C964 Computer Science Capstone: Clothesifier

Andrew Staus

Western Governors University

010312389

Table of Contents

[Part A: Project Proposal for Business Executives 4](#_Toc1515453419)

[Letter of Transmittal 4](#_Toc1044087116)

[Project Recommendation 7](#_Toc1721750458)

[Problem Summary 7](#_Toc205640471)

[Application Benefits 7](#_Toc739883765)

[Application Description 8](#_Toc1967641075)

[Data Description 8](#_Toc1580263866)

[Objectives and Hypothesis 8](#_Toc1084915425)

[Methodology 9](#_Toc1455093807)

[Funding Requirements 9](#_Toc1716359054)

[Data Precautions 10](#_Toc821607173)

[Developer’s Expertise 10](#_Toc742297278)

[Part B: Project Proposal 11](#_Toc1516064891)

[Problem Statement 11](#_Toc602847551)

[Customer Summary 11](#_Toc411678477)

[Existing System Analysis 11](#_Toc1668713332)

[Data 11](#_Toc1911399030)

[Project Methodology 12](#_Toc1307313575)

[Project Outcomes 12](#_Toc1967172832)

[Implementation Plan 13](#_Toc306674583)

[Evaluation Plan 14](#_Toc849271423)

[Resources and Costs 15](#_Toc1695791095)

[Timeline and Milestones 17](#_Toc787992355)

[Part C: Application 19](#_Toc1767968421)

[Part D: Post-implementation Report 19](#_Toc22579049)

[A Business (or Organization) Vision 20](#_Toc1496854501)

[Datasets 20](#_Toc645781746)

[Data Product Code 22](#_Toc157285634)

[Objective (or Hypothesis) Verification 24](#_Toc1315546441)

[Effective Visualization and Reporting 24](#_Toc656892821)

[Accuracy Analysis 24](#_Toc1956352798)

[Application Testing 24](#_Toc871280264)

[Application Files 24](#_Toc1639588139)

[User Guide 24](#_Toc1125023771)

[Summation of Learning Experience 24](#_Toc1702287338)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

December 23, 2022

William Hamilton, CTO

Clothier Company

2130 21 Street NE

Calgary, AB T2T 1T2

Dear Mr. Hamilton,

Clothier is a rapidly growing e-commerce business, serving tens of thousands of customers each month with a wide selection of high-quality clothing items. As the business and its customer base have grown, so has the inventory. However, customers have reported difficulty finding the items they are interested in purchasing, leading to a high rate of abandonment and lost sales.

The Clothier website already includes the functionality to sort items by category, but this feature is not being utilized due to the lack of categorization for the existing and growing inventory, and manually labelling each item in the inventory would be cost-prohibitive. To address this issue, I propose implementing a machine learning system to automatically categorize all items in the Clothier inventory, as well as provide categorization for new items as they are added. A machine learning system can be implemented to automatically categorize all items in the Clothier inventory and provide the ability to also provide categorization for new items as they are added. The system will be deployed as an API that can be integrated directly into Clothier’s existing inventory system, providing seamless integration for business users.

The categorization of clothing will provide a simpler customer experience, and lead to more conversion as customers easily find the items they are looking for. Categorization will also benefit the product and marketing teams by providing the ability to generate better analyze which types of products are popular, and which are not.

The development and deployment of the machine learning system will require a time investment of 75 working hours for one developer at a rate of $100 per hour, totalling $7,500. This investment represents a cost-effective solution for improving inventory management and customer experience for the Clothier business. The service will be developed using software development tools and libraries that are available at no cost and with no licensing fees. This means that Clothier will not incur any additional costs beyond the time investment for the developer. The service will be deployed on a serverless cloud platform, eliminating the need for Clothier to manage and maintain infrastructure. This will further reduce costs and improve the overall efficiency of the system. The serverless platform does incur a minimal usage fee of $0.20 per one million requests, but this limit far exceeds Clothier's current needs. As a result, costs for the foreseeable future will effectively be zero.

Overall, the development and deployment of the machine learning system will provide significant benefits for the Clothier business and its customers, with minimal costs and no maintenance required.

I have the necessary education and experience to successfully develop and implement the automated categorization system for the Clothier e-commerce platform. My education, culminating in a Bachelor of Computer Science, in addition to a decade of industry experience in software development, project management, and data science has provided me with a strong foundation in software development and data science.

Sincerely,

## Project Recommendation

### Problem summary

**.** The Clothier e-commerce business is facing a significant challenge in terms of inventory management. Users have reported difficulty navigating the website due to a lack of categorization, leading to high rates of abandonment and lost sales. The e-commerce platform being used by Clothier includes the ability to sort items by category, but this feature is not being utilized due to the lack of category labels for the existing and future inventory.

Manually reviewing and categorizing each item in the inventory would be cost and time prohibitive, due to the large volume of items. An automated system would provide a quick and accurate solution, allowing the entire inventory to be categorized in minutes rather than the months it would take humans to perform the same task.

By resolving the categorization problem, the Clothier business will be able to improve the customer experience and provide business stakeholders with the ability to analyze relevant information.

### Application benefits

**.** The implementation of the Clothesifier automated categorization system will bring several benefits to the Clothier e-commerce business.

First, the categorization of items on an e-commerce site is expected by customers. Research shows that 80% of customers browsing the Clothier website leave without purchasing an item, citing difficulties in navigating the site and finding the desired products. By integrating the Clothesifier system directly into the inventory system, customers will be able to search for products by type, leading to higher conversions and a better user experience.

Additionally, business stakeholders in the sales, marketing, and product teams have reported difficulty in understanding customer trends due to a lack of category information. The automated categorization system will provide these users with better insights, enabling them to create more targeted marketing campaigns and improve the overall customer experience.

### Application description

**.** The Clothesifier automated categorization system will be able to accept an image of an inventory item and return the category for that item, along with a confidence metric indicating the model's level of certainty about the prediction. The system will be deployed as a web API, allowing for seamless integration with the existing Clothier e-commerce systems. This will allow the Clothier to easily categorize their inventory and improve efficiency and accuracy in their business operations.

### Data description

**.** The Clothesifier system will be trained to recognize the difference between clothing items using a freely available dataset called Fashion MNIST. The dataset consists of 7,000 examples of 28 by 28 images of clothing articles, along with categorical labels for each image. The data is of high quality and does not contain any outliers. However, it is possible that the images in the Clothier inventory may not be of the same standard, which could limit the performance of the model. To address this potential limitation, we will perform data augmentation to make the dataset more representative of real-world data and improve the model's ability to perform well in any circumstance.

### Objectives and hypothesis

**.** Upon completion of this project, the Clothesifier system will be able to accept an image from the Clothier inventory and return the category for that item with 85% accuracy. In addition, the model's confidence levels for the classification will be made available for examination by business users.

The hypothesis is that a model trained on the dataset will be able to generalize to the Clothier data and perform accurate image classification. This will allow the Clothier to more easily categorize and manage their inventory, improving efficiency and accuracy in their business operations.

### Methodology

**.** A waterfall methodology will be followed to develop the system, following a linear, sequential approach. This methodology involves dividing the development process into distinct phases, with each phase building upon the previous one. The phases to be followed are:

* Requirements – Stakeholders will be engaged to gather the specifications needed for the final system. Integration specialists with knowledge of the Clothier e-commerce platform will provide input.
* Design – Interface, deployment architecture, model, and data will be detailed to ensure that all systems integrate.
* Implementation – The code will be written and the systems deployed and integrated.
* Verification – The system will be tested to ensure that it behaves correctly with both valid and invalid input.
* Maintenance – Any bugs, issues, or additional features missed during verification will be assessed and resolved.

### Funding requirements

**.** The application will be implemented using a serverless architecture provided by AWS Lambda. This will eliminate the need for infrastructure management and maintenance. While there is a usage fee of $0.20 per one-million requests, the first one-million requests will be free. Based on the size of the Clothier inventory, it is highly unlikely that there will be any costs associated with running the service.

For the project development, we will use development tools and software libraries that are available at no cost and do not require any licensing fees. The estimated development time for this project is 75 working hours for two software developers at a rate of $100 per hour, for a total cost of $7,500.

### Data precautions

**.** The data used to train the model is publicly available and does not contain any sensitive or proprietary information. In addition, the nature of the system being developed ensures that the data is never stored in an accessible database.

### Developer’s expertise

**.** We propose to assign the project to a developer with a decade of experience in developing and deploying software, managing projects, and data science. This developer holds a Bachelor of Science in Computer Science, which complements their practical experience. We believe that this individual is an excellent candidate for developing the system, given their combination of education and experience.

# Part B: Project Proposal

## Problem Statement

The Clothier e-commerce platform is experiencing low conversion rates and suboptimal customer experiences due to its inadequate categorization system for the products it offers. To address this issue, we propose the implementation of an automated classification system deployed as a REST API. This system will accept images of clothing as input and accurately return the corresponding category for the item.

## Customer Summary

The Clothier e-commerce platform serves as a retail destination for customers looking to purchase clothing and accessories. However, the platform's current categorization system is inadequate, leading to poor customer experiences and low conversion rates. Our proposed application, a data product in the form of a REST API endpoint, aims to address this problem by providing a reliable and accurate method for categorizing both new and existing inventory on the platform. The API can be easily integrated into the platform's inventory pipeline, allowing for the automatic categorization of newly added items and streamlining the inventory management process for the platform's developers. Additionally, the API can be utilized by Clothier's analysis team to accurately categorize the current inventory, improving the overall user experience and conversion rates on the platform.

## Existing System Analysis

The Clothier e-commerce platform currently lacks a functional categorization system, leading to poor customer experiences and low conversion rates. While the platform does support the integration of REST APIs, the lack of a reliable and accurate categorization system remains a significant issue. Our proposed solution, a data product in the form of a REST API endpoint, aims to address this problem by providing a comprehensive and automated method for categorizing both new and existing inventory on the platform. The API can be easily integrated into the platform's existing infrastructure, allowing for the seamless implementation of a reliable categorization system and improving the overall user experience on the platform.

## Data

The proposed application will utilize the Fashion MNIST dataset, a publicly available and free source, to train the classification model. This dataset consists of 60,000 training images and 10,000 test images of clothing and accessories, each labelled with the corresponding category.

Any outliers or incomplete data within the Fashion MNIST dataset will be identified and handled during the preprocessing phase. Outliers may be removed from the dataset or treated in an appropriate manner depending on the specific circumstances and impact on model performance.

## Project Methodology

Throughout the development of this project, we will follow the Waterfall project management methodology. This involves a sequential and linear progression through the following phases: requirements gathering, design, implementation, testing, and maintenance.

The Waterfall methodology is a cost-effective and efficient approach for small projects and those where requirements are expected to remain stable. It allows for a clear and structured development process, with each phase building upon the previous one.

During the requirements gathering phase, we will engage with the Clothier IT team to understand their needs and requirements for the Clothifier application. This will involve discussions about the desired features and functionality, as well as any specific requirements or constraints.

Once the requirements have been gathered and clarified, we will move into the design phase, where we will determine the architecture of the application. This will involve deciding on the technologies and frameworks to be used, as well as the overall structure and design of the application.

Next, we will move into the implementation phase, where we will set up the programming environment, and begin coding the application. We will follow industry best practices for coding and testing to ensure that the application is of high quality and meets the requirements outlined in the earlier phases.

Once the implementation is complete, we will move into the testing phase, where we will verify the proper behaviour of the application across both intended and unintended inputs. We will allow for extra time in this phase to resolve any bugs that are discovered.

Finally, we will move into the maintenance phase, where we will address any remaining bugs or regulatory concerns and ensure that the application is operating optimally. Ongoing maintenance will be performed as needed to ensure the continued success of the Clothifier application.

## Project Outcomes

Upon completion of the project, the API endpoint will be available for use by the customer. To demonstrate the functionality of the API, a browser-based GUI will be provided that allows a user to upload a single image and test the results. The demonstration GUI will include documentation on its proper usage.

The API endpoint will adhere to the constraints of the REST architectural style and will include documentation on its proper use. The customer will be able to access the API endpoint to integrate its functionality into their systems and applications. The API endpoint and associated documentation will provide clear instructions on how to use the API consistently and predictably.

## Implementation Plan

The proposed approach for implementing this project is a top-down approach, in which each step is completed in order. This approach is being adopted to ensure that functionality is verified after each step, ensuring that all modules will be able to integrate into a viable product. The steps to be followed are as follows:

1. Conduct exploratory data analysis (EDA) using the Pandas, Matplotlib, Sci-Kit Learn, and Plotly Python libraries to understand the characteristics of the Fashion MNIST dataset and determine the best approach for developing the machine learning model.
2. Reshape the data using the Numpy library and augment it with additional examples using Numpy and Sci-Kit Learn.
3. Compile a convolutional neural network (CNN) model using the Keras TensorFlow interface.
4. Train the CNN on the augmented Fashion MNIST dataset for 100 epochs to achieve sufficient accuracy for the requirements.
5. Analyze model performance using Matplotlib and Sci-Kit Learn to verify that accuracy meets the requirements.
6. Export the model in a TensorFlow Lite (tflite) format.
7. Write the API code using the FastAPI, Pydantic, Numpy, Pillow and TFLite\_Runtime Python libraries. Use the Mangum library to wrap the app in an event handler for compatibility with AWS Lambda.
8. Package and deploy the API code as a function on AWS Lambda.
9. Compile the required libraries for the API to run on an Amazon Linux environment, package them, and deploy them as a layer on AWS Lambda.
10. Develop a basic frontend for API functionality testing using HTML5, CSS, and JavaScript. Use the Jquery library to access the API hosted on the AWS Lambda site, the Bootstrap CSS and JavaScript libraries to provide a responsive grid interface, and the Chart.JS library to provide visualizations of the results.
11. Store the project on a GitHub repository and deploy the webpage using Github Sites.

## Evaluation Plan

To ensure the successful and accurate functioning of the Clothifier application, we will implement a thorough evaluation plan to validate the performance of each module at multiple stages throughout the development process.

### Model module testing

**.** Before training, 1000 clothing images will be reserved for validation. After training the model on the remaining 6000 images, we will assess its accuracy on the validation set to determine its performance on items that the model was not trained on. A successful test will have the model score over 80% accuracy on the validation set.

### API module testing

**.** We will test the API to confirm that it responds with an appropriate classification when provided with a valid Base64 image file at the correct endpoint. In the event of any exceptions, the API should return a message containing the exception information for debugging purposes.

### Webpage module testing

**.** The webpage will be tested to ensure that it accepts the upload of an image file and displays the classification results from the API to the user. Appropriate error alerts should be displayed to the user if any exceptions arise during the classification process.

### System testing

**.** In addition to testing the individual modules, we will also conduct system testing to ensure that the entire Clothifier application is functioning correctly and seamlessly as a whole. This will involve testing the integration and communication between the different modules, as well as the overall user experience and performance.

### Post-deployment evaluation

**.** After the Clothifier application is deployed and in use, we will conduct ongoing evaluations to assess its performance and effectiveness. This will include monitoring the accuracy of the classification results, as well as tracking any issues or errors that may arise in the API or webpage. We will also solicit feedback from users of the application to gather insights on its usability and user experience. Based on the results of these evaluations, we will make any necessary updates or improvements to the Clothifier application to ensure its continued success and value for the Clothier e-commerce platform.

## Resources and Costs

The deployment and operation of the Clothifier API will have an initial cost of $0.00, as it will be hosted on the serverless cloud platform AWS Lambda. AWS Lambda offers a free tier for the first 1 million requests per month, with a tiered pricing system for additional requests.

The Clothifier application is lightweight and has a normal compute duration of fewer than 0.002 GB-seconds, utilizing the minimum memory tier of 128 MB. Given the current size of the Clothier inventory, it is expected that usage will remain within the free tier of 1 million requests per month. However, if the inventory size increases and additional requests are made exceeding the free tier, the cost for each additional 1 million classifications is expected to be less than $0.25.

**Table 1**

*AWS Lambda Compute Pricing*

|  |  |  |
| --- | --- | --- |
| Usage Cost Per: |  |  |
| Billion GB-seconds / month | Duration (GB-second) | 1M requests |
| First 6 | $0.0000166667 | $0.20 |
| Next 9 | $0.000015 | $0.20 |
| Over 15 | $0.0000133334 | $0.20 |

**Table 2**

*AWS Lambda Memory Pricing*

|  |  |
| --- | --- |
| Memory (MB) | Price per 1ms |
| 128 | $0.0000000021 |
| 512 | $0.0000000083 |
| 1024 | $0.0000000167 |

## Timeline and Milestones

Development and deployment are expected to require 75 hours of labour over the course of 2 weeks. This schedule is the result of constraints of the assigned software engineer’s schedule. Completion of the project is dependent on the milestones described below.

**Table 3**

*Timeline and Milestones*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mile-stone | Prere-quisites | Activity | Resource Assigned | Hours | Start | End |
| 1 | - | Requirements approval | Project Manager | 8 | 2022-12-05 | 2022-12-05 |
| 2 | 1 | Architecture Design | Software Engineer | 8 | 2022-12-06 | 2022-12-06 |
| 3 | 1 | Development Environment configuration | Software Engineer | 2 | 2022-12-07 | 2022-12-07 |
| 4 | 3 | Exploratory Data Analysis | Software Engineer | 3 | 2022-12-07 | 2022-12-07 |
| 5 | 4 | Model Creation | Software Engineer | 12*\** | 2022-12-07 | 2022-12-08 |
| 6 | 5 | Model Module Testing | Software Engineer | 1 | 2022-12-08 | 2022-12-08 |
| 7 | 2 | API Development | Software Engineer | 8 | 2022-12-08 | 2022-12-09 |
| 8 | 6, 7 | API Deployment | Software Engineer | 2 | 2022-12-09 | 2022-12-09 |
| 9 | 8 | API Module Testing | Software Engineer | 1 | 2022-12-12 | 2022-12-12 |
| 10 | 2 | Frontend Website Creation | Software Engineer | 8 | 2022-12-12 | 2022-12-13 |
| 11 | 10 | Frontend Website Deployment | Software Engineer | 2 | 2022-12-13 | 2022-12-13 |
| 12 | 11 | Website Module Testing | Software Engineer | 1 | 2022-12-13 | 2022-12-13 |
| 13 | 8, 12 | Module Integration | Software Engineer | 1 | 2022-12-13 | 2022-12-13 |
| 14 | 13 | System Testing | Quality Assurance | 20 | 2022-12-14 | 2022-12-16 |
| 15 | 14 | Final project delivery | Software Engineer | 8 | 2022-12-19 | 2022-12-19 |

*\*Model creation will only require 2 hours of developer time, the remaining 10 hours consist of training time, which will be performed autonomously outside of business hours.*

# Part C: Application

## Application Files

The application files will be organized within the three development modules, (a) Model Training; (b) API; (c) Webpage; as described in the table below.

**Table 4**

*List of Application Artifacts*

|  |
| --- |
| Directory/Filename…………….…….Description |
| \clothsifier  \API………………………………*Files related to the back-end API*  \layers\\*………………..…………*AWS Linux compiled libraries for main.py*  aws-lambda-fuction.zip………….*Packaged API code and model for AWS Lambda*  main.py…………………………..*API Source Code*  model.tflite………………………*The Frozen* *CNN Model in .tflite format*  requirements.txt…………………*Required libraries for API code*  \model training…………………. ..*Files related to the training of the model*  \data………………………….…..*Data for training and testing*  boot.jpg………………………… *An image of a boot*  Fashion MNIST.zip…………….*The Fashion MNIST dataset*  pullover.jpg……………………..*An image of a sweater*  \model\\*…………………………*The CNN model in .pb format*  notebook.ipynb…………………. *The notebook for EDA and Model training*  requirements.txt…………………*The required libraries for the notebook*  \webpage…………………………*Files to run the front-end webpage*  index.html………………..*The homepage of the website*  scripts.js……………….....*The JavaScript*  shirt.svg…………………..*A vector image of a shirt*  styles.css……………….…*The style sheet* |

# Part D: Post-implementation Report

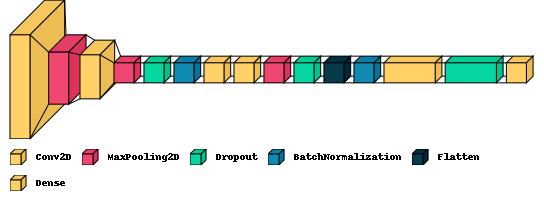
## Business Vision

The Clothier e-commerce site had a large and diverse catalogue of clothing items for sale, but no system in place for categorizing them. This made it difficult for customers to find specific items and for the company to effectively manage and organize their inventory. The Clothsifier system was developed to address this problem.

Using the Python programming language and open-source libraries, a convolutional neural network was trained in a Jupyter notebooked that could accurately assign categories to clothing items.

**Figure 1**

*A representation of the CNN model*



*Image retrieved from the training Jupyter notebook at https://github.com/AndrewStaus/Clothesifier/blob/main/model%20training/notebook.ipynb*

The trained model was integrated into a REST API service that was created using the Python programing language, and open-source libraries then deployed on a serverless AWS Lambda platform with an x86 instruction set, running Amazon Linux as a for integration into other systems.

**Figure 2**

*The welcome message on the root endpoint of the API*

Text

Description automatically generated

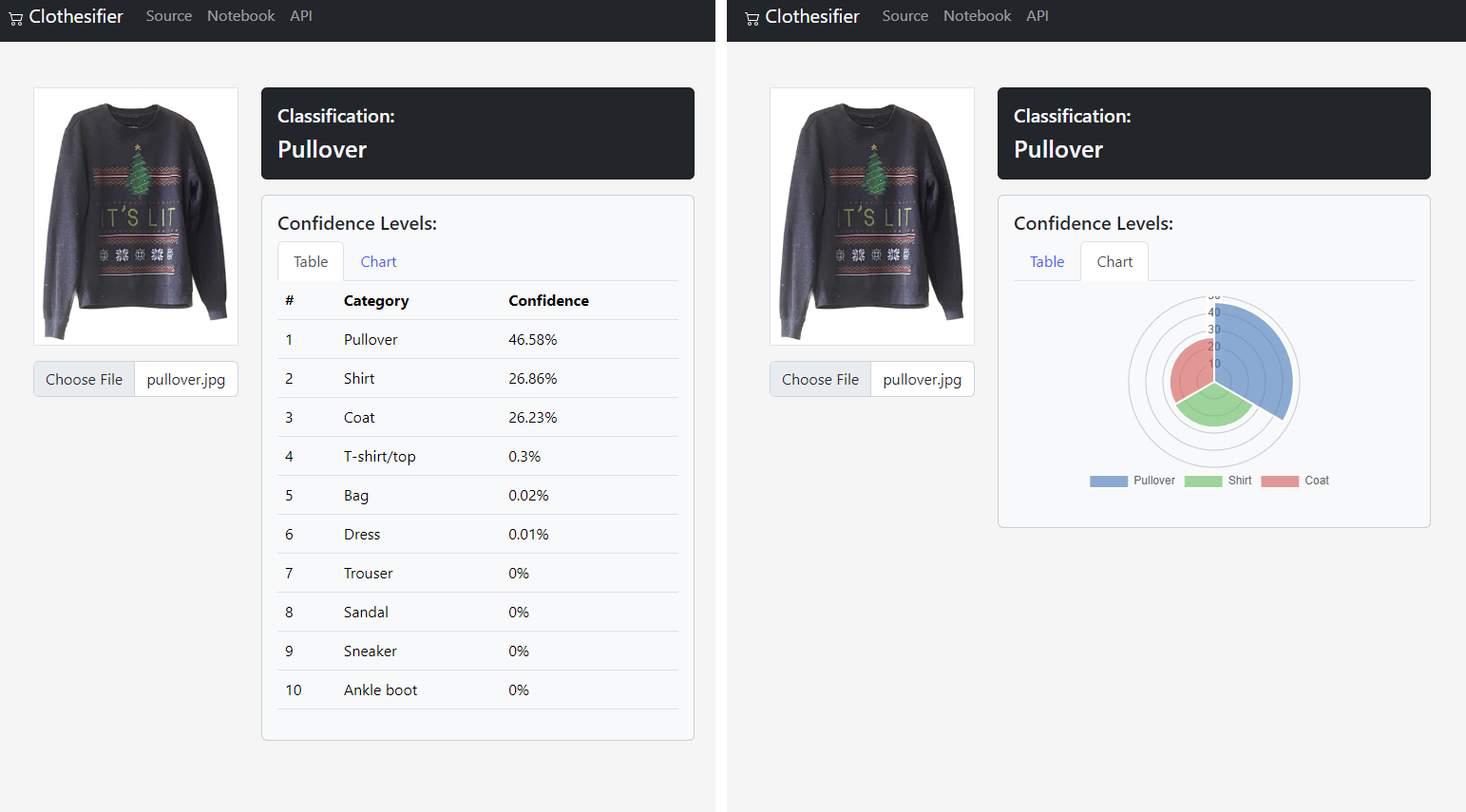
*Image captured from the API root endpoint at https://xxlkbgor75nvr7qw256z2xnrdm0ppqai.lambda-url.us-east-2.on.aws/*

While the API was the main deliverable for the customer, a front-end website was developed using HTML, CSS, and JavaScript languages, along with open-source libraries. The website allows users to easily test the system's functionality by uploading an image of clothing and receiving the predicted category in return.

To use the Clothsifier system, users can visit the front-end website and upload an image of a piece of clothing. The system will then process the image and return the predicted category for the item.

**Figure 3**

*Clothsifier Classifing an Image of a Pullover*



*Image captured from the Clothesifier webpage at https://andrewstaus.github.io/Clothesifier/*

## Datasets

The convolutional neural network (CNN) used in the application was trained using the Fashion MNIST dataset, which was sourced from Kaggle.com. This dataset is provided in the project's Github repository in the \model training\data\Fashion MNIST.zip file and are also available on Kaggle. The dataset consists of two CSV files: fashion-mnist\_train.csv and fashion-mnist\_test.csv, containing image data and labels for 6000 training examples and 1000 validation examples, respectively. These files are considered the raw datasets for the CNN.

**Table 5**

*Example of Raw Data from Fashion-mnist\_train.csv*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| label | pixel1 | pixel2 | pixel3 | … | pixel784 |
| 2 | 0 | 0 | 0 | … | 0 |
| 9 | 0 | 0 | 0 | … | 0 |
| 6 | 0 | 0 | 0 | … | 0 |
| … | … | … | … | … | … |
| 7 | 0 | 0 | 0 | … | 0 |

Note: the images have a flat background with values of zero.

**Figure 4**

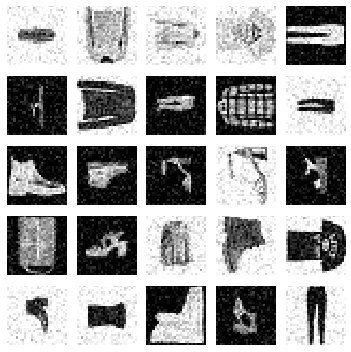
*Images Representations Extracted from the Raw Data*



To further enrich the dataset, data augmentation was performed using multiple Python libraries in a Jupyter notebook. The augmented dataset was then used to train the CNN model. However, the final application does not utilize this dataset, as the model is deployed in a trained state on the API.

**Figure 5**

*Images Representations Extracted from the Augmented Data*



## Data Product Code

TODO

## Objective (or Hypothesis) Verification

TODO

## Effective Visualization and Reporting

TODO

## Accuracy Analysis

TODO

## Application Testing

TODO

## User Guide

TODO

## Summation of Learning Experience

TODO