Journal of Informatics, Computer Science, Data Science And Arificial Intelligence
Volume 2 No. 2 June 2024, ISSN : 3031-9145

JICDA
ISSN : 3031-9145
Journal of Informatics, Computer Science, Data Science, and Artificial Intelligence

DENTAL DISEASE CLASSIFICATION USING IMAGE DATA AND MACHINE LEARNING MODELS

Agus Fahmi Limas Ptr[1], Filipus Naibaho[2], Samudra Fadhillah[3]

[1]Deli Sumatera University, agusfahmilimasptr@gmail.com, [2]Deli Sumatera University, pilipusnaibaho@gmail.com, [3]Deli Sumatera University, fadhillahsamudra@gmail.com

ARTICLE INFO

ARTICLE HISTORY:
Received : 04 June 2025
Revised : 13 June 2025
Accepted : 20 June 2025

Keywords:
Dental Disease Classification
Machine Learning
Image Embeddings
Neural Network
Multiclass Classification

ABSTRACT

This study investigates the application of machine learning techniques for the classification of dental diseases based on image data. Two models—Naive Bayes and Neural Network—were evaluated using a publicly available dataset containing 13,839 annotated images across ten dental disease categories. Image embeddings were extracted using the pre-trained Inception v3 model to convert raw images into structured feature vectors. These features were then used to train and evaluate both classifiers using standard performance metrics, including AUC, precision, recall, and F1-score. The results indicate a significant performance gap between the two models. The Neural Network outperformed Naive Bayes across all metrics, achieving an AUC of 0.932 and an F1-score of 0.669, while Naive Bayes performed poorly with near-zero precision and recall. Confusion matrix analysis further highlighted the Neural Network's superior ability to handle multiclass classification, although it still struggled with underrepresented classes such as Caries 2, Caries 3, and Caries 4. These findings suggest that deep learning-based approaches, when combined with robust image embeddings, are more effective for dental disease classification tasks and offer strong potential for supporting automated diagnostic systems in dentistry.

INTRODUCTION

Oral health plays a fundamental role in maintaining overall human well-being, with dental diseases among the most prevalent health issues worldwide. According to the

World Health Organization (WHO), nearly 3.5 billion people are affected by oral conditions such as dental caries, periodontal disease, tooth discoloration, and oral ulcers, all of which can cause pain, discomfort, and reduced quality of life if not detected and treated promptly (Jain et al., 2024). Among these, untreated dental caries alone affect more than 2.5 billion people globally (Qin et al., 2022). Therefore, timely and accurate diagnosis of oral diseases is essential for preventing complications and improving patient outcomes.

In clinical settings, diagnosis typically relies on the expertise of dental professionals interpreting visual and radiographic data. However, this process can be time-consuming, subjective, and dependent on the availability of trained personnel. These challenges have prompted growing interest in automated diagnostic systems based on artificial intelligence (AI), particularly those that leverage machine learning (ML) for image-based disease detection (Schwendicke et al., 2020).

Machine learning models, especially deep learning approaches like convolutional neural networks (CNNs), have demonstrated remarkable performance in various medical imaging tasks, such as skin cancer detection (Sharma et al., 2022), diabetic retinopathy analysis (Wang et al., 2025), and tumor segmentation (Liu et al., 2023). CNNs are capable of learning rich hierarchical features from complex image data, making them highly suitable for classification problems in healthcare (Salehi et al., 2023). Despite these advances, the adoption of AI in dental diagnostics remains limited, particularly in multiclass classification tasks involving visually similar dental conditions (Katsumata, 2023).

This study aims to investigate the performance of two machine learning algorithms—Naive Bayes and Neural Network—in classifying dental disease images into ten diagnostic categories. The classification process utilizes a pre-trained Inception v3 model to generate image embeddings, which serve as feature representations for the classifiers. All experiments are conducted using the Orange Data Mining platform, which facilitates visual programming and reproducible workflows.

The primary objectives of this study are to evaluate and compare the classification performance of Naive Bayes and Neural Network models in detecting various dental diseases from image data, to assess the effectiveness of image embeddings in improving model accuracy, and to identify the limitations and challenges associated with classifying visually similar or underrepresented disease categories. By assessing these aspects, this research contributes to the growing field of AI-based dental diagnostics and offers insights for developing more accurate and efficient image classification systems for oral healthcare.

RESEARCH METHODS

This study employed a machine learning-based image classification approach to detect and categorize dental diseases using visual data. The entire experimental workflow was implemented using the Orange Data Mining software, which provides a block-based visual interface for data processing and model training. The methodological steps are described in the following subsections:

3.1 Data Collection

The dataset used in this study is the "Oral Diseases" image collection, publicly available on Kaggle (Sajid, 2024). It contains approximately 13,839 labeled dental images, each categorized into one of the following six disease classes:

1. Caries (tooth decay)
2. Calculus (tartar build-up)
3. Gingivitis (gum inflammation)
4. Hypodontia (congenital tooth absence)
5. Tooth Discoloration (changes in tooth coloration)
6. Ulcer (oral mucosal sores)

The dataset exhibits considerable variability in terms of image resolution, lighting conditions, and camera angles, making it suitable for developing models with strong generalization capabilities.

3.2 Preprocessing and Feature Extraction

1. Image Resizing

   All input images were resized to a uniform resolution of 299×299 pixels to comply with the input requirements of the Inception v3 model used during image embedding. This step ensures consistency in input shape and reduces computational complexity.

2. Image Embedding with Inception v3

   To extract meaningful features, a pre-trained Inception v3 model (trained on ImageNet) was used to convert each image into a fixed-length numerical feature vector (Pardede et al., 2023). This deep embedding method captures high-level semantic features relevant for classification (Hidayat et al., 2023).

3. Feature Selection and Normalization

   Only significant features from the embedding output were retained through built-in feature selection tools in Orange (Cengel et al., 2024). The feature vectors were further normalized to standardize the input space, enhancing model learning efficiency.

3.3 Classification Algorithms

This study evaluated two popular ensemble learning algorithms for image classification:

1. Random Forest (RF)

   A bagging-based ensemble method that constructs multiple decision trees using randomly sampled subsets of the training data. The final prediction is based on a majority vote across all trees, which helps in reducing overfitting and improving stability (Rabah et al., 2024).

2. Gradient Boosting (GB)

   A boosting technique that builds models sequentially, with each new tree correcting the errors made by its predecessors. This method is highly effective in capturing complex patterns and typically yields superior accuracy, especially in imbalanced or noisy datasets (Joly et al., 2025).

These algorithms were selected due to their robustness in handling high-dimensional, embedded image features, and their strong performance in prior image classification studies.

The dataset was divided into 70% training, 15% validation, and 15% testing subsets. Additionally, k-fold cross-validation was applied (where applicable) within the Orange platform to further evaluate the model’s performance and mitigate overfitting risks (Pardede et al., 2024).

3.4 Classification Workflow

The classification pipeline was implemented using the Orange Data Mining platform in a visual, modular workflow. The key steps include:

1. Input of embedded image features
2. Feature normalization
3. Model selection (Random Forest and Gradient Boosting)
4. Training on labeled data
5. Prediction on unseen data
6. Model validation through cross-validation techniques

Each model was trained to learn the relationships between the embedded features and the disease labels, allowing the system to classify new dental images into the appropriate category.

3.5 Model Evaluation

To comprehensively assess the classification performance of the models, several evaluation metrics were employed. These include precision, recall, F1-score, confusion matrix, and Receiver Operating Characteristic (ROC) curves. Together, these metrics provide a well-rounded view of how well each model performs, especially when dealing with multi-class image classification tasks. Below is a brief explanation of each metric along with the corresponding formula:

1. Precision

   Precision indicates how many of the predicted positive cases were actually correct. It focuses on the quality of positive predictions (Ichsan et al., 2024).

   $$Precision = \frac{TP}{TP+FP} \tag{1}$$

2. Recall (Sensitivity or True Positive Rate)

   Recall shows how many actual positive cases were correctly identified by the model (Firmansyah et al., 2022).

   $$Recall = \frac{TP}{TP+FN} \tag{2}$$

3. F1-Score

   The F1-score is the harmonic mean of precision and recall. It balances both metrics, especially when classes are imbalanced (Pardede et al., 2022).

   $$F1Score = 2\,x\,\frac{Precision\,X\,Recall}{Precision+Recall} \tag{3}$$

Model evaluation was carried out using k-fold cross-validation (via the Orange platform), which helps estimate generalization performance and reduce overfitting. ROC curves were used to visualize the sensitivity and specificity across different disease classes.

During the evaluation process, some metrics—such as precision or recall—were occasionally unavailable for specific classes. This occurred when certain disease categories were underrepresented or absent in the validation folds, resulting in undefined scores for those classes. Despite this, the overall model performance remained measurable through global metrics such as accuracy and confusion matrix, which consistently provided meaningful insights into classification effectiveness.

RESULTS AND DISCUSSION

4.1 Model Performance Summary

This study compared the classification performance of two machine learning models: Naive Bayes and Neural Network, using evaluation metrics including Area Under the ROC Curve (AUC), F1-score, precision, and recall. The performance results are summarized in Table 1.

Table 1. Performance comparison of Naive Bayes and Neural Network models

Model	AUC	F1-Score	Precision	Recall
Naive Bayes	0.094	0.016	0.009	0.094
Neural Network	0.932	0.669	0.672	0.666

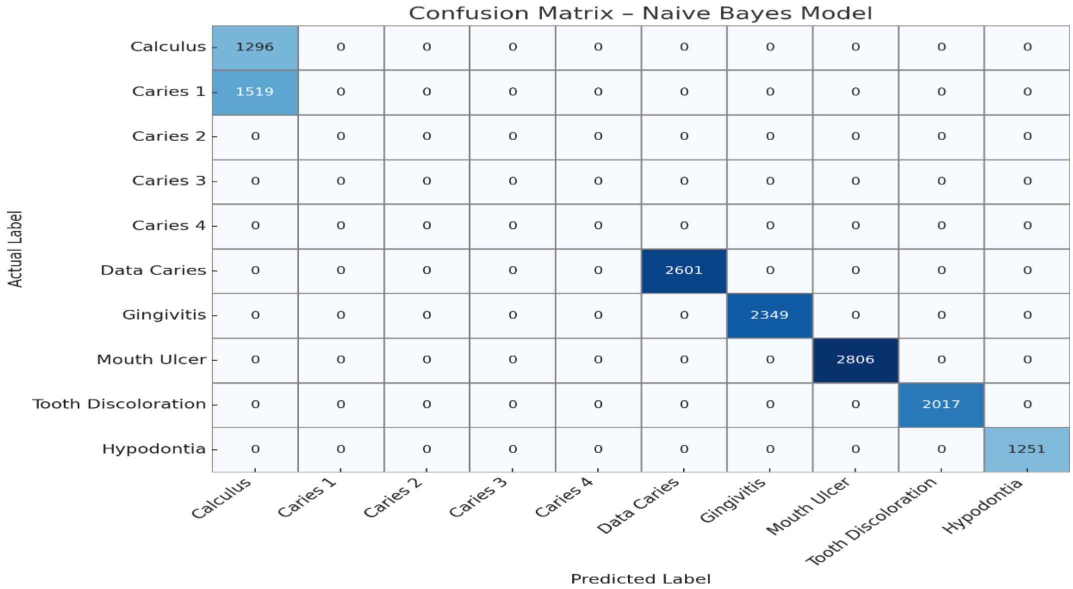
As shown in Table 2, the Neural Network model significantly outperformed Naive Bayes across all evaluation metrics. The Neural Network achieved an AUC of 0.932, indicating excellent discrimination capability between the disease classes. It also demonstrated high and balanced scores in F1-score (0.669), precision (0.672), and recall (0.666), suggesting effective classification performance, particularly in handling class imbalances.

In contrast, the Naive Bayes model yielded poor results, with an AUC of only 0.094 and near-zero values in precision, recall, and F1-score. This indicates that Naive Bayes was unable to generalize effectively on the multiclass classification task involving dental disease images.

These findings highlight the superiority of the Neural Network model in learning complex feature representations extracted from image embeddings, making it more suitable for the automated classification of oral diseases.

4.2 Confusion Matrix Analysis

The performance of the Naive Bayes classifier was evaluated using a multiclass confusion matrix, as shown in Figure 1. The dataset includes ten disease classes, with a total of 13,839 labeled dental images.



	Calculus	Caries 1	Caries 2	Caries 3	Caries 4	Data Caries	Gingivitis	Mouth Ulcer	Tooth Discoloration	Hypodontia
Calculus	1296	0	0	0	0	0	0	0	0	0
Caries 1	1519	0	0	0	0	0	0	0	0	0
Caries 2	0	0	0	0	0	0	0	0	0	0
Caries 3	0	0	0	0	0	0	0	0	0	0
Caries 4	0	0	0	0	0	0	0	0	0	0
Data Caries	0	0	0	0	0	2601	0	0	0	0
Gingivitis	0	0	0	0	0	0	2349	0	0	0
Mouth Ulcer	0	0	0	0	0	0	0	2806	0	0
Tooth Discoloration	0	0	0	0	0	0	0	0	2017	0
Hypodontia	0	0	0	0	0	0	0	0	0	1251

Figure 1. Confusion matrix of the Naive Bayes model

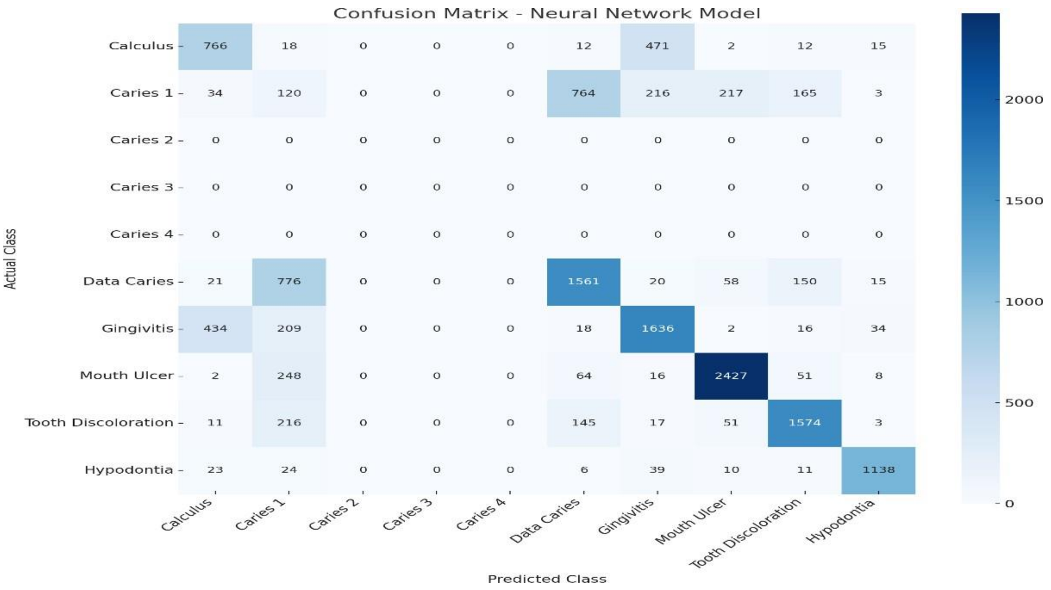
The confusion matrix reveals a major limitation of the Naive Bayes classifier in this context: it failed to make any predictions for several classes (Caries 2, 3, 4), and completely misclassified Caries 1 as Calculus. While it performed well on distinct classes such as Data Caries, Gingivitis, Mouth Ulcer, Tooth Discoloration, and Hypodontia, the inability to correctly distinguish between the various types of caries significantly reduced the model's overall reliability.

This imbalance in prediction indicates that Naive Bayes is not well-suited for handling complex, high-dimensional feature representations extracted from image data in this task. The poor performance is further reflected in its evaluation metrics:

1. AUC: 0.094
2. F1-Score: 0.016
3. Precision: 0.009
4. Recall: 0.094

These results suggest that the model is overly biased toward a few dominant classes and struggles to generalize in a multiclass classification setting involving visually similar categories. Naive Bayes was only able to correctly classify a limited number of classes, such as "Calculus", "Data Caries", "Gingivitis", "Mouth Ulcer", "Tooth Discoloration", and "Hypodontia". It failed to identify any samples from "Caries 2", "Caries 3", or "Caries 4", suggesting its limitations in handling complex and visually similar categories.

Below is the confusion matrix for the Neural Network model, as shown in Figure 2.



	Calculus	Caries 1	Caries 2	Caries 3	Caries 4	Data Caries	Gingivitis	Mouth Ulcer	Tooth Discoloration	Hypodontia
Calculus	766	18	0	0	0	12	471	2	12	15
Caries 1	34	120	0	0	0	764	216	217	165	3
Caries 2	0	0	0	0	0	0	0	0	0	0
Caries 3	0	0	0	0	0	0	0	0	0	0
Caries 4	0	0	0	0	0	0	0	0	0	0
Data Caries	21	776	0	0	0	1561	20	58	150	15
Gingivitis	434	209	0	0	0	18	1636	2	16	34
Mouth Ulcer	2	248	0	0	0	64	16	2427	51	8
Tooth Discoloration	11	216	0	0	0	145	17	51	1574	3
Hypodontia	23	24	0	0	0	6	39	10	11	1138

Figure 2. Confusion matrix of the Neural Network model

While the Neural Network model performed far better, the confusion matrix reveals several misclassifications, particularly between visually similar classes such as "Caries 1" and "Data Caries", as well as some confusion between "Tooth Discoloration" and "Gingivitis".

Despite some overlapping classifications, the Neural Network model demonstrates strong generalization and robustness, making it a promising tool for automated dental disease detection. However, it still faces challenges with underrepresented or visually ambiguous categories such as "Caries 2", "Caries 3", and "Caries 4", which were not predicted at all—indicating the need for better-balanced datasets or fine-tuned model training.

4.3 Discussion

The comparative analysis of Naive Bayes and Neural Network models highlights significant differences in classification performance, particularly in handling the complexity of multiclass dental disease image data. The results clearly demonstrate that the Neural Network model substantially outperformed Naive Bayes in all key performance metrics, including AUC, F1-score, precision, and recall (as summarized in Table 1). This suggests that deep learning-based approaches are more adept at learning abstract and high-dimensional representations extracted through image embeddings, such as those generated by the Inception v3 model used in this study.

The Naive Bayes model, which relies on probabilistic assumptions and feature independence, showed substantial limitations. Its confusion matrix (Figure 1) indicates a severe inability to classify many disease classes—most notably Caries 2, 3, and 4, for which it made no correct predictions. Additionally, it misclassified all instances of Caries 1 as Calculus, revealing the model's inability to discriminate between visually similar categories. This outcome suggests that Naive Bayes is not suitable for high-dimensional image data, where pixel and feature correlations are critical for accurate classification.

In contrast, the Neural Network model demonstrated much stronger generalization, correctly classifying most instances across major disease categories (Figure 2). However, the confusion matrix still reveals certain weaknesses, particularly in distinguishing between overlapping or closely related conditions such as Caries 1 vs. Data Caries, and Gingivitis vs. Tooth Discoloration. These misclassifications can likely be attributed to visual similarities between certain classes, insufficient inter-class variability, or dataset imbalance—issues that are common in medical image classification tasks.

Furthermore, although the Neural Network showed excellent performance in high-sample classes like Mouth Ulcer and Gingivitis, it completely failed to detect underrepresented classes such as Caries 2, Caries 3, and Caries 4. This suggests the presence of class imbalance in the dataset, which could bias the model toward dominant categories. Addressing this limitation may require strategies such as data augmentation, class reweighting, or collecting more balanced data samples for underrepresented classes.

In summary, the Neural Network model exhibits clear advantages in this task due to its capacity to learn complex patterns from image embeddings. Nevertheless, its performance still depends heavily on the quality and distribution of training data. Future work should explore more balanced datasets, advanced augmentation techniques, or class-aware training objectives to further improve classification across all dental disease categories.

CONCLUSION

This study explored the application of machine learning models—Naive Bayes and Neural Network—for the classification of dental disease images using a publicly available dataset of oral conditions. Image embeddings generated via the Inception v3 architecture were used to extract meaningful features from raw image data, followed by classification using the two selected algorithms within the Orange Data Mining environment. The results revealed a significant performance gap between the models. The Neural Network demonstrated strong classification capabilities, achieving high AUC (0.932), F1-score (0.669), precision (0.672), and recall (0.666). It was able to generalize well across multiple disease classes and effectively distinguish between several visually similar categories. In contrast, the Naive Bayes model showed extremely poor performance across all evaluation metrics, failing to identify multiple classes and yielding near-zero precision and recall. The confusion matrix analysis further confirmed that while the Neural Network model achieved broader class coverage and better generalization, it still struggled with underrepresented and visually ambiguous categories such as Caries 2, 3, and 4. These challenges highlight the importance of balanced datasets and class-aware training strategies in improving performance for multiclass medical image classification tasks. In conclusion, the Neural Network model proved to be a more suitable and robust approach for the automated classification of dental disease images, offering a promising foundation for the development of AI-assisted diagnostic tools in dental care. Future work should focus on enhancing class balance, incorporating more diverse image samples, and refining model architectures to improve recognition of minority classes.

REFERENCES

Cengel, T. A., GENCTURK, B., Yasin, E., YILDIZ, M. B., CINAR, I., Özbek, O., & KOKLU, M. (2024). Classification of Orange Features for Quality Assessment Using Machine Learning Methods. Selcuk Journal of Agricultural and Food Sciences, December. https://doi.org/10.15316/sjafs.2024.036

Firmansyah, I., Samudra, J. T., Pardede, D., & Situmorang, Z. (2022). Comparison Of Random Forest And Logistic Regression In The Classification Of Covid-19 Sufferers Based On Symptoms. JOURNAL OF SCIENCE AND SOCIAL RESEARCH, 5(3), 595. https://doi.org/10.54314/jssr.v5i3.994

Hidayat, T., Astuti, I. A., Yaqin, A., Tjilen, A. P., & Arifianto, T. (2023). Grouping of Image Patterns Using Inceptionv3 for Face Shape Classification. International Journal on Informatics Visualization, 7(4), 2336–2343. https://doi.org/10.30630/joiv.7.4.1743

Ichsan, A., Riyadi, S., & Pardede, D. (2024). Analysis of Logistic Regression Regularization in Wild Elephant Classification with VGG-16 Feature Extraction. Journal of Computer Networks, Architecture and High Performance Computing, 6(2), 783–793. https://doi.org/10.47709/cnahpc.v6i2.3789

Jain, N., Dutt, U., Radenkov, I., & Jain, S. (2024). WHO's global oral health status report 2022: Actions, discussion and implementation. Oral Diseases, 30(2), 73–79. https://doi.org/10.1111/odi.14516

Joly, N. A., Munchi, R., Mohalder, R. D., Reya, F. M., Islam, M. S., & Chowdhury, S. (2025). A Majority Hard Voting Based Ensemble Approach to Predict

Cardiovascular Disease. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025, June. https://doi.org/10.1109/ECCE64574.2025.11013017

Katsumata, A. (2023). Deep learning and artificial intelligence in dental diagnostic imaging. Japanese Dental Science Review, 59(July), 329–333. https://doi.org/10.1016/j.jdsr.2023.09.004

Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y., & Zhou, H. (2023). Deep learning based brain tumor segmentation: a survey. Complex & Intelligent Systems, 9(1), 1001–1026. https://doi.org/10.1007/s40747-022-00815-5

Pardede, D., Firmansyah, I., Handayani, M., Riandini, M., & Rosnelly, R. (2022). Comparison Of Multilayer Perceptron's Activation And Optimization Functions In Classification Of Covid-19 Patients. JURTEKSI (Jurnal Teknologi Dan Sistem Informasi), 8(3), 271–278. https://doi.org/10.33330/jurteksi.v8i3.1482

Pardede, D., Ichsan, A., & Riyadi, S. (2024). Enhancing Multi-Layer Perceptron Performance with K-Means Clustering. Journal of Computer Networks, Architecture and High Performance Computing, 6(1), 461–466. https://doi.org/10.47709/cnahpc.v6i1.3600

Pardede, D., Wanayumini, & Rosnelly, R. (2023). A Combination Of Support Vector Machine And Inception-V3 In Face-Based Gender Classification. International Conference on Information Science and Technology Innovation (ICoSTEC), 2(1), 34–39. https://doi.org/10.35842/icostec.v2i1.30

Qin, X. F., Zi, H., & Zeng, X. J. (2022). Changes in the global burden of untreated dental caries from 1990 to 2019: A systematic analysis for the Global Burden of Disease study. Heliyon, 8(9), e10714. https://doi.org/10.1016/j.heliyon.2022.e10714

Rabah, M. A. O., Drid, H., Medjadba, Y., & Rahouti, M. (2024). Detection and Mitigation of Distributed Denial of Service Attacks Using Ensemble Learning and Honeypots in a Novel SDN-UAV Network Architecture. IEEE Access, 12(July), 128929–128940. https://doi.org/10.1109/ACCESS.2024.3443142

Sajid, S. (2024). Oral Diseases. https://www.kaggle.com/datasets/salmansajid05/oral-diseases

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A Study of CNN and Transfer Learning in

Medical Imaging: Advantages, Challenges, Future Scope. Sustainability, 15(7), 5930. https://doi.org/10.3390/su15075930

Schwendicke, F., Samek, W., & Krois, J. (2020). Artificial Intelligence in Dentistry: Chances and Challenges. Journal of Dental Research, 99(7), 769–774. https://doi.org/10.1177/0022034520915714

Sharma, A. K., Tiwari, S., Aggarwal, G., Goenka, N., Kumar, A., Chakrabarti, P., Chakrabarti, T., Gono, R., Leonowicz, Z., & Jasinski, M. (2022). Dermatologist-Level Classification of Skin Cancer Using Cascaded Ensembling of Convolutional Neural Network and Handcrafted Features Based Deep Neural Network. IEEE Access, 10, 17920–17932. https://doi.org/10.1109/ACCESS.2022.3149824

Wang, H., Li, L., Wang, W., Li, Z., Jian, T., Yang, X., Song, B., Li, S., Xu, F., Liu, S., & Li, Y. (2025). Development and validation of a deep learning image quality feedback system for infant fundus photography. Scientific Reports, 15(1), 1–12. https://doi.org/10.1038/s41598-025-10859-5