International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Automated Detection and Classification of Oral Precancerous Stages from White Light Images Using LightGBM

Dr.Amita Johar[1], Muddam Shivani[2], C Bramarambika[3], Mourya Archana[4], Gottipati Koteswara Rao[5]

[1] Assistant Professor, Department Of Computer Science And Engineering(AI& ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5]btech Students ,Department Of Computer Science And Engineering(AI& ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*— Cancer is one of the major health challenges across the world and continues to cause a large number of deaths every year. Among different types, oral cancer is quite common and often goes unnoticed until it reaches an advanced stage. This delay in detection is a key reason for its high mortality rate. Identifying the disease at an early or pre-cancerous stage can greatly improve treatment outcomes and patient survival. In this work, a method is proposed to differentiate between non-cancerous and cancerous oral lesions while also identifying their early stages. The approach uses different color spaces to capture variations in images and extracts important color and texture features. These features are then analyzed using the LightGBM algorithm for classification. The results show strong performance across multiple evaluation measures, making the method both effective and efficient for practical use in early oral cancer detection.**

***Keywords*— Oral Cancer Detection, Precancerous Lesions, White Light Imaging, LightGBM, Machine Learning, Image Classification, Feature Extraction, Medical Image Analysis**

I. INTRODUCTION

Cancer is a condition in which abnormal cells grow uncontrollably and spread to different parts of the body [1]. Oral cancer develops in areas such as the gums, tongue, lips, floor of the mouth, and palate, and may also extend to the throat region [2]. Most oral cancers originate from epithelial cells and are commonly associated with precancerous conditions like leukoplakia, erythroplakia, and erythroleukoplakia [3], [4]. These lesions indicate early cellular changes that may eventually develop into Oral Squamous Cell Carcinoma (OSCC), which accounts for the majority of oral cancer cases [5]. Risk factors such as tobacco use and alcohol consumption significantly increase the likelihood of developing oral cancer [7]. Since early-stage lesions are often painless and difficult to identify, diagnosis is frequently delayed, especially in countries like India where many cases are detected at advanced stages [8], [9]. Early detection can significantly improve survival rates and reduce mortality [10].

Traditional diagnostic approaches mainly rely on clinical examination and biopsy, which depend on the experience and expertise of healthcare professionals. These methods can be subjective and may lead to inconsistent results. To overcome these limitations, automated image-based analysis has gained attention in recent years. Color and texture features extracted from medical images play an important role in identifying abnormal tissue patterns. Texture represents repeated patterns within an image, while color provides valuable information about tissue variations [11], [12]. Various methods such as statistical, structural, and learning-based approaches have been used for feature extraction. Combining both color and texture features improves the effectiveness of classification systems and helps in better identification of abnormal regions [13], [14].

With the advancement of technology, machine learning and deep learning techniques have become widely used in medical image analysis. These approaches enable automated systems to learn patterns from large datasets and provide accurate predictions. Several models, including Convolutional Neural Networks (CNNs), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), have been applied to oral cancer detection. While many of these models achieve good performance, they often face challenges such

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

as high computational complexity, limited dataset size, and lack of generalization [16]–[18]. Some methods also require complex image acquisition procedures, making them less practical for real-time clinical use.

To address these challenges, this study proposes an efficient approach using color and texture features combined with the LightGBM classifier. The method focuses on achieving high accuracy while maintaining low computational cost. By enabling early detection and reducing dependence on manual diagnosis, the proposed system can support healthcare professionals in making faster and more reliable decisions.

II. RELATED WORK

[1] S. Sarkar et al. (2013), this study provides an overview of cancer development from an epigenetic perspective, focusing on how gene expression changes without altering DNA sequences. The authors explain that cancer progression is not only driven by genetic mutations but also by epigenetic modifications such as DNA methylation and histone changes. These modifications influence how genes are activated or suppressed, playing a critical role in tumor formation and growth. The paper highlights how environmental factors, lifestyle, and cellular stress can trigger these epigenetic alterations. It also discusses how understanding these mechanisms can help in early diagnosis and targeted therapy. The findings are important because they shift the focus from purely genetic causes to a broader biological context. This research supports the idea that early detection methods, including image-based classification systems, can benefit from understanding underlying cellular changes associated with cancer development.

[2] P. H. Monter and S. G. Patel (2015), this research focuses on cancers affecting the oral cavity and provides detailed insights into their clinical characteristics, diagnosis, and treatment options. The authors describe the anatomical regions where oral cancer commonly occurs and emphasize the importance of identifying early symptoms such as ulcers, patches, or abnormal tissue growth. The study highlights that delayed diagnosis is a major factor contributing to poor survival rates. It also explains the role of clinical examination, imaging, and biopsy in confirming the presence of cancer. Treatment strategies such as surgery, radiation therapy, and chemotherapy are discussed, depending on the stage of the disease. The paper underlines the importance of early screening and awareness among patients and healthcare professionals. This reference is useful for understanding the medical background of oral cancer and reinforces the need for automated systems that assist in early detection and classification of oral lesions.

[3] S. Warnakulasuriya et al. (2007), this paper discusses the classification and terminology of potentially malignant disorders of the oral cavity. The authors aim to standardize the definitions of conditions that may lead to oral cancer, such as leukoplakia and erythroplakia. By providing clear classifications, the study helps clinicians better identify and monitor high-risk lesions. It explains how these disorders may not be cancerous initially but have a significant chance of progressing into malignant conditions if left untreated. The research emphasizes the importance of early diagnosis and regular monitoring of such lesions. It also highlights the challenges in distinguishing between harmless and potentially dangerous lesions through visual examination alone. This work is important because it lays the foundation for developing automated diagnostic tools. Machine learning systems can use such standardized classifications to improve accuracy in detecting and categorizing precancerous stages from medical images.

[11] A. Humeau-Heurtier (2019), this study presents a comprehensive survey of texture feature extraction techniques used in image processing. The author explains different methods for analyzing texture, including statistical, structural, and model-based approaches. Texture features help in identifying patterns and variations within images, which are essential for classification tasks. The paper also discusses widely used techniques such as Gray Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP). It highlights the importance of selecting appropriate texture features depending on the application. In medical imaging, texture analysis plays a crucial role in detecting abnormalities that may not be visible to the human eye. The study provides valuable insights into how texture features can improve classification accuracy. This reference supports the use of texture-based analysis in oral cancer detection systems, where identifying subtle differences in tissue patterns is essential for early diagnosis.

[13] L. Liu et al. (2019), This paper reviews the evolution of texture representation methods over

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

the past two decades, moving from traditional techniques to modern deep learning approaches. The authors compare classical methods such as Bag of Words (BoW) with advanced Convolutional Neural Networks (CNNs). The study highlights how deep learning models automatically learn complex features from images, reducing the need for manual feature extraction. However, it also points out that traditional methods still have advantages in terms of simplicity and lower computational requirements. The paper provides a balanced understanding of both approaches and their applications in image classification. In medical imaging, combining traditional and modern techniques can lead to better performance. This reference is useful for understanding the role of feature extraction in classification tasks and supports the idea of integrating efficient algorithms like LightGBM with meaningful feature representations for improved oral cancer detection.

III. DATASET DETAILS

The dataset used in this work consists of oral cavity images collected for the purpose of detecting cancer and its precancerous stages. Each image represents a sample of either cancerous or non-cancerous tissue captured under white light conditions. The dataset contains approximately 950 images, including around 500 cancer cases and 450 non-cancer cases. These images reflect variations in texture, color, and structural patterns of oral tissues. The dataset is organized in image format, where each sample acts as an input for feature extraction and classification. Instead of tabular data, image-based features such as color components and texture descriptors are derived from each image. These extracted features are then converted into numerical form for further analysis. Proper labeling is maintained to distinguish between different classes, ensuring reliable training and evaluation. This structured dataset supports accurate classification of oral cancer conditions using machine learning techniques.

Before training the models, several preprocessing steps are applied to improve the quality and consistency of the dataset. Initially, all images are loaded and checked to ensure proper formatting and readability. Feature extraction techniques are then applied to derive meaningful color and texture attributes from multiple color spaces. Missing or irrelevant data, if present, is handled carefully to avoid inconsistencies. The extracted features are normalized to bring them into a common scale, which helps improve model performance. After preprocessing, the dataset is divided into training and testing sets, where 80% of the data is used for training and the remaining 20% is used for evaluation. This split ensures that the model is tested on unseen data, providing a fair estimate of its performance. These preprocessing steps enhance the reliability of the dataset and ensure that the models learn meaningful patterns effectively.

The prepared dataset is used to train and evaluate different machine learning and deep learning models, ensuring a consistent and unbiased comparison. Both the LightGBM model and the hybrid MobileNetV2 with SVM model are trained using the same dataset, allowing for fair performance evaluation. The dataset supports feature-based as well as deep feature extraction approaches, making it suitable for multiple classification techniques. Performance metrics such as accuracy, precision, recall, and F1-score are calculated using predictions on the test dataset. The consistency in dataset preparation ensures that improvements in results are due to model efficiency rather than data variations. By providing a balanced and well-processed dataset, the system achieves reliable classification results. Overall, the dataset plays a vital role in enabling accurate detection of oral cancer and its precancerous stages using image-based analysis.

IV. PROPOSED METHODOLOGY

The proposed methodology focuses on developing an efficient and structured framework for the classification of oral cancer and its precancerous stages using white light images. The overall workflow begins with dataset acquisition, followed by preprocessing, feature extraction, model training, evaluation, and prediction. Initially, the oral image dataset is uploaded into the system through a user-friendly interface, allowing smooth interaction with the application. During preprocessing, images are analyzed and converted into suitable formats. Important operations such as normalization and feature scaling are applied to improve model performance. The dataset is then divided into training and testing sets using an 80:20 ratio to ensure proper validation. This step helps in preventing overfitting and ensures that the models perform well on unseen data. The methodology is designed in a step-by-step manner, where each stage is connected logically. This structured pipeline ensures consistent and reliable classification results for oral cancer detection.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

After preprocessing, feature extraction is performed to capture meaningful information from the images. Both color and texture features are extracted using multiple color spaces, which help in identifying variations in oral tissues. These features are then converted into numerical form and used as input for machine learning models. The Light Gradient Boosting Machine (LightGBM) is first applied as a baseline model due to its fast processing and high efficiency. It builds multiple decision trees and combines them to improve classification accuracy. The model is trained using the training dataset and evaluated on the testing dataset using performance metrics such as accuracy, precision, recall, and F1-score. This baseline evaluation helps in understanding the capability of traditional machine learning techniques in handling the classification task. Although LightGBM performs well, it may not fully capture complex patterns present in image data, which motivates the use of advanced models.

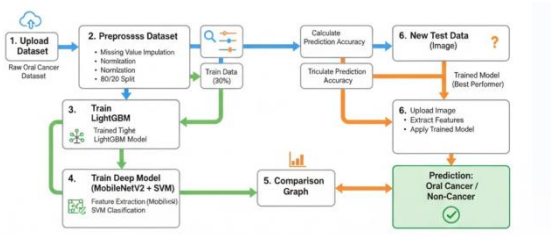

Figure [1]: System Architecture for Oral Cancer Detection

In figure [1] the system architecture represents a structured workflow for the classification of oral cancer using white light images. The process begins with dataset upload, where raw oral cancer images are loaded into the system. The dataset then undergoes preprocessing, which includes handling missing values, normalization, and splitting into training and testing sets. The processed data is used to train two models: LightGBM and a hybrid deep learning model combining MobileNetV2 with SVM. The LightGBM model focuses on feature-based learning, while the hybrid model extracts deep features for improved classification. Both models are evaluated, and their performance is compared using a comparison graph based on accuracy metrics. The best-performing model is selected and used in the prediction phase. In the final stage, new test images are uploaded, features are extracted, and the trained model predicts whether the image represents oral cancer or non-cancer. This architecture ensures an efficient, accurate, and systematic approach for disease classification.

To overcome the limitations of feature-based methods, a deep learning approach is introduced using MobileNetV2 combined with a Support Vector Machine (SVM) classifier. MobileNetV2 is used to extract deep features directly from images, capturing complex patterns and structural information. Additionally, Histogram of Oriented Gradients (HOG) is applied to enhance feature representation by focusing on edge and texture details. The extracted features are further optimized using dimensionality reduction techniques such as Principal Component Analysis (PCA), which reduces computational complexity while preserving important information. The SVM classifier is then used for final classification due to its effectiveness in handling high-dimensional data. This hybrid approach improves the overall accuracy and robustness of the system. By combining deep learning and traditional classification techniques, the methodology achieves better performance compared to individual models.

The final stage of the methodology focuses on model evaluation, comparison, and prediction. Both LightGBM and the hybrid MobileNetV2 + SVM model are compared using evaluation metrics such as accuracy, precision, recall, and F1-score. Visualization techniques, such as comparison graphs, are used to analyze model performance clearly. Based on the results, the best-performing model is selected for final prediction. The system includes a prediction module where users can upload new oral images, and the trained model classifies them as cancerous or non-cancerous. This feature makes the system practical and suitable for real-time applications. The results demonstrate that the hybrid model outperforms the baseline model in terms of accuracy and reliability. Overall, the proposed methodology provides a complete solution for oral cancer classification by integrating preprocessing, feature extraction, model comparison, and prediction into a single framework.

V.RESULT AND DISCUSSION

The experimental results demonstrate the effectiveness of machine learning and deep learning models in classifying oral cancer and its precancerous stages using white light images. Two models were implemented and evaluated, namely LightGBM and a hybrid MobileNetV2 + SVM model. Among these, the hybrid model achieved the highest accuracy of 92%, showing superior

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

performance in capturing complex image patterns and structural variations. The LightGBM model achieved an accuracy of 86%, indicating its capability in handling feature-based classification efficiently. Evaluation metrics such as precision, recall, and F1-score further confirm the reliability of the models, where the hybrid model consistently outperformed LightGBM. The use of dimensionality reduction techniques such as PCA helped improve performance while reducing computational complexity. The results clearly indicate that combining deep learning with traditional classifiers enhances classification accuracy and provides more reliable predictions for oral cancer detection.

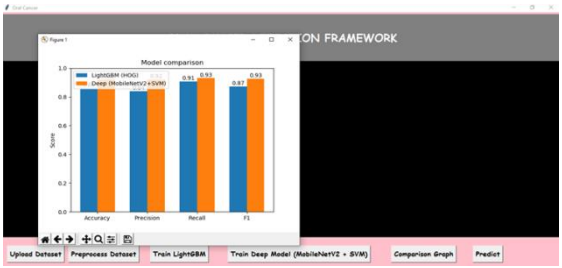

Figure [2]: Performance Comparison of Algorithms

Figure [2] illustrates the comparison of accuracy, precision, recall, and F1-score for both models. The hybrid model clearly outperforms LightGBM in all evaluation metrics, demonstrating its ability to capture complex image features effectively.

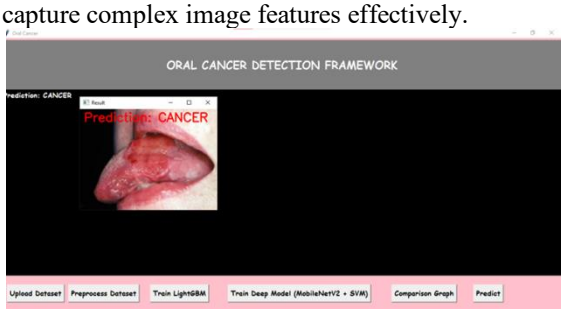

Figure [3]: Oral Cancer Prediction Output

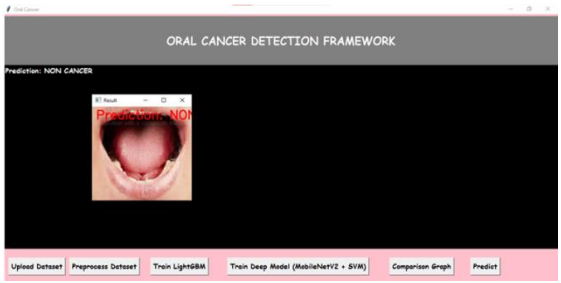

Figure [4]: Oral Cancer Prediction Output

Figure [4] and Figure [5] show the prediction results of the system. When a cancerous image is uploaded, the system correctly classifies it as cancer, while non-cancerous images are identified accurately. The output is displayed clearly through the interface, making it easy for users to interpret results.

DISCUSSION

The results indicate that selecting an appropriate model plays a critical role in achieving accurate classification of oral cancer. The hybrid MobileNetV2 + SVM model outperforms the LightGBM algorithm due to its ability to extract deep features and capture complex patterns from images. LightGBM, although efficient and faster, relies on hand-crafted features, which may limit its ability to identify intricate variations in medical images. The use of multiple color spaces and texture features improves the overall performance of the system. Additionally, preprocessing steps such as normalization and feature extraction contribute significantly to model accuracy and consistency.

The evaluation using accuracy, precision, recall, and F1-score confirms that deep learning-based approaches are more effective for image classification tasks. The system also provides a practical prediction module, allowing users to upload images and obtain results instantly. This enhances usability in real-world scenarios. Overall, the integration of machine learning and deep learning techniques creates a robust and efficient system for early oral cancer detection, with potential for further improvement using larger datasets and advanced hybrid models.

VI. CONCLUSION

In this work, an effective system for the classification of oral cancer and its precancerous stages using white light images has been developed. The approach focuses on extracting meaningful color and texture features and applying machine learning techniques to achieve accurate predictions. The LightGBM algorithm demonstrated reliable performance with efficient computation, making it suitable for practical applications. Additionally, the hybrid model combining MobileNetV2 and SVM further improved classification accuracy by capturing complex image patterns.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The system successfully differentiates between cancerous and non-cancerous conditions, providing consistent results across multiple evaluation metrics. The inclusion of preprocessing steps and feature normalization ensures data quality and enhances model performance. The comparison between models shows that combining traditional and deep learning methods can lead to better outcomes.

This study highlights the importance of automated systems in medical diagnosis, especially for early detection of diseases like oral cancer. By reducing dependence on manual examination, the proposed system can assist healthcare professionals in making faster and more accurate decisions. Overall, the developed model offers a practical, efficient, and scalable solution for oral cancer classification.

REFERENCES

1. S. Sarkar, G. Horn, K. Moulton, A. Oza, S. Byler, S. Kokolus, and M. Longacre, ‘‘Cancer development, progression, and therapy: An epigenetic overview,’’ Int. J. Mol. Sci., vol. 14, no. 10, pp. 21087–21113, Oct. 2013.
2. P. H. Monter and S. G. Patel, ‘‘Cancer of the oral cavity,’’ Surgical Oncol. Clinics North Amer., vol. 24, no. 3, pp. 491–508, Apr. 2015.
3. S. Warnakulasuriya, N. W. Johnson, and I. Van Der Waal, ‘‘Nomenclature and classification of potentially malignant disorders of the oral mucosa,’’ J. Oral Pathol. Med., vol. 36, no. 10, pp. 575–580, Nov. 2007.
4. B. W.Neville and T. A. Day, ‘‘Oral cancer and precancerous lesions,’’ CA, A Cancer J. Clinicians, vol. 52, no. 4, pp. 195–215, Jul. 2002.
5. J. Bagan, G. Sarrion, and Y. Jimenez, ‘‘Oral cancer: Clinical features,’’ Oral Oncol., vol. 46, no. 6, pp. 414–417, Jun. 2010.
6. S.Sudha,A.Veluthattil, S. Kandasamy, and S.Chakkalakkoombil, ‘‘Effect of hypofractionated, palliative radiotherapy on quality of life in late-stage oral cavity cancer: A prospective clinical trial,’’ Indian J. Palliative Care, vol. 25, no. 3, p. 383, 2019.
7. U. Mangalath, S. Aslam, A. H. Abdul Khadar, P. Francis, M. K. Mikacha, andJ. Kalathingal, ‘‘Recent trends in prevention of oral cancer,’’ J. Int. Soc. Preventive Community Dentistry, vol. 4, no. 6, p. 131, 2014.
8. I. van der Waal, R. de Bree, R. Brakenhoff, and J. Coebergh, ‘‘Early diagnosis in primary oral cancer: Is it possible?’’ Medicina Oral Patología Oral y Cirugia Bucal, vol. 16, no. 3, pp. e300–e305, 2011.
9. J. Seoane, B. Takkouche, P. Varela-Centelles, I. Tomás, and J. M. Seoane-Romero, ‘‘Impact of delay in diagnosis on survival to head and neck carcinomas: A systematic review with meta-analysis,’’ Clin. Otolaryngology, vol. 37, no. 2, pp. 99–106, Apr. 2012.
10. M. A. Weinberg and D. J. Estefan, ‘‘Assessing oral malignancies,’’ Amer. Family Physician, vol. 65, no. 7, pp. 84–1379, Apr. 2002.
11. A. Humeau-Heurtier, ‘‘Texture feature extraction methods: A survey,’’ IEEE Access, vol. 7, pp. 8975–9000, 2019.
12. T. Song, J. Feng, S. Wang, and Y. Xie, ‘‘Spatially weighted order binary pattern for color texture classification,’’ Expert Syst. Appl., vol. 147, Jun. 2020, Art. no. 113167.
13. L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, ‘‘From BoW to CNN: Two decades of texture representation for texture classification,’’ Int. J. Comput. Vis., vol. 127, no. 1, pp. 74–109, Jan. 2019.
14. L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, ‘‘Local binary features for texture classification: Taxonomy and experimental study,’’ Pattern Recognit., vol. 62, pp. 135–160, Feb. 2017.
15. X. Qi, G. Zhao, L. Shen, Q. Li, and M. Pietikäinen, ‘‘LOAD: Local orientation adaptive descriptor for texture and material classification,’’ Neurocomputing, vol. 184, pp. 28–35, Apr. 2016.
16. S. Wang, Q. Wu, X. He, J. Yang, and Y. Wang, ‘‘Local N-ary pattern and its extension for texture classification,’’ IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 9, pp. 1495–1506, Sep. 2015.
17. J. Zhang, J. Liang, C. Zhang, and H. Zhao, ‘‘Scale invariant texture rep resentation based on frequency decomposition and gradient orientation,’’ Pattern Recognit. Lett., vol. 51, pp. 57–62, Jan. 2015.