11.11
Params size (MB): 0.23
Estimated Total Size (MB): 11.63
----------------------------------------------------------------
```

Fig. 5: CNN Architecture

D. Model Evaluation

To assess model performance, we conducted rigorous evaluations on accuracy, particularly focusing on individual classes. After training the XGBoost Classifier and CNN model on the Fashion MNIST dataset the following results have been obtained.

For XGBoost,

Accuracy on training set - 98.6%

Accuracy on test set - 89.43%

For CNN,

Accuracy on training set - 95.07%

Accuracy on test set - 91.26%

The accuracy has also been evaluated for each class separately using CNN and below are the results -

T-shirt/Top: 87.70% Trouser: 97.70%

Pullover: 88.00% Dress: 90.60%

Coat: 86.40% Sandal: 98.10%

Shirt: 73.80% Sneaker: 95.60%

Bag: 98.60% Ankle Boot: 96.20%

Fig. 6 and Fig. 7 below represents the confusion matrix for XGBoost and CNN on test set.

Fig. 8 shows the trend of model accuracy across increasing epochs.

E. Model Explainability

In the context of interpreting black box models, model explainability methods such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have been employed to gain insights into the decision-making processes of XGBoost and CNN models on the Fashion MNIST dataset. SHAP helps identify key factors influencing predictions, while LIME provides insights into local interpretability.

Fig. 6: XGBoost Predictions Confusion Matrix

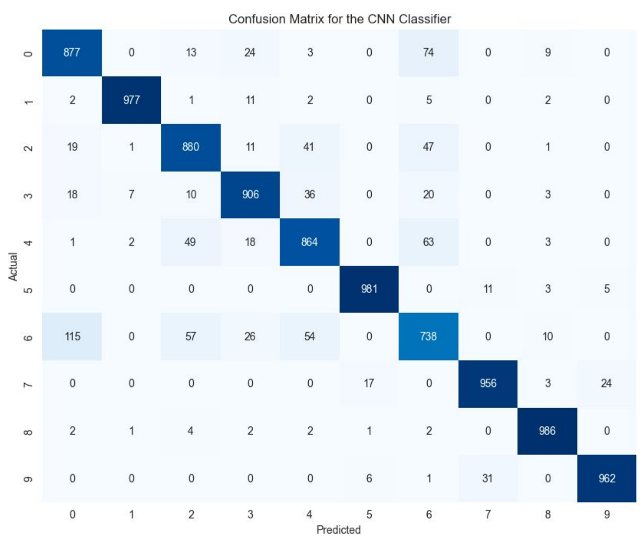

Fig. 7: CNN Predictions Confusion Matrix

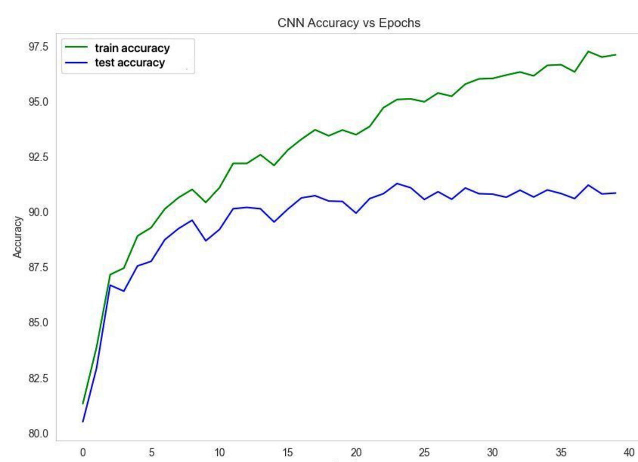

Fig. 8: CNN model training - accuracy across epochs

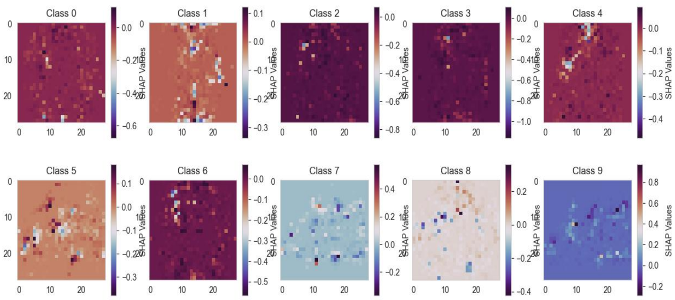

Fig. 9: SHAP Explanation for XGBoost

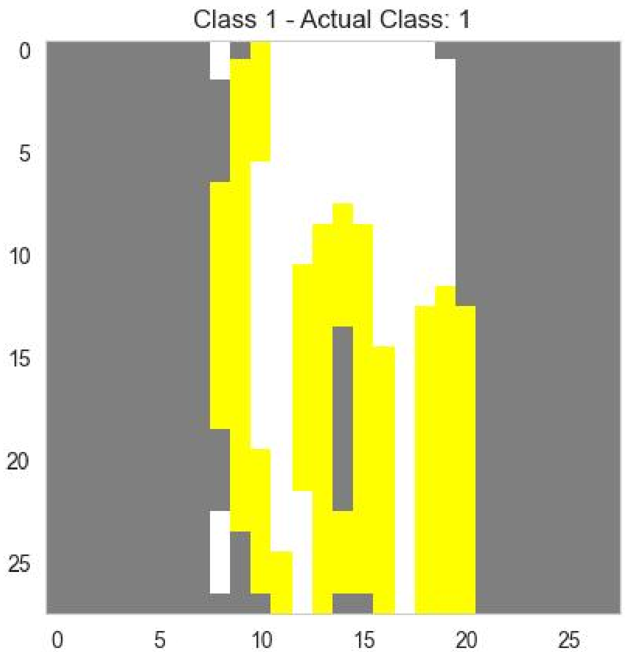

Fig. 10: LIME Explanation for XGBoost

Fig. 9 and Fig. 10 suggests that the predictions of XGBoost model for fashion items heavily rely on pixels encapsulating the shape of the garment.This insight aligns with expectations, as the outer shape is a crucial characteristic for distinguishing any garment within the dataset.

It can also be inferred that this reliance on the outer shape is the reason why the model tends to misclassify shirts, t-shirts, and pullovers, Coats as they share similar shape as observed in confusion matrices in Fig. 6 and Fig. 7

Fig. 11 suggests that the CNN model has also considered the pixels constituting the shape of the trouser to identify the given image.

Utilizing the fine-tuned CNN, predictions were obtained for an out-of-sample image(distinct from the test set) to validate the model's efficacy in real-world scenarios. Illustrated in Fig. 12, the CNN accurately predicted the bag with a commendable accuracy of 98.7%. Furthermore, Fig. 13 showcases the interpretability of the model's decision-making process through LIME.

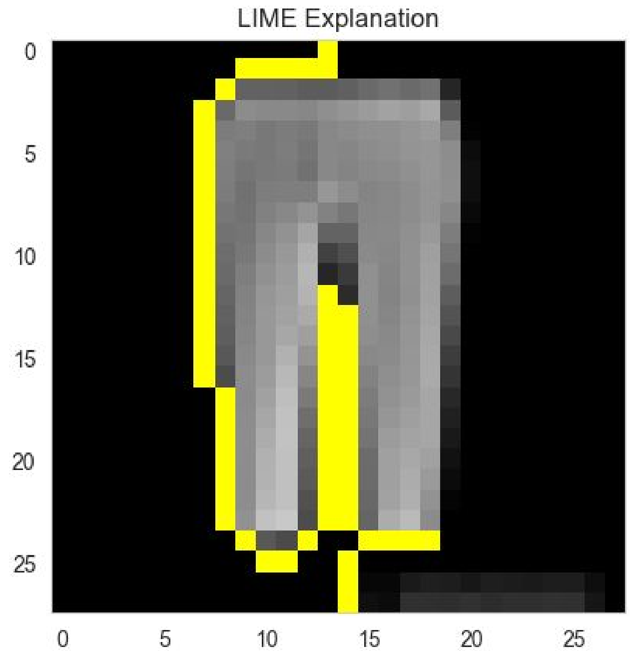

Fig. 11: LIME Explanation for CNN

Fig. 12: Out of Sample Image

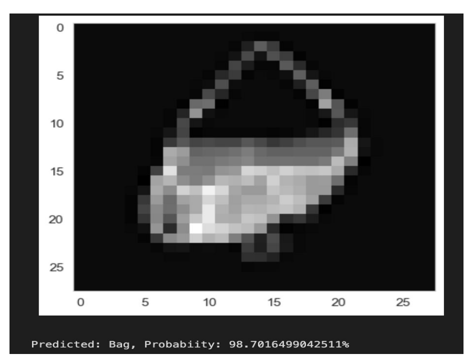

Fig. 13: Out of Sample Image

IV. Conclusion

In the exploration of the Fashion MNIST dataset, a comparative study between Convolutional Neural Networks (CNN) and XGBoost has unveiled intriguing insights into their performance. After observing the performance of two models, it is clear that CNN outperforms XGBoost. To validate the real-world applicability of our findings, an out-of-sample image sourced from Google was subjected to CNN. The CNN, being a more complex and adaptive architecture, demonstrated superior performance, reaffirming its robustness in handling diverse and nuanced visual patterns.

Classes such as 't-shirt/top,' 'pullover,' and 'shirt' exhibited lower accuracy. This trend prompted a deeper investigation into the interpretability of model decisions. The human intuition that these classes share similar shapes was validated through advanced interpretability tools such as SHAP and LIME. This comprehensive analysis not only showcases the superiority of CNN over XGBoost but also delves into the intricacies of model decisions for specific classes. The fusion of quantitative metrics, interpretability tools, and real-world validation collectively provides a holistic understanding of the models' strengths and limitations in the context of Fashion MNIST classification.

The models can be incorporated in the fashion industry to enhance quality control by accurately identifying defects during manufacturing, elevate customer experience through precise product categorization, and streamline inventory management on e-commerce platforms. This technological integration underscores a commitment to excellence, efficiency, and a superior customer journey in the dynamic landscape of fashion.

The future work is to focus on refining the model to better differentiate between certain classes with lower accuracy, notably 't-shirt/top,' 'pullover,' and 'shirt.' This effort aims to enhance the model's precision within these specific categories, contributing to an overall improvement in classification accuracy.

References

[1] N. Jmour, S. Zayen and A. Abdelkrim, "Convolutional neural networks for image classification," 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET), Hammamet, Tunisia, 2018, pp. 397-402, doi: 10.1109/ASET.2018.8379889.

[2] Xiao, H., Rasul, K. and Vollgraf, R., 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv preprint arXiv:1708.07747.

[3] N. Hamilton et al., "Enhancing Visualization and Explainability of Computer Vision Models with Local Interpretable Model-Agnostic Explanations (LIME)," 2022 IEEE Symposium Series on Computational Intelligence (SSCI), Singapore, Singapore, 2022, pp. 604-611, doi: 10.1109/SSCI51031.2022.10022096.

[4] S. Bhatnagar, D. Ghosal and M. H. Kolekar, "Classification of fashion article images using convolutional neural networks," 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, India, 2017, pp. 1-6, doi: 10.1109/ICIIP.2017.8313740.

[5] E. Xhaferra, E. Cina and L. Toti, "Classification of Standard FASHION MNIST Dataset Using Deep Learning Based CNN Algorithms," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2022, pp. 494-498, doi: 10.1109/ISMSIT56059.2022.9932737.

[6] F. Ren, Y. Liu, J. Zhang, J. Qian, P. Gao and H. Li, "Research on garment image classification algorithm based on machine learning," 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 2021, pp. 264-268, doi: 10.1109/IAECST54258.2021.9695531.

[7] W. Di, "A comparative research on clothing images classification based on neural network models," 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, Weihai, China, 2020, pp. 495-499, doi: 10.1109/ICCASIT50869.2020.9368530.

[8] M. Kayed, A. Anter and H. Mohamed, "Classification of Garments from Fashion MNIST Dataset Using CNN LeNet-5 Architecture," 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), Aswan, Egypt, 2020, pp. 238-243, doi: 10.1109/ITCE48509.2020.9047776.

[9] H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2684-2691, doi: 10.1109/IJCNN.2017.7966185.

[10] A. El-Sawy, E. B. Hazem, and M. Loey, "CNN for Handwritten Arabic Digits Recognition Based on LeNet-5," 2016 International Conference on Advanced Intelligent Systems and Informatics, Springer, Cham, 2016, pp. 566-575.

The github repo for the implementation can be found here - link
The website/blog link can be found here - link