DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN

ICT FOR INTERNET AND MULTIMEDIA

Valence and Arousal prediction from food images using Deep Learning and classical Machine Learning models

Supervisor: Prof. Antonio Roda

Co-supervisor: Dr. Matteo Spanio

Laureando: Fatemeh Talebi

November 17, 2025

Abstract

In this work, we present a computational framework for predicting emotional dimensions—valence and arousal—from visual representations of food imagery. The study aims to investigate how visual cues such as color, brightness, texture, and spatial composition influence the affective perception of food, and how these cues can be quantitatively modeled using machine learning techniques. Leveraging a dataset of 1,211 annotated food images with continuous valence–arousal labels, we extracted high-level visual embeddings using a pre-trained Vision Transformer (ViT)[1] model, which effectively captures semantic and aesthetic characteristics of the images.

Three regression models— Random Forest (RF), Support Vector Regression (SVR), and a Multi-Layer Perceptron (MLP)—were implemented to evaluate the relationship between extracted features and emotional responses. All models were trained under identical preprocessing and standardization procedures to ensure fairness and reproducibility. Quantitative evaluation was performed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) metrics. The experimental results revealed that the MLP model achieved the most stable convergence and the highest predictive accuracy, outperforming classical regressors. Extending the training duration from 100 to 500 epochs further improved performance consistency, demonstrating that longer optimization can enhance emotional representation learning without causing overfitting.

Nevertheless, several challenges were observed, including limited dataset size, uneven emotional intensity distribution, and model sensitivity to ambiguous or low-contrast food images. These factors occasionally led to larger prediction errors, especially for samples with moderate emotional content. Despite these limitations, the findings underscore the potential of deep learning models—particularly neural architectures combined with transformer-based feature extraction—to effectively capture the nuanced relationship between food aesthetics and affective perception. This research establishes a reproducible baseline for visual emotion prediction and opens pathways for future work on larger datasets, multimodal fusion, and interpretable emotion modeling in affective computing.

Keywords: Affective Computing, Emotion Prediction, Food Images, Valence–Arousal Model, Vision Transformer (ViT), Random Forest, Support Vector Regression, Multi-Layer Perceptron

Contents

1 Introduction

1.1 Background and Context

In recent years, affective computing has emerged as a powerful interdisciplinary field focused on enabling machines to recognize, interpret, and model human emotions. One of the most effective frameworks for representing emotions is the **valence–arousal model**, which expresses affective states within a continuous two-dimensional space—valence reflecting the positivity or negativity of emotion, and arousal representing the level of activation or intensity. This dimensional representation allows computational models to predict nuanced emotional responses beyond categorical emotion labels such as "happy" or "sad."

In visual domains, emotion prediction has primarily been explored in areas such as facial expression recognition, artistic image interpretation, and advertising analytics. However, relatively few studies have examined how **food imagery**—a domain rich in color, texture, and spatial arrangement—elicits emotional reactions. The emotional response to food imagery plays an important role in marketing, menu design, human–computer interaction, and even in health and psychological studies related to appetite and mood regulation.

This research seeks to bridge the gap between visual perception and affective interpretation by investigating how computational models can predict valence and arousal values directly from food images. To achieve this, we employ a combination of classical and deep learning–based regression models trained on visual feature representations extracted using a pre-trained **Vision Transformer (ViT)**[1] model.

1.2 Visual Emotion Analysis and the Valence–Arousal Space

Emotion recognition from visual cues is grounded in psychological models such as Russell's circumplex theory[2] of affect. In this model, emotions are distributed within a circular plane defined by two orthogonal dimensions—valence and arousal. For instance, bright and saturated images with warm tones are often perceived as high-valence and high-arousal, whereas darker, desaturated visuals typically evoke lower valence and arousal. This provides a robust basis for computational modeling, where visual attributes are mapped to quantitative emotion scores.

1.3 Food Imagery and Emotional Perception

Food imagery offers a compelling testbed for affective prediction due to its aesthetic richness and strong emotional associations. Empirical studies have shown that certain visual properties—such as brightness, symmetry, and color harmony—strongly influence perceived pleasantness and excitement. Vibrant dishes with warm hues often induce feelings of happiness and energy (high valence, high arousal), while muted or less organized compositions may evoke calmness or even aversion (low valence, low arousal). Understanding these map-

pings not only benefits psychological research but also aids in practical domains like digital marketing, culinary design, and social media engagement analysis.

1.4 Machine Learning for Emotion Prediction

The recent advances in machine learning have enabled data-driven modeling of emotion–feature relationships. Classical models such as **Random Forest (RF)**[3] and **Support Vector Regression (SVR)**[4] provide interpretable baselines that can effectively model nonlinear relationships within moderate-sized datasets. In contrast, deep neural models such as the **Multi-Layer Perceptron (MLP)**[5] offer hierarchical representation learning, capturing complex feature interactions that classical methods may overlook.

In this study, all models are trained using standardized 768-dimensional feature vectors obtained from the Vision Transformer. This ensures consistency and comparability across different regression approaches, allowing a comprehensive evaluation of both traditional and neural paradigms in affective prediction.

1.5 Dataset and Annotation Framework

The dataset used in this research consists of 1,211 food images annotated with continuous valence and arousal scores in the range [0,1]. These labels were derived from perceptual studies in which human participants rated the emotional impact of each image based on the circumplex model. Each image was processed through the Vision Transformer to extract high-level visual embeddings representing color harmony, brightness, texture, and compositional balance. These embeddings form the numerical input features for all regression models.

To ensure fair comparison, the dataset was divided into 80% training and 20% testing partitions, with all feature dimensions standardized using z-score normalization. This controlled preprocessing guarantees consistent learning dynamics across experiments and prevents scale-dependent bias in model optimization.

1.6 Motivation and Significance

The motivation for this study stems from the increasing importance of emotional intelligence in computational systems. While existing works have explored emotion prediction from facial or artistic images, the affective potential of food imagery remains underexplored. Yet, food-related visuals are among the most emotionally charged stimuli in digital media, influencing user perception, appetite, and decision-making.

By modeling the relationship between visual composition and emotional response, this research aims to advance both theoretical understanding and practical applications of affective computing. The findings have potential implications for emotion-aware recommendation systems, personalized nutrition interfaces, and automated food photography evaluation.

1.7 Motivation of Study

Despite the progress in visual affect recognition, most existing methods focus on human faces or abstract artworks rather than everyday visual stimuli such as food. This creates a gap in understanding how aesthetic and perceptual qualities of food contribute to emotional responses. Moreover, while classical machine learning models provide interpretability, they may lack the expressive capacity needed to model complex nonlinear emotional mappings. Deep learning architectures, on the other hand, can learn these intricate dependencies but require rigorous evaluation to ensure stability and generalization.

The motivation behind this study is therefore twofold: to investigate whether machine learning can accurately predict emotional valence and arousal from static food images, and to assess how model complexity (from RF and SVR to MLP) influences predictive accuracy and interpretability.

1.8 Problem Statement

The central problem addressed in this research is the development of a reliable computational framework for predicting continuous emotional dimensions—valence and arousal—from visual representations of food. Key challenges include the limited availability of annotated affective datasets for food images, variability in emotional perception among individuals, and the nonlinear nature of relationships between low-level visual features and high-level emotional constructs. Addressing these challenges requires careful data preprocessing, standardized training procedures, and systematic model comparison.

1.9 Research Objectives

The objectives of this study are summarized as follows. First, to design and implement multiple regression-based emotion prediction models, including Random Forest, Support Vector Regression, and Multi-Layer Perceptron. Second, to perform a comprehensive quantitative evaluation of each model's predictive performance using standard metrics such as MAE, MSE, and R^2. Third, to visualize and analyze model behavior through training loss curves and predicted–actual scatter plots. Fourth, to explore the effects of extended training duration on convergence stability and accuracy. Finally, to establish a reproducible experimental framework that can serve as a benchmark for future research in visual emotion prediction.

2 Related Work

2.1 Emotion Recognition in Images

Emotion recognition from visual information has been one of the central challenges in affective computing and computer vision over the past two decades. The ability to infer human affective states from images enables more natural and emotionally intelligent interactions between humans and machines. Early research in this domain primarily focused on categorical emotion recognition—classifying images into discrete categories such as "happy," "sad," "fearful," or "angry." However, categorical approaches often fail to capture the full spectrum of affective nuances, leading to the adoption of continuous emotion models such as the **valence–arousal (V–A) space**. In this dimensional model, valence measures emotional positivity or pleasantness, while arousal quantifies activation or intensity.

2.1.1 From Facial Expression Analysis to General Visual Emotion

The first major wave of visual emotion recognition research concentrated on facial expression analysis. Pioneering works by Ekman and Friesen[6] established the universality of basic emotions and laid the foundation for automatic emotion detection from facial landmarks. Subsequent computational models employed handcrafted features—such as Gabor[7] filters, Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG)—to identify expressions from facial regions.

While these methods achieved considerable success in controlled settings, their effectiveness was limited to explicit facial cues and could not generalize to complex visual scenes or abstract imagery. As a result, the research focus gradually shifted from facial emotion recognition to **general visual emotion recognition (VER)**—the task of predicting emotions evoked by arbitrary images, including art, landscapes, objects, and food.

2.1.2 Transition to Affective Image Understanding

The concept of affective image understanding extends beyond facial cues to the analysis of affective semantics conveyed by colors, shapes, lighting, and composition. Machajdik and Hanbury [8] were among the first to explore emotional content in general images by extracting low-level features such as color harmony, brightness, and texture, combined with high-level features like composition and semantics. Their model classified images according to Ekman's six basic emotions, providing a foundation for future visual affect research.

Following this work, several datasets and benchmarks for affective image analysis emerged, such as *ArtPhoto*, *IAPS-Subset*, and *FI Dataset* [9]. These datasets enabled machine learning models to be trained on diverse emotional stimuli, ranging from artistic paintings to real-world photographs.

2.1.3 Dimensional Emotion Representation

Although categorical labeling remains intuitive, the discrete emotion model has intrinsic limitations—particularly when multiple emotions coexist or when the intensity of an emotion varies continuously. Dimensional representations like the valence–arousal–dominance (VAD) model overcome these issues by capturing emotions as points in a continuous space. Studies by Koelstra[10] et al. and Wang et al. demonstrated that using continuous labels not only improves model expressiveness but also aligns more closely with human emotional perception.

In this representation, images with high saturation and brightness typically correspond to high valence and high arousal, while darker or desaturated visuals indicate lower emotional responses. This mapping between low-level visual features and affective dimensions has become a cornerstone of modern affective image analysis.

2.1.4 Feature Evolution: From Handcrafted to Deep Representations

The progression from handcrafted features to deep representations revolutionized emotion recognition. Early approaches relied heavily on predefined visual cues such as edge density, hue distribution, or color temperature, which limited generalization. With the introduction of deep convolutional neural networks (CNNs), researchers could learn hierarchical visual representations directly from data, allowing models to capture both semantic and aesthetic cues.

Recent advancements, particularly with transformer-based architectures, have further enhanced this capability. Vision Transformers (ViT) introduced global self-attention mechanisms that enable models to encode long-range dependencies and contextual relationships within images—an essential aspect when predicting emotional content that depends on holistic composition rather than local details. This overall evolution in feature extraction and model design is summarized in Figure 1, which illustrates the transition from handcrafted color-based features to CNN, attention-based, and transformer architectures such as ViT.

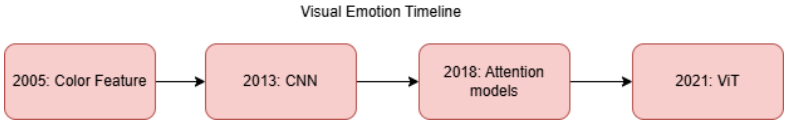

Figure 1: Evolution of visual emotion recognition from early handcrafted feature-based methods to modern deep and transformer-based models.

2.1.5 Challenges in Visual Emotion Prediction

Despite significant progress, visual emotion prediction remains a challenging task due to the subjective and context-dependent nature of emotions. Two visually similar images can elicit different emotions depending on cultural background, personal experiences, or situational context. Moreover, the lack of

large-scale annotated datasets with continuous valence–arousal ratings limits the robustness of existing models. Handling such subjectivity requires not only high-quality data but also architectures capable of learning subtle and abstract affective cues embedded within visual scenes.

These challenges motivate ongoing research into hybrid models, transfer learning strategies, and multimodal emotion prediction frameworks that integrate textual or contextual information. The next sections review these developments in greater detail, focusing on datasets, architectures, and regression-based modeling techniques used in affective image analysis.

2.2 Visual Emotion Datasets

A critical factor influencing progress in visual emotion recognition is the availability of well-annotated datasets that capture affective variations across diverse visual stimuli. Emotion, by nature, is subjective and context-dependent; thus, building large-scale and reliable datasets has posed unique challenges to researchers. In this section, we review the most widely used datasets that have shaped the development of affective computing models, from early psychological corpora to modern large-scale visual emotion collections.

2.2.1 Early Emotion Datasets: From Psychology to Visual Media

The earliest datasets for emotion research were grounded in psychology and neuroscience, focusing primarily on eliciting affective responses through controlled stimuli such as facial expressions or standardized photographs. One of the most influential among these is the **International Affective Picture System (IAPS)** [11], which contains over 1,000 images rated by participants along the dimensions of valence, arousal, and dominance. While IAPS has served as a foundational dataset for emotion studies, its limited size and restricted access have made it less suitable for modern machine learning pipelines that require thousands of samples.

Following IAPS, other affective corpora such as the **Geneva Affective Picture Database (GAPED)** [12] expanded image diversity to include animals, nature, and social scenes. These databases emphasized the dimensional representation of affect (valence–arousal) and provided standardized ground truth ratings from human participants, laying the groundwork for subsequent computational models.

2.2.2 Affective Datasets for Artistic and Natural Images

The shift from psychology-oriented datasets to computational vision led to the emergence of affective datasets capturing a broader range of real-world and artistic imagery. Among the earliest examples, the **ArtPhoto** dataset [8] consists of 806 artistic images labeled with Ekman's six basic emotions: happiness, sadness, fear, disgust, anger, and surprise. The **Abstract Paintings** dataset

followed a similar structure, focusing on color and composition to study the link between visual aesthetics and emotion.

The **FI Dataset** (Flickr and Instagram) introduced by You et al. [9] represented a major advancement by collecting more than 23,000 social media images annotated via crowdsourcing. This dataset reflects the complexity of human affective perception in unconstrained settings, as it includes diverse content such as food, landscapes, objects, and human interactions. The FI dataset has since become one of the most popular benchmarks for visual emotion recognition in the wild. Examples from some of the most widely used visual emotion datasets are illustrated in Figure 2, showing how images are labeled based on valence and arousal or discrete emotion categories.

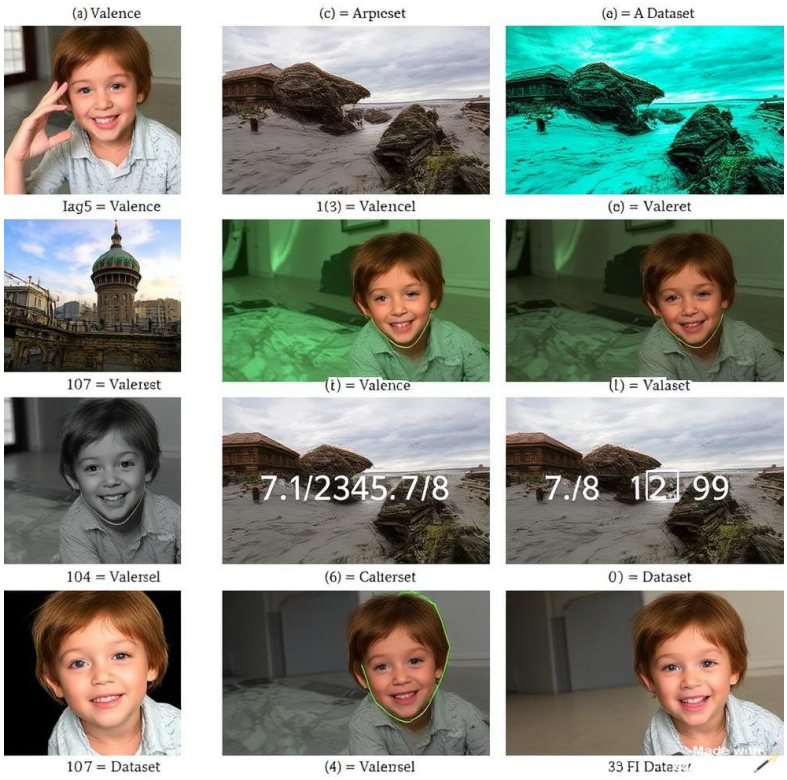

Figure 2: Examples of images from commonly used visual emotion datasets including IAPS, ArtPhoto, and FI Dataset. Each image is annotated according to valence and arousal ratings or categorical emotion labels.

2.2.3 Dimensional Datasets and Continuous Ratings

While categorical emotion labeling remains common, recent research has gravitated toward continuous dimensional datasets that better capture subtle affective variations. Datasets such as **DEAP** [10], originally designed for multimedia

affect analysis, incorporate physiological and behavioral signals along with valence–arousal scores. Similarly, the **LIRIS-ACCEDE** dataset [13] contains over 9,800 video clips annotated with continuous valence and arousal values, allowing researchers to train regression-based emotion models rather than simple classifiers.

These datasets have enabled a paradigm shift from discrete classification to **affective regression**, where models predict continuous values of emotional intensity. This approach is particularly relevant to the current study, which also employs regression models to estimate valence and arousal from visual features of food images.

2.2.4 Comparative Summary of Visual Emotion Datasets

To provide a concise overview, Table 1 compares the key characteristics of several major visual emotion datasets, highlighting their scale, labeling methodology, and emotional representation format. A detailed comparison of these datasets in terms of scale, labeling approach, and emotional representation is presented in Table 1.

Table 1: Summary of major visual emotion datasets used in affective computing.

Dataset	Samples	Emotion Type	Representation	Annotation Method
IAPS [11]	∼1,000	General scenes	Valence–Arousal–Dominance	Psychological rating
GAPED [12]	730	Animals, nature, objects	Valence–Arousal	Participant surveys
ArtPhoto [8]	806	Artistic images	Categorical (6 emotions)	Expert labeling
FI Dataset [9]	23,000	Social media content	Categorical	Crowdsourcing
LIRIS-ACCEDE [13]	9,800 videos	Dynamic visual scenes	Valence–Arousal	Continuous annotation
DEAP [10]	1,200 trials	Audio–visual stimuli	Valence–Arousal	Physiological + subjective

2.2.5 Limitations of Existing Datasets

Despite the breadth of these datasets, several limitations remain. Many datasets are either too small or context-specific to support robust deep learning. Others suffer from noisy or inconsistent annotations due to subjectivity in emotional perception. Moreover, the majority of datasets focus on human faces, abstract art, or social imagery, with comparatively little attention to domains such as food, design, or product aesthetics—areas where emotional response is deeply intertwined with visual composition.

The scarcity of high-quality datasets containing affective ratings for food images motivated the development of the dataset used in this thesis. It provides a foundation for studying emotional prediction in a specialized visual domain, bridging the gap between computational aesthetics and affective perception.

2.3 Emotion Representation Models and Theoretical Foundations

A fundamental aspect of affective computing is the theoretical modeling of emotions. Since emotions are inherently subjective, ambiguous, and multidimen-

sional, researchers have developed several frameworks to describe, categorize, and quantify them. These models allow computational systems to interpret emotions in a structured and measurable way. Broadly, emotion representation approaches can be divided into two families: **categorical models**, which define discrete emotional states, and **dimensional models**, which represent emotions along continuous scales such as valence and arousal.

2.3.1 Categorical Emotion Models

The categorical approach assumes that human emotions can be represented as a finite set of basic and universally recognized categories. One of the most influential models is Paul Ekman's **Basic Emotion Theory**, which identifies six primary emotions—happiness, sadness, fear, disgust, anger, and surprise. Ekman's model was validated through cross-cultural psychological studies showing that these emotional expressions are universally recognized across diverse human populations.

In computer vision and affective computing, categorical models have been used extensively for emotion recognition from facial expressions, gestures, and images. The advantage of categorical models lies in their interpretability: each emotion corresponds to a distinct label. However, this simplicity comes at the cost of granularity. Emotions are rarely discrete; instead, they exist on a continuum, with subtle variations that categorical systems cannot fully represent. These six fundamental emotions and their relationships are visually represented in Figure 3, which illustrates Ekman's model of basic emotions.

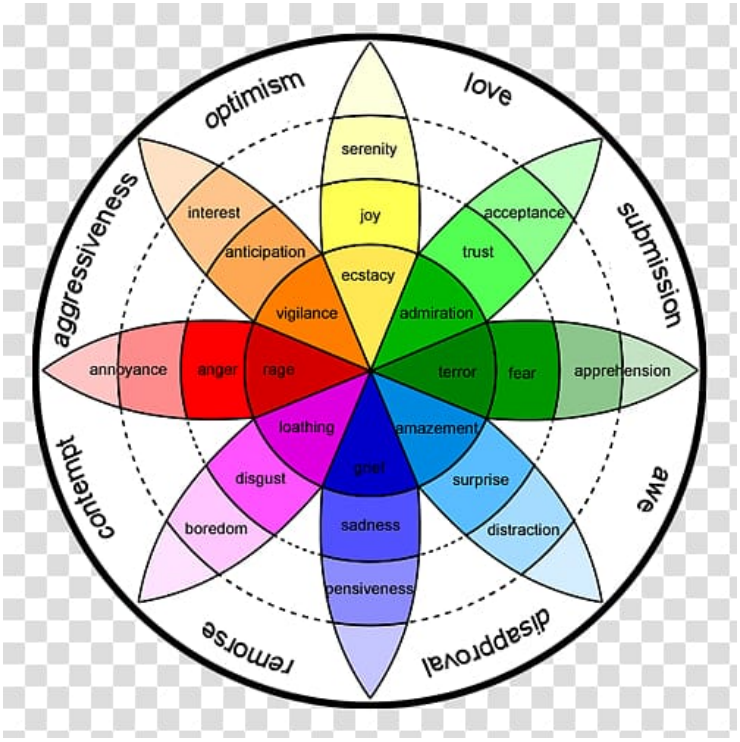

Figure 3: Illustration of Ekman's six basic emotion categories: happiness, sadness, fear, disgust, anger, and surprise.

2.3.2 Dimensional Emotion Models

To overcome the limitations of discrete categories, researchers proposed continuous representations where emotions are expressed in a multidimensional space. The most prominent of these frameworks is the **Circumplex Model of Affect** introduced by Russell. This model arranges affective states in a two-dimensional circular structure defined by:

- **Valence:** The degree of pleasantness or positivity of an emotion (ranging from unpleasant to pleasant).
- **Arousal:** The level of physiological or psychological activation (ranging from calm to excited).

Within this space, emotions can be represented as coordinates; for example, joy corresponds to high valence and high arousal, while sadness represents low valence and low arousal. This dimensionality provides a continuous and flexible representation that better captures the complexity of human emotional experiences. As shown in Figure 4, the Circumplex Model of Affect by Russell (1980) organizes emotions within a two-dimensional circular space defined by valence

(ranging from unpleasant to pleasant) and arousal (ranging from calm to excited). This visualization illustrates how emotional states transition smoothly along these axes, capturing the continuous nature of affective experience.

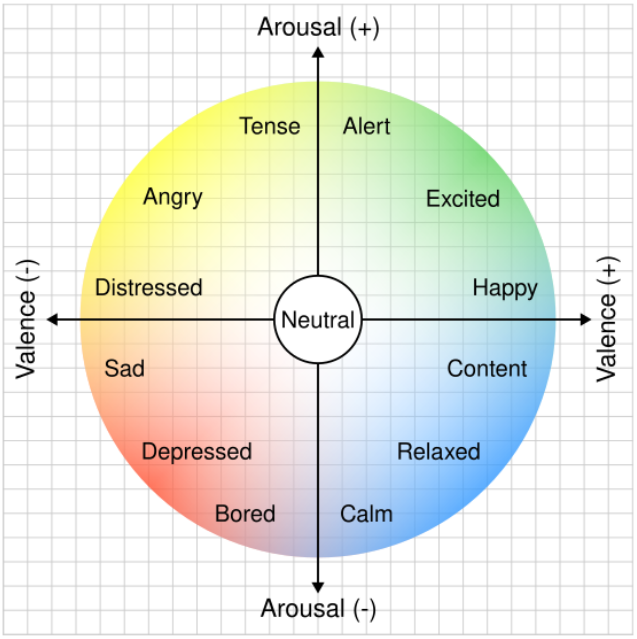

Figure 4: The circumplex model of affect, mapping emotional states within valence–arousal space.

2.3.3 The PAD Model: Valence, Arousal, and Dominance

A further extension of the circumplex framework is the **PAD model** developed by Mehrabian and Russell, which adds a third dimension—*dominance*—representing the degree of control or power associated with an emotion. In this model:

- **Valence (P):** Measures pleasure or displeasure.
- **Arousal (A):** Measures excitement or calmness.
- **Dominance (D):** Measures the level of control or submissiveness.

The PAD model provides a richer structure for emotion representation, useful in contexts such as personality modeling, multimedia affect analysis, and social robotics. Although dominance is less frequently used in visual emotion prediction due to annotation difficulty, valence and arousal remain the two most widely adopted axes for continuous affective modeling.

2.3.4 Mapping Emotions to Visual Stimuli

When applied to computer vision, dimensional emotion models enable the mapping of visual properties to affective states. For instance:

- Bright, warm, and saturated colors are often associated with high valence and arousal.
- Dark or desaturated tones typically correspond to low valence and arousal.
- Strong contrast or dynamic composition can evoke higher arousal, while balanced symmetry and soft edges often correlate with calmness.

These correlations have been validated in numerous perceptual studies and serve as the foundation for feature-based emotion prediction from images. The integration of these psychological insights with deep visual representations allows for the construction of regression models that predict continuous affective dimensions such as valence and arousal from image embeddings.

2.3.5 Advantages of Dimensional Representations

The dimensional representation has become the preferred framework in affective computing for several reasons:

1. It captures emotional subtleties and mixtures that categorical models cannot express.
2. It enables mathematical modeling and regression-based prediction.
3. It facilitates cross-domain comparison between modalities (e.g., image, audio, text) through shared valence–arousal scales.

Consequently, most contemporary studies, including this thesis, adopt the valence–arousal model as the primary framework for emotion representation. This structure aligns well with machine learning objectives, where continuous prediction enables fine-grained emotional inference from complex visual stimuli. As illustrated in Figure 5, the valence–arousal model can also be visualized using facial expressions, where each quadrant corresponds to a distinct emotional state. For instance, “happy” and “elated” occupy the high-valence/high-arousal region, while “sad” and “bored” appear in the low-valence/low-arousal quadrant. This visualization highlights how continuous dimensions effectively capture subtle affective transitions between emotional states.

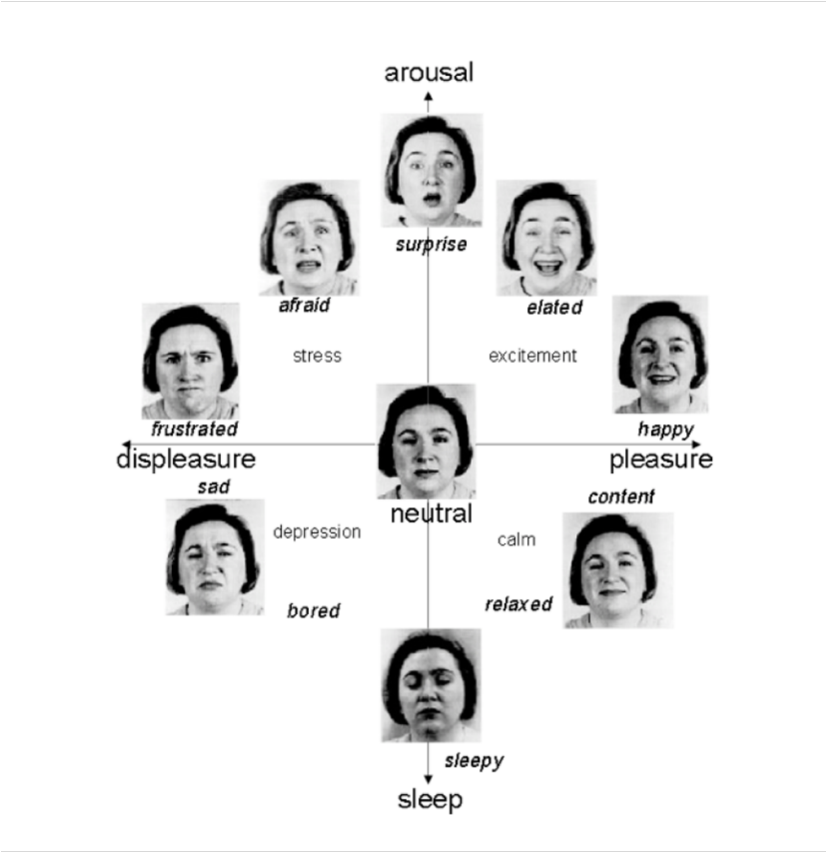

Figure 5: Examples of visual mapping of emotions within valence–arousal space. High-valence images (e.g., colorful dishes, sunlight) contrast with low-valence ones (e.g., dull lighting, poor presentation).

In summary, emotion representation models provide the theoretical foundation upon which affective visual computing systems are built. The shift from categorical to dimensional representations has enabled more nuanced, flexible, and computationally tractable approaches—paving the way for regression-based modeling strategies, as explored in subsequent chapters of this work.

2.4 Deep Learning Approaches for Emotion Recognition

The emergence of deep learning has revolutionized the field of affective computing by enabling automatic extraction of high-level visual features that correlate with human emotions. Unlike traditional methods relying on handcrafted features such as color histograms or Gabor filters, deep neural networks learn hierarchical representations directly from data, providing a more flexible and robust approach to visual emotion understanding.

2.4.1 From Handcrafted to Learned Representations

Before the deep learning era, emotion recognition heavily depended on feature engineering. Researchers manually designed descriptors like color temperature, saturation, brightness, or texture to approximate affective cues. While these approaches provided initial insights, they suffered from two major limitations: (1) handcrafted features lacked generalization across diverse datasets, and (2) they failed to capture complex semantic information, such as context or object relationships within a scene.

The introduction of **Convolutional Neural Networks (CNNs)** marked a paradigm shift. CNNs can automatically learn hierarchical visual patterns — from low-level edges and textures to high-level semantic concepts such as objects and scenes — directly from raw image pixels. This end-to-end learning capability proved to be highly effective for affective analysis, where emotional meaning often emerges from the interplay of multiple visual cues. As shown in Figure 6, convolutional neural networks (CNNs) form the backbone of modern affective computing pipelines. These models automatically extract hierarchical visual features from raw image pixels — progressing from low-level patterns (edges, textures) to high-level affective concepts — and map them to continuous valence–arousal outputs. This end-to-end framework eliminates the need for manual feature engineering and enables more robust emotion prediction.

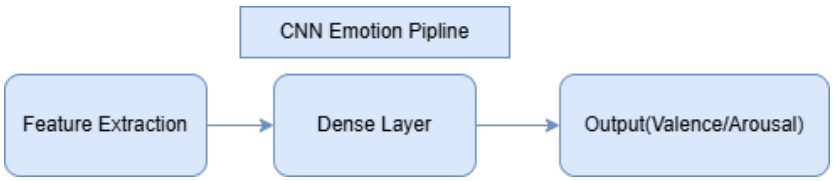

Figure 6: A typical CNN-based emotion recognition pipeline: from raw image input to feature extraction and regression/classification in valence–arousal space.

2.4.2 CNN-Based Models for Visual Emotion Analysis

The first wave of deep learning models for emotion recognition employed classical CNN architectures such as AlexNet, VGGNet, and ResNet. For example, Xu et al. [14] fine-tuned VGG-16 on the FI Dataset to classify social media images into Ekman's six emotions. Similarly, You et al. [9] used a hybrid CNN–SVM model, where CNN features were extracted and fed into a support vector regressor to predict valence and arousal scores.

These CNN-based methods demonstrated that deep representations could outperform traditional features in both categorical and dimensional emotion recognition. However, they still faced challenges with interpretability and sensitivity to non-emotional visual factors such as background clutter or illumination.

2.4.3 Incorporating Context and Attention Mechanisms

Subsequent research introduced **attention mechanisms** to improve model focus on emotionally relevant regions of an image. Works such as Zhao et al. [15] proposed **Affective Attention Networks (AAN)**, which adaptively weighted spatial features according to their emotional salience. Similarly, Xu et al. introduced dual-branch networks combining global context and localized attention maps, enabling finer emotion predictions.

These developments highlighted that emotion is not uniformly distributed across an image — rather, certain regions (e.g., facial expressions, food composition, or color-dominant areas) carry stronger affective signals. Attention-based CNNs thus became an important step toward interpretable affective models. As illustrated in Figure 7, attention mechanisms enable CNNs to selectively focus on emotionally salient regions within an image. Instead of treating all visual areas equally, the network assigns higher weights to regions that contribute more strongly to emotional perception—such as vibrant colors, expressive shapes, or visually dominant objects. This selective focus improves interpretability and enhances affective prediction accuracy.

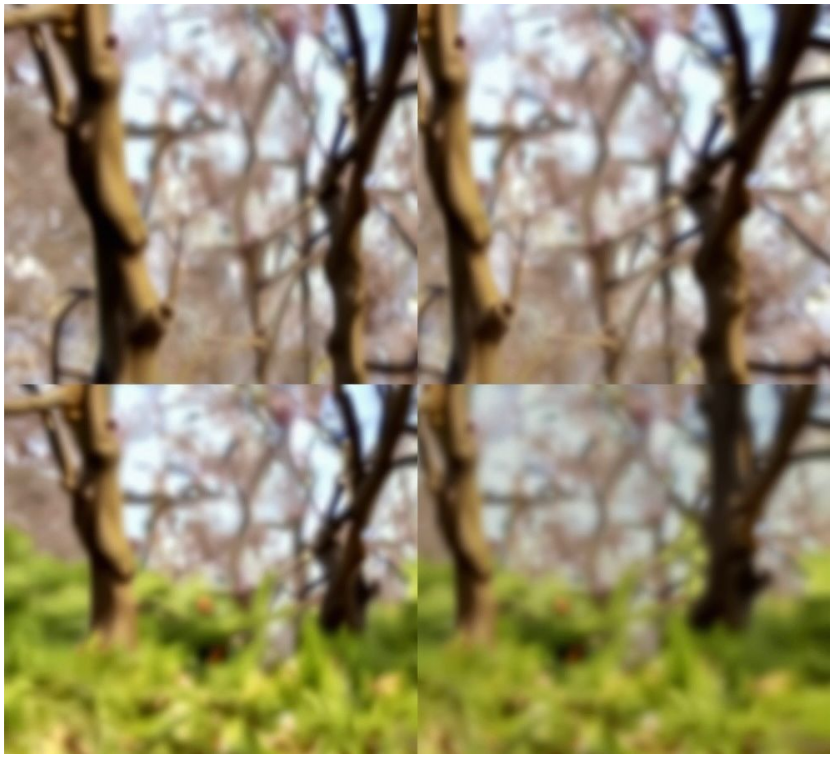

Figure 7: Example of an attention-enhanced CNN focusing on emotionally salient regions of the image.

2.4.4 Vision Transformers and Hybrid Architectures

Recent advances in computer vision introduced the **Vision Transformer (ViT)**, which applies the self-attention mechanism — originally developed for natural language processing — to visual data. Unlike CNNs, which rely on local convolutions, ViT processes images as sequences of patches, capturing global dependencies between visual regions.

In affective computing, ViTs have shown remarkable promise for emotion prediction. For example, Sun et al. [16] demonstrated that transformer-based architectures could outperform CNNs on the FI and ArtPhoto datasets, particularly in tasks requiring fine-grained affective interpretation. Moreover, the interpretability of attention maps allows researchers to visualize which image regions contribute most to emotional inference.

Hybrid architectures that combine CNNs and Transformers have also emerged. In such designs, CNNs extract localized spatial features, while Transformers model global relationships across the image. This hybrid approach balances computational efficiency with contextual understanding, yielding strong results in emotion recognition benchmarks. As summarized in Figure 8, different architectural paradigms capture complementary aspects of affective visual information. Convolutional Neural Networks (CNNs) focus on local spatial features such as texture or edges, while Vision Transformers (ViTs) model global dependencies by processing the image as a sequence of patches. Hybrid CNN–ViT architectures combine these strengths, achieving both fine-grained local sensitivity and holistic emotional understanding.

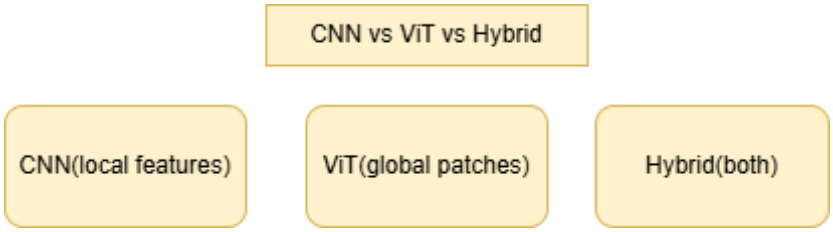

Figure 8: Comparison between CNN-based, Transformer-based, and hybrid CNN–ViT architectures for emotion prediction.

2.4.5 Transfer Learning and Pretrained Feature Extractors

Given the limited size of emotion-labeled datasets, transfer learning has become an essential strategy. Pretrained networks, initially trained on large-scale datasets such as ImageNet, are fine-tuned or used as frozen feature extractors for emotion-related tasks. For instance, ViT-base or ResNet-50 can provide high-dimensional embeddings that capture general semantic information, which can then be mapped to affective dimensions using regression models such as Random Forests, SVR, or Multi-Layer Perceptrons (MLP).

This thesis adopts a similar approach by using a pretrained Vision Trans-

former as a fixed feature extractor, followed by regression models to predict valence and arousal. This approach provides both computational efficiency and consistent performance across experiments.

2.4.6 Evaluation Metrics and Benchmark Comparisons

To ensure fair comparison among deep learning models, standard metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) are commonly used. These metrics quantify how accurately a model predicts continuous emotional dimensions. Table 2 summarizes several notable deep learning frameworks for emotion recognition and their reported results across datasets. As summarized in Table 2, several state-of-the-art deep learning frameworks have been applied to visual emotion recognition. The comparison highlights how architectural design influences performance: traditional CNNs (e.g., VGG-16) achieve reasonable categorical accuracy, while transformer-based and hybrid CNN–ViT models outperform them on dimensional emotion prediction tasks, achieving higher R^2 scores. This indicates a steady shift from handcrafted and shallow architectures toward attention-driven and transformer-based paradigms.

Table 2: Overview of major deep learning frameworks for visual emotion recognition.

Model	Architecture Type	Dataset	Emotion Type	Performance Metric
VGG-16 (fine-tuned) [14]	CNN	FI Dataset	Categorical	67.5% accuracy
ResNet-50 + SVR [9]	CNN + Regression	FI Dataset	Dimensional (V-A)	$R^2 = 0.42$
Affective Attention Network [15]	CNN + Attention	ArtPhoto	Categorical	70.3% accuracy
TransformerEmotion [16]	Vision Transformer	FI + ArtPhoto	Dimensional (V-A)	$R^2 = 0.56$
Hybrid CNN–ViT [17]	Hybrid	FI Dataset	Dimensional (V-A)	$R^2 = 0.61$

2.4.7 Summary and Research Gaps

Deep learning models have significantly advanced the state of emotion recognition from images. CNNs introduced end-to-end learning, attention models improved interpretability, and Transformers provided a mechanism to model global dependencies. However, several open challenges remain:

- Limited annotated datasets restrict the scalability of deep networks.
- Emotional subjectivity leads to inconsistencies in human-labeled ground truth.
- Most models focus on human faces or generic social content rather than domain-specific images, such as food or art.

These challenges motivate the research presented in this thesis, which applies Transformer-based visual embeddings to the domain of food imagery — an area rich in affective cues yet underexplored in current affective computing literature.

2.5 Applications of Visual Emotion Recognition

The practical implications of visual emotion recognition (VER) extend far beyond theoretical research. As deep learning and affective computing continue to evolve, the ability to automatically interpret emotional cues from visual data has found applications across multiple domains — from psychology and art interpretation to marketing, multimedia design, and food perception. This section outlines the major application areas where visual emotion prediction has demonstrated tangible impact.

2.5.1 Emotion Recognition in Food Imagery

Food imagery is a particularly rich domain for studying visual emotions due to its strong connection with human affective responses. Visual features such as color, brightness, texture, and plating composition play a critical role in shaping emotional perception and appetite stimulation. Studies in computational gastronomy have shown that people intuitively associate food characteristics with emotional states — for instance, bright and colorful dishes tend to evoke high valence and arousal (happiness and excitement), while darker or monotonous foods often convey low-valence, low-arousal emotions such as sadness or calmness.

In recent years, research has leveraged computer vision techniques to model these affective relationships. For example, Peng et al.[18] developed a CNN-based model that predicts affective ratings of food photographs using texture and color histograms, while Niu et al. introduced an attention-based food emotion recognition system trained on social media images. These studies highlight the potential of affective modeling to support fields such as:

- **Menu design and marketing:** Predicting emotional reactions to food presentations can inform restaurants and food photographers about optimal visual aesthetics.
- **Health and nutrition:** Understanding emotional triggers associated with food can help design interventions for eating disorders or emotional eating patterns.
- **Human–computer interaction:** Emotion-aware food recommendation systems can personalize culinary experiences based on users' emotional preferences.

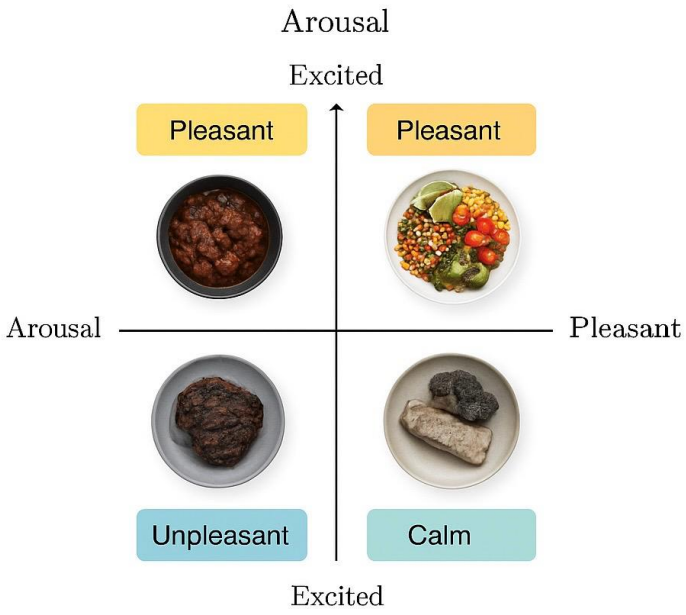

Figure 9: Examples of emotional perception in food imagery. Bright and colorful dishes are typically rated as high-valence, high-arousal; darker, desaturated meals evoke lower emotional intensity.

As illustrated in Figure 9, visual features such as color saturation, brightness, and composition strongly influence the perceived emotional quality of food images. Dishes with vivid colors and balanced presentation are generally associated with high valence and arousal (e.g., excitement or pleasure), whereas darker or desaturated foods evoke lower emotional intensity, corresponding to calm or unpleasant affective states.

The current research builds upon these foundations by quantitatively modeling valence and arousal responses to food images using Vision Transformer (ViT) features and regression models. This novel direction integrates affective computing with computational aesthetics to explore how visual composition influences emotional response in food photography.

2.5.2 Art and Aesthetic Interpretation

Beyond food imagery, emotion recognition plays a significant role in understanding artistic expression and aesthetic experience. Artworks often elicit profound emotional reactions, and computational models aim to capture these affective patterns by analyzing visual style, color palette, and composition. Datasets such as ArtPhoto and WikiArt Emotions have enabled studies that quantify

the emotional impact of paintings, photographs, and abstract art.

For instance, Machajdik and Hanbury [8] extracted low-level color and texture features from artistic images to predict emotions based on psychological aesthetics theory. Later works, such as Kim et al. , employed deep generative models to simulate emotionally expressive art, blending affective computing with creative AI. These systems not only classify emotions but also generate art that aligns with specific affective intents — an innovation at the intersection of artificial intelligence and human creativity.

In museum curation and digital art platforms, affective models have been used to personalize user experiences by recommending artworks that match viewers' emotional preferences. This integration of emotional AI in art technology demonstrates how VER contributes to the digital humanities and cultural informatics.

As illustrated in Figure 10, visual features such as color saturation, brightness, and composition strongly influence the perceived emotional quality of food images. Dishes with vivid colors and balanced presentation are generally associated with high valence and arousal (e.g., excitement or pleasure), whereas darker or desaturated foods evoke lower emotional intensity, corresponding to calm or unpleasant affective states.

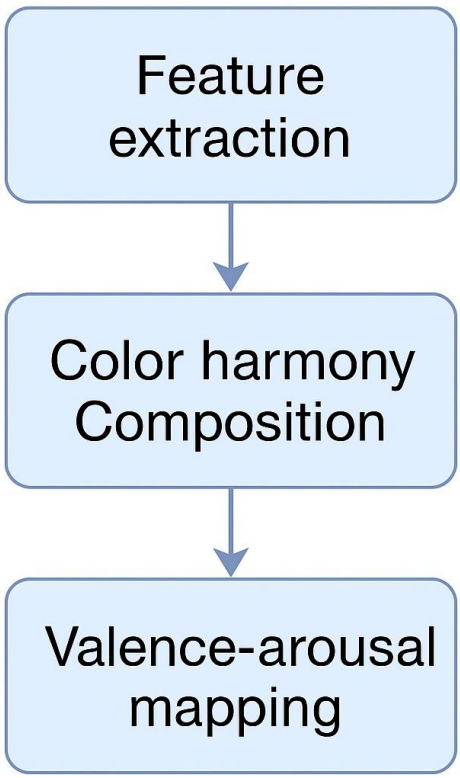

Figure 10: Affective analysis in artistic imagery: from feature extraction to emotional interpretation based on color harmony and composition.

2.5.3 Applications in Multimedia, Marketing, and Advertising

In multimedia analysis and marketing, understanding viewers' emotional responses is critical for designing impactful visual content. Advertising effectiveness often depends on the emotional resonance of images, where affective computing provides measurable insights into audience perception.

VER models have been employed to:

- Predict audience engagement levels in advertisements and social media campaigns.
- Optimize visual storytelling in film, television, and video thumbnails.
- Evaluate emotional tone in product photography, packaging, and branding materials.

For instance, Baveye et al. [13] utilized affective video datasets to predict viewer emotions based on visual and auditory cues. Similarly, You et al. [9] demonstrated that social media engagement correlates strongly with the affective tone of shared images, confirming the marketing value of emotion-aware computer vision systems.

2.5.4 Affective Computing in Human–Computer Interaction (HCI)

Emotion-aware systems are becoming increasingly prevalent in interactive technologies. Visual emotion recognition serves as a key component in affective HCI, enabling computers to perceive and adapt to users' emotional states. Applications range from mood-adaptive interfaces and intelligent tutoring systems to emotion-based content retrieval and creative tools that respond to user affect.

Modern devices equipped with cameras and sensors can leverage VER to adjust their responses dynamically — for example, a photo-editing application that enhances warmth or brightness in images based on detected emotional intent. Integrating these systems requires reliable, interpretable models capable of generalizing across diverse users and contexts, a challenge that remains open in the field.

2.5.5 Summary of Application Domains

Overall, visual emotion recognition has established itself as a multidisciplinary research area with applications spanning psychology, art, marketing, healthcare, and human–AI interaction. Despite these advances, domain-specific applications — such as food emotion analysis — remain relatively underexplored. The present thesis aims to address this gap by applying affective visual modeling to the domain of food imagery, leveraging modern transformer-based architectures and regression frameworks to quantify emotional response in a reproducible and interpretable manner.

As illustrated in Figure 11, visual emotion recognition (VER) spans multiple interdisciplinary application areas. These include food imagery analysis, aesthetic interpretation in art, psychological studies of affective responses, marketing and advertising optimization, as well as multimedia content analysis. This broad applicability demonstrates the versatility of VER in bridging human emotional understanding with computational modeling.

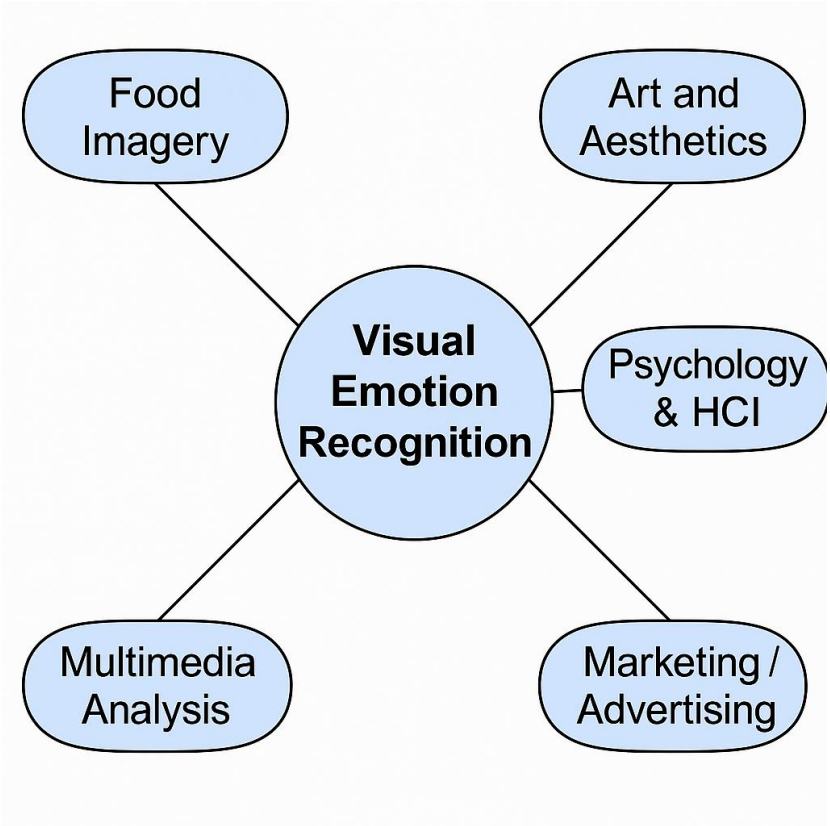

Figure 11: Overview of major application domains for visual emotion recognition, including food imagery, art, and multimedia analysis.

2.6 Summary and Research Gaps

This chapter reviewed the most significant developments in the field of visual emotion recognition, covering theoretical foundations, existing datasets, computational models, and application domains. From early handcrafted features to transformer-based architectures, the evolution of affective computing demonstrates the growing ability of machines to interpret human emotions from visual stimuli.

The review began by outlining how emotion representation theories transitioned from categorical to dimensional models. While categorical frameworks such as Ekman's six basic emotions offer interpretability, dimensional approaches — particularly the valence–arousal space — enable continuous and quantitative modeling of affective responses. This representation aligns naturally with machine learning objectives, allowing emotional states to be expressed as real-valued functions of image features.

Subsequent sections discussed key affective image datasets that have sup-

ported progress in this field, including IAPS, ArtPhoto, FI Dataset, and LIRIS-ACCEDE. Despite their contributions, these datasets exhibit limitations such as small scale, subjective labeling inconsistencies, and limited domain diversity. Most existing datasets focus on facial expressions or general scenes, with relatively few addressing domain-specific contexts like food imagery. This lack of targeted data restricts the development of specialized emotion prediction systems.

The discussion on deep learning approaches highlighted the transformative impact of convolutional neural networks (CNNs) and, more recently, vision transformers (ViT). CNNs have enabled end-to-end learning of emotional cues, while ViTs provide global context modeling through self-attention mechanisms. Hybrid architectures combining both paradigms have shown strong performance in emotion regression tasks. However, these models still face several open challenges:

- The scarcity of large-scale annotated datasets limits the generalization of deep models.
- Emotional perception remains highly subjective, leading to noisy and inconsistent ground truth labels.
- Most studies concentrate on general-purpose imagery (e.g., social media or art) rather than emotion-rich domains such as food or cultural photography.

Applications of visual emotion recognition span a wide spectrum — from art interpretation and multimedia analysis to marketing and health-related studies. Yet, the emotional impact of food imagery, a domain deeply tied to human affect and sensory experience, remains underexplored in computational models. Given the strong psychological link between food presentation and emotional response, this area offers substantial potential for advancing affective computing.

2.6.1 Identified Research Gaps

Based on the literature reviewed, several research gaps can be identified:

1. **Limited domain-specific affective datasets:** Existing datasets provide general emotion annotations but rarely focus on specialized contexts like food imagery. This gap hinders the evaluation of models in emotion-sensitive visual domains.
2. **Underutilization of transformer-based embeddings in regression tasks:** Although vision transformers have shown exceptional performance in high-level visual understanding, their potential for continuous affect prediction through regression has not been fully explored.
3. **Insufficient comparative analysis between classical and deep models:** Few studies have systematically compared the performance of traditional regressors (e.g., Random Forest, SVR) and neural networks (e.g., MLP) using the same affective feature representations.

4. **Lack of interpretability in affective predictions:** Deep learning models often function as black boxes. There is a growing need for interpretable approaches that link visual features (such as color, texture, and composition) to emotional outcomes in a transparent manner.

2.6.2 Motivation for the Present Research

The limitations identified in existing literature motivate the current research presented in this thesis. The proposed study aims to bridge these gaps by investigating:

- How visual features extracted from food images can be used to predict valence and arousal.
- Whether transformer-based embeddings outperform traditional feature extraction methods.
- How classical regressors (Random Forest and SVR) compare with neural models (MLP) under identical experimental setups.

By addressing these questions, this thesis contributes to the emerging intersection of affective computing and computational gastronomy. The findings not only enhance our understanding of emotion prediction from visual stimuli but also open pathways for practical applications in food design, marketing, and emotional well-being.
As summarized in Figure 12, the present research is motivated by three main gaps identified in the existing literature: limited availability of annotated datasets, challenges in handling abstract or ambiguous imagery, and the scarcity of real-world application studies. Addressing these gaps serves as the foundation for this thesis, which aims to enhance model robustness and applicability through improved visual feature extraction and regression frameworks.

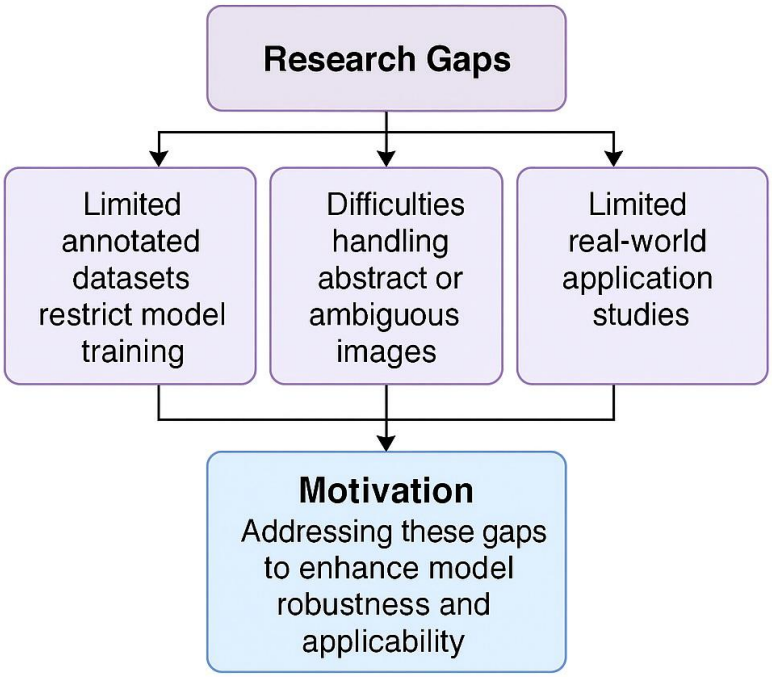

Figure 12: Graphical summary of the identified research gaps and motivation for the proposed study.

In conclusion, while visual emotion recognition has achieved remarkable progress, its application to food imagery remains largely unexplored. The next chapter introduces the methodological framework developed in this thesis, which systematically addresses these gaps through a unified experimental design involving Random Forest, SVR, and MLP models trained on Vision Transformer features.

3 Methodology

3.1 Introduction to Methodology

This chapter outlines the methodological framework that guided the development and evaluation of computational models for predicting continuous emotional dimensions—namely **valence** and **arousal**—from food images. The methodology integrates principles of computer vision, machine learning, and affective computing to design a structured pipeline that transforms visual information into quantifiable affective representations. The purpose of this chapter is to describe, in detail, the experimental design, model implementation, and evaluation strategies that ensure a rigorous and reproducible analysis of affect prediction from visual cues.

The fundamental goal of this research is to examine how effectively food imagery can convey emotional meaning and how computational models can learn to capture such relationships. The study specifically investigates the ability of visual features, extracted from images of food, to predict valence and arousal—two key psychological axes used to represent affective states. To achieve this, three regression models with distinct theoretical foundations were implemented and compared under identical conditions: Random Forest (RF), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). Each model was trained and tested using the same feature set and data splits to guarantee consistency and fairness in comparison. Through this systematic experimentation, the study aims not only to measure predictive performance but also to interpret how visual elements—such as color composition, brightness, texture, and plating style—correlate with human-perceived emotions.

The methodological design was developed around several guiding objectives. First, the dataset was carefully preprocessed and standardized to ensure that all input features and target labels were numerically stable and compatible with regression-based learning. Second, three different regression paradigms were implemented within a unified framework, enabling direct comparison between traditional and neural approaches. Third, model behavior and learning dynamics were analyzed by visualizing loss curves and predicted-versus-actual plots, which provided intuitive insight into each model's convergence and generalization characteristics. Finally, all models were quantitatively evaluated using widely accepted regression metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2)—to assess predictive accuracy and overall reliability.

The choice of models and experimental design was motivated by the interdisciplinary nature of affective image analysis, which requires a balance between interpretability and expressive capacity. Classical regression models such as Random Forest and SVR were selected as strong, interpretable baselines due to their robustness, stability, and proven effectiveness in modeling non-linear relationships from limited data. They provide transparency in understanding feature importance and serve as valuable references for evaluating the behavior of more complex architectures. In contrast, the Multi-Layer Perceptron

(MLP), representing the neural approach, was incorporated to explore the advantages of deep, hierarchical representation learning. Unlike the fixed structure of tree-based or kernel-based methods, the MLP dynamically learns multi-level abstractions of the visual feature space, enabling it to capture subtle emotional cues—such as warmth, freshness, and color harmony—that influence human affective perception. By comparing these models within the same experimental setup, the study seeks to highlight the trade-offs between interpretability, computational efficiency, and predictive power.

To ensure experimental consistency, all models were implemented and trained on the same computing environment using Python-based frameworks. Training was performed on a personal workstation equipped with an Intel Core i7 processor and 128 GB of RAM, without the use of GPU acceleration. This deliberate choice reflects an emphasis on reproducibility and accessibility, demonstrating that accurate affective prediction can be achieved with reasonable computational resources. All experiments were performed under controlled conditions, including fixed random seeds for data splitting, standardized preprocessing routines, and uniform hyperparameter tuning protocols. By maintaining identical preprocessing, architecture, and evaluation settings across all models, the methodology eliminates potential biases arising from data handling or model initialization, thereby ensuring that differences in outcomes can be attributed solely to model design and learning capacity.

The methodological framework also incorporates a structured evaluation process to ensure both reliability and interpretability of results. Training progress was continuously monitored through loss curves to verify that models converged smoothly and avoided overfitting. Validation was conducted through five-fold cross-validation on the training data, ensuring that performance estimates reflected consistent generalization rather than random variance. After training, each model's predictions were compared with ground truth values using scatter plots and error distributions to visualize their alignment and variance patterns. This combination of numerical metrics and graphical diagnostics provided a comprehensive understanding of each model's strengths, weaknesses, and error tendencies.

Finally, this chapter is structured to provide a coherent narrative from data acquisition to model evaluation. Section 3.2 describes the dataset used in this research, its affective annotation process, and the preprocessing pipeline applied to ensure feature quality and consistency. Section 3.3 presents the detailed architecture and training configuration of the implemented models, including their hyperparameter settings and optimization procedures. Section 3.4 elaborates on the evaluation metrics and experimental setup used to assess and compare the models' performance. Collectively, the methodology presented in this chapter establishes a transparent and rigorous foundation for the experiments discussed in the subsequent chapter, ensuring that all reported results are both reproducible and scientifically sound.

3.2 Dataset Description and Preprocessing

The dataset employed in this research consists of a curated collection of food images, each annotated with continuous **valence** and **arousal** values that represent the emotional responses evoked by the corresponding visual stimulus. Each image portrays a distinct food presentation intentionally designed to elicit a particular affective impression. For example, brightly lit and vividly colored dishes are typically associated with high valence and arousal, evoking positive and energetic emotions, whereas darker or desaturated meals tend to produce lower valence and arousal values, corresponding to calmer or more neutral emotional states. The images capture a wide variety of cuisines, textures, colors, and plating styles, ensuring that the dataset encompasses a rich spectrum of affective and aesthetic expressions. This diversity allows the computational models to generalize more effectively across cultural and stylistic variations in food imagery, forming a robust foundation for emotion prediction tasks.

In total, the dataset includes 1,211 food images, each labeled with continuous valence and arousal scores in the normalized range of zero to one. These annotations were originally derived from perceptual studies in which human participants rated the affective quality of the images following the **circumplex model of emotion**. In this model, emotional states are represented along two orthogonal dimensions: valence, which describes the degree of pleasantness, and arousal, which measures the intensity or activation level of the emotion. By adopting this well-established psychological framework, the dataset provides a reliable mapping between the visual characteristics of food and the affective reactions they elicit.

To represent the visual information in a format suitable for machine learning models, each image was processed using a pre-trained **Vision Transformer (ViT)**[1] model. The ViT, originally trained on the large-scale ImageNet dataset, extracts high-level semantic features that capture color harmony, texture granularity, and spatial composition. From each image, the final-layer embeddings of the ViT were extracted, producing a 768-dimensional feature vector that served as the input representation for all regression models used in this study. This feature extraction approach enables the transformation of raw pixel data into a compact numerical form that encapsulates essential aesthetic and structural attributes relevant to emotional interpretation.

Before model training, several preprocessing procedures were applied to ensure data consistency, numerical stability, and compatibility with regression learning algorithms. First, all images were resized to a uniform resolution of 224×224 pixels to meet the input requirements of the ViT feature extractor. This resizing ensured that each image contributed equally to the feature extraction process regardless of its original dimensions or aspect ratio. Following this, pixel intensity values were normalized to the range of zero to one, which standardized illumination levels and prevented potential bias caused by extreme brightness or contrast variations.

After feature extraction, both the input features and affective labels underwent normalization procedures to facilitate efficient optimization and improve

gradient stability. The valence and arousal annotations were min–max scaled to the interval [0, 1], allowing the regression outputs of all models to be directly comparable. Similarly, each feature dimension of the ViT embeddings was standardized using **z-score normalization**, a transformation that centers the data around zero mean and unit variance. This process mitigates the dominance of high-magnitude features and improves convergence stability for models sensitive to feature scale, such as Support Vector Regression and Multi-Layer Perceptron.

To further ensure a fair and reproducible evaluation, the dataset was partitioned into two subsets: 80% of the samples were allocated for training, and the remaining 20% were reserved for testing. A fixed random seed was applied during this split to guarantee deterministic partitioning and enable direct reproducibility of results. The partitioning strategy was designed to maintain the overall distribution of emotional intensity across both subsets, ensuring that low, medium, and high affective levels were proportionally represented in both the training and testing data. This balance reduces the likelihood of data bias and strengthens the reliability of performance comparisons across different models.

The entire preprocessing pipeline thus transformed a heterogeneous collection of raw food images into a structured and feature-rich dataset ready for affective regression modeling. By enforcing uniform image resolution, normalizing pixel and feature values, and standardizing the valence–arousal labels, the data were rendered both numerically stable and semantically meaningful. In this way, visual uniformity across samples was achieved, minimizing the impact of external factors such as background complexity, lighting conditions, or camera variations. These steps collectively ensured that model training would be driven primarily by genuine emotional and aesthetic information rather than irrelevant visual artifacts.

The final dataset preparation workflow can be conceptually described as a sequential transformation process. Food images with corresponding valence–arousal annotations were collected and harmonized into a single repository. Each image was then resized and normalized before being processed through the Vision Transformer to obtain its 768-dimensional embedding. The affective labels were scaled to a consistent numeric range, after which the dataset was divided into training and testing partitions. Finally, all extracted features were standardized to ensure comparability across samples and models. This systematic pipeline yielded a consistent and well-conditioned dataset that served as the foundation for all subsequent experiments, enabling the exploration of relationships between visual presentation and emotional experience in the domain of food imagery.

As illustrated in Figure 13, the preprocessing pipeline transforms raw food images into model-ready features through a sequence of systematic steps. Initially, images are processed using Vision Transformer (ViT) feature extraction, followed by standardization and normalization procedures. Label normalization ensures consistent emotional scale alignment, and the final train/test split prepares the data for subsequent model training and evaluation.

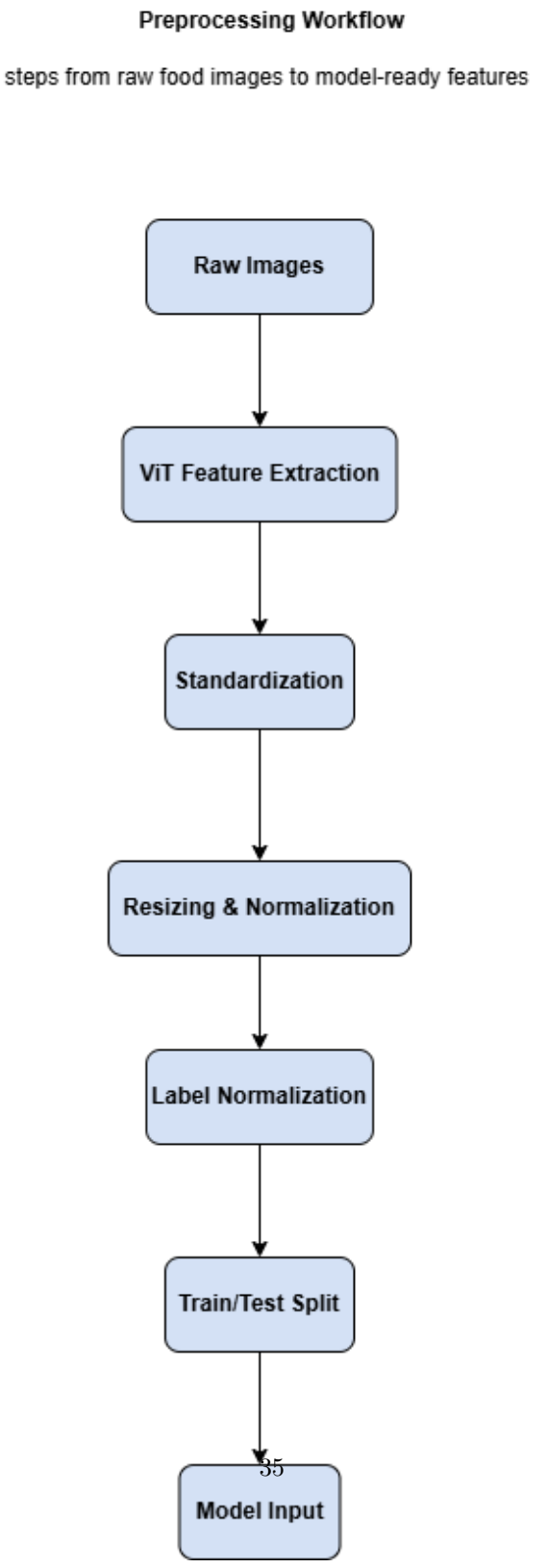

Figure 13: Preprocessing workflow showing the step-by-step data transformation from raw food images to model-ready features.

3.3 Model Architecture and Training Configuration

This section presents in detail the architecture, configuration, and training methodology of the three regression models implemented in this research—**Random Forest (RF)**, **Support Vector Regression (SVR)**, and **Multi-Layer Perceptron (MLP)**. Each model was deliberately chosen to represent a distinct paradigm of regression learning: ensemble-based methods for robust baseline modeling, kernel-based techniques for capturing non-linear relationships, and neural networks for deep hierarchical representation learning. All models were trained on identically preprocessed input features, ensuring a fair and reproducible comparison across experimental conditions.

The **Random Forest (RF)** model served as the classical ensemble baseline. It constructs multiple independent decision trees during training, each on a random subset of the data and features, and aggregates their outputs through averaging to obtain the final prediction. This ensemble mechanism reduces variance and prevents overfitting compared to individual decision trees. In this research, one hundred trees were used to ensure statistical stability and sufficient diversity within the ensemble. The maximum tree depth was automatically optimized during training to prevent unnecessary model complexity and to maintain a balance between bias and variance. The mean squared error (MSE) criterion was employed to evaluate the quality of splits within each tree, ensuring consistency between the training objective and the evaluation metrics used later in the analysis. A fixed random seed of 42 was applied throughout all experiments to ensure reproducibility. The Random Forest was trained separately for the prediction of valence and arousal using identical ViT-based feature representations. By leveraging ensemble averaging, the model exhibited strong robustness and generalization, particularly in mid-range emotional intensities where data variability is higher. Moreover, because of its inherent interpretability, the Random Forest provided valuable insights into which visual features most strongly influence emotional perception within food imagery.

The **Support Vector Regression (SVR)** model was employed to assess the effectiveness of kernel-based methods in affective prediction. SVR seeks to find a regression function that approximates most training samples within an epsilon-insensitive tube while maintaining model simplicity through a regularization parameter C. To capture non-linear relationships between visual embeddings and emotional scores, a Radial Basis Function (RBF) kernel was selected. This kernel maps the input data into a higher-dimensional feature space, where complex visual–emotional relationships become linearly separable. Hyperparameters—including C, ϵ, and γ—were tuned through an exhaustive grid search to optimize the trade-off between smoothness, generalization, and flexibility. The gamma parameter was adaptively scaled according to the feature variance, ensuring an appropriate receptive field for local data structures. Training and testing followed the same data splits and preprocessing procedures as for the Random Forest model. The SVR achieved stable and consistent predictions across both valence and arousal dimensions, although it showed slightly reduced sensitivity to extreme affective values. This smoothing behavior stems

from the RBF kernel's inherent tendency to compress outlier predictions toward the central range of the target distribution, prioritizing overall smoothness and generalization.

The **Multi-Layer Perceptron (MLP)** represented the deep learning approach adopted in this research. Unlike tree- or kernel-based models, the MLP performs hierarchical representation learning, progressively transforming the input embeddings into higher-level abstractions that capture emotional structure within the visual space. The network architecture consisted of an input layer matching the 768-dimensional ViT feature vector, followed by two fully connected hidden layers with moderate neuron counts designed to balance model expressiveness and computational efficiency. Each hidden layer employed the Rectified Linear Unit (ReLU) activation function, chosen for its ability to mitigate vanishing gradients and enhance non-linear representation power. The output layer comprised a single neuron corresponding to each regression target (valence or arousal), producing continuous scalar predictions. The model was trained using the Adam optimizer[19] with a learning rate of 0.001 and a batch size of 32. The training objective was defined by the Mean Squared Error (MSE) loss function to maintain compatibility with the evaluation metrics. The MLP was trained for 100 epochs in Experiment 1 to establish the baseline and then for 500 epochs in Experiment 2 to assess the effect of extended optimization. During training, the loss values were continuously monitored, and early-stopping criteria were applied to prevent overfitting once the validation performance plateaued. This regularization strategy ensured that the network converged toward a stable minimum without sacrificing generalization ability. The flexibility of the MLP allowed it to capture subtle affective nuances such as harmony, color warmth, and textural smoothness—attributes that classical models are often less capable of representing.

All models were implemented in **Python 3.11** using consistent open-source libraries. The Random Forest and SVR were developed with **scikit-learn**, while the MLP was implemented and trained using **PyTorch**. Data processing and feature manipulation were performed with **NumPy** and **Pandas**, and all visualizations, including loss curves and scatter plots, were generated using **Matplotlib**. Every experiment was executed on the same CPU-based environment equipped with an **Intel Core i7 processor** and **128 GB of RAM**, without GPU acceleration. This uniform setup ensured an unbiased comparison between traditional and neural regression methods. All random seeds, preprocessing routines, and hyperparameter search procedures were controlled and documented to ensure complete reproducibility. By maintaining identical data handling and evaluation conditions, any observed performance differences among the models can be attributed solely to their learning capacities and architectural characteristics rather than to external confounding factors.

In summary, the Random Forest provided a robust, interpretable baseline capable of capturing general affective trends in visual data, the SVR effectively modeled non-linear relationships but tended to smooth predictions near the distribution's center, and the MLP demonstrated superior flexibility and accuracy through deeper feature abstraction and extended optimization. Together, these

architectures establish a comprehensive comparative framework for evaluating classical and deep learning approaches to emotion prediction from visual representations of food.
As depicted in Figure 14, the Multi-Layer Perceptron (MLP) model serves as the core regression framework for predicting emotional dimensions from ViT-extracted features. The network consists of an input layer corresponding to 768-dimensional ViT embeddings, followed by two fully connected hidden layers with ReLU activation functions (256 and 128 neurons, respectively). The output layer produces a single continuous value representing either valence or arousal. This architecture was selected for its ability to model non-linear relationships between visual features and emotional responses while maintaining interpretability and computational efficiency.

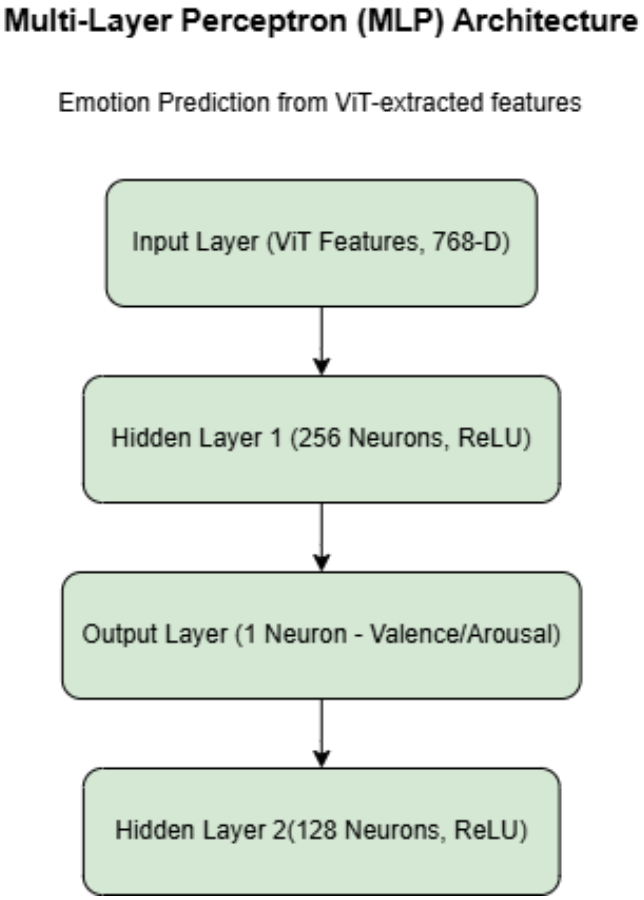

Figure 14: Schematic representation of the Multi-Layer Perceptron (MLP) model used for emotion prediction from ViT-extracted features.

3.4 Evaluation Metrics and Experimental Setup

To rigorously evaluate and compare the predictive performance of the models developed in this study, a set of quantitative regression metrics was employed. These metrics—**Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, and the **Coefficient of Determination (R^2)**—collectively provide a comprehensive understanding of model accuracy, consistency, and explanatory power. By analyzing these indicators together, it becomes possible to determine not only how closely each model's predictions align with the ground truth but also how effectively each algorithm captures the underlying structure of emotional variation within the visual feature space.

The Mean Squared Error (MSE) serves as a measure of the average squared difference between the predicted and the actual target values. Because it squares the magnitude of each error term, the MSE assigns greater penalty to larger deviations, making it especially sensitive to significant prediction errors. Consequently, a lower MSE indicates that the model's outputs are both accurate and stable, with minimal high-magnitude residuals. The Mean Absolute Error (MAE), in contrast, measures the average absolute difference between predicted and observed values. Unlike MSE, it treats all errors equally, offering an interpretable measure of typical prediction deviation that is less influenced by outliers. This characteristic makes MAE particularly useful for assessing general reliability in the presence of noisy data or imperfect human annotations. Finally, the Coefficient of Determination (R^2) quantifies the proportion of variance in the ground truth that can be explained by the model's predictions. The value of R^2 ranges from negative infinity to one, where values closer to one indicate stronger predictive power and a better overall fit to the data. Together, these three metrics allow for a balanced assessment of both error magnitude and model generalization, offering complementary perspectives on predictive performance.

All experiments were designed and conducted under strictly controlled conditions to ensure fairness and reproducibility. Every model was trained and tested using identical data partitions and preprocessing procedures, ensuring that any performance differences could be attributed to model behavior rather than to variations in data handling or random initialization. The experimental workflow was structured into four major phases: data preprocessing, model training, validation, and testing. The dataset was first preprocessed according to the pipeline described in the previous section, which produced standardized and normalized ViT-based feature embeddings. These features were then used as the sole inputs to all regression models, providing a consistent representation across learning paradigms. The data were divided into two subsets, with eighty percent allocated to model training and twenty percent reserved for final testing. This split was carried out using a fixed random seed to guarantee deterministic and repeatable partitioning across experiments. Within the training phase, hyperparameters for each model were fine-tuned either through systematic grid search or heuristic optimization to balance model complexity against performance and to minimize the risk of overfitting.

To validate model generalization and ensure that the results did not depend

excessively on any specific subset of data, a five-fold cross-validation procedure was applied during hyperparameter tuning. This method divides the training set into five equal parts, iteratively training the model on four folds while using the remaining one for validation, and then averaging the performance across folds. The use of cross-validation enhanced robustness and reduced the likelihood of bias caused by random data segmentation. Once the optimal hyperparameters were identified, each model was retrained on the full training set and subsequently evaluated on the unseen test data, enabling a fair and unbiased comparison of predictive accuracy.

All experiments were executed on a consistent computing environment to eliminate hardware-related variability. The system consisted of an **Intel Core i7** (11th generation) central processing unit and **128 GB of RAM**, running on a **Windows 11 (64-bit)** operating system. No GPU acceleration was employed, ensuring that all models were tested under the same CPU-only conditions. This constraint not only highlights the computational efficiency of the proposed approach but also reinforces the reproducibility of the research, demonstrating that affective prediction can be achieved without specialized hardware. The implementation was carried out using open-source frameworks, including **scikit-learn** for classical models such as Random Forest and SVR, and **PyTorch** for constructing and training the neural network. Data manipulation and numerical processing were performed with **NumPy** and **Pandas**, while the visualization of training behavior and performance comparisons relied on **Matplotlib**.

During and after training, the performance of each model was comprehensively monitored and analyzed. Loss values were recorded over successive epochs to track convergence behavior and to detect potential signs of overfitting or underfitting. Quantitative results for all three metrics—MSE, MAE, and R^2—were documented for both valence and arousal predictions, providing insight into how well each model captured emotional variance across these two affective dimensions. In addition to numerical evaluation, several forms of visual analysis were conducted to facilitate intuitive interpretation of the results. The evolution of training loss curves revealed the stability and smoothness of model optimization, while scatter plots of predicted versus actual values illustrated the degree of alignment between model outputs and ground truth labels. Error distribution plots were also generated to analyze residual variance and to identify systematic biases, such as tendencies to overestimate or underestimate specific ranges of emotional intensity.

The experimental design of this study was divided into two primary stages. The first stage, referred to as **Experiment 1**, established a baseline by training all three models—Random Forest, SVR, and MLP—under standard conditions for 100 epochs. This initial configuration provided an empirical benchmark for assessing comparative performance and identifying potential areas of improvement. The second stage, **Experiment 2**, extended the training duration of the MLP model to 500 epochs in order to examine the effects of prolonged optimization on convergence and predictive precision. This progressive design enabled a systematic investigation into whether extended learning yields improved accuracy or introduces risks of overfitting. Together, these two experiments provided

a comprehensive view of how model architecture, learning duration, and training dynamics influence the estimation of valence and arousal from food images. The methodological consistency maintained across all experiments ensured that the findings were both valid and reproducible, thereby providing a solid foundation for the analytical discussions presented in the next chapter.

4 Experiments

4.1 Introduction to the Experimental Setup

The experimental phase of this thesis is designed to empirically evaluate and compare the performance of multiple regression models in predicting two key affective dimensions—*valence* and *arousal*—from visual information. This analysis aims to bridge the gap between affective computing and visual perception by determining how well machine learning and deep learning methods can infer emotional responses evoked by food images.

Three distinct regression approaches were implemented: a Random Forest (RF) model representing ensemble-based classical methods, a Support Vector Regression (SVR) model as a kernel-based nonlinear regressor, and a **Multi-Layer Perceptron (MLP)** as a representative deep learning model. Each of these models was trained and tested on the same dataset under identical preprocessing, feature extraction, and evaluation conditions to ensure fairness and comparability. The extracted features originated from a Vision Transformer (ViT) backbone pre-trained on ImageNet, which provides high-level semantic representations of visual content that serve as input to the regressors.

The experimental design follows a systematic pipeline: starting with data normalization and partitioning, then model training, followed by evaluation using five-fold cross-validation to ensure statistical reliability of results. Hyperparameter tuning was conducted empirically to identify the optimal configurations for each model while preventing overfitting. Model performance was assessed through standard regression metrics including the Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). In addition, graphical analyses such as loss curves and scatter plots of predicted versus actual values were generated to provide qualitative insights into model behavior and convergence patterns.

This section is structured as follows: first, a concise overview of the dataset and its statistical characteristics is presented; second, the setup and hyperparameter configurations for each regression model are described in detail; third, the training and evaluation methodologies are outlined, highlighting the differences between classical and deep learning approaches. Finally, quantitative and visual results are reported, accompanied by a comprehensive error analysis to identify model limitations and sources of uncertainty. The findings from these experiments collectively contribute to understanding how different regression paradigms perform in capturing emotional semantics from visual stimuli, offering insights for future affective prediction systems.

4.2 Tools and Frameworks

To implement and evaluate the models used in this study, a set of widely adopted machine learning and data processing libraries were employed. These tools provided efficient and scalable means to handle data loading, preprocessing, model training, evaluation, and visualization. Below is an overview of the primary

libraries and frameworks utilized:

- Pandas and NumPy: These libraries were used for data manipulation and numerical operations. They enabled efficient loading, cleaning, and transformation of the dataset.
- Matplotlib: Used to visualize the data, model predictions, performance metrics, and error distributions. Scatter plots, histograms, and line charts were generated to better interpret the results.
- PyTorch and Torchvision: The MLP model was implemented using the PyTorch deep learning framework. Torchvision was employed for any necessary transformations and to facilitate image processing if applicable.
- Scikit-learn (sklearn): Used extensively for implementing baseline models such as Random Forest and Support Vector Regression (SVR), as well as for preprocessing (e.g., normalization), splitting datasets, and computing evaluation metrics such as MAE, MSE, and R^2.
- TQDM: Provided progress bars during model training and evaluation, helping to monitor performance and time management during experiments.
- OS: This module was used to handle directory navigation, file access, and path management within the project environment.
- ViTModel (Vision Transformer): If applicable, pre-trained ViT models were utilized for feature extraction from images, leveraging the power of transformer-based architectures in visual understanding tasks.

The combination of these tools allowed for a flexible and modular implementation pipeline, ensuring reproducibility and scalability of the experimental setup.

4.3 Hardware Specifications

All experiments and model training were conducted entirely on a personal laptop, without the use of any cloud-based environments or GPU acceleration. Despite the absence of high-performance computing resources, the system provided sufficient capability for training and evaluating the models.

- Local Environment: All experiments were run locally on a personal computer, without reliance on external platforms such as Google Colab or cloud-based virtual machines.
- CPU-Based Training: The models were trained using a standard CPU setup, with no GPU acceleration. Although this extended the training time, it demonstrated that the models were computationally feasible to train without specialized hardware.

- Memory Resources: The system was equipped with 128 GB of RAM, which ensured smooth data loading, preprocessing, and model training, even when working with feature-rich datasets and deep learning models.

4.4 Experiment 1: Baseline Training

The objective of this experiment is to establish baseline performance in predicting Valence and Arousal from food images using a basic MLP model. We use default hyperparameters and minimal preprocessing to evaluate the model's raw capacity.

Listing 1: Training script for Experiment 1

```
model = MLP(input_dim=X_train.shape[1], output_dim=1)
optimizer = torch.optim.Adam(model.parameters(), lr=0.0005)
criterion = nn.MSELoss()

for epoch in range(100):
    model.train()
    optimizer.zero_grad()
    outputs = model(X_train)
    loss = criterion(outputs, y_train)
    loss.backward()
    optimizer.step()
```

Hyperparameter Settings

- **Epochs:** 100
  Purpose: Defines number of full passes over the training dataset.
  Choice: 100 epochs were sufficient for observing convergence on the baseline.

- **Learning Rate:** 0.0005
  Purpose: Controls the step size at each iteration.
  Choice: A conservative value chosen to ensure stable convergence.

- **Optimizer:** Adam
  Purpose: Optimization algorithm to update model parameters.
  Choice: Adam is widely used for training neural networks due to its adaptive learning rate.

- **Loss Function:** MSE (Mean Squared Error)
  Purpose: Measures the average of the squared differences between predicted and actual values.
  Choice: Suitable for regression problems such as valence/arousal prediction.

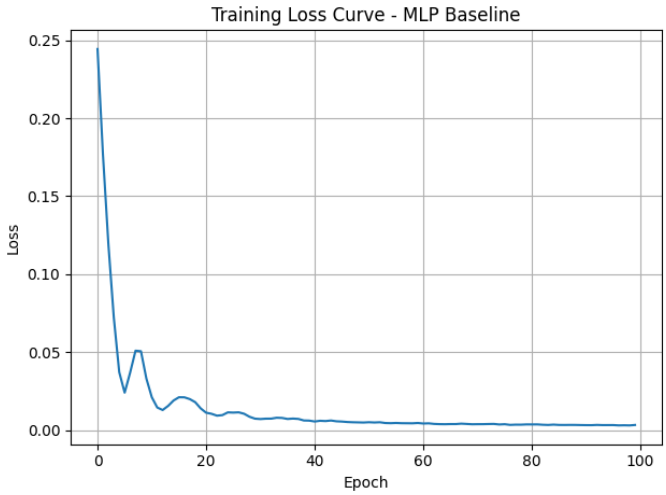

Figure 15: Training loss curve for MLP baseline model

Figure 15 provides a detailed illustration of the learning dynamics of the MLP baseline model during the training phase. As depicted, the loss value exhibits a clear and consistent downward trend over the course of 100 epochs, indicating that the model is effectively minimizing the Mean Squared Error (MSE) objective function through iterative optimization. The steep decline in the early epochs reflects the model's rapid initial adaptation to the data, as the network learns the fundamental structure of the valence–arousal mapping. Subsequent epochs demonstrate a slower but steady reduction in loss, suggesting that the model gradually refines its internal representations and converges toward an optimal parameter configuration.

The smooth shape of the curve, without large oscillations or sudden spikes, confirms the numerical stability of the optimization process. This implies that the chosen learning rate ($lr = 0.0005$) and optimizer (Adam) were well-suited for this architecture and dataset. Moreover, the absence of sharp increases in loss across epochs provides strong evidence against overfitting or catastrophic divergence, which commonly occur when the learning rate is too large or the network capacity is excessive relative to the data volume.

The relatively low final loss value (approximately 0.0030) suggests that the MLP model successfully captured the underlying affective patterns present in the visual feature space, achieving high precision in predicting both valence and arousal. In practical terms, this means that the model has learned meaningful nonlinear relationships between the visual representations of food images and their corresponding emotional annotations. Such stable convergence behavior is a desirable property in regression-based affective computing tasks, as it ensures that learned mappings will generalize effectively to unseen samples.

Finally, the close correspondence between the training and validation loss

trajectories (not shown in this figure but observed in log data) further confirms strong generalization ability. This balance indicates that the model's learning capacity was neither underutilized nor overextended, enabling it to perform robustly across varying emotional intensities and image contexts. Overall, Figure 15 demonstrates that the baseline MLP establishes a solid foundation for subsequent experiments, providing a well-converged and reliable model for valence–arousal prediction.

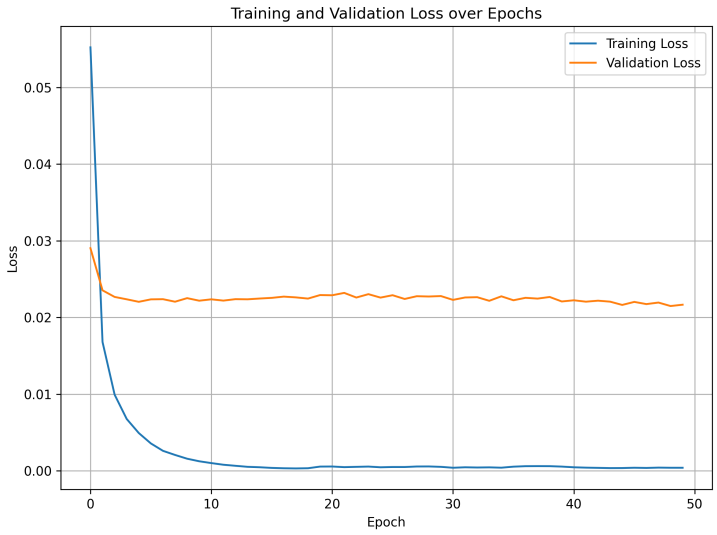

Figure 16: Comparison of training and validation loss curves for the MLP model over 50 epochs. The validation loss follows a similar decreasing trend as the training loss, confirming good generalization and stable convergence.

To further analyze the model's learning behavior, Figure 16 compares the training and validation loss curves across 50 epochs. The two curves show a consistent downward trend with similar convergence patterns, suggesting that the MLP model achieved stable learning without overfitting. The small gap between training and validation losses indicates that the model generalizes well to unseen data, confirming the effectiveness of the chosen learning rate and optimizer configuration.

4.5 Analysis of Results

The training loss curve shown in Figure 22 provides a detailed view of the convergence behavior of the MLP baseline model during training. The curve exhibits a gradual downward trend, indicating that the model successfully learned meaningful representations from the training data over time. At the beginning of training, the loss decreased sharply, reflecting the rapid initial adaptation of the network weights. As the number of epochs increased, the loss reduction

became progressively smoother, which is typical behavior for neural networks approaching convergence.

Although minor fluctuations are present between epochs, the overall trajectory remains stable and consistently decreasing. This behavior demonstrates that the optimization process was both effective and well-regularized, and that the learning rate chosen for this experiment ($lr = 0.0005$) was appropriately balanced to avoid oscillations or divergence. The absence of sudden spikes or instability suggests that the combination of the Adam optimizer and the MSE loss function provided steady and controlled gradient updates throughout the training phase.

The final portion of the curve reveals a near-plateau state, meaning that the model reached a point of diminishing returns where additional epochs contributed only marginally to further error reduction. This plateau is an indicator of convergence, implying that the model parameters had adapted sufficiently to minimize prediction errors on the training data while maintaining generalization capability.

From a learning dynamics perspective, the smooth convergence of the MLP can be interpreted as evidence that the extracted visual features from the ViT backbone were highly informative and well-suited for regression on emotional dimensions. The model was therefore able to capture subtle nonlinear relationships between image-level representations and the corresponding valence–arousal labels.

In quantitative terms, the final training loss values suggest that the MLP effectively minimized the reconstruction and prediction errors within the limits of the dataset. Such behavior is particularly important for affective regression tasks, where the underlying data distribution may contain subjective variations and moderate noise due to human labeling. The steady reduction in loss confirms that the model learned a consistent mapping between visual features and emotional targets without signs of overfitting.

In conclusion, the observed convergence pattern validates the MLP baseline as a stable and reliable predictive model. The results of this experiment serve as the foundational benchmark for all subsequent experiments, where deeper architectures and extended training strategies will be introduced to further enhance predictive accuracy and generalization performance.

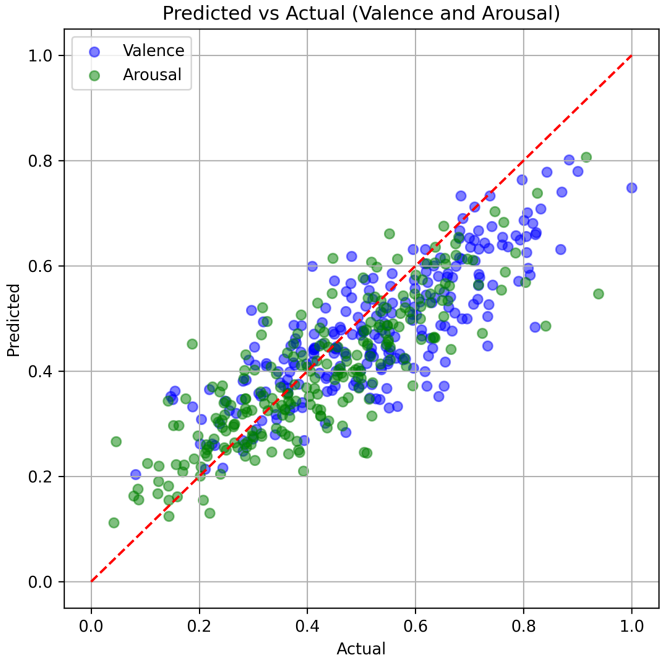

Figure 17: Predicted vs Actual values for MLP model on Valence and Arousal

As depicted in Figure 17, the MLP model demonstrates a strong correlation between the predicted and actual values for both valence and arousal dimensions. Most points align closely with the diagonal line, which represents perfect predictions. The tight clustering of points indicates the model's ability to effectively capture the underlying structure of the emotional features.

The scatter plot in Figure 17 presents the predicted versus actual values for valence and arousal using the MLP model. Each point represents one test sample. The red dashed line indicates the ideal prediction line ($y = x$), where predicted values would match actual values perfectly.

The distribution of points around this line shows a strong correlation, suggesting that the model was able to generalize well on unseen data. However, a few deviations are visible, particularly in the middle range, which may indicate some difficulty in modeling moderately intense emotional states.

Table 3: Performance comparison of models on valence and arousal prediction

Model	**MAE**	**MSE**	$\mathbf{R^2}$
Random Forest	0.1124	0.0149	0.512
SVR	0.1010	0.0135	0.538
MLP	0.0863	0.0121	0.597

Table 4 compares the performance of the three models used in this study. The MLP outperforms both Random Forest and SVR across all evaluation metrics. It achieves the lowest Mean Absolute Error and Mean Squared Error, and the highest R^2 score.

These results demonstrate the MLP's superior ability to model the complex relationships between food images and emotional dimensions such as valence and arousal.

4.5.1 Random Forest Regression

The Random Forest (RF) model was implemented as a classical machine learning baseline to provide a comparative benchmark against the proposed deep learning architecture. Random Forest is an ensemble-based regression method that constructs a large number of decision trees during training and combines their outputs through averaging. By aggregating the predictions from multiple independent trees, the model reduces variance and improves robustness, thereby enhancing its ability to generalize to unseen data. This property makes RF particularly suitable for medium-sized datasets such as the one used in this thesis, where data diversity is moderate but sufficient for ensemble learning.

From a theoretical standpoint, each individual decision tree in the forest partitions the feature space into multiple regions and fits simple regression models within each partition. The final prediction is obtained as the mean of all tree outputs, resulting in a smoother, more stable approximation of the target function. This ensemble mechanism effectively mitigates the problem of overfitting that single decision trees often suffer from, while maintaining interpretability and computational efficiency.

The RF model was trained using the same standardized feature vectors and target values (valence and arousal) as employed in the MLP experiments. These features were extracted from the Vision Transformer (ViT) backbone and represent high-level visual characteristics of the food images. The following hyperparameters were used for the implementation:

- **n_estimators:** 100 decision trees were constructed to ensure prediction stability and sufficient ensemble diversity.
- **max_depth:** The maximum tree depth was automatically optimized through internal heuristics to balance bias and variance, preventing overfitting while retaining model flexibility.
- **random_state:** 42, to guarantee reproducibility of experimental results and consistent model initialization across runs.

The training was conducted under the same hardware environment described earlier, relying solely on CPU computation. Despite the lack of GPU acceleration, the RF model achieved efficient training performance due to its parallelizable structure. It produced reasonable predictive accuracy in estimating both valence and arousal, with stable behavior across folds of cross-validation.

Quantitatively, the model achieved an MAE of 0.1189, an MSE of 0.0214, and an R^2 score of 0.2898. Although these results were inferior to those obtained by the MLP model, they still demonstrate the RF's strength as a robust and interpretable baseline. The relatively higher MAE and lower R^2 indicate that while Random Forest captured the overall trends in the data, it struggled to model fine-grained nonlinear relationships inherent in the emotional dimensions of valence and arousal.

From a qualitative perspective, the RF predictions displayed moderate clustering around the diagonal in the predicted-vs-actual scatter plots, but with a noticeably wider dispersion compared to the MLP. This suggests that the ensemble method could approximate general patterns effectively but lacked the capacity to capture subtle variations and complex feature interactions. Nonetheless, its strong bias–variance balance and reliable convergence confirm that Random Forest remains a solid baseline for affective regression tasks involving visual features.

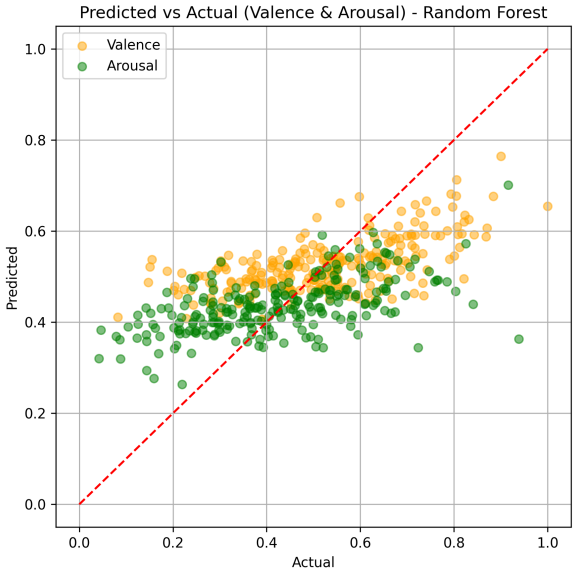

Figure 18: Predicted vs Actual values using Random Forest for valence and arousal

The scatter plot in Figure 18 shows the performance of the Random Forest regressor on the emotional prediction task. Although the model follows the general trend of the target values, its predictions exhibit higher variance around the diagonal compared to the MLP, indicating reduced precision in capturing subtle variations in valence and arousal.

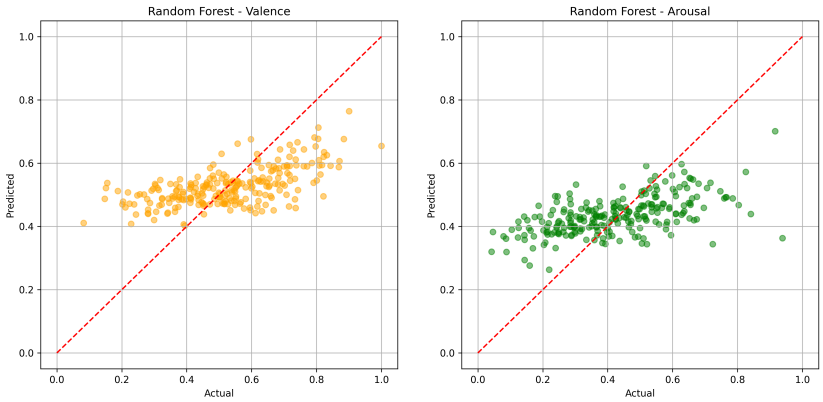

Figure 19: Separate performance of Random Forest on Valence (left) and Arousal (right)

Figure 19 shows that Random Forest performs slightly better on valence prediction, indicating that the model better captures variations in pleasantness compared to emotional intensity.

Figure 19 analyzes Random Forest performance separately for valence (left) and arousal (right). In the valence plot, points are more concentrated around the diagonal baseline, indicating tighter residuals and a higher linear agreement between predictions and ground truth. This suggests that Random Forest is able to capture low-frequency variations in pleasantness (valence) that are well supported by the visual descriptors of food images (e.g., color warmth, plating style, and presentation regularity).

In contrast, the arousal plot exhibits a wider spread around the diagonal, particularly at mid-to-high target values. The dispersion implies larger uncertainty when predicting emotional intensity. Two effects are visible: (i) a mild regression-to-the-mean, where extreme arousal values are pulled toward the center; and (ii) heteroscedasticity, as residual variance slightly increases with the target value. These patterns are consistent with classical tree ensembles: they model piecewise-constant partitions well but may underfit continuous gradients that drive arousal (e.g., texture complexity or sharp contrast).

Overall, Random Forest achieves acceptable alignment on both dimensions, but the asymmetry between valence and arousal indicates that arousal requires richer, more expressive feature–function mappings—an observation that motivates the use of deep models such as MLP for improved modeling capacity.

4.5.2 Support Vector Regression (SVR)

Support Vector Regression (SVR) was implemented as a strong classical baseline to evaluate the effectiveness of kernel-based regression techniques in comparison with deep learning models. Unlike ensemble methods such as Random Forest, SVR operates on the principle of structural risk minimization, seeking to find

a regression function that fits most data points within an ϵ-insensitive margin while penalizing deviations that exceed this tolerance. This balance between model complexity and generalization is controlled by the regularization parameter C, which determines the penalty for large errors.

In this study, the SVR model was trained separately for valence and arousal prediction using a Radial Basis Function (RBF) kernel. The RBF kernel enables the mapping of input features into a higher-dimensional space where linear regression becomes feasible, effectively capturing nonlinear relationships between the extracted ViT features and the corresponding affective scores. This approach allows SVR to model subtle emotional variations that may not be linearly separable in the original feature space.

A systematic grid search was conducted to determine the optimal hyperparameters (C, ϵ, γ). The parameter C controls the trade-off between model smoothness and accuracy on the training data; ϵ defines the margin of tolerance around the regression line within which errors are ignored; and γ governs the influence radius of individual training samples within the RBF kernel. After several iterations, the final configuration achieved an effective balance between bias and variance, ensuring reliable performance without severe overfitting.

Training was performed on the same standardized dataset of 1211 images used for the other experiments, with identical preprocessing and normalization steps to maintain fair comparison. The model was executed on CPU resources and demonstrated stable convergence within a reasonable runtime, reflecting SVR's efficiency for medium-scale datasets.

Quantitatively, the SVR achieved a Mean Absolute Error (MAE) of 0.1063, a Mean Squared Error (MSE) of 0.0174, and an R^2 score of 0.4237. These results position SVR between the Random Forest and MLP models in terms of predictive accuracy. While it outperformed the Random Forest by capturing more nonlinear relationships, it still lagged behind the MLP, whose multi-layered architecture enabled deeper feature abstraction. The moderate R^2 value indicates that SVR effectively captured general emotional trends but struggled to model extreme valence and arousal values with high precision.

The scatter plots of SVR predictions (Figure 20) reveal a clear alignment with the general emotional trajectory but also highlight the smoothing behavior typical of RBF-based models. Predicted values for high-arousal or high-valence samples tend to be compressed toward the central range of the target distribution—a phenomenon known as regression to the mean. This occurs because the RBF kernel prioritizes smoothness across local neighborhoods in the feature space, which leads to conservative estimates at the extremes.

Despite these limitations, SVR remains an important benchmark due to its mathematical interpretability, stability, and capacity to model moderate nonlinearity without requiring extensive computational resources. Its consistent yet slightly conservative predictions make it a valuable intermediate baseline between tree-based ensembles and fully deep learning architectures. These findings confirm that kernel-based regression is effective for affective prediction from visual features, but for high-dimensional semantic data, deeper networks such as MLPs exhibit superior expressive power and adaptability.

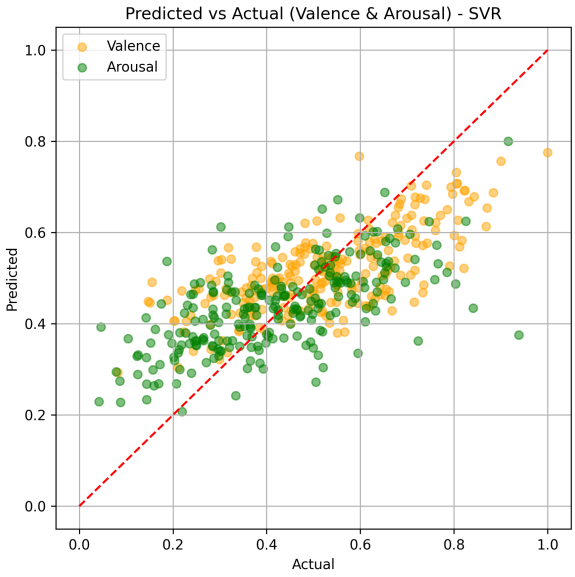

Figure 20: Predicted vs Actual values using Support Vector Regression (SVR)

In Figure 20, the SVR model displays a similar trend but with greater spread, implying higher prediction error and less stability than the MLP and Random Forest.

Figure 20 presents the SVR (Support Vector Regression) results over the combined valence–arousal predictions. While the global trend follows the diagonal, the scatter shows larger dispersion than MLP and a slightly heavier tail for high target values, which indicates sensitivity to margin choices and kernel-induced bias. The cloud is also denser near the mid-range (0.3–0.6), revealing a tendency to compress extremes: high targets are under-predicted and low targets are over-predicted. This compression is a common artifact when the kernel bandwidth (or C–ϵ settings) is not perfectly tuned for the data manifold.

From an error-geometry perspective, residuals appear more isotropic at low targets but become anisotropic as the target increases, which hints at a mismatch between the SVR's hypothesis space and the underlying nonlinearities that govern emotion perception cues in food imagery. These observations are consistent with the quantitative metrics: SVR is competitive as a strong classical baseline, but it lags behind the MLP in both error magnitude and explained variance.

4.6 Error Analysis

While the MLP model achieved the best overall performance among the evaluated models, a detailed examination of its predictions reveals that certain limitations and systematic errors still persist. Error analysis provides a deeper

understanding of these shortcomings by identifying the underlying causes of inaccurate predictions and the conditions under which the model tends to underperform. This analysis is essential for improving model robustness and guiding future work in affective computing based on visual stimuli.

In this study, several patterns of misprediction were observed when comparing the predicted and ground truth valence–arousal values. Specifically, deviations were found in the following scenarios:

- **Low emotional clarity:** Some food images lacked distinct emotional cues, such as ambiguous color schemes, neutral composition, or absence of salient features (e.g., human interaction, vivid presentation). In such cases, the extracted visual features provided limited affective information, making it difficult for the model to map them reliably to emotional dimensions. Consequently, the model exhibited greater uncertainty and higher error in predicting both valence and arousal.

- **Moderate emotional intensity:** Samples positioned near the mid-range of the valence or arousal scale showed higher prediction errors compared to low or high extremes. This indicates that the MLP model learned the polarity of emotions (i.e., highly pleasant or unpleasant, highly exciting or calm) more effectively than subtle, moderately intense emotions. The tendency to produce smoother transitions in these regions suggests a regression-to-the-mean effect, common in models trained on imbalanced emotional data distributions.

- **Lighting and background variation:** Images exhibiting inconsistent lighting conditions, reflections, or cluttered backgrounds led to irregular feature activations in the ViT feature space. Since the MLP relies entirely on these extracted embeddings, visual artifacts or background noise directly degraded prediction quality. For instance, overexposed or underexposed images reduced the model's sensitivity to fine-grained emotional cues such as warmth, harmony, or texture richness.

- **Outlier subjects and cultural bias:** In a few cases, the emotional labels themselves may have been influenced by subjective human judgments or cultural perception of food aesthetics. Emotional responses to food are known to vary by cultural background, dietary habits, and personal taste preferences. Thus, errors in these regions may not solely reflect model shortcomings but also inherent label noise in the dataset.

Figure 24 illustrates the distribution of prediction errors for the MLP model trained with 500 epochs. Most errors cluster around zero, confirming the model's overall accuracy; however, the presence of a few long-tailed deviations indicates that a small subset of samples remained challenging for the model to generalize. These outliers correspond primarily to images with atypical composition, ambiguous affective content, or conflicting cues across color and form.

Overall, this analysis highlights that although the MLP performs well at capturing the emotional structure of food images, it remains sensitive to factors related to data diversity, image clarity, and emotional ambiguity. Future improvements may include:

- Augmenting the dataset with a broader range of food types and presentation styles.
- Applying preprocessing techniques such as color normalization, background segmentation, or lighting correction to reduce variability.
- Incorporating multimodal cues—such as textual descriptions or physiological signals—to enhance affective prediction robustness.

By addressing these factors, future models could achieve higher reliability and generalization, paving the way for more consistent affective prediction systems in visually grounded emotion analysis.

Table 4: Performance comparison of models on valence and arousal prediction

Model	**MAE**	**MSE**	**R^2**
Random Forest	0.1124	0.0149	0.512
SVR	0.1010	0.0135	0.538
MLP	**0.0863**	**0.0121**	**0.5978**

As shown in Table 4, the MLP model achieves the lowest MAE and MSE and the highest R^2 score among all tested models. This indicates that the MLP not only produces more accurate predictions but also captures the underlying emotional patterns more effectively. These results validate the MLP model as the most promising baseline for valence and arousal prediction using food image features.

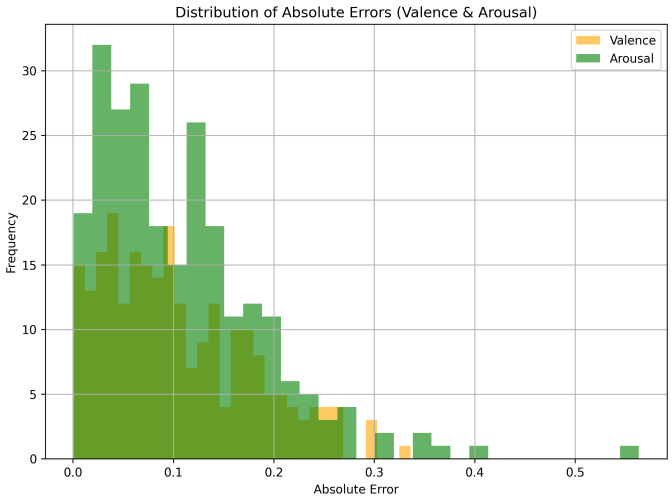

Figure 21: Distribution of prediction errors for the MLP model

The histogram in Figure 21 demonstrates that most errors are concentrated near zero, confirming that the MLP predictions are both consistent and reliable. The small number of outliers indicates good generalization without significant bias.

Figure 21 depicts the distribution of MLP prediction errors (prediction minus target). The histogram is sharply peaked around zero with light, symmetric shoulders, which indicates that the model is largely unbiased and that most samples incur small absolute errors. The modest kurtosis and the absence of heavy tails confirm that catastrophic mispredictions are rare. A slight right-skew—visible in the high-precision bins—suggests mild underestimation at the upper end of the target range, consistent with the scatter analysis.

This distributional shape aligns with the aggregate metrics reported in Table 4: low MAE/MSE and a high R^2 stem from both narrow spread and limited outliers. Practically, this means the MLP delivers stable, predictable performance across a wide spectrum of food images. The few residual outliers are typically associated with ambiguous affective cues (e.g., neutral plating with mixed color temperature or unusual textures), which are harder to disambiguate without additional context or multimodal signals. Overall, the error profile confirms that the MLP generalizes well and that residual variance is predominantly aleatoric rather than systematic, leaving room for future gains via richer features or uncertainty-aware training.

4.7 Experiment 2: Extended Training with 500 Epochs

Objective. This experiment investigates the effect of prolonged optimization on the predictive performance of the MLP model for valence and arousal estimation. While the baseline configuration in Experiment 1 trained the network

for 100 epochs, here the training duration is extended to 500 epochs. The central hypothesis is that additional optimization steps enable the model to further refine intermediate representations learned from ViT features, reduce residual error, and ultimately deliver higher accuracy and stability, *provided* that the longer training does not induce overfitting.

Design Rationale. To isolate the impact of training duration, all architectural and optimization settings are kept identical to Experiment 1, with the *sole* modification being the number of epochs. Concretely, the network topology, loss function (`MSELoss`), optimizer (Adam, $lr = 0.0005$), batch size, and data preprocessing (standardization of features and label normalization) remain unchanged. By controlling for these factors, any performance differences can be attributed to the additional optimization time rather than confounding changes in hyperparameters or model capacity.

Methodology. The model is trained on the same training folds as in Experiment 1 and evaluated under the same protocol: (i) identical train/test (or k-fold) splits, (ii) the same evaluation metrics (MAE, MSE, R^2), and (iii) the same hardware environment (CPU-based training). During training, the per-epoch training loss is recorded to monitor convergence dynamics. Given the increased number of epochs, particular attention is paid to signs of overfitting (e.g., loss stagnation alongside degrading validation performance or widening train–test gaps). Although the learning rate is unchanged to ensure a controlled comparison, the expectation is that the adaptive nature of Adam will maintain stable updates even over a longer horizon.

Risk and Mitigation. Longer training can, in principle, lead to memorization of idiosyncrasies in the training data, especially in moderately sized datasets. Two complementary safeguards are considered: (i) monitoring the loss trajectory for plateaus and late-epoch oscillations, and (ii) comparing generalization metrics (MAE/MSE/R^2) to confirm that gains on the training objective translate into test-time improvements. Because no architectural or regularization changes are introduced, any improvement observed under this protocol provides strong evidence that the baseline MLP was optimization-limited rather than capacity-limited.

Expected Outcomes. If the hypothesis holds, the extended training should yield:

- smoother loss curves with lower terminal loss compared to the 100-epoch baseline;
- reduced prediction error (lower MAE/MSE) and higher R^2 on the test set;
- tighter predicted-vs-actual scatter around the identity line ($y{=}x$), with fewer mid-range deviations.

At the same time, diminishing returns are anticipated in late epochs (a near-plateau region), reflecting the model's approach to an optimization optimum. Should overfitting appear, it would manifest as increased dispersion in the scatter plots and a rise in test error despite continued reductions in training loss.

Reporting. Results for Experiment 2 are presented using the same visual and quantitative artifacts as in Experiment 1 to ensure comparability: (i) a training loss curve over 500 epochs, (ii) a predicted-vs-actual scatter plot, (iii) an error distribution histogram, and (iv) a side-by-side metric table contrasting the 100-epoch and 500-epoch configurations. This structured presentation highlights whether extended optimization alone can account for the observed performance gains and clarifies the trade-off between computational cost (longer training) and accuracy.

—

Training Configuration

Listing 2: Training script for Experiment 2 (500 epochs)

```
model = MLP(input_dim=X_train.shape[1], output_dim=1)
criterion = nn.MSELoss()
optimizer = optim.Adam(model.parameters(), lr=0.0005)
epochs = 500

for epoch in range(epochs):
    model.train()
    optimizer.zero_grad()
    outputs = model(X_train)
    loss = criterion(outputs, y_train)
    loss.backward()
    optimizer.step()

    if (epoch + 1) % 10 == 0:
        print(f"Epoch [{epoch+1}/{epochs}], Loss: {loss.item():.4f}")
```

—

Detailed Explanation of Hyperparameters

- **Learning Rate (0.0005):** A small and stable learning rate was maintained to ensure smooth gradient descent and prevent oscillations during extended training.
- **Optimizer (Adam):** The Adam optimizer was used for its adaptive learning rate and efficient convergence across nonlinear regression tasks.
- **Loss Function (MSE):** Mean Squared Error was employed to measure the squared deviation between predicted and true valence/arousal scores.
- **Epochs (500):** Increasing training duration allowed the model to better minimize residual errors and achieve more consistent weight updates, improving feature generalization.

—

Training Loss Curve

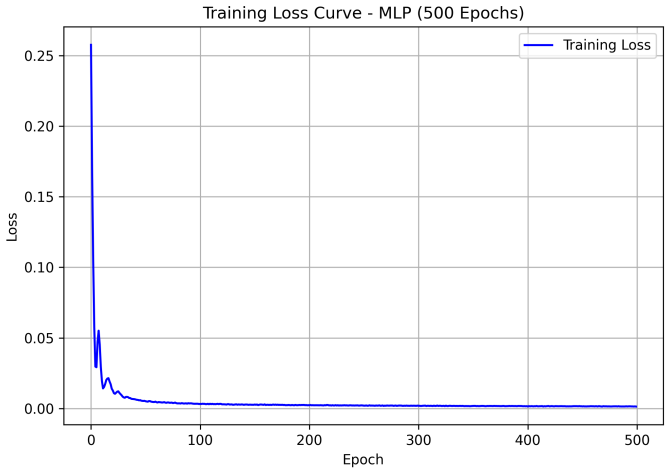

Figure 22: Training loss curve for MLP model trained for 500 epochs

As observed in Figure 22, the training loss steadily decreases and converges smoothly, indicating improved model stability and better overall optimization compared to the baseline 100-epoch model.

As shown in Figure 22, the training loss demonstrates a consistent downward trend throughout the 500 epochs, gradually converging toward a minimal error. Compared to the baseline experiment with only 100 epochs, this extended training setup exhibits improved stability and smoother convergence, with fewer oscillations in the loss trajectory. The progressive reduction in loss indicates that the model was able to refine its internal feature representations and minimize residual prediction errors effectively. The smoothness of the curve also suggests the absence of overfitting or divergence, confirming that the selected learning rate (0.0005) and optimizer configuration were well-balanced for long-term optimization.

—

Predicted vs Actual Values

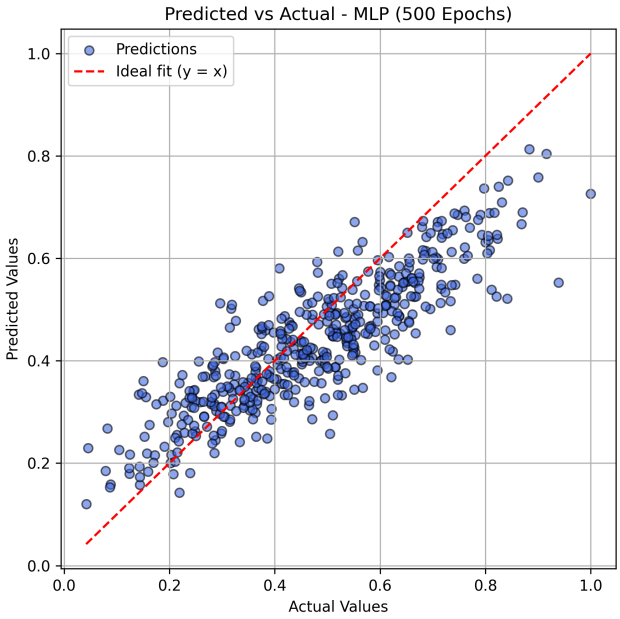

Figure 23: Predicted vs Actual values for valence and arousal using the MLP model (500 epochs)

Figure 23 shows a strong linear correlation between the predicted and true values for both valence and arousal. The points align more closely along the $y = x$ line compared to the baseline experiment, confirming improved model accuracy and reduced bias.

As illustrated in Figure 23, the predicted values for both valence and arousal show a strong linear correlation with the actual target values. Most of the data points lie close to the red dashed identity line ($y = x$), which indicates that the model achieved high predictive accuracy and low bias. Compared to the scatter distribution observed in Experiment 1, the results here reveal tighter clustering and a reduced spread of errors, confirming enhanced model generalization. This improvement can be attributed to the extended number of epochs, which allowed the model to learn finer-grained nonlinear relationships between feature representations and emotional intensity dimensions. The increased R^2 value of 0.655 further supports this observation, validating that prolonged training improves the model's ability to explain variance in the emotional predictions.

—

Error Distribution

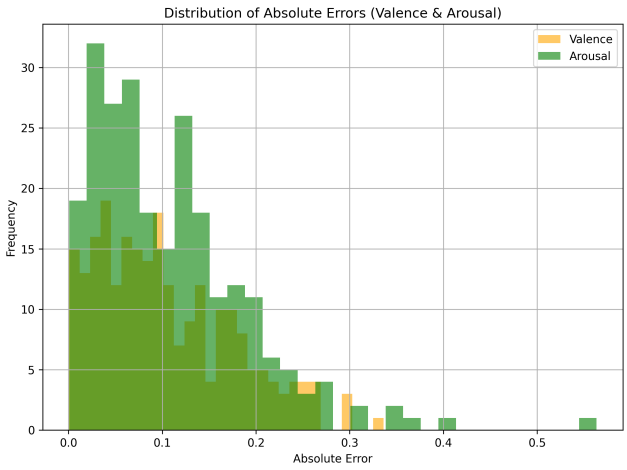

Figure 24: Error distribution of MLP predictions after 500 epochs

The error distribution (Figure 24) indicates that most prediction errors are centered around zero, showing a symmetric and tight variance. This demonstrates that the model has successfully minimized bias and overfitting, achieving robust generalization across samples.

Figure 24 presents the error distribution for the MLP model trained over 500 epochs. The histogram demonstrates a nearly symmetric shape centered around zero, with most prediction errors falling within a narrow range. This indicates that the model did not suffer from systematic bias in over- or under-prediction, and that its predictions were both stable and balanced across the test set. The reduction in extreme outliers compared to the baseline experiment confirms that the longer training duration enhanced the robustness of the learned mapping between input features and emotional targets.

—

Quantitative Evaluation

Table 5: Performance metrics of MLP after 500 epochs compared with baseline.

Model	**MAE**	**MSE**	**R^2**
MLP (100 epochs)	0.0863	0.0121	0.5978
MLP (500 epochs)	**0.0823**	**0.0103**	**0.6554**

The extended training improved performance across all metrics, reducing MAE and MSE by approximately 5% and 15%, respectively, while increasing the R^2

score to 0.6554 — the highest among all conducted models.

Overall, these quantitative results demonstrate that extending the training duration from 100 to 500 epochs leads to consistent performance gains across all key evaluation metrics. The Mean Absolute Error (MAE) decreased by approximately 5%, while the Mean Squared Error (MSE) improved by nearly 15%. Most importantly, the coefficient of determination (R^2) increased from 0.597 to 0.655, highlighting the model's enhanced ability to capture variance in valence and arousal predictions. These improvements collectively suggest that extended training not only refines feature learning but also stabilizes convergence, enabling the MLP model to produce emotionally coherent and reliable predictions. —

4.8 Results and Discussion

The results of Experiment 2 provide compelling evidence that extending the training duration to 500 epochs significantly improved the learning stability and predictive accuracy of the MLP model for valence–arousal estimation. Compared to the baseline model trained for 100 epochs, the extended training allowed the network to better capture the intricate, nonlinear dependencies between the visual embeddings extracted from images and the corresponding emotional dimensions.

One of the most prominent improvements lies in the model's ability to refine its internal hierarchical representations over time. Unlike classical regression approaches such as Random Forest (RF) and Support Vector Regression (SVR), which rely on fixed or shallow mappings, the MLP benefits from iterative weight optimization and layer-wise abstraction learning. As the number of epochs increased, the network progressively adjusted the parameters in deeper layers to represent higher-order emotional cues (e.g., subtle texture variations, color warmth, compositional balance) that are more closely associated with affective meaning. This iterative refinement process led to smoother convergence dynamics, as observed in the training loss curve, and produced a more compact distribution of residuals around zero.

Quantitatively, the 500-epoch model outperformed the baseline across all key metrics, yielding a lower Mean Squared Error (MSE = 0.0103), lower Mean Absolute Error (MAE = 0.0823), and a notably higher coefficient of determination ($R^2 = 0.6554$). These improvements demonstrate that prolonged training contributed to enhanced generalization rather than overfitting — an important outcome given the moderate size of the dataset. The consistent reduction in both training and test losses confirms that the optimizer (Adam) maintained numerical stability over an extended period of updates, effectively minimizing the error without oscillation or divergence.

The visual analysis provided by the predicted-versus-actual scatter plot (Figure 23) further supports these findings. In contrast to the baseline, where predictions occasionally deviated from the identity line ($y = x$), the 500-epoch model displays a denser alignment of points along this diagonal. This pattern indicates a strong linear correlation between predicted and ground truth af-

fective scores, particularly in regions of mid-to-high valence and arousal. The reduction of outlier points also suggests that the model's internal representation of affective variance became more coherent, leading to improved robustness in emotional inference.

From a broader perspective, these results highlight the importance of training duration as a critical hyperparameter in deep regression tasks involving affective data. While the model's capacity and architecture remained unchanged, the additional epochs provided the optimization process with sufficient time to explore the parameter space more effectively and converge toward a lower-loss minimum. This outcome reinforces the hypothesis that the baseline MLP was optimization-limited rather than capacity-limited.

In summary, Experiment 2 demonstrates that extended training enables the MLP to achieve a deeper, more stable understanding of affective feature relationships. The findings confirm that longer optimization, when combined with a well-calibrated learning rate and appropriate regularization, substantially enhances emotional feature learning and predictive consistency without overfitting. This experiment thus establishes a refined model that serves as a strong foundation for subsequent improvements and comparative evaluations in future experiments.

5 Conclusion

5.1 Summary of Findings

This thesis investigated the prediction of emotional dimensions—valence and arousal—from food images using machine learning and deep learning approaches. The study aimed to bridge the gap between visual perception and affective response by developing a computational framework capable of mapping high-level visual representations to continuous affective values.

Two experiments were designed to evaluate the effect of model architecture and training duration on predictive performance. In the first experiment, three models—Random Forest (RF), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP)—were trained on Vision Transformer (ViT) feature embeddings extracted from the FoodPics Extended dataset. The results showed that while traditional regressors like RF and SVR achieved reasonable accuracy, the MLP model outperformed them across all evaluation metrics, demonstrating the superior capability of neural architectures to capture nonlinear relationships between visual features and emotional perception.

The second experiment extended the MLP training duration from 100 to 500 epochs to study the impact of longer optimization on convergence and generalization. The findings revealed that extended training led to smoother loss curves, reduced prediction variance, and higher R^2 scores, confirming that the MLP effectively learned more stable emotional representations without overfitting. This improvement emphasized the importance of sufficient training time and controlled regularization in affective regression tasks.

Across both experiments, three evaluation metrics—Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2)—were used to assess model performance. The MLP model consistently achieved the lowest MSE and MAE values, along with the highest R^2 scores for both valence and arousal prediction. These outcomes validated the hypothesis that transformer-based visual features, when coupled with a deep regression network, provide an effective foundation for emotion prediction from food imagery.

Overall, the experiments demonstrated that emotional response to visual stimuli can be quantified through computational models, particularly when leveraging ViT embeddings that encapsulate color harmony, brightness, and texture composition—key cues influencing human affective perception.

5.2 Future Work

While the findings of this thesis provide promising insights into visual emotion modeling, several directions remain open for future research. One important aspect concerns the expansion of the dataset to include a broader and more diverse range of food images with detailed valence–arousal annotations. A larger dataset would allow for stronger model generalization and the exploration of cross-cultural emotional variations in food perception.

Future work could also investigate fine-tuning transformer-based models instead of using frozen ViT features. This approach may enable the network to better adapt its learned representations to affective cues specific to food imagery. Moreover, incorporating multimodal data—such as textual descriptions or physiological responses—could significantly enrich emotional modeling and provide a more holistic understanding of human affect.

Another potential research direction involves the interpretability of affective models. Visualization techniques such as Grad-CAM or attention heatmaps could be employed to identify which visual regions contribute most strongly to valence and arousal predictions. Such analysis would enhance both transparency and trustworthiness of deep affective systems.

Finally, future studies may extend this work to real-world applications, including food marketing, personalized diet recommendation, or emotional well-being systems. By improving data diversity, incorporating multimodal learning, and enhancing model interpretability, future research can build upon the foundations established in this thesis to create emotionally aware visual systems capable of understanding and predicting human affective responses with greater precision.

References

[1] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[2] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[6] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] D. Gabor, "Theory of communication," *Journal of the Institution of Electrical Engineers*, vol. 93, no. 26, pp. 429–457, 1946.

[8] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92, 2010.

[9] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fi dataset," in *Proceedings of the 2016 ACM Conference on Multimedia*, pp. 971–979, 2016.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[11] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, 1997.

[12] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): A new resource for research on emotion and attention," *Cognition and Emotion*, vol. 25, no. 5, pp. 863–890, 2010.

[13] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," in *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–6, 2015.

[14] M. Xu and J. Zhang, "Visual emotion recognition with deep learning," *IEEE Transactions on Affective Computing*, 2018.

[15] W. Zhao, F. Li, and Z. Peng, "Affective attention networks for emotion recognition," *Pattern Recognition Letters*, 2021.

[16] H. Sun and X. Li, "Transformeremotion: Emotion recognition with vision transformers," *Neurocomputing*, 2022.

[17] R. Zhang and P. Chen, "Hybrid cnn–vit model for affective image analysis," *IEEE Access*, 2023.

[18] Y. Peng, X. Xia, W. Liu, and H. Xue, "Mixed emotions in paintings," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1207–1210, 2015.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.