Journal of Informatics and Web Engineering
Vol. 5 No. 2 (June 2026) eISSN: 2821-370X

Design of a CNN–NLP Based Visual Chatbot for e-Commerce Systems

Sidra Tul Kubra[1*], Salma Jahan Nisha[2], Nabhan Salih[3], Amalina Ibrahim[4], Wan-Noorshahida Mohd-Isa[5]**

[1,2,4,5]Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
[3]Department of Computer Science and Engineering, Ohio State University, Dreese Laboratories, 395, 2015 Neil Ave, Columbus, OH 43210, United States
[4,5]Centre for Image and Vision Computing, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
**corresponding author: (sidra.tul.kubra@student.mmu.edu.my; ORCiD: 0009-0000-7893-6018)*
***corresponding author: (salih.43@osu.edu; ORCiD: 0000-0001-7906-4071)*

Abstract - E-commerce platforms have limitations to assist interactive and highly personalized experiences using conventional text or voice-based chatbots, particularly in context-aware areas where product discovery and visual identification are necessary. Hence, we propose to integrate an image as a visual-based chatbot. We present a conceptual framework that combines Convolutional Neural Networks (CNN) for image recognition with Natural Language Processing (NLP) for dynamic, context-aware search in a chatbot. The CNN model enables the recognition of products from user-uploaded images, while the NLP component processes and generates appropriate responses to enhance the shopping experience by providing suggestions. This hybrid system will be developed using the Python programming language, which uses libraries such as Keras, OpenCV, and Flask. By allowing users to interact with a chatbot that uses visual inputs, this system aims to create a more intuitive, personalized e-commerce experience that could lead to high engagement and customer satisfaction. The evaluation metrics are measured in terms of the usage of this system by three users in real-world e-commerce applications. This small-scale test may offer insights into how image-based communication can revolutionize the online shopping experience by analysing usability, interaction efficiency, user satisfaction, engagement behaviour, and system responsiveness in practical scenarios during testing.

Keywords—Deep Learning, Artificial Intelligence, Object Detection, Natural Language Processing, Convolutional Neural Network

Received: 23 June 2025; Accepted: 19 November 2025; Published: 16 June 2026

1. INTRODUCTION

In our current technological world, image recognition has gained major popularity in various areas. As per various researchers, it is a type of recognition method that uses Artificial Intelligence (AI) with various deep learning models, and augmented reality to provide virtual input by users, visual search options, and other applications to upgrade online

purchasing [1]. Despite these advancements, our current chatbot technologies mainly depend on text format inputs, limiting their capability to process and respond to visual user inputs such as images or videos.

This reduces the chatbot's ability to provide accurate and context-aware product suggestions based on images, which is considered a critical feature in many e-commerce scenarios based according to studies found [1]. This limitation is especially noticeable in e-commerce, where user experience works as an important driving factor for areas such as customer sales and customer satisfaction [2].

It is important to interact with customers in real-time, answering their input queries and addressing their questions is significant. Chatbots are increasingly being used on a variety of platforms to connect with customers, particularly in the e-commerce space [3].

Current chatbot systems' limits include their dependency on text-based methods, which decreases their capability to talk with customers in visual areas, such as e-commerce. E-commerce platforms use image-based product descriptions and rely heavily on image-based material for product searches and decision-making from a user perspective, especially in industries like electronics, fashion, and home decor. Although image recognition technology has made significant strides based on research, however, chatbot systems for real-time product suggestions still do not fully use it for practical purposes. [1], [3]. This gap that we see in e-commerce between the image and text searches creates a gap within the chatbots' ability to provide results with accuracy for products. This happens because they cannot visually understand user input just from text. Thus, it impacts the user experience in terms of searching using a chatbot. This puts the need for a multimodal chatbot framework in the e-commerce area that can handle both text and image inputs as part of a conversation with the users and provide accurate results. This process can be achieved using deep learning methods to create a chatbot that uses Natural Language Processing (NLP) for user queries and Convolutional Neural Networks (CNNs) for image analysis. This way, the accuracy and applicability of finding products in e-commerce engines will be highly improved by the development of such a framework, which provides a more context-aware response, thus increasing user satisfaction in product buying, and eventually boosting the e-commerce experience. In order to enhance natural interactions, researchers are increasingly looking into multimodal chatbots, which combine text, images, and even speech or gestures. Multiple input modes can greatly improve contextual understanding and user experience, as demonstrated by the applications of such systems in fields such as assistive technology, education, and healthcare [4], [5]. Specifically, chatbots with sign language capabilities show how NLP can be combined with nonverbal communication modalities to create inclusive interactions [6]. These developments show the feasibility and advantages of multimodal approaches, and their extension to e-commerce platforms is a logical and significant step forward.

This study was done with the goal of addressing the increasing demand for more efficient and user-friendly customer experiences on e-commerce platforms during shopping or other activities. Regular chatbots, the majority of which use text input conversations, have a lot of disadvantages, especially in times like now, where AI is dominating with various methods such as image recognition, video recognition, etc. In areas like electronics, fashion, and cosmetics, image inputs play an important role in finding products and making decisions for better accuracy and experience. Despite the significant advancements in image recognition technologies, there are still limitations in the development of these technologies into chatbot systems for real-time product recommendations.

This paper intends to contribute in many important ways to the research field. The first to take note of is the development of the multimodal chatbot framework that combines NLP for text-based understanding and also CNNs for image recognition. This method of using NLP and CNN together enables chatbots to process both text and visual inputs at the same time based on users' input queries on products. There are limitations that we have highlighted from our research on regular text-based chatbots that can be overcome by this multimodal approach that we have taken for this paper. In this paper, we have also highlighted how it provides context-aware responses, basically based on the context of the user, and provides the matched information, in real-time, on product recommendations. This is a recent area of development in the field of AI, and researchers are working on it further. Apart from these, we have provided results based on standard metrics calculations of individual accuracy of the multimodal chatbot's performance, tested by users. We discussed the initial prototype development process; technologies used and determined its efficiency with a small dataset using feedback from real-time user interactions and product data. The study finally shows the possibility of a robust image and text-based chatbot to improve e-commerce platforms by increasing recommendation accuracy and positive user feedback, thus changing the user experience for the better. Lastly, the framework's practical ability to integrate with current e-commerce platforms is also a factor to contribute. This framework provides a solution that can help to increase user engagement, improve product discovery using images, and improve the overall shopping experience by adjusting to both visual and contextual data.

2. LITERATURE REVIEW

Image recognition using NLP integration for text-based conversation in the e-commerce domain has resulted in great advancements in recent times to modify existing user experience in E-commerce and other areas for better results using chatbots. Early work on chatbots has been performed based on research. One such example comes from the first chatbot called ELIZA [7], which laid the foundation for conversational agents from a long time ago, focusing on text-based communication only initially. Throughout the years, systems like ALICE and CleverBot were introduced. They used AIML to improve conversational responses for better clarity [8]. These systems, over time, evolved towards more dynamic approaches for more versatile interactions. This was achieved with the use of Latent Semantic Analysis (LSA), enabling better understanding of user inputs [8]. Nowadays, in e-commerce, intelligent assistants have become crucial components for improving shopping and service experiences by offering real-time product recommendations. Systems, for example, Inktomi and mySimon developed price comparison features, thus laying the groundwork for modern recommendation engines [9], [10]. More recent advancements in personal assistants provided by big giants include chatbots like Siri and Google's Alexa. They have incorporated NLP and speech recognition techniques for chatbot agents, ensuring context-aware interactions with users [11].

Since the beginning of the first chatbot development, despite these advancements, the integration of image recognition features within e-commerce chatbot assistants remains quite underdeveloped. Early systems like Google Goggles and SnapTell displayed the possibility for visual search options. However, they were still limited in scope, especially in categories such as apparel and electronics [12], [13]. Another research suggests that CamFind and Slyce added product category recognition and price comparison features yet still fell behind in broader terms of e-commerce applications, particularly in the cases of personalized recommendations [14], [15].

Recent research shows methods to develop image recognition recommendation systems to enhance personalized shopping experiences within the e-commerce domain. A CNN-based model was proposed by [16] for product recommendations to identify the image and find relevant products, achieving accuracy between 95.2% and 96.5% and a recall of 85% to 91%. This model used product images from categories such as clothing and home decor. Another research by [17] developed a virtual shopping assistant that integrated image recognition with APIs like CloudSight. This achieved a recognition accuracy of 97.95%. To enhance user experience, a system development was made by [18] which combines CNN for image recognition for e-commerce, achieving a 93% accuracy for image classification. A reverse image search feature was developed by [19] within a recommendation system, effectively providing personalized suggestions. However, in terms of scalability and robustness, the system did not show promising results.

Kuo et al. [20] used Mask R-CNN for retail product detection, which resulted in achieving a mean average precision of 98.92%. A fashion image retrieval system was proposed by [21] to give feedback using NLP. This system achieved a high dialog retrieval accuracy of 93.5% for dresses, 92.92% for shirts. Another fashion recommendation feature was developed by [22] using the YOLOv5 model for object detection, which resulted in an accuracy of 96%. A personalized product recommendation system developed by combining AI and image processing achieved an accuracy of 99.45% by [23].

In addition, Generative AI models such as Variational Autoencoders (VAEs) have demonstrated success in personalized recommendations based on research we have found. Zheng et al. [24] developed a meal recommender system using VAEs, which achieved a high NDCG score of 0.963 based on metrics. However, the system's RMSE (47.77) and MAE (36.28) indicated that further improvements could be made in terms of accuracy, which can be extended to the e-commerce domain as well to enhance product recommendations. Another study introduced a hybrid recommender system for academic peer review, which combines content-based and collaborative filtering approaches to achieve it. Their system showed high results as mentioned, without specifics on NDCG and precision metrics. This particular paper suggested that hybrid approaches could improve e-commerce recommendation systems for a better e-commerce experience, especially when combining image recognition and NLP [25].

In the healthcare domain, [26] proposed an image-based recommendation system for health-related products recognition, achieving a similarity score of 0.9932 using the ResNet50 framework. A proposed framework integrates CNN for skin classification to recommend skincare products, demonstrating efficient skin classification with the help of facial image recognition by [27].

The development of various types of image recognition followed by recommendation systems has seen significant progress in recent years; however, challenges still remain in achieving higher accuracy for these systems to apply to a broader range of products and also to ensure robustness and scalability in terms of features. As technology

progresses, future researchers should focus on overcoming these challenges, particularly by focusing on and working with diverse datasets and experimenting with cross-platform compatibility.

In our study, we are proposing a way to create a framework that combines image recognition and recommendation algorithms to deliver a more personalized e-commerce experience. This study also aims to address existing gaps in research by providing dynamic product recommendations based on user-uploaded images. This approach is intended to provide better user engagement and contribute to the growing field of intelligent chat assistants and image-based search in e-commerce.

The objective of this paper is to design a framework for a chatbot with image detection for the e-commerce domain. Overall, the literature review suggests that e-commerce is a demanding field, and an increasing demand for good shopping experiences is necessary for further improvement of digital technology, promoting economic growth and creativity in the digital economy. At this current stage of the research, an object detection system has been conceptualized with the use of OpenCV, CNN, and NLP. At the end of the paper, we present an early result of this feature that is integrated within a chatbot component. Table 1 gives an overview of related studies in this area, including their methods, datasets, and main findings.

Table 1. Summary of Related Literature Review

Author	Purpose	Methodology	Dataset	Accuracy	Limitations
Zhang et al., 2024 [16]	Improve image recognition and product recommendation in e-commerce.	A CNN-based model integrated with product recommendation algorithms, evaluated using accuracy and recall.	E-commerce product images (clothing, electronics, home decor, etc.).	95.2%-96.5%, Recall: 85%-91%	Limited dataset diversity and performance variability under different conditions.
Dahal et al. (2024) [18]	To enhance user experience on an e-commerce site by incorporating both image processing and voice recognition.	CNN for image classification, ResNet-18 for facial detection, using Ibug 300W and Fashion MNIST datasets; backend with NodeJS, frontend with React.	ibug 300w large face landmark dataset and Fashion MNIST dataset. net	Image: 93%, Landmark loss: 0.0013.	Integration challenges, dataset limitations.
Badave et al. (2022). [19]	Provide personalized product recommendations using reverse image search, a recommendation system, and a chatbot interface.	Applied collaborative and content-based filtering; used CNN for reverse image search feature extraction; developed a system with Flask backend and chatbot frontend.	Myntra dataset (15,000 rows, 24 columns, including product details).	Not explicitly stated; highlights effective personalized recommendations.	Dataset size may limit scaling; challenges expanding to larger, more diverse datasets.
Wu et al. (2021) [21]	Develop an interactive fashion image retrieval system using natural language feedback.	Implemented a transformer-based model for dialog-based retrieval.	Fashion IQ Dataset (Dresses: 19,087 images, Shirts: 31,728 images, Tops & Tees: 26,869 images	Recall@10: Dresses – 93.50%, Shirts – 92.92%, Tops & Tees – 93.25%.	Limited generalization; high cost of human-annotated captions.
Jain et al. (2024) [22]	Develop a fashion recommendation system using image-based detection and classification.	Used YOLOv5 for object detection, CNN for pattern classification, and web scraping for product details; data stored in MongoDB for filtering.	70,000 images, DeepFashion2 (801,000 images, 13 categories), Dress Pattern (~500 images/class, 10 styles).	YOLOv5 model: 96%.	Fine-tuning needed for pattern classification; optimization required for web scraping speed; real-time recommendations.
Li et al. (2024) [23]	Develop a personalized product recommendation system combining AI and image processing.	Graph Convolutional Networks (GCNs) for multi-label classification and a similarity-based recommendation algorithm with image and behaviour data	Fashion MNIST (60% training, 20% validation, 20% testing); self-built dataset (70% training, 15% validation, 15% testing).	Multi-label classification: mAP 83.5, OF1 80.9, CF1 78.5; Recommendation accuracy: 99.45%.	High computational overhead; challenges with complex product categories and data-scarce scenarios.
Zheng, B et al (2025). [24]	Develop a meal recommender system using VAEs for personalized meal suggestions.	VAEs	Meal planning dataset, nutritional data	NDCG score of 0.963, RMSE (47.77), MAE (36.28), MSE (2282.32)	Accuracy can be improved, particularly in nutritional predictions.

Lim, Y et al.(2025) [25]	Improve reviewer assignment in academic peer review using hybrid recommender models.	Hybrid recommender models (content-based & collaborative filtering)	Peer review datasets, user interactions	Strong NDCG and precision metrics	Scalability issues in real-world applications.
Natadirja et al. (2023) [26]	Develop an image-based recommendation system for Kalcare.com.	Eight CNN models (ResNet50, VGG16, VGG19, MobileNet, MobileNetV2, NasNetMobile, InceptionV3, Xception.	Kalcare.com product image dataset (~5,256 images, resized to 240x240 pixels).	ResNet50 achieved the highest similarity score (0.9932).	Dataset limited to Kalcare.com products; user poll limited to employees.
Hanchinal et al. (2024) [27]	Develop an AI-driven skincare recommendation system integrating skin type analysis and price comparisons.	Used CNN for skin classification; applied collaborative and content-based filtering; web scraping for real-time pricing.	Trained dataset for skin classification; user-uploaded facial images processed via edge detection, grayscale, and median blur.	Training vs validation accuracy graph demonstrates CNN efficiency.	Dataset size and diversity may impact scalability; challenges in integrating multiple e-commerce platforms for price tracking.

3. PROPOSED FRAMEWORK

By combining NLP and image recognition, the suggested visual, context-aware chatbot system for e-commerce is intended to offer real-time product recommendations. The framework for the Image recognition Chatbot System is designed using the Object-Oriented Design (OOD) model by [28]. By using the OOD method, we have used some key performance metrics as variables to determine the final results of this system. This combines other entities such as design, properties, and their relationships. For technologies in development, we used CNN for image recognition and NLP for text processing to understand and answer user questions.

The conceptual framework using the OOD model design is given in steps below.

A (Image Recognition Module) = (E, Properties(E), Rel(E))

- **I**: Represents the Image Detector module
- **E**: Represents the system's smaller entities, which are product categories, images, queries
- **Properties(E)**: Describes the properties related to the above entities, which are visual features, text representations
- **Rel(E)**: Represents the relationships between the entities, for example, how text and image data are integrated to generate recommendations.

The key components of the framework include 3 modules is as follows.

- Image Recognition: This module processes user-uploaded images using CNNs to extract visual characteristics by applying small filters and reading them that let the system recognize products or product categories as a whole picture.
- Text Processing: This refers to the text-based module. User queries are sent to the chatbot to understand using the NLP module and converted into a text format in the form of a response that the system can understand and display in the User Interface (UI).
- Recommendation Engine Logic: This logic combines the results of the text processing and image recognition modules as a whole to create the product suggestions that match closely or fully with the user's questions or images.

3.1. System Design

System design is a fundamental concept of any application. In this part, we created the system's architecture and drew out how the two main modules, NLP and image recognition using CNNs, will interact with each other as well as the frontend UI. We established the modular structure of the system as a diagram below and made sure that every part works well together in the whole picture of the system.

3.1.1 System Architecture

It consists of a modular pipeline that manages both text and image inputs and forms the foundation of the system's overall high-level architecture. While the NLP module detects user needs by processing text queries, the image recognition module uses a CNN to identify products from user-uploaded images. A multimodal embedding process is then used to combine the data from both modules, guaranteeing that the system can process both input types at once and produce precise product recommendations. Figure 1 shows the overall system architecture and how the main components of the chatbot interact with each other.

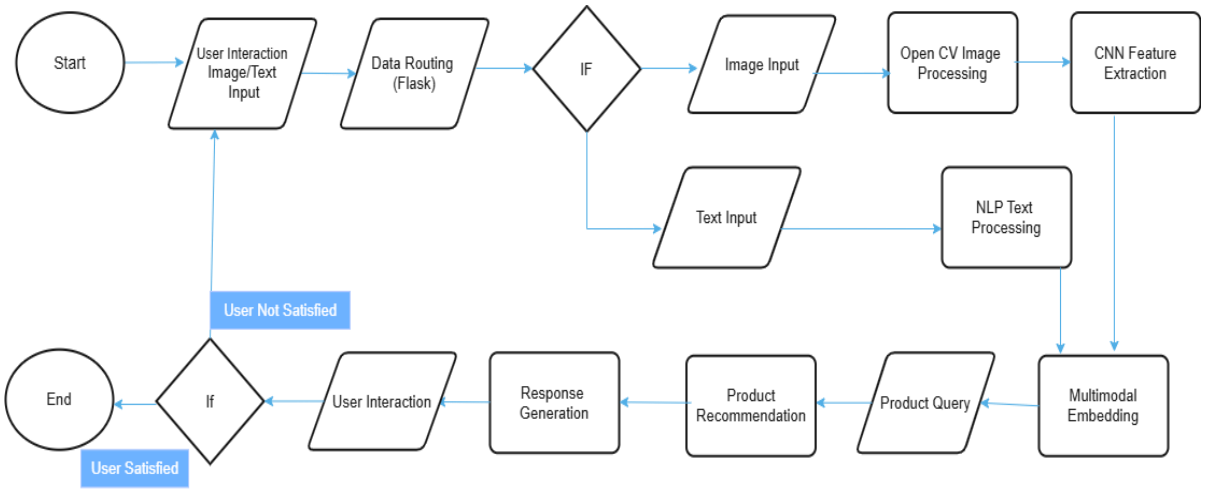


Figure 1. System Architecture Flowchart

The diagram illustrates the architecture of the system, showing the interaction between image recognition and text processing modules to provide real-time product recommendations, including data routing, multimodal fusion, and recommendation output.

The design follows the direction of OOD principles by treating each component as a separate entity with separate properties and relationships. This modular design ensures that the system is easy to maintain, scalable, and flexible for users. One module's replacement or update will not affect the other components.

3.1.2 Core Components and Interaction

- Image Recognition Module
- Text Processing Module
- Multimodal Fusion

3.1.3 Data Flow and Interaction

In this part, we determine how data flow happens from user input to recommendation output. The user inputs whether text or an image will be sent through the chatbot system frontend, after that will be proceed the required modules will proceed and be combined into a single representation when a user interacts with the chatbot.

3.1.4 Tools and Technologies

It is important to choose the correct tools and technologies for the system. While BERT and other Python libraries are used for text processing, OpenCV is used to build and preprocess the inputs using the CNN model for image recognition.

3.1.5 System Integration and Communication Pipeline

The proposed system's integration ensures proper communication between the backend, multimodal layer, text processing module, and image recognition module. The general workflow is given as follows.

1. User Input: Through the chatbot interface, the user can enter an image, a text query, or a combination of both.
2. Routing and Preprocessing: The type of input is determined by the Flask backend. While text queries are sent to the NLP module for processing, images are sent to the CNN-based image recognition module.
3. Feature Extraction:
   - The input image's visual feature vectors are extracted by the image recognition module.
   - The NLP module uses BERT to convert the user's query into contextual embeddings.
4. Multimodal Fusion: A single multimodal feature vector that represents both textual and visual context is created by concatenating the features that were extracted from both modules.
5. Recommendation Generation: The product database is compared with the fused representation. The most matched products are retrieved, ranked, and formatted as well to provide recommendations by the system.
6. Response Delivery: The backend compiles the ranked suggestions into a structured response and sends it back to the chatbot interface for real-time user communication for product search and queries.

This integration method confirms that textual and visual data are considered together when making recommendations by enabling real-time communication between all the different types of modules. By explicitly connecting the recognition, NLP, and backend components, the system creates a workflow for multimodal interaction, overcoming the limitations set by traditional text-only chatbots.

4. RESEARCH METHODOLOGY

For the methodology, a systematic approach has been used in the development of the visual, context-aware chatbot system for e-commerce to make sure that the NLP and image recognition components work together to fulfil the objective of real-time image detection and provide product recommendations. Development and evaluation are the two main stages of the process in this part of the research.

System Development Phase: During this phase, we implemented the system using the Python programming language and related libraries. We have used BERT, which is a pre-trained language model provided by Google for using NLP to understand text context, and the Keras Application Programming Interface (

API) framework for CNN networks. Each module was built by us to accept text and image inputs based on searches and integrate them to generate product recommendations.

Evaluation Phase: This phase evaluates the system's performance based on many key metrics to provide accuracy and results. In this phase, we also include its current ability to give accurate product recommendations on the dataset we have used to train and test it. To determine the system’s efficiency by the results percentage, we used performance metrics like recall, accuracy, precision, and user feedback to calculate the results and percentage.

4.1 System Development

The following six-step procedure describes the system's development and evaluation in order to make the difference between the methodology and the proposed framework above in more clarity. The methodology outlines the actual procedures and process flow that we used to construct, integrate, and test the system, whereas the framework (Figure 1) offers a conceptual overview of the architecture.

1. Dataset Preparation
   For data preparation, we have paired image sequences to match them with the text queries and create a multimodal dataset. Afterward, we trained our recognition model using this newly created dataset and tested it.
2. Model Training
   We have used the CNN architecture to train the image recognition model. In the training process, image

preprocessing techniques, augmentation methods, and iterative optimization were used so that we can achieve good accuracy.

3. NLP Component Setup
   The NLP model mentioned above has been used to process text-based questions and give meaningful replies. In this module, we have also implemented intent classification and embeddings for semantic representation.
4. Multimodal Fusion
   This step is basically the combination of the image module and text module outputs for the interaction of the chatbot with the users and provides related search results.
5. Prototype Implementation
   The minimal frontend implementation was done, with which the backend CNN model for image recognition and the NLP module for text recognition were integrated to complete the system's overall functionality.
6. System Evaluation
   The system evaluation was done using qualitative and quantitative methods to detect the usability accuracy and also performance metrics received from user testing and user feedback.

4.2 Image Recognition Module

User-uploaded images must be processed and analysed by the image recognition module in order to extract visual characteristics that allow for product identification. A pre-trained ResNet-50 model, which has been optimized for the particular product categories pertinent to the e-commerce platform, is used by the system to extract features using CNN.

4.2.1 CNN Layer Architecture

A refined ResNet-50 CNN model is used in the image recognition module. Before entering the network, input images are normalized and resized to 224×224×3 (RGB). For multimodal recommendations, the CNN combines text embeddings with discriminative visual features that are extracted. Table 2 shows the CNN layer structure of the fine-tuned ResNet-50 model used for image recognition.

Table 2. CNN Layer Architecture (ResNet-50 Fine-tuned)

Layer Type	**Parameters / Filters**	**Output Dimension**
Input Layer	224×224 RGB image	224×224×3
Conv Layer 1	64 filters, 7×7 kernel, stride 2	112×112×64
Max Pooling	3×3 kernel, stride 2	56×56×64
Residual Block (×3)	$\text{Conv}(1\times1, 64) \rightarrow \text{Conv}(3\times3, 64) \rightarrow \text{Conv}(1\times1, 256)$	56×56×256
Residual Block (×4)	$\text{Conv}(1\times1, 128) \rightarrow \text{Conv}(3\times3, 128) \rightarrow \text{Conv}(1\times1, 512)$	28×28×512
Residual Block (×6)	$\text{Conv}(1\times1, 256) \rightarrow \text{Conv}(3\times3, 256) \rightarrow \text{Conv}(1\times1, 1024)$	14×14×1024
Residual Block (×3)	$\text{Conv}(1\times1, 512) \rightarrow \text{Conv}(3\times3, 512) \rightarrow \text{Conv}(1\times1, 2048)$	7×7×2048
Global Average Pooling	–	1×1×2048
Fully Connected Layer	Dense (128 neurons, ReLU)	128
Output Layer	Softmax (N product classes)	N

4.2.2 Key Steps in Image Recognition

- Image Preprocessing – Resize to 224×224, normalize pixel values (OpenCV).
- Feature Extraction – Forward pass through ResNet-50 layers to capture shapes, textures, and high-level patterns.
- Fine-Tuning – Pre-trained weights adapted to product categories (fashion, cosmetics, electronics).
- Output Vector – The CNN outputs a 2048-D feature vector, later reduced via a fully connected layer and fused with text embeddings for multimodal processing.

4.3 Text Processing Module

The task of translating user inquiries into a meaningful representation that the system can utilize to produce product recommendations falls to the text processing module. Text data is subjected to NLP techniques such as tokenization, embedding, and semantic understanding. Key steps in the text processing module include the following techniques.

- Text Preprocessing: Any extraneous characters, like punctuation or special symbols, are eliminated from user queries. Text normalization, such as changing text to lowercase, is also included in this step.
- Tokenization: To make processing easier, the cleaned text is tokenized into smaller units (words or subwords). Python NLP libraries like NLTK are used for this step.
- Embedding: The BERT-based model, which captures the semantic meaning of words and their relationships in context, is used to transform the tokenized text into numerical vectors. BERT is especially helpful for comprehending user inquiries about features and descriptions of products.
- Output: For more precise product recommendations, the multimodal fusion process combines the generated text embeddings with the image features.

4.4 Multimodal Fusion

Multimodal fusion that has been used merges the different features extracted from text and image inputs by the user to create a single representation. This is an essential step in the procedure. In this step, we make sure the system is able to detect both the text and image inputs and analyse them so that it can provide a relevant and more personalized product output that matches the user's request. The main steps in the multimodal fusion process are three, and they are described below.

- Feature Combination: The small, filtered feature vectors are achieved as a CNN output from the image recognition module after applying CNN layers. The text-based outputs received from the BERT model and the CNN output are combined to form a single vector as the final output from the backend. This makes it possible for the system to use both the text and image data together as a search within the chatbot for conversational purposes and recommendations.
- Finding Matches: Vector search is a good method to find from a database. In this system, we have used that for product matches with proper relevance from within our small database. The search method is used for additional analysis to measure the similarity of a product based on user-uploaded images, text input, or both.
- Recommendation Engine: The most relevant products are found by querying through the product database to match with user-requested data searched by using the unified feature vector. These products are then ranked in terms of 1,2,3, and so on according to similarity and sent back to the UI to the user as recommendations asked by users based on the product image they uploaded.

4.5 Backend Development

Flask is used in backend development due to its being a lightweight Python web framework. It is used in the management of communication between the different modules and the e-commerce platform frontend. Incoming user queries are handled by Flask, which also processes the data using text processing and image recognition modules before returning the suggestions.

Key steps in the backend development are described as follows.

- Routing: Incoming user inputs are routed by Flask to the relevant processing module. After recognizing the input type, if the input is text or an image, the system sends that input to the relevant processing module based on the text or image.
- API Design: RESTful APIs that enable communication between the chatbot and the e-commerce platform have been made with Flask. User input is received by the API endpoints, which then process it and return the suggested products.
- Data Handling: In addition to making sure the system is scalable and capable of processing numerous user requests at once, the backend is in charge of controlling the data flow between the modules.
- Response Generation: The backend returns the results in an easy-to-use format to the UI after the product recommendations have been generated.

4.6 Tools and Technologies

The following tools and technologies were used during the development phase.

- Python: This is the main programming language used to create the system backend with various libraries.
- Keras: A high-level neural network API used to construct and optimize CNN models.
- OpenCV: Used for preprocessing images, such as augmentation, normalization, and resizing.
- BERT: For semantic understanding, it converts user queries into meaningful text embeddings.
- Flask: The web framework used to construct the backend, which manages module-to-module communication and data routing.
- Text tokenization and other NLP tasks are accomplished with the help of the NLTK library.

4.7 Challenges and Solutions

During the development phase, several challenges were encountered by us.

- Challenge 1: Ensuring that the image recognition module could accurately classify products with the limited dataset provided.
  - Solution: The accuracy of the system's image classification was upgraded by fine-tuning the pre-trained ResNet-50 model on the product categories that were chosen as the dataset.
- Challenge 2: Handling noisy or incomplete user text queries.
  - Solution: Tokenization and stop-word removal methods were used as text preprocessing techniques to process the text and increase the NLP module's accuracy for better text-based output results.
- Challenge 3: Combining textual and visual elements to create a single, cohesive representation.
  - Solution: By using multimodal fusion techniques, the system was able to efficiently combine both feature types in order to generate product recommendations.

The actual process of creating and integrating the system's essential modules is covered in the Development Phase section of this paper. Flexibility was made possible by the modular approach, which guaranteed that every part could be created and improved separately before being incorporated into the bigger system.

5. EVALUATION

5.1 Results

For the initial testing and accuracy check, we used a small dataset of 20 products from various common e-commerce categories. The dataset was split 60-40, with 12 images for training and 8 images reserved for testing the model’s classification accuracy.

In our study, we evaluated the visual chatbot’s ability to classify products in three main categories: fashion, cosmetics, and electronics. We used common metrics like accuracy, precision, recall, and F1-score. Rather than giving formal definitions, we explained these metrics in practical terms: recall shows the percentage of relevant items the model found, accuracy reflects how often the model was correct overall, precision tells us how many of the retrieved items were actually relevant, and the F1-score balances precision and recall giving an overall performance measure. To ensure fairness, we used a weighted average so that categories with more test samples had a larger impact on the final score.

The model performed very well in the fashion and cosmetics industries, achieving over 90% accuracy in both. It showed high recall rates of 92% for fashion and 88% for cosmetics, meaning it identified the most relevant products. The precision was also high, at 94% for fashion and 92% for cosmetics, indicating the model’s predictions were usually correct. These results were based on the help of three people who tested 8 images to support the evaluation. Table 3 presents the results more clearly, based on the support values and calculations.
To supplement quantitative evaluation, we carried out a small-scale user study in which three users uploaded photographs and submitted text inquiries in real-time. On a 5-point Likert scale, participants assessed the advice's applicability, usability, and general satisfaction.

Every participant expressed great satisfaction with the search experience and found the interface easy to use, as reflected in their perfect ease-of-use scores. One user suggested adding dynamic filters for options like price range or brand to help refine recommendations even more. Overall, these results show that the system's multimodal reasoning meets user expectations for both text-based and visual product searches.

However, there are a few limitations to consider. Future research should involve a larger and more diverse group of users to capture a wider range of interaction patterns, as the small sample size limits how broadly we can apply these findings. To address these challenges, we plan to enhance the CNN backbone with self-supervised pre-training, add real-time filtering options so users can fine-tune recommendations as needed, and expand the training dataset, especially for electronics and other complex categories. Additionally, offline reinforcement learning and ongoing A/B testing will be implemented to continuously adapt the purchasing experience based on user behaviour. Table 3 shows the accuracy results of the image detection framework.

Table 3. Accuracy Rate for Image Detection Framework

Category	Precision	Recall	F1-Score	Support
Fashion	0.94 (94%)	0.92(92%)	0.93 (93%)	9
Cosmetics	0.92 (92%)	0.88 (88%)	0.90 (90%)	9
Electronics	0.78 (78%)	0.83 (83%)	0.81 81%)	6
Overall	0.88 (88%)	0.87 (87.67%)	0.88 (87.83%)	24
Average	0.88 (88%)	0.87 (87.67%)	0.88 (87.83%)	
Macro Avg	0.88 (88%)	0.87 (87.67%)	0.88 (87.83%)	
Weighted Avg	0.89(89.25%)	0.88 (88.25%)	0.89 (88.75%)	

The performance of the conceptual visual chatbot system was compared to existing work in the field. CNN-based models for classifying e-commerce products show accuracies between 85% and 99% for categories like fashion, cosmetics, and electronics. Few instances are given as follows.

- Zhang et al. [13] reported 95.2%-96.5% accuracy with recall rates between 85%-91% for product categories like clothing and home décor.
- Jain et al. [19] achieved 96% accuracy using YOLOv5 for fashion item detection.
- Dahal et al. [15] achieved 93% accuracy for image classification, focusing on a more simplified dataset.

Unlike other systems, the conceptual visual chatbot combines image recognition and NLP to offer real-time, context-based recommendations. This multimodal method creates a more personalized and flexible shopping experience, even though its accuracy is a bit lower.

- Electronics, being a more complex category, showed 80% accuracy in our results. This matches the difficulty of classifying detailed images of electronic products. Table 4 summarizes the main performance metrics obtained in this study.

Table 4. Performance Metrics for This Study

Study	Methodology	Dataset	Accuracy	Recall	Other Metrics
Zhang et al. (2024) [13]	CNN-based product recommendation	E-commerce products (Clothing, Home decor)	95.2%-96.5%	85%-91%	-
Jain et al. (2024) [19]	YOLOv5 for object detection in fashion	Fashion datasets (DeepFashion)	96%	N/A	Focused on fashion items
Dahal et al. (2024) [15]	CNN for image classification	Fashion MNIST, Ibug 300W	93%	N/A	CNN with facial detection
Li et al. (2024) [20]	AI & image processing for recommendations	Fashion MNIST, self-built dataset	99.45%	N/A	Personalized system, multi-label classification
This Study	CNN + NLP for multimodal recommendations	Fashion, Cosmetics, Electronics	88%	87%	Real-time, contextual recommendations using text + image

Although the overall accuracy of 88% might seem lower than in some studies, it shows the challenge of working with both text and image inputs together. With more data and better model tuning, we expect these results to improve.

5.2 User Feedback

We carried out a user study with 30 participants, gathering feedback through a 5-point Likert scale to evaluate different parts of the system's performance. Participants were asked questions to rate usability, recommendation accuracy, and overall satisfaction.

1. "How effective did you find the system in understanding your image and text input?"
    - This question was designed to assess how well the system processes both image and text inputs to generate accurate recommendations.
2. "How likely are you to recommend this system to others for e-commerce product search?"
    - This question helped gauge overall user satisfaction and the likelihood of users sharing their positive experiences with others.
3. "How satisfied are you with the variety of products the system can recommend?"
    - This question was designed to evaluate the diversity of the product recommendations and whether users felt the system provided a comprehensive range of options. Table 5 presents the user ratings collected using the Likert scale.

Table 5. Likert Scale Ratings

Question	Average Rating	Interpretation
Effectiveness in understanding image and text input	4.5/5	Most users found the system effective at processing both image and text inputs.
Likelihood of recommending the system to others	4.6/5	High user satisfaction with the system and its recommendation capabilities.
Satisfaction with the variety of recommended products	4.4/5	Users were generally satisfied with product variety, but some suggested expanding the categories.

5.3 Quantitative Results

The feedback results are summarized below based on a 5-scale set by us.

- Ease of Use: After trying the system, participants gave the system an average score of 4.7 out of 5, showing that most users found the interface easy to use and navigate, and used the features.
- Accuracy of Recommendation: Based on user testing, the system received an average rating of 4.5 out of 5 for the recommendation of related matched products upon the search they used. Users received good output responses based on both text and image inputs, though there's still room for improvement, especially for more complex categories like electronics.
- Overall: In an overall system-level usage, most users were satisfied and gave an average score of 4.6 out of 5, indicating a positive experience during their conversation.

5.4 Qualitative Feedback

Feedback based on user acceptance is as follows.

- Real-time Recommendation Recognition: We tested the system with 30 participants for the real-time recommendation module. Their feedback was based on the text and image input, and they mentioned the product finding was fairly easier and faster due to real-time recommendations. The output received from the chatbot was related to their search.
- Multimodal Conversation Capability: We have checked with 25 participants for the multimodal query. Based on their feedback, we can understand that the system was able to handle the text and image inputs well. It was also mentioned that the system was easier to use, especially due to the fact that it can use both text and image as search options, since the other systems usually contain text-based search. Table 6 summarizes the feedback provided by users during the evaluation.

Table 6. User Response

Feedback Area	Participants' Ratings	Number of Participants
Ease of Use	4.7/5	23 participants
Recommendation Accuracy	4.5/5	21 participants
Overall Satisfaction	4.6/5	25 participants
Suggestions for Dynamic Filters	4.3/5	18 participants
Speed Improvements Needed	3.9/5	12 participants

The feedback that we gathered from the users for the study shows that users are very satisfied with the system, especially with how easy it is to use. The accuracy of the recommendations matching their products and overall experience has also been praised. Participants liked the multimodal approach method combining text and image input queries, because, according to the initial users, it is intuitive and offered personalized suggestions, which made the experience easy.

6. CONCLUSION

The paper summarizes the confirmed results achieved and states that combining textual and visual interactions makes e-commerce more dynamic and user-friendly, thus making it more feasible for users. By making use of an advanced NLP model with CNN-based image recognition, the system can handle a wide range of product inquiries from users and deliver relevant recommendations in real time in the interface.

Although the system displays a promising result, further development is necessary to make it robust and scalable, so that its accuracy across all product categories can be improved, and it is able to manage increasingly complex user requests from the e-commerce end. There are some areas that can benefit from future improvement, including a bigger dataset for testing and refining the recommendation algorithms further. In the future, we plan to add reinforcement learning, which could also allow the system to learn from user input and give better responses, resulting in a more personalized and responsive shopping experience.

In summary, the conceptual framework outlined in this paper discusses the foundation for the future development of intelligent e-commerce systems, which have the potential to change the future of how consumers discover and purchase products online. This work is aimed at contributing to the recently growing field of AI-driven personalized shopping, which has the capabilities of using better chatbots through improved image recognition and NLP. It also opens the door for further research and implementation within the e-commerce sector.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Sidra Tul Kubra: Conceptualization, Data Curation, Methodology, Investigation, Writing – Original Draft Preparation;
Salma Jahan Nisha: Data Curation, Methodology, Investigation, Writing;
Nabhan Salih: Methodology, Supervision, Writing – Review & Editing;
Amalina Ibrahim: Supervision, Writing – Review & Editing;
Wan-Noorshahida Mohd-Isa: Methodology, Supervision, Validation, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow the Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

DATA AVAILABILITY

- The data that support the findings of this study are available from the corresponding author upon reasonable request.
- Derived data supporting the findings of this study are available from the corresponding author on request.

REFERENCES

[1] D. Patil, "Artificial intelligence in retail and e-commerce: enhancing customer experience through personalization, predictive analytics, and real-time engagement," Nov. 26, 2024. [Online]. Available: http://dx.doi.org/10.2139/ssrn.5057420.

[2] J. Sidlauskiene, Y. Joye, and V. Auruskeviciene, "AI-based chatbots in conversational commerce and their effects on product and price perceptions," *Electronic Markets*, vol. 33, no. 1, pp. 24, 2023, doi: 10.1007/s12525-023-00633-8.

[3] R. Bansal, A. H. Ngah, A. Chakir, and N. Pruthi, "Leveraging ChatGPT and artificial intelligence for effective customer engagement," in *Advances in Human-Computer Interaction and Systems Design*, IGI Global, Hershey, PA, USA, pp. 1–27, 2024, doi: 10.4018/979-8-3693-0815-8.

[4] M. F. McTear, "The rise of the conversational interface: A new kid on the block?" in *Future and Emerging Trends in Language Technology: Machine Learning and Big Data - 2nd International Workshop, FETLT 2016, Revised Selected Papers*, J. F. Quesada, F. J. Martin Mateos, and T. Lopez-Soto, Eds. Cham, Switzerland: Springer Verlag, 2017, pp. 38–49, 2017, doi: 10.1007/978-3-319-69365-1_3.

[5] D. Adiwardana, M. T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, and Q. V. Le, "Towards a human-like open-domain chatbot," *arXiv*, arXiv:2001.09977, 2020.

[6] R. Kaur, A. Kumar, and P. Singh, "SignBot: A multimodal sign language-enabled chatbot for the hearing impaired," *Internaltional. Journal of Human-Computer Interaction*, vol. 37, no. 17, pp. 1592–1606, 2021.

[7] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1966, doi: 10.1145/365153.365168.

[8] S. La Bua, "Latent semantic analysis and its application in conversational chatbots," unpublished, 2015.

[9] mySimon, *MySimon price comparison website*, 2006. [Online]. Available: https://web.archive.org

[10] Inktomi Corp., *Inktomi search engine*, 2006. [Online]. Available: https://web.archive.org

[11] J. Hauswald, M. Laurenzano, Y. Zhang, C. Li, A. Rovinski, A. Khurana *et al.*, "Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers", *ACM SIGARCH Computer Architecture News*, vol. 43, no. 1, pp. 223-238, 2015, doi: 10.1145/2786763.2694347.

[12] Google Goggles, "Google Goggles overview and requirements," Google Inc., 2012. [Online]. Available: https://en.wikipedia.org/wiki/Google_Goggles

[13] L. Rao, "Image recognition startup SnapTell acquired by Amazon subsidiary A9.com," *TechCrunch*, Jun. 16, 2009. [Online]. Available: https://techcrunch.com/2009/06/16/image-recognition-startup-snaptell-acquired-by-amazon-subsidiary-a9com/

[14] Nokia Point & Find, "Nokia Point & Find," *Wikipedia, the free encyclopedia*, 2012. [Online]. Available: https://en.wikipedia.org/wiki/Nokia_Point_%26_Find

[15] Slyce, "Slyce image recognition and barcode scanning," Slyce Inc., 2013. [Online]. Available: https://en.wikipedia.org/wiki/Slyce

[16] N. Zhang, "To improve image recognition and product recommendation in e-commerce systems," in Proceedings of E-commerce conferences, 2024.

[17] N. Goel, "Shopbot: An image-based search application for e-commerce domain," *Master's Projects*, 516, 2017, doi: 10.31979/etd.r7a5-6dzf.

[18] R. Dahal, S. Dhakal, R. Timalsina, and S. Neupane, "Re-commerce site with image processing and voice recognition," in *Innovations in Retail AI*, 2024.

[19] P. Badave, B. Bhomaj, B. Bindu, R. Shivarkar, and P. N. Dhavase, "E-commerce website with recommendation system including chatbot and reverse image search," *Ijraset Journal For Research in Applied Science and Engineering Technology*, 2022, doi: 10.22214/ijraset.2022.46904.

[20] M. Kuo, H.-T. Chan, and C.-H. Hsia, "Study on Mask R-CNN with data augmentation for retail product detection," in *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS),* Hualien City, Taiwan*,* pp. 1-2, 2021, doi: 10.1109/ISPACS51563.2021.9651028.

[21] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman *et al.*, "Fashion IQ: A new dataset towards retrieving images by natural language feedback", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11302-11312, 2021, doi: 10.1109/cvpr46437.2021.01115.

[22] D. Jain, E. M. Thazhathu, I. Adiraju, J. Bhattacharya, and D. Singh, "FashionAI: image-based clothing detection and shopping recommendation," *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, pp. 1-6, 2024, doi: 10.1109/INOCON60754.2024.10512219.

[23] C. J. Li, "A personalized product recommendation system for e-commerce platforms based on artificial intelligence and image processing technologies", *Traitement Du Signal*, vol. 41, no. 6, pp. 2961-2971, 2024, doi: 10.18280/ts.410615.

[24] Z. B. Ter, P. Naveen, and J. Jayapradha, "Generative AI-based meal recommender system," *Journal of Informatics and Web Engineering*, vol. 4, no. 2, pp. 20–35, 2025, doi: 10.33093/jiwe.2025.4.2.20.

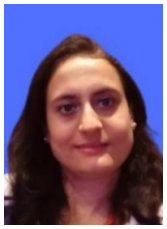
[25] Y.-X. Lim, S.-C. Haw, and J. Jayapradha, "Optimizing reviewer assignment with recommender systems: Models, related work, and evaluation," *International Journal on Robotics Automation and Sciences*, vol. 7, no. 2, pp. 56-76, 2025, doi: 10.33093/ijoras.2025.7.2.6.

[26] T. Natadirja, H. Akbar, G. Firmansyah, and B. Tjahjono, "E-commerce product image-based recommendation system Kalcare.com using deep learning," *Jurnal Indonesia Sosial Teknologi*, vol. 4, no. 8, pp. 930–940, 2023, doi: 10.59141/jist.v4i8.669.

[27] T. K. Hanchinal, V. D. Bhavani, and V. B. Mindolli, "Intelligent beauty product recommendation using deep learning," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), Bhubaneswar, India*, pp. 1–5, 2024, doi: 10.1109/IC-CGU58078.2024.10530808.

[28] A. Serban and F. Bota, "A conceptual framework for software fault prediction using neural networks," in Simian, D., Stoica, L. (eds) *Modelling and Development of Intelligent Systems*. MDIS 2019. *Communications in Computer and Information Science*, vol 1126. Springer, Cham., vol. 1126, Cham, Switzerland: Springer, pp. 137–148, 2020, doi: 10.1007/978-3-030-39237-6_12.

BIOGRAPHIES OF AUTHORS

	**Sidra Tul Kubra** is currently pursuing a Master's degree by research in Information Technology at MMU, specializing in Artificial Intelligence. Her research focuses on AI-driven chatbots and visual interaction systems, aiming to merge cutting-edge technology with user-centered design principles. Sidra holds a Bachelor's degree in Software Engineering, which has provided her with a solid foundation for her professional work. As a Web Developer, she has successfully built websites across various sectors, including e-commerce platforms, non-profit organizations, for-profit businesses, and content management systems (CMS). Sidra's work exemplifies her ability to balance research and hands-on experience in the technology field. She can be contacted by email: sidra.tul.kubra@student.mmu.edu.my or sidra7472@gmail.com.
	**Salma Jahan Nisha** is currently pursuing her Master of Information Technology (Research) at Multimedia University, Malaysia. She holds a Bachelor's degree in Software Engineering with Multimedia and is a Software Engineer with professional experience in web development and project management. Her research focuses on *Sign Language Chatbot in the E-commerce Domain*. She has delivered technology projects in education, including the Training Management Portal and the Brain Based Learning application, and also taught a UX/UI course as a guest lecturer in Singapore for Educare Global Academy. Her research interests include sign language recognition, human–computer interaction, natural language processing, and real-time AI systems. She recently presented her work *"Sentence-based Sign Language Recognition using CNN and NLP"* at MECON 2025. She can be contacted by email at: salmajn359@gmail.com or 1221404902@student.mmu.edu.my.
	**Nbhan D. Salih** is currently a Lecturer at Ohio State University, USA, with prior academic experience as an Assistant Professor in the Faculty of Computing and Informatics at Multimedia University (MMU), Malaysia. He holds a Ph.D. in Engineering, and his research interests lie at the intersection of computer vision, digital image processing, and machine learning. Dr. Salih has been actively involved in advancing technologies that enable intelligent image analysis and pattern recognition, contributing to both the academic and applied aspects of visual computing. He can be contacted at email: salih.43@osu.edu.
	**Amalina Ibrahim** currently serves as a lecturer at Multimedia University's Faculty of Computing and Informatics (FCI). In her role, she primarily focuses on teaching students in the areas of Computing Fundamentals and Mathematics. Additionally, she supervises students during their final projects and internships. Ms. Amalina earned her Master of Science (IT) degree from Multimedia University. Her research interests revolve around Image & Signal Processing, Computer Vision, and she also specializes in underwater imaging. Outside of academia, Ms. Amalina holds a Malaysian Amateur Radio license under the call sign 9MHER, and she is also a member of the National Association for Amateur Radio with the call sign K2MAL. She can be contacted by email at: amalina.ibrahim@mmu.edu.my.

Wan-Noorshahida Mohd-Isa is currently an Assistant Professor at the Faculty of Computing and Informatics, Multimedia University, Cyberjaya campus, Malaysia. She is also the Chairperson for the Research Centre for Image and Vision Computing of Multimedia University. She is an engineering alumnus of Vanderbilt University, USA with a Bachelor's degree in Electrical Engineering (magna cum laude) who then ventured into the computing field. She completed her Master's degree from the University of Southampton in Great Britain and she went on to pursue her Doctorate degree at Multimedia University. She was a Laureate of the France-Malaysia Collaboration Programme for Joint Research 2022 (MyTIGER 2022), which is a research programme under the Embassy of France to Malaysia. Her specific research area is on visual analytics, where she designs and develops machine learning models for videos and images. She can be contacted by email at: wan.noorshahida.isa@mmu.edu.my.