National College of Ireland

Optimizing Footwear Recommendations Using CNN-Based Visual Similarity and Health-Fit Filtering

MSc Research Project
MSc in Data Analytics

Siddhesh Patil
Student ID: x23309512

School of Computing
National College of Ireland

Supervisor: Prof. Hamilton Niculescu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Siddhesh Patil
Student ID:	x23309512
Programme:	MSc Data Analytics **Year:** 2025
Module:	MSc Research Project
Supervisor:	Prof. Hamilton Niculescu
Submission Due Date:	11-08-2025
Project Title:	Optimizing Footwear Recommendations using CNN based Visual Similarity and Health-Fit Filtering
Word Count:	8052 **Page Count** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Siddhesh Patil
Date:	11-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Footwear Recommendation
Using CNN-Based Visual Similarity and Health-Fit Filtering

Siddhesh Patil
x23309512

Abstract

The increased popularity of personalized and comfort-focused recommendations in fashion e-commerce has shown that the current visual recommendation systems have severe limitations due to prioritizing aesthetic appeal over practical functionality. The study describes a hybrid footwear recommendation system composed of three-phase framework combining CNN-based models and rule-based health-fit filtering. The experiment uses a curated version of the UT Zappos50K dataset, where four ergonomic properties, namely arch support, cushioning, breathability, and wide-feet compatibility, are characterized by means of metadata-based tagging and domain-specific heuristics. Phase 1 involves extracting visual embeddings that are based on pretrained CNN models such as ResNet50, VGG16, and EfficientNetB0 in order to provide unimodal similarity findings. In Phase 2, extension of the current retrieval setting to multimodal retrieval is performed by adding a filter of a binary health-fit tag to the embeddings. In Phase 3, health-fit attributes are predicted by a multi-label classification model that is trained based on footwear images. The evaluation was done based on Precision@k, macro F1-score, cosine similarity, and health-tag match rate. All models show good results, but out of all, EfficientNetB0 emerges as the most efficient. It performs better for comfort-related features like breathability and wide feet. Its capacity of producing compact and semantically dense representations makes it a good fit for hybrid rule-based filtering systems.

The given research introduces a usable, transparent, and ergonomically balanced recommendation framework that combines deep visual learning and medical-grade personalization in the domain of fashion technologies.

1 Introduction

Over the past few years, artificial intelligence (AI) and the fashion e-commerce niche have been colliding and changing how users can discover and interact with products. Visual recommendation systems have now been developed with the help of deep learning models, and especially the Convolutional Neural Networks (CNNs), which make recommendations on the basis of aesthetic similarity according to user inputs, i.e., uploaded images or curated preferences. It has been especially relevant in the field of footwear, where the most minimalistic style-specific differences, such as the shape of the toe, the thickness of the sole, or the texture of the material, can greatly affect the decision of the customer. However, while these systems excel in identifying aesthetic similarity, they fail in addressing a practical concern- functional comfort and foot health. When it comes to footwear, recommendations in the present context seldom take into account the personal ergonomic requirements, although consumers are increasingly becoming conscious about the comfort and orthotic positioning. For people with certain kinds of anatomical conditions, e.g., plantar fasciitis, overpronation, or wide feet, the selection of proper footwear can be directly associated with musculoskeletal well-being. Nevertheless, true integration of such

health-fit qualities in most recommender systems remains a work in progress, thus producing style-focused but health-agnostic ideas (Umar *et al.*, 2022)

Visual recommendation pipelines are centred around CNN models like VGG16, ResNet50, and EfficientNetB0 that have performed better than others in object detection, picture retrieval, and fine-level classification. In other cases, such as models pretrained on images like ImageNet, they can be customized to produce dense semantic embeddings of footwear images. These embeddings are commonly compared against others with a similarity measure like cosine distance or constructed into an efficient index such as FAISS to perform a scalable indexing task (Johnson *et al.*, 2017). Although a lot of literature exists in the evaluation of CNNs over classification tasks, very little research has been done in regard to their application over health-aware hybrid recommendations systems, and more specifically, those that yield combinations of visual similarity with lower downstream combinations of rules, to determine aspects of comfort and ergonomics.

This research addresses that gap by proposing a three-phase recommendation framework:
i. Phase 1 to extract the visual embeddings using CNN architectures, serving to retrieve similarities in a unimodal process.
ii. Phase 2 will also present multimodal filtering where the results of visual search will be narrowed down with rules-based logic based on binary health-fit labels that include arch support, breathability, cushioning, and wide-feet compatibility.
iii. Phase 3 conducts prediction through training the multi-label classification model that will be able to directly learn to predict these ergonomic attributes based on images.

A curated subset of UT Zappos50K data is used to develop the system and is manually annotated to health-fit labels through medically informed criteria. Evaluation metrics of model performance are macro F1-score, Precision@K, Health Tag Match Rate, cosine similarity, and t-SNE, UMAP plots that could provide some visual intuition of the quality of different CNN architectures to cluster similar shoes in the latent space.

The research question that guides the study is:
What is the comparative performance of ResNet50, VGG16, EfficientNetB0 CNN architectures with respect to visual embedding that aids in recommending health-aware footwear through hybrid filtering and multi-label classification?

This work will provide a visually expressive, individualized, and functionally associated approach to the recommendation framework, which consists of visual representation learning and health-oriented tagging. The results lead to the development of fashion-tech systems that focus on their stylish and comfortable elements, as well as contribute to an inclusive user experience in areas where technology design is incorporated with well-being Tan and Le, (2019).

2 Related Work

To contextualize the contributions of this study, prior literature is categorized into four core themes: (1) deep CNN architectures in fashion tasks, (2) visual embedding and retrieval methods, (3) health-fit tagging and multi-label classification, and (4) interpretability via embedding space visualization.

2.1 CNN Architectures for Visual Embedding in Fashion and Footwear

CNNs have played a pivotal role in most tasks in computer vision, especially with regard to image classification and feature extraction. EfficientNet is one of the most powerful architectures. Proposed by Tan and Le, (2019), the EfficientNet suggests the compound scaling approach that

balances depth, width, and resolution more substantially compared to the previous ones. This makes it precise and effective, especially in comparison to large networks such as ResNet. Nonetheless, the paper obtained top results on ImageNet, but its intended scope was limited to classification accuracy, and it failed to assess the suitability and quality of the embeddings for tasks such as retrieval or tagging. ResNet, produced by He *et al.*, (2016), is another milestone in architecture. ResNet introduced a way to train very deep networks without the learning difficulties that accompany the use of deep networks, such as the vanishing gradients. Such models as ResNet50 have become common patterns in different fields. Despite the fact that the original ResNet work was generic and not footwear-specific, Shen *et al.*, (2023) implemented ResNet50 on shoe classification in a surveillance environment. Even with a poor resolution of their input images, the model was still able to achieve accuracy well above 90 percent, which implies that ResNet is still a dependable model even in the presence of noise during real-life applications. Still, they did not compare architectures or study the quality of the feature embeddings that can be generated, both of which are essential to retrieval. Another popular architecture is VGGNet proposed by Simonyan and Zisserman, (2014), which has a simple design characterized by a stacking of small convolution filters that are simultaneously small. It gets used a lot in transfer learning because of its clean underlying structure, though VGG16 is often quite heavy in parameters and prone to overfit on smaller datasets.

The performance of these CNNs has been studied in the fashion domain, where a few studies have been carried out. As an example, Alishev, (2024) tested EfficientNet and VGG to classify fashion and retrieve the images. The paper attempted to improve the dissimilarity among embeddings by either contrastive loss or hyperparameter tuning. Although the results were encouraging, the model's basics were not applied to the UT Zappos50K and did not apply some health-oriented labels, e.g., arch support or cushioning. On a similar note, Liu *et al.*, (2024) proposed not only very lightweight CNNs but also models that could be used to predict such single product tags, as material or fit in this case. Their models were efficient but not on multi-label tasks, and not in testing the performance on retrieval-based systems. An even more directly applicable contribution can be found in Suvarna *et al.*, (2024), who performed a fine-tuning of ResNet50, VGG16, and EfficientNetB0 on the UT Zappos50K dataset, specifically with a footwear classification task in mind. In their research, they found EfficientNetB0 to be the most well-balanced in accuracy and speed. Even though they are helpful when selecting a model backbone, they did not consider how appropriate the learned embeddings were to use in either of the two central aspects of this project, retrieval and tag prediction.

2.2 CNN-based Visual Retrieval and Fashion Recommender Systems

CNNs are generally applied in classification solutions, but they also have a significant performance in visual retrieval and recommendation systems, as well as in fashion. This is one of the first instances of which is presented by R. *et al.*, (2024), who created a pipeline on a CNN model that was capable of extracting features in product images and qualifying them by their visual relevance. This way of doing it contributed to making retrieval precision and computation performance better on large-scale foundations. Nevertheless, the system did not consider both the needs and preferences of users, at least the health-related preferences such as arch support or foot width, which diminishes its applicability in providing more user-specific or ergonomic suggestions. A more vision-oriented application is where Chang *et al.*, (2023) used CNNs during the shopping for shoes and tried to locate visually similar items in a catalogue. They used classification and nearest-neighbour retrieval using embeddings. It performed effectively in visual similarity, though it was never able to provide an explanation as to how the recommendation was being made, and without post-filtering based upon comfort to the user, or functionality. Suvarna *et al.*, (2024), in their turn, also evaluated the retrieval performance with the embeddings of VGG16, ResNet50, and EfficientNetB0, building on their previous studies concerning classification. Again, EfficientNetB0 was the most accurate one. Nevertheless, neither

their system contained the features of interpretability nor any form of filtering based on health-fit interest, which is paramount in case we desire the suggestions to become really practical. Retrieval, considered by Alishev, (2024), is the classification criterion helpful to support the implementation of this work. With the help of contrastive learning, the objective of the study was to make embedding spaces semantically significant. This aided in the increased accuracy of the retrieval of fashion datasets. Nonetheless, the system could not support non-visual contents such as support or fit, nor could it visualize how structures were composed in the embedded elements, and this restricts the interpretability and applicability in the real world. By contrast, Abdulnabi *et al.*, (2015) suggested an even more user-centred approach to the recommendation system, which combined not only CNN-based embeddings of images but also user-made metadata. They employed the idea of interpretability through Grad-CAM in their model so that the user or the developer would be aware of what the model regarded during retrieval. It was also successful in comfort-seeking recommendation tasks. With that said, they had quite a modest dataset and did not test the scalability of the system with filtering on more complex health-fit tag groups.

Overall, the study indicates that there has been a definite progression in the research in terms of starting with basic image similarity systems to evolving into a hybrid structure that attempts to introduce explainability and context. Nonetheless, most of them fall short in the area of incorporating attributed features such as arch support, or they offer no good reasons as to why some objects were retrieved. That is precisely the territory that our study aims to research more deeply.

2.3 Multi-Label Classification and Health-Fit Tagging

Over the past years, the role of multi-label classification in the fashion and e-commerce studies has received more attention, particularly in light of the products of many types or with multiple properties. This is addressed by Baldrati *et al.*, (2021)who proposed a tag-aware self-supervised retrieval network contrastive learning complemented with attribute supervision. Multiple fashion tags are associated with their model as their model can effectively learn embeddings that match well with those tags, and this enhances retrieval accuracy. Nevertheless, they did not consider aesthetic labels such as colour and fabric. They did not incorporate important ergonomic or health-fit qualities (e.g., arch support, wide feet, cushioning & breathability) into their dataset, and the model could not be as readily applied to functional advice. Corresponding to this Li *et al.*, (2023) introduced a contrastive learning framework in order to come up with entity embeddings that focusses on the product attributes. Their approach minimised the use of such labelled data, which will be important during scaling, as well as enhanced clustering of styles similar to one another. Nevertheless, their tags were visual or fashionable, as was the case with Baldrati *et al.*, (2021). The presence of functional tags that influence health and comfort was not taken into account and tested, and these factors restrict the value of using them in activities such as tutorials on health-conscious shoe recommendations. This gap was also filled by Liu *et al.*, (2024), who devised lightweight CNN structures that are able to perform under low-resource conditions. They foresee the personal products' attributes, such as fit or fabric, in their models without making the models very big in size. Though this has practical usefulness in the context of on-device use, the models not only failed to provide support for multiple tags but also were never provided with the use of retrieval systems, two features important in making extensive recommendations in any practical scenario. In the health-related perspective, Umar *et al.*, (2022) performed a clinical study that involved the effects of footwear characteristics on plantar fasciitis. They discovered that the use of shoes with adequate arch support decreased pain levels significantly among users, which further confirms the premise that a shoe choice can indeed affect one's health in real-life terms. Although that study was not AI-based, it showed the importance of the health-fit tags and justified the necessity of integrating these tags into the recommenders. In the material aspect, R. *et al.*, (2024) surveyed the new advancements in smart footwear material, including breathable

mesh and shock-absorbing soles. Such characteristics relate directly to the health-fit labels employed in this project, such as cushioning and breathability. Even though the overview did not relate these concepts to computer vision or deep learning, it is a solid background on the rationale of the importance of these features. Alcacer *et al.*, (2021) focused on personalization differently and offered shoes of the required size with the help of a hybrid system incorporating both classification and collaborative filtering. They used the past user interactions as the basis of their approach to predicting the fit and comfort. Although this works well in classical recommendation systems, it does not apply in a cold-start scenario where the system only has an image at its disposal. In addition, it has not exploited CNNs or visual information; hence, it somewhat restricts adapting to our content-based approach.

In summary, though the tag prediction and attribute learning areas have been extensively studied, most of them concentrate on either the visual aesthetics or metrics based on user history. There are not many that attempt to meet at the confluence of graphic content and functional health-fit labels, particularly under a multi-label scheme. Closing this gap is among the main objectives of the present study.

2.4 Interpretability and Embedding Space Visualization

With increasingly sophisticated machine learning models, it is more and more crucial to understand the decisions they take. Transparency can be particularly crucial in application domains that involve a user, such as a product recommendation engine. T-SNE is one of the common ways to visualize deep learning feature spaces. Kobak and Berens, (2019) studied the ways of enhancing the capabilities of t-SNE when representing high-dimensional data, especially in the biological data context. The ideas can be applied even though their work did not pertain to the fashion or computer vision arena. To take a concrete example, when we would like to know how our footwear embeddings cluster by tag or by functional role, t-SNE or UMAP may provide us with an intuitive guide as to what the model has learned. Regarding large-scale retrieval, (Johnson, Douze and Jégou, 2017) have proposed a very fast similarity search library, namely FAISS, which can be easily executed through GPU-enabled applications. FAISS uses fast and scalable nearest neighbour search in very large embedding spaces and is frequently applied to real-time recommendation systems. It constitutes the core of the retrieval system that is used in this project. Nevertheless, FAISS alone is performance-oriented and lacks implementation support to explain why the obtained items may be returned, which is what this work is trying to change. (Stanković *et al.*, 2018) have considered the task of ergonomic shoe fitting within the context of non-deep-learning and studied 3D shape scans of 3D feet to enhance the fit. They used geometric similarity of feet to cluster and to recommend shoe matches. Although this adds weight to the idea of relevance of clustering in footwear design, the study employed neither image-based technique nor embedding, so it is not as applicable to our CNN-based architecture. Because of that, it has been mentioned here as a contextual example but not in the sense of methodological reference. Chang *et al.*, (2023) developed CNN embeddings in their visual recommendation system in the comparison with the product metadata, including type and material. Although this helps to justify the similarity of clustering in footwear design, the study could not rely on image-based approaches or embeddings, which is why it is less applicable in our CNN-based investigation. That is why it is referred to here primarily as one of the contextual rather than methodological examples. Chang *et al.*, (2023) in their visual recommendation system, generated CNN embeddings and matched them with product metadata like type and material. Although this was effective with visual similarity, it did not extend to how the visualization of embeddings may fit or how certain products were grouped together, which could be used to increase the level of trust and transparency on AI-powered tools. Abdulnabi *et al.*, (2015) had a multitask learning strategy where attributes and categories were classified using the same CNN model. Their framework augmented classification uniformity by means of shared backbones and multiple

output heads. They however, did not focus on ergonomic features but on e-commerce properties described by sellers (such as product category or brand). They further did not discuss the embedding of visualization, which constrains the interpretability of the learned representations. Generally, these works note that there is a space where deep learning system practitioners need to explore better visualization and explainability despite the existence of some tools regarding the understanding of deep learning systems. The study hopes to bridge that by pooling together elements of CNN-based embedding extraction with visualizations and dimensionality-reduction methods such as t-SNE and UMAP to better inform about how the models behave, especially in relationship to measures of functional health-fit attributes.

2.5 Summary and Research Gap

Although the past literature has taken steps in CNN-based fashion recommendation and visual embedding, multimodal networks themselves, although, there still exists a dire need for systems that:

- Personalize for health-fit needs
- Include rule-based ergonomic filtering
- Quantitatively compare CNNs on both visual and functional similarity
- Use embedding-space interoperability with medical properties

The novelty of the study presents that unique gap by building a hybrid footwear recommendation system composed of CNN-based embeddings, multi-label identification of health-fit tags, and rule-filtered personalization, with quantitative and visual analysis. Refer the summary of related work in table 1.

Title	Authors	Key Focus	Relevance to Research
EfficientNet: Rethinking Model Scaling.	(Tan and Le, 2019)	Introduced compound scaling for CNNs	Model architecture selection (Efficient Net)
Sports Shoe Classification Using CNN	(Li *et al.*, 2023)	Fine-tuned CNNs for sports footwear recognition	Model architecture selection (VGG, ResNet) and classification benchmarking
Multi-Modal & Health-Fit Tagging	(Baldrati *et al.*, 2021)	Introduced a multimodal, tag-informed retrieval system	Closely related to tag embedding and classification
FAISS: Billion-scale Similarity Search with GPUs	(Johnson, Douze and Jégou, 2017)	Developed an efficient indexing technique for high-dimensional vectors	Backbone tool used for visual similarity retrieval in this research

Table 2.1. Summary of Important Related Work

3 Research Methodology

This study is aimed at establishing a health-aware footwear recommendation system. The procedure consists of five basic components, namely dataset preparation, dataset preprocessing, visual embedding extraction, multimodal embedding fusion, and multi-label classification, followed by final evaluation. In further sections, each of these phases will be explained.

3.1 Dataset and Data Preprocessing

The foundation of this study is the UT-Zappos50K dataset, a publicly available resource comprising 50,025 footwear images and rich, corresponding metadata UT (Yu and Grauman, 2014; Yu and Grauman, 2017) was accessed via the Kaggle distribution by Shah (2021). The characteristics of this dataset make it highly applicable in research studies, as it also contains valuable information in the form of detailed textual descriptions alongside the categories of the attributes like Insole, Material, ToeStyle, and HeelHeight that are important for functionality extraction.

A rule-based heuristic is created to convert this raw metadata into a structured, machine-readable format to be used by the hybrid system. This data preprocessing pipeline designs the programmatic parsing of the textual description in order to generate a set of four binary tags related to health-fit, arch support, cushioning, breathability, and wide feet. This heuristic is recovered from the orthopaedic and podiatric literature from (Umar *et al.*, 2022). The 'arch support' tag is assigned a value of 1 if the Heel height column contains '1 inch' or '1 ¾ inch' or otherwise, the Insole column contains keywords like 'orthotic', 'contour', 'arch', or 'poron'. Similarly, the 'cushioning' tag was identified by keywords such as 'padded', 'cushion', 'memory foam', and 'gel', and breathability was derived at keywords in the material column as mesh, knit, canvas, perforated, and woven. Finally, the wide feet tag was evaluated on the basis of ToeStyle such as wide, round, square, or open toe. This was a very orderly and medically consistent tagging procedure, which led to uniform and medically reasoned labelling of the dataset; ultimately giving a multi-label training set.

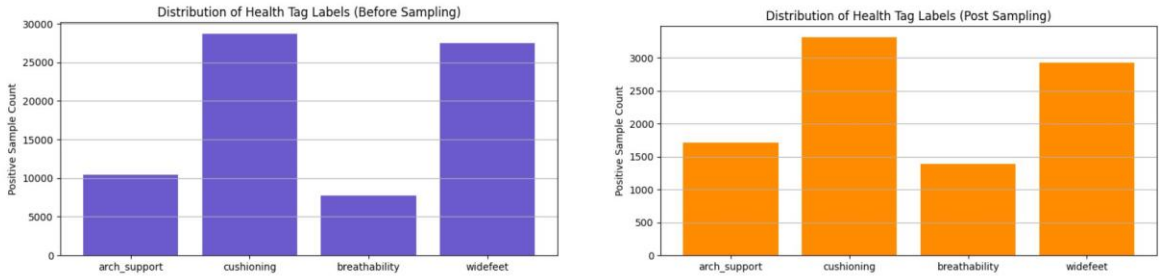

Fig.1. Health-tag Distribution Before & After Sampling

However, to train deep learning models using the entire dataset, the model training became too computationally constrained, considering that it involved finetuning numerous CNN-based architectures and conducting repetitive and multiple inferences on the different modules. The stratified sampling strategy was used to ensure that the representational diversity was not negatively affected in terms of this limitation. That is, 1,000 images were selected randomly to at least have each of the four target health-fit tags, producing an initial pool of 4,000 images. Because tags were not mutually exclusive, overlaps naturally occurred (e.g., footwear with both arch support and breathability), and these were retained to preserve authentic co-occurrence patterns. Following this, 1,000 images were randomly drawn from the remainder of the dataset to add generic miscellaneous footwear with mixed tags, resulting in a final set of 4,897 unique images without additional deduplication. This subset reflects realistic multi-label distributions while maintaining overall category representation, allowing a reduced risk of severe label bias or under-representation. The impact of such sampling strategy can be seen by comparing the distributions of the health-tags before and after the sampling strategy, as shown in Fig.1, where the balance between categories has been improved.

3.2 CNN Architecture Overview

To embed a health-sensitive footwear recommendation framework that takes the balance between semantic visual comprehension and ergonomic interpretability into consideration, this research chose the implementation of the ResNet 50, VGG16 and EfficientNet B0 as their main visual backbones at each stage. These architectures exemplify a wide philosophy on design and form a meaningful contrast of depth, performance and feature learning over visually grounded retrieval challenges.

ResNet50 by (He *et al.*, 2016) is popular as the residual learning framework. In contrast to traditional deep CNNs, ResNet does not suffer the problem of vanishing gradients on deeper layers due to the ability of identities (skip connections) to allow gradients to percolate through deeper layers. Every residual block is taught the difference F(x) between the input and output, which causes effective training and improved generalization in deep networks. As demonstrated on domain-specific retrieval tasks in recent work (Shen *et al.*, 2023), ResNet50 is especially effective at learning hierarchical feature representations--for example, learning to represent the extent, pattern, or overall ergonomic focus of sole curvature or mesh texture in conjunction with ergonomic aspects, such as arch support or breathability. Its depth and abstraction ability qualifies it to be a good candidate of fine-tuned classification (Phase 3) and semantic retrieval (Phase 1).

VGG16, introduced by Simonyan and Zisserman, (2014) is a 16-layer deep CNN with a uniform architecture and small size filter 3x3 filters in all layers of the network. Although VGG16 has more parameters and, therefore, requires more computational power, it generates dense and uniform feature maps, which is useful in major tasks, such as the categorization and footwear identification, where the unique shape descriptions, such as the open toes, heel, or broad forefoot, are keen to be captured. Its application in structured product recommendation Suvarna *et al.*, (2024) and in footwear retrieval based on surveillance (Alishev, 2024) illustrates that VGG16 still remains a hardy solution in fashion-oriented matrices. It proves to be efficient locally and in general as a baseline in the embedded in Phase 1 and Phase 3 (embedding-extraction and classification) in the bottom layer and especially in smaller regime datasets.

EfficientNetB0 by Tan and Le, (2019) reinvents the design of CNNs by compound scaling, which changes depth, width and resolution with a fixed scaling coefficient in tandem with each other. This render EfficientNetB0 to be more parameter-efficient and inference-friendly, and does not do it on the expense of expression capacity. It applies MBConv blocks (with inverted residual bottlenecks, also referred to as inverted residual bottlenecks and depthwise separable convolutions) in its internal architecture to eliminate redundant computation. Such design enhancements allow it to create small and semantically rich embeddings with very little overfitting. As shown by Abdulnabi *et al.*, (2015), EfficientNetB0 can be used as an effective embedding method to predict medical tags and learn product attributes but can be used in limited compute conditions that deeper models cannot. It is suitable to our Phase 2 multi modal fusion since here speed and representational alignment in binary health-fit tags are of importance.

The inclusion of these three architectures enables the study to reach an understanding of the impact of different depths, techniques and their effects on the capability of CNN to acquire ergonomic semantics from images.

3.3 Phase One: Visual Embedding Extraction

Driven by the fashion recommendation study by Li *et al.*, (2023) and Alishev, (2024), this paper used a similar idea of using visual similarity in the first phase to explore the possibility of using such architectures to achieve health-aware retrieval without explicit supervision in the first place. This process was performed to: 1) compare the ability of the pretrained CNNs to cluster the ergonomically identical shoes based on appearance only, and 2) to create a feasible visual suggestion pipeline that would involve scalable retrieval procedures such as FAISS. This provided a baseline to compare against the subsequent classifying multimodal and classification-based enhancements as well. This is completed by employing three pretrained convolutional neural networks in the combination of ResNet50, VGG16, and EfficientNetB0. These embeddings encode visual information on tags like shape, material, design, which may be correlated with their ergonomic property e.g. breathability or arch support. FAISS is proven to be a mechanism that scales to handle the size of the database of visually similar shoes. This step enabled examining whether appearance, on its own, can then be used as a significant surrogate of health-fit traits. Also, dimensionality reduction such as t-SNE was applied to investigate the effectiveness of the visual features clustering based on ergonomic tag, which gave early indication as to the effectiveness of visual-only retrieval.

3.4 Phase Two: Multimodal Embedding + Health Tag Fusion

Although visual similarity picks up structural and stylistic attributes, it fails to take on functional or ergonomic requirements as such. The next step includes multimodal embeddings fusion, where visual information along with favouring health-tag data is fused in order to create richer embeddings that are semantically anchored.

This step is motivated by the recent advancements of a tag-aware retrieval mechanism Baldrati *et al.*, (2021)**,** where the labels referring to health features (like arch support, cushioning, breathability, and wide feet) should be added to the embedding space to increase semantic compatibility between user needs and the provided recommendations. In particular, binary tag representation was concatenated with the pretrained CNN embedding of every single picture to form a hybrid representation that captures not only the appearance, but also the ergonomic usefulness. The multimodal vectors thus generated were subsequently indexed through FAISS to facilitate the fast top-k retrieval. This fusion solution enabled the system to prioritize recognizing products not only that were similar in practice but also were more suited to the user-based health-fit constraints than visual-only retrieval.

To determine the impact of this fusion, we could do tests on the retrieval performance, both based on Precision@5 and a HealthTagMatch@5 measure, which counts the goodness of the retrieved items in comparison to the health-fit tags of the queries. Examination by the t-SNE plots further tended to show superior clustering in a functional axis proving the ability of multimodal fusion to occupy a visual aesthetic-ergonomic relevance niche.

3.5 Phase Three: Multi-Label Classification

The current stage aims at predicting the ergonomic properties of footwear based on a multi-label classification platform. This step has been implemented to improve the practical applicability of the system further. As discussed in the work of Baldrati *et al.*, (2021) learning multiple product attributes concurrently can increase the generality of deep neural networks in e-commerce tasks by stimulating the model to learn distributed visual semantics. Taking learnings out of the study, the method develops the forecast of four health-fit labels of arch support, cushioning, breathability, and wide feet to be described by multi-label classification.

Three fine-tuned pretrained convolutional neural network architectures ResNet50, VGG16, and EfficientNetB0 with the corresponding weights, were trained on the task using curated dataset with custom heuristic labelling. This was aimed at acquiring the mappings of image pixels to ergonomic indicators which are medically relevant and they worked to ensure that the recommendation system becomes not only visually conscious but also functionally intelligent. These fine-tuned models are also used to generate embeddings that form a layer used in downstream and multimodal fusion and thereby, in evaluation. Multi-label predictions quality was calculated across the macro-averaged F1-scores and per-tag results. Additional structure in learned embedding space could also be confirmed with t-SNE and UMAP visualizations providing the qualitative evaluation of the way models showing ergonomic properties distinction.

3.6 Evaluation Metrics

Multiple evaluations were implemented to rectify the performance and reliability of all the modules of the proposed system depending on the specific goals of the different modelling phases. In case of the Phase 1 (visual embedding and retrieval systems), Precision@5 was adopted to determine the ergonomic preservation efficacy of visually similar items. For Phase 2 (multi modal retrieval) addition of 'HealthFitMatch@5' was presented as one of the custom metrics that evaluate the overlap between predicted health-fit tags on query and top-5 retrieval results, which gives a tag-level semantic measure of quality of recommendations. For the Phase 3 (multi-label classification) stage, all typical classification measures, including macro-averaged F1-score, and precision, and recall were determined on all four health-fit categories to address the class imbalances. The separability and clustering behaviour of high-dimensional embeddings by using t-SNE and UMAP plots also added an interpretability level to model inspections. Through this multi-pronged strategy, both functional and semantic correctness of the pipeline was thoroughly tested.

4 Design Specification

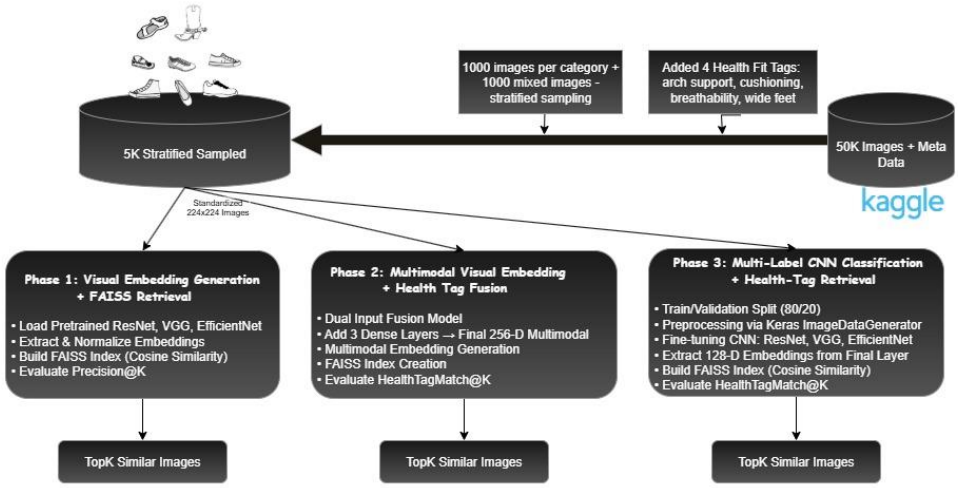

Fig. 2. System Architecture

The orchestrated system is formulated into a three-phase pipeline that sequentially improves the recommendations of footwear trying to match visual similarity and health-fit relevance. Following the UT-Zappos50K dataset, stratified sampling was carried out to generate 5,000 images that ensured four manually tagged health-fit labels, arch support, cushioning,

breathability and wide feet balanced. The pretrained convolutional neural networks (CNN): ResNet50, VGG16, EfficientNetB0 were applied to the shoe collection in Phase 1 and their visual embedding was updated using FAISS to query the top K visually similar shoes based on cosine similarity. Phase 2 developed a multimodal fusion model that fused visual embeddings with health-tag vectors to achieve richer encoded 256-dimensional representations to facilitate more semantically receptive retrievals that were assessed with HealthTag Match@K. Lastly, the Phase 3 conducted the multi-label classification whose aim was to predict the four health-fit tags on the image directly with the fine-tuned CNNs. Embeddings at the last layer of classification models were also indexed and retrieved, enabling model comparison in terms of both prediction performance (characterized by conventional measures such as macro and micro F1-score) and retrieval according to tag. Such a modular design facilitates the entire recommendation system, which takes into consideration the aesthetics as well as ergonomic functionality.

5 Implementation

5.1 Data Preprocessing and Health-Tag Engineering

Steps	Description	Details
1. Import Libraries	Import all the necessary libraries required	Importing numpy, pandas, os, keras & tensorflow libraries
2. Data Loading & Extraction	Load the zip file containing the dataset and the metadata and unzip the images in a local directory.	zipfile.ZipFile, zip_ref.extractall(), pd.read_csv('meta-data.csv')
3. Path & Filename Mapping	Create a mapping between the CID from the metadata and the actual image file paths.	Created a cid_filename column by cleaning the CID and a image_full_path column by iterating over files with os.walk.
4. DataFrame Merging	Merge the metadata with the image paths to create a single, unified DataFrame for all subsequent operations.	pd.merge(meta_df, image_df, on='cid_filename', how='left')
5. Metadata Normalization	Convert the important metadata columns (Heel Height, Insole, Material, ToeStyle) to lower-case strings to unify the keyword matching.	merged_df[col].fillna('').astype(str).str.lower()
6. Health-Tag Engineering	Four Binary health-fit tags are programmatically generated based on a series of keyword rules extracted from the medical literature.	keyword checks for 'arch_support', 'cushioning', 'breathability', and 'widefeet' in the normalized metadata columns.
7. Dataset Sampling	Use a stratified sampling technique to sample a smaller, computationally feasible, and balanced part of the entire data set	Selected a subset of 1,000 images for each health-fit tag and a random sample of 1,000 for other footwear.

Table 5.1: Data Preprocessing and Health-Tag Engineering

5.2 Phase 1 Implementation - Visual Embedding and Unimodal Retrieval

Steps	Description	Details
1. Model Instantiation	Load the pre-trained CNN models as ResNet50, VGG16, and EfficientNetB0 via Keras.	Used Tensorflow.keras.applications to load ResNet50(...), VGG16(...), EfficientNetB0(...)
2. Feature Extraction Backbone	Use as a feature extractor adjust each model by stripping off the last classification layers.	Created a new Model with Input and GlobalAveragePooling2D as the final output.
3. Image Preprocessing	Specify certain preprocessing functions per each model to have correct format of images and pixel range normalized.	resnet_preprocess, vgg_preprocess, effnet_preprocess
4. Embedding Generation	Loop through the sampled dataset, load & preprocess each of the images and feed them to the feature extractor such that they generate dense embedding	Use tqdm for progress tracking. Store embeddings in a NumPy array.
5. FAISS Index Creation	Create an index on a FAISS to allow searching for nearest neighbours efficiently among the high-dimensional embedding vector	faiss.IndexFlatIP(dimension), index.add(embeddings)
6. Retrieval Function	Method where input is a query image and a FAISS index and the output would be the K most visually related images.	index.search(query_embedding, k=5)

Table 5.2. Phase 1 - Visual Embedding and Unimodal Retrieval

5.3 Phase 2 Implementation - Hybrid Multimodal Filtering

Steps	**Description**	**Details**
1. Multimodal Model Architecture	Design a new model architecture of every CNN backbone that takes as inputs an image tensor and a one-hot encoded health-tag tensor. These two inputs are merged in the process of going through their individual base layers.	tensorflow.keras.layers.Input for image_input (224, 224, 3) and tag_input (4). tensorflow.keras.layers.Concatenate() is used to fuse the outputs from the GlobalAveragePooling2D layer and a small Dense layer for the tags.
2. Multimodal Embedding Generation	Apply the new multimodal model to produce fused embedding of each item in the dataset. This needs both the image and the relevant health-fit tags fed to the model.	Function generate_multimodal_embeddings() iterates through the data and uses model.predict([batch_images, batch_tags]).
3. FAISS Indexing	Create a new FAISS index of each of the three CNN backbones with the newly generated multimodal embeddings.	build_faiss_index(embeddings) function, calls faiss.IndexFlatIP on the multimodal embeddings

4. Hybrid Retrieval Evaluation	Compare the health retrieval performance of the new multimodal embeddings (using a special-purpose metric that measures the intersection of health tags between an item of interest and its top-K retrieved neighbours).	health_tag_match_at_k(df, query_index, topk_indices, threshold=2) calculates the number of matching tags and checks against a threshold. This is called within evaluate_model_batch.
5. Batch-wise Performance Check	Mean HealthTag Match@K score is obtained on a random basis sampling of N items selected systematically out of the data set (N will be large enough to provide accurate ensemble scores, which represent the average of the scores of all the multimodal models).	evaluate_model_batch("Model Name", index, embeddings, df, samples=100). *Formula for HealthTag@K* = Number of correct health-tag matches in the top K results ÷ K
6. Per-Tag Precision Score	Finding the Precision@5 score on each of the four health tags separately in the three models to obtain a detailed insight of the performance of each.	precision_at_k(embeddings, labels, k=5) is called for each tag (arch_support, cushioning, breathability, widefeet).

Table 5.2. Phase 2 - Hybrid Multimodal Filtering

5.4 Phase 3 Implementation - Multi-Label Classification

Steps	Description	Details
1. Model Fine-Tuning	Strengthen the pre-trained CNN models with a classifier head of your own and train it on the health-fit tagged data. In this method, it takes advantage of the ability of the base model to extract visual features and learns to predict the new, domain-specific tags.	A tensorflow.keras.Model is created with a pre-trained CNN Model base, followed by GlobalAveragePooling2D and a Dense output layer with sigmoid activation for multi-label prediction. The model is compiled with binary_crossentropy loss.
2. Training with Image Generators	Use ImageDataGenerator to load and preprocess images of the training and validation DataFrames used to implement more effective creation of models without loading large batches of images in memory at the same time.	tensorflow.keras.preprocessing.image.ImageDataGenerator is configured with preprocessing_function and flow_from_dataframe to generate batches of images and their corresponding raw multi-label tags.
3. Model Evaluation	Test the model performance in the validation set in terms of standard multi-label classification metrics after the model is trained.	model.predict(val_gen) is used to get predictions. sklearn.metrics.f1_score and sklearn.metrics.classification_report are used to calculate and display the

		macro F1-score, precision, and recall for each health tag.
4. Embedding Extraction	Extract embeddings of the fine-tuned classification model last but one layer. Then, the subsequent visualization and qualitative analysis are applied to such embeddings.	A new tensorflow.keras.Model is built with the same inputs but with the output layer set to the second-to-last layer (model.layers[-2].output), which corresponds to the dense embedding layer before the final sigmoid activation.
5. Dimensionality Reduction & Visualization	Dimensionality across the fine-tuned embeddings should be reduced to 2D using UMAP and t-SNE to plot the critical data. This enables one to do a visual check of the quality of how the model has learned to cluster similar items in the latent space based on their health-fit tags.	umap.UMAP and sklearn.manifold.TSNE are used to project the embeddings. seaborn and matplotlib.pyplot are then used to create scatter plots, with points colored by their respective health tags.
6. Qualitative Retrieval Analysis	Query on the fine-tuned set and show the query image and its top K most similar images, and their health tags. This gives meaning of whether model performs well or not in a way that is interpretable by a human.	faiss.IndexFlatIP is used to create a retrieval index. show_similar_images_with_tags() queries the index and uses matplotlib to plot the query and retrieved images with their corresponding tags.

Table 5.4: Phase 3 - Multi-Label Classification

6 Evaluation

The section presents the overall analysis of the performance of the proposed system tested in various experimentation arrangements. The outcomes are also evaluated based on the pertinent measures so as to confirm the actual effectiveness of the model, as far as the research goals are concerned.

6.1 Phase 1: Visual Embedding + FAISS Retrieval

The features computed by the pretrained ResNet50, EfficientNetB0, and VGG16 models were reduced to 128-dimension using embeddings. The performance in terms of retrieval was measured in Precision@5 per health tag. Compared to the others, EfficientNetB0 performed the best especially with regards to arch support (0.7155), wide feet (0.6981) with VGG16 second best in the breathability (0.8305) and cushioning (0.6119) aspects as indicated in Table 6.1. Such results imply that the compound scaling aspect of EfficientNet has a more powerful representation capacity.

Model	Arch Support (Precision)	Cushioning (Precision)	Breathability (Precision)	Wide Feet (Precision)
ResNet50	0.6966	0.6087	0.8285	0.6876
VGG16	0.7026	0.6119	0.8305	0.6950
EfficientNetB0	0.7155	0.6160	0.8325	0.6981

Table 6.1. Phase 1- Evaluation Scores

Fig. 3. Phase 1- Visual Similarity Results

By observations, the ResNet50 and VGG16 give greater scores of cosine similarity hence the visual similarity is much better in the output displayed. Nevertheless, EfficientNetB0 is better in Precision@K, implying that it has embeddings that are not as visually tight as the others, but more functional in regard to health-fit tags. This brings out one major understanding: visual similarity and ergonomic relevance do not necessarily exhibit a direct relationship. To find health-aware recommendation, models like EfficientNetB0 may offer more tag-sensitive representations even though having slightly lower visual embedding precision.

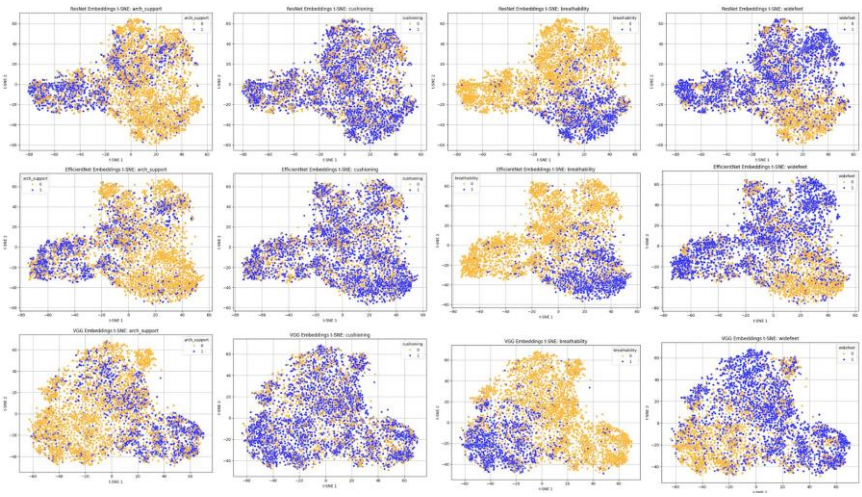
Fig.4. t-SNE Plots for Phase 1: Unimodal embeddings

The t-SNE plot in Fig. 4 shows that ResNet50 and EfficientNetB0 produce more semantically meaningful embeddings, as they can be seen in the visualized clear clustering of health-fit tags such as breathability and arch support. This implies that latent representations become increasingly more in line with functional attributes. Conversely, the VGG16 embeddings are less separable, which characterizes poor sensitivity of tags. The above observations justify downstream multimodal fusion as the architecture and embedding-space interpretability as the important evaluation metric.

6.2 Phase 2: Multimodal Embedding + Health-Tag Fusion

Multimodal fusion was introduced by concatenating CNN visual embeddings with binary health tag vectors, forming a 256-D representation. Retrieval was again performed using

FAISS (cosine similarity) and evaluated via Precision@5 and HealthTagMatch@5, defined as the average match rate of health tags between query and retrieved items across ‘100 samples.

Fig. 5. Phase 2 - Visual Similarity Results (Efficient Net)

Model	HealthTag Match@5	Arch Support (Precision)	Cushioning (Precision)	Breathability (Precision)	Wide Feet (Precision)
ResNet	0.60	0.6797	0.6050	0.8058	0.6714
VGG	0.67	0.6722	0.5999	0.8097	0.6727
EfficientNet	0.68	0.7070	0.6214	0.8230	0.7163

Table 6.2. Phase 2- Evaluation Scores

Table 6.2 indicates that the architecture EfficientNetB0-MM became the most effective one with HealthTagMatch@5 = 0.68. HealthTag@K is calculated by the number of correct health-tag matches in the top K results ÷ K. This finding depicts the capacity of the model to integrate the semantic health-tagged information into the visual search space. Both VGG16-MM and EfficientNetB0-MM showed a relatively similar outcome in improving Precision@5 compared to their unimodal base i.e. in case of tags like arch support and breathability. These gains demonstrate that multimodal fusion has the potential of improving the access to the ergonomically relevant footwear. Conversely, ResNet50-MM has caused a marginal decreased dimension in certain tag-specific precision scores implying the hypothesis that the power of fusion can differ across backbone designs. Interestingly, although ResNet50 is good at learning visual features, the multimodal version did not correspond to better alignment of health tags. Along with this, the top-5 retrieval of EfficientNetB0-M visually shows an increase in semantic alignment, or in other words, incorporates visual aesthetics similarity and health-fit tag relevance in response results. (Refer Fig 5)

The t-SNE visualization, Fig 6, showed clear clusters in the fused embeddings representing health-fit tags such as arch support and breathability. It goes to show that multimodal representations learned by EfficientNetB0-MM do not lose significant structure in the learned embedding space. The separation is more distinguishable than in the unimodal variant, which means that fusion assists in grouping the feature space to functional labels. Conversely, other models such as the ResNet50-MM had a lower cluster and retrieval diversity, which indicated the possibility of incompatibility between its visual attributes with the health tag vectors during fusion.

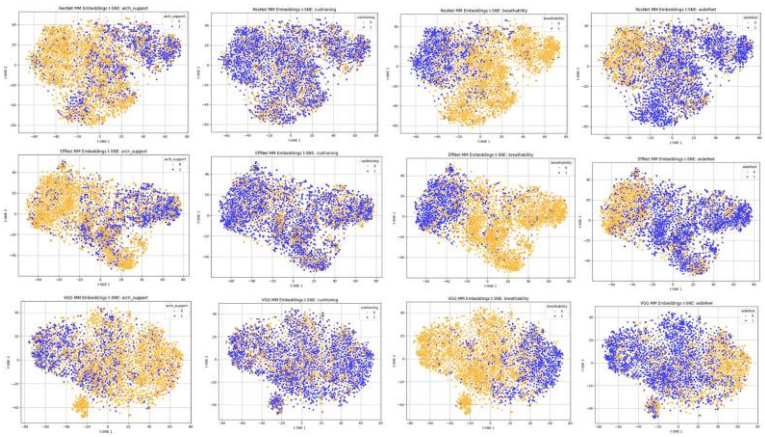

Fig 6. tSNE Plots for Phase 2 – Multimodal Embeddings

6.3 Phase 3 - Multi-label Classification Metrics

The CNN backbones (ResNet50, VGG16, EfficientNetB0) were initialized with ImageNet-pretrained weights and fine-tuned end-to-end along with a newly added dense classification head, allowing all convolutional layers to update during training for optimal adaptation to the health-fit classification task. The models were evaluated using Macro F1-score, individual tag F1, and classification reports.

Model	Macro F1	Arch Support (F1 Score)	Cushioning (F1 Score)	Breathability (F1 Score)	Wide Feet (F1 Score)
ResNet	0.771	0.741	0.699	0.812	0.831
VGG	0.809	0.758	0.829	0.817	0.833
EfficientNet	0.792	0.699	0.814	0.820	0.835

Table 6.3. Phase 3- Evaluation Scores

At the final step of multi-label classification, VGG16 was found to be the top-ranking model on Macro F1-score with 0.809 right after EfficientNetB0 with 0.792 and ResNet50 with 0.771 (refer table 6.3). VGG16, as well, scored the best per-tag F1 on cushioning (0.829) and arch support (0.809) with EfficientNetB0 again slightly ahead on wide feet (0.835) and breathability (0.820). ResNet50 failed in overall performance but showed decent results in measuring breathability (0.812) and wide feet (0.831). The findings indicate that the three models are capable of learning health-related features well, even though, VGG16 exhibited the most stably performing tags across the board.

Fig 7. Top 5 Visually Similar for Phase 3

Top-5 shoes recalled in Phase 3, Fig 7, not only have significant visual similarity with the query image but also overlap well in health-fit aspects and therefore, there is substantial convergence of visuals with utility which is satisfactory.

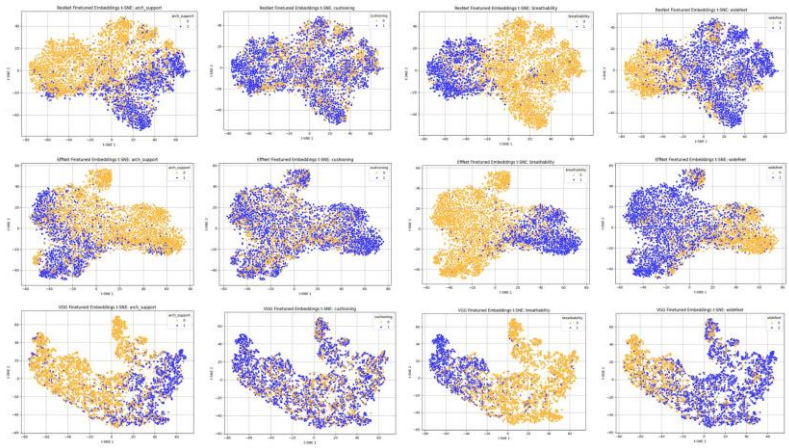

Fig 8. tSNE Plots for Phase 3 – MultiLabel Classification

From a qualitative perspective, both t-SNE, Fig. 8, and UMAP, Fig 9, visualizations of the final-layer embeddings revealed that EfficientNetB0 exhibited the clearest cluster separation for key health tags, confirming that its representations are semantically well-aligned with functional categories. VGG16 showed moderate visual grouping, supporting its strong numerical performance, while ResNet50 clusters were more diffuse, despite having high cosine similarity in Phase 1.

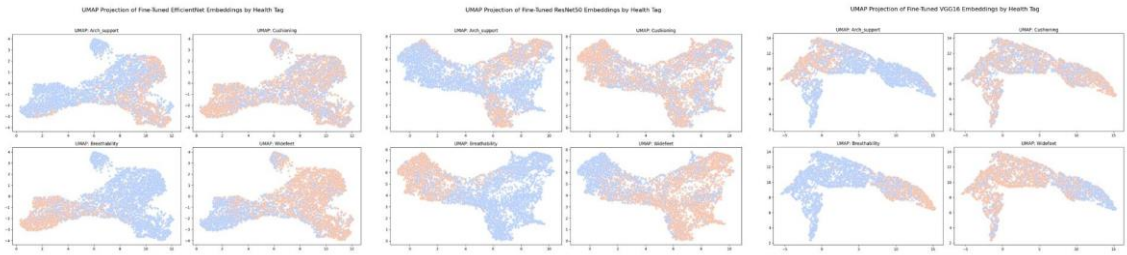

Fig 9 . UMAP Plots for Phase 3 – Multi Label Classification

Overall, while VGG16 delivered the best Macro F1-score, EfficientNetB0 model strikes the best balance between visual alignment and health-tag predictiveness, validating its suitability for ergonomic-aware footwear recommendation.

6.4 Discussion

The research aim was to understand the comparative performance of ResNet50, VGG16, EfficientNetB0 CNN architectures with respect to visual embedding that aids in recommending health-aware footwear through hybrid filtering and multi-label classification. The three experimental stages sequentially enhanced the functionalities of the system: In Phase 1, the unimodal image search was implemented according to visual similarity; in Phase 2, the multimodal embedding fusion with health-tag vectors was applied; and, finally, in Phase 3, a direct multi-label classification of health-fit attributes was accomplished. In different stages, various measuring scores and plots have been used to measure the numerical efficiency as well as semantic consistency of the embeddings.

Phase 1 showed that ResNet50 achieved the highest scores of cosine similarity values having the t-SNE clusters clear and reflecting very structural visual consistency. Nevertheless, it failed to produce the most favourable retrieval performance in respect to Precision@5 on health-fit tags. EfficientNetB0 showed the best Precision@5 on arch support and

breathability, even though its cosine scores were a little lower, which may imply that it represented a higher degree of functional similarity in its embeddings compared with the visually similar ones. The given finding aligns with the previous study of Suvarna et al. (2024) which underlined the trade-off of EfficientNetB0 between efficiency and semantic relevance. In visual clustering and precision results, VGG16, on the other hand, underperformed relative to the others, overall, on representation learning.

Phase 2 involved a multimodal strategy, in which the visual embedding was added together with bitwise health-tag vectors to create a 256-dimensional representation. In this case, EfficientNetB0-MM proved to be the most effective achieving the best HealthTagMatch@5 of 0.68. It resulted being effective showing good results in Precision@5 in the following categories arch support (0.7070), breathability (0.8230) and wide feet (0.7163) The separation of clusters in its t-SNE plots was better in comparison to its unimodal version, which proved the addition of more semantic information by the fusion. Interestingly, ResNet-MM only experienced a marginal or in some tags, negative improvements, indicating that simple concatenation of vectors is not necessarily beneficial to all architecture choices. That means that more complex fusion strategies, say attention mechanisms or comparison layers, that lead to potentially more consistency across modalities would be necessary. The findings strengthen the results of (Baldrati *et al.*, 2021) and define that efficient tag-aware retrieval is confined not only to powerful architecture but also to compatibility of input modalities.

Phase 3 progressed to direct classification that makes use of a fine-tuned CNN and dense layers to predict multiple tags of health-fit. The evaluation showed that VGG16 surprisingly delivered the highest Macro F1-score (0.809), driven largely by its exceptional performance on cushioning (F1: 0.829). EfficientNetB0, while scoring slightly lower on macro average, achieved the best F1 for wide feet and breathability, the two most ergonomically significant tags. This is consistent with earlier studies conducted by (Liu *et al.*, 2024), who pointed out the importance of having lightweight models to strike a balance in various product qualities. The consistent pattern in the plot of UMAP and t-SNE also affirmed that EfficientNetB0 and ResNet50 embeddings have robust discrimination at a tag level, particularly arch support and breathability. VGG16, despite its numerical power, exhibited less clear-cut separation in visualizations- which implies that its decisions depend on memorization rather than the structured feature representations.

Although the results are quite encouraging, the design of this system is not without limitations. In the process of Phase 2, the fusion process only depended on simple concatenation without the weighting and learning of interaction between visual and tag modalities. Also, health-tag alignments were evaluated strictly by rule and given as a binary, whereas in the future, the model can involve user preferences or a grading system to be more applicable to the more nuanced context in the world. The sampled approach is associated with a stratified sampling technique but a few categories of outliers or combination of tags that would not be presented might have affected generalizability of the model.

This project makes some contributions as viewed in the literature as a whole. It offers a complete end-to-end benchmarking setting throughout CNNs in either aesthetic or ergonomic advice, composes clarity-related benchmarking into contact diagnostics using t-SNE and UMAP, and the significance of coordinating visual likeness with functional significance. The results validate and generalize the previous efforts on fashion embeddings (Alishev, 2024); (R. *et al.*, 2024) and, in particular, cut a niche in health-friendly personalization, which is a relatively unexplored aspect.

Generally, there is room to improve or rather refine the models despite the good match between the result in visual embeddings and health-fit tags provided. Future versions may consider transformer-based architectures, learnable fusion layers or user studies to test on the satisfaction of recommendations. However, the staged design and overall evaluation used in the present work forms a strong basis in terms of health-centric fashion recommendation systems.

7 Conclusion and Future Work

The study was largely successful in achieving its objectives. Within the assessed models, EfficientNetB0 was always the most balanced in terms of good Precision@5, good Macro F1 scores, and clear clusters than others. Multimodal testing showed that simple techniques such as concatenating vectors can be used to successfully project functional semantics into the space of visual retrieval as well. Interestingly, VGG16 gave the best performance on direct tag prediction but its embedding clusters were less distinct, which implies its overfitting tendency. Conversely, ResNet50 had the highest visual continuity of embeddings, but its tag-accuracy was also lower, indicating a trade-off between visual coherence of embeddings and specifically-relevant accuracy.

For future, attention-based models, cross-modal transformers, and joint embedding projections are more expressive fusion mechanisms that can be used to enrich the interactions between visual and health tag modalities. Defining the graded health scores rather than binary labels, including real user comments, or even using wearable sensor data, would further personalize and refer the system to the real-world needs. In the business sector, this project can by extension be the foundation of a possible mobile application or web-based platform that enables the user to leave an image of a shoe or a foot profile and receive a set of recommendations in the order of style as well as ergonomic appropriateness. The inclusion of a user satisfaction feedback loop had the potential to turn the system into a highly adaptive health-sensitive recommendation assistant, as well.

References

Abdulnabi, A.H. et al. (2015) ‘Multi-task CNN model for attribute prediction’, IEEE Transactions on Multimedia, 17(11), pp. 1949–1959. doi:10.1109/TMM.2015.2477680.

Alcacer, A. et al. (2021) ‘Combining classification and user-based collaborative filtering for matching footwear size’, Mathematics, 9(7), p. 771. doi:10.3390/math9070771.

Alishev, A. (2024) Transfer learning and hyperparameter optimisation with convolutional neural networks for fashion style classification and image retrieval. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-225621 (Accessed: 10 August 2025).

Baldrati, A. et al. (2021) ‘Conditioned image retrieval for fashion using contrastive learning and CLIP-based features’, ACM Multimedia Asia. MMAsia ’21: ACM Multimedia Asia, Gold Coast, Australia: ACM, pp. 1–5. doi:10.1145/3469877.3493593.

Chang, C.-C. et al. (2023) ‘A shoe shopping system based on convolutional neural network image recognition’, 2023 IEEE 5th Eurasia Conference on IoT, Communication and Engineering (ECICE), Yunlin, Taiwan: IEEE, pp. 305–309. doi:10.1109/ECICE59523.2023.10383162.

He, K. et al. (2016) ‘Deep residual learning for image recognition’, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, pp. 770–778. doi:10.1109/CVPR.2016.90.

Johnson, J., Douze, M. and Jégou, H. (2017) ‘Billion-scale similarity search with GPUs’, arXiv. doi:10.48550/arXiv.1702.08734.

Kobak, D. and Berens, P. (2019) ‘The art of using t-SNE for single-cell transcriptomics’, Nature Communications, 10(1), p. 5416. doi:10.1038/s41467-019-13056-x.

Li, B. et al. (2023) ‘Application of deep learning neural networks in sports shoe brand classification task’, 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI), Guiyang, China: IEEE, pp. 152–158. doi:10.1109/ICCBD-AI62252.2023.00034.

Liu, H.-I. et al. (2024) ‘Lightweight deep learning for resource-constrained environments: A survey’, arXiv. doi:10.48550/arXiv.2404.07236.

R., P.G. et al. (2024) ‘Recent innovations in footwear sensors: Role of smart footwear in healthcare – A survey’, arXiv. doi:10.48550/arXiv.2402.01645.

Shen, E. et al. (2023) ‘ResNet50-based classification of footwear in nuclear power plants surveillance images’, 2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Taizhou, China: IEEE, pp. 1–5. doi:10.1109/CISP-BMEI60920.2023.10373252.

Simonyan, K. and Zisserman, A. (2014) ‘Very deep convolutional networks for large-scale image recognition’, arXiv. doi:10.48550/arXiv.1409.1556.

Stanković, K. et al. (2018) ‘Three-dimensional quantitative analysis of healthy foot shape: A proof of concept study’, Journal of Foot and Ankle Research, 11(1), p. 8. doi:10.1186/s13047-018-0251-8.

Suvarna, B. et al. (2024) ‘Optimizing footwear image classification with hyperparameter tuning: Insights from the UTZappos dataset’, 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), Gwalior, India: IEEE, pp. 1034–1039. doi:10.1109/AIC61668.2024.10731099.

Tan, M. and Le, Q.V. (2019) ‘EfficientNet: Rethinking model scaling for convolutional neural networks’, arXiv. doi:10.48550/arXiv.1905.11946.

Umar, H. et al. (2022) ‘Impact of routine footwear on foot health: A study on plantar fasciitis’, Journal of Family Medicine and Primary Care, 11(7), pp. 3851–3855. doi:10.4103/jfmpc.jfmpc_637_21.

Yu, A. and Grauman, K. (2017) ‘Semantic jitter: Dense supervision for visual comparisons via synthetic images’, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 2017, pp. 5570–5579.