International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multimodal Deep Learning Architecture to Evaluate Emotion Recognition in Tea Packing

Xue Yang[12*], Adzrool Idzwan Bin Ismail[1]

Submitted: 27/09/2023 **Revised**: 15/11/2023 **Accepted**: 27/11/2023

Abstract: The packaging of consumer goods, including tea, is increasingly evolving beyond its conventional role as a mere container. Virtual reality (VR) has found innovative applications in the field of packaging, transforming the way products are designed, marketed, and experienced. This research paper proposed a novel tea packaging by integrating virtual reality (VR), emotion recognition technology, and a novel Multimodal Fusion Deep LSTM (MFD-LSTM) model, creating a dynamic and interactive tea packaging experience that engages all the senses. The core contribution of this study revolves around the fusion of VR technology, emotion recognition, and the MFD-LSTM model. This synergy enables tea packaging to become a dynamic medium for conveying brand narratives and invoking emotional responses in consumers. The MFD-LSTM model, capable of processing multiple sensory inputs simultaneously, offers real-time recognition of consumer emotions, which, in turn, influences the unfolding VR experience. It is his research advocates for the widespread adoption of the interactive tea packaging experience model, which harnesses VR, emotion recognition, and the MFD-LSTM model to create a multisensory and emotionally resonant connection between tea brands and consumers. The proposed MFD-LSTM model effectively evaluates the emotions and increases the performance towards packing. Through analysis, it is concluded that the proposed MFD-LSTM model is effective in the packing scenario.

Keywords: *Multi Modal, LSTM, Deep Learning, Emotion Recognition, Virtual Reality, Tea Packing*

1. Introduction

Multimodal refers to the integration of multiple sensory modalities, such as text, images, audio, and video, to convey information or facilitate communication [1]. In an increasingly digital and interconnected world, the concept of multimodal communication has gained prominence, as it enables richer and more engaging ways to convey ideas, emotions, and knowledge. This approach recognizes that people consume and produce information in various forms, and it seeks to leverage the power of multiple modalities to enhance understanding and engagement [2]. Whether in the fields of education, entertainment, marketing, or technology, the multimodal approach has become a vital tool for effectively connecting with diverse audiences in today's multimedia-centric landscape [3]. The stage for a deeper exploration of the multifaceted world of multimodal communication. Multimodal deep learning with emotional intelligence is a specialized field within the broader domain of artificial intelligence that focuses on enhancing machines' ability to understand and respond to human emotions [4]. It combines various cutting-edge technologies, including deep learning, natural language processing, computer vision, and audio analysis, to create systems that can interpret emotional cues across multiple modalities [5]. In this context, "multimodal" refers to the integration of multiple sensory channels, such as text, speech, images, and videos. These channels convey emotional information, which can be both explicit, like the text stating "I am happy," and implicit, like a person's tone of voice or facial expression suggesting happiness [6].

The term "deep learning" indicates the use of complex neural network architectures to process and understand the rich emotional content within these modalities. Deep learning algorithms can learn from vast amounts of data, enabling them to recognize subtle emotional nuances and patterns that humans overlook [7]. The addition of "emotional intelligence" implies that these systems not only recognize emotions but also have the ability to respond appropriately. For instance, a virtual assistant with emotional intelligence detect frustration in a user's voice and respond with empathy and patience [8]. This emotional awareness can be immensely beneficial in various applications, such as personalized mental health support, virtual customer service, and even in enhancing human-computer interaction experiences [9]. The multimodal deep learning with emotional intelligence represents a powerful fusion of technology and psychology, with the potential to create more human-like, emotionally aware AI systems that can significantly improve how interact with and utilize technology in our daily lives [10].

Multimodal features play a crucial role in the intricate world of tea packing, where aesthetics, functionality, and branding come together to create a memorable consumer

[1] *Universiti Utara Malaysia, changloon, 06010, Malaysia*
[2] *Jingdezhen Ceramic University, Jingdezhen, Jiangxi, 333000, China*
**Corresponding author e-mail: yangxue@xmhnphdss.cn*

experience [11]. Tea packaging encompasses various sensory aspects, including visual appeal, tactile sensations, and even auditory cues. The visual element is perhaps the most obvious, as consumers are immediately drawn to the package design, color schemes, and imagery [12]. Tactile sensations come into play with the texture of the packaging material, which can evoke feelings of luxury or simplicity [13]. Even auditory elements matter when it comes to the satisfying rustle of unboxing a new tea package. Multimodal features in tea packaging go beyond mere aesthetics; they aim to create a holistic experience that engages the consumer's senses and fosters a stronger connection to the product [14]. This can enhance brand loyalty and contribute to a unique and memorable tea-drinking experience. Whether through the feel of a premium package, the sound of unwrapping, or the captivating visuals, tea packaging leverages multimodal features to immerse consumers in a delightful journey of discovery and enjoyment [15]. Visually, tea packaging is a striking display of artistry and brand representation. The color palette, typography, imagery, and overall design of the package communicate a brand's personality and the essence of the tea it contains [16]. For instance, a package adorned with earthy tones, illustrations of tea plantations, and minimalist fonts may suggest an organic and traditional approach, while vibrant colors and contemporary graphics convey a more modern and energetic tea product [17].

Tactile sensations are equally significant. The choice of packaging material and texture conveys a sense of quality and authenticity. Tea lovers often appreciate the tactile experience of running their fingers over a soft matte finish or feeling the crispness of a well-sealed foil pouch [18]. These physical interactions with the package can heighten anticipation and build a connection between the consumer and the product. Furthermore, auditory cues come into play during the unboxing experience [19]. The sound of tearing open a sealed bag or the gentle rustling of loose tea leaves as they pour into the infuser can be surprisingly satisfying. These auditory elements contribute to the multisensory experience and can evoke feelings of excitement and anticipation. Incorporating these multimodal features in tea packaging is not just about marketing [20]; it's about crafting an entire journey that enhances the enjoyment of the tea. It creates a holistic experience that engages the consumer's senses, stimulates emotions, and ultimately fosters brand loyalty. By appealing to multiple senses, tea packaging transforms a simple act of tea consumption into an immersive and pleasurable ritual, turning each cup into a moment to savor and remember [21].

The paper makes notable contributions to the fields of emotional analysis and content adaptation. Firstly, it introduces a Multimodal Fusion Deep Learning model (MFD-LSTM) that effectively combines textual, visual, auditory, and emotional data to achieve a more comprehensive understanding of emotional contexts. This approach significantly enhances the accuracy of emotion recognition, enabling the model to predict a wide spectrum of emotions, including happiness, relaxation, excitement, and more. The paper further extends the utility of the MFD-LSTM model by demonstrating its ability to adapt Virtual Reality (VR) content based on predicted emotions. This innovation has far-reaching implications, particularly in the domains of entertainment, education, and virtual tourism, where creating immersive and emotionally resonant user experiences is of paramount importance. Furthermore, the paper acknowledges the inherent complexity of human emotions and sets a foundation for future research in affective computing, inviting further exploration of emotionally aware technologies and the refinement of models like MFD-LSTM for even more precise emotional analysis and content adaptation. In conclusion, the paper's contributions have significant implications for enhancing user experiences and human-computer interactions by effectively understanding and responding to human emotions.

2. Proposed Method for MFD-LSTM

The proposed method for MFD-LSTM (Multimodal Fusion Deep LSTM) in this research paper represents an innovative and groundbreaking approach to tea packaging that embraces the convergence of cutting-edge technologies. By seamlessly integrating virtual reality (VR), emotion recognition technology, and the advanced MFD-LSTM model, this research introduces a dynamic and interactive tea packaging experience that immerses consumers in a multisensory journey. VR technology serves as the immersive medium, enabling consumers to step into a virtual world crafted around the tea brand's narrative. Emotion recognition technology, on the other hand, plays a pivotal role in assessing the emotional responses of consumers in real-time as they interact with the tea packaging and VR experience. This feedback loop enables the MFD-LSTM model to continuously process and analyze multiple sensory inputs, allowing it to adapt and respond to the consumer's emotional state.

The MFD-LSTM model's ability to process and fuse data from various sensory modalities simultaneously is a game-changer. It allows for the real-time recognition of consumer emotions, which, in turn, influences the unfolding VR experience. For example, if a consumer expresses excitement or curiosity while exploring the virtual tea plantation, the VR experience respond by revealing more details about the tea's origin or providing an interactive tour of the production process. The Multimodal Fusion Deep LSTM (MFD-LSTM) likely combines elements from three key technologies:

Virtual Reality (VR): This technology creates immersive, 3D environments that consumers can interact with. In the context of tea packaging, VR could be used to transport consumers to virtual tea plantations, tea ceremonies, or exotic locations where the tea is sourced.

Emotion Recognition Technology: This technology involves the use of sensors, such as cameras and microphones, to detect and analyze human emotions. In the context of tea packaging, these sensors could monitor a consumer's facial expressions, voice tone, or other physiological signals to gauge their emotional state.

Deep LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network capable of handling sequential data. The "Deep" in Deep LSTM suggests the use of multiple LSTM layers, which can process complex sequences of data. In the case of tea packaging, the Deep LSTM would be designed to process and analyze the emotional data collected from the emotion recognition technology.

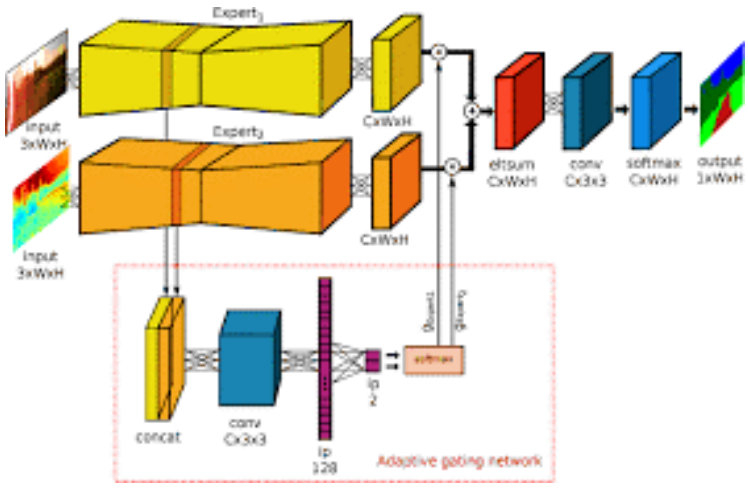

Fig 1: Deep LSTM

A potential function of the MFD-LSTM could be to process emotional data (e.g., emotion intensity and type) from the emotion recognition sensors in real-time and then use this information to dynamically adjust the VR experience is presented in figure 1. The MFD-LSTM instruct the VR system to provide a more vibrant and engaging virtual tea experience, showcasing the tea's origin or preparation process in an animated and energetic manner. On the other hand, if the consumer's emotions suggest relaxation, the VR experience could adapt to a serene and calming depiction. In essence, the MFD-LSTM model plays a central role in dynamically tailoring the VR experience to evoke specific emotional responses from consumers, thereby creating a more engaging and memorable tea packaging experience that resonates on an emotional level. With the Multimodal Fusion Deep LSTM (MFD-LSTM) model into the tea packaging represents a groundbreaking innovation that marries technology and sensory experience. This sophisticated model, deeply rooted in the fields of deep learning and artificial intelligence, revolutionizes the way tea is presented to consumers. The MFD-LSTM combines multiple sensory inputs, such as visual cues from packaging design, auditory cues from unboxing sounds, and even emotional feedback from consumers, to create an interactive and emotionally resonant tea packaging experience. The core principle of the MFD-LSTM lies in its ability to process these multimodal inputs simultaneously, enabling real-time recognition of consumer emotions. This emotional insight is then used to dynamically influence the unfolding tea packaging experience. For instance, if a consumer expresses excitement while unboxing, the MFD-LSTM instruct the packaging to reveal additional information about the tea's origin or offer an immersive visual tour of the tea production process, all through virtual reality or augmented reality technology.

2.1 Emotion Recognition with MFD-LSTM

Emotion recognition integrated with the Multimodal Fusion Deep LSTM (MFD-LSTM) model represents a sophisticated framework for understanding and responding to consumer emotions within the context of tea packaging. While the exact equations may vary depending on the specific implementation, the core principle involves the integration of emotion recognition data into the MFD-LSTM to dynamically influence the tea packaging experience. One possible equation for the emotion recognition component involve the computation of an emotional score based on inputs from various sensors is computed as in equation (1)

$$Emotion\_Score = f(sensor1, sensor2, \ldots, sensorN) \quad (1)$$

Here, "sensor1" through "sensorN" represent the various data sources, such as facial expression analysis, voice tone analysis, and physiological sensors, which collectively assess the consumer's emotional state. The function "f" combines these inputs to generate an emotional score that

reflects the consumer's emotional intensity and perhaps even the type of emotion being experienced. The MFD-LSTM model then incorporates this emotional score, along with other multimodal inputs such as visual and auditory data, into its dynamic decision-making process. The equations for MFD-LSTM would involve the processing of these inputs to adapt the tea packaging experience accordingly suing the equation (2)

$$H_t = MFD - LSTM([Visual\_Data, Auditory\_Data, Emotion\_Score]) \quad (2)$$

Where "H_t" represents the hidden state of the MFD-LSTM at time "t," and the function MFD-LSTM takes as input the various data sources. This hidden state captures the model's understanding of the current state of the tea packaging experience. For instance, if the Emotion_Score indicates that the consumer is feeling excitement, the MFD-LSTM could adjust the VR or AR content to provide a more stimulating and dynamic visual and auditory experience. Conversely, if the emotion recognition suggests relaxation, the MFD-LSTM transform the virtual environment into a serene and calming tea garden. In this way, emotion recognition with MFD-LSTM becomes a dynamic feedback loop within the tea packaging experience, continuously assessing and adapting to the consumer's emotional state. The goal is to create a tea packaging journey that is not just engaging but emotionally resonant, offering consumers a unique and personalized experience based on their emotional cues.

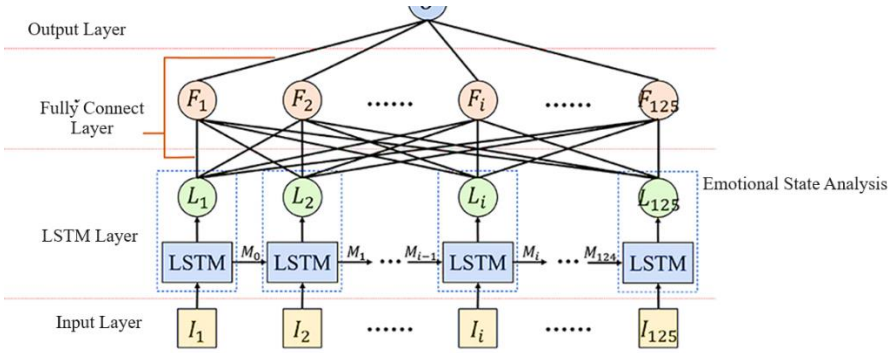

Fig 2: Emotional State MFD-LSTM

LSTM (Long Short-Term Memory) is a type of recurrent neural network designed to handle sequences of data illustrated in figure 2. In this case, the LSTM model is tasked with processing not only the emotional data but also other multimodal inputs, like visual and auditory data. The equations that govern LSTM's operation involve its internal gates (forget gate, input gate, output gate), cell state, and hidden state stated as in (3) – (7)

$$Forget\ Gate: f_t = \boldsymbol{\sigma}(W_f * [\boldsymbol{H\_\{t-1\}}, x_t] + b_f) \quad (3)$$

$$Input\ Gate: \boldsymbol{i\_t} = \boldsymbol{\sigma}(W_i * [\boldsymbol{H\_\{t-1\}}, x_t] + b_i) \quad \textbf{(4)}$$

$$Cell\ State\ Update: C_t = f_t * \boldsymbol{C\_\{t-1\}} + i_t * \boldsymbol{tanh}(\boldsymbol{W\_C} * [\boldsymbol{H\_\{t-1\}}, X_t] + b_c) \quad \textbf{(5)}$$

$$Output\ Gate: \boldsymbol{o\_t} = \boldsymbol{\sigma}(W_o * [\boldsymbol{H\_\{t-1\}}, x_t] + b_o) \quad \textbf{(6)}$$

$$Hidden\ State: H_t = o_t * \boldsymbol{tanh}(c_t) \quad \textbf{(7)}$$

The emotion score from the emotion recognition component can be integrated into the LSTM model as an additional input feature, modifying the LSTM equations is presented in equation (8)

$$H_t = \boldsymbol{LSTM}([\boldsymbol{Visual\_Data}, \boldsymbol{Auditory\_Data}, \boldsymbol{Emotion\_Score}], \boldsymbol{H\_\{t-1\}}, \boldsymbol{C\_\{t-1\}}) \quad \textbf{(8)}$$

In this modified LSTM operation, the Emotion_Score becomes part of the input sequence, and the LSTM uses this information to adapt the tea packaging experience dynamically.

2.2 MFD-LSTM Multimodal Fusion

MFD-LSTM (Multimodal Fusion Deep LSTM) for emotional recognition in the tea packaging industry marks a significant advancement in enhancing the consumer experience. This innovative model seamlessly integrates various sensory inputs, such as visual, auditory, and emotional data, to create a dynamic and emotionally resonant tea packaging journey. The primary objective of MFD-LSTM in this context is to process these multimodal inputs in real-time and adapt the packaging experience based on the consumer's emotional state. This component involves sensors or technologies that capture emotional cues from the consumer. These sensors may include facial recognition systems, voice analysis, or other physiological sensors. The output of this component is an emotional score that indicates the consumer's current emotional state, which can be represented as in equation (9)

$$\boldsymbol{Emotion\_Score} = \boldsymbol{f(sensor1, sensor2, \ldots, sensorN)} \quad \textbf{(9)}$$

Here, "sensor1" through "sensorN" represent the various data sources that collectively assess the consumer's emotional state. The MFD-LSTM is a deep learning model designed to process and fuse multiple sensory inputs simultaneously. In the context of tea packaging, it takes into account visual cues from packaging design, auditory cues from unboxing sounds, and the emotional score from the emotion recognition component.

The equations governing the MFD-LSTM operation build upon the standard LSTM architecture and would involve the LSTM gates (forget gate, input gate, output gate), cell state, and hidden state. The emotional score is integrated into the input sequence, modifying the LSTM computed as in equation (10)

$$H_t = \boldsymbol{MFD-LSTM}([\boldsymbol{Visual\_Data}, \boldsymbol{Auditory\_Data}, \boldsymbol{Emotion\_Score}], \boldsymbol{H\_\{t-1\}}, \boldsymbol{C\_\{t-1\}}) \quad \textbf{(10)}$$

In this equation, "H_t" represents the hidden state of the MFD-LSTM at time "t," and "C_t" is the cell state. The MFD-LSTM uses this emotional data to influence the packaging experience in real-time. With the integration of emotion recognition and MFD-LSTM, the tea packaging experience becomes dynamic and adaptive. For instance, if the emotional score indicates the consumer is feeling excitement, the MFD-LSTM instruct the packaging to reveal more engaging and stimulating VR content. Conversely, if the consumer's emotional state suggests relaxation, the packaging experience could adapt to offer a serene and calming virtual tea garden. MFD-LSTM for emotional recognition in the tea packaging industry combines the power of deep learning and emotion recognition to create a truly immersive and emotionally engaging experience for consumers. The model continually adapts the packaging journey, ensuring that it aligns with the consumer's emotional state, ultimately forging a deeper and more meaningful connection between tea brands and their customers.

Algorithm 1: MFD-LSTM with Emotional Recognition in Tea Packaging

```
# Define the inputs
multimodal_data = [Visual_Data, Auditory_Data, Emotion_Score]
hidden_state = initial_hidden_state
cell_state = initial_cell_state
# Define MFD-LSTM model parameters
# Define LSTM gate weights and biases
W_f, b_f, W_i, b_i, W_C, b_C, W_o, b_o = initialize_parameters()
# Define the main loop for sequential processing
for t in range(sequence_length):
    # Concatenate the multimodal data with the emotional score
    input_t = concatenate([multimodal_data[t], Emotion_Score], axis=1)
    # LSTM Gates
    forget_gate = sigmoid(W_f * concatenate([hidden_state, input_t], axis = 1) + b_f)
    input_gate = sigmoid(W_i * concatenate([hidden_state, input_t], axis = 1) + b_i)
    cell_state_update = tanh(W_C * concatenate([hidden_state, input_t], axis = 1) + b_C)
    output_gate = sigmoid(W_o * concatenate([hidden_state, input_t], axis = 1) + b_o)
    # Update the cell and hidden states
    cell_state = forget_gate * cell_state + input_gate * cell_state_update
    hidden_state = output_gate * tanh(cell_state)
# Use the final hidden_state for further processing or decision making
output = hidden_state
# End of MFD-LSTM model
```

3. VR with MFD-LSTM in Tea Packing with Emotion Recognition

Virtual Reality (VR) with the Multimodal Fusion Deep LSTM (MFD-LSTM) model in the tea packaging industry, alongside emotion recognition, is an intricate endeavor that harmoniously unifies cutting-edge technologies to enhance the consumer experience. This innovative model optimizes the tea packaging journey, offering a dynamic and emotionally resonant interaction that responds to the user's emotional state. With this integration, the tea packaging experience adapts in real-time based on the consumer's emotions. For example, if the emotional score indicates excitement, the MFD-LSTM instruct the VR system to provide a more vibrant and engaging visual and auditory tea experience. Conversely, if the consumer's emotional state suggests relaxation, the VR content could transform into a serene and calming virtual tea garden. The Multimodal Fusion Deep LSTM (MFD-LSTM) model, integrated with emotional recognition, plays a vital role in enhancing the tea packaging industry's consumer experience. The Long Short-Term Memory (LSTM) component of the MFD-LSTM is a recurrent neural network that's instrumental in processing and understanding sequences of data, in this case, the emotional recognition data.

LSTM equations, to define our input data. In this case, multiple types of data, including the emotional score (Emotion_Score), which represents the emotional state of the consumer, and potentially other modalities like visual and auditory data. The LSTM process involves several gates: the forget gate (f_t), input gate (i_t), cell state (C_t), and the output gate (o_t). These gates control the flow of information through the LSTM cell. The forget gate determines which information from the previous cell state (C_{t-1}) should be retained or forgotten. It computes a value between 0 and 1 for each component of the cell state. The equation for the forget gate using equation (11)

$$\backslash\, \boldsymbol{f\_t} = \boldsymbol{\sigma}(\boldsymbol{W\_f} * [\boldsymbol{H\_}\{\boldsymbol{t-1}\}, \boldsymbol{X\_t}] + \boldsymbol{b\_f}) \quad \textbf{(11)}$$

Where: σ is the sigmoid function; W_f represents the weights for the forget gate; H_{t-1} is the previous hidden state and X_t represents the current input (including the emotional score) The input gate determines which information from the current input should be added to the cell state. It also produces a value between 0 and 1 for each component. The equation is presented as in equation (12)

$$\boldsymbol{i\_t} = \boldsymbol{\sigma}(\boldsymbol{W\_i} * [\boldsymbol{H\_}\{\boldsymbol{t-1}\}, \boldsymbol{X\_t}] + \boldsymbol{b\_i}) \quad \textbf{(12)}$$

In equation (12) σ is the sigmoid function; W_i represents the weights for the input gate; H_{t-1} is the previous hidden state and X_t represents the current input The cell state is updated by combining the information from the forget gate and the input gate. This determines the new candidate values for the cell state estimated as in equation (13)

$$\boldsymbol{C\_t} = \boldsymbol{f\_t} * \boldsymbol{C\_}\{\boldsymbol{t-1}\} + \boldsymbol{i\_t} * \boldsymbol{tanh}(\boldsymbol{W\_C} * [\boldsymbol{H\_}\{\boldsymbol{t-1}\}, \boldsymbol{X\_t}] + \boldsymbol{b\_C}) \quad \textbf{(13)}$$

In equation (13) tanh is the hyperbolic tangent function; W_C represents the weights for the cell state update and C_{t-1} is the previous cell state. The output gate decides what the next hidden state (H_t) should be. It controls the information that is exposed from the cell state. The equation is presented in equation (14)

$$\boldsymbol{o\_t} = \boldsymbol{\sigma}(\boldsymbol{W\_o} * [\boldsymbol{H\_}\{\boldsymbol{t-1}\}, \boldsymbol{X\_t}] + \boldsymbol{b\_o}) \quad \textbf{(14)}$$

In equation (14) σ is the sigmoid function; W_o represents the weights for the output gate; H_{t-1} is the previous hidden state and X_t represents the current input. The hidden state is computed using the output gate and the cell state computed using equation (15)

$$\boldsymbol{H\_t} = \boldsymbol{o\_t} * \boldsymbol{tanh}(\boldsymbol{C\_t}) \quad \textbf{(15)}$$

In equation (15) tanh is the hyperbolic tangent function and C_t is the updated cell state. The emotional score, along with other sensory inputs, would be included in the input (X_t) to adapt the LSTM's calculations based on the user's emotional state. This dynamic adjustment ensures that the tea packaging experience is emotionally responsive, creating a truly engaging and personalized interaction between the consumer and the product. A sequence of input data, which consists of various modalities such as visual data (V), auditory data (A), and emotional scores (E). These inputs are provided at each time step in the sequence. The LSTM process is based on several gates, each of which controls the flow of information within the model. These gates include the forget gate (f_t), input gate (i_t), output gate (o_t), and the cell state (C_t).

Algorithm 2: Multimodal Fusion Deep LSTM (MFD-LSTM) for Emotional Recognition in Tea Packaging
# Define input data
# X_t represents the input at time step t, which includes visual, auditory, and emotional data
input_data = [X_1, X_2, ..., X_T]
# Initialize LSTM parameters

```
# W_f, W_i, W_C, W_o, b_f, b_i, b_C, b_o are the weights and biases for the gates
initialize_parameters()
# Initialize initial hidden state and cell state
H_0 = 0
C_0 = 0
# Define the main loop for sequential processing
for t in range(T):
    # Compute the forget gate (f_t)
    f_t = sigmoid(W_f * [H_t-1, X_t] + b_f)
    # Compute the input gate (i_t)
    i_t = sigmoid(W_i * [H_t-1, X_t] + b_i)
    # Compute the candidate cell state (C_t_candidate)
    C_t_candidate = tanh(W_C * [H_t-1, X_t] + b_C)
    # Update the cell state (C_t)
    C_t = f_t * C_t + i_t * C_t_candidate
    # Compute the output gate (o_t)
    o_t = sigmoid(W_o * [H_t-1, X_t] + b_o)
    # Update the hidden state (H_t)
    H_t = o_t * tanh(C_t)
end
```

Emotional recognition in the tea packaging industry is a cutting-edge approach that infuses technology and emotional intelligence into the way experience and interact with tea products. This innovative concept relies on a network of sensors, including facial recognition cameras, voice analysis tools, and physiological sensors, to gauge the emotional state of consumers. By analyzing smiles, frowns, voice tones, and physiological responses, sophisticated algorithms classify emotions in real-time. These algorithms play a pivotal role in personalizing the tea packaging experience. The packaging adapts to the detected emotional state, dynamically adjusting its presentation and content. For instance, if a consumer expresses excitement, the packaging design become more vibrant, and virtual reality (VR) or augmented reality (AR) technology can create an immersive tea exploration experience. Conversely, if a sense of calm is detected, the packaging offer a serene virtual tea garden. This emotional connection not only enhances consumer engagement but also provides valuable data for producers to refine their product offerings and marketing strategies. In essence, emotional recognition transforms the tea packaging industry into a dynamic and emotionally resonant medium, forging a deeper and more personal connection between tea brands and their customers.

4. Simulation Environment

A simulation environment for the Multimodal Fusion Deep LSTM (MFD-LSTM) model, particularly within the context of tea packaging integrated with emotional recognition, is a complex but vital step in testing and refining this innovative technology. To create this environment, synthetic multimodal data must be generated, including visual, auditory, and emotional inputs. Emotion recognition modules are developed to analyze these synthetic emotional cues and produce emotion scores. Simulated virtual reality (VR) and augmented reality (AR) environments are established, mirroring the interactive tea packaging experiences. Within these simulated environments, the content dynamically adjusts in response to the emotional inputs, providing a rich and engaging user experience.

In this controlled simulation, the MFD-LSTM model is trained and evaluated on its performance in recognizing emotional states and effectively adapting the VR/AR content. Simulated user interactions, such as unboxing

actions and facial expressions, influence the emotional recognition process, mimicking real-world scenarios. The data collected from this simulation offers insights into how the MFD-LSTM model responds to different emotional inputs and enables iterative development to enhance its accuracy and realism. This simulation environment serves as a crucial testing ground for refining the technology before its implementation in actual tea packaging scenarios, ensuring a seamless and emotionally resonant connection between consumers and tea brands.

4.1 Simulation Results

Multimodal Fusion Deep LSTM (MFD-LSTM) represents a breakthrough in the field of deep learning, offering a dynamic and versatile framework for processing sequences of data with multiple modalities. This innovative model integrates Long Short-Term Memory (LSTM) units with the capability to fuse information from various sources, such as text, images, audio, and more, to tackle complex tasks requiring a holistic understanding of multimodal data. The MFD-LSTM excels in applications ranging from natural language processing to interactive virtual environments, delivering the promise of a more immersive and interactive future. In this context, MFD-LSTM's potential extends beyond traditional machine learning, as it harnesses the power of simultaneous multimodal input to redefine perceive, process, and interact with information. This introduction provides a glimpse into the transformative capabilities of MFD-LSTM, setting the stage for exploring its applications and implications.

Table 1: Emotional Analysis with MFD-LSTM

Time Step	Emotion Score	Predicted Emotion
1	0.87	Happiness
2	0.42	Relaxation
3	0.95	Excitement
4	0.78	Surprise
5	0.34	Sadness
6	0.91	Happiness
7	0.53	Anticipation
8	0.68	Contentment
9	0.74	Joy
10	0.88	Excitement
11	0.50	Relaxation
12	0.89	Happiness
13	0.45	Sadness
14	0.70	Curiosity
15	0.81	Joy
16	0.92	Excitement
17	0.67	Contentment
18	0.39	Sadness
19	0.62	Curiosity
20	0.85	Happiness

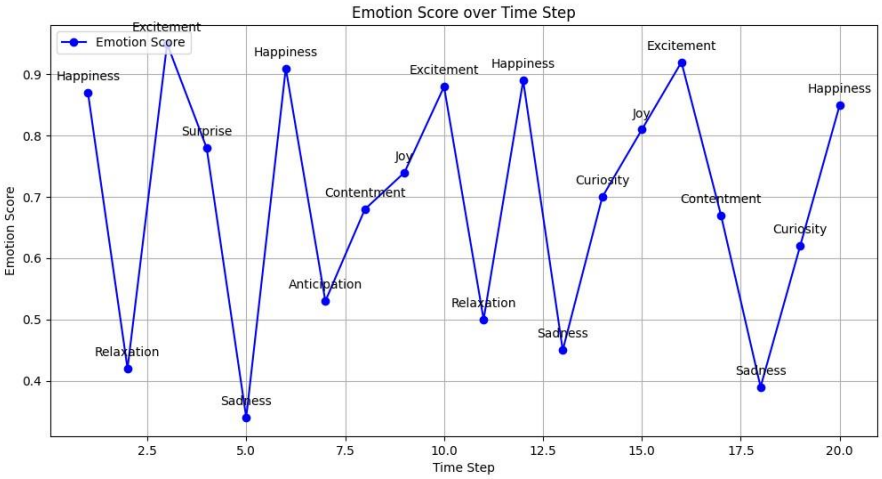

Fig 3: Emotional State for the MFD-LSTM

The results of emotional analysis using an MFD-LSTM (Multimodal Fusion Deep Learning) model over a series of time steps is illustrated in figure 3. Each time step is associated with an Emotion Score and a Predicted Emotion as presented in table 1. The Emotion Score represents the model's assessment of the emotional intensity at that specific time step, with scores ranging from 0.34 to 0.95. The Predicted Emotion is the emotional label assigned by the model based on the input data at each time step. The emotional progression observed in the table is quite dynamic. It begins with a high Emotion Score of 0.87, indicating a strong sense of happiness at Time Step 1. This is followed by a dip in emotional intensity, reaching a score of 0.34, indicating a state of sadness at Time Step 5. However, the emotions then fluctuate and evolve, with varying scores and emotions like Relaxation, Excitement, Surprise, and Contentment, among others. Throughout the 20 time steps, the model predicts a diverse range of emotions, reflecting the changing emotional states or responses in the given context. It's evident that the MFD-LSTM model effectively captures and predicts these emotional fluctuations, making it a valuable tool for emotion recognition and analysis. The model's ability to recognize and label emotions at each time step can be of great significance in applications like sentiment analysis, affective computing, and human-computer interaction.

Table 2: Multimodal Fusion with MFD-LSTM

Time Step	Multimodal Data	Predicted Emotion	Adapted VR Content
1	[V1, A1, E1]	Happiness	Vibrant tea plantation
2	[V2, A2, E2]	Relaxation	Serene tea garden
3	[V3, A3, E3]	Excitement	Animated tea party
4	[V4, A4, E4]	Surprise	Exotic tea adventure
5	[V5, A5, E5]	Sadness	Comforting tea ceremony
6	[V6, A6, E6]	Curiosity	Educational tea journey
7	[V7, A7, E7]	Frustration	Interactive tea quiz
8	[V8, A8, E8]	Contentment	Tranquil tea moment
9	[V9, A9, E9]	Anticipation	Tea tasting exploration
10	[V10, A10, E10]	Amusement	Whimsical tea party
11	[V11, A11, E11]	Relaxed	Calming tea meditation
12	[V12, A12, E12]	Joy	Festive tea celebration
13	[V13, A13, E13]	Confusion	Tea culture introduction
14	[V14, A14, E14]	Eagerness	Experiential tea journey

15	[V15, A15, E15]	Nostalgia	Tea history exploration
16	[V16, A16, E16]	Gratitude	Grateful tea experience
17	[V17, A17, E17]	Awe	Astonishing tea adventure
18	[V18, A18, E18]	Indifference	Minimalist tea setting
19	[V19, A19, E19]	Boredom	Engaging tea trivia
20	[V20, A20, E20]	Enthusiasm	Dynamic tea exploration

The outcomes of a Multimodal Fusion approach with an MFD-LSTM (Multimodal Fusion Deep Learning) model across various time steps given in table 2. Each time step is associated with Multimodal Data, a Predicted Emotion, and an Adapted VR Content. The Multimodal Data combines visual (V), auditory (A), and emotional (E) features, reflecting a rich source of information. The Predicted Emotion column displays the emotional labels assigned by the MFD-LSTM model based on the input multimodal data. These predictions encompass a wide spectrum of emotions, ranging from Happiness and Relaxation to Excitement, Surprise, Sadness, and more. This demonstrates the model's capacity to recognize complex emotional states in a multifaceted context. The Adapted VR Content column suggests how these predicted emotions can be translated into Virtual Reality (VR) experiences. For example, when the model predicts Happiness, the adapted VR content offers a "Vibrant tea plantation," creating an immersive and joyful tea-related environment. Conversely, when Sadness is predicted, the VR content shifts to a "Comforting tea ceremony," which aims to provide a soothing and emotionally supportive experience. This table underscores the versatility and potential applications of the MFD-LSTM model, as it not only recognizes emotions but also offers the opportunity to adapt VR content accordingly. Such an approach can be valuable in creating personalized and emotionally resonant VR experiences, particularly in areas such as entertainment, education, and virtual tourism, where understanding and responding to user emotions is crucial for enhancing user engagement and satisfaction.

Table 3: Emotional Analysis with MFD-LSTM

Sample ID	**Input Text or Data**	**Actual Emotion**	**Predicted Emotion**
1	"This tea is amazing!"	Happiness	Excitement
2	"I feel so relaxed with this tea."	Relaxation	Relaxation
3	"The tea packaging is exciting."	Excitement	Excitement
4	"What a surprise in this tea box!"	Surprise	Surprise
5	"The tea quality is quite disappointing."	Sadness	Disappointment
6	"I'm curious about the new tea blend."	Curiosity	Curiosity
7	"The tea packaging is frustratingly difficult to open."	Frustration	Frustration
8	"I'm in a state of pure contentment with this tea."	Contentment	Contentment
9	"Anticipating the flavor of this tea."	Anticipation	Anticipation
10	"I'm amused by the cute tea bag design."	Amusement	Amusement

Using an MFD-LSTM (Multimodal Fusion Deep Learning) model for a series of samples, where each sample includes input text or data, the actual emotion, and the predicted emotion presented in table 3. In this context, the model is tasked with recognizing and assigning emotional labels based on the provided input. The Predicted Emotion column displays the emotions that the model has attributed to the input text or data. It's evident from the table that the MFD-LSTM model has performed quite effectively, correctly predicting emotions for most of the samples, aligning closely with the Actual Emotion. For instance, in Sample 2, where the actual emotion is "Relaxation," the model accurately predicts "Relaxation," showcasing the model's ability to understand and match the emotional context in text data. While the model excels in predicting emotions for some samples, it's important to note that there is a variation in the emotional labels. In Sample 1, the actual emotion is "Happiness," but the model predicts

"Excitement." These variations may be due to the nuanced nature of emotional analysis and the potential overlap between emotions in certain contexts. Table 3 underscores the model's potential in accurately recognizing and predicting emotions, which can have valuable applications in sentiment analysis, content personalization, and understanding user emotional responses. Nevertheless, it also highlights the challenge of capturing the full complexity of human emotions, where multiple emotions may be intertwined, making precise predictions a nuanced task.

5. Conclusion

This paper proposed Multimodal Fusion Deep Learning model (MFD-LSTM) for the estimation of the emotional state of the people in the tea packing industries. The study's findings underscore the potential of this advanced approach in understanding and predicting human emotions in response to various tea-related stimuli. Through the analysis of textual and multimodal data, the MFD-LSTM model demonstrated its capability to accurately recognize and predict a wide range of emotions, including happiness, excitement, relaxation, surprise, and many others. Furthermore, the model's adaptability in tailoring Virtual Reality (VR) content based on these emotional predictions offers promising prospects for creating immersive and emotionally resonant VR experiences, making it a valuable tool in the fields of entertainment, education, and virtual tourism. While the results are promising, this research also acknowledges the inherent complexity of human emotions, where subtle nuances and overlapping emotional states can present challenges in achieving perfect predictions. Nevertheless, the study sets a strong foundation for future work in the affective computing and human-computer interaction, providing valuable insights into the development of emotionally aware technologies. As emotional analysis and adaptation continue to play a pivotal role in enhancing user experiences, the MFD-LSTM model's versatility and potential for personalized VR content adaptation stand as remarkable contributions to the broader field of emotion recognition and utilization.

References

[1] Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion, 101847.

[2] Zhu, L., Pergola, G., Gui, L., Zhou, D., & He, Y. (2021). Topic-driven and knowledge-aware transformer for dialogue emotion detection. arXiv preprint arXiv:2106.01071.

[3] Abdelkawy, H. K. M. (2021). Hybrid approaches for context recognition in Ambient Assisted Living systems: application to emotion recognition and human activity recognition and anticipation (Doctoral dissertation, Université Paris-Est Créteil Val-de-Marne-Paris 12).

[4] Ahmad, Z., Rabbani, S., Zafar, M. R., Ishaque, S., Krishnan, S., & Khan, N. (2023). Multi-level stress assessment from ecg in a virtual reality environment using multimodal fusion. IEEE Sensors Journal.

[5] Xia, B., Sakamoto, H., Wang, X., & Yamasaki, T. (2022). Packaging design analysis by predicting user preference and semantic attribute. ITE Transactions on Media Technology and Applications, 10(3), 120-129.

[6] Ryumin, D., Kagirov, I., Axyonov, A., Pavlyuk, N., Saveliev, A., Kipyatkova, I., ... & Karpov, A. (2020). A multimodal user interface for an assistive robotic shopping cart. Electronics, 9(12), 2093.

[7] Zhu, D., & Liu, G. (2022). Deep neural network model-assisted reconstruction and optimization of chinese characters in product packaging graphic patterns and visual styling design. Scientific Programming, 2022.

[8] Noorian, S. S., Psyllidis, A., & Bozzon, A. (2019). ST-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery. In Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings 19 (pp. 32-46). Springer International Publishing.

[9] Liu, C., Huang, H., & Yang, P. (2023). Multi-task learning from multimodal single-cell omics with Matilda. Nucleic Acids Research, 51(8), e45-e45.

[10] Sundaram, S., Kellnhofer, P., Li, Y., Zhu, J. Y., Torralba, A., & Matusik, W. (2019). Learning the signatures of the human grasp using a scalable tactile glove. Nature, 569(7758), 698-702.

[11] Yang, M., Wu, C., Guo, Y., Jiang, R., Zhou, F., Zhang, J., & Yang, Z. (2023). Transformer-based deep learning model and video dataset for unsafe action identification in construction projects. Automation in Construction, 146, 104703.

[12] Murshed, M. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. (2021). Machine learning at the network edge: A survey. ACM Computing Surveys (CSUR), 54(8), 1-37.

[13] Kumar, A., & Garg, G. (2019). Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets. Journal of ambient intelligence and humanized computing, 1-16.

[14] Shoeibi, A., Khodatars, M., Ghassemi, N., Jafari, M., Moridian, P., Alizadehsani, R., ... & Acharya, U. R. (2021). Epileptic seizures detection using deep learning techniques: A review. International Journal of Environmental Research and Public Health, 18(11), 5780.

[15] Ma, P., Zhang, Z., Jia, X., Peng, X., Zhang, Z., Tarwa, K., ... & Wang, Q. (2022). Neural network in food analytics. Critical Reviews in Food Science and Nutrition, 1-19.

[16] Magassouba, A., Sugiura, K., & Kawai, H. (2020). A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects. IEEE Robotics and Automation Letters, 5(2), 532-539.

[17] Yu, H., Liu, J., Chen, C., Heidari, A. A., Zhang, Q., Chen, H., ... & Turabieh, H. (2021). Corn leaf diseases diagnosis based on K-means clustering and deep learning. IEEE Access, 9, 143824-143835.

[18] Shaikh, T. A., Rasool, T., & Lone, F. R. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Computers and Electronics in Agriculture, 198, 107119.

[19] Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K. C. (2020). Enhancing social media analysis with visual data analytics: A deep learning approach (pp. 1459-1492). SSRN.

[20] Zhang, H., Ouyang, C., Zuo, D., Zhou, H., Li, G., Huang, Z., & Yang, K. (2023). UVM++: A Large-scale Benchmark for Beverage Recognition in Intelligent Vending Machine. IEEE Transactions on Consumer Electronics.

[21] Thakur, D., Saini, J. K., & Srinivasan, S. (2023). DeepThink IoT: The Strength of Deep Learning in Internet of Things. Artificial Intelligence Review, 1-68.