International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

BeatLens: A Context-Aware Vision-to-Music Framework for Image-Based Song Recommendations

[1]Aditya Arolkar, [2]Dhaval Smart, [3]Gaurav Waghmare, [4]Pratham Atale, [5]Prof. Sonali Despande

[1,2,3,4]Student, Smt. Indira Gandhi College of Engineering, Ghansoli, New Mumbai, Maharashtra, India

[5]Professor, Smt. Indira Gandhi College of Engineering, Ghansoli, New Mumbai, Maharashtra, India

***Abstract* - BeatLens is a song recommendation engine based on AI that is created to boost social media storytelling through the automation of music selection for Instagram stories. It solves the typical problem of taking too much time to select songs by using uploaded images via sophisticated computer vision models such as YOLO (for object detection) and CLIP (for scene classification) to decipher visual context. The system then uses Large Language Models (LLMs) like LLaMA 3, LLaVA, and Mistral to suggest songs based on the mood, theme, and setting of the image. For maximum accessibility, BeatLens is available in 14 languages, namely English, Marathi, Hindi, Spanish, Punjabi, Bhojpuri, Korean, German, Portuguese, Japanese, Tamil, Telugu, Kannada, and Malayalam. This multilingual functionality, paired with its AI-powered analysis, turns song choosing into an intuitive, streamlined process—improving user experience and minimizing decision fatigue.**

Keywords: AI-powered Recommendation, Image Analysis, Object Detection, Scene Recognition, LLM, User Experience, YOLO.

I. INTRODUCTION

BeatLens is an AI-driven song recommendation platform that aims to transform social media storytelling through automated music selection for Instagram stories. It solves the universal problem of discovering songs that exactly fit the mood and context of visual content. Through the use of sophisticated computer vision models such as YOLO for object detection and CLIP for scene classification, BeatLens identifies key visual features and contextual information from uploaded images. This information is passed through Large Language Models (LLMs) like LLaMA 3, LLaVA, and Mistral to suggest tracks that match the mood, tone, and theme of the image. Beyond this, 14 languages—English, Marathi, Hindi, Spanish, Punjabi, Bhojpuri, Korean, German, Portuguese, Japanese, Tamil, Telugu, Kannada, and Malayalam—are supported by BeatLens to cater to a multilingual audience. By streamlining the time-consuming task of song selection, BeatLens makes it an effortless and intelligent process. This new concept not only adds value to the user experience but also creates new possibilities for research on cross-domain personalization algorithms, providing an unprecedented combination of visual and auditory modalities that can revolutionize multimedia content generation on social media.

1.1 Project Aims and Objectives

Project Aim:

BeatLens aims to develop an AI-driven song recommendation engine that uses computer vision to analyze images and suggest music, enhancing user experience on social media platforms by automating the song selection process.

Project Objectives:

The major objective of BeatLens is to create a novel AI-driven song recommendation engine that maximizes user experience on social media platforms through the automation of visual content music selection. The following are five major objectives:

1. Computer Vision and AI Integration: BeatLens seeks to integrate computer vision models such as YOLO and CLIP with Large Language Models (LLMs) in a seamless manner to examine images and recommend music. The integration will allow the system to comprehend both the visual context and the user's preferences.
2. Personalized Music Suggestions: By using LLMs to understand visual signals, BeatLens will offer personalized music suggestions that align with the mood, atmosphere, and theme of the images uploaded. This method guarantees that users get customized suggestions that enrich their multimedia content.
3. User Experience Enhancement: The initiative aims to redefine the annoying experience of searching manually for music as a seamless and smart one. By eliminating song choice decisions, BeatLens tries to minimize decision fatigue and enhance user satisfaction in social media environments.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

4. Scalability and Accessibility: BeatLens is built to be scalable on different social media websites and hence is accessible to many users. Scalability will provide the advantages of AI music recommendation to the masses.
5. Innovation in Cross-Modal Recommendations: Through innovating a vision-to-music recommendation system, BeatLens plays a role in the innovation of cross-modal personalization algorithms. This innovation offers new research opportunities in multimedia content consumption and production, potentially revolutionizing user experiences with social media sites.

1.2 System Objectives

The BeatLens system is built with a number of core goals that seek to transform the way users choose music for their social media posts. Below are four main system goals:

1. Visual Context Analysis: BeatLens utilizes sophisticated computer vision models like YOLO for object detection and CLIP for scene classification. The models inspect uploaded images to identify valuable visual cues such as objects, surroundings, and activities. It is an important step in comprehending the context and theme of the visual data.
2. Personalized Music Suggestions: Through the use of Large Language Models (LLMs) such as LLaMA 3, LLaVA, and Mistral, BeatLens converts the visual insights to personalized music suggestions. These models map the visual elements identified to musical genres, mood, and cultural movements in a way that the recommended songs match the mood and topic of the image.
3. User Experience Enhancement: One of the main objectives of BeatLens is to streamline the process of choosing songs automatically and, thus, minimize decision fatigue and increase user satisfaction on social media websites. Through appropriate music suggestions, the users will spend less time on searching for that perfect song, leaving them to concentrate on designing compelling content.
4. Innovation in Cross-Modal Recommendations: BeatLens innovates a vision-to-music recommendation model, advancing the development of cross-modal personalization algorithms. The new approach presents novel research and development directions in multimedia content production and consumption, with the potential to revolutionize how users interact with social media.

1.3 Project Background

The rise of social media has changed the way we share moments from our lives, with social sites like Instagram now being part of our everyday routines. Perhaps the most used feature on such sites is the feature to add music to stories, which maximizes the emotional appeal and engagement of visual content. Nevertheless, choosing the appropriate song that appeals to the mood and theme of an image can be frustrating and time-consuming for users.

Classic music recommendation methods mainly depend on collaborative filtering or content-based filtering, targeting user behavior or audio attributes. These methods overlook the deep contextual signals inherent in visual media, resulting in a gap between the visual and audio components of multimedia content.

BeatLens fills this gap by introducing a new vision-to-music recommendation paradigm. With the help of state-of-the-art computer vision models such as YOLO for object recognition and CLIP for scene identification, BeatLens inspects images uploaded to provide insightful visual signals. These visual signals are further interpreted as customized music suggestions by Large Language Models (LLMs) such as LLaMA 3, LLaVA, and Mistral.

This novel method not only improves the user experience through automated song choice but also introduces new research opportunities for cross-modal personalization algorithms. By fusing visual and auditory modalities, BeatLens presents a scalable solution that can be used on many social media platforms, and could potentially revolutionize multimedia content creation and usage by users.

II. SOFTWARE COMPONENTS

The BeatLens framework has a few core software modules:

- Computer Vision Models: CLIP for scene classification and YOLO for object detection.
- Large Language Models (LLMs): OLLAMA models to convert visual understanding into music recommendations.
- Caching Mechanisms: For performance optimization by storing and loading data in an efficient manner.
- User Interface: Streamlit for building an interactive user interface to upload the image and display recommendations.

III. METHODOLOGY

1. Data Collection Techniques

Primary Data: Pictures were shared on an AI system that leveraged YOLO and CLIP models to detect objects and scenes. Object detection within pictures was carried out by YOLO, and CLIP executed scene detection through similarity scores derived from text-image features.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

Secondary Data: A literature review of past research and work helped provide context for observations against prevailing theories and knowledge structures.

2. Tools and Equipment

The study employed state-of-the-art AI models, such as YOLO for object detection and CLIP for scene classification. Streamlit was used to develop an interactive data input and visualization interface. Python libraries like Torch, NumPy, PIL, and OpenCV were used for image processing and integrating the models.

3. Data Analysis Methods

Objects and scenes detected were processed with similarity scoring algorithms to determine image content patterns. Confidence levels were computed to rank scenes into categories, for guaranteed classification. Data extracted was also processed with natural language models (e.g., Ollama) for song recommendation generation based on detected features.

4. Bias Mitigation Strategies

To prevent AI model prediction biases, several fallback models (e.g., Llama3, Llava) were utilized to cross-validate the results. Also, varied datasets were utilized while training the models to increase generalizability.

5. Justification of Methodology

The mixed-methods strategy was adopted to utilize the strengths of both qualitative contextual knowledge and quantitative accuracy. Utilizing AI models provides scalability and replicability, while fallback mechanisms provide solutions for possible errors in single-model outputs.

6. Multilingual Text-Image Alignment

Integrated Orion-14B's multilingual capabilities1, trained on 2.5T tokens across 14 languages (English, Chinese, Japanese, Korean, etc.), to analyze text-image pairs. This enabled cross-lingual scene classification via CLIP's similarity scoring, leveraging knowledge transfer from dominant languages observed in Orion-14B's training

7. Scene Analysis

Scene analysis was conducted using CLIP's similarity scoring mechanism, which matched textual descriptions to visual features in the images. This process enabled the identification of overarching themes and contexts within the scenes, such as urban landscapes, natural environments, or cultural settings. The multilingual capabilities of Orion-14B further enriched scene analysis by interpreting scene descriptions in 14 languages, ensuring contextual accuracy across diverse linguistic inputs.

8. Object Analysis

Object analysis was performed using the YOLO (You Only Look Once) model, a real-time object detection system. YOLO divides images into grids and predicts bounding boxes and class probabilities for objects within each grid cell. This approach ensures fast and accurate detection of multiple objects in a single image. The detected objects were further analyzed for their spatial relationships, sizes, and contextual relevance to the scenes. Additionally, the integration of advanced models like Phi-3, Gemma, and Mistral enhanced object recognition accuracy by providing cross-validation and reducing biases in predictions.

IV. MATHEMATICAL LOGIC

1. Propositional Logic

- Deals with propositions (statements that are either true or false).
- Basic Operators:
  - Negation: ¬P (not P)
  - Conjunction: P ∧ Q (P and Q)
  - Disjunction: P ∨ Q (P or Q)
  - Implication: P → Q (if P, then Q)
  - Biconditional: P ↔ Q (P if and only if Q)
- Truth Tables: Define the meaning of the operators.

P	Q	P ∧ Q	P ∨ Q	P → Q	P ↔ Q
True	True	True	True	True	True
True	False	False	True	False	False
False	True	False	True	True	False
False	False	False	False	True	True

Laws:

- De Morgan's Laws:
  - ¬(P ∧ Q) ≡ (¬P) ∨ (¬Q)
  - ¬(P ∨ Q) ≡ (¬P) ∧ (¬Q)
- Distributive Laws:
  - P ∧ (Q ∨ R) ≡ (P ∧ Q) ∨ (P ∧ R)
  - P ∨ (Q ∧ R) ≡ (P ∨ Q) ∧ (P ∨ R)

2. Predicate Logic

- Extends propositional logic to include predicates, variables, and quantifiers.
- Quantifiers:

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

- Universal Quantifier: ∀x P(x) (for all x, P(x) is true)
- Existential Quantifier: ∃x P(x) (there exists an x such that P(x) is true)

- Formulas: Combine predicates, variables, quantifiers, and logical operators.
  - Example: ∀x (Man(x) → Mortal(x)) (All men are mortal)
- Rules:
  - Quantifier Negation:
    ○ ¬(∀x P(x)) ≡ ∃x ¬P(x)
    ○ ¬(∃x P(x)) ≡ ∀x ¬P(x)

3. Set Theory

- A foundation for mathematics, dealing with sets and their properties.
- Basic Notation:
  - ∈: Element of (e.g., x ∈ A means x is an element of set A)
  - ⊆: Subset of (e.g., A ⊆ B means A is a subset of B)
  - ∪: Union (A ∪ B is the set of elements in A or B or both)
  - ∩: Intersection (A ∩ B is the set of elements in both A and B)
  - : Set difference (A \ B is the set of elements in A but not in B)
- Axioms:
  - Axiom of Extensionality: Two sets are equal if they have the same elements.
    - ∀A ∀B (∀x (x ∈ A ↔ x ∈ B) → A = B)
- Operations:
  - Power Set: P(A) is the set of all subsets of A.

4. Model Theory

- Deals with the relationship between formal languages and their interpretations.
- Key Concepts:
  - Structures: Interpretations of formal languages.
  - Satisfaction: A formula is satisfied in a structure if it is true under the interpretation.
  - Logical Consequence: Γ |= φ (φ is a logical consequence of Γ)
- Theorems:
  - Completeness Theorem: A formula is provable if and only if it is logically valid.

5. Proof Theory

- Studies the structure of mathematical proofs.
- Key Concepts:
  - Formal Systems: Axioms and inference rules.
  - Derivations: Sequences of formulas derived from axioms using inference rules.
- Example:
  - Modus Ponens: If P and P → Q, then Q.

V. RESULT

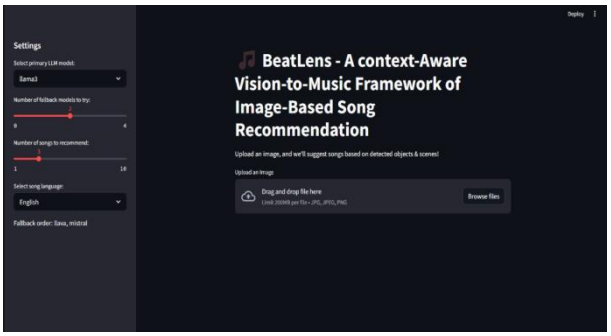

Figure 5.1: Home Page

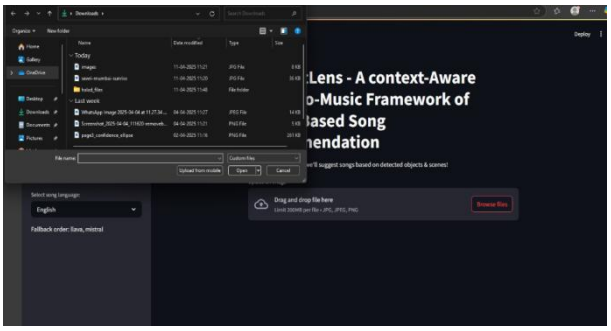

Figure 5.2: Upload Menu

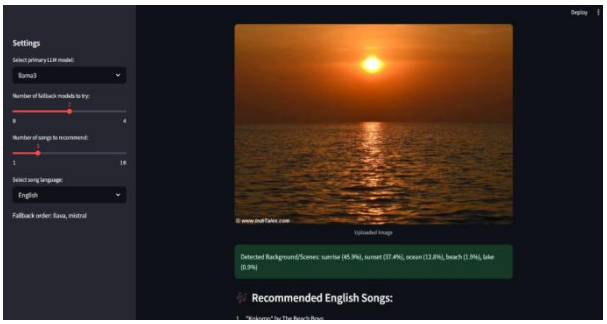

Figure 5.3: Image Scenario Detected

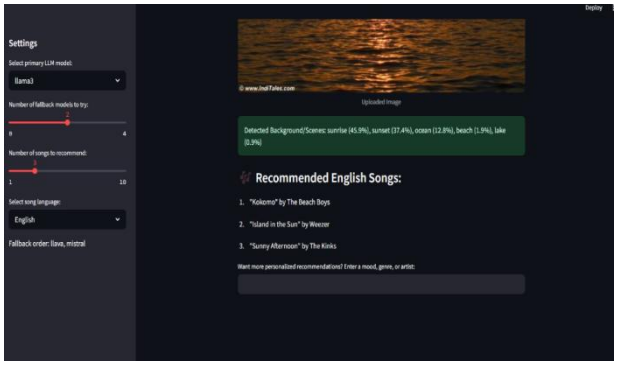

Figure 5.4: Recommended Songs

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

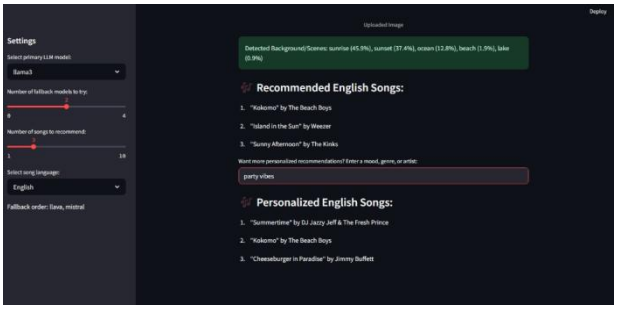

Figure 5.5: Personalized Song Recommendation

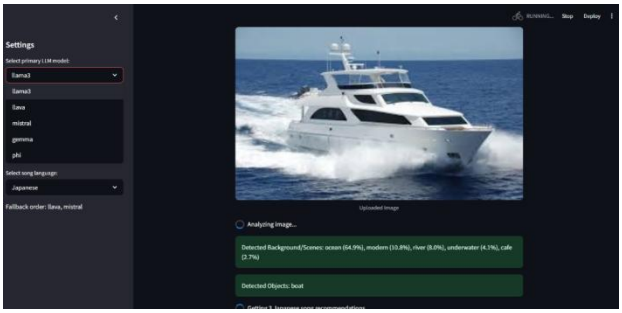

Figure 5.6: Object and Scenario Detection with five model support

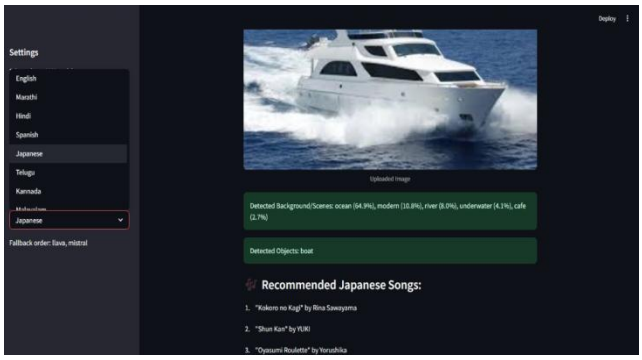

Figure 5.7: Multilanguage Support (14-Languages)

VI. CONCLUSION

In conclusion, this AI system efficiently suggests songs by processing image content through YOLO and CLIP models for object and scene recognition. The system uses several language models to create personalized playlists, providing a novel intersection of visual knowledge and musical preference.

VII. FUTURE SCOPE

BeatLens can upgrade with better models such as YOLOv8, emotional detection, more language support, and real-time recommendations. Smart device integration and music theory can further improve personalization, and it would be a tool with immense power to facilitate smooth music discovery.

1. Improved Object and Scene Recognition: Moving to higher, more powerful models such as YOLO's largest model (eg. YOLOv11x) and future releases of CLIP for higher object and scene recognition accuracy.
2. Utilizing bigger, more varied datasets to train models that can generalize better to different contexts in images.
3. Emotion Recognition and Personalization: Including facial emotion recognition to examine user mood from images, allowing more personalized song suggestions. Including user-specific tastes like favorite artists, genres, or moods to make recommendations even more specific.
4. Real-Time Recommendations: Adding functionality to offer real-time song recommendations during live video streaming or recording, generating dynamic playlists in response to shifting scenes.
5. Integration with Smart Devices: Integrating with smart home appliances (e.g., Alexa, Google Home) to generate ambient music experiences based on environments sensed by smart cameras. Synchronizing with fitness and wellness devices for mood-driven music recommendations while exercising or unwinding.
6. Music Theory Integration: Integration of music theory concepts to harmonize harmonic components of music with the sensed scene or objects to provide an enriched audio experience.
7. Multilingual: Adding more language support to accommodate additional regional and international languages, being inclusive and addressing diverse audiences across the globe.
8. Commercial Applications: Utilizing BeatLens in promotional campaigns to craft music playlists according to brand images or advertisements. Deploying the system in creative arts such as filmmaking or photography for soundtrack recommendations on the basis of visual materials.
9. Scalability and Accessibility: Creating mobile applications and web systems for increased accessibility, enabling direct image upload from devices. Improving computational efficiency for increased speed on low-resource platforms.
10. Collaborative Features: Supporting collaborative playlist building where several users can add images to create shared song suggestions.

These developments could establish BeatLens as a top instrument in AI-powered music discovery and personalization.

ISSN (online): 2581-3048
Volume 9, Issue 4, pp 140-146, April-2025
https://doi.org/10.47001/IRJIET/2025.904021

ACKNOWLEDGEMENT

The authors would like to extend their sincere gratitude to everyone who helped my study project be completed successfully. Before anything else, we want to express our deepest gratitude to the technical team for all of their help and support during the project. They overcame a number of technical obstacles and produced the expected results thanks to their knowledge and commitment. We are also grateful to our professors, research guide, supervisor, and mentor, who provided us with valuable direction and input during the research process. Their sage advice and insightful critiques helped us refine our ideas and raised the bar on our work. Furthermore, we would like to thank our institute and our beloved HOD madam for providing us with the resources and facilities we needed to complete this project. Their assistance and inspiration were crucial in the accomplishment of our research. Last but not least, we would like to express our gratitude to the project team for their cooperation, dedication, and hard work. Their assistance was essential in attaining the project's goals and finishing on schedule. Finally, we would like to express our sincere gratitude to everyone who has supported us on this trip. We could not have accomplished it without their help and contributions, which are priceless.

REFERENCES

[1] YOLO (You Only Look Once): Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788.

[2] CLIP (Contrastive Language-Image Pre-training): Radford, A., Kim, J. W., Xu, C., McLeavey, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv*, abs/2103.00020.

[3] Streamlit: Streamlit documentation. Retrieved from https://streamlit.io/

[4] Ollama: Ollama documentation. Retrieved from https://ollama.com/

[5] Transformers Library: Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *ArXiv*, abs/1909.08053.

[6] PyTorch: PyTorch documentation. Retrieved from https://pytorch.org/

[7] NumPy: Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature* 585, 357–362 (2020).

[8] PIL (Pillow): Pillow documentation. Retrieved from https://pillow.readthedocs.io/en/stable/

[9] OpenCV: OpenCV documentation. Retrieved from https://opencv.org/

[10] Llama3, Llava, Mistral, Gemma, Phi: Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv*:2307.09288.

AUTHORS BIOGRAPHY

Aditya Arolkar, Student, Smt.Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India.

Dhaval Smart, Student, Smt.Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India.

Gaurav Waghmare, Student, Smt.Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India.

Pratham Atale, Student, Smt.Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India.

Prof. Sonali Deshphande, Professor, Smt.Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 4, pp 140-146, April-2025

https://doi.org/10.47001/IRJIET/2025.904021

Citation of this Article:

Aditya Arolkar, Dhaval Smart, Gaurav Waghmare, Pratham Atale, & Prof. Sonali Despande. (2025). BeatLens: A Context-Aware Vision-to-Music Framework for Image-Based Song Recommendations. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(4), 140-146. Article DOI https://doi.org/10.47001/IRJIET/2025.904021

****

