Arsenic Skin Detection using Machine Learning

Amit Chandra Das (2014242042), Tanushree Das (2212225042), Ashraful Islam (2212669042)

Department of Computer Science and Engineering, North South University, Dhaka, Bangladesh

Abstract— Arsenic contamination is a major health problem in Bangladesh, particularly in rural areas where groundwater is the main source of drinking water. Long-term exposure to arsenic leads to visible skin changes such as pigmentation, dark spots, and keratosis. This project aims to develop an image-based machine learning model that can automatically detect early signs of arsenic poisoning from skin images. Using a dataset of labeled skin images, the system will extract visual features and classify whether a person is affected by arsenic poisoning. This solution could help rural healthcare workers quickly identify patients needing further diagnosis and treatment.

Index Terms—Arsenic detection, machine learning, computer vision, CNN, Bangladesh.

I. Introduction

Chronic exposure to arsenic through contaminated groundwater is a widespread environmental health problem affecting millions of people globally, with Bangladesh being one of the most severely impacted nations. The initial and most common clinical manifestations of this long-term arsenic ingestion, known as arsenicosis, are dermatological. These visible indicators include distinct abnormalities such as melanosis (hyperpigmentation), dark spots, and keratosis, particularly on the palms and soles.

While these signs are clear markers for early diagnosis, their reliable identification requires trained medical professionals, specifically dermatologists, who are often unavailable in the remote, rural communities where arsenic contamination is most prevalent. This diagnostic gap leads to delayed treatment, allowing the progression of arsenicosis to more severe conditions, including debilitating skin cancers and internal organ damage.

Recent advances in machine learning (ML) and computer vision, however, have created new opportunities for automated medical image analysis. These technologies, especially deep learning models like Convolutional Neural Networks (CNNs), have demonstrated high efficacy in detecting patterns and features from dermatological images. This project aims to harness these technologies to develop a low-cost, scalable, and automated screening tool. By analyzing digital images of skin, the system can provide a preliminary diagnosis, flagging individuals who require further medical assessment and helping to bridge the healthcare gap in resource-limited settings.

II. Problem Statement

The core problem is the lack of accessible, timely, and affordable diagnostic solutions for arsenic-induced skin lesions in the rural and underserved regions of Bangladesh. The manual diagnosis of arsenicosis depends on visual inspection by trained dermatologists, who are critically scarce in these areas. This severe shortage of specialized medical expertise creates a significant barrier to early detection.

Consequently, many individuals remain undiagnosed until the condition has progressed to more severe, often irreversible stages. This diagnostic delay allows for the development of debilitating health outcomes, including an increased risk of skin cancer, internal cancers, and other systemic organ damage. The absence of a scalable screening method means that the true prevalence of early-stage arsenicosis is likely underestimated, preventing proactive public health interventions.

This project directly addresses this critical healthcare gap by proposing the development of an automated, image-based system that uses machine learning to accurately identify the dermatological signs of arsenic poisoning. The proposed solution aims to serve as a reliable and cost-effective screening tool for community health workers and local clinicians, enabling widespread and early detection to mitigate the long-term consequences of chronic arsenic exposure.

III. Objectives

The primary objective of this project is to develop and evaluate an automated, image-based system for the early detection of arsenic-induced skin lesions by comparing the performance of multiple machine learning models. The specific objectives to achieve this goal are as follows:

1. **To Develop a Curated Image Dataset:** To collect, organize, and meticulously label a comprehensive dataset of digital skin images, comprising both arsenic-affected (e.g., melanosis, keratosis) and healthy skin samples, which will serve as the foundation for training and validating the machine learning models.
2. **To Implement Image Preprocessing and Feature Extraction Protocols:** To design and apply a standardized pipeline for image preprocessing—including resizing, color normalization, and noise reduction—and to extract discriminative features required for traditional machine learning models.
3. **To Train and Validate Multiple Machine Learning Classifiers:** To build, train, and rigorously evaluate a suite of machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and Convolutional Neural Networks (CNN), to determine their respective efficacies in this diagnostic task.
4. **To Conduct a Comparative Performance Analysis:** To systematically assess and compare the performance of each trained model using a held-out test dataset. The evaluation will be based on standard metrics, including accuracy, precision, recall, and F1-score, to identify the most effective and reliable model for this specific classification problem.

IV. Literature Review

The application of machine learning and computer vision in dermatology is a well-established and rapidly advancing field of research. Numerous studies have successfully demonstrated the potential of these technologies to classify a wide range of skin conditions with high accuracy. Convolutional Neural Networks (CNNs), for instance, have achieved dermatologist-level performance in identifying skin cancers, such as melanoma, from digital and thermoscopic images. This highlights the power of deep learning models to learn intricate patterns directly from pixel data.

Alongside deep learning, traditional machine learning models have also proven effective, particularly when combined with robust feature extraction techniques. Research has shown that algorithms like Support Vector Machines (SVM) and Random Forest can successfully classify conditions such as eczema and psoriasis by analyzing features related to texture, color, and lesion morphology. A 2022 study, for example, underscored the efficacy of texture-based features for dermatological classification, a finding that is highly relevant to this project, as the manifestations of chronic arenicolids—melanosis and keratosis—are primarily abnormalities of skin pigmentation and texture.

Despite the broad success of machine learning in dermatology, a significant research gap exists in the automated detection of arsenic-induced skin lesions. While the foundational methodologies are well-developed for other dermatological diseases, they have not been extensively applied to this specific, high-impact public health problem. This project aims to bridge that gap by

adapting and applying these proven classification techniques to build a reliable screening tool for early stage arenicolids, thereby addressing a critical need in affected communities.

V. Proposed Methodology

The methodology for this project is designed as a systematic, multi-phase process to develop and evaluate an automated system for arsenic-induced skin lesion detection. The approach emphasizes a comparative analysis of different machine learning models to identify the most effective classifier for this specific task.

A. Data Collection and Preprocessing

The foundation of this project is a high-quality, labeled dataset. The initial phase will involve sourcing images from publicly available dermatological atlases and medical repositories. If feasible, collaboration with healthcare institutions will be sought to acquire anonymized clinical images to enhance dataset diversity. Once collected, all images will undergo a rigorous preprocessing pipeline to ensure uniformity and quality. This includes:

- **Resizing:** Standardizing all images to a consistent dimension to ensure uniform input for the models.
- **Normalization:** Scaling pixel values to a standard range (e.g., 0 to 1) to aid in model convergence.
- **Image Enhancement:** Applying filters to reduce noise and enhance contrast, which helps in making lesion features more prominent.

B. Feature Extraction

For the traditional machine learning models (SVM and Random Forest), a comprehensive set of discriminative features will be extracted from the preprocessed images. This step is crucial for enabling these models to interpret visual data. The feature set will include:

- **Color Features:** Analysis of color distribution using histograms and mean color values in various color spaces (e.g., RGB, HSV).
- **Texture Features:** Quantification of skin texture using statistical methods derived from the Gray-Level Co-occurrence Matrix (GLCM), such as contrast, correlation, energy, and homogeneity.

C. Model Training and Validation

This core phase focuses on training and validating three distinct machine learning classifiers to facilitate comparative study:

1. **Support Vector Machine (SVM):** A powerful algorithm effective in high-dimensional feature spaces, making it well-suited for complex classification tasks.
2. **Random Forest:** An ensemble learning method known for its robustness against overfitting and its ability to handle many features.
3. **Convolutional Neural Network (CNN):** A deep learning architecture that can automatically learn hierarchical features directly from raw image data, potentially offering superior performance if the dataset is sufficiently large.

The dataset will be split into training (80%) and testing (20%) sets. Techniques such as k-fold cross-validation will be employed during training to ensure the models' robustness and prevent bias.

D. Comparative Performance Evaluation

The final phase involves a rigorous and systematic evaluation of the trained models to compare their effectiveness. The performance of each classifier on the held-out test dataset will be measured using a standard set of evaluation metrics:

- **Confusion Matrix:** To visualize classification results, including true

positives, false positives, true negatives, and false negatives.

- **Accuracy:** To measure the overall proportion of correct classifications.
- **Precision and Recall:** To assess the model's ability to correctly identify positive cases (arsenic-affected skin) while minimizing false alarms.
- **F1-Score:** The harmonic mean of precision and recall, providing a single score that balances both metrics.

The results will be analyzed to determine which model provides the most reliable and accurate performance for detecting arsenic-induced skin lesions.

VI. Expected Outcomes and Deliverables

Upon successful completion, this project will yield the following key outcomes and deliverables:

- **A Suite of Trained Machine Learning Models:** A set of fully trained and evaluated machine learning models (SVM, Random Forest, and CNN). The project will deliver a detailed comparative analysis of their performance, identifying the most effective model for classifying arsenic-induced skin lesions, with a target accuracy of over 85% for the best-performing model.
- **A Curated and Labeled Image Dataset:** A well-documented and organized dataset of skin images, containing both arsenic-affected and healthy samples. This dataset will be made available to the research community to facilitate future studies on arenicolids.
- **A Comprehensive Project Report:** A final report documenting the entire project lifecycle, including the data collection process, preprocessing steps, feature extraction methods, model architectures, and a thorough analysis of the comparative results. The report will conclude with insights into the most promising computational approaches for this diagnostic challenge.
- **A Public Code Repository:** A GitHub repository containing all source code for image preprocessing, feature extraction, model training, and evaluation, along with clear documentation to ensure reproducibility of the results.

VII. Conclusion

This project directly addresses a critical public health crisis in Bangladesh by proposing an innovative, technology-driven solution for the early detection of chronic arenicolids. By harnessing the power of machine learning to analyze skin images, this research aims to develop a cost-effective, accurate, and accessible screening tool for use in rural and underserved communities. The successful development of this system has the potential to empower community healthcare workers, facilitate timely diagnosis and intervention, and ultimately reduce the severe long-term health consequences of arsenic poisoning. Furthermore, this work contributes to the vital and growing field of medical AI, demonstrating its practical application in tackling neglected environmental health challenges in developing nations.

VIII. References

[1] S. Paul, et al., “Skin Disease Classification Using Convolutional Neural Networks,” *IEEE Access*, vol. 8, pp. 157675–157687, 2020.

[2] M. Rahman, et al., “Texture-Based Machine Learning Model for Dermatological Disease Detection,” in *Advances in Intelligent Systems and Computing*, vol. 1406, Springer, 2022, pp. 115–125.

[3] World Health Organization, “Arsenic,” Fact Sheet, May 2018. [Online]. Available:

https://www.who.int/news-room/fact-sheets/detail/arsenic