International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 1 (2026)
https://doi.org/10.31436/ijpcc.v12i1.642

A Conceptual Framework for a Lightweight AI System for Skin Disease Risk Prediction Using Epidemiological Data in Rural Bangladesh

Mohammad Raihanul Islam[1], Andi Fitriah binti Abdul Kadir[1], Syazwan Aizat Ismail[2]
[1]*Department of Computer Science, International Islamic University Malaysia,* Kuala Lumpur, 53100, Malaysia
[2]*National Poison Centre, University Sains Malaysia, Penang, Malaysia.*

*Corresponding author: raihanulmcse@gmail.com
(Received: 1st December 2025; Accepted: 9th January, 2026; Published on-line: 30th January, 2026)

Abstract— Skin disease remains a significant public health issue in rural Bangladesh, where limited access to dermatologists and inadequate diagnostic facilities often delay accurate assessment and treatment. To address these constraints, this conceptual paper presents a lightweight AI-based framework for predicting skin disease risks using structured epidemiological data gathered from hospital visits and interviews with patients and healthcare staff. The framework incorporates environmental, occupational, hygiene-related, and living-condition factors to model individual risk profiles. Preliminary experiments conducted on an existing dataset demonstrate that conventional machine learning algorithms, particularly K-Nearest Neighbors (KNN) and Random Forest, achieve strong predictive performance, with accuracy reaching up to 88% in train–test evaluations and 80% in 10-fold cross-validation. These results confirm the viability of achieving high diagnostic reliability without image-based tools, relying solely on patient and environmental attributes. The findings further support the practical feasibility of deploying the proposed model in resource-limited rural clinics to aid early risk identification and more efficient allocation of healthcare resources. Privacy protection is incorporated as a core component to ensure secure and ethical handling of patient information.

Keywords— Skin disease risk prediction, epidemiology, lightweight AI, rural healthcare, machine learning.

I. Introduction

Skin diseases are a major health concern in Bangladesh, especially in rural communities. Based on observations from local hospitals, clinics, health centers, and the patient's scenarios, these conditions are very common. People living in crowded areas, dirty environments, with poor water sources, low income, and poor sanitation are more affected compared to those living in developed areas. Similarly, frequent chemical exposure, irregular bathing and laundry practices, poor household environments, shared clothing, limited use of soap, flood-prone areas, and household pets also contribute to the increase in skin problems. In some areas near rivers and ponds, where there is no tube well or clean water facilities,-people are fully dependent on these water sources and often use them directly for bathing and washing. They also share clothing and bedding and face difficulties in accessing medical care when symptoms appear. In many communities, skin problems become not only a health issue but also an essential part of daily life. According to reports from community health camps, skin issues represent about 15–20% of all patient visits. Furthermore, it is estimated that more than 60% of people experience some form of skin problem during their lifetime [1], [2]. Skin conditions like scabies, tinea, vitiligo, urticaria, acne, ringworm, and impetigo, eczema are very common, particularly among children and senior citizens. Chronic conditions such as eczema, urticaria, tinea, and psoriasis frequently persist in adulthood, leading to long-term health and social challenges [3], [4].

A large number of families live in underserved areas in Bangladesh, typically in small and crowded houses. It is not unusual to find more than six people sleeping under one roof. When people live close proximity, skin problems can spread easily from one person to another. Children often share the same bed and wear each other's clothes, which also contribute to the transmission of disease. In addition, many people bathe in the same pond or canal. For this reason, contagious diseases such as scabies re-occur every year [3]. The key—factors are closely linked to the environment and unhygienic lifestyle of farming communities [5].

In addition, frequent floods and high humidity also create perfect conditions to create germs. Combined with poor drainage systems, stagnant water, and animal waste, diseases spread even more rapidly [6]. A persistent lack of

safe water and sanitation makes these problems more critical. Few households have access to safe piped water or toilets, and among poor families and those living in flood-prone areas, access is minimal [1] Knowledge about hygiene within community is also very limited, particularly among women, who are primarily responsible for household cleaning and related tasks. This gap, partially responsible for educational and cultural challenges, contributes to the continued spread of diseases.

The healthcare system in Bangladesh is gradually improving, but underserved areas such as villages, continue to face significant challenges. Usually, a single dermatology specialist is responsible for many patients. A report has shown that the majority of people living outside big cities do not have access to skin specialists, or proper patient records in hospital [7]. Villagers often seek hospital care only when conditions become severe. Approximately one-third of skin problems are treated without expert practitioners, relying instead on village pharmacies or traditional methods. However, misinformation, violations of medical guidelines and self-medication increase the risk of both contiguous and incontiguous diseases.

Advancements in healthcare, particularly those driven by artificial intelligence (AI), have shown significant progress in addressing gaps within the healthcare sectors. As a result, false information, weak health regulations, and self-medication increase the risk of long-term illness and drug resistance. New digital health tools, especially those leveraging AI, have the potential to mitigate these issues and improve access to appropriate care.

Within the domain of dermatology, the utilization of deep learning algorithms and image-based models has significantly improved the accuracy of disease detection, classification, and prioritization in settings with ample resources [8], [9]. Updated technologies are developed using large sets of high-quality skin images. Importantly, they can perform as well as, or even better than, experienced dermatologists. However, they rely on expensive technology, steady electricity, skilled personnel, and reliable internet access. Consequently, people in poor and undeserved areas of Bangladesh often cannot access these advanced solutions.

The most effective approach, therefore, is to leverage readily accessible epidemiological data variables such as age, gender, occupation, household structure, hygiene practices, water availability, and population density, and scrutinize these using interpretable, artificial intelligence frameworks. Supervised machine learning methodologies can yield actionable risk assessments based on structured data variables. This approach aligns with the World Health Organization's concept of social determinants of health which emphasizes that communities' health is shaped by their personal income, household environment, and social conditions. The primary goal of this framework is to make healthcare more accessible, user-friendly, build trust, and strengthen local skills enabling artificial intelligence to better support global health [11].

Although the sector continues to undergo reforms, these evident constraints in the real world have motivated the present research to design a model that is closely aligned with the community's actual conditions and capable of practical grassroots implementation.

II. Literature Review

Several epidemiology studies have documented that skin disease is a burden in Bangladesh. A clinic-based study by [2] reported that about 58% of patients with skin disease have fungal infections. In comparison, scabies and contagious diseases occur in more than 20% of cases, together with bacterial and viral infections. During the rainy season, disease and infections increase due to poor-quality water. The highest-risk populations are children and seniors [5].

Institutional field studies have reported that the prevalence of scabies is between 18% and 34% among students. This is mostly due to living in overcrowded houses and poor clothing hygiene [6]. Transmission is more rapid among family members, suggesting that social and environmental variables are important for skin issues [3].

Common chronic skin diseases like eczema, urticaria, and psoriasis cause both physical and psychological morbidity. These types of problems are connected with mental stress, depression, and reduced professional or educational performance [4]. In Bangladesh, chronic skin disease has been reported to affect approximately 10% to 20% of people. Severity worsens due to environmental and individual factors such as climate change, poor diet, family stress, and delays in obtaining a proper medical treatment [10].

Both infectious and non-infectious skin diseases can be identified within the Social Determinants of Health (SDH) model. Social and structural factors such as education, occupation, income, living environment, and access to healthcare play an important role in determining whether people are healthy or sick [12]. Other studies have shown that water facilities, hygiene, household gatherings, and contact with animals are strong predictors of skin disease, even after adjustments for age and gender [1].

In the past five years, many scholars have used artificial intelligence (AI) to predict disease risks with community and survey data. This approach is very effective in regions such as Africa and Asia, where it is still limited to image datasets and digital health records [13], [14]. For example, one study of [15] proposed a supervised classification method using health-related and climatic factors to predict skin problems utilizing KNN, SVM, and Random Forest. Their findings show

https://doi.org/10.31436/ijpcc.v12i1.642

that epidemiology-based prediction is feasible without using clinical image data. Most of those models used tabular data rather than clinical skin images. They are designed to be lightweight, interpretable, and suitable for use in public health programs.

AI-based analysis of community and survey data still necessitates consideration of fundamental ethical principles, such as informed permission, confidentiality, and appropriate data governance in public health programs, even though this work does not specifically address privacy-preserving strategies [15].

Several image-based deep learning systems have classified skin lesions with extremely high accuracy. For instance, [9] trained deep CNNs to achieve dermatologist-level performance on dermoscopic pictures, whereas [15], [8] employed hybrid models, ResNet, and DenseNet to get AUC values near 0.99. These methods rely on high-quality clinical photos and computing resources, which are challenging to implement in clinics located in rural Bangladesh.

Overall, the existing literature demonstrates that skin disease is prevalent in Bangladesh among children, students, and senior citizens and is heavily influenced by social and environmental variables. However, most AI-based literature on skin diseases either relies on relatively small, survey-based models or clinical skin images and well-recorded datasets, which are hard to gather in rural settings. Even though it is clear that factors such as water source, hygiene practices, household crowding, education, income, and animal contact are responsible for spreading the disease, there is still a lack of a lightweight, simple prediction model that uses the epidemiological data to identify the skin risk for rural communities. This study addressed the gap by creating and evaluating an epidemiological data model for skin disease risk prediction specifically adapted to the rural Bangladeshi context and appropriate for incorporation into community health initiatives.

TABLE I
SUMMARY OF RELATED STUDIES ON SKIN DISEASE RISK PREDICTION

Year	Author	Method Used	Accuracy	Research Gap	Epidemiology data	Dataset
2025	Abbas et al.	Transfer Learning (DL), Explainable AI	CNN - 98%, ResNet-50: 84% DenseNet-121: 89% accuracy	Fully image-based	Not used	Large image datasets
2025	Hoque et al.	Epidemiological survey analysis	Identified major predictors.	Not ML based;	Partially analysis	Field survey data (Bangladesh)
2025	Islam et al.	Cross-section asses	58% fungal infections, >20% scabies cases	No predictive modelling;	Epidemiological analysis only	Clinic-based records
2024	Hasan et al.	Risk factor analysis.	Scabies prevalence 18–34%	Lacks ML prediction;	Yes	Field survey (Madrasahs)
2024	Wan et al.	ML models on (EHR).	AUC ≈ 0.82–0.83 (high performance)	Fully digital EHR system;	No	Large-scale EHR dataset.
2024	Yusra et al.	Hybrid ML	99.26% skin-disease detection	ML-based diagnosis	No	Image datasets
2024	Panwar et al.	ML models	Used simple ML models	Limited dataset	Yes	Survey dataset
2024	Vayadande et al.	ML for risk prediction.	Effective for health surveys	Not dermatology-specific;	Yes	General health datasets
2022	Meena et al.	KNN, SVM, Random Forest.	97% (RF)	No privacy, hygienic also not rural	Partially	Hospital dataset
2022	Chouhan et al.	Economic impact study	Environment risk related	Not ML modelling	Epidemiological	Livestock dataset
2021	Samiul Huq et al.	Community based	Find major skin disease	No predictive	Yes	Clinical survey data

2017	Esteva et al.	CNN, Deep learning	Dermatologist-level accuracy	Not usable in low-resource settings;	No	ISIC image dataset

Many studies have used deep learning or ensemble machine learning techniques to detect skin diseases using infected skin images or well-recorded clinical datasets. These methods frequently rely on high-quality imaging, lab data, or urban lifestyle questionnaires, which are challenging to maintain in rural settings in Bangladesh. In the framework of lightweight, epidemiology-based risk prediction, Table I thus classifies the current literature into broad categories and identifies its primary shortcomings. [24] used ensemble models to obtain 97% accuracy for structured tabular data on the UCI dermatology dataset, however the features are specialized biopsy attributes rather than community epidemiology. Although they don't focus on rural skin conditions, epidemiology-based machine learning research such as [26] for parasitic infections and [19] for EHR-based melanoma risk demonstrate that survey data can support prediction. Although they did not concentrate on dermatology specifically, [13], [14] demonstrated lightweight ML on general health surveys.

TABLE II
LIMITATIONS OF EXISTING AI-BASED APPROACHES

Type	**Model**	**Limitations**	**Related work**
Image-based	CNN / transfer learning on dermoscopic images (ResNet-50, DenseNet-121, sequential CNN, etc.)	Requires dermatoscopes or high-quality clinical images, GPUs, and stable internet; not feasible for most rural Bangladeshi clinics	[20], [21], [22]
Image-based	Advanced deep models and ensembles (Xception, Inception-v3, Inception-ResNet-v2, MobileNet, multi-CNN	Optimised for large, multi-class image datasets; high computational cost; no integration of hygiene, socio-economic, or environmental variables.	[20], [21]
Image-based	Classical ML with texture features (GLCM, color statistics) + DT, SVM, KNN on ISIC / HAM10000	Depends on careful preprocessing (hair removal, segmentation, denoising) and good dermoscopic images; unsuitable where only tabular clinic data exist.	[19]
Image-based (mobile / app)	Mobile / app-based systems combining CNNs with ensemble and data-mining algorithms	Improves accessibility but still relies on smartphone cameras and connectivity; does not exploit routine epidemiological records from rural facilities.	[21], [23]
Tabular clinical (dermatology)	Ensemble data-mining on UCI dermatology dataset.	Uses biopsy and histopathology attributes collected in specialist hospitals;	[24]
lifestyle (skin)	ML on survey-based lifestyle and treatment data (LR, DT, RF, CatBoost, GBC, LightGBM) for chronic skin diseases	Focuses on symptom improvement in urban specialty clinics; does not model infectious vs non-infectious skin-disease risk in rural populations such as Bangladesh.	[25]
Epidemiology ML (non-skin infection)	ML-based risk-factor analysis using epidemiological survey data for intestinal parasitic infections	Shows how ML can use socio-demographic, environmental and haematological features for infection risk, but targets intestinal parasites (not skin diseases) and an Ethiopian context.	[26]
Epidemiology (skin prevalence & QoL)	Community prevalence and DLQI / CDLQI studies of skin disease in rural populations.	Quantify burden and quality-of-life impact but remain descriptive and do not propose ML-based risk-prediction tools for frontline workers.	[25], [28]
System-level (skin ML)	Reviews of ML/DL for skin-lesion recognition (traditional ML + many	Summarise image-centric pipelines and big datasets; provide little guidance on lightweight, interpretable, epidemiology-based models for low-resource settings.	[21], [29]

	CNN/UNet variants; multiple public image datasets)		

Table II highlighted that most current research is either image-based or depends on structured clinical datasets. Although they are descriptive rather than predictive, recent reviews and quality of life studies like [25], [4] further emphasize the social and psychological burden of skin disease. The primary image-based and epidemiology-based research mentioned above are summarized in Table I, and a summary of their technological limitations is given in Table II. However, only a small number of studies use epidemiological survey variables to create lightweight, comprehensible risk-prediction models appropriate for rural Bangladesh.

TABLE III
COMPARATIVE ANALYSIS OF THE PROPOSED EPIDEMIOLOGY-BASED FRAMEWORK WITH PREVIOUS SKIN DISEASE PREDICTION STUDIES

Study	**Data type**	**Methods**	**Best performance**	**Context vs this work**
Abbas et al. (2025)	Dermoscopic images	Sequential CNN, ResNet, DenseNet	98% accuracy, 99% AUC	High-resource, image-based; no rural epidemiology
Verma et al. (2019)	UCI dermatology dataset	Ensemble data-mining models	97% accuracy	Specialist clinic attributes; no hygiene/living-condition variables
Zafar et al. (2022)	Survey data (intestinal parasites)	LR, SVM, RF, XGBoost with SMOTE	AUC > 0.8	Epidemiology-based but not skin diseases; Ethiopian setting
Park et al. (2024)	Lifestyle data (chronic skin)	LR, RF, boosting models	High F1-scores	Urban specialty clinics; focuses on symptom control
This study	Epidemiological survey (skin risk)	KNN, RF, LR, NB	88% train–test, 80.2% 10-fold accuracy	Lightweight, interpretable model for rural Bangladesh

Table III demonstrates that most of the previous research either uses well recorded clinical datasets or dermoscopic pictures, which limits its applicability to rural settings with limited resources

However, the suggested framework, is more appropriate for rural Bangladeshi settings since it only uses lightweight models and epidemiological survey variables.

III. Research Methodology

A. THEORETICAL FRAMEWORK

These days, people in rural Bangladesh are aware of skin disease in rural Bangladesh, and they can understand that it happens for many reasons. As they are linked to social, economic, and environmental problems. According to the World Health Organization (WHO), this is explained through the Social Determinants of Health idea. This says that human health is not only about our body but also about the world around us [11]. Importantly, education, cleanliness, money, housing, and gender can also change a person's health. When the community does not have good conditions or low income, they are more likely to have skin problems [17].

The Social Determinants of Health (SDH) framework, which emphasizes that individual health outcomes are impacted not only by biological factors but also by education, income, family, water and sanitation, and larger environmental conditions, serves as the overall foundation for this study. In rural Bangladesh, where overpopulation, polluted water, and inadequate sanitary facilities significantly impact disease risk, these social and structural factors are particularly significant for skin diseases. Simultaneously, the work adopts a Supervised Machine Learning (SML) approach to risk prediction, wherein a model learns a mapping from epidemiological input features (hygiene, environmental, and demographic variables) to an output label representing the type or degree of skin-disease risk. This combination of SDH and SML offers an empirical basis for identifying community members who are most at risk and could thus benefit from early intervention utilizing epidemiology data.

Computer science domain, especially artificial intelligence, machine learning, is now used to help identify health risks.

There are some normal or lightweight models, like decision KNN, logistic regression, and random forests can find and explain the links between many variables. These models use patient data from surveys or health center records to study people's health [18].

Also, the benefits of using more models help to read data about people, families, and their lives. This makes risk accuracy better and helps to target those who need it most. For example, it can also help to detect which areas of communities have a high risk of getting a certain skin issue. Skin disease prevalence is significantly influenced by environmental and behavioral variables, including household obstruction of personal hygiene, and water quality [19]. As a result, these factors are included in the study's epidemiological dataset.

Thus, the theoretical conceptual paper figures on two interconnected frameworks that connect with public health and artificial intelligence.

1. Social Determinants of Health (SDH): Skin diseases are spreading due to socio-economic, environmental, and household factors.
2. Supervised Machine Learning (SML): Like Logistic Regression, Naïve Bayes, KNN, and Random Forest learn patterns from provided epidemiological datasets.

For this study, frameworks will provide a strong, clear base. Without using clinical images, epidemiological data can be used securely in the model to predict the risk of skin diseases in rural communities.

B. Conceptual Framework

The proposed conceptual framework is designed with a structured approach. By using epidemiological data and lightweight machine learning models to detect the risk of dermatological problems in rural communities in Bangladesh. The framework consists of four key stages: (a) data collection, (b) data preprocessing, (c) modelling and analysis, and (d) evaluation. Each stage is important for achieving practical accuracy in underserved areas with limited resource facilities.

C. *Data Collection*

The initial plan of data collection is to record epidemiological data from Bangladeshi rural clinics and hospitals. The variable will be the patient demographics (age, gender, marital status, education, income), sanitation and hygienic behavior (frequency of bathing, hand washing, soap use), environmental features (source of water, crowded household, pet contact), and related to work (chemicals, sunshine). During the physical clinical visits inside the doctor's consultation room. The expert dermatologist assigns the dependent variable to three classes: 0 - Not a skin disease, 1 - Contagious, 2 - Not Contagious.

To ensure the ethical and responsible handling of patient data, privacy protection was maintained throughout the data collection phase. All records were stored using anonymized identifiers, and directly identifying attributes were kept in a separate, access-restricted file. Clinical data were encrypted at rest and only deidentified or aggregated datasets were used for analysis. These measures were implemented to prevent unauthorized disclosure, maintain confidentiality, and uphold ethical standards for handling health information.

TABLE IV
SAMPLE OF EPIDEMIOLOGICAL DATASET

Id	Age	Gender	Occupation	Water	Birth	Household	Pet Contact	Skin Status
R001	30	Male	Farmer	Pond	3-5 times / week	>6 persons/room	Yes	Contagious
R002	45	Female	Housewife	Tube well	Daily	4–5 persons/room	No	No disease
R003	18	Male	Student	River	1-2 times / week	>6 persons/room	Yes	Contagious
R004	55	Female	Farmer	Pond	Daily	3–4 persons/room	No	Non-
R005	27	Male	Labor	Tube well	2-3 times / week	4–5 persons/room	Yes	Contagious

D. Data Pre-Processing

The collected dataset uses several preprocessing techniques before the ML model is used:

1. Filled the missing Values: To avoid any kind of bias, missing values are either input or deleted.
2. Outlier Detection: To identify any human errors.
3. Encode the categorical Variables: Encode the numerical types to nominal variables like gender, occupation, educational level, and disease types.
4. Classes Balancing: Uses the resampling techniques, like supervised sampling or SMOTE, where appropriate. These techniques are employed to increase the class distribution, for example, the class-2 ratio.
5. Scaling the features: algorithms like KNN, normalizing, or standardization are used.

So, these preprocessing steps enhance the prediction performance and model accuracy.

E. Modelling and Analysis

The models were used with four supervised learning algorithms: K-Nearest Neighbors (KNN), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR). KNN is a non-parametric classifier, like an instance-based classifier, that assigns a new sample a class label based on the majority class of its k nearest training samples in feature space. In this study, we used Euclidean distance and fixed k = 5 neighbors. RF is a collection of decision trees trained on bootstrap samples. Each tree is constructed using a random subset of characteristics and determined by majority voting among the trees. Moreover, we employed 100 trees with at least two samples per leaf and a maximum depth of, say, 10. We used the Gaussian Naïve Bayes classifier, a probabilistic classifier that applies Bayes' theorem under the presumption of conditional independence amongst predictors. We trained a multinomial logistic regression model with L2 regularization penalty and regularization strength C = 1.0. LR is a linear model that uses the logistic function to predict the likelihood of belonging to each class.

To achieve the best accuracy, every model is trained using '10-fold cross-validation and train–test splitting (60:40, 70:30, 80:20, 90:10). The most important features of the dataset (sun exposure, chemical contact, and bathing frequency) can be identified using feature importance analysis.

According to initial performance, KNN and Random Forest frequently outperform LR and NB, achieving 88% in the 90:10 split and 79–80% in 10-fold cross-validation.

F. Evaluation & performance metrics

There are several metrics used to get the best model performance:

1. The highest precision indicates the minimal false positives. Precision estimates the percentage of positive cases that are correct.
2. Cohen's kappa statistic measures the degree of agreement between the model's prediction accuracy and the actual labels; values nearer 1 denote stronger deal.
3. The confusion matrix was used to generate several common classification measures that were used to evaluate the model's performance.
4. The percentage of real positive cases that the model correctly detects is called recall. A high recall indicates few false negatives. When classes are unbalanced, the F1-score, which is the harmonic mean of precision and recall, provides a single metric that balances both.
5. The model's ability to distinguish between classes across all potential classification limits is summarized by the Area Under the Curve (ROC-AUC); greater AUC values suggest better overall separability between positive and negative classes.

The metrics are computed as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Cohen's Kappa K} = \frac{p_o - p_e}{1 - p_e}$$

Here p_o is an observed agreement (accuracy) and p_e is expected agreement.

$AUC = \int_0^1 TPR(FPR)\, d(FPR)$ Here, TPR = $\frac{TP}{TP+FN}$, FPR= $\frac{FP}{FP+TN}$

Here, TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively, with the positive class denoting the risk of skin diseases.

Based on the results, the highest Kappa values (0.60–0.62) were obtained by KNN and Random Forest, suggesting moderate to good agreement. However, the Naïve Bayes performed worse (~63%) than Logistic Regression, which generated moderate accuracy (~70%).

This prediction confirms that this framework is workable for skin disease risk in underserved areas. The framework is visualized in Figure 1.

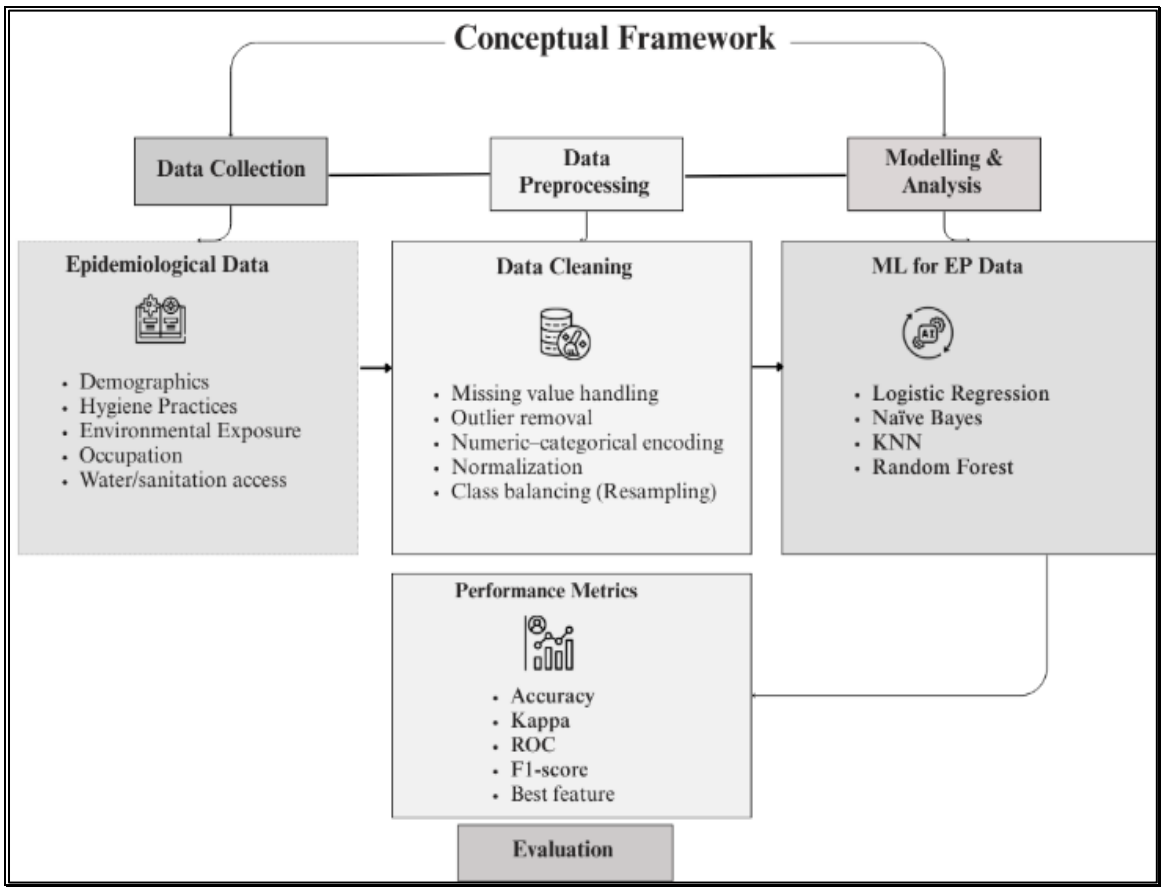

Fig. 1 Research methodology for epidemiology-based AI skin disease risk prediction.

IV. Experimental Analysis

To review a 500-epidemiology data set, the data were pre-processed in many ways, like addressing missing values, dropping the outliers, encoding numerical and categorical factors, leveling feature scales, and using resampling techniques. After preprocessing, several supervised machine learning models like Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest classifiers were assessed for their capacity to predict outcomes from the dataset. This allowed for a methodical comparison of performance across various algorithmic families.

A. Classification performance of the model:

According to 10-fold cross-validation, Random Forest performed the best overall, with 80.2% accuracy and a significant Kappa value of 0.6216, suggesting strong agreement beyond chance. KNN came in second with 79% accuracy, whereas Naïve Bayes and Logistic Regression had lower accuracies of 63% and 70.6%, respectively, demonstrating that instance-based and tree-based approaches outperformed linear and probabilistic models for this dataset. The accuracy output is visualized in Table V.

TABLE V
10-FOLD CROSS-VALIDATION ACCURACY OF ML MODELS ON THE EPIDEMIOLOGICAL DATASET

Algorithm	Accuracy (%)	Kappa	ROC
Logistic Regression	70.6	0.44	0.77
Naïve Bayes	63	0.28	0.75
KNN	79	0.60	0.82
Random Forest	80.2	0.62	0.88

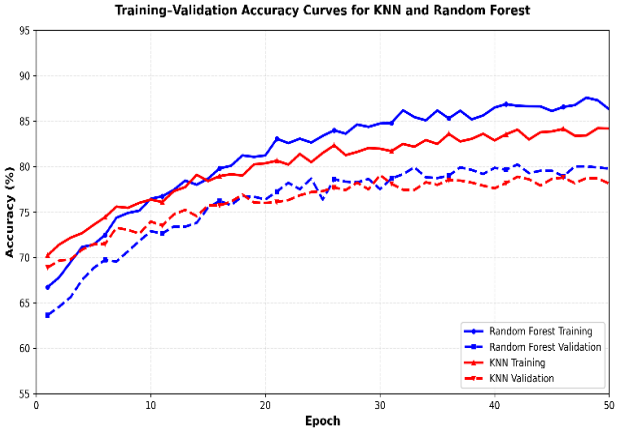

Fig. 2 Training–validation accuracy curve.

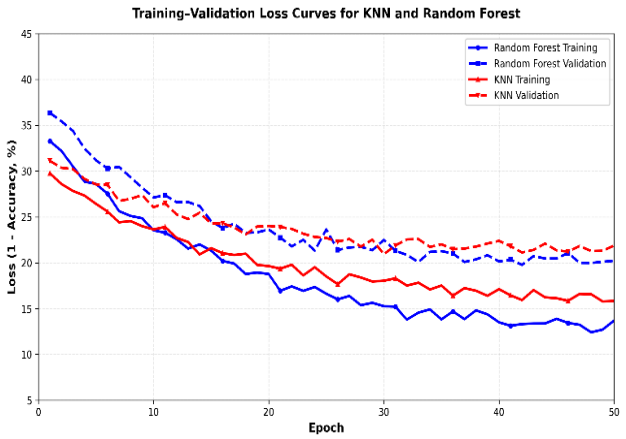

Fig. 3 Training–validation loss curve.

In the Figure 2 shown the training curves with validation accuracies stabilizing between 79 and 80% without significant gaps from, both KNN and Random Forest achieve smooth convergence.

However, the models generalize well and do not show significant overfitting on the epidemiology dataset, as shown by the associated loss curves in Figure 3, which drop slowly and do not diverge.

B. Train and test splitting:

Across various train-test splits, KNN and Random Forest produced the most consistent and often higher accuracies; performance improved as training size grew. KNN's accuracy peaked at 88% at the 90:10 split, while Random Forest's accuracy peaked at 84%. Both algorithms outperformed Naïve Bayes and Logistic Regression, which displayed greater variability and lower scores. The train-split accuracy output is visualized in Figure 4.

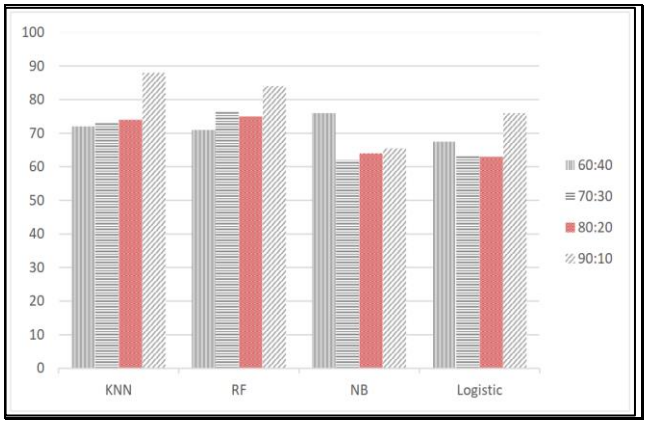

Fig. 4 Train–Test Split Analysis

C. Best Algorithm:

The accuracy scores of four machine learning models were tested to predict skin diseases. The best accuracy was 88% for KNN, 84% for Random Forest, and 76% for both Naïve Bayes and Logistic Regression. This graphic demonstrates that KNN is the dataset's most efficient algorithm, surpassing the competitors in predicting. The best algorithm accuracy is visualized in Figure 5.

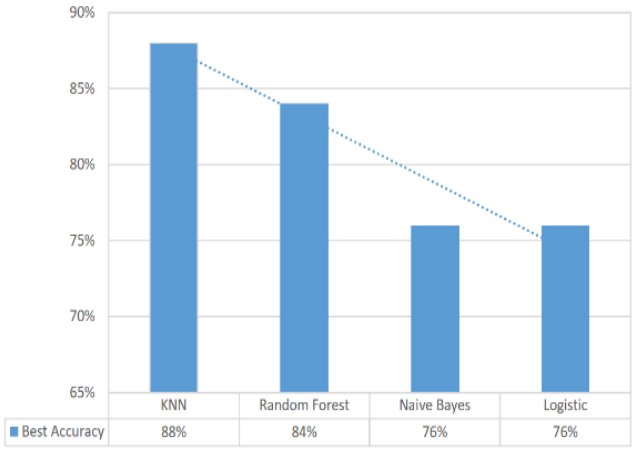

Fig. 5 Best Algorithm Accuracy

D. Best Fit Features:

The categories and importance of various factors influencing skin disease risk. Sun exposure, chemical exposure, and bathing frequency are rated as highly important and span environmental, occupational, and hygiene categories. Shared clothing and family history have medium to high significance within behavioural and genetic domains, while pet contact and household environment play a moderate role as environmental factors. This highlights the multidimensional nature of risk, covering demographic, behavioural, and ecological aspects. The best features shown is in table VI.

TABLE VI
THE BEST EPIDEMIOLOGICAL FEATURES FOR SKIN-DISEASE RISK PREDICTION.

Feature	Category	Importance level
Sun exposure	Environmental	High
Chemical exposure	Occupational	High
Bathing frequency	Hygiene	High
Shared clothing	Behavioural	Medium–high
Household environment	Environmental	Moderate
Pet contact	Environmental	Moderate

V. Discussion

Although this study is conceptual, preliminary experiments were conducted using an epidemiology-based dataset to assess the feasibility of developing lightweight machine learning models for rural Bangladesh. This conceptual framework aims to present a lightweight machine learning method for predicting skin disease risks using

epidemiological factors. Despite the study's conceptual nature, preliminary testing was conducted to assess the viability of different conventional machine learning methods using epidemiological data linked to skin diseases.

These initial experiments provided insights into the potential applicability of conventional machine learning methods in resource-constrained settings.

Across all experimental runs, KNN and Random Forest consistently outperformed Naïve Bayes and Logistic Regression in terms of model accuracy. However, KNN obtained 79% accuracy with a kappa value of 0.605 in 10-fold cross-validation, while Random Forest obtained 80.2% accuracy with a kappa of 0.6216. These kappa values show moderate to large agreement, demonstrating that the algorithms can learn significant epidemiological patterns and outperform chance.

About 70% accuracy was attained by logistic regression, indicating that while linear correlations do exist, they are insufficient on their own to fully capture the intricacy of skin-disease patterns. The independence assumption does not match well with coupled epidemiological variables, as evidenced by the lowest performance of Naïve Bayes (~63%). When multiple train–test splits were tested, the most stable performance was also observed in KNN and Random Forest. Accuracy increased when the training size increased (90:10 split), reaching up to 88% (KNN) and 84% (Random Forest). This trend indicates that these models benefit from larger training samples and can generalize well to unseen data.

However, it is important to recognize a few limitations. First, the current results may not accurately reflect the diversity of real rural communities in Bangladesh because they are based on a single dataset with a small sample size and class imbalance. Second, it is impossible to establish a causal association between risk factors and skin disease because the variables are taken from cross-sectional survey data. Third, the models' performance may alter when used in new contexts, such as various districts or primary-care institutions with distinct population characteristics, as they have not yet been prospectively verified in standard clinical workflows. Therefore, using locally gathered epidemiology data, additional validation is needed from various rural areas.

Lastly, although basic privacy and encryption techniques were implemented in the data collection and preprocessing phase, it will be crucial to incorporate improved privacy-preserving strategies to guarantee the safe and moral handling of patient data during large-scale deployment. Overall, these limitations suggest that the current findings should be considered preliminary; yet they offer a valuable basis for creating interpretable, cost-effective, and privacy-conscious prediction frameworks for the risk of skin diseases in underserved areas.

VI. Conclusion

This conceptual paper proposes a lightweight, epidemiology-driven model for predicting skin disease risk in underserved rural areas of Bangladesh. Early experiments conducted on an existing dataset indicate that conventional machine learning models, particularly KNN and Random Forest, achieve reliable predictive performance with accuracy reaching approximately 88% under train–test evaluation and 80% under 10-fold cross-validation. These results demonstrate that high-accuracy prediction can be made without the need for image-based diagnostic tools, utilizing only structured patient and environmental data. The outcomes validate the feasibility of the suggested framework for practical implementation in rural clinics. Additionally, privacy protection is another important factor to secure patient information In addition, privacy protection was integrated as a core design element to ensure secure and ethical handling of patient information. Collectively, these contributions provide a foundation for an interpretable and context-appropriate risk-prediction model that may strengthen early disease detection capabilities in low-resource communities Future work will involve acquiring epidemiological data directly from rural Bangladeshi populations, integrating privacy-preserving techniques into the risk-detection process, and validating the proposed framework within real-world healthcare workflows.

Acknowledgment

The authors hereby thank the IJPCC reviewers for their assistance in reviewing the manuscript and determining that it is suitable for publication.

Conflict Of Interest

The authors declare that there is no conflict of interest

Authors contribution statement

All authors contributed to the conception and design of the study, the development of the conceptual framework, and the interpretation of the outcomes. The first author led the literature review, data preprocessing, the model development, and drafting of the manuscript, while the co-authors provided supervision, domain expertise, model evaluations and critical revisions. All authors read and approved the final version of the manuscript.

Data availability statement

In this study, the dataset is not publicly available due to patient confidentiality and health centers restrictions. The dataset and analysis scripts can be obtained from the corresponding author upon reasonable request and subject to institutional and ethical approval.

ETHICS STATEMENT

The epidemiological data set used in this study was separated and anonymized before analysis. The original data collection followed institutional and authorities' ethical guidelines for research involving patient participants, and informed consent was obtained in the primary studies where required. Clinical records were stored with encrypted identifiers to maintain confidentiality.

REFERENCES

[1] T. Hoque, Md. R. Islam, and A. Akter, "Common skin diseases in children: A cross-sectional study from a semi-urban community in Bangladesh," *Int. J. Pediatr. Neonatol.*, vol. 7, no. 1, pp. 66–70, Jan. 2025, doi: 10.33545/26648350.2025.v7.i1b.122.

[2] K. Islam, M. I. Islam, T. Jahan, M. A. Yusuf, S. H. Chowdhury, and F. T. Zohora, "Pattern of Skin Diseases among Rural Adult Patients Attending the Dermatology OPD at A Tertiary Care Hospital Bangladesh," *Bangladesh J. Med. Microbiol.*, vol. 19, no. 1, pp. 54–59, Apr. 2025, doi: 10.3329/bjmm.v19i1.80603.

[3] N. Islam and M. I. H. Shakil, "Epidemiology of scabies among resident school and madrasah children: An observational study," *Int. J. Dermatol. Sci.*, vol. 7, no. 1, pp. 06–09, Jan. 2025, doi: 10.33545/26649772.2025.v7.i1a.44.

[4] Mohammad Samiul Huq, Abu Hena Chowdhury, Towhida Noor, and Saleheen Huq, "Psycho-social determinants and magnitude of public health problems of psoriasis in Bangladesh," *World J. Adv. Res. Rev.*, vol. 10, no. 2, pp. 108–118, May 2021, doi: 10.30574/wjarr.2021.10.2.0207.

[5] C. S. Chouhan *et al.*, "Epidemiology and economic impact of lumpy skin disease of cattle in Mymensingh and Gaibandha districts of Bangladesh," *Transbound. Emerg. Dis.*, vol. 69, no. 6, pp. 3405–3418, Nov. 2022, doi: 10.1111/tbed.14697.

[6] M. J. Hasan, M. A. Rafi, T. Choudhury, and M. G. Hossain, "Prevalence and risk factors of scabies among children living in Madrasahs (Islamic religious boarding schools) of Bangladesh: a cross-sectional study," *BMJ Paediatr. Open*, vol. 8, no. 1, p. e002421, June 2024, doi: 10.1136/bmjpo-2023-002421.

[7] N. Ahmed, M. Islam, and S. Farjana, "Pattern of Skin Diseases: Experience from a Rural Community of Bangladesh," *Bangladesh Med. J.*, vol. 41, no. 1, pp. 50–52, May 2014, doi: 10.3329/bmj.v41i1.18784.

[8] S. Abbas, F. Ahmed, W. A. Khan, M. Ahmad, M. A. Khan, and T. M. Ghazal, "Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence," *Sci. Rep.*, vol. 15, no. 1, p. 1746, Jan. 2025, doi: 10.1038/s41598-024-83966-4.

[9] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[10] R. Parvin *et al.*, "Clinical Epidemiology, Pathology, and Molecular Investigation of Lumpy Skin Disease Outbreaks in Bangladesh during 2020–2021 Indicate the Re-Emergence of an Old African Strain," *Viruses*, vol. 14, no. 11, p. 2529, Nov. 2022, doi: 10.3390/v14112529.

[11] "WHO Initiative on artificial intelligence for skin conditions." Accessed: Nov. 28, 2025. [Online]. Available: https://www.who.int/initiatives/who-initiative-on-artificial-intelligence-for-skin-conditions

[12] A. Krumeich and A. Meershoek, "Health in global context; beyond the social determinants of health?," *Glob. Health Action*, vol. 7, no. 1, p. 23506, Dec. 2014, doi: 10.3402/gha.v7.23506.

[13] P. Panwar, S. Bangwal, U. Pasbola, A. Kumar, A. Sar, and T. Choudhury, "Diagnosis and Prediction of Skin Diseases Using Deep Learning for Rural Healthcare," in *2024 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS)*, Dehradun, India: IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ISTEMS60181.2024.10560209.

[14] K. Vayadande, A. A. Bhosle, R. G. Pawar, D. J. Joshi, P. A. Bailke, and O. Lohade, "Innovative approaches for skin disease identification in machine learning: A comprehensive study," *Oral Oncol. Rep.*, vol. 10, p. 100365, June 2024, doi: 10.1016/j.oor.2024.100365.

[15] N. Yusra, K. K, S. A, and D. V, "Interpretable machine learning for dermatological disease detection: Bridging the gap between accuracy and explainability," *Comput. Biol. Med.*, vol. 179, Sept. 2024, doi: 10.1016/j.compbiomed.2024.108919.

[16] E. B. Weiner, I. Dankwa-Mullan, W. A. Nelson, and S. Hassanpour, "Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice," *PLOS Digit. Health*, vol. 4, no. 4, p. e0000810, Apr. 2025, doi: 10.1371/journal.pdig.0000810.

[17] M. Marmot, R. Bell, and P. Goldblatt, "Action on the social determinants of health," *Rev. DÉpidémiologie Santé Publique*, vol. 61, pp. S127–S132, Aug. 2013, doi: 10.1016/j.respe.2013.05.014.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[19] G. Wan, "Title Individualized melanoma risk prediction using machine learning with electronic health records," 2024, doi: https://doi.org/10.1101/2024.07.26.24311080.

[20] M. Ahammed, Md. A. Mamun, and M. S. Uddin, "A machine learning approach for skin disease detection and classification using image segmentation," *Healthc. Anal.*, vol. 2, p. 100122, Nov. 2022, doi: 10.1016/j.health.2022.100122.

[21] N. Fatima, S. A. M. Rizvi, and M. S. B. A. Rizvi, "Dermatological disease prediction and diagnosis system using deep learning," *Ir. J. Med. Sci. 1971 -*, vol. 193, no. 3, pp. 1295–1303, June 2024, doi: 10.1007/s11845-023-03578-1.

[22] Muddasar Abbas, Muhammad Imran, Abdul Majid, and Nadeem Ahmad, "Skin Diseases Diagnosis System Based on Machine Learning," *J. Comput. Biomed. Inform.*, vol. 4, no. 01, Dec. 2022, doi: 10.56979/401/2022/53.

[23] B. Suman, N. Harika, B. Sruthi, and B. Bhagyasree, "Prediction of Skin Diseases Using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 8, pp. 791–796, Aug. 2022, doi: 10.22214/ijraset.2022.46138.

[24] A. K. Verma, S. Pal, and S. Kumar, "Comparison of skin disease prediction by feature selection using ensemble data mining techniques," *Inform. Med. Unlocked*, vol. 16, p. 100202, 2019, doi: 10.1016/j.imu.2019.100202.

[25] C.-Y. Park, J. Joo, O.-H. You, S. Yi, C.-Y. Kim, and A.-R. Jo, "Development of a predictive model for managing lifestyle behaviors among patients with chronic skin diseases: Using machine learning techniques," *Inform. Med. Unlocked*, vol. 48, p. 101528, 2024, doi: 10.1016/j.imu.2024.101528.

[26] A. Zafar *et al.*, "Machine learning-based risk factor analysis and prevalence prediction of intestinal parasitic infections using epidemiological survey data," *PLoS Negl. Trop. Dis.*, vol. 16, no. 6, p. e0010517, June 2022, doi: 10.1371/journal.pntd.0010517.

[27] C. I. Wootton *et al.*, "Assessing skin disease and associated health-related quality of life in a rural Lao community," *BMC Dermatol.*, vol. 18, no. 1, p. 11, Dec. 2018, doi: 10.1186/s12895-018-0079-8.

[28] R. R. Yotsu *et al.*, "Skin disease prevalence study in schoolchildren in rural Côte d'Ivoire: Implications for integration of neglected skin diseases (skin NTDs)," *PLoS Negl. Trop. Dis.*, vol. 12, no. 5, p. e0006489, May 2018, doi: 10.1371/journal.pntd.0006489.

[29] J. Sun *et al.*, "Machine Learning Methods in Skin Disease Recognition: A Systematic Review," *Processes*, vol. 11, no. 4, p. 1003, Mar. 2023, doi: 10.3390/pr11041003.

[30] M. K, N. N. K. Veni, B. S. Deepapriya, P. A. H. Vardhini, B. Kalyani, and S. L, "A Novel Method for Prediction of Skin Diseases Using Supervised Classification Techniques," Apr. 08, 2022, *In Review*. doi: 10.21203/rs.3.rs-1509955/v1.