# Threads of Understanding: A CNN Approach to Fashion MNIST Image Classification

1st Uma Sivakumar
*Texas A&M University*
College Station, USA
umas0697@tamu.edu

2nd Priyadharshini Ramesh Kumar
*Texas A&M University*
College Station, USA
priyadarshini01.r@gmail.com

3rd Hasitha Varada
*Texas A&M University*
College Station, USA
hasitha_16@tamu.edu

***Abstract*—The ever-changing field of machine learning applications presents researchers with ongoing problems in creating resilient models that can identify complex patterns in a variety of datasets. In-depth analysis of machine learning techniques is covered in this study, with a focus on how well they work with various fashion classification datasets. Knowing how flexible and effective these models are on broader datasets is crucial, as the fashion industry grows more and more data-driven. In this study, we describe our joint experience building and assessing machine learning models for the Fashion MNIST dataset. Above and beyond achieving maximum precision, we also tried to comprehend model interpretability, the effects of architectural decisions, and the wider ramifications for practical uses.**

***Index Terms*—Machine Learning, Neural Networks, Image Processing, Convolution Neural Network, Hyperparameters**

## I. Introduction

From retail to medical diagnostics, image classification appears to be a basic difficulty in our research of computer vision and machine learning. The Fashion MNIST dataset—a carefully selected set of grayscale photos showcasing a variety of fashion items in ten categories—was the subject of our scholarly inquiry. This dataset offered a pertinent baseline for assessing machine learning model performance in fashion categorization, acting as a modern replacement for the classic MNIST dataset. Our main goal was to use machine learning to build models that could correctly categorize the variety of fashion products in the dataset. This led to investigations into different model architectures, optimization strategies, and hyperparameter fine-tuning.The aim of this endeavor is a reflection of the actual dynamics of the digital fashion and e-commerce domains, where automated solutions play a crucial role in user experience enhancement, recommendation, and classification. An important turning point in our knowledge of the complexity of picture classification in the rapidly changing field of machine learning education was reached through our investigation of the Fashion MNIST dataset.

We became aware of the task's practical ramifications as we dug deeper into its complexities. In the current digital era, automated classification algorithms permeate every aspect of life, impacting everything from tailored fashion recommendations to online shopping experiences. Thus, our goal was not just to fulfill an academic requirement; rather, it was to equip us for the complex problems that arise when machine learning and the dynamic fashion industry come together.

Our objective in this academic endeavor was to navigate the complexities of model selection, training, and evaluation. Because of the diversity included in the Fashion MNIST dataset, it was necessary to carefully weigh the advantages and disadvantages of different machine learning architectures. Prioritizing different evaluation metrics and gaining a deeper understanding of the elements driving model performance was our main objective when we set out on this trip. This will help us make well-informed decisions in practical applications down the road.

## II. About the Data-set

The Fashion-MNIST dataset [8], which includes 70,000 fashion items across ten categories, is the one that was used. This dataset resamples the original image at various resolutions to effectively serve various frontend components. It is based on the assortment found on Zalando's website.To create Fashion-MNIST, it utilize the thumbnails of 70,000 distinct goods. The conversion process is shown graphically.The dataset is kept in the same file format as the MNIST data collection and is separated into training and test sets.

The Fashion MNIST dataset's exploratory data analysis (EDA) provided insightful information about the fundamental qualities of the various fashion products it contains. After carefully analyzing the structure of the information, we were able to identify 60,000 grayscale photos that are categorized into ten different fashion groups. The balanced nature of the dataset was shown by descriptive statistics, which gave an insight into the distribution of class labels, from class disparities as shown in Figure 1.

A greater comprehension of the intricacy of the information was made possible by visualizations, which included everything from class distributions to sample photos as shown in Figure 2. Pie charts and histograms showed the relative sizes of each fashion category, providing guidance for the construction of subsequent models by drawing attention to possible issues arising

Principal Component Analysis (PCA) is a crucial method for effective visualization while examining and comprehending the Fashion MNIST dataset. Direct visualization is hampered by the high-dimensionality of the grayscale images, which were all initially recorded as a grid of pixel values. In order to overcome this, principal component analysis (PCA) is used

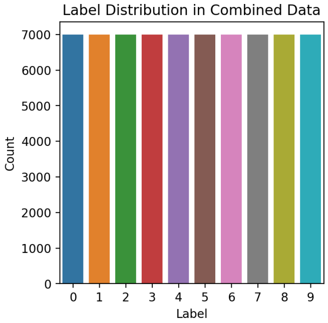

Fig. 1. Class Distribution of Fashion MNIST dataset

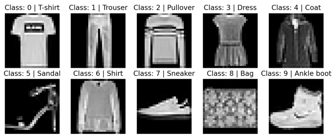

Fig. 2. Sample images of Fashion MNIST dataset

to reduce the dimensionality of the dataset by turning it into a collection of uncorrelated characteristics.

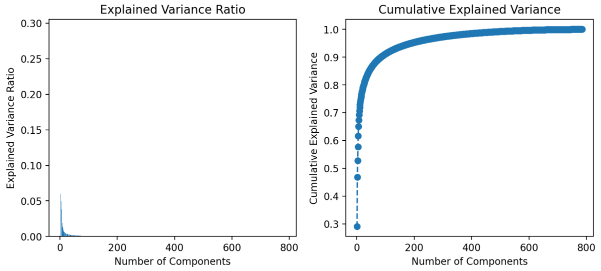

Fig. 3. variance explanation of Fashion MNIST dataset

PCA provides a more lucid representation of the dataset by projecting the data onto a lower-dimensional subspace, usually in 2D or 3D. This enables researchers to discern the prevailing patterns and structures within the fashion-related photos. This decrease contributes to data compression, which may accelerate later machine learning algorithms in addition to helping to understand the distribution of various apparel categories. Through the use of PCA, the Fashion MNIST dataset may be visualized in a way that makes it easier to read and explore, leading to a greater comprehension of the underlying patterns present in the varied collection of fashion photos.

## III. Review of the Literature

With the main objective of attaining accurate classification with minimum training, several CNN models, including cnn-dropout-1, cnn-dropout-2, cnn-dropout-3, and cnn-simple, were introduced in the work of [2]. The results show a significant increase in accuracy, surpassing the greatest result achieved using Support Vector Machines (SVM) at 89.7% to an astounding 99.1%. This demonstrates how well CNNs handle the complexities of datasets containing fashion products.

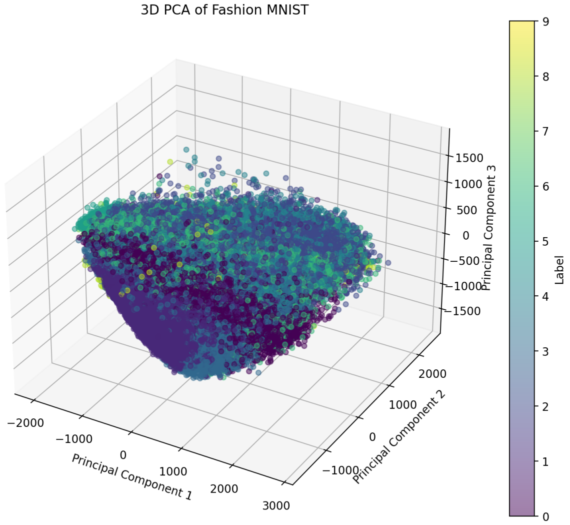

Fig. 4. PCA visualization of Fashion MNIST dataset

CNN optimization for fashion product categorization heavily relies on architectural considerations and hyperparameter adjustment. Different designs with different convolutional layers, filter sizes, and fully connected layers were investigated, as reported in [3]. Notably, Architecture 3 outperformed the others with an accuracy rate on the MNIST dataset that exceeded 99%. This emphasizes how crucial careful architectural design is to attaining high accuracy rates.

Additional comparison evaluations, as reported in [4], demonstrate how well LeNet-5-based CNNs outperform conventional classifiers like Evolutionary Deep Learning (EDEN) and Support Vector Classifiers (SVC). On the Fashion MNIST dataset, the CNN model demonstrated over 98% accuracy, proving its supremacy over the most advanced setups. Furthermore, [6] presented a 15-layer Multiple Convolutional Neural Network that outperformed previous research in Fashion-MNIST classification accuracy, obtaining an astounding 94.04%. [7] presented a model that uses two fully connected layers with dropout layers, max-pooling layers, three deep convolution layers, and other layers to solve the multiclass classification problem. The study showed that the Adam optimizer improved test accuracy compared to alternative CNN configurations.

## IV. Models Experimented

A systematic strategy customized to the fashion categorization objectives is necessary to choose the best model for the Fashion MNIST dataset. Model designs, particularly Convolutional Neural Networks (CNNs), are selected based on their capacity to capture spatial hierarchies in grayscale images, following a thorough exploratory data analysis (EDA) to comprehend the complexities of the dataset. Transfer learning improves the procedure, and optimization involves data augmentation and hyperparameter adjustment. Model performance is assessed using task-specific evaluation criteria, which guarantee strong generalization via cross-validation. The model selection is further refined by taking application requirements, computational resource impact, and interpretability and complexity into account. Researchers are able to customize their decisions to Fashion MNIST's distinct features thanks to this methodical methodology.

The choice of base model in the comparative examination of Fashion MNIST model performance is determined by a strategic assessment of interpretability and complexity. As a baseline, an Artificial Neural Network (ANN) is used to illustrate the benefits and limitations of the selected Convolutional Neural Network (CNN). CNNs are superior at identifying spatial hierarchies, whereas ANNs are more straightforward and comprehensible. This comparison approach highlights the distinctiveness of architecture in image-based datasets and offers subtle insights into how model complexity affects fashion categorization tasks. By means of thorough review and methodical experimentation, the research seeks to provide insightful information about the neural network's suitability for fashion image classification.

As a part of preprocessing the dataset was split into training, testing and validation in the ratio of 60:20:20.Then the X_train, X_val, X_test data was normlized by scaling the pixel values to a range between 0 and 1. This normalization helps the model converge faster during training.X_train, X_val, and X_test represent the pixel values of the images in the training, validation, and test sets, respectively. The division by 255.0 scales the pixel values to the range [0, 1], as the original pixel values are in the range [0, 255].

### *A. Artificial Neural Network*

A key component of machine learning is the Artificial Neural Network (ANN), which is modeled after the complex networks found in the human brain. Neural Networks (ANNs) are composed of linked layers of nodes that extract complex patterns from data. The input, hidden, and output layers are made up of learning through iterative weight adjustments. Showcasing their adaptability in challenging applications, ANNs perform exceptionally well in a variety of tasks like pattern identification, natural language processing, and image and audio recognition.

With the aid of the Keras library, the Artificial Neural Network (ANN) architecture is developed in a sequential fashion as shown in Figure 5. To ensure compatibility with later dense layers, the model starts with a Flatten layer that converts the input data, which consists of 28x28 grayscale images, into a one-dimensional array. The network becomes non-linear in the first dense layer, which has 64 neurons and uses the rectified linear unit (ReLU) activation function. Using the flattened input, this layer acts as a first feature extractor, picking out pertinent patterns. The network is then able to capture higher-level representations thanks to the subsequent refinement of the learned features by a second dense layer consisting of 128 neurons and ReLU activation.

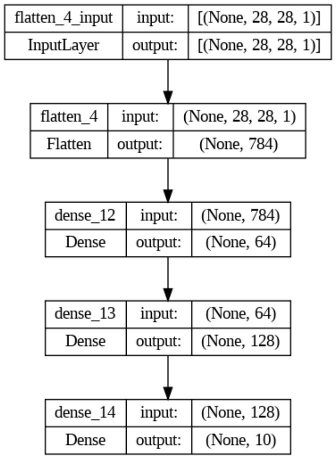

Fig. 5. ANN Architecture

The model can generate probability distributions for each of the 10 fashion categories included in the Fashion MNIST dataset thanks to the use of the softmax activation function in the last dense layer, which consists of 10 neurons. This architecture strikes a balance between interpretability and network depth, making it a simple yet efficient solution for picture categorization problems. The network's input flows systematically thanks to the layers' sequential organization, and the activation functions' selection makes it easier to learn intricate patterns that are necessary for precise categorization.

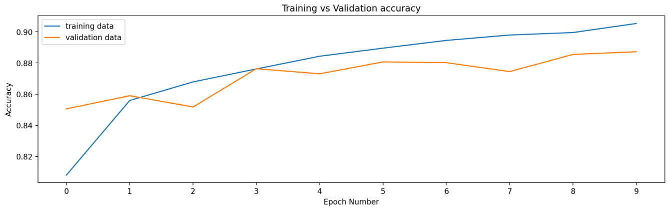

Fig. 6. Accuracy plot for ANN

A visual depiction of a machine learning model's performance during training is the training vs. validation accuracy plot. This graphic, which is typically seen during the iterative training phase, shows how well the model generalizes to new, unknown data and learns from the training set. In a perfect world, there would be no overfitting and a steady increase in both training and validation accuracy. The two curves' close convergence indicates how well the model generalizes to a range of data points. But, if a discernible gap appears, where the training accuracy exceeds the validation accuracy, this may indicate overfitting—that is, the model has learned the training data by heart but finds it difficult to apply that knowledge to novel situations.To balance accurate learning and efficient generalization, hyperparameters, model complexity, or regularization strategies must be carefully considered and adjusted. This can be done by tracking and analyzing the training vs. validation accuracy plot. From Figure 6 we find out that the ANN model is not overfitting or underfitting and the model generalizes well to new and unknown data.

### B. Convolutional Neural Network

In the context of visual data processing, Convolutional Neural Networks (CNNs) offer notable improvements over standard Artificial Neural Networks (ANNs) and signal a paradigm shift in the field of image-based machine learning. CNNs perform exceptionally well on tasks like image classification because they are extremely good at capturing local patterns, spatial hierarchies, along with Automatic feature extraction and translation invariance within images. CNNs use convolutional layers to methodically examine localized features, in contrast to ANNs, which interpret input data as a flat vector and may find it difficult to maintain spatial links. CNNs may learn hierarchical representations by identifying complex patterns and textures found in photos thanks to this architectural difference. CNNs are a better option because of their inherent strengths when it comes to the Fashion MNIST dataset, which contains grayscale photos of clothing items and requires a sophisticated grasp of visual aspects.

Here we have modeled three different CNN's in the order of increasing model complexity and have compared their training vs validation accuracy graphs to further understand the working of neural network on the Fashion MNIST dataset.

### C. Model Comparisons

All the models are a version of CNN. Let's examine the details of each model in more detail:

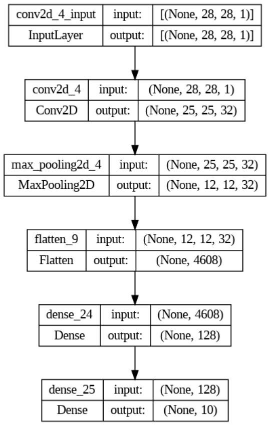
conv2d_4_input | input: | [(None, 28, 28, 1)]
InputLayer | output: | [(None, 28, 28, 1)]
conv2d_4 | input: | (None, 28, 28, 1)
Conv2D | output: | (None, 25, 25, 32)
max_pooling2d_4 | input: | (None, 25, 25, 32)
MaxPooling2D | output: | (None, 12, 12, 32)
flatten_9 | input: | (None, 12, 12, 32)
Flatten | output: | (None, 4608)
dense_24 | input: | (None, 4608)
Dense | output: | (None, 128)
dense_25 | input: | (None, 128)
Dense | output: | (None, 10)

Fig. 7. Architecture of Model 1

*1) Model 1:* Model 1 is a basic Convolutional Neural Network (CNN) with a 4x4 kernel and a single convolutional layer with 32 filters for image classification. Downsampling is done via max-pooling with a 2x2 pool size, and the flattened output is connected to a dense layer of 128 neurons via the ReLU activation. The last layer uses softmax activation to classify into the Fashion MNIST dataset's ten fashion categories.

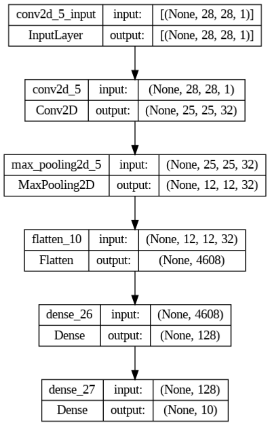
conv2d_5_input | input: | [(None, 28, 28, 1)]
InputLayer | output: | [(None, 28, 28, 1)]
conv2d_5 | input: | (None, 28, 28, 1)
Conv2D | output: | (None, 25, 25, 32)
max_pooling2d_5 | input: | (None, 25, 25, 32)
MaxPooling2D | output: | (None, 12, 12, 32)
flatten_10 | input: | (None, 12, 12, 32)
Flatten | output: | (None, 4608)
dense_26 | input: | (None, 4608)
Dense | output: | (None, 128)
dense_27 | input: | (None, 128)
Dense | output: | (None, 10)

Fig. 8. Architecture of Model 2

*2) Model 2:* Model 2, a variant of Model 1, modifies the spatial sampling strategy by introducing explicit strides of [1, 1] in the convolutional layer. This alteration affects the way the convolutional layer scans the input data. The rest of the architecture, which includes flattening, max-pooling, and thick layers for later feature extraction and classification, is consistent with Model 1.

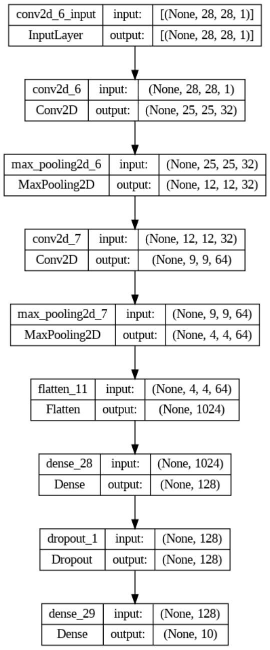
conv2d_6_input | input: | [(None, 28, 28, 1)]
InputLayer | output: | [(None, 28, 28, 1)]
conv2d_6 | input: | (None, 28, 28, 1)
Conv2D | output: | (None, 25, 25, 32)
max_pooling2d_6 | input: | (None, 25, 25, 32)
MaxPooling2D | output: | (None, 12, 12, 32)
conv2d_7 | input: | (None, 12, 12, 32)
Conv2D | output: | (None, 9, 9, 64)
max_pooling2d_7 | input: | (None, 9, 9, 64)
MaxPooling2D | output: | (None, 4, 4, 64)
flatten_11 | input: | (None, 4, 4, 64)
Flatten | output: | (None, 1024)
dense_28 | input: | (None, 1024)
Dense | output: | (None, 128)
dropout_1 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_29 | input: | (None, 128)
Dense | output: | (None, 10)

Fig. 9. Architecture of Model 3

*3) Model 3:* Model 3 includes two convolutional layers, which adds more complexity. Max-pooling comes after the first convolutional layer with 32 filters, and a second convolutional layer with 64 filters is added after that. Max-pooling is used once again, and then dense layers and flattening. Notably, following the initial thick layer comes a dropout layer with a dropout rate of 0.2. During training, dropout randomly deactivates a portion of neurons to assist prevent overfitting.

The classifier is the last dense layer with softmax activation.

When contrasting the three models created for the Fashion MNIST dataset, it is evident that each one has unique architectural elements that are intended to improve image categorization capabilities. Model 1 sets the standard with a max-pooling and dense layer in front of a single convolutional layer with 32 filters. This simple approach establishes a strong basis for classification problems by capturing basic spatial hierarchies within the grayscale images. Model 2 is almost exactly the same as Model 1, except it makes explicit changes to the convolutional layer, which affects the spatial sampling strategy. This modification makes it possible for the model to handle incoming data in a slightly different way, maybe capturing various spatial relationships. Model 2 investigates the effect of a subtle convolutional layer arrangement on classification accuracy while keeping things simple.

Model 3 incorporates extra layers and regularization strategies as complexity increases in order to improve feature extraction and avoid overfitting. Expanding upon Model 1's architecture, Model 3 adds a second convolutional layer with 64 filters, which improves the model's ability to recognize complex patterns. In order to improve generalization, dropout is incorporated after the first thick layer. This creates a regularization process by randomly deactivating a portion of neurons during training. In order to guarantee that the model may be adjusted to the complex and varied fashion categories found in the Fashion MNIST dataset, this strategic augmentation attempts to achieve a compromise between complexity and interpretability. The transition from Model 1 to Model 3 is a thoughtful process that provides a range of options to take into account depending on the particular requirements of classes.

Now comparing the accuracy plot for each model we can see that Model 1, as shown in Figure 10 has a comparatively wider gap in terms of the accuracy plot which can be due to the simplicity of the model

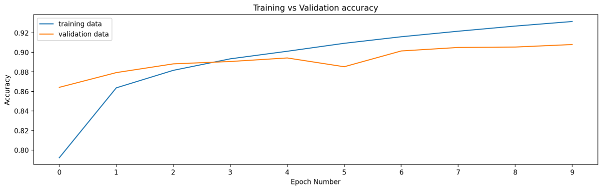

Fig. 10. Accuracy plot for Model 3

Now when comparing Figure 11 and Figure 12 we can notice that Superior performance is demonstrated by CNN Models 2 and 3, as well as by recall, precision, and F1-score. Model 2 performs competitively despite having a simpler design with only two convolutional layers and a faster compilation time. The large difference between Model 2's high training accuracy of 97.4% and its somewhat lower testing accuracy of 90.6%, which indicates a 7% accuracy divergence, however, points to overfitting. On the other hand, Model 3, which has two pairs of convolutional layers, performs better overall and is the better option when it comes to addressing overfitting issues.

```
F1-score: 0.9030001359566023
```

Fig. 11. F1 - score for Model 3

## V. Conclusion

Finally, within the framework of the Fashion MNIST dataset, this paper has provided a comprehensive comparative examination of three different convolutional neural network (CNN) models: Model 1, Model 2, and Model 3. After a thorough analysis of precision, recall, and F1-score measures, all models had similar overall accuracies of 90%. Establishing a baseline with simplicity, Model 1 serves as a core architecture, whereas Model 2 introduces subtle modifications with stated strides. In an attempt to alleviate overfitting issues, Model 3 adds another convolutional layer and dropout, but at the expense of more complexity. Notably, we also developed an artificial neural network (ANN) for comparison, so our study goes beyond CNNs.The ANN's inclusion provides a useful baseline that draws attention to the special benefits and difficulties that CNN architectures have when it comes to classifying fashion images. As the field develops, this research offers subtle insights into the trade-offs associated with various models, providing practitioners and researchers with important things to think about when trying to improve the generalization and accuracy of neural networks in a variety of image classification tasks. ( source code: Click Here for colab and Click Her for Blog post )

## References

[1] Duan, C., Yin, P., Zhi, Y., Li, X. (2019). Image classification of Fashion-MNIST data set based on VGG network. In *Proceedings of 2019 2nd International Conference on Information Science and Electronic Technology (ISET 2019)* (Vol. 19). International Informatization and

[2] Henrique, A. S., Fernandes, A. M. R., Lyra, R., Leithardt, V. R. Q., Correia, S. D., Crocker, P., Dazzi, R. L. S. (2021). Classifying Garments from Fashion-MNIST Dataset Through CNNs. *Advances in Science, Technology and Engineering Systems Journal, 6*(1), 989-994.

[3] Kadam, S. S., Adamuthe, A. C., Patil, A. (2020). CNN Model for Image Classification on MNIST and Fashion-MNIST Dataset. *Journal of Scientific Research.*

[4] Kayed, M., Anter, A., Mohamed, H. (2020). 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)* (pp. 238-243). Aswan, Egypt. doi:10.1109/ITCE48509.2020.9047776.

[5] K V, Greeshma, K., Sreekumar. (2019). Fashion-MNIST classification based on HOG feature descriptor using SVM. *International Journal of Innovative Technology and Exploring Engineering, 8*, 960-962.

[6] Nocentini, O., Kim, J., Bashir, M. Z., Cavallo, F. (2022). Image Classification Using Multiple Convolutional Neural Networks on the Fashion-MNIST Dataset. *Sensors, 22*(23), 9544.

[7] Saranya, M. S., Geetha, P. (2022). Fashion Image Classification Using Deep Convolution Neural Network. In A. (Eds.) *Computer, Communication, and Signal Processing. ICCCSP 2022.* IFIP Advances in Information and Communication Technology (Vol. 651). Springer, Cham.

[8] Xiao, Han et al. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms." ArXiv abs/1708.07747 (2017): n. pag.

[9] Yian Seo, Kyung-shik Shin. (2019). Hierarchical convolutional neural networks for fashion image classification. *Expert Systems with Applications, 116*, 328-339.