



Faculty of Engineering

**Personalized Prediction Clothing Algorithm Based on Multi-modal
Feature Fusion**

Liu Rong

**Master of Engineering
2024**

Personalized Prediction Clothing Algorithm Based on Multi-modal Feature Fusion

Liu Rong

A thesis submitted

In fulfillment of the requirements for the degree of Master of Engineering

(Electrical and Electronic)

Faculty of Engineering
UNIVERSITI MALAYSIA SARAWAK
2024

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

.....
Signature *Liu Rong*

Name: Liu Rong

Matric No.: 22020275

Faculty of Engineering

Universiti Malaysia Sarawak

Date : *7/11/2024*

ACKNOWLEDGEMENT

My sincere thanks go out to the following individuals during the composition of this graduation thesis:

First and foremost, I wish to extend my heartfelt appreciation to my supervisor, Dr. Annie Anak Joseph. Through the entire research period, I greatly appreciated her patience and invaluable advice. I had the privilege of being mentored and supervised by her throughout my academic career. In addition to propelled my academic growth, her guidance inspired me to contribute to personalized clothing prediction and give back to society. Moreover, I would like to express my sincere gratitude to my co-supervisors, Assoc. Prof. Ir Dr David Boon Liang and Ir Dr Ngu Sze Song. Under tight time constraints, they stressed the importance of being a diligent researcher, maintaining high standards, and ensuring impeccable results. It has been an honor for me to be guided academically by these two highly esteemed mentors.

I would like to express my gratitude to my parents, especially, for their unwavering support and understanding. This allowed me to devote myself to my research. The constant encouragement they provided and their belief in me gave me the confidence and strength to overcome obstacles. It is their love and support that I will always be grateful for.

Finally, thanks to all Faculty of Engineering and Centre for Graduate Studies staff for their assistance and guidance. It has been a privilege to receive their meticulous guidance and support throughout my academic career. I sincerely thank them for all their contributions to my growth and academic development.

ABSTRACT

Information about clothing products has shown a steady growth trend in recent years, thanks to advances in information technology and material standards. Fashion consumers struggle to choose clothing that meets their needs from massive data due to the surge in clothing products. Previous research predicted single clothing attributes, and the model's generalization ability was weak. This results in low personalized prediction accuracy, making it difficult to meet fashion consumers' personalized needs. Therefore, this thesis proposes a multi-modal fusion algorithm for predicting personalized clothing. The algorithm is designed to help consumers make more informed purchasing decisions by analysing their personal preferences and predicting fashion clothing categories. An analysis of Nunalie's sales from October 2016 to December 2019 is presented using the publicly available real sales dataset, Visuelle. There are 5355 clothing products in this data set, along with 45MB of sales data. In addition to unstructured image data, structured data consists of 21 columns, including 1-12 weeks of clothing sales data, as well as details of season, color, fabric, day, week, month, year, label, category. This thesis proposes four deep convolutional neural network (CNN) models—TCN-ResNet, TCN-2DCNN, 1DCNN-ResNet, and 1DCNN-2DCNN—that integrate the multimodal features of clothing images and sales text data. A comparison of model prediction accuracy reveals that the 1DCNN-2DCNN and 1DCNN-ResNet models demonstrate superior performance in clothing prediction. To assess the generalization ability of the two models, cross-validation was performed. The experimental results indicate that the 1DCNN-2DCNN model exhibits superior generalization, achieving a recall rate of 97.20%, an F1 score of 98.60%, a macro average of 98.62%, a weighted average of 98.63%, and model accuracy of 98.59%. Personalized prediction and clothing classification have been achieved. Through the

analysis of hybrid models, the superiority of the proposed model in solving personalized preferences is demonstrated.

Keywords: Fashion consumers, image data, clothing categories, personalized, multimodal

Algoritma Pakaian Ramalan Peribadi Berdasarkan Gabungan Ciri Pelbagai Modal

ABSTRAK

Maklumat tentang produk pakaian telah menunjukkan trend pertumbuhan yang stabil dalam beberapa tahun kebelakangan ini, hasil daripada kemajuan teknologi maklumat dan standard bahan. Pengguna menghadapi kesukaran memilih pakaian yang memenuhi keperluan mereka akibat peningkatan produk pakaian. Penyelidikan terdahulu hanya meramalkan atribut pakaian tunggal, dan keupayaan generalisasi model adalah lemah. Ini menyebabkan ketepatan ramalan yang diperibadikan rendah, menjadikannya sukar untuk memenuhi keperluan peribadi pengguna. Oleh itu, kajian ini mencadangkan algoritma penggabungan multi-mod untuk meramalkan pakaian yang diperibadikan. Algoritma ini direka untuk membantu pengguna membuat keputusan pembelian yang lebih bermaklumat dengan menganalisis keutamaan peribadi mereka dan meramalkan kategori pakaian. Satu analisis penjualan Nunalie dari Oktober 2016 hingga Disember 2019 dipersembahkan menggunakan set data jualan sebenar yang boleh diakses secara umum, Visuelle. Terdapat 5355 produk pakaian dalam set data ini, bersama dengan 45MB data jualan. Selain daripada data imej tidak berstruktur, data berstruktur terdiri daripada 21 lajur, termasuk data jualan pakaian selama 1-12 minggu, serta butiran mengenai musim, warna, fabrik, hari, minggu, bulan, tahun, label, dan kategori. Tesis ini mencadangkan empat model rangkaian neural konvolusi dalam (CNN) — TCN-ResNet, TCN-2DCNN, 1DCNN-ResNet, dan 1DCNN-2DCNN — yang mengintegrasikan ciri multi-mod imej pakaian dan data teks jualan. Perbandingan ketepatan ramalan model menunjukkan bahawa model 1DCNN-2DCNN dan 1DCNN-ResNet menunjukkan prestasi yang lebih baik dalam ramalan pakaian. Untuk menilai keupayaan generalisasi kedua-dua model, kesahan silang dijalankan. Hasil eksperimen menunjukkan bahawa model 1DCNN-2DCNN mempunyai

keupayaan generalisasi yang unggul, mencapai kadar ingatan semula sebanyak 97.20%, skor F1 sebanyak 98.60%, purata makro sebanyak 98.62%, purata tertimbang sebanyak 98.63%, dan ketepatan model sebanyak 98.59%. Ramalan peribadi dan pengelasan pakaian telah berjaya dicapai. Melalui analisis model hibrid, keunggulan model yang dicadangkan dalam menyelesaikan keutamaan peribadi telah dibuktikan.

Kata kunci: *Pengguna fesyen, data imej, kategori pakaian, diperibadikan, multi-mod*

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
<i>ABSTRAK</i>	v
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiv
CHAPTER 1: INTRODUCTION	1
1.1 Study Background	1
1.2 Problem Statement	4
1.3 Research Questions and Hypothesis	5
1.4 Objective	6
1.5 Project Scope	7
1.6 Research Contribution	7
1.7 Chapter Outline	9

CHAPTER 2: LITERATURE REVIEW	11
2.1 Overview	11
2.2 Application and Evolution of Neural Networks in Classification	11
2.3 Traditional Clothing Image Classification Method	16
2.3.1 Clothing Classification Method Based on Local Features	17
2.3.2 Clothing Classification Based on Global Features	20
2.4 Clothing Classification Method Based on Deep Learning	24
2.4.1 Clothing Classification Based on Deep Network	24
2.4.2 Clothing Classification Based on Supervised Learning	28
2.5 Research Status of Multi-modal Feature Fusion	30
2.5.1 Classification of Multimodal Fusion Architectures	31
2.5.2 Research on Multi-modal Fusion Methods	32
2.6 Research on the Combination of Clothing Classification Prediction and Personalized Recommendation	34
2.6.1 Clothing Classification Prediction Algorithm	34
2.6.2 Personalized Clothing Recommendation Algorithm	35
2.7 Research on Clothing Prediction Algorithm Based on Multi-modal Fusion	35
2.7.1 Temporal Convolutional Network (TCN)	36
2.7.2 Residual Network (ResNet)	37
2.7.3 One-dimensional Convolution(1DCNN)	39

2.7.4	Two-dimensional Convolution(2DCNN)	42
2.8	Research Gap	43
2.9	Chapter Summary	45
	CHAPTER 3: METHODOLOGY	47
3.1	Overview	47
3.2	Dataset	47
3.2.1	Image Data	49
3.2.2	Text Data	50
3.3	Study Materials	53
3.3.1	Hardware	53
3.3.2	Software	54
3.4	Proposed Methodological Framework	56
3.5	Model Analysis and Construction	57
3.6	Personalized Prediction Clothing Algorithm Based on Multi-modal Feature Fusion	59
3.7	Optimization of Evaluation Indexes for Multi-Classification	64
3.7.1	Multi-Classification	64
3.7.2	Confusion Matrix and Evaluation	66
3.7.3	Speed Indicator	68
3.7.4	Application of Cross-validation	70

3.8	Proposal Model	71
3.8.1	Proposal TCN-ResNet Model	71
3.8.2	Proposal 1DCNN-ResNet Model	74
3.8.3	Proposal TCN-2DCNN Model	77
3.8.4	Proposal 1DCNN-2DCNN Model	80
3.9	Model Optimization	82
3.10	Chapter Summary	83
CHAPTER 4: RESULTS AND DISCUSSION		85
4.1	Overview	85
4.2	Comparison between TCN-ResNet model and 1DCNN-ResNet model	85
4.3	Comparison between TCN-ResNet model and TCN-2DCNN model	88
4.4	Comparison between TCN-ResNet model and 1DCNN-2DCNN model	90
4.5	Cross Validation	93
4.6	Chapter Summary	97
CHAPTER 5: CONCLUSION AND RECOMMENDATIONS		98
5.1	Conclusion	98
5.2	Recommendations	99
REFERENCES		101
APPENDICES		112

LIST OF TABLES

	Page
Table 2.1 The development history of CNN (Bhatt et al., 2021)	14
Table 2.2 Local feature clothing classification method	19
Table 2.3 Global feature clothing classification method	22
Table 2.4 Deep network clothing classification method	25
Table 2.5 Supervised learning clothing classification method	29
Table 2.6 Research on multi-modal fusion methods	33
Table 2.7 Accuracy Evaluation of Different Models for Tuning on the Same Validation Set(Mathur, 2020)	38
Table 2.8 Research Gap Analysis	44
Table 3.1 Confusion matrix	67
Table 4.1 Comparison between TCN-ResNet model and 1DCNN-ResNet model	87
Table 4.2 Comparison between TCN-ResNet model and TCN-2DCNN model	89
Table 4.3 Comparison between TCN-ResNet model and TCN-2DCNN model	92
Table 4.4 Changes after adding repeated cross-validation	96

LIST OF FIGURES

	Page
Figure 1.1 Seasonal Color Analysis for Fashion Styling(Su et al., 2024)	2
Figure 1.2 E-commerce transactions in China from 2011 to 2022(National Bureau of Statistics,2022)	3
Figure 2.1 The development process of a neural network	13
Figure 2.2 CNN structure diagram	15
Figure 2.3 Image classification method	16
Figure 2.4 Mapping process of multi-modal features	31
Figure 2.5 Basic process of clothing classification	34
Figure 2.6 TCN model diagram	37
Figure 2.7 ResNet residual block	39
Figure 2.8 Computation of one-dimensional convolution	40
Figure 2.9 Convolution operation	41
Figure 3.1 Sample images representing various product categories within the VISUELLE dataset (Skenderi et al., 2022)	50
Figure 3.2 Clothing sales data	52
Figure 3.3 Color data distribution	52
Figure 3.4 Anaconda'running interface	55
Figure 3.5 Python 3.9 running interface	55
Figure 3.6 Proposed a methodological framework	56
Figure 3.7 Model construction steps	59
Figure 3.8 A multimodal feature fusion framework for clothing sales forecasting	59
Figure 3.9 Detailed flow of multimodal feature extraction and fusion	61
Figure 3.10 Fusion of visual features and text features	63
Figure 3.11 Multi-classification in machine learning	66

Figure 3.11	Multi-classification in machine learning	66
Figure 3.12	TCN-ResNet model	73
Figure 3.13	1DCNN-ResNet model	76
Figure 3.14	TCN-2DCNN Model	79
Figure 3.15	1DCNN-2DCNN model	81
Figure 4.1	Model 1DCNN-2DCNN Training and Validation Loss	91
Figure 4.2	Model 1DCNN-2DCNN Training and Validation Accuracy	91
Figure 4.3	Model 1DCNN-ResNet Training and Validation Loss	91
Figure 4.4	Model 1DCNN- ResNet Training and Validation Accuracy	91
Figure 4.5	Cross-Validation of 1DCNN-2DCNN Model at 30 and 300 Epochs	94
Figure 4.6	Model Improvement Effect of 1DCNN-2DCNN Architecture	94

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
TCN	Temporal Convolutional Network
ResNet	Residual Network
MCP	McCulloch-Pitts
MLP	Multilayer Perceptron
SDAE	Stacked Denoising AutoEncoder
AEs	Autoencoders
DAE	Denoising AutoEncoder
SAE	Sparse AutoEncoder
RBM	Restricted Boltzmann Machine
DBN	Deep Belief Network
DBM	Deep Boltzmann Machine
ReLU	Rectified Linear Unit
ZFNet	Zeiler and Fergus Net
SENet	Squeeze-and-Excitation Network
EnCaps	Enhanced Capsules
GAN	Generative adversarial networks
MTL	Multi-task learning
SCN	Stochastic Configuration Networks
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
HRNet	High-Resolution Network
CSV	Comma-separated values

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
TPR	True positive rate
FPR	False positive rate
FLOPs	Floating point operations
FPS	Frames per second
AI	Artificial Intelligence

CHAPTER 1

INTRODUCTION

1.1 Study Background

As Dunlap (1928) points out, clothing is the most important part of living. As time progresses, consumers' demands for quality of life and personalization rise. For example, Figure 1.1 integrates the four seasons color theory with users' biological characteristics to create personalized clothing recommendations (Su et al., 2024). This approach highlights how data-driven methods can enhance the personalization of the user's dressing experience. For Chinese retailers and brand companies to meet consumers' diverse needs, they must stay abreast of consumer trends in their domestic market and rapidly understand the opportunities and challenges they face (Shocker et al., 1994). At present, a new technological revolution represented by artificial intelligence is rapidly emerging around the world (Wang & Rasheed, 2014). E-commerce uses advanced digital technology to continuously innovate management, operation, and marketing models to create new business growth points and further enhance the user experience (French & Storey, 2021).

Although this thesis primarily focuses on the Chinese market, the models and methodologies employed possess broad applicability and can be extended to markets in other countries. Consequently, while the current analysis emphasizes China, the insights gained are equally relevant and valuable for understanding markets in other regions.



Figure 1.1: Seasonal Color Analysis for Fashion Styling (Su et al., 2024)

In the past 10 years, the global apparel market has remained at the trillion-dollar level. According to the National Bureau of Statistics of China, China's e-commerce transaction volume reached 4.383 billion yuan in 2022, an increase of 3.5% over 2021. Enterprises above designated size in the clothing industry have produced more than 23 billion pieces of clothing. As shown in Figure 1.2, China's e-commerce sales from 2011 to 2022 reflect the development of China's e-commerce market and the increase in consumer online shopping (National Bureau of Statistics, 2022). Compared to the development before 2021, the entire e-commerce industry is gradually transforming into a fiercely competitive stock market. In the current environment of rising labour costs, the industry is paying close attention to any products that can save costs and improve efficiency. AI (Artificial Intelligence) technology, with its potential to emulate and even surpass human skills in certain areas, is highly anticipated and valued. To understand the impact of China's major commodity exports on the global trade market in 2023, attention should be paid to their quantity, value, and growth rate. (National Bureau of Statistics, 2023).

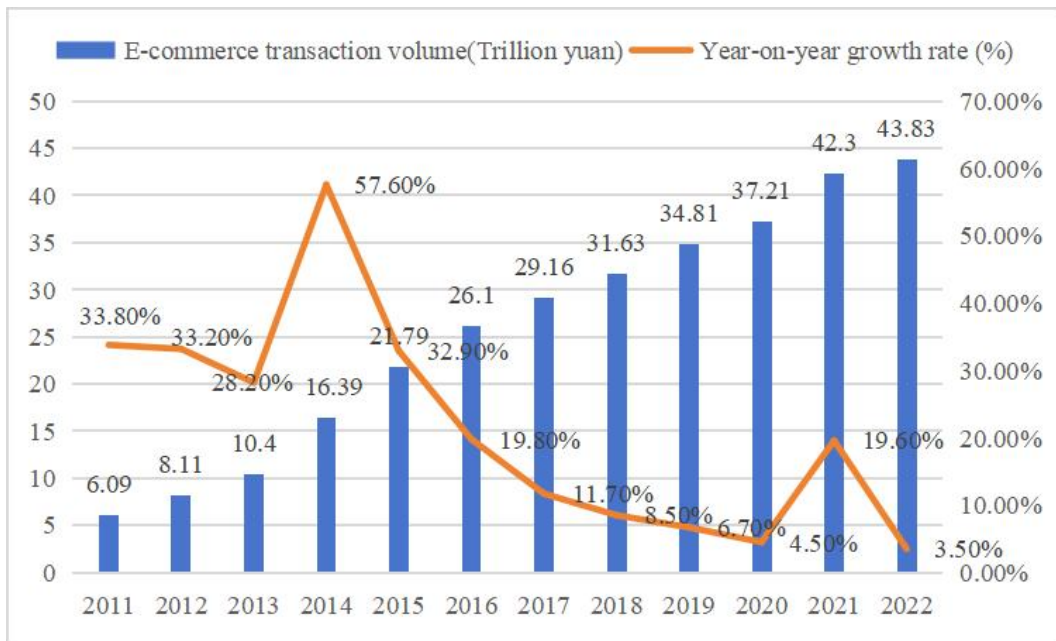


Figure 1.2: E-commerce transactions in China from 2011 to 2022 (National Bureau of Statistics, 2022)

According to Chinese residents' per capita consumption, clothing and daily necessities remain relatively equal (National Bureau of Statistics, 2023). Clothing accounts for 5% of the per capita consumption expenditure and composition of Chinese residents in 2023. This is based on a per capita consumption of 1,497 yuan. Daily needs account for 6% of consumption, and per capita consumption is 1,526 yuan. Consumers' clothing expenditure cannot be ignored, even though daily necessities are indispensable in daily life. As a result, Chinese consumers place equal value on clothing in their consumption structure as they do on daily necessities. The two maintain a balance between consumption and expenditure.

With advanced data analysis and other technologies, customized clothing forecasts can be made. By analysing consumer purchase history, preferences, and other information, clothing styles, colors, and styles can be predicted. Retailers and brands can increase sales efficiency and profitability by using personalized forecasts to better predict trends and

optimize production and inventory management. This results in higher customer satisfaction.

With technological advancements, traditional clothing classification algorithms depend on manually crafted feature extraction and classifier implementations. However, with the development of deep learning, deep learning-based methods have gradually become more effective in processing complex information in the mainstream of images.

1.2 Problem Statement

In personalized clothing classification prediction, accurately predicting fashion preferences involves leveraging diverse data sources such as text and images. Current research primarily focuses on two key challenges.

First, multimodal data need to be effectively represented and integrated in a way that exploits the complementarity of various modes. This requires developing methods to effectively fuse information from different modes to enable models to better understand clothing features. For example, while textual descriptions offer insights into clothing seasons and types, images provide detailed visual characteristics and labels. Integrating these sources provides a more comprehensive clothing description, improving personalized classification accuracy.

Second, research must address data conversion challenges between different modes. In personalized clothing classification prediction, it is crucial to convert text descriptions into image features or image features into text descriptions. This research aims to address these challenges by developing practical algorithms capable of learning mapping relationships between patterns and integrating information seamlessly. This transformation

process poses a significant challenge considering the heterogeneity of the data and the subjective relationships between patterns. However, overcoming this obstacle allows for more efficient utilization of multimodal data, thus enhancing personalized classification comprehensiveness. Based on these challenges, an optimization of parameter classification is proposed.

1.3 Research Questions and Hypothesis

In response to the two challenges mentioned in this thesis, this thesis raises the following questions:

- i. How to solve the problem of diverse consumer clothing needs?
- ii. How can the clothing recommendation experience be enhanced within customized clothing classification?
- iii. How is the performance of the proposed method?

Currently, personalized clothing classification forecasting in the apparel industry primarily relies on a single data source. This refers to prediction algorithms that depend solely on one type of data, such as a user's purchase history, a specific body measurement, or basic color preferences. The drawback of this approach is that it fails to fully capture the complexity of the user's profile, leading to less accurate and less personalized predictions.

To address this limitation, the research incorporates information from single data sources and enhances it by learning the mapping relationships between different patterns, thus overcoming the challenges of data transformation across various modes. This

approach improves the accuracy of personalized classification predictions. In contrast, the multi-modal feature fusion algorithm proposed in this thesis goes further by effectively extracting, jointly representing, and integrating features. This enables more precise customized clothing category predictions, significantly enhancing the accuracy and practical application of personalized clothing forecasts.

1.4 Objective

This study aims to develop a method for improving personalized clothing classification by combining image data with clothing sales information. The method will address challenges in data conversion across different modes by learning the mapping relationships between them. Additionally, it seeks to enhance the accuracy of personalized classification predictions. The study proposes a personalized clothing prediction algorithm that incorporates multi-modal features to achieve accurate and customized clothing category predictions, thereby improving both the accuracy and practicality of personalized clothing forecasts.

- i. To design an algorithm that integrates information from a single data source and effectively integrates structured text and unstructured data images.
- ii. To enhance the accuracy and practicality of personalized clothing classification predictions, with the goal of meeting user needs and providing a better clothing recommendation experience.
- iii. To evaluate the performance of the proposed method with the state-of-the-art methods.

1.5 Project Scope

This project investigates and implements personalized clothing prediction algorithms based on multi-modal fusion, addressing key limitations in the field. A major challenge is the effective representation and integration of diverse data sources, such as user preferences, body measurements, and fashion trends, to deliver accurate and customized clothing recommendations. Existing models often struggle to fully leverage the complementary nature of multimodal data and encounter difficulties in converting between data types due to the inherent heterogeneity and complexity of mapping subjective relationships.

To address these challenges, the proposed approach involves developing a customized model that adapts the algorithm to individual user characteristics. The project encompasses stages of data collection, cleaning, reconstruction, algorithm design, and comprehensive evaluation. Comparative analyses with mainstream models will be conducted to benchmark the performance of the proposed model. The algorithm aims to enhance the fusion and conversion of multimodal data, thereby improving the accuracy and usability of personalized clothing predictions. This thesis will conduct rigorous testing and validation to ensure the application of the proposed algorithm in personalized clothing prediction scenarios.

1.6 Research Contribution

This research not only enhances users' shopping experience and satisfaction but also drives technological advancements in the clothing industry while broadening its application scope, such as in smart wear, e-commerce platforms, and personalized

recommendations. The outcomes of this project are anticipated to introduce new innovations and technological improvements to the business sector.

- i. A multi-modal weighted combination method based on visual information is proposed to achieve a more comprehensive representation of fashion items through the weighted fusion of multi-modal features, including graphic and textual information. This method contributes to the development of the fashion and retail industries, particularly in the areas of smart shopping, product recognition, and personalized recommendations, enabling consumers to enjoy a more convenient and accurate shopping experience.
- ii. A personalized clothing prediction algorithm based on multi-modal feature fusion is proposed. The algorithm consists of four main components: feature extraction, feature joint representation, feature fusion, and customized clothing category prediction. The implementation of this algorithm can significantly improve users' shopping experience on e-commerce platforms, boosting consumer satisfaction and loyalty.
- iii. A hybrid model that combines one-dimensional convolution (1DCNN) and two-dimensional convolution (2DCNN) is introduced to enhance the performance of the fashion prediction model. This model has been compared with current mainstream models, such as the Residual Network (ResNet) and Temporal Convolutional Network (TCN). The innovation of this model provides new technological methods for clothing design and sales forecasting. Through more accurate predictions, consumers can receive personalized clothing recommendations based on their preferences

using the proposed model, further enhancing their overall shopping experience.

1.7 Chapter Outline

This thesis is divided into 5 chapters, namely introduction, literature review, research methods, results and discussion, and conclusion.

Chapter 1 introduces the research background, purpose, and significance of the thesis. It introduces research questions and research methods and provides readers with an overall research framework.

Chapter 2 reviews the relevant literature in the field. It covers traditional clothing classification methods, deep learning clothing classification methods, multi-modal feature fusion methods, and fusion research on clothing prediction and personalized recommendations. Based on the summary of previous studies' main ideas, research methods and conclusions, some gaps in current research are pointed out. This provides a theoretical and empirical basis for this thesis.

Chapter 3 describes in detail the experimental data set used in the study, including the source, scale and characteristics of the data. It then describes the experimental settings and processes, including the data set division method and evaluation indicators. Finally, the loss function and the software and hardware environment used are introduced to ensure experiment reliability.

Chapter 4 conducts experiment and compares the proposed research methods. The model analysis and construction process are explained in detail, including a multi-modal feature extraction module, feature joint representation, feature fusion and personalized

clothing prediction. Then, a comparative analysis and discussion are conducted on the proposed model. The performance and effect of each module are discussed in depth.

Chapter 5 is the summary module of this thesis. In this section, the experimental results and findings are summarized, the advantages and innovations of the model are highlighted, and possible limitations and future research directions are discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter explores the latest research progress in multi-modal fusion and personalized clothing prediction. Section 2.2 discusses the current research status in clothing classification and the evolution of its methods. This is to gain a deeper understanding of the development of relevant methods in clothing classification. Section 2.3 introduces traditional clothing classification methods in detail and describes classification methods based on local and global features. Section 2.4 explores deep learning applications to clothing classification. Section 2.5 discusses recent advances in multimodal feature fusion. Section 2.6 discusses clothing classification and personalization prediction algorithms. It reviews related classification prediction algorithms and personalized clothing recommendation algorithms. Section 2.7 introduces the latest research progress in clothing prediction algorithms based on multi-modal fusion, including TCN, ResNet, and various convolution methods. Section 2.8, the research gap is highlighted, focusing on the current challenges in clothing personalization prediction. It provides a comprehensive summary of the content of this chapter to provide inspiration and direction for the research in this thesis.

2.2 Application and Evolution of Neural Networks in Classification

The research of Rosenblatt (1958) introduced a neural network composed of two layers of neurons and used the McCulloch-Pitts (MCP) model to classify multidimensional data. This innovation enables the neural network to automatically learn and update weights from training samples. It is also seen as a significant attempt to reveal intelligence

mysteries. Rumelhart et al. (1986) proposed the backpropagation algorithm, which is the basis of Multilayer Perceptron (MLP) training. In addition, neural network pioneer Hinton and others also proposed the (Backpropagation) BP algorithm, which introduced nonlinearity to the multi-layer perceptron. Therefore, 1986 is considered a critical period for the MLP model development and improvement.

Hinton and Salakhutdinov (2006) proposed autoencoders (AEs) for learning compact representations of data, including the encoding and decoding stages. Denoising AutoEncoder (DAE) extracts more robust abstract features by introducing a degradation process. Sparse AutoEncoder (SAE) improves feature quality and generalization ability through sparsity constraints. Stacked Denoising AutoEncoder (SDAE) further enhances deep learning capabilities. In the development process, from AE to DAE, SAE to SDAE, each improvement is aimed at improving model performance and the scope of application. This includes simple data compression algorithms to deep learning models for processing complex data.

Rumelhart et al. (1986) proposed the Restricted Boltzmann Machine (RBM), a randomly generated neural network for learning probability distributions, which can be used to pretrain other models. A Deep Belief Network (DBN) is stacked with multiple RBMs, and the output of each RBM becomes the input of the next RBM to capture complex features in the data. Deep Boltzmann Machine (DBM) is also an undirected graph model that can quickly initialize model parameters through bottom-up propagation and handle data uncertainty issues combined with top-down feedback. DBN is suitable for tasks that gradually extract deep features from data, such as image recognition. DBM is more suitable for tasks dealing with high-dimensional and complex data structures and

better handles data uncertainty. The development process of neural networks is shown in Figure 2.1.

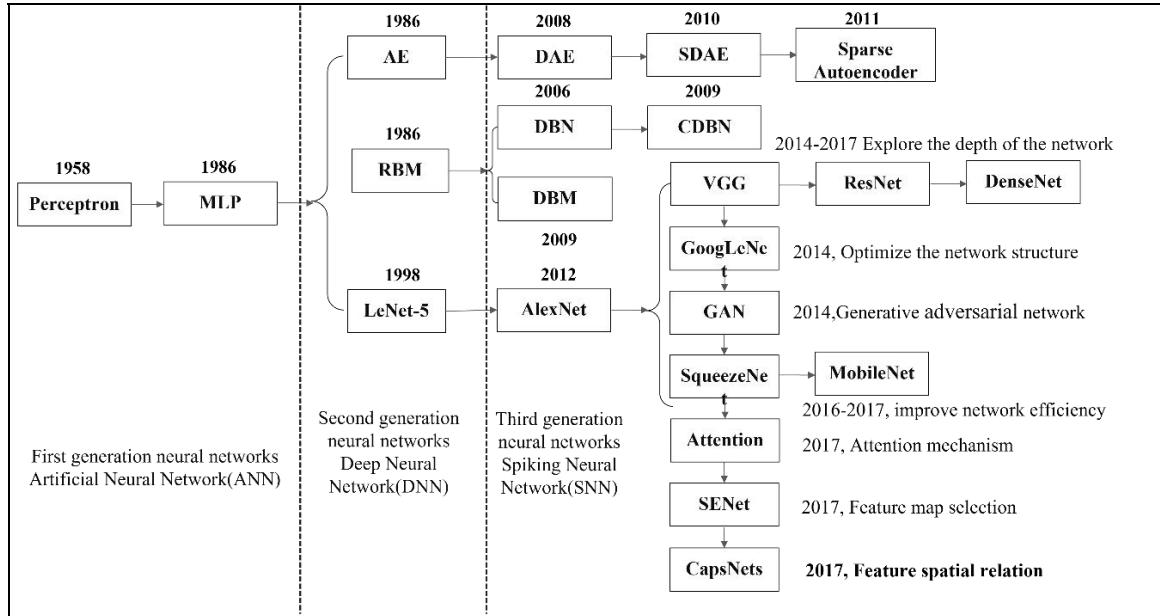


Figure 2.1: The development process of a neural network

Das (2019) summarizes the relevant content, and the development history of Convolutional Neural Network (CNN) is shown in Table 2.1. Lecun et al. (1998) introduced LeNet specifically for handwritten digit recognition. Krizhevsky et al. (2012) found that AlexNet introduced the Rectified Linear Unit (ReLU) activation function and Dropout technology to promote deep learning. Simonyan and Zisserman (2015) proposed the VGG network to improve performance by increasing the number of convolutional layers. Szegedy et al. (2015) propose that GoogleNet introduced the Inception module to improve efficiency by using filters of different sizes in parallel. He et al. (2016) proposed that ResNet solves deep network training problems through residual learning, so that deeper networks can be trained better. The above network has achieved remarkable results in image recognition tasks and has had a significant impact on deep learning research and applications.

Table 2.1: The development history of CNN (Bhatt et al., 2021)

Year	CNN	Developed by	Place	Top-5 error rate	No. of parameters
1998	LeNet (8)	Yann LeCun	-	-	60 thousand
2012	Alex Net (7)	Alex Krizhevsky, Geoffrey Hinton, Ilya Sutskever	1st	15.30%	60 million
2013	Zeiler and Fergus Net (ZFNet)	Matthew Zeiler and Rob Fergus	1st	14.80%	-
2014	Google Net (19)	Google	1st	6.67%	4 million
2014	VGG Net (16)	Simonyan, Zisserman	2nd	7.30%	138 million
2015	ResNet (152)	Kaiming He	1st	3.60%	-
2016	SqueezeNet	DeepScale	-	19.58%	1.2 to 1.4 million

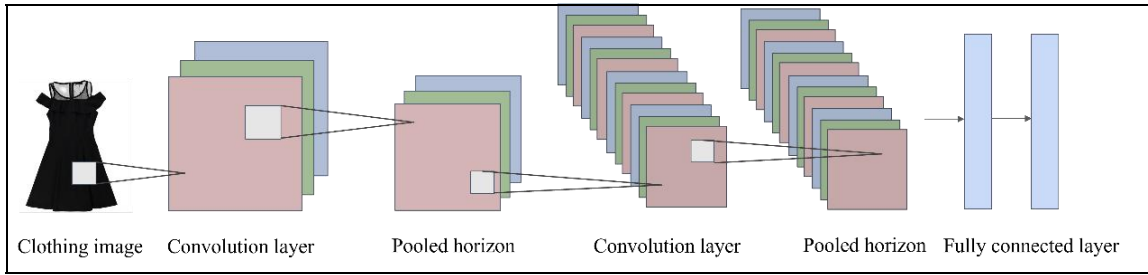


Figure 2.2: CNN structure diagram

CNN structure diagram, as shown in Figure 2.2. Capsule Networks proposed by Sabour et al. (2017) enhance the processing ability of feature location information by improving CNN structure. The (Squeeze-and-Excitation Network) SENet was proposed by Hu et al. (2018) optimizes CNN features representation and utilization by introducing a channel attention mechanism. Vaswani et al. (2017) presented Self-Attention as an innovative module. Further enhances the CNN model's performance. The above technology solves traditional CNN problems by introducing novel structures and mechanisms. This improves model performance and application scope.

Image-based classification methods can be divided into traditional image methods and deep learning methods, as shown in Figure 2.3. Traditional image methods include two categories: image features and classification models, while deep learning methods are divided into two categories: learning styles and deep networks. Traditional methods use color, visual elements, textures, etc. For classification, deep learning methods use deep learning models such as convolutional neural networks.

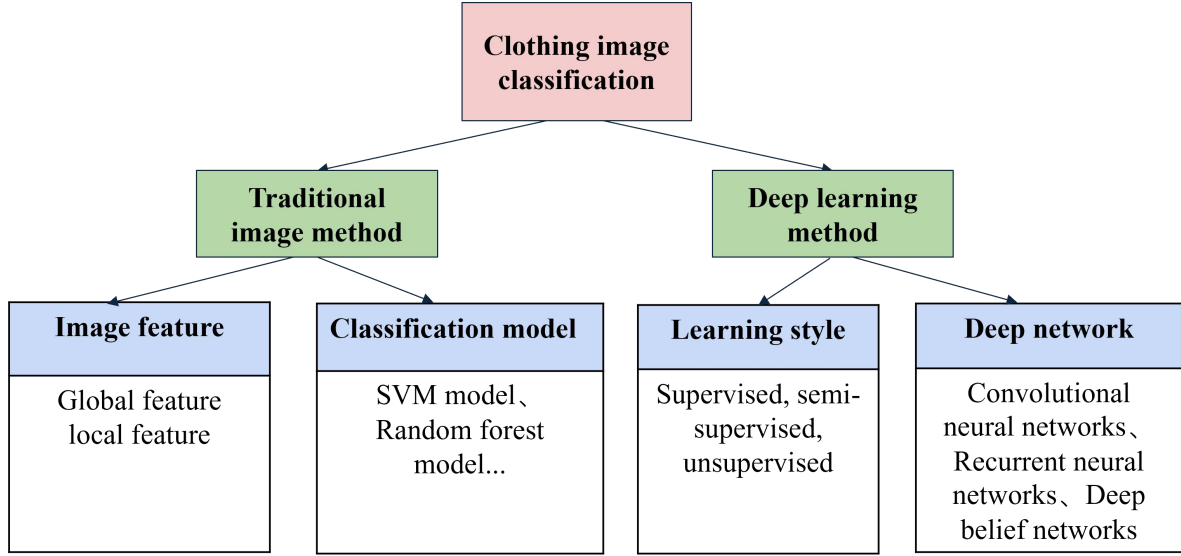


Figure 2.3: Image classification method

2.3 Traditional Clothing Image Classification Method

There are many types of clothing, and each type shows unique style and characteristics, adding colourful colors to the entire clothing world. Clothing image classification methods can be classified from multiple perspectives, including season, style, and special functions (Sang & Shen, 2019; Xu & Xin, 2020). According to gender, clothes can be divided into men's clothing and women's clothing. Due to the varying seasons, clothing is classified into spring, summer, autumn, and winter clothing to adapt to various climate conditions. Further, they are categorized into different styles, such as deep clothes, skirts, and robes. Each style has its own design and style. In addition, they are classified according to special functions and divided into flight suits, diving suits, fire suits, police uniforms, etc. These clothes play a crucial role in specific scenes or occupations. These classification methods reflect people's different clothing needs in a variety of cultures and eras. They also show the diversity of aesthetic pursuits. Traditional image classification methods usually require manually designed feature extraction algorithms, which may not perform well when processing complex images. To understand and describe clothing

diversity and characteristics, the traditional clothing image classification method is divided into a local feature classification method and a global feature classification method to introduce a variety of classification dimensions more comprehensively.

2.3.1 Clothing Classification Method Based on Local Features

Local feature classification methods classify specific elements or components of clothing. For example, different styles of clothing, such as shirts, jackets, dresses, etc., can be distinguished based on local features such as collars, cuffs, and skirts.

Wang and Zhao (2023) proposed an improved traditional clothing image classification algorithm using gradient discrimination and dictionary learning. This method sparsely reorganizes the data and uses a linear classifier in the last layer to sparsely reorganize the data. Experiments have shown that on the Ming and Qing Dynasty clothing pattern data set, the classification accuracy of this algorithm reached 82.17%, achieving the most accuracy among similar algorithms.

Chen et al. (2019) propose to adopt an innovative pre-training strategy, similar to loss-based learning, specifically for classifying clothing image attributes. To solve the problem of not considering the semantic level and granularity of clothing attributes when processing clothing classification. Zhang et al. (2019) propose that traditional clothing multi-classification methods usually rely on manual extraction of features such as color, texture, and edges of images. In contrast, deep residual networks achieve higher recognition accuracy by increasing neural network depth and have been widely used in various fields.

Wang et al. (2023) adopt a multi-stage feature fusion method to extract features from different parts and compare these feature values at the pixel level to select features that retain the maximum value. This method can reduce noise and better handle local feature extraction tasks in complex scenes. Surakarin and Chongstitvatana (2015) introduced traditional clothing classification methods, which rely on manually extracting obvious features from images, such as color, texture, and edges. However, these artificial feature extraction methods are cumbersome and have a low feature recognition rate. The research proposes an improved method based on random erasure and residual networks to solve this problem. Through this method, the improved residual model successfully enhanced clothing data classification accuracy by 2.43%.

By locating the key points in clothing images, local features can be effectively extracted, thereby improving the classification effect of the model. As shown in Table 2.2, by extracting local features from the image, complex image data is converted into a representation with a lower dimension. This facilitates the classification of clothing.

Table 2.2: Local feature clothing classification method

Authors	Classification tasks	Data set	Scene complexity	Evaluation index	Advantage	Limitations	Application
(Lu et al., 2019)	Picture classification, Extract features	6 free image databases; High		Accuracy 73.57%	Improved robustness, Enhanced discriminative ability	High computational complexity and parameter adjustment issues	Handling structurally inconsistent data sets
(Nazir et al., 2021)	image classification, Texture feature extraction	Art Fabric; Low		Accuracy 97.60%	Multi-feature fusion、 PCA dimensionality reduction	Performance is limited, feature selection depends on	Textile design analysis, image retrieval and classification
(Mo et al., 2020)	Feature extraction	FERET、AR、 CMU PIE、Yale, and USPS; High		F1 score 100%	Overcoming feature extraction limitations、 Good robust performance	The computational complexity is high	Face recognition, data mining, pattern recognition
(Hoang & Rebhi, 2018)	Clothing defect classification	Fabric defect database; Medium		Accuracy 92.10%	The classification results increased by 8%.	Color space has limitations	Fabric defect detection、 Image processing field

There are challenges and limitations in many aspects of the traditional local characteristic clothing classification method, but it also has certain advantages. These methods may show higher accuracy and performance in specific scenarios and data sets, providing feasible solutions for specific application scenarios. In addition, some of the methods in Table 2.2 have high robustness and stability. They maintain high performance in different environments, have strong practicability and operability, and are easy to apply to actual production and applications. However, these methods also face many challenges such as data set limitations, high computational complexity, insufficient performance evaluation, limited feature extraction, and limited application range. They need to be further improved and optimized. Therefore, although the traditional local characteristic clothing classification method has certain advantages, it still needs further research and improvement to solve the existing problems. This is to improve its performance and effect in practical applications.

2.3.2 Clothing Classification Based on Global Features

Feature global classification focuses on analyzing the overall appearance and style of clothing, taking into account characteristics such as shape, texture, and color matching. Recent studies emphasize that enhancing the external neural network structure can significantly improve clothing classification accuracy. For instance, Ren et al. (2022) proposed an improved convolutional neural network (CNN) algorithm based on the Xception model. This algorithm addresses the challenges of global classification by decomposing the problem into high, medium, and low local layers, thereby refining the classification process. Zhang et al. (2023) further highlighted the importance of global features in clothing classification by proposing a model based on transfer learning. This

model effectively extracts global features from clothing images, demonstrating the critical role these features play in accurate classification.

In addition, Li et al. (2022) introduced a lightweight CNN-based clothing classification method. Traditionally, CNN algorithms use the entire image as input, extracting global features through multiple convolutional layers for classification. However, Li's study innovates by incorporating a key point positioning algorithm that focuses on significant local details, such as collars, to enhance classification accuracy.

Finally, Huang (2023) proposed a multi-category clothing recognition and retrieval method that integrates both local and global features. This approach has proven to be effective in capturing detailed information within clothing images, leading to more precise classification outcomes. Table 2.3 summarizes the global clothing feature classification methods discussed.

Table 2.3: Global feature clothing classification method

Authors	Classification tasks	Data set	Scene complexity	Evaluation index	Advantage	Limitations	Application
(Yang et al., 2014)	single feature, Image classification	City College of New York Clothing Pattern; High		Accuracy 92.55%	Automatic recognition of complex patterns and colors of clothing	Equipment such as cameras needs to be installed, adding complexity and cost.	Shopping and dressing guide for the visually impaired
(Yang et al., 2022)	Multi-scale global features, Image detection	Passive Millimeter Wave image; Medium		Precision 88.1%	Adaptive label assignment optimizes the number of positive samples.	Difficulty detecting small objects	Suitable for most scenes with good penetration of clothing
(Mo et al., 2020)	Multiple global features, feature extraction and classification	(Augmented Reality) AR Face、 Small-Scale、 Yale、 Large-Scale Database; High		Accuracy 98.71%	Overcoming feature extraction limitations、 Good robust performance.	Multiple feature extraction is attractive and high-dimensional	Face recognition and other fields, processing complex local information and structures

Table 2.3 continued

(Liu et al., 2016)	Single feature, Identify classification	Fabric defect database; Medium	Accuracy 76.4%	Data set detailed annotation	Magnified images have lower detection rates	Clothing retrieval task
(Li et al., 2023)	Multiple features global, Image classification	Deep Fashion; High	Accuracy 87.78%	Comprehensive feature extraction	High computational cost, complex model, and labeling requirements	E-commerce platform, clothing recommendation
(Hu & Jiang, 2023)	Multi-scale inspection, defect detection	Defect database of 3,524 knitted fabric images; High	Precision 72.32% Recall 78.87%	Deformable convolution module	Data acquisition is difficult and calculation costs are high.	Fabric defect detection, medical imaging, etc.
(Septiarini et al., 2021)	Image processing analyzes color and texture characteristics	150 sarong images; Low	Accuracy 100%	Repeated cross-validation ensures model robustness	A dataset lacking diversity was used for training.	Detect manufacturing defects, inspect fabrics with real-time images

As seen in Table 2.3, the traditional global feature clothing classification method shows advantages, existing problems, and challenges. First, some methods show high levels of accuracy and adopt techniques such as adaptive label assignment to improve the model's robustness. Some methods are capable of multi-feature extraction, attractiveness, and multi-dimensionality, improving the classification effect of the model. Secondly, some methods require additional equipment support, such as installing cameras, which increases the complexity and cost of the system. Some methods have high computational costs, complex models, and high labelling requirements. Therefore, traditional global feature extraction faces challenges in terms of equipment, data sets, and computational costs.

2.4 Clothing Classification Method Based on Deep Learning

Deep learning plays a crucial role in clothing classification, mainly using deep learning models such as CNN to learn clothing image characteristics automatically. Through continuous optimization of data sets, model architectures, and evaluation methods, clothing classification accuracy and efficiency are improved.

2.4.1 Clothing Classification Based on Deep Network

Deep learning plays a crucial role in clothing classification. This research aims to achieve accurate classification and search of clothing images by establishing and training deep learning models. For example, Helwan (2021) explored the use of deep learning technology to classify different types of clothing images, and Mohanty et al. (2022) proposed a modified multi-category classification method for different types of clothing. Table 2.4 illustrates the application of neural networks to clothing classification.

Table 2.4: Deep network clothing classification method

Authors	Tasks	Data set	Scene complexity	Evaluation index	Advantage	Limitations	Application
(Yang, 2022)	Multi-label classification algorithms and decision tree algorithms, Identification of clothing styles	Multi-label classified clothing dataset; High		Precision rate 73%, Recall rate 43% and F1 score 55%	Improved recommendation accuracy	High resource consumption and reliance on a large amount of annotated data	Scenarios for intelligent recommendation of clothing styles
(Shin et al., 2023)	CNN model and improved convolutional and pooling layer image classification technology	Fashion-MNIST; Medium		Precision 93%	Multi-label classification algorithm improves user satisfaction	Resource consumption is high. Model training is complex. Depend on data quality	E-commerce platform, personalized services, fashion applications
(Yang et al., 2021)	Local global attention mechanism, multi-scale densely connected units, image classification	Fashion-MNIST; High		Accuracy 95.18%	Multi-scale feature extraction, high recognition accuracy	Computing resource consumption and large demand	Digitize, classify and identify ethnic costumes

Table 2.4 continued

(Liu et al., 2016)	Single feature, Identify classification	Fabric defect database; Medium	Accuracy 76.4%	Data set detailed annotation	Magnified images have lower detection rates	Clothing retrieval task
(Li et al., 2021)	Extreme learning; Online Sequential Extreme Learning Machine (OSELM)	Deep Fashion; High	Accuracy 93.98%	The model has no outliers	High model complexity	Clothing image classification
(Wang & Chen, 2024)	Transfer learning; genetic algorithm optimization, XGBoost	E-commerce datasets; High	Precision 95%	High-precision, comprehensive utilization of multi-feature information	Computing resource consumption, requiring a large amount of labelled data	CAD systems for clothing design, textile and fashion industry
(Yu et al., 2021)	Convolutional Neural Network (CNN); Enhanced Capsule (EnCaps) Network	Deep Fashion; High	Accuracy 84.2%	High precision and good robustness	Requires a large amount of training data and computing resource consumption	Online clothing shopping platform

Table 2.4 continued

(Zeng et al., 2020)	Generative adversarial network; dual-path attention model	Deep Fashion; High	-	Improved quality of generated images	There are generation errors for complex images	Generate clothing image scene
(Danopoulos et al., 2021)	Image reconstruction using generative adversarial networks (GAN)	Deep Fashion; High	Average reconstruction time is 0.013 MS, 43 dB PSNR	Improve speed and efficiency	-	Scenarios that require fast, efficient, high-quality image reconstruction
(Truong & Yanushkevich, 2021)	Construct new datasets and propose soft attention units; Mask RCNN and ResNet101V2	Character clothing pairs and segmentation masks; Medium	Accuracy 98.55%±0.35%	High accuracy and good generalization ability	Applicability limitations and data set limitations	Workplace safety monitoring, retail operations, etc.

Referring to Table 2.4, deep network clothing classification methods utilize advanced algorithms and multi-feature extraction techniques to improve recommendation accuracy. Several methods demonstrate high accuracy and robustness in evaluation, enabling precise clothing item identification. However, challenges such as resource consumption, model complexity, dataset limitations, scenario constraints, and algorithmic limitations exist. Addressing these challenges requires additional research to enhance the applicability and effectiveness of deep networks in clothing classification.

2.4.2 Clothing Classification Based on Supervised Learning

Clothing classification is based on supervised learning, including supervised, unsupervised, and semi-supervised learning. The difference between the three is whether to use labelled data to process related data. Supervised learning relies on labelled data for training. Unsupervised learning focuses on discovering structures and patterns from unlabelled data. Semi-supervised learning tries to combine the advantages of both, using a small amount of labelled data and a large amount of unlabelled data. Labelled data to improve the learning efficiency and accuracy of the model.

For example, Liu et al. (2023) proposed a multi-modal matching framework based on generative adversarial networks (GAN). Using supervised learning signals, including images and text descriptions, the generator produces realistic and compatible images with the input clothing. Rasmus et al. (2015) A semi-supervised method that combines supervised and unsupervised learning. Through backpropagation, the model simultaneously minimizes the sum of the supervised and unsupervised cost functions, avoiding hierarchical prediction. Table 2.5 shows related work on supervised learning.

Table 2.5: Supervised learning clothing classification method

Authors	Learning style	Data set	Scene complexity	Evaluation index	Advantage	Limitations	Application
(Han et al., 2023)	Semi-supervised learning	Colourful fashion parsing dataset (22 clothing categories and 1 background category), medium		Pixel Accuracy 92.77%	Self-supervised and unsupervised contrastive learning patterns and colours	Inefficient for large data sets	Semantic segmentation task on small-scale and unevenly distributed clothing datasets
(Wang, 2021)	Semi-supervised learning; unsupervised learning	Female clothing image dataset (10,200 training set images and 3300 test set images), high		Average precision 76.85%, Average 72.88%	Improve performance and reduce labour costs when labelled data is lacking.	Dependence on large-scale data	Suitable for tasks that lack labelled data
(Yoshino et al., 2022)	Supervised learning	Gait recognition, medium		Accuracy 94.4%	Generate high-quality gait images	Insufficient recognition accuracy when wearing different clothes	Contactless person identification required

The supervised learning clothing classification method presents a series of characteristics and application aspects in the table. First, these methods use various data sets and learning styles, such as semi-supervised learning, semi-automatic encoding transformation, data augmentation techniques, etc. Second, these methods show high accuracy and performance in terms of evaluation metrics, such as pixel accuracy and average accuracy. Advantages include improved performance, reduced labour costs, improved classifier robustness, etc. However, these methods also have some shortcomings and limitations, such as strong dependence on large-scale data, possible inefficiency on specific tasks and data sets, and the introduction of noise.

2.5 Research Status of Multi-modal Feature Fusion

Multimodal feature fusion aims to integrate information from different sources to improve model performance and decision-making capabilities. This process involves a variety of methods, including concatenation, feature-level fusion, dynamic fusion, hypergraph-based self-supervised multi-modal representation learning, and dual attention networks. This thesis focus on the splicing method of multi-modal fusion, aiming to improve the performance of the algorithm, enhance the interpretability of the model, and enhance classification accuracy. In the process of multi-modal fusion, there are key issues such as how to effectively process and integrate data from different modes and how to deal with heterogeneity between different modes. Figure 2.4 shows the mapping process for multi-modal features.

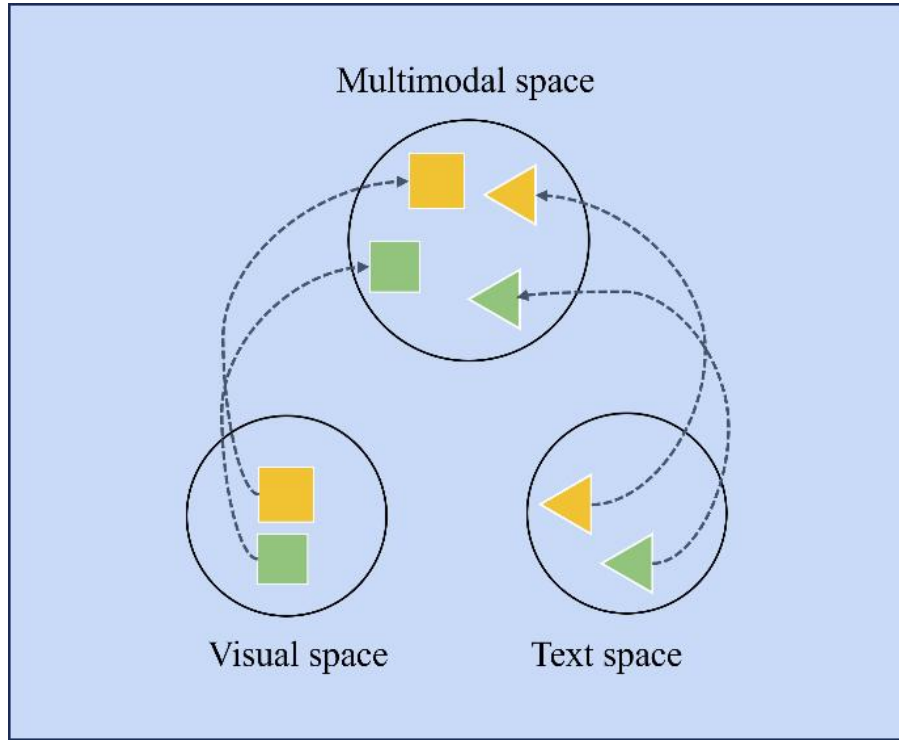


Figure 2.4: Mapping process of multi-modal features

2.5.1 Classification of Multimodal Fusion Architectures

The multi-modal fusion architecture covers serial, parallel, and deep fusion structures. In clothing classification prediction, the serial structure focuses on improving model performance by processing data or features from different modes sequentially. Correspondingly, while parallel structures process data from different modes simultaneously, transfer learning is an application of parallel structures that assists modes that lack data. The flexible sorting method used in deep multi-task learning (MTL) shares layers more effectively and meets the needs of multiple tasks (Meyerson & Miikkulainen, 2018).

In addition, Nagrani et al. (2021) use the Transformer structure to process multiple modes, such as multi-modal fusion and classification through video frames and sound

features. For the multi-model-specific Transformer architecture, multi-modal pre-training can improve classification performance Xu et al. (2023).

Overall, the classification of multimodal architectures covers all aspects of consideration, from model design to training methods. This highlights the diversity and complexity of multimodal learning.

2.5.2 Research on Multi-modal Fusion Methods

Research on multi-modal fusion methods primarily focuses on feature fusion and model fusion. Feature fusion typically involves techniques like splicing and addition to combine features from different data sources into a single representation. Model fusion, on the other hand, often utilizes methods such as Stacking and Blending to combine the outputs of multiple models, thereby enhancing overall accuracy.

Table 2.6 shows the application of various learning methods and data sets of multi-modal fusion to process multi-modal data. These methods involve different learning technologies, reconstruction algorithms, multi-scale fusion, reinforcement learning, etc., and use multiple data sets at the same time, such as Deep Fashion, Fashion MNIST, etc. These methods make full use of features from multiple categories, reduce background interference, and improve image recognition and analysis capabilities. However, these methods also have some shortcomings and limitations, such as difficulty adjusting model parameters and high computational complexity.

Table 2.6: Research on multi-modal fusion methods

Authors	Learning style	Data set	Scene complexity	Evaluation index	Advantage	Limitations	Application
(He et al., 2020)	Multi-category fusion	Deep Fashion, Fashion MNIST;	High	Accuracy 95.00%	Make full use of multi-category features、Reduce background interference	It is difficult to adjust model parameters.	Scenes with complex backgrounds
(Zhang et al., 2022)	Multi-feature fusion method	Slice-level millimetres wave radar data;	Medium	-	Improved the imaging system's ability to identify different dangerous targets	The specific performance parameters studied were not mentioned	Security detection field, non-contact detection scenarios
(Yu et al., 2023)	Reinforcement learning	Clothing-8, Deep Fashion;	High	Accuracy 94.62%	Combining frequency domain and spatial information	High computational complexity	Fashion e-commerce platform, clothing design and production, security field
(Li et al., 2023)	Multimodal feature classification	Deep Fashion;	High	Accuracy 87.78%	Improve image recognition and analysis capabilities.	For complex classification tasks	Combine global and local features for classification

2.6 Research on the Combination of Clothing Classification Prediction and Personalized Recommendation

Research on clothing classification and personalized recommendations centers on two main technologies: using (Artificial Intelligence) AI, like deep learning, to classify clothing images and predict attributes, and analyzing user behavior and preferences to generate personalized recommendations. Combining these technologies can improve user experience and the accuracy of recommendation systems.

2.6.1 Clothing Classification Prediction Algorithm

Clothing classification prediction mainly uses deep learning models to classify and attribute predictions of fashion categories and subcategories to improve the identification and recommendation of fashion items (Amin et al., 2023). Hu et al. (2023) proposed a clothing image classification prediction model based on feature extraction and sine and cosine algorithm optimization (Stochastic Configuration Networks) SCN and improved classification accuracy by introducing multi-channel parallel convolution, pooling, and batch normalization operations. The classification diagram of clothes is shown in Figure 2.5.

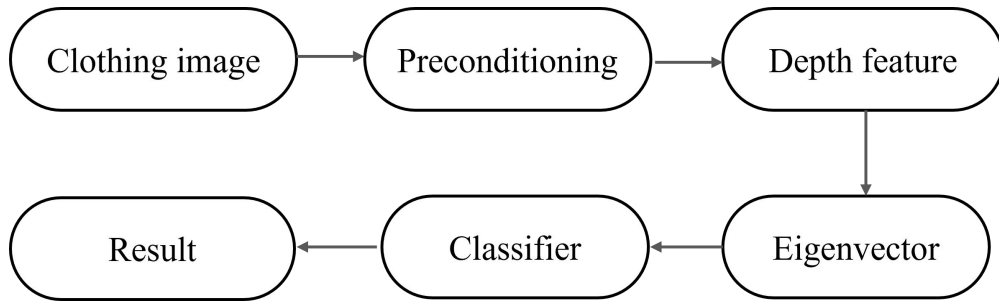


Figure 2.5: Basic process of clothing classification

2.6.2 Personalized Clothing Recommendation Algorithm

The personalized recommendation algorithm provides personalized recommendation results and experiences for the recommendation system by analysing information such as user attributes, behaviour, and context. Recommendation algorithms are the core technology behind individualized recommendations. Wang et al. (2019) proposed a method based on users' historical behaviour's and preferences, which can effectively predict clothing users may be interested in, demonstrating efficiency and accuracy in processing large-scale data sets. Sivaranjani et al. (2023) proposed a fashion clothing recommendation architecture based on machine learning and summarized and analysed related research work on key technologies such as compatibility prediction and diagnosis, and intelligent clothing generation. He and McAuley (2016) combine visual recognition technology to extract dressing styles and habits by analysing pictures uploaded by users. This is done to improve personalized clothing matching accuracy.

2.7 Research on Clothing Prediction Algorithm Based on Multi-modal Fusion

In clothing prediction algorithms, multi-modal fusion methods can comprehensively utilize information from different data forms (such as images, text, user behaviour, etc.) to improve data processing efficiency and accuracy. This section will introduce several common deep learning models, including Temporal Convolutional

Network (TCN), Resnet, one-dimensional convolution and two-dimensional convolution, and their application in multi-modal fusion clothing prediction algorithms.

2.7.1 Temporal Convolutional Network (TCN)

Bai et al. (2018) proposed a multivariate time series forecasting method that utilizes temporal convolutional neural networks (TCN) to process sequence data. This method extracts information from the time dimension through convolution operations. It uses dilated convolution to capture temporal relationships and long-term dependencies more effectively in time series data. With this approach, the model can predict more accurately future trends and patterns in multivariate time series. In the task of clothing prediction, TCN has the following advantages for clothing prediction.

First, TCN uses convolution operations to parallelize calculations. This allows the algorithm to better utilize modern hardware's parallel computing capabilities when processing large-scale data, thereby improving training and inference efficiency (Yamamoto et al. 2020). Second, compared with typical recurrent networks (such as LSTM and GRU), the TCN architecture is more accurate, but also simpler and clearer. This structural advantage helps the model better understand and process data (Lea et al., 2016). Finally, combined with the multi-modal feature fusion method, TCN can be input in a single form, such as keywords, related text, etc. This flexibility enables the algorithm to adapt to more application scenarios and needs (Han, 2023).

TCN, the model diagram is shown in Figure 2.6, where l is equal to input length, k is equal to kernel size, b is proportional to dilation base, $k \geq b$, and for complete historical coverage n is the minimum number of residual blocks, where n can be calculated from other values.

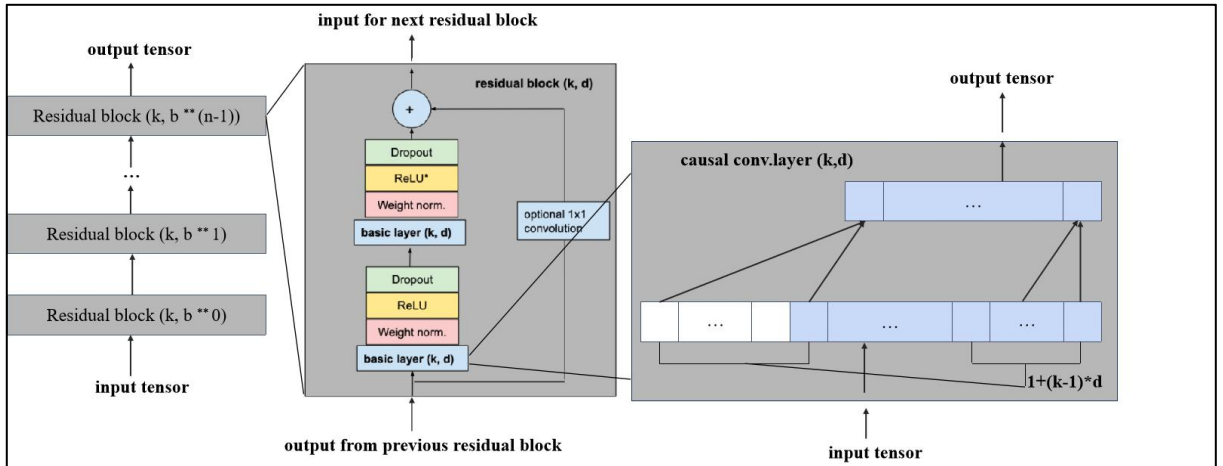


Figure 2.6: TCN model diagram (Bai et al., 2018)

2.7.2 Residual Network (ResNet)

A residual network (ResNet) was proposed by He et al. (2016) to avoid the degradation phenomenon during deep network training. Network training and optimization become more challenging as the number of layers increases. Deep networks will perform worse if gradient information disappears or explodes. For this problem, ResNet introduces the residual module, which learns the network residual mapping to make training easier. In addition, ResNet uses small convolution kernels and deeper network structures to improve expression capabilities and learning efficiency. To compare ResNet's advantages in classification performance, the thesis analyzes mainstream models and lightweight classification models, as shown in Table 2.7.

Table 2.7: Accuracy Evaluation of Different Models for Tuning on the Same Validation Set (Mathur, 2020)

Authors	Model	Network	Accuracy	Advantage	Limitations	Application
(Hiraoka et al., 2023)	HRNet	High-Resolution	80.32%	High-resolution features	Large parameter size, high computational cost	Image segmentation, pose estimation
(Chen et al., 2022)	ShuffleNet	Shuffle-based	88.31%	Low parameter count	Performance inferior to deep networks	Object detection
(Sharma et al., 2023)	EfficientNet	Compound Scaling	93.17%	High performance	Large model parameters, high computational cost	Image classification, object detection, mobile applications
(Fan et al., 2021)	AlexNet	Convolutional	47.22%	Breakthrough success		
(Altaf et al., 2023)	DenseNet	Dense Connections	84.61%	Dense connections, feature reuse	Large model parameters	
(Kulkarni et al., 2023)	MoblieNet V2	Depth wise Separable	92.82%	Low parameter count	Performance inferior to complex networks	
(Sahithi and Vigneshwari, 2023)	VGG16	Parameters138M	71.70%	High fitting ability	Many parameters, the training time is slow	Feature extraction, transfer learning
(Krueangsai and Supratid, 2022)	ResNet50	23M	95.60%	Avoid gradient disappearance	Overfitting risks and complexity increase	Object detection, image segmentation

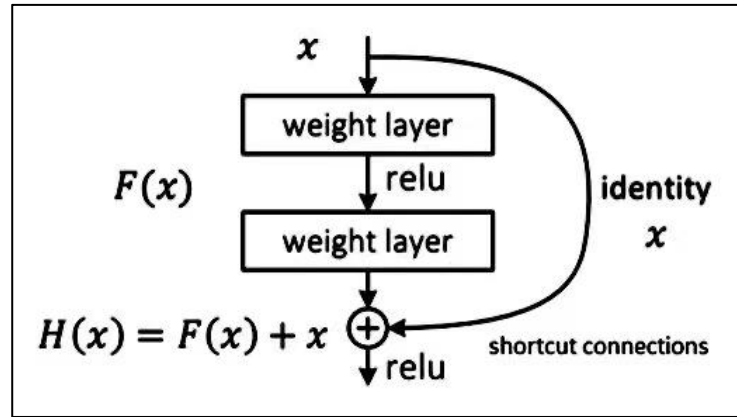


Figure 2.7: ResNet residual block

ResNet’s residual blocks are a crucial part of the network, also known as “identity mapping”. As shown in Figure 2.7. These residual blocks directly pass the input x to the output through “shortcut connections”, so that the output result can be expressed as $H(X) = F(x) + x$. In this equation, if the residual $f(x)$ is equal to 0, then $H(x)$ is equal to the input x , which is the identity mapping. Therefore, the ResNet learning goal is no longer simply to learn a complete output but the difference between the target value $H(x)$ and the input x , that is, the residual $F(x) = H(x) - x$. This design allows the network to learn residuals more easily and train deep neural networks more efficiently. This alleviates problems such as vanishing or exploding gradients.

2.7.3 One-dimensional Convolution(1DCNN)

The 1DCNN model is designed to process one-dimensional sequence data like text or time series. It captures local relationships in sequences more effectively than traditional fully connected networks, making it ideal for tasks such as speech recognition, natural language processing, and time series prediction. The model uses a convolutional kernel that slides over the input sequence, extracting features at each step. Figure 2.8 illustrates a diagram for performing one-dimensional convolution.

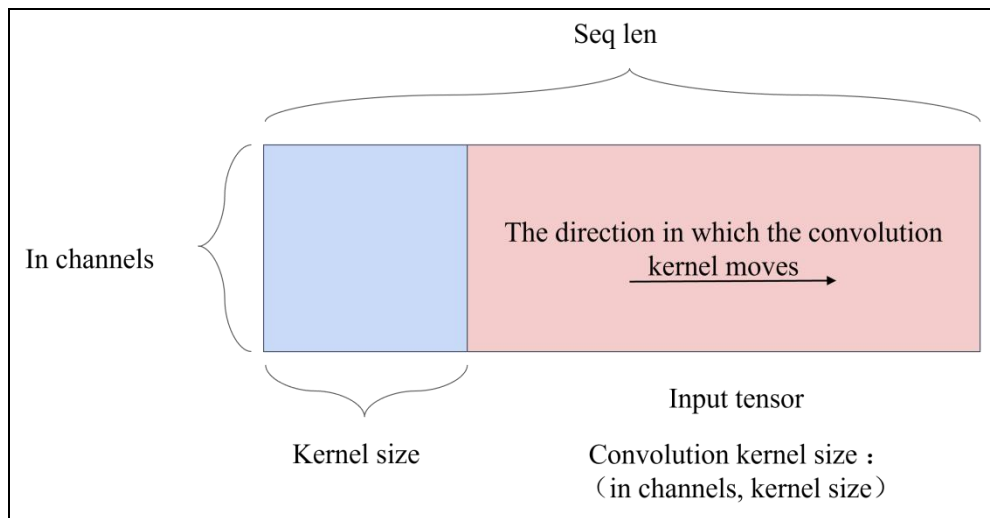


Figure 2.8: Computation of one-dimensional convolution

Johnson and Zhang (2017) proposed that one-dimensional convolution is used for text classification in text classification tasks. The core of one-dimensional convolution lies in its way of processing data: sliding a sliding window (also called a convolution kernel) on the input sequence, calculating the sum of the products of the data in the window and the convolution kernel, and generating a new output sequence. This operation is not only applicable to image data but also to text and time series data. As shown in Figure 2.9, the result is obtained after a convolution operation is performed between an area in the picture and a convolution kernel.

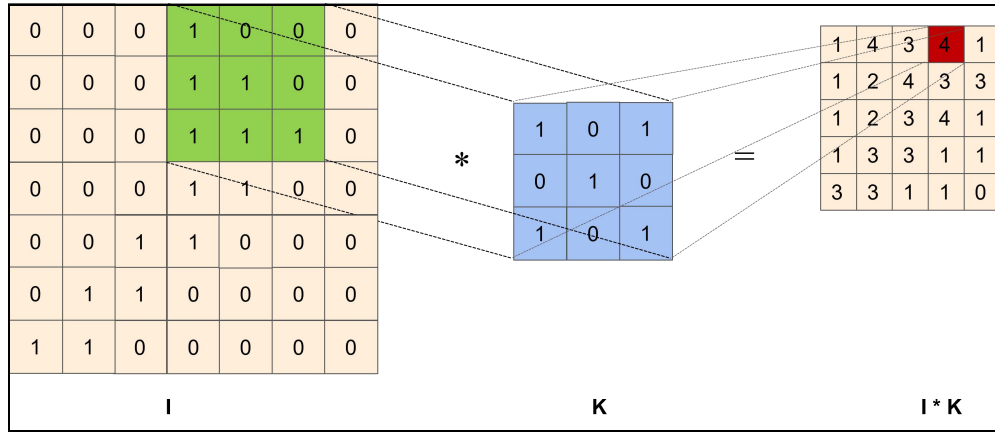


Figure 2.9: Convolution operation

To obtain the result, a value corresponding to the corresponding position of the area is multiplied by a value corresponding to the corresponding position of the convolution kernel. This is done to obtain the value related to the area position. Finally, the result is calculated after multiplying all the results and summing them up. The formula for calculating one-dimensional convolution is as follows:

$$n \cdot n \cdot n_c * f \cdot f \cdot n_c \cdot n'_c \rightarrow \left[\frac{n + 2p - f}{s} + 1 \right] \cdot \left[\frac{n + 2p - f}{s} + 1 \right] \cdot n'_c \quad \text{Equation 2.1}$$

n : Input data size;

n_c : Number of channels for input data;

f : Convolution kernel size;

n'_c : Number of convolution kernels

p : Padding size

s : Stride

$*$: Convolution operation

2.7.4 Two-dimensional Convolution(2DCNN)

In image processing, 2D convolution is a commonly used technique. It uses a smaller matrix (called a waveform kernel or filter) to perform local damage at various positions of the input data to extract important characteristic information from the data. By using different waveform definitions, multiple feature maps can be produced, each capturing and expressing different aspects of the input data. Such feature maps can be used to represent differences in input data. Furthermore, to improve computational efficiency, the two-dimensional surface algorithm also includes some fast algorithms to speed up the processing process by reducing the multiplicative invasion of noise (Cheng & Parhi, 2020). These convolution kernels slide on the input data and are multiplied element by element to finally generate an output feature map. This process is not limited to the field of image processing but has also been applied to other fields, such as multi-task learning of time series data (Yeh et al., 2023).

In image processing, two-dimensional convolution is not only used in specific application cases such as feature extraction and image filtering but is also a key technology to achieve complex image recognition and classification tasks. This can also be done by converting convolution calculations into matrix multiplication and using heterogeneous convolution filters. Convolution improves deep learning models' computational efficiency (Singh et al., 2019).

For example, the formula for two-dimensional convolution represents the input image matrix and is convolved with the matrix to obtain a new matrix, which represents the output image. Here, the indices are related to the image matrix, while the indices are related to the kernel matrix.

$$y[i, j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h[m, n] \cdot x[i - m, j - n] \quad \text{Equation 2.2}$$

2.8 Research Gap

Existing research has certain limitations in taking full advantage of modal fusion complementarity and compatibility. Zhang et al. (2020) proposed an improved framework for clothing matching based on item compatibility (Radford et al., 2021). It does not rely on visual features but exclusively uses text descriptions, especially title sentences, to build basic features. A long-short-term memory (LSTM) network generates feature embeddings from title sentences. By integrating these embeddings in a compatibility matrix characterized by a style compatibility space, a style compatibility space is created. The framework is evaluated on three large-scale clothing datasets to demonstrate its effectiveness compared with baseline methods.

Chen et al. (2022) proposed a personalized clothing compatibility prediction model that encodes visual features of users' clothing preferences and measures attention and clothing compatibility in different regions. Finally, the two modules are jointly optimized to enhance the model's predictive ability. However, the comprehensiveness of multi-modal fusion has not been mentioned, especially in terms of the model's effective integration and balance of text and visual information (Kim et al., 2021). Table 2.8 provides an analysis of the gaps between current literature and unresolved issues in the field of research.

Table 2.8: Research Gap Analysis

Authors	Methodology	Existing Solutions	Limitations	Research Gap
(Zhang et al.,2020)	Focuses on visual information, integrating text and visual features for a comprehensive representation.	Improved framework using LSTM for feature embeddings from text descriptions (titles)	Relies exclusively on text descriptions without utilizing visual features.	Inadequate exploitation of modal fusion complementarity and compatibility
(Chen et al., 2022)	Uses 2D convolution for image and 1D convolution for text to capture essential information from both modalities.	Personalized model encoding visual features and measuring attention in clothing regions.	Does not mention the effective integration and balance of text and visual information.	Lack of comprehensive multi-modal fusion in predicting clothing compatibility
(Radford et al., 2021)	Aims to provide a more balanced and integrated approach by leveraging weighted fusion of multi-modal features.	Framework based on item compatibility using title sentences for feature embedding in a style space.	Does not utilize visual features, potentially limiting the model's effectiveness in multi-modal environments.	Limited balance and integration of text and visual features in fashion item analysis
(Kim et al., 2021)	Introduces a personalized clothing prediction algorithm based on comprehensive multi-modal feature fusion to improve predictive accuracy.	Discussed challenges in modal fusion but did not propose a comprehensive multi-modal integration approach.	Insufficient exploration of effective multi-modal integration techniques.	Evaluation based on limited feature sets, leading to less effective predictions

To solve the above problems, this thesis proposes an innovative method: the multi-modal weighted combination method. This method mainly relies on visual information and aims to provide a more comprehensive representation of fashion items through the weighted fusion of multi-modal features (including visual information and text information) (Lu et al. 2021) . In addition, this thesis proposes a personalized clothing prediction algorithm based on multi-modal feature fusion (Shi et al., 2021). The algorithm consists of four main parts. The feature extraction module uses two-dimensional convolution for image feature extraction and one-dimensional convolution for text feature extraction. Its goal is to capture essential information embedded in images and descriptions.

2.9 Chapter Summary

This chapter conducts a comprehensive review of related research about clothing classification and personalized prediction. First, the literature review starts with traditional clothing classification methods and discusses classification methods based on local and global features. It also introduces classification methods based on deep learning, including convolutional neural networks, recurrent neural networks, and supervised learning methods. Following this, the literature review discusses the current research status of multi-modal feature fusion in detail, including a classification of multi-modal fusion architectures, a discussion of multi-modal fusion methods, and an overview of multi-modal fusion applications.

Then, the literature review explores research on clothing prediction and personalized recommendation. This includes a review of individualized clothing recommendation algorithms, a review of clothing sales prediction algorithms, and a combination method of personalized recommendations and clothing sales prediction.

Finally, the literature review also analyses research on clothing classification and customized prediction algorithms.

Based on these studies, this thesis proposes a personalized clothing prediction algorithm based on multi-modal feature fusion to fill the research gap in the current research field. Through comprehensive discussion, this chapter summarizes previous research results and provides a theoretical and practical basis for subsequent research.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter outlines the methodology used in this research, providing a comprehensive framework for the dataset, materials, and models involved. Section 3.2 describes the dataset in detail, with subsections 3.2.1 and 3.2.2 focusing on image and text data, respectively, explaining their structure and how they are processed. Section 3.3 presents the study materials, including the hardware and software tools utilized for the experiments. In Section 3.4, the proposed methodological framework is introduced, serving as the foundation for the subsequent model development. Section 3.5 discusses the process of model analysis and construction, leading into Section 3.6, where a personalized clothing prediction algorithm based on multi-modal feature fusion is proposed. Section 3.7 focuses on optimizing evaluation indexes for multi-classification tasks, covering topics such as multi-classification methods, confusion matrix evaluation, speed indicators, and the use of cross-validation. Section 3.8 presents the proposed models, including detailed descriptions of the TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN models, each of which is further elaborated in subsections 3.8.1 to 3.8.4. Section 3.9 covers the optimization of the models to enhance performance. Finally, Section 3.10 summarizes the content of this chapter, emphasizing the significance of the methodologies and models discussed.

3.2 Dataset

The study utilized sales records from the Italian fast fashion company Nunalie from October 2016 to December 2019 as its dataset (Ekambaram et al., 2020). This dataset

encompasses sales data for 5,355 items, providing extensive multi-modal information, including product descriptions, images, and prices. The dataset's diversity enhances the comprehensive analysis and prediction of fashion product sales performance.

The ratio of training, test, and validation sets in the dataset is 8:1:1, with image data divided into six folders. They are AI17, AI18, AI19, PE17, PE18, and PE19. In the fashion industry, AI typically stands for Autumn-Winter (from the Italian or French term "Autunno-Inverno"), while PE stands for Spring-Summer (from the Italian term "Primavera-Estate"). These folders contain 864, 999, 1,084, 955, 776, and 899 fashion clothing pictures. In addition to unstructured image data, structured data consists of 21 columns, including 1-12 weeks of clothing sales data, as well as details of season, color, fabric, day, week, month, year, label, category. The data collection method entails organizing image data within a folder containing the "images" directory and a comma-separated values (CSV) file listing the image file name. Textual data is stored within the columns of this CSV file.

Validation sets are employed throughout model training to monitor and adjust model performance. The model checkpoint callback function saves the model parameters that exhibit the best performance on the validation set. Upon completion of training, the best-performing model is loaded and evaluated using the test dataset. Confusion matrices and classification reports are generated to assess the model's performance and generalization ability. Within the dataset, picture category labels are replaced with numerical labels; for instance, "AI19" is replaced with "2" and "PE19" with "5". The algorithm implementation utilizes lambda and map functions.

3.2.1 Image Data

In the Visuelle’s dataset, image data is categorized by year and season, labelled as AI17, AI18, AI19, PE17, PE18, and PE19. Specifically, AI17, AI18, and AI19 correspond to the Autumn-Winter seasons of 2017, 2018, and 2019, while PE17, PE18, and PE19 represent the Spring-Summer seasons of the same years. These six categories reflect different fashion styles, with each style corresponding to a time window equivalent to a six-week prediction horizon.

Each product is associated with an RGB image, with resolutions ranging from 256 to 1,193 pixels in width and 256 to 1,172 pixels in height, having median values of 575 and 722 pixels, respectively. These images were captured in a controlled environment to ensure color accuracy and minimize potential prediction bias. Each image depicts a clothing item against a white background and includes a binary foreground mask to aid in distinguishing the item from the background. Figure 3.1 illustrates clothing images.

The preprocessing of the image data involves loading image files, resizing, and normalization. Subsequently, the image data is stored as NumPy arrays and loaded into variables for training and testing purposes. Each image array has a shape of (64, 64, 3), representing the width, height, and number of RGB channels.



Figure 3.1: Sample images representing various product categories within the VISUELLE dataset (Skenderi et al., 2022)

3.2.2 Text Data

Each product is linked to six folders (AI17, AI18, AI19, PE17, PE18, and PE19), which represent six distinct, non-consecutive time periods. Each period is segmented by weeks, showcasing a unique fashion style, and these were extracted and validated by the Nunalie team. Products are sold during a specific season and released on a particular day. This temporal information is recorded as a text string. During training, the code snippet initially removes image paths, release dates, and structural details from the training and test dataset. The non-numeric data is labelled and converted into an array format suitable for model input.

In the text data, Figure 3.2 shows the number of purchases by customers from January to December and the ID renderings of clothing sold by the store. Clothing colour data is included in the data set of more than 770,000 clothing sales texts. Clothing colour

data includes black, blue, brown, green, grey, red, violet, white, and yellow. As shown in Figure 3.3.

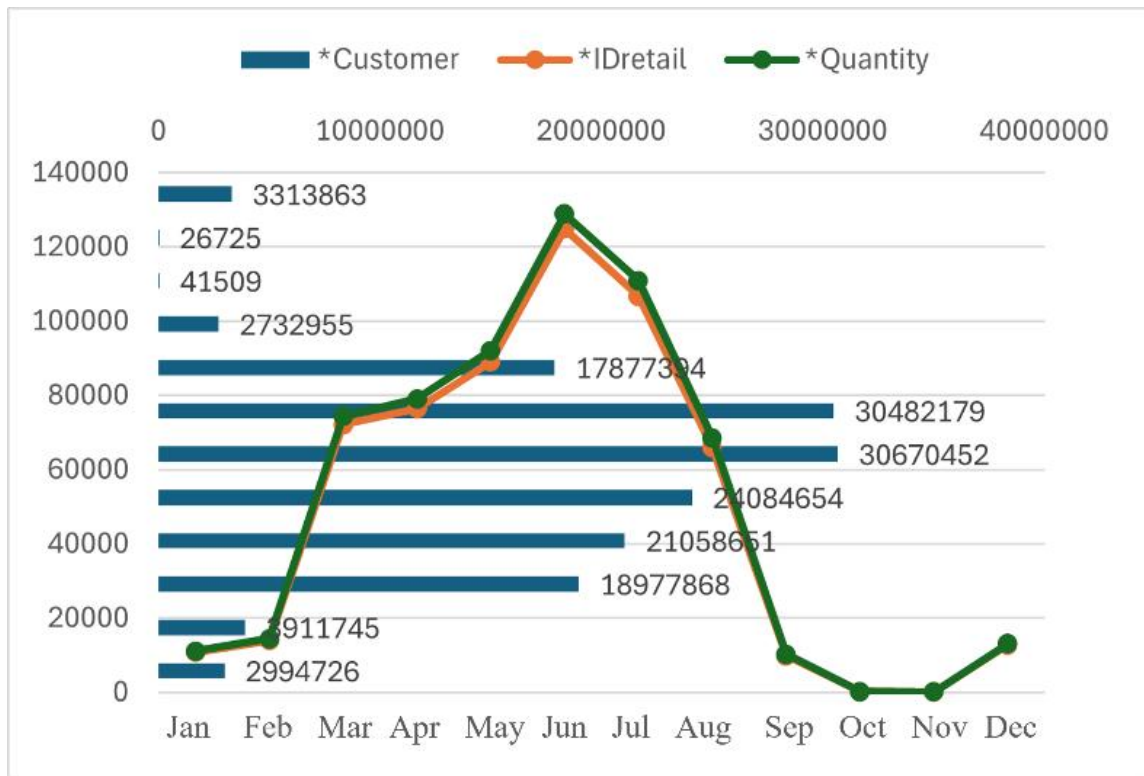


Figure 3.2: Clothing sales data (Skenderi et al., 2022)

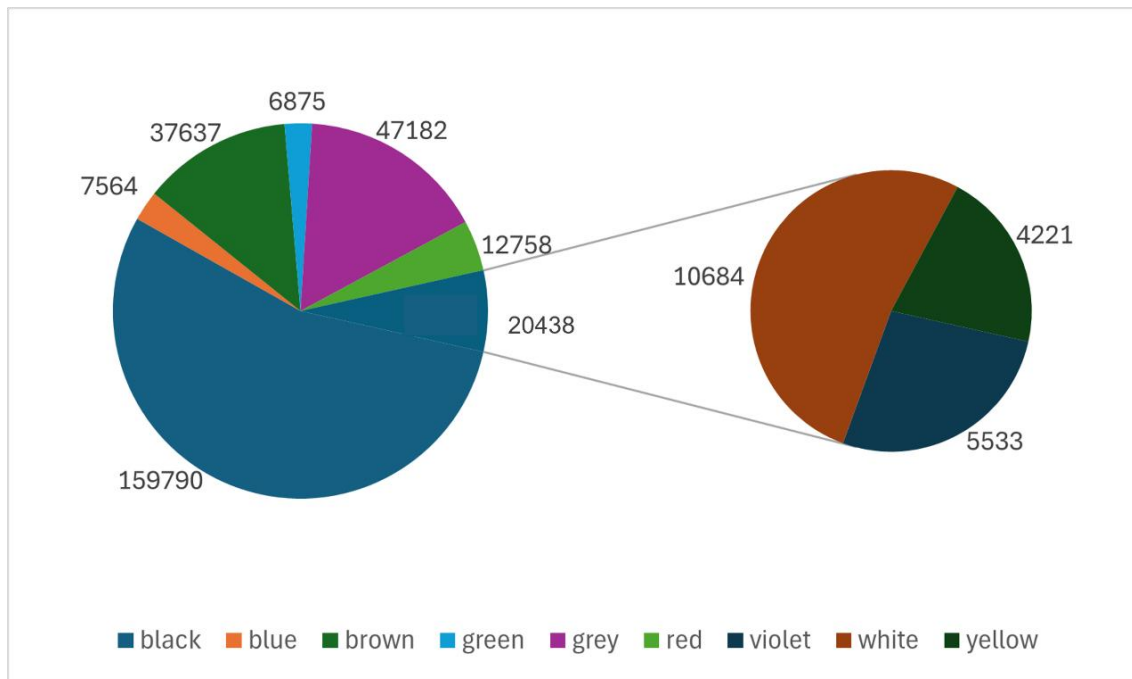


Figure 3.3: Color data distribution (Skenderi et al., 2022)

3.3 Study Materials

There are several key components to implement a multimodal fusion personalized clothing prediction algorithm. The first step in processing large-scale data and performing complex calculations is to utilize high-performance computing devices. Several algorithms are used to process these data, including pre-processing, identifying features, and recognizing images. Large-scale data can be stored and managed using storage and management systems, which provide backup, disaster recovery, access control, and other features. Personalized clothing predictions are conducted using the deep learning framework Keras. This new model is then subjected to algorithmic and model optimization. For example, parameter tuning, feature engineering, model structure adjustment, hardware optimization, etc.

3.3.1 Hardware

- i. Processor: Intel64 Family 6 Model 142 Stepping 9 processor, clocked at approximately 2701 MHz.
- ii. Memory: Physical memory overview is 12,187 MB, available physical memory is 6,056 MB. Peak virtual memory is 16,155 MB, available virtual memory is 7,352 MB.
- iii. The model was implemented using Keras and trained on an NVIDIA A100 (40GB) GPU.
- iv. BIOS version: Dell Inc. 1.5.4, dated 2018/1/31.
- v. Network Adapters: There are 4 network adapters installed in the system, including Bluetooth device, Realtek PCIe GBE Family Controller (variable bandwidth), Intel(R) Dual Band Wireless-AC 3165 (wireless bandwidth), and Microsoft Wi-Fi Direct Virtual Adapter.

To complete the research, it is necessary to have a computer with high performance. The computer requires a certain amount of memory and peak virtual memory and comes with multiple network adapters, including wired and wireless connectivity options. The computer has an older BIOS version that still supports the normal operation of this hardware configuration.

3.3.2 Software

Anaconda includes the Python interpreter and many commonly used scientific computing and data science libraries, such as NumPy, Pandas, and Keras, which typically uses TensorFlow as its backend API. It also provides a convenient package management system (Conda). The installation of Anaconda simplifies the configuration and management of the Python environment, making the setup and use of libraries like Keras and TensorFlow more convenient and efficient. The interfaces of Anaconda and PyCharm during operation are shown in Figures 3.4 and 3.5.

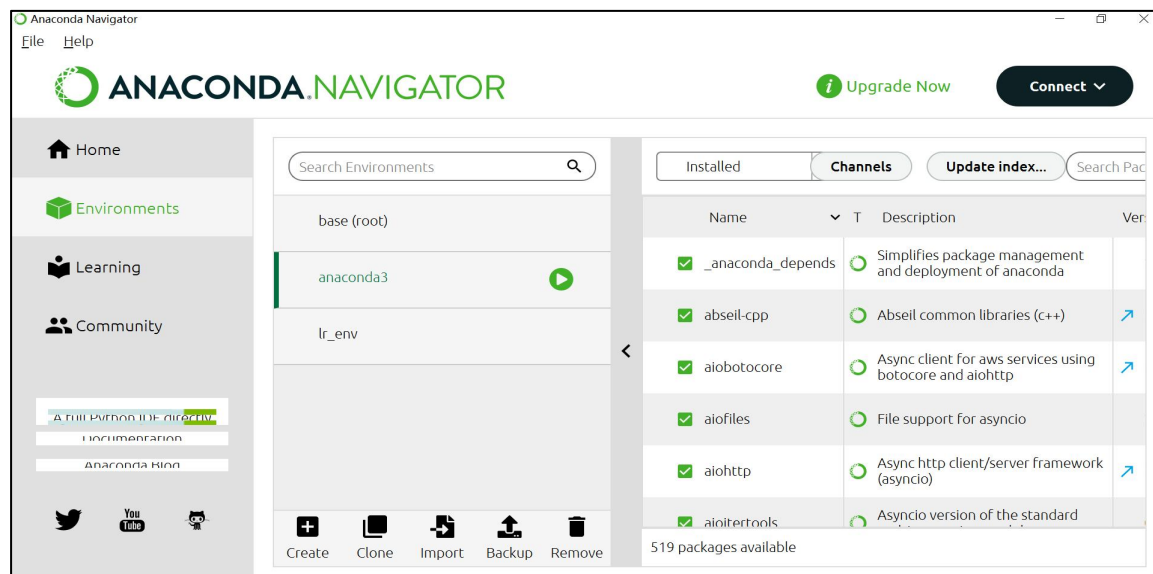


Figure 3.4: Anaconda's running interface

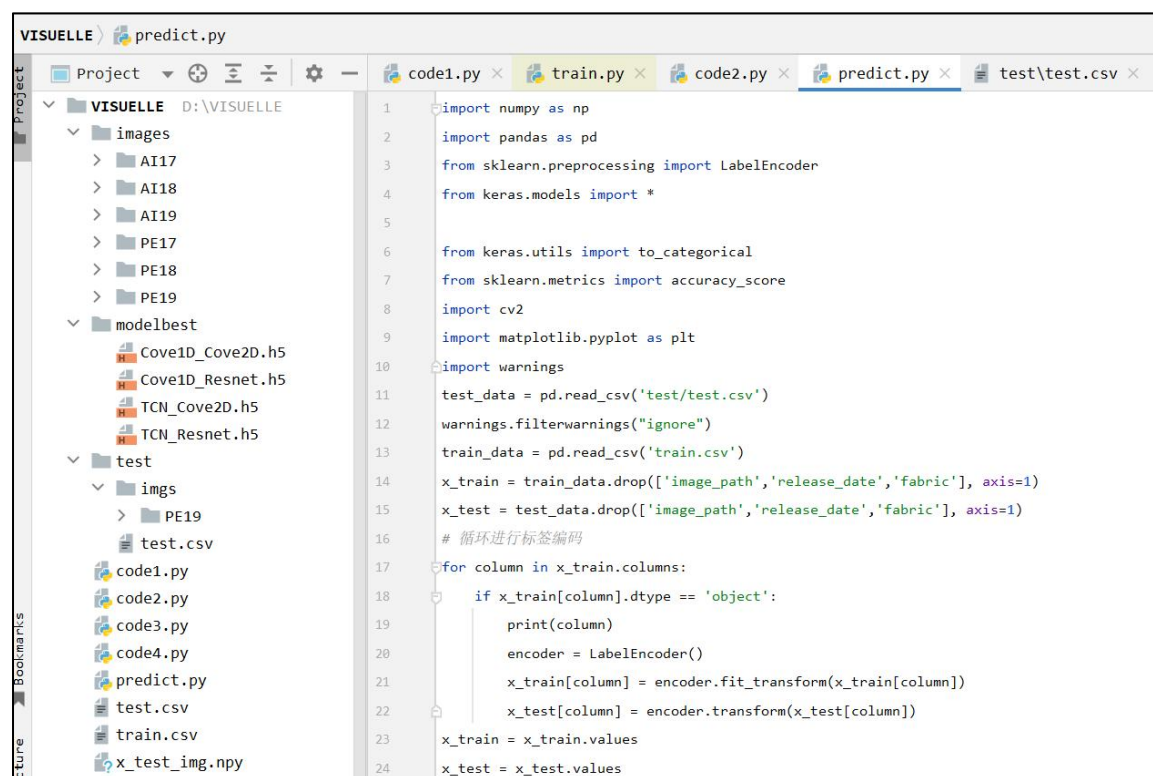


Figure 3.5: Python 3.9 running interface

3.4 Proposed Methodological Framework

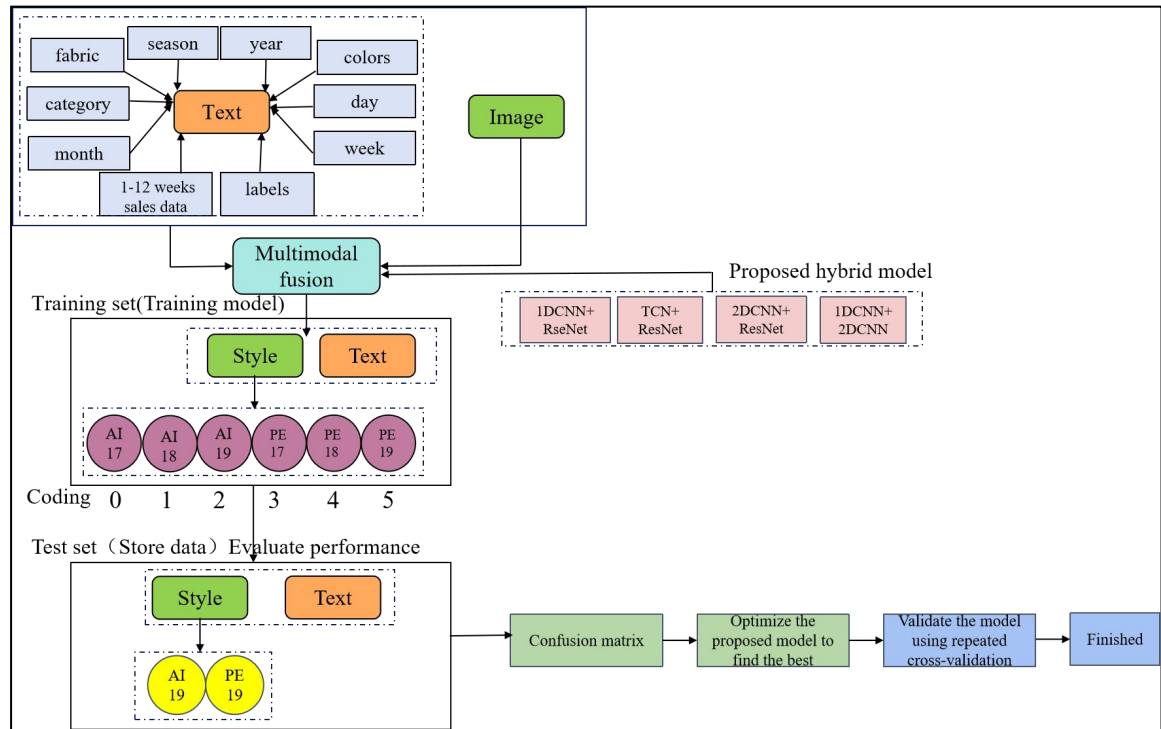


Figure 3.6: Proposed a methodological framework

The section outlines the comprehensive process of data preparation, model training, and final testing in the research on personalized clothing prediction. Figure 3.6 illustrates the overall framework of the proposed method.

Initially, the data input consists of text and image data. The text data includes attributes such as season, color, and other descriptive features extracted from product descriptions, which characterize each fashion item. The image data is categorized by year and season, labeled as AI17, AI18, AI19, PE17, PE18, and PE19, representing six distinct fashion styles corresponding to time windows equivalent to a six-week forecast period. These unique styles capture the evolving nature of consumer preferences in the fast fashion industry.

Subsequently, the process of data feature extraction and multi-modal fusion is discussed. Text and image features are processed separately: image data features are extracted using four different models (1DCNN - ResNet, TCN - ResNet, 2DCNN - ResNet, 1DCNN - 2DCNN). These extracted visual features are then fused with text features in a multimodal fusion block, generating a comprehensive feature representation that enhances the predictive ability of the model.

During the training phase, the model is trained using data from the AI17, AI18, AI19, PE17, PE18, and PE19 styles. These datasets encompass different years and seasons, enabling the model to learn diverse data patterns. In the testing phase, AI19 and PE19, which are closest to current fashion trends, are selected to verify the model, ensuring the relevance and accuracy of its predictions.

Finally, the output from the test phase is analyzed using a confusion matrix to evaluate the model's accuracy. Based on this analysis, the model is further optimized to enhance its predictive capabilities. The optimized model is validated through cross-validation to ensure robustness and reliability across different data subsets.

3.5 Model Analysis and Construction

The model analysis section comprehensively evaluates and compare four different models. First, a series of performance evaluation metrics, including accuracy, precision, recall, and F1 score, will be used to measure the performance of each model. Second, the experimental design, including each model's selection and tuning process of hyperparameters, is described below. Then, the four models will be compared in detail to analyze their advantages and disadvantages. Through in-depth analysis of experimental results, explore the performance of each model under specific conditions, as well as its

stability and generalization ability. Finally, visualization tools such as charts and curves will be used to visually demonstrate the performance and behaviour of the models, facilitating a clearer understanding of the differences, strengths, and weaknesses between them. The steps of model construction are illustrated in Figure 3.7.

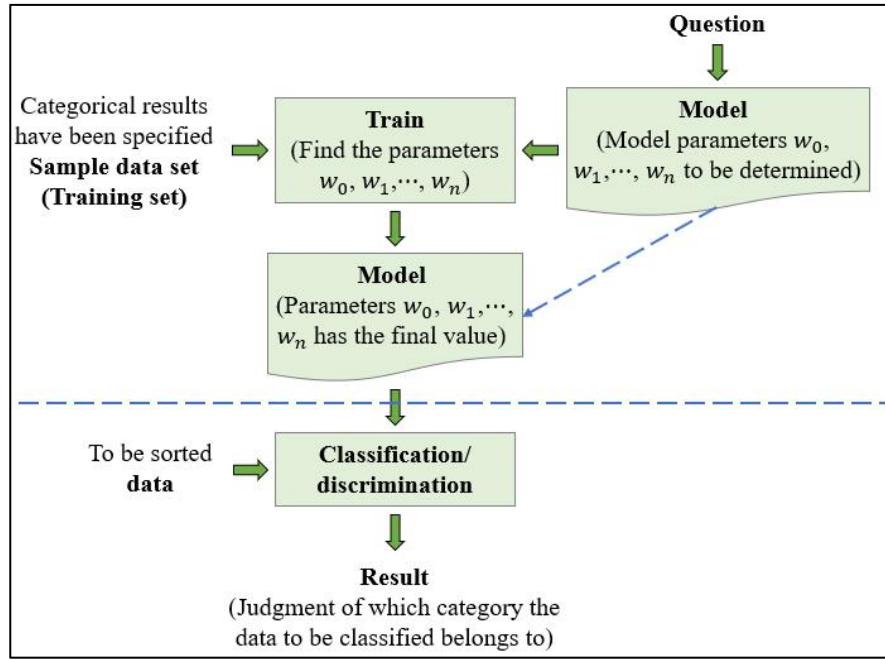


Figure 3.7: Model construction steps

3.6 Personalized Prediction Clothing Algorithm Based on Multi-modal Feature Fusion

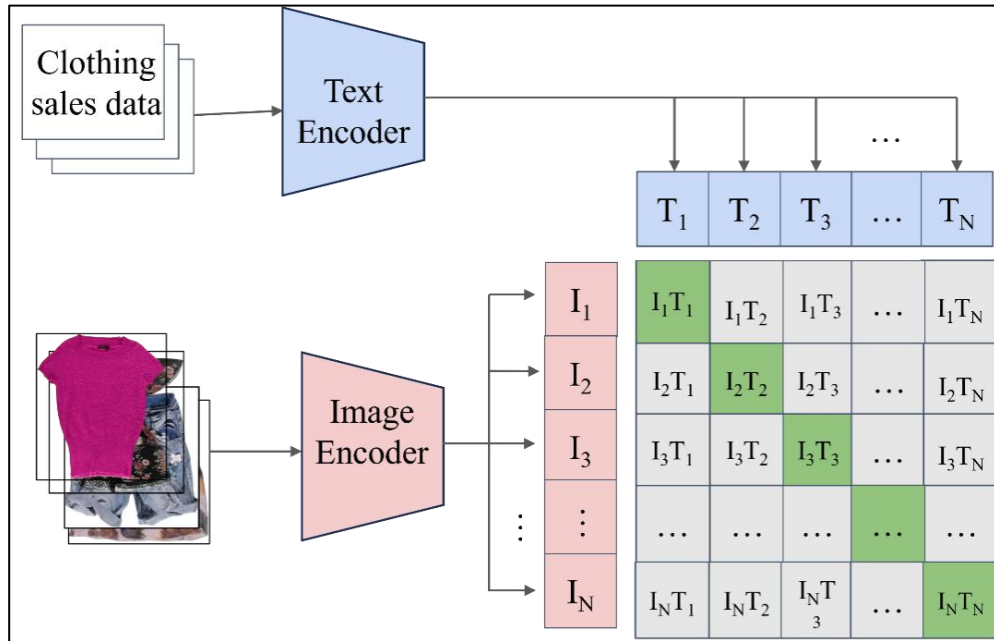


Figure 3.8: A multimodal feature fusion framework for clothing sales forecasting

Figure 3.8 illustrates a multi-modal feature fusion framework for predicting clothing sales by integrating text and image data. The framework includes a Text Encoder that processes clothing sales data, such as season, category, color, and sales figures, to extract meaningful textual features $(T_1, T_2, \dots, T_N)$. Simultaneously, an Image Encoder processes images of clothing items to extract visual features $(I_1, I_2, \dots, I_N)$, capturing the appearance, style, and design elements. These text and image features are then combined in a fusion block, where each cell represents a combination of text and image features (*e.g.*, I_1T_1, I_2T_2), forming a comprehensive feature matrix. This fusion allows the model to learn from both modalities, enhancing its ability to make personalized clothing predictions by leveraging a more complete view of each product.

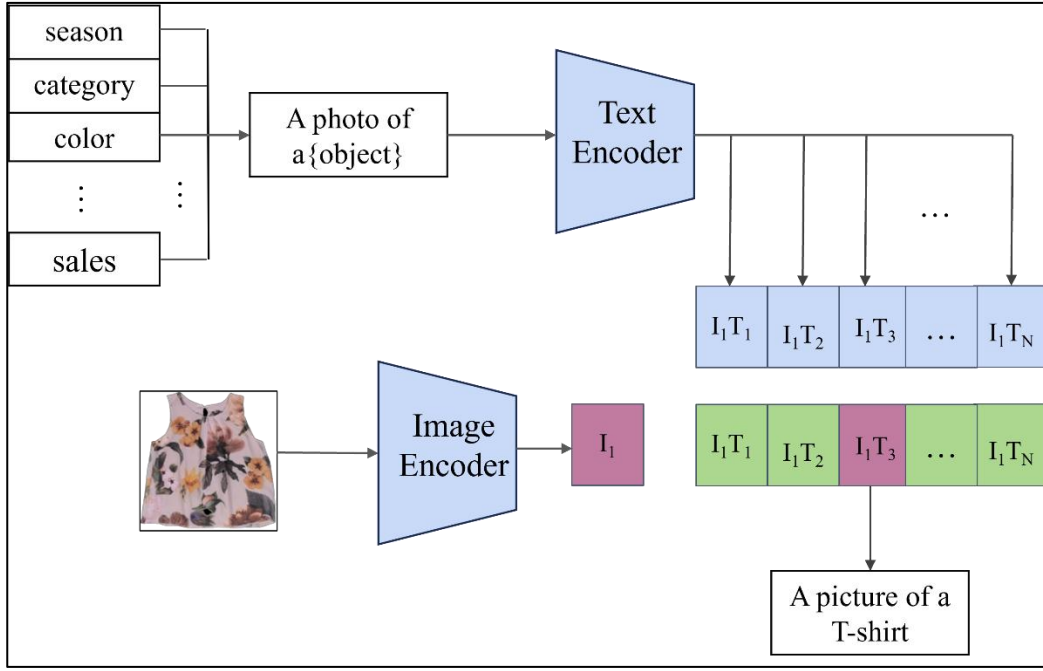


Figure 3.9: Detailed flow of multimodal feature extraction and fusion

Figure 3.9 demonstrates a detailed example of the multi-modal feature fusion process for personalized clothing prediction. The framework starts with structured text data inputs, including season, category, color, and sales information, which are transformed into a descriptive sentence like "A photo of a {object}" and processed by a Text Encoder to generate textual features ($T_1, T_2, \dots, T_N$). Concurrently, the Image Encoder extracts visual features (I_1) from the clothing image. These features are then fused into a combined representation ($I_1T_1, I_1T_2, \dots, I_1T_N$), which integrates both the descriptive text and visual elements of the clothing item. This fused representation is used to enhance the model's prediction accuracy, allowing it to provide personalized recommendations, such as identifying a picture of a T-shirt among other items.

In response to the limitations of traditional single-factor models in fully utilizing time series-related information, resulting in poor accuracy and reliability in time series prediction, a time series prediction model based on multi-modal information fusion is

proposed. For prediction, this model fuses textual and numerical data into modal data. G. Skenderi et al. (2022) proposed visual recognition involves supervised learning of human-annotated image label data and language-image contrastive learning of network-extracted image-text pairs. A novel formula is introduced by combining image labels and text into a common image-text-label space. This thesis integrates the perspectives proposed in these two studies to combine time series table data and image label data. This is for predicting fashion and clothing categories. It aims to recommend fashion categories to consumers. As shown in Figures 3.8 and 3.9, it is the study of multi-modal fusion (Radford et al., 2021).

During the multi-modal fusion process, the algorithm incorporates textual data such as user preferences, purchase history, and season to enrich the model's training dataset. As a result of this integration, user preferences and trends can be reflected in the visual characteristics of clothing.

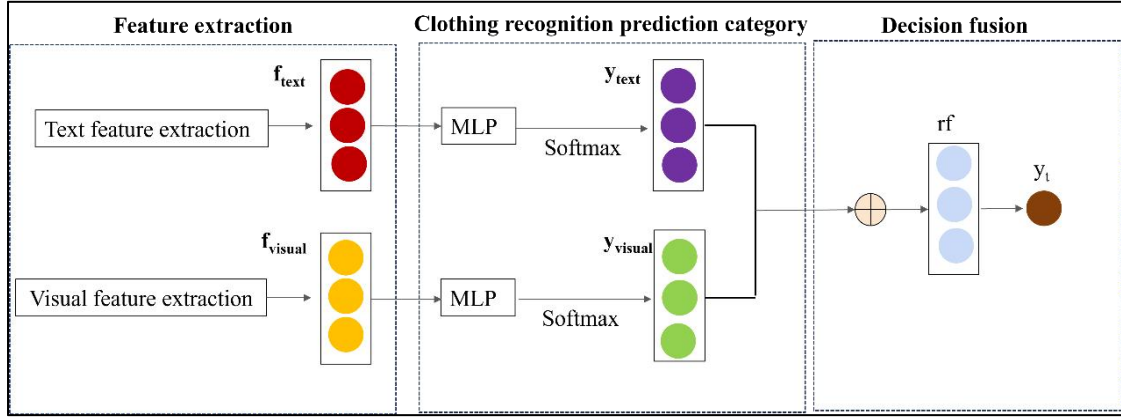


Figure 3.10: Fusion of visual features and text features

Figure 3.10 illustrates the feature processing and decision-level fusion strategy in a multi-modal clothing recommendation system. It details how features are extracted from both text and visual data, processed through multi-layer perceptrons (MLP), and classified via Softmax layers, with the results fused at the decision level to generate the final recommendation. For text feature extraction, methods like Temporal Convolutional Network (TCN) or 1D Convolutional Neural Network (1D CNN) are used to process input data such as clothing sales, categories, and colors. These models, including TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN, convert the text data into context embedding vectors (f_{text}) that capture semantic features. These vectors are then fed into an MLP for further transformation, resulting in classification probabilities (y_{text}) through a Softmax layer. Similarly, visual feature extraction is performed using ResNet or 2D Convolutional Neural Networks (2D CNN) on clothing images, producing visual feature vectors (f_{visual}) that are also processed by an MLP and classified via Softmax to generate probabilities (y_{visual}). The decision-level fusion combines these independent classification outputs (y_{text} and y_{visual}) to produce the final recommendation (y_t), and the performance is evaluated using confusion matrices to assess predictive accuracy across different categories, enhancing the system's personalized recommendation capabilities.

3.7 Optimization of Evaluation Indexes for Multi-Classification

Based on multi-modal fusion, this part discusses how to optimize the multi-classification evaluation index. It will first be introduced how multi-classification works, as well as how it applies to customized clothing forecasting. Afterwards, a confusion matrix will be discussed in depth and its role in evaluating classification performance. After completing the algorithm development, common multi-classification evaluation indicators such as accuracy, recall, and F1 score were used to optimize its performance. To ensure algorithm robustness and reliability, discuss repeated cross-validation and its application to evaluating model performance and parameter tuning.

3.7.1 Multi-Classification

For personalized clothing predictions, the model uses a multi-modal fusion approach, integrating both image data and textual attributes (such as season, category, color and fabric etc). This data is used to train the model in a multi-class classification task, categorizing fashion items into distinct classes based on trends across different years and seasons.

The model is trained on six fashion styles (AI17, AI18, AI19, PE17, PE18, PE19), which capture evolving trends and consumer preferences. By fusing image and textual features, the model learns to represent each item comprehensively, improving classification accuracy by capturing the relationships between visual and descriptive features.

During testing, the model predicts preferences for the most recent trends (AI19, PE19), which are aligned with current fashion. The model's performance is evaluated using a confusion matrix to measure its accuracy in categorizing items into the correct fashion categories.

The model produces two key outputs: it predicts various fashion attributes for each item, such as season, style, and color, by analyzing diverse trend data (AI17, AI18, AI19, PE17, PE18, PE19), offering a detailed depiction of how an item fits into different fashion categories. Additionally, the model is evaluated using data from AI19 and PE19 to ensure accurate trend predictions, with a confusion matrix employed to assess its classification accuracy.

As shown in Figure 3.11, the multi-classification framework is fundamental to the model's operation. Supervised learning, specifically multi-class classification, is applied to handle the multiple categories of fashion items. This ensures that the model not only understands and predicts fashion trends but also provides accurate, personalized recommendations based on users' preferences, derived from their browsing and purchase history.

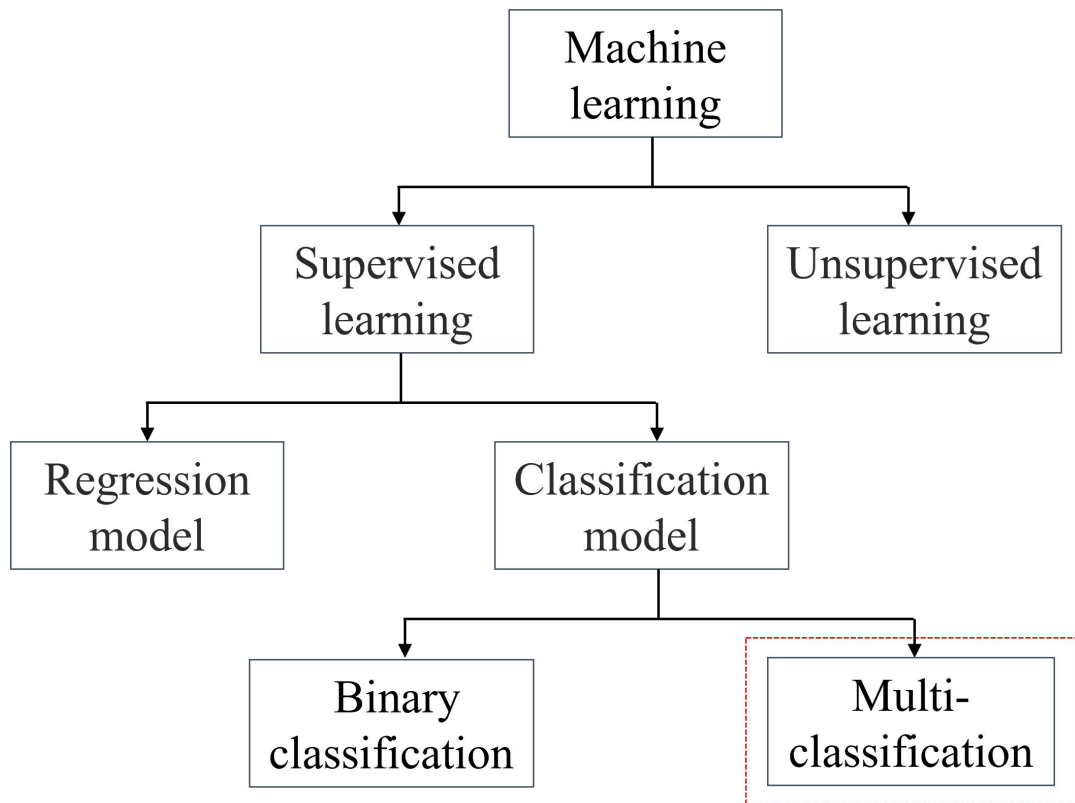


Figure 3.11: Multi-classification in machine learning

3.7.2 Confusion Matrix and Evaluation

The Confusion Matrix, also called an error matrix, is a valuable tool for evaluating models' performance in classification problems in supervised learning. In this thesis, the confusion matrix clearly represents the disparities between the algorithm model's classification outcomes and the actual scenario. It is primarily applied to multi-classification problems. Table 3.1 shows the confusion matrix, which provides detailed explanation and analysis. True Positive (TP) denotes the count of samples accurately predicted as positive by the model. False Positive (FP) signifies the count of samples erroneously classified as positive by the model, despite being false positive. True Negative (TN) represents the count of samples accurately predicted by the model in the negative

class. False Negative (FN) indicates the count of samples incorrectly classified as negative by the model, although they are correct.

Table 3.1: Confusion matrix

Class	Predict positive class	Predict negative class
Actual positive class	TP	FN
Actual negative class	FP	TN

Through the confusion matrix, a variety of performance indicators can be calculated to evaluate the performance of the classification model, such as accuracy, precision, recall or true positive rate (TPR), false positive rate (FPR) and F1 score, etc. The advantage of the confusion matrix is that it not only provides the number of misclassifications but also illustrates which categories have the most prediction errors, which is very helpful for improving classification algorithms and model tuning.

Accuracy is the most intuitive performance indicator, and its calculation formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Equation 3.1}$$

Precision focuses on how many of the samples predicted as positive are positive. The calculation formula is:

$$Precision = \frac{TP + TN}{TP + FP} \quad \text{Equation 3.2}$$

Recall or true rate (TPR) focuses on how many of all TP samples are predicted as positive by the model. The calculation formula is:

$$Recall = \frac{TP + TN}{TP + FN} \quad \text{Equation 3.3}$$

The F1 score is the harmonic mean of precision and recall and is calculated as:

$$F1 \text{ score} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad \text{Equation 3.4}$$

Macro avg, which is the average precision of all classes, where n represents the class, is calculated by:

$$Macro \text{ avg} = \frac{1}{n} \sum_{i=1}^n P_i \quad \text{Equation 3.5}$$

Weighted avg is an improvement on the macro average, considering the proportion of the number of samples of each category in the total sample:

$$Weighted \text{ avg} = \sum_i w_i P_i \quad \text{Equation 3.6}$$

3.7.3 Speed Indicator

In this thesis, factors evaluated for the model's running speed included computational effort floating point operations (FLOPs), runtime frames per second (FPS), and the number of parameters (Params). To comprehensively assess performance, the experiment adopted a deep learning architecture with convolutional, pooling, and fully connected layers. Layers were added with different numbers of nodes and layers. Rectified linear unit (ReLU) activation functions and the Adam optimizer enhanced generalization

and stability. An appropriate initial learning rate was set, and Dropout layers were added to mitigate overfitting and improve regularization. To enhance generalization and stability, ReLU activation functions and the Adam optimizer were used. An appropriate initial learning rate was set, and dropout layers were added to mitigate overfitting and improve regularization.

In data preprocessing, input data is encoded, features are selected, and image data is scaled. Additionally, pre-trained model parameters are loaded. Initially, the model's accuracy is assessed, and once a certain accuracy threshold is reached, a speed index is introduced to evaluate performance. Despite utilizing CPU-only training without GPU support, hardware resource demands for forward propagation are still considered. Moreover, Params post-training reflects the model's memory footprint. As convolutional operations entail substantial computation during forward propagation, FLOPs are employed to calculate floating point operations per second. This enables fine-tuning of the model structure and comparing computational complexity among different models. C_i , C_0 represents input and output channels, K represents the size of the convolution kernel, and H and W are the size of the output feature map.

$$FLOPs = (2 \times C_i \times K^2 - 1) \times H \times W \times C_0 \quad \text{Equation 3.7}$$

$$(2 \times C_i \times K^2 - 1) = (C_i \times K^2) + (C_i \times K^2 - 1) \quad \text{Equation 3.8}$$

While FLOPs estimate computational complexity, the execution time can be influenced by factors like hardware performance and data size. Therefore, a time function was incorporated to calculate training time, representing training speed and providing a direct understanding of the model's performance on CPU hardware.

3.7.4 Application of Cross-validation

Cross-validation involves dividing a dataset into smaller parts and repeatedly training and validating the model to assess its generalization performance and stability. In this thesis, the dataset was split into training (80%), validation (10%), and test sets (10%) in an 8:1:1 ratio. The training set was used for cross-validation and model training, the validation set for interim evaluation and parameter tuning, and the test set for final model evaluation.

First, a 10-fold cross-validation was performed on the training set. The training set was divided into 10 equal parts (folds). In each iteration, one fold was used as the validation fold, and the remaining nine folds were used for training. This process was repeated 10 times, ensuring that each fold was used for validation once. Performance metrics were recorded at each iteration. After completing the first round of cross-validation, the training set was reshuffled, and a second round of 10-fold cross-validation was performed. This repeated validation ensured the model was evaluated on different data partitions, providing a more comprehensive and impartial performance assessment. Performance metrics such as accuracy were recorded for each iteration of the two cross-validation rounds. After both rounds were completed, the results were averaged to provide a stable and reliable estimate of model performance. The optimal model was selected based on the average performance across all iterations. Next, the validation set (10%) was used to further fine-tune the model and adjust hyperparameters. The validation set remained independent of the training set and was not used during cross-validation.

Finally, the test set (10%) was used to evaluate the final model, ensuring it performed well on unseen data and fully validating its generalization ability. Thus, the

repeated 10-fold cross-validation method combined with the use of validation and test sets ensured the model's robustness, reliability, and accuracy. This comprehensive evaluation of the model's generalization significantly reduced the risk of overfitting, resulting in an optimized, stable model suitable for future prediction tasks.

3.8 Proposal Model

In this chapter, four models—TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN—are proposed to enhance personalized clothing prediction through multi-modal data fusion. These models combine temporal and spatial feature extraction using convolutional and residual networks, aiming to improve prediction accuracy and provide a comprehensive approach to fashion data analysis. Each model is specifically designed to tackle different aspects of the prediction task, offering a robust framework for forecasting fashion trends.

3.8.1 Proposal TCN-ResNet Model

The proposed TCN-ResNet model combines the strengths of Temporal Convolutional Networks (TCN) and ResNet for personalized clothing prediction by integrating both sequential text data and image data.

TCN is employed for processing sequential data, such as fashion-related attributes (e.g., season, category, color, fabric). By using dilated convolutions and residual connections, TCN captures long-range dependencies in the input sequences efficiently, outperforming traditional RNN models such as LSTM and GRU. In this thesis, TCN handles a 21×1 text feature sequence, expanding the model's receptive field without significantly increasing the parameter count, making it well-suited for sequential patterns in fashion trends. ResNet, renowned for its deep learning architecture, is used to extract

spatial features from clothing images. By incorporating residual connections, ResNet enables the training of deeper networks without suffering from the vanishing gradient problem, making it ideal for extracting detailed visual information from images. The ResNet portion in this model captures crucial visual features that aid in distinguishing subtle differences between clothing items. The TCN-ResNet model fuses both types of data by connecting the TCN and ResNet outputs via a concatenate layer. This hybrid approach combines temporal patterns from text data and spatial features from image data, which are then fed into a fully connected layer with a Softmax activation function for personalized clothing classification predictions. Figure 3.12 shows the architecture of the proposed TCN-ResNet model for personalized clothing prediction.

The model was implemented using the Keras functional API, and both the TCN and ResNet components were trained from scratch. Image and sequential data were preprocessed, and the ResNet50V2 network was utilized to handle image inputs, while TCN processed the sequential text features. The model training involved careful tuning of parameters such as epoch count and batch size, with the use of callbacks to save the best-performing models. By leveraging the complementary strengths of TCN for text data and ResNet for image data, the TCN-ResNet model effectively addresses the challenge of personalized clothing prediction with a balance between model complexity and performance.

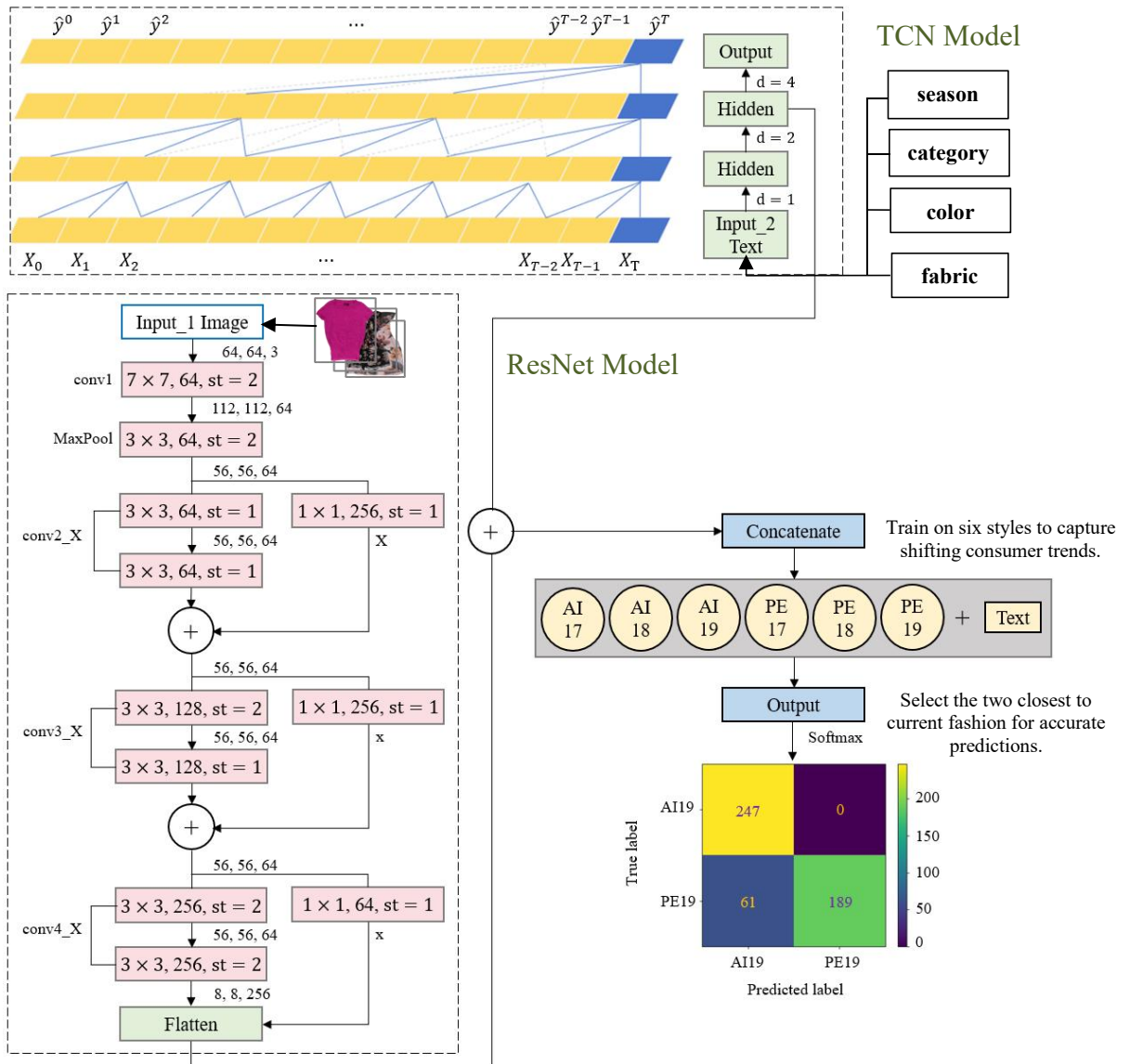


Figure 3.12: TCN-ResNet model

3.8.2 Proposal 1DCNN-ResNet Model

The 1DCNN-ResNet model is designed for personalized clothing classification by combining one-dimensional convolutional neural networks (1DCNN) for text data and the ResNet architecture for image data. This hybrid model processes both text and image inputs, fusing the extracted features to make classification predictions.

The 1DCNN component is responsible for handling textual data, such as clothing descriptions and categorical attributes. Its structure includes multiple 1D convolutional layers followed by pooling layers, efficiently capturing local dependencies in sequential data. This lightweight architecture processes text inputs (e.g., a 21×1 feature sequence) and outputs a feature vector representing key patterns in the text. The ResNet component processes image data, using convolutional and identity blocks to extract deep spatial features. These features are essential for identifying subtle visual differences in clothing items. After passing through several layers, the image data are transformed into high-dimensional feature vectors. Once the text and image features are extracted, the outputs from the 1DCNN and ResNet branches are concatenated to form a unified feature vector. This combined vector is then passed through a fully connected layer for final classification using a Softmax activation function. Figure 3.13 shows the architecture of the proposed 1DCNN-ResNet model for personalized clothing prediction.

The model was built using the Keras functional API, with CSV files used to load and preprocess the datasets. During training, the model checkpoint callback saved the best model, and training performance was monitored via loss and accuracy curves. After training, the best model was evaluated using confusion matrices and classification reports to assess prediction accuracy. This fusion of 1DCNN and ResNet leverages both textual

and visual data, offering a comprehensive and efficient solution for personalized clothing prediction tasks.

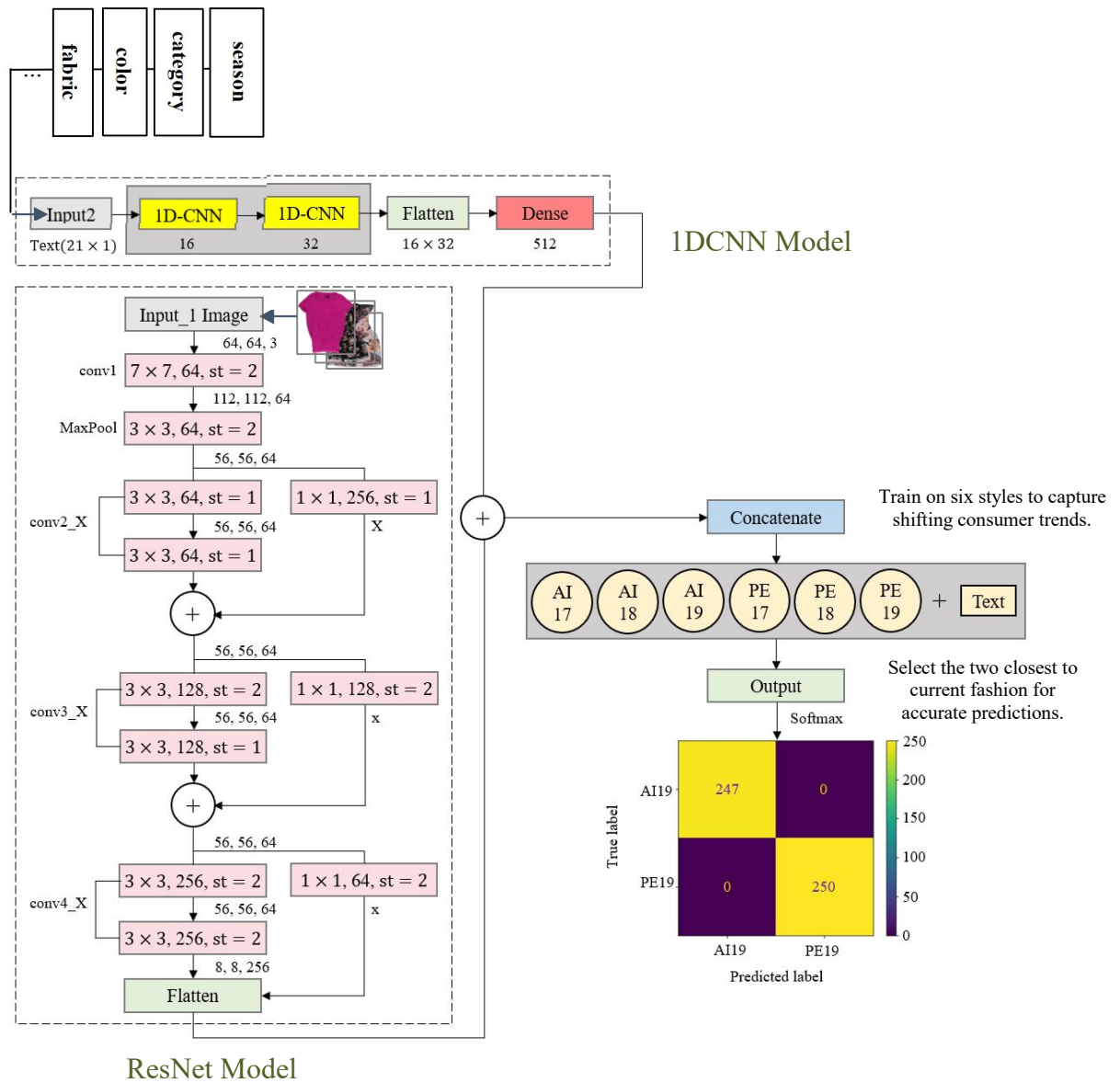


Figure 3.13: 1DCNN-ResNet model

3.8.3 Proposal TCN-2DCNN Model

The TCN-2DCNN model integrates both Temporal Convolutional Networks (TCN) for handling sequential data and 2D Convolutional Neural Networks (2DCNN) for processing image data, aiming to improve personalized clothing prediction tasks by leveraging multi-modal data.

The TCN component, designed for time series data, employs residual blocks (ResBlocks) to address common challenges in deep network training, such as vanishing and exploding gradients. Each ResBlock includes two Conv1D layers with residual connections, allowing the model to learn residual information and train more effectively. This design helps in capturing temporal dependencies within the sequential data, such as clothing attributes, while improving the model's feature representation and overall generalization. For the image data, the 2DCNN component extracts essential spatial features by applying multiple convolutional and pooling layers. This architecture is crucial for capturing fine-grained visual patterns in clothing items, ensuring the model can distinguish subtle differences in visual structures. The image input is processed into feature maps that gradually reduce in spatial resolution but increase in feature depth, resulting in a rich hierarchy of visual information. Figure 3.14 shows the architecture of the proposed TCN-2DCNN model for personalized clothing prediction.

The model fuses the features extracted from both the sequential and image data streams and generates predictions through fully connected layers with Softmax activation. The training process includes regular validation, with the best model saved for final testing. Performance is evaluated using confusion matrices and classification reports to ensure accuracy. By combining temporal and spatial feature extraction, the TCN-2DCNN model

offers a robust solution for personalized clothing prediction, effectively utilizing both text and image data to improve prediction accuracy while maintaining computational efficiency.

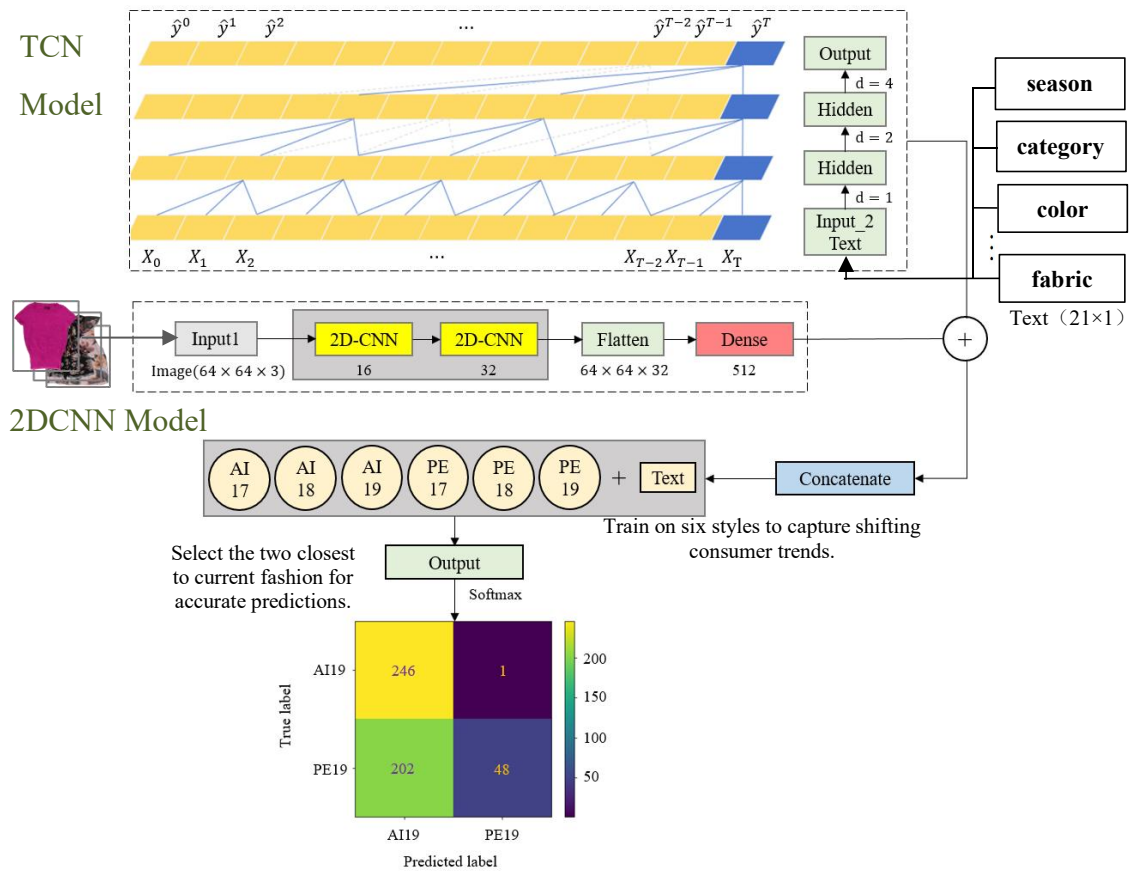


Figure 3.14: TCN-2DCNN Model

3.8.4 Proposal 1DCNN-2DCNN Model

First, the data from the training and test sets are pre-processed and labeled. The 1DCNN-2DCNN model consists of two inputs: image data and textual time series data. Image data undergoes processing through convolutional layers and fully connected layers, while temporal data is processed through one-dimensional convolutional layers and fully connected layers. The features extracted from these two data streams are merged, and predictions are generated through fully connected layers and SoftMax activation functions. Finally, confusion matrices and classification reports evaluate the model's performance.

The algorithm is based on the 1DCNN-2DCNN model, as shown in Figure 3.15. Two input methods are used: image data and time series data. 2DCNN is used as input 1 to process time series data, and 1DCNN is used as input 2 to process image data.

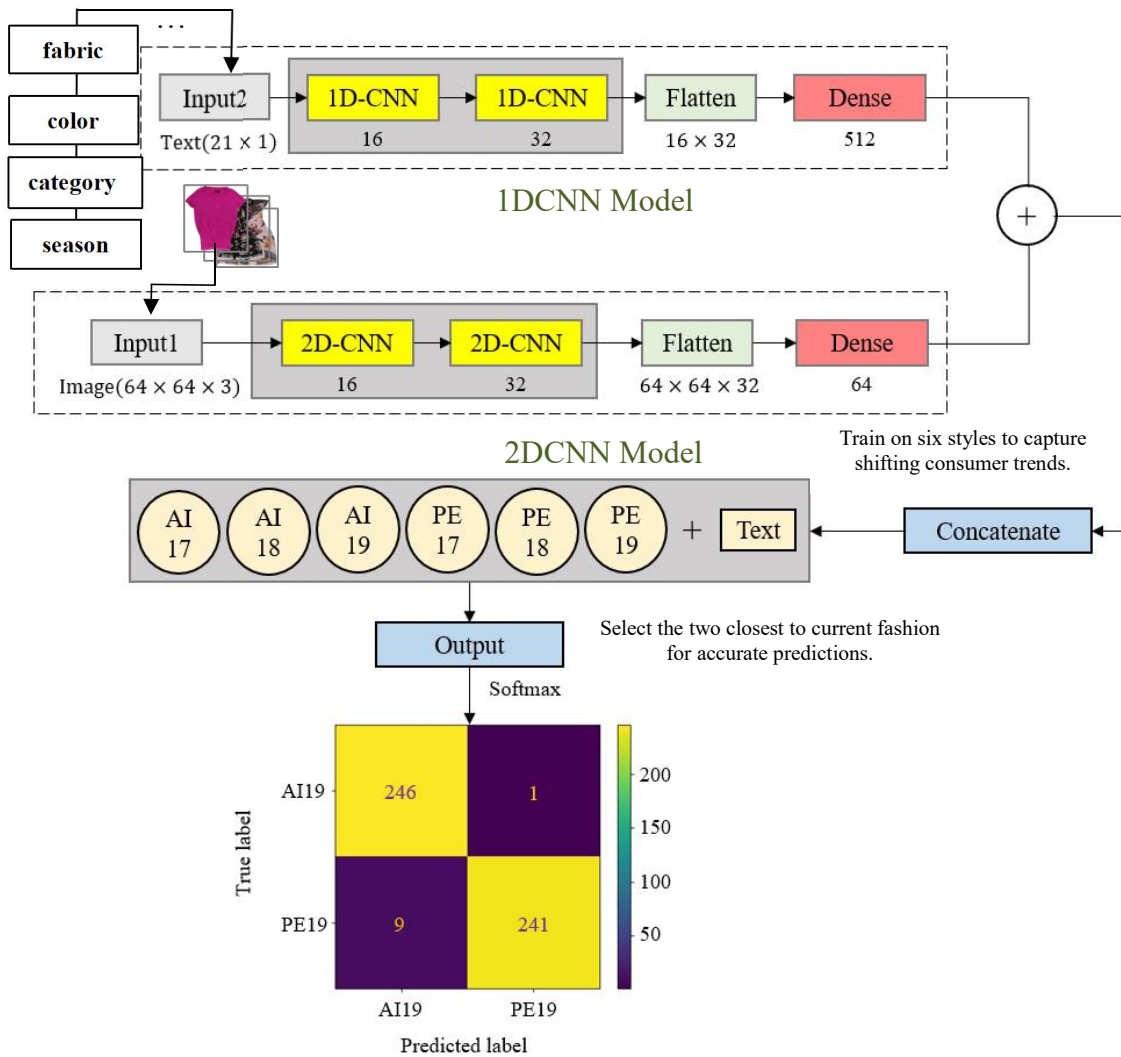


Figure 3.15: 1DCNN-2DCNN model

3.9 Model Optimization

Model optimization of the multi-modal fusion personalized clothing prediction algorithm should comprehensively consider the following aspects:

First, massive amounts of structured and unstructured data need to be effectively processed and analysed, including the use of technologies such as image recognition, annotation, and fusion of tag data. This step is the basis for the algorithm optimization of the thesis and can provide personalized clothing design. When one-dimensional convolution is combined with two-dimensional convolution, the case of processing input images with multiple channels is considered. Assume that the input image is X , and its size is $H \times W \times C$, where H represents the height of the image, W represents the width of the image, and C represents the number of channels of the input image. The convolution kernel is W , and its size is $M \times N \times C$, where M represents the height of the convolution kernel, N means the width of the convolution kernel, and C represents the number of channels in the input image. Then, the calculation formula for the combination of one-dimensional and two-dimensional convolution is as follows, $(X \times W)[i, j]$ representing the pixel value of the output image $i = 1, 2, 3, \dots, H - M + 1, j = 1, 2, \dots, W - N + 1$.

$$(W \times W)[i, j] = \sum_{u=1}^M \sum_{v=1}^N \sum_{c=1}^C X[i+u-1, j+v-1, c] \cdot W[u, v, c]$$

Equation 3.9

Secondly, on the basis of feature extraction from clothing pictures, the relevant information of structured data text and unstructured data images is integrated through feature space to more accurately capture the user's personalized needs and preferences. In order to improve the convergence speed and generalization ability of the model, this thesis uses the gradient descent method, Adam optimization algorithm and learning rate

scheduling to optimize the proposed model. Among them, the gradient descent method minimizes the loss function by updating parameters along the opposite direction of the gradient, while the Adam algorithm combines the first-order and second-order moment estimates of the gradient to achieve adaptive learning rate adjustment. Among them, learning rate scheduling is a method of dynamically adjusting the learning rate to balance the convergence speed and model performance during the training process.

Finally, regularization is used in the study to reduce overfitting of the model. Reduce the complexity of the model by adding a regularization term to the loss function to limit the size of the model parameters. At the same time, batch normalization technology is used to accelerate model training and improve the generalization ability of the model. This method performs a normalization operation on the input of each hidden layer, so that the input of each layer has a similar distribution, making it easier to train and optimize the model.

3.10 Chapter Summary

This chapter presented a comprehensive overview of the methodology employed in this thesis. It began by detailing the dataset, which forms the basis of the research, with a focus on both image and text data, along with their respective structures and processing techniques. The study materials, including the hardware and software tools used throughout the research, were also described. Following this, the proposed methodological framework was outlined, leading into a thorough discussion on the model analysis and construction processes.

A key contribution of this chapter is the introduction of a personalized clothing prediction algorithm based on multi-modal feature fusion, where its design principles and

implementation were explained in detail. In addition, the chapter addressed the optimization of evaluation metrics for multi-classification tasks, including multi-classification techniques, confusion matrix analysis, speed indicators, and the application of cross-validation. Moreover, four models—TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN—were proposed and discussed in terms of their structures and roles in enhancing prediction accuracy. Model optimization strategies were also explored to improve both the performance and efficiency of these models.

In conclusion, this chapter emphasized the significance of the proposed algorithms and models in advancing the field. Future research directions and challenges were highlighted, particularly regarding further performance improvements and potential broader applications of the models. This lays a solid foundation for the subsequent chapters and future advancements in this domain.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

This research comprehensively evaluates and compares the four proposed models (TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN). Sections 4.2 to 4.4, experiments are conducted on the proposed model, and comparative analysis is performed to obtain the experimental results. Section 4.5, a cross-validation step is added to verify the generalization ability and robustness of the proposed model. Section 4.6, the entire experiment is comprehensively summarized. The results show that the 1DCNN-ResNet and 1DCNN-2DCNN models show the best performance.

4.2 Comparison between TCN-ResNet model and 1DCNN-ResNet model

According to the experimental results in Table 4.1, the TCN-ResNet model and 1DCNN-ResNet model are compared and analyzed. The table shows traversal training for six different styles of clothing, where 0 indicates AI17, 1 indicates AI18, 2 indicates AI19, 3 indicates PE17, 4 indicates PE18, and 5 indicates PE19. In this thesis, the test set for all seasons contained two styles, the "2" and "5" categories, representing AI19 and PE19, respectively.

The model was trained on six unstructured data folders and customer purchase history text. The reason for training on six types of data but only testing on two is to ensure that the model captures a broad range of fashion trends (AI17, AI18, AI19, PE17, PE18, PE19) over time. However, the test set focuses on the two most recent and relevant styles, AI19 and PE19, which are the most representative of current fashion trends. This

approach ensures the model can provide accurate and up-to-date predictions that align with consumer preferences in today's market.

Examples of unstructured data folders include images of clothing items from different seasons, such as AI19 or PE19, which serve as the input for the model. Customer purchase history text includes both structured and unstructured data, such as purchase dates, product descriptions, customer preferences (such as color, fabric), and transaction history. This data helps the model learn patterns in consumer behavior and preferences.

After the training model was obtained, the personalized prediction structure was generated using images and customer records from the store's AI19 and PE19 test set folders. It was observed that the 1DCNN-ResNet model performed significantly better than the TCN-ResNet model in terms of prediction accuracy, indicating that the 1DCNN-ResNet model is more effective for the clothing classification prediction task. Specifically, the prediction accuracy, precision, recall, etc., all showed a high performance, reaching 100%. However, despite its higher accuracy, the 1DCNN-ResNet model was slower than the TCN-ResNet model, with a processing time of 4.952 seconds for the given dataset.

Table 4.1: Comparison between TCN-ResNet model and 1DCNN-ResNet model

Model	Clothing Style	Precision	Recall	Model (F1 score)	Support	Macro avg	Weighted avg	Confusion matrix	Model Accuracy	Prediction data	Speed
TCN-ResNet	2	1.0000	0.8901	0.8901	247	0.9010	0.8780	0.8756	0.8773	497	4.952s
	5	1.0000	0.7560	0.8610	250	0.9016	0.8773	0.8755			
1DCNN-ResNet	2	1	1	1	247	1	1	1	1	497	11.6s
	5	1	1	1	250	1	1	1			

4.3 Comparison between TCN-ResNet model and TCN-2DCNN model

In this thesis, the model was trained using unstructured data from six folders and text from customer shopping records. The six types of data (AI17, AI18, AI19, PE17, PE18, and PE19) were used to ensure that the model could learn from diverse data patterns across different years and seasons. This variety enabled the model to generalize more effectively by exposing it to various styles and trends. After training, the study focused on styles related to the latest fashion—AI19 and PE19—to assess the model's ability to predict current fashion trends. By selecting these specific styles, we ensured that the model's performance was evaluated based on current market conditions, making the predictions more applicable and accurate for real-world use. The test set contained these two styles, AI19 and PE19, represented by the "2" and "5" categories, respectively.

According to the results in Table 4.2, the TCN-ResNet model and TCN-2DCNN model were comprehensively evaluated. It was observed that the accuracy of the TCN-ResNet model (87.73%) was significantly higher than that of the TCN-2DCNN model (59.15%). However, as shown in Table 4.1, the evaluation metrics of the TCN-ResNet model were lower than those of the 1DCNN-ResNet model, indicating that 1DCNN-ResNet still outperforms it in certain aspects.

Table 4.2: Comparison between TCN-ResNet model and TCN-2DCNN model

Model	Clothing Style	Precision	Recall	Model (F1 score)	Support	Macro avg	Weighted avg	Confusion matrix	Model Accuracy	Prediction data	Speed
TCN-ResNet	2	1.0000	0.8901	0.8901	247	0.9010	0.8780	0.8756	0.8773	497	4.952s
	5	1.0000	0.7560	0.8610	250	0.9016	0.8773	0.8755			
TCN-2DCNN	2	0.5491	0.9960	0.7079	247	247	0.7643	0.5940	0.5915	497	5.03s
	5	0.9796	0.1920	0.3211	250	250	0.7656	0.5915			

4.4 Comparison between TCN-ResNet model and 1DCNN-2DCNN model

During the experiment, the 1DCNN-2DCNN model demonstrated excellent performance in both training and validation. As shown in Figures 4.1 and 4.2, the training loss and validation loss (orange and blue curves, respectively) steadily decrease throughout the training process, with only a minimal gap between them. This indicates that the model performs similarly on both the training and validation sets, with no significant signs of overfitting. Additionally, the training accuracy and validation accuracy remained highly stable, further confirming the strong generalization ability of the 1DCNN-2DCNN model.

In contrast, Figures 4.3 and 4.4 reveal that although the 1DCNN-ResNet model performs well in terms of loss and accuracy on the training set (orange curve), its performance on the validation set (blue curve) exhibits considerable fluctuations. Particularly during the mid-to-late stages of training, the validation loss increases significantly, and validation accuracy declines, suggesting potential overfitting issues with the model.

Thus, the 1DCNN-2DCNN model not only surpasses the 1DCNN-ResNet model in training speed, but also demonstrates superior generalization ability, and greater stability in both loss and accuracy. As presented in Table 4.3, the experimental results indicate that the 1DCNN-2DCNN model achieved a prediction accuracy of 97.99% and an F1 score of 98.01% in the clothing sales forecasting task, significantly outperforming the TCN-ResNet model. Additionally, compared to the 1DCNN-ResNet model, the 1DCNN-2DCNN model offers clear advantages in both running speed and generalization ability.

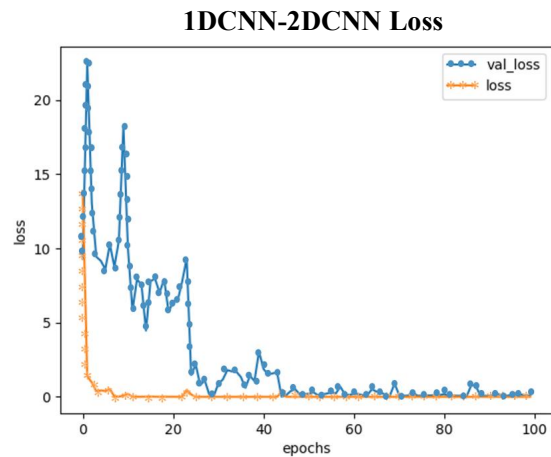


Figure 4.1: Model 1DCNN-2DCNN Training and Validation Loss

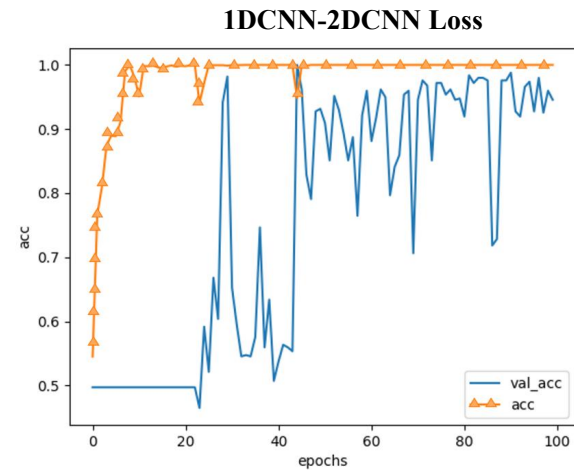


Figure 4.2: Model 1DCNN-2DCNN Training and Validation Accuracy

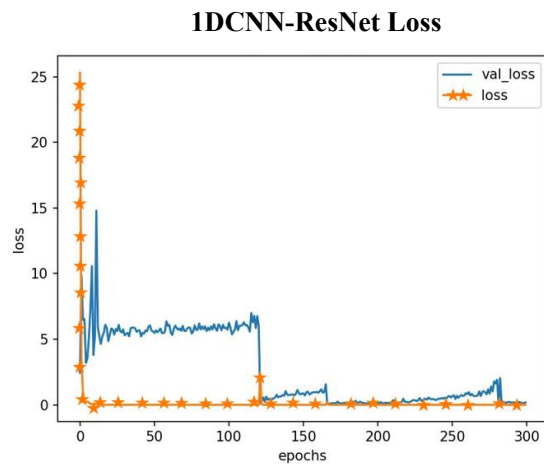


Figure 4.3: Model 1DCNN-ResNet Training and Validation Loss

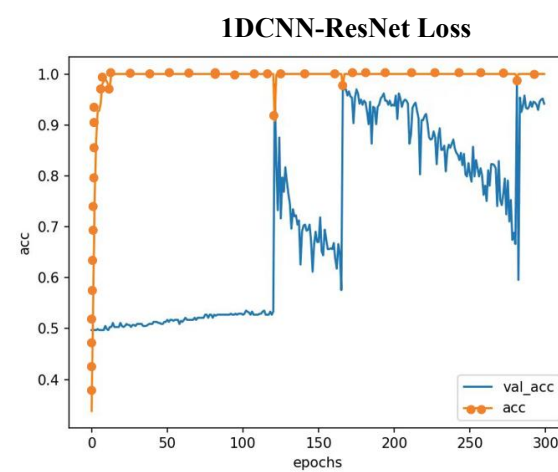


Figure 4.4: Model 1DCNN-ResNet Training and Validation Accuracy

Table 4.3: Comparison between TCN-ResNet model and 1DCNN-2DCNN model

Model	Clothing Style	Precision	Recall	Model (F1 score)	Support	Macro avg	Weighted avg	Confusion matrix	Model Accuracy	Prediction data	Speed
TCN-ResNet	2	1.0000	0.8901	0.8901	247	0.9010	0.8780	0.8756	0.8773	497	4.952s
	5	1.0000	0.7560	0.8610	250	0.9016	0.8773	0.8755			
1DCNN-2DCNN	2	0.9647	0.9960	0.9801	247	0.9803	0.9800	0.9801	0.9799	497	4.90s
	5	0.9959	0.9640	0.9797	250	0.9804	0.9799	0.9797			

4.5 Cross Validation

In this thesis, duplicate 10-fold cross-validation was used, with each fold containing ten subsets and repeated twice. The model is trained and validated on current fold data. The training results for each fold are then saved, and a fold result chart is printed, as shown in Figure 4.5. This approach allows for a more robust assessment of the model's performance by mitigating the effects of random variability and providing a more reliable estimate of its generalization ability. The randomness of data segmentation, the randomness of model initialization, and the imbalance of data samples should all be taken into account when improving model performance. These factors help to train the model with different initial parameters in each repetition, resulting in different feature representations and reducing the risk of overfitting.

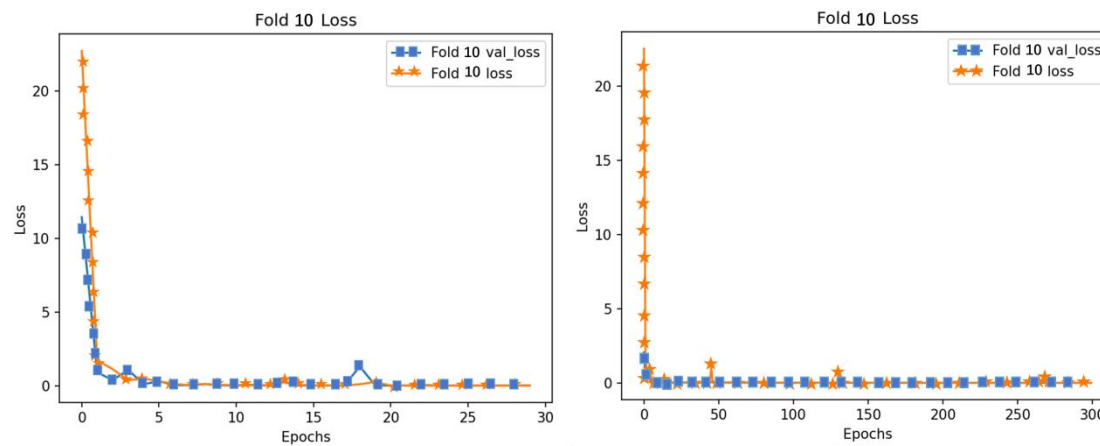


Figure 4.5: Cross-Validation of 1DCNN-2DCNN Model at 30 and 300 Epochs

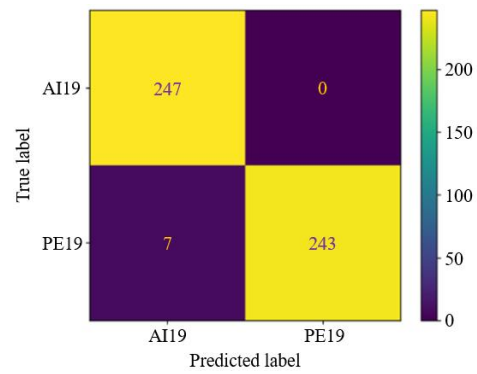


Figure 4.6: Model Improvement Effect of 1DCNN-2DCNN Architecture

As shown in the results in Table 4.4, the 1DCNN-2DCNN model performs well on both the training set and the test set, and there is no underfitting. Therefore, the cross-validation method was used to divide the data set several times and repeated validation was adopted to evaluate the generalization performance of the model more accurately. During cross-validation, the model parameters were adjusted and the network structure optimized, increasing the accuracy from 97.99% to 98.59%, a 0.60% improvement. The prediction results are shown in Figure 4.6.

Table 4.4: Changes after adding cross-validation

Model	Clothing Style	Precision	Recall	Model (F1 score)	Support	Macro avg	Weighted avg	Confusion matrix	Model Accuracy	Prediction data
1DCNN-	2	0.9724	1	0.9860	247	0.9862	0.9863	0.8755	0.9859	497
2DCNN	5	1	0.9720	0.9858	250	0.9860	0.9859	0.9797	0.9859	497

4.6 Chapter Summary

This chapter presented the overall results of the study, focusing on the development of four hybrid models—TCN-ResNet, 1DCNN-ResNet, TCN-2DCNN, and 1DCNN-2DCNN—for personalized clothing prediction. The models leverage multi-modal feature fusion by integrating structured and unstructured data, combining temporal and spatial feature extraction techniques. Compared to mainstream models such as TCN and ResNet, the proposed hybrid models demonstrate lower computational complexity and simpler architectures, making them efficient for fashion prediction tasks.

TCN was applied to textual data, capturing temporal dependencies in fashion-related attributes, while ResNet served as the backbone for deep spatial feature extraction from clothing images. The 1DCNN model efficiently processed text data, focusing on local patterns, and the 2DCNN model was employed for extracting critical visual features from images. Each of these components contributed to the comprehensive multi-modal approach.

The hybrid models outperformed traditional methods by balancing accuracy with computational efficiency. Repeated K-fold cross-validation further enhanced the reliability of the evaluation. Future research can build upon these models to explore more efficient and accurate deep learning architectures for fashion prediction, providing valuable insights and guidance for ongoing developments in this field.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This thesis adopts a multi-modal fusion method to integrate different data sources (image features and user behavior data), enhancing clothing classification prediction accuracy and reliability. The innovation lies in clothing classification prediction for individual users, considering their needs and preferences, which is closer to users' actual requirements than traditional unified models. This thesis draws several vital conclusions based on the experimental and model comparative analyses. First, after evaluating the performance of four different models, it was found that the 1DCNN-ResNet and the 1DCNN-2DCNN model showed excellent accuracy, precision, recall, and F1 score. Secondly, while the 1DCNN-ResNet model showed slightly slower speed than the 1DCNN-2DCNN model, its superior accuracy renders it a viable alternative worth considering. The experimental findings indicate the absence of significant overfitting during the model training process, with minimal disparity between verification and training losses, underscoring the model's robust generalization capability. Although methods such as confusion matrices have achieved good results for model evaluation, this thesis introduces cross-validation to further prove the model's accuracy. This evaluation technique enhances the accuracy of the 1DCNN-2DCNN model to 98.59%, demonstrating the model's reliability and practicality. For personalized clothing prediction with multi-modal fusion, future research directions may explore how to adapt large-scale models to vertical fields better to improve further the effectiveness and practicality of the model in personalized clothing prediction.

5.2 Recommendations

This thesis demonstrates that the proposed 1DCNN-2DCNN multi-modal fusion model can achieve personalized clothing prediction with high accuracy, and the verification results also indicate strong performance. Based on experimental and model analysis, future research can be directed towards the following aspects:

First, future research can deeply explore the performance of different models to better understand the advantages and disadvantages of various approaches in clothing classification prediction, thereby enhancing the accuracy and reliability of predictions. Second, it is important to continue exploring cross-validation techniques; the introduction of repeated cross-validation has proven beneficial in assessing model accuracy. Future research could compare different cross-validation methods to evaluate model performance across various datasets, ensuring consistent and reliable results in diverse scenarios.

Additionally, the proposed 1DCNN-2DCNN model is currently limited to a specific dataset and does not incorporate dataset expansion or transfer learning techniques, which limits its flexibility in adapting to new or broader datasets. This restricts the model's generalizability and applicability to other contexts or larger-scale applications. To address these limitations, future research should consider incorporating transfer learning to enhance the model's ability to generalize across diverse datasets and contexts. Expanding the training to include broader datasets and employing data augmentation techniques could further improve the model's adaptability and performance across various applications, thus overcoming the current constraint of being confined to a specific dataset.

Finally, extending the model to vertical domains is crucial. As personalized clothing prediction with multi-modal fusion gains increasing attention, adapting large-scale

models to specific vertical fields becomes essential. Therefore, future research should explore new methods to enhance the effectiveness and practicality of personalized clothing prediction models within specific vertical domains.

REFERENCES

- Altaf, Y., Wahid, A., & Kirmani, M. M. (2023). Deep Learning Approach for Sign Language Recognition Using DenseNet201 with Transfer Learning. In *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, (pp. 1-6).
- Amin, M. S., Wang, C., & Jabeen, S. (2023). Fashion Sub-categories and Attributes Prediction Model Using Deep Learning. *The Visual Computer*, 39(9), 3851-3864.
- Beljadid, A., Tannouche, A., & Balouki, A. (2020). Application of Deep Learning for the Detection of Default in Fabric Texture. In *2020 IEEE 6th International Conference on Optimization and Applications (ICOA)*, (pp. 1-5).
- Bellemare, J. (2018). Fashion Apparel Industry 4.0 and Smart Mass Customization Approach for Clothing Product Design. In *Customization 4.0: Proceedings of the 9th World Mass Customization & Personalization Conference (MCPC 2017), Aachen, Germany, November 20th-21st*, (pp. 619-633).
- Bhatt, D., Patel, C.I., Talsania, H.N., Patel, J., Vaghela, R., Pandya, S., Modi, K.J., & Ghayvat, H. (2021). CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electronics*, 10(20), 2470.
- Cheng, C., & Parhi, K. K. (2020). Fast 2D Convolution Algorithms for Convolutional Neural Networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(5), 1678-1691.
- Chen, H., Zhang, A., Gong, C., Liang, W., & Wang, Z. (2022). Fault Diagnosis Method for Photovoltaic Panels Based on Improved Shufflenet V2 and Infrared Images. In

- 2022 7th International Conference on Power and Renewable Energy (ICPRE), (pp. 447-451).
- Chen, Y., & Meng, Z. (2019). FashionAI Clothing Attribute Recognition Based on Object Detection Algorithm. *Computer Systems & Applications*, 28(8), 170-175.
- Chen, Y., Zhou, Z., Lin, G., Chen, X., & Su, Z. (2022). Personalized Outfit Compatibility Prediction based on Regional Attention. In *2022 9th International Conference on Digital Home (ICDH)*, (pp. 75-80).
- Das, S. (2019). CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and More. Analytics Vidhya.
- Danopoulos, K., Anagnostopoulos, C., Kachris, D., & Soudris, D. (2021). FPGA Acceleration of Generative Adversarial Networks for Image Reconstruction. In *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCASST)*, (pp. 1-5).
- Dunlap, K. (1928). The Development and Function of Clothing. *The Journal of General Psychology*, 1(1), 64-78.
- Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S. S. K., Dwivedi, S., & Raykar, V. (2020). Attention Based Multi-modal New Product Sales Time-series Forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 3110-3118).
- Fan, J., Lee, J. H., & Lee, Y. (2021). Application of Transfer Learning for Image Classification on Dataset with Not Mutually Exclusive Classes. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, (pp. 1-4).

- French, A. M., & Storey, V. C. (2021). The Journal of Organizational Computing and Electronic Commerce: An Analysis of its Impact on the Information Systems Field in the 21st Century. *Journal of Organizational Computing and Electronic Commerce*, 31(2), 151-169.
- Han, M., Ding, H., Li, L., & Huang, Z. (2023) “Hybrid Learning Method with Class-based Color Augmentation for Clothing Parsing,” in *2023 8th International Conference on Information Systems Engineering (ICISE)*, (pp. 600–603).
- Han, X. (2023). Research on Clothing Personalized Recommendation Algorithm Based on Improved Collaborative Filtering Algorithm. In *3rd International Conference on Internet of Things and Smart City (IoTSC 2023)*, (pp. 92-98).
- He, R., & McAuley, J. (2016). VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).
- Helwan, A. (2021). Deep Learning for Fashion Classification. CodeProject. Retrieved October 1, 2023, from <https://www.codeproject.com/Articles/5297322/Preparing-Data-for-AI-Fashion-Classification?display=Print>
- He, Z., Li, Y., Shi, X., Li, P., & Huang, W. (2020). Multi-Deep Features Fusion Algorithm for Clothing Image Recognition. *Proc. Int. Conf. Digit. Home, ICDH*, (pp. 104–109).
- Hiraoka, T., Kamiya, T., & Aoki, T. (2023). Automatic Segmentation of Finger Bone Regions from CR Image Using the Improved HRNet. In *2023 23rd International Conference on Control, Automation and Systems (ICCAS)*, (pp. 1498-1501).
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504-507.

- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 7132-7141).
- Huang, J. (2024). Manifold Embedded Global and Local Discriminative Features Selection for Single-shot Multi-categories Clothing Recognition and Retrieval. *International Journal of Intelligent Computing and Cybernetics*, 17(2), 363-394.
- Hu, Y., & Jiang, G. (2023). Swin BN A Swin Transformer-based Model for Knitted Fabric Defect Detection. *Journal of Silk*, 60(1), 59–69.
- Johnson, R., & Zhang, T. (2017). Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, (pp. 562-570).
- Kim, D., Saito, K., Mishra, S., Sclaroff, S., Saenko, K., & Plummer, B. A. (2021). Self-supervised Visual Attribute Learning for Fashion Compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 1057-1066).
- Krueangsai, A., & Supratid, S. (2022). Effects of Shortcut-level Amount in Lightweight ResNet of ResNet on Object Recognition with Distinct Number of Categories. In *2022 International Electrical Engineering Congress (IEECON)*, (pp. 1-4).
- Kulkarni, U., Gurlahosur, S. V., Babar, P., Muttagi, S. I., Soumya, N., Jadekar, P. A., & Meena, S. M. (2023). Facial Key points Detection using MobileNetV2 Architecture. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, (pp. 1-6).
- Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal Convolutional Networks: A Unified Approach to Action Segmentation. *ECCV Workshops*, (pp. 47-54).

- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Li, L., Wang, S., Huang, Z., Han, M., Ding, H., & Shi, X. (2023). Clothing Classification Based on Channel Attention and Token Labeling. In *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, (pp. 961-964).
- Li, M., Chen, H., & Cheng, Z. (2022). A Lightweight End-to-end Network for Wearing Mask Recognition on Low-resolution Images. In *2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, (pp. 38-44).
- Liu, L., Zhang, H., Li, Q., Ma, J., & Zhang, Z. (2023). Collocated Clothing Synthesis with GANs Aided by Textual Information: A Multi-Modal Framework. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(1), 1-25.
- Li, J., Shi, W., & Yang, D. (2021). Clothing Image Classification with a Dragonfly Algorithm Optimised Online Sequential Extreme Learning Machine. *Fibres and Textiles in Eastern Europe*, 29(3), 91–96.
- Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1096–1104).
- Li, L., Wang, S., Huang, Z., Han, M., Ding, H., & Shi, X. (2023). Clothing Classification Based on Channel Attention and Token Labeling. *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, (pp. 961–964).
- Lu, S., Zhu, X., Wu, Y., Wan, X., & Gao, F. (2021). Outfit Compatibility Prediction with Multi-layered Feature Fusion Network. *Pattern Recognition Letters*, 147, 150-156.

- Meyerson, E., & Miikkulainen, R. (2017). Beyond Shared Hierarchies: Deep Multitask Learning Through Soft Layer Ordering. *arXiv preprint arXiv:1711.00108*.
- Mohanty, D. K., Gupta, P. D., Dey, R., & Pattnaik, S. (2022). Modified Convolutional Neural Network for Fashion Classification. In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, (pp. 1-9).
- Morris, R. G. M. (1989). Parallel Distributed Processing: Implications for Neurobiology. *Trends in Neurosciences*, 12(2), 51-54.
- Mo, D., Lai, Z., Wang, X., & Wong, W. (2020). Jointly Sparse Locality Regression for Image Feature Extraction. *IEEE Transactions on Multimedia*, 22(11), 2873–2888.
- National Bureau of Statistics of the People's Republic of China. (2022). Statistical Communique on National Economic and Social Development of the People's Republic of China in 2022.
- National Bureau of Statistics of the People's Republic of China. (2023). Statistical Communique on National Economic and Social Development of the People's Republic of China in 2023.
- Ren, S., Sun, J., He, K., & Zhang, X. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Vol. 3, pp. 770-778).
- Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations By Back-propagating Errors. *Nature*, 323(6088), 533-536.
- Sahithi, B., & Vigneshwari, S. (2023). Exploring the Performance of VGG16 and Efficient Net Models for Plant Disease Classification: A Comparative Approach. In *2023*

First International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV), (pp. 1-6).

Septiarini, A., Saputra, R., Tejawati, A., Wati, M., Hamdani, H., & Puspitasari, N. (2021, December). Analysis of Color and Texture Features for Samarinda Sarong Classification. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, (pp. 102-107).

Shocker, A. D., Srivastava, R. K., & Ruekert, R. W. (1994). Challenges and Opportunities Facing Brand Management: An Introduction to the Special Issue. *Journal of Marketing Research*, 31(2), 149-158.

Shin, S. Y., Jo, G., & Wang, G. (2023). A Novel Method for Fashion Clothing Image Classification Based on Deep Learning. *Journal of Information and Communication Technology*, 22(1), 127–148.

Shi, J. W., Song, X. M., Liu, Z. X., & Nie, L. Q. (2021). Fashion Graph-enhanced Personalized Complementary Clothing Recommendation. *Journal of Cyber Security*, 6(5), 181-198.

Sharma, R., Gangrade, J., Gangrade, S., Mishra, A., Kumar, G., & Gunjan, V. K. (2023). Modified EfficientNetB3 Deep Learning Model to Classify Cooler Fundus Images of Eye Diseases. In *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, (pp. 632-638).

Singh, P., Verma, V. K., Rai, P., & Namboodiri, V. P. (2019). Hetconv: Heterogeneous Kernel-based Convolutions for Deep Cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 4835-4844).

Sivaranjani, L., Rachamadugu, S. K., Reddy, B. S., Sakthivel, M., & Depuru, S. (2023, September). Fashion Recommendation System using Machine Learning. In *2023*

- 4th International Conference on Smart Electronics and Communication (ICOSEC)*, (pp. 1367-1374).
- Skenderi, G., Joppi, C., Denitto, M., Scarpa, B., & Cristani, M. (2022). The Multi-Modal Universe of Fast-Fashion: The Visuelle 2.0 Benchmark. Retrieved from <https://arxiv.org/pdf/2204.06972.pdf>.
- Su, X., Duan, J., Ren, J., Feng, Q., Zhang, H., & Chen, X. (2024). Personalized Clothing Recommendation Fusing the 4-season Color System and Users' Biological Characteristics. *Multimedia Tools and Applications*, 83(11), 12597–12625.
- Surakarin, W., & Chongstitvatana, P. (2015, November). Classification of Clothing with Weighted SURF and Local Binary Patterns. In *2015 International Computer Science and Engineering Conference (ICSEC)*, (pp. 1-4).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1-9).
- Truong, T., & Yanushkevich, S. (2021, January). Relatable Clothing: Detecting visual Relationships Between People and Clothing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, (pp. 4070-4076).
- Oliveira, W.D., Penatti, O.A., & Berton, L. (2020). A Comparison of Graph-based Semi-supervised Learning for Data Augmentation. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, (pp. 264-271).
- Wang, W., & Chen, L. (2024). Design and Implementation of Aerospace Arisa Fashion Design CAD System Based on Deep Convolution Neural Network. *Computer-Aided Design and Applications*, 21(S1), 132–145.

- Wang, H., & Rasheed, K. (2014). Artificial Intelligence in Clothing Fashion. *In Proceedings on the International Conference on Artificial Intelligence (ICAI)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
- Wang, N., He, Z., Huang, Z., Zhou, Y., & Wu, X. (2019). Research on Improved Collaborative Filtering Algorithm in Personalized Clothing Recommendation. *Journal of Hunan Institute of Engineering (Natural Science Edition)*, 29(1), 33–36.
- Wang, Q., Liu, Z., Wang, M., & Zhao, Y. (2023). Pose-Driven Local Feature Alignment for Pedestrian Re-Identification. *Computer Systems & Applications*, 32(4), 268-273. (In Chinese)
- Wang, Z. W., Pu, Y. Y., Wang, X., Zhao, Z. P., Xu, D., & Qian, W. H. (2020). Accurate Retrieval of Multi-scale Clothing Images Based on Multi-feature Fusion. *Jisuanji Xuebao/Chinese Journal of Computers*, 43(4), 740–754.
- Wang, Y., Liu, L., Fu, X., & Liu, L. (2024). MCCP: Multi-modal Fashion Compatibility and Conditional Preference Model for Personalized Clothing Recommendation. *Multimedia Tools and Applications*, 83(4), 9621-9645.
- Wang, Y. (2021). Clothing Attribute Recognition with Semi-supervised Learning. In *2021 IEEE International Conference on Electronic Technology, Communication and Information (ICETCI)*, (pp. 507-511).
- Wu, M., Zhang, G., & Jin, C. (2022). Time Series Prediction Model Based on Multimodal Information Fusion. *Journal of Computer Applications*, 42(8), 2326-2332.
- Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10), 12113–12132.

- Xu, X., & Liu, X. (2020). Recognition Technology of Clothing Attribute Label Based on Deep Learning. *Computer Science and Application.*, 10(4), 619-628.
- Yamamoto, R., Song, E., & Kim, J. M. (2020). Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6199–6203).
- Yang, B., Xu, D., Zhang, H., & Luo, H. (2021). Minority Clothing Recognition Based on Improved DenseNet-BC. *Journal of Zhejiang University, Science Edition*, 48(6), 676-683.
- Yang, X., Yuan, S., & Tian, Y. (2014). Assistive Clothing Pattern Recognition for Visually Impaired People. *IEEE Transactions on Human-Machine Systems*, 44(2), 234–243.
- Yang, H., Zhang, D., Hu, A., Liu, C., Cui, T.J., & Miao, J.H. (2022). Transformer-Based Anchor-Free Detection of Concealed Objects in Passive Millimeter Wave Images. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-16.
- Yang, B. (2022). Clothing Design Style Recommendation using Decision Tree Algorithm Combined with Deep Learning. *Computational Intelligence and Neuroscience*, 2022(1), 5745457.
- Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., et al. (2022). Unified Contrastive Learning in Image-Text-Label Space. Retrieved from <https://arxiv.org/pdf/2204.03610.pdf>.
- Yoshino, K., Nakashima, K., Ahn, J., Iwashita, Y., & Kurazume, R. (2022) “Gait Recognition Using Identity-Aware Adversarial Data Augmentation,” in *2022 IEEE/SICE International Symposium on System Integration (SII)*, (pp. 596–601).

- Yu, F., Li, H., Shi, Y., Tang, G., Chen, Z., & Jiang, M. (2023). FFENet: Frequency-spatial Feature Enhancement Network for Clothing Classification. *PeerJ Computer Science*, 9, e1555.
- Yu, F., Du, C., Hua, A., Jiang, M., Wei, X., Peng, T., & Hu, X. (2021). EnCaps: Clothing Image Classification Based on Enhanced Capsule Network. *Applied Sciences*, 11(22), 11024.
- Zeng, W., Zhao, M., Gao, Y., & Zhang, Z. (2020). TileGAN: Category-oriented Attention-based High-quality Tiled Clothes Generation from Dressed Person. *Neural Computing and Applications*, 32(23), 17587–17600.
- Zhang, X., Shen, M., Li, X., & Wang, X. (2023). AABLSTM: A novel Multi-task Based CNN-RNN Deep Model for Fashion Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1), 1-18.
- Zhao, H., & Wang, Z. (2023). Traditional Costume Image Classification Algorithm Based on Multi-layer Discriminative Dictionary Learning. *Journal of Beijing University of Posts and Telecommunications*, 46(2), 104. (In Chinese)
- Zhang, H., Huang, W., Liu, L., & Chow, T. W. (2019). Learning to Match Clothing from Textual Feature-based Compatible Relationships. *IEEE Transactions on Industrial Informatics*, 16(11), 6750-6759.
- Zhang, M., He, R., Zhan, W., Li, M., & Chen, J. (2019). Clothing Label Attribute Recognition Based on Improved Residual Neural Network. *Computer and Digital Engineering*, 47(4), 933–939.
- Zhang, H., Zong, H., & Qiu, J. (2022). Active and Passive Mixed Millimeter Wave Imaging Target Recognition Method Based on Multi-feature Fusion. *International Journal of RF and Microwave Computer-Aided Engineering*, 32(12), e23524.

APPENDICES

Appendix A: Journal Publications

1. **Rong Liu**, Annie Anak Joseph, Miaomiao Xin, Hongyan Zang, Wanzhen Wang, & Shengqun Zhang. (2024). Personalized Clothing Prediction Algorithm Based on Multi-modal Feature Fusion. *International Journal of Engineering and Technology Innovation*, 14(2), 216–230.
2. **Rong Liu**, Annie Anak Joseph, & Wanzhen Wang. (2024). Preliminary Study of E-commerce Sale Prediction Based on Machine Learning Methods. *International Engineering Conference, UNIMAS*.

Appendix B: Image data

› VISUELLE › images › AI17



AI17 Image data for clothing

› VISUELLE › images › AI18



AI18 Image data for clothing

» VISUELLE » images » AI19



AI19 Image data for clothing

» VISUELLE » images » PE17



PE17 Image data for clothing

› VISUELLE › images › PE18



PE18 Image data for clothing

› VISUELLE › images › PE19



PE19 Image data for clothing