




# Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study

Anurag Kumar Verma<sup>1</sup> · Saurabh Pal<sup>1</sup>  · Surjeet Kumar<sup>1</sup>

Received: 26 April 2019 / Accepted: 5 July 2019 /

Published online: 27 July 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

Nowadays, skin disease is a major problem among peoples worldwide. Different machine learning techniques are applied to predict the various classes of skin disease. In this research paper, we have applied six different machine learning algorithm to categorize different classes of skin disease using three ensemble techniques and then a feature selection method to compare the results obtained from different machine learning techniques. In the proposed study, we present a new method, which applies six different data mining classification techniques and then developed an ensemble approach using bagging, AdaBoost, and gradient boosting classifiers techniques to predict the different classes of skin disease. Further, the feature importance method is used to select important 15 features which play a major role in prediction. A subset of the original dataset is obtained after selecting only 15 features to compare the results of used six machine learning techniques and ensemble approach as on the whole dataset. The ensemble method used on skin disease dataset is compared with the new subset of the original dataset obtained from feature selection method. The outcome shows that the dermatological prediction accuracy of the test dataset is increased compared with an individual classifier and a better accuracy is obtained as compared with subset obtained from feature selection method. The ensemble method and feature selection used on dermatology datasets give better performance as compared with individual classifier algorithms. Ensemble method gives more accurate and effective skin disease prediction.

**Keywords** Skin disease · Dermatology · Extra tree classifier · Radius neighbors classifier · Passive aggressive classifier

---

✉ Saurabh Pal  
drsaurabhpal@yahoo.co.in

Surjeet Kumar  
<https://scholar.google.co.in/citations?hl=en&user=u2ogX5wAAAAJ>

<sup>1</sup> MCA Department, VBS Purvanchal University, Jaunpur 222002 Uttar Pradesh, India

## Introduction

The application of machine learning algorithms is widely used in the area of the medical field. Various disease diagnosis classification algorithms are developed to find the high accuracy for predicting the disease. Many machine learning algorithms are developed for predicting the various type of disease at early level after examining the various attributes of the disease. These algorithms are widely applicable in breast cancer, kidney diseases, thyroid disease, diabetes, cancer diseases, erythematous diseases, and many more. In this research paper, we choose erythematous disease for analysis. Various classification algorithms are applied and then ensemble methods are applied in this study. Another approach using feature selection is applied with these classification algorithms to obtain the accuracy of the prediction for the application to make an expert system [1, 2]. Various studies have been done in this field and some are discussed below.

Polat and Güneş [3] achieved 96.71% classification accurate rate on the diagnosis of erythematous diseases using a novel hybrid intelligence method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problem.

Immagulate and Vijaya [4] focus on non-melanoma skin cancer and classify types, using support vector machines (SVM) to accurately predict disease types. The chrominance and texture features are extracted pre-processed training datasets.

Chang and Chen [5] discussed the decision tree combined with neural network classification methods to construct the best predictive model of dermatology. The learning predicted and analyzed six common skin conditions. All classification techniques can predict disease fairly accurately, and the neural network model has the highest accuracy of 92.62%.

Ramya and Rajeshkumar [6] discussed the Gray-Level Co-Occurrence Matrix (GLCM) technique for finding features from segmented disease and classifying skin disease based on fuzzy classification, which is more accurate than existing ones.

Übeyli and Doğdu [7] find the results of a study where they deployed a k-means clustering approach to classifying erythematous diseases dataset. The results of the study indicate approximately 94% overall classification accuracy rate when using 5 out of the 6 decision classes (excluding Pityriasis rubra pilaris—20 instances).

Güvenir et al. [8] discussed the VF15 classification algorithm which represents a concept description by a set of feature value intervals. The classification of a new instance is based on voting among the classifications made by the value of each feature separately.

Ahmed et al. [9] discussed clusters of pre-processed data, using k-means clustering algorithms to separate related and unrelated data into skin disease. Frequent patterns were evaluated using the MAFIA algorithm. Decision tree and AprioriTid algorithms are used to extract frequent patterns from clustered datasets.

Fernando et al. [10] discussed a disease prediction method, DOCAID, to predict malaria, typhoid fever, jaundice, tuberculosis, and gastroenteritis based on patient symptoms and complaints using the naïve Bayesian classifier algorithm. The authors reported an accuracy rate of 91% for predicting disease.

Jaleel et al. [11] extracted these features using a 2D wavelet transform method and then classified them using a back propagation neural network (BPNN). They classify the dataset as cancer or non-cancer.

Theodoraki et al. [12] developed a predictive model to predict the final outcome of a seriously injured patient after an accident. The investigation includes a comparison of data mining techniques using classification, clustering, and association algorithms. Using this

analysis, they obtained results in terms of sensitivity, specificity, positive predictive value, and negative predictive value, and compared results between different predictive models.

Sharma and Hota [13] used SVM and ANN data mining techniques, to classify various types of erythema-squamous diseases. They used a confidential weighted voting scheme to combine the two technologies to achieve the highest accuracy of 99.25% in the training and 98.99% in the testing phases.

Rambhajani et al. [14] used a Bayesian classification to classify the erythematous-squamous disease dataset. The author used the best first search feature selection technology technique, and they removed 20 features from the dermatology dataset collection collected by the University of California Irving repository and then used the Bayesian technology to achieve 99.31% accuracy.

Bakpo and Kabari [15] used ANN for diagnosis of different skin diseases and they achieved 90% accuracy. There are a few unique features for skin cancer regions.

Manjusha et al. [16] predict different skin diseases using the naïve Bayesian algorithm. Automatic identification of circulatory disease dermatological features was extracted from the local binary pattern from affected skin images and used for classification.

Yadav and Pal [17] discussed about women thyroid prediction using data mining techniques. They used two ensemble techniques. The first ensemble technique generated by decision tree and the second was generated by bagging and boosting techniques. They observed dataset for thyroid symptom and find better accuracy results.

Tuba et al. [18] applied the automatic erythematous-squamous diseases detection method by optimized support vector machine. Parameters of the support vector machine were tuned by the recent swarm optimization algorithm, elephant herding optimization. They tested on a standard dataset that contains data for 366 patients with one of six different erythematous-squamous diseases. The accuracy achieved by them is 99.07%.

Zhang et al. [19] applied a hybrid approach that uses granular computing (GrC) and supports vector machines (SVM). The authors reviewed and evaluated most of the past artificial intelligence systems used for the diagnosis of erythematous-squamous disease and tabulated the classification results of all these algorithms. The results achieved averages of sensitivity and specificity as 98.43% and 99.71%, respectively.

In this study, an attempt is done to use machine learning methods to ensemble six different classifiers, which are the following:

### **Passive Aggressive Classifier**

Passive aggressive algorithms are a family of online learning algorithms (for both classification and regression) proposed by Crammer et al. The passive aggressive classifier (PAC) algorithm is perfect for classifying massive streams of data. It is easy to implement and very fast but does not provide global guarantees like the support vector machine (SVM). Passive: if correct classification, keep the model; Aggressive: if incorrect classification, update to adjust to this misclassified example.

### **Linear Discriminant Analysis**

Linear discriminant analysis (LDA) is closely related to the analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. LDA is also closely related to principal component

analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data.

### Radius Neighbors Classifier

Radius neighbors classifier (RNC) is very similar to a k-neighbors classifier with the exception of two parameters. First, in radius neighbors classifier, we need to specify the radius of the fixed area used to determine if an observation is a neighbor using radius. Unless there is some substantive reason for setting radius to some value, it is best to treat it like any other hyper-parameter and tune it during model selection. The second useful parameter is outlier-label, which indicates what label to give an observation that has no observations within the radius—which itself can often be a useful tool for identifying outliers.

### Bernoulli Naïve Bayesian

Bernoulli naïve Bayesian (BNB) implements the naïve Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, the decision rule for Bernoulli naïve Bayes is based on:

$$P(x_i|y) = P(i|y)x_i + (1-P(i|y))(1-x_i)$$

which differs from the multinomial NB's rule in that it explicitly penalizes the non-occurrence of a feature  $i$  that is an indicator for class  $y$ , where the multinomial variant would simply ignore a non-occurring feature.

### Gaussian Naïve Bayesian

Gaussian naïve Bayesian (NB) implements the Gaussian naïve Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The parameters  $\sigma_y$  and  $\mu_y$  are estimated using maximum likelihood.

### Extra Tree Classifier

An “extra trees” classifier (ETC), otherwise known as an “Extremely randomized trees” classifier, is a variant of a random forest. Unlike a random forest, at each step, the entire sample is used and decision boundaries are picked at random, rather than the best one. In real-world cases, performance is comparable with an ordinary random forest, sometimes a bit better.

Three different ensemble techniques, bagging classifier, AdaBoost classifier, and gradient boosting classifier, are applied to predict skin disease.

## Methods

Figure 1 demonstrates the whole structure of the methodology used in this research paper. The figure demonstrates the different data mining methods (I) PAC, (II) LDA, (III) RNC, (IV) BNB, (V) NB, and (VI) ETC used in this study. The approach used in this paper is completely data driven. In this paper, we have applied the six classification algorithms to measure the accuracy and sensitivity of the predicted values of skin disease classes. The obtained values are then improved using ensemble technique by using the bagging classifier, AdaBoost classifier, and gradient boosting classifier. The same techniques are again applied on the skin disease dataset by using feature selection—in this selection, we obtained 15 important attributes using the feature importance method. Now, we reduce the dataset and take only 15 attributes and 366

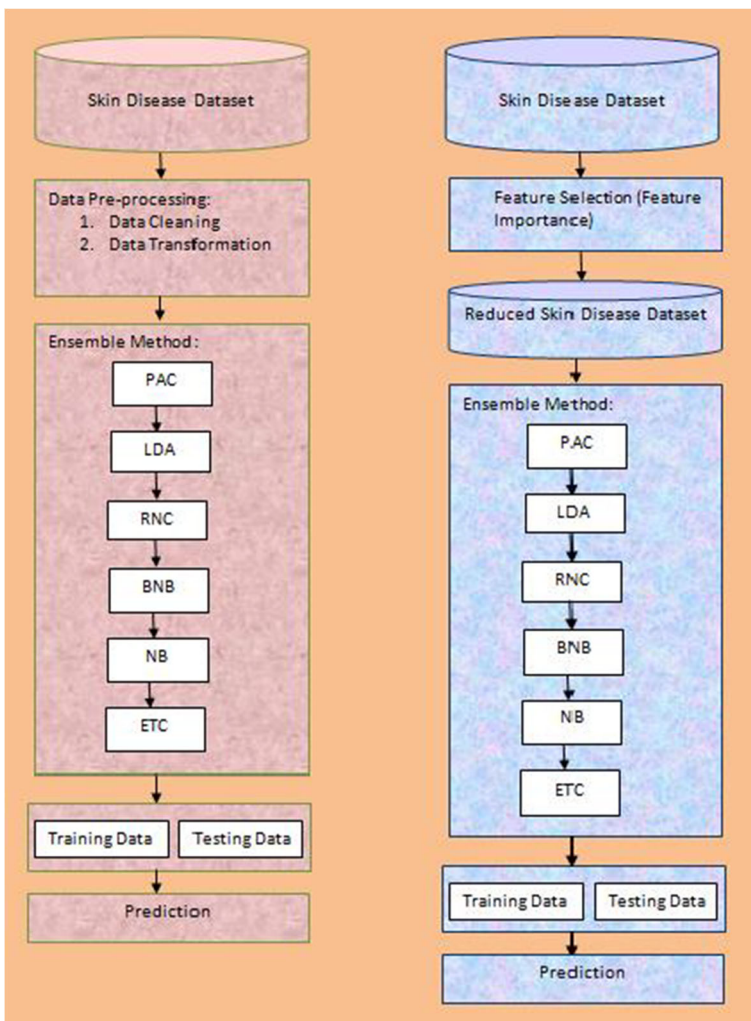


Fig. 1 Methodological approach for skin disease

instances of the dataset and again evaluate the accuracy of the prediction of skin disease dataset. A comparative study is then performed to evaluate the best prediction.

## Dataset Analysis

The database used in this study is taken from the UCI machine learning repository (<http://archive.ics.uci.edu/ml>). Briefly, this dataset was formed to examine skin disease and classify the type of erythemato-squamous diseases. This dataset contains 34 variables; in this dataset, 33 variables are linear and 1 variable is nominal. These six classes of skin disease include the following: C1, psoriasis; C2, seborrheic dermatitis; C3, lichen planus; C4, pityriasis rosea; C5, chronic dermatitis; C6, pityriasis rubra. Biopsy is one of the basic treatments in diagnosing these diseases. A disease may also contain the properties of another class of disease in the initial stage, which is another difficulty faced by dermatologists when performing the different class of diagnosis of these diseases. Initially, patients were first examined with 12 clinical features, after which the assessment of 22 histopathological attributes was performed using skin disease samples. Histological features were identified by analyzing the samples under a microscope. If any of the diseases is found in the family, the family history attribute in the dataset constructed for the domain has a value of 1 (one), and if not found, the value is 0 (zero). The age of the patient is used to indicate age characteristics. All other attributes (clinical and histopathological both) were assigned a value in the range from 0 to 3 (0 = absence of features; 1, 2 = comparative intermediate values; 3 = highest value). There are six classes of

**Table 1** Skin disease dataset [8]

Classes (no. of instances)	Clinical	Histopathological attributes
C1: psoriasis (112)	f1: erythema	f12: melanin incontinence
C2: seborrheic dermatitis (61)	f2: scaling	f13: eosinophils in the infiltrate
C3: lichen planus (72)	f3: definite borders	f14: PNL infiltrate
C4: pityriasis rosea (49)	f4: itching	f15: fibrosis of the papillary dermis
C5: chronic dermatitis (52)	f5: Koebner phenomenon	f16: exocytosis
C6: pityriasis rubra (20)	f6: polygonal papules	f17: acanthosis
	f7: follicular papules f18: hyperkeratosis	f19: parakeratosis
	f8: oral mucosal	f20: clubbing of the rete ridges involvement
	f9: knee and elbow	f21: elongation of the rete ridges
	f10: scalp involvement	f22: thinning of the suprapapillary epidermis
	f11: family history	f23: spongiform pustule
	f34: age	f24: munro microabscess
		f25: focal hypergranulosis
		f26: disappearance of the granular layer
		f27: vacuolization and damage of basal layer
		f28: spongiosis
		f29: saw-tooth appearance of rete ridges
		f30: follicular horn plug
		f31: perifollicular parakeratosis
		f32: inflammatory mononuclear infiltrate
		f33: band-like infiltrate

erythemato-squamous disease, with 366 instances and 34 attributes in the domain. Table 1 summarizes the contents of the attributes.

### Feature Selection

Feature selection is a process to automatically select only those features from the dataset which contribute most to the prediction variable or output. Irrelevant attributes in the dataset can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression. Three benefits of performing feature selection before modelling dataset are the following:

1. Reduces over fitting
2. Improves accuracy
3. Reduces training time

There are many feature selection techniques (univariate selection, recursive feature elimination, principal component analysis, and feature importance) available, but in this research paper, we have applied the feature importance method to choose 15 most important attributes from the skin disease dataset. The importance of attributes is shown in Fig. 2. Figure 2 shows an important score for each attribute where the larger the score, the more important the attribute.

We can get the feature importance of each feature of the dataset by using the feature importance property of the model. Feature importance gives a score for each feature of the dataset, the higher the score, the more important or relevant is the feature towards the output variable. The feature importance is calculated based on following algorithm:

1. Train an FR suing  $D_{train}$  (all  $k$  features)
2. Compute average RMSE of a model for cross-validation data (C-V)
3. Rank performance by  $VI_k = \sum_{\theta}^{\varphi} \frac{(\epsilon_{\theta k} - \epsilon_{\theta nk})}{\varphi \delta}$
4. For each subset of  $k$   $k_i = k - 1, k - 2, k - 3, \dots, 1$  do
5. Train a new forest from  $k_i$  features, highest VI
6. Calculate the average RMSE of the model on C-V set
7. Re-rank the features ( $k$ )
8. Find  $j_i$  with smallest RMSE

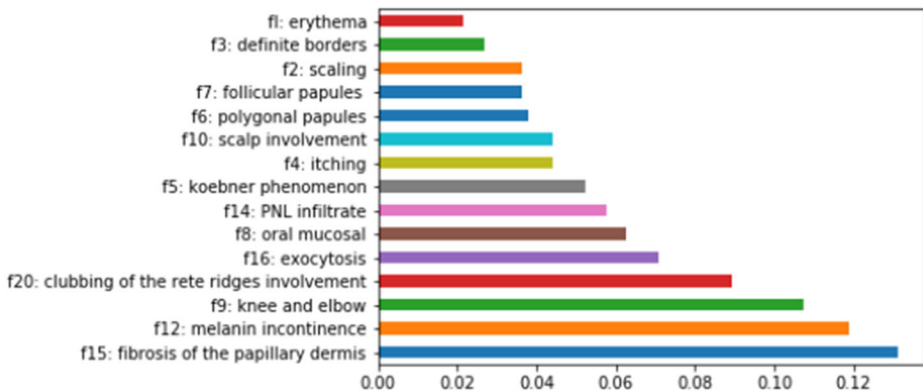


Fig. 2 Important attributes

## Ensembles Method

In this research paper, the ensemble method is used as a method to find the accuracy of the skin disease dataset to improve the performance of algorithms. We will apply an ensemble method to combine six different machine learning algorithms using bagging classifier, AdaBoost classifier and gradient boosting classifier.

1. *Bagging*. Bootstrap aggregating, also known as bagging, is a machine learning model aggregation technique designed to improve the stability and accuracy and to reduce variance to avoid over fitting of machine learning algorithms applied in regression and classification methods.

- A. Takes original dataset  $D$  with  $N$  training examples
- B. Creates  $M$  copies  $\{\widetilde{D}_m\}_{m=1}^M$ 
  - a. Each  $\widetilde{D}_m$  is generated from  $D$  by sampling with replacement.
  - b. Each dataset  $\widetilde{D}_m$  has the same number of examples as in dataset  $D$ .
  - c. These datasets are reasonably different from each other.
- C. Trains models  $h_1, \dots, h_M$  using  $\widetilde{D}_1, \dots, \widetilde{D}_M$ , respectively
- D. Uses an averaged model  $h = \frac{1}{M} \sum_{m=1}^M h_m$  as the final model

2. *AdaBoost*, also known as adaptive boosting, is a machine learning meta-algorithm which uses the concept of combining independent individual hypothesis in a sequential order to improve the accuracy. Basically, boosting algorithms convert the weak learners into strong learners. It is well designed to address the bias problems.

- A. Given training data  $(x_1, y_1), \dots, \dots, \dots, (x_N, y_N)$  with  $y_n \in \{-1, +1\}, \forall n$
- B. Initialize weight for each example  $(x_n, y_n) : D_1(n) = \frac{1}{N}, \forall n$
- C. For round  $t = 1:T$

- a. Learn a weak  $h_t(x) \rightarrow \{-1, +1\}$  using training data weighted as per  $D_t$ .
- b. Compute the weighted fraction of errors of  $h_t$  on this training data

$$\epsilon_t = \sum_{n=1}^N D_t(n) \mathbb{I}[h_t(x_n) \neq y_n].$$

- c. Set “importance” of  $h_t : \alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ .
- d. Update the weight of each example.

$$D_{t+1}(n) \propto \begin{cases} D_t(n) \times \exp(\alpha_t) & \text{if } h_t(x_n) = y_n \\ D_t(n) \times \exp(-\alpha_t) & \text{if } h_t(x_n) \neq y_n \end{cases} \\ = D_t(n) \times \exp(\alpha_t y_n h_t(x_n))$$

- e. Normalize  $D_{t+1}$  so that it sums to

$$1: D_{t+1}(n) = \frac{D_{t+1}(n)}{\sum_{m=1}^N D_{t+1}(m)}$$

- D. Output the boosted final hypothesis

$$H(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$$



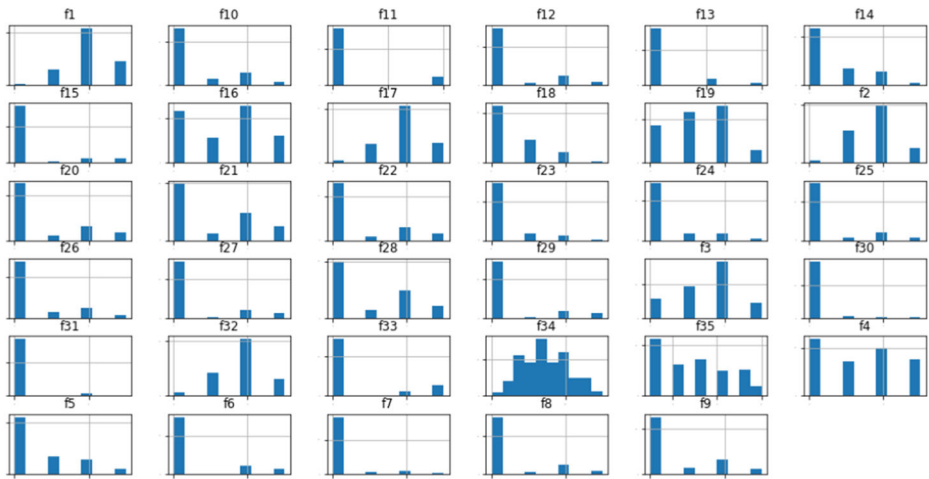


Fig. 3 Visualization of skin disease dataset

3. *Gradient boosting.* It is an example of a generalized boosting algorithm. Gradient boosting machine learning technique works for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

- A. Given training data  $(x_1, y_1), \dots, (x_N, y_N)$ .
- B. Initialize  $\hat{f}_0$  with a constant.
- C. For  $t = 1$  to  $M$  do the following:
  - a. Compute the negative gradient  $g_t(x)$ .
  - b. Fit a new base-learner function  $h(x, \theta_t)$ .
  - c. Find the best gradient descent step size

$$\rho_t: \rho_t = \operatorname{argmin}_{\rho} \sum_{i=1}^N \Psi [y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)].$$

- d. Update the function estimate  $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$ .

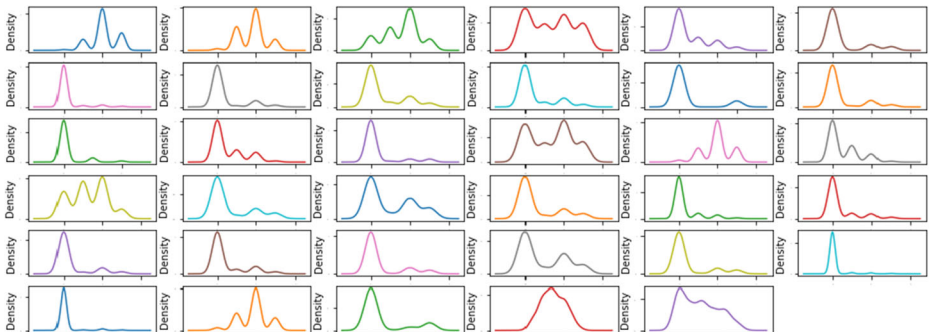


Fig. 4 Density map of skin disease dataset

**Table 2** Mean and standard deviation of classifiers

Algorithms	Mean value	Standard deviation	Accuracy
PAC	0.4627	0.0950	94.59%
LDA	0.5547	0.0802	96.57%
RNC	0.5513	0.0915	58.67%
BNB	0.5409	0.0585	98.64%
NB	0.4659	0.0694	91.89%
ETC	0.4860	0.0759	93.24%

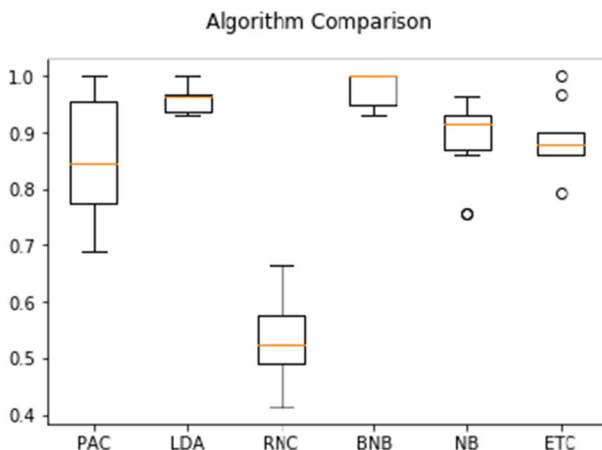
## Results

We have conducted two experiments to obtain the prediction results for skin disease datasets. In the first experiment, we have used the skin disease dataset obtained from the UCI machine repository which consists 34 variables and one target variable “Class.” Then we have applied an ensemble method to obtain the predicted values. While in the second experiment, we have used feature selection method with the same classifier and ensemble methods to check the results obtained are better or not. Same six classifiers are again used in the feature selection dataset to obtain the predictions.

### Experiment 1

Before analyzing the dataset, we first visualize the distribution of values as shown in Figs. 3 and 4. Figure 3 depicts the distribution of values of skin disease used in our study containing 366 instances and 35 attributes. Each feature shows the distribution of frequency among the 4 values (0 to 3) except the feature f34 (which is age) and feature f35 (which is targeting variable class).

The density map is a smooth continuous version of the smoothed graph estimated from the data. The most common form of estimation is called the kernel density estimation. In this method, a continuous curve (core) is drawn at each individual data point, and then all of these curves are added together for a single smoothed density estimate. The most commonly used

**Fig. 5** Accuracy of different classifier algorithms

**Table 3** RMSA values for ensemble methods

Ensemble methods	Root mean square error (RMSE)					
	PAC	LDA	RNC	BNB	NB	ETC
Bagging	0.0651	0.0751	0.0564	0.0698	0.0852	0.0752
AdaBoost	0.0785	0.0698	0.0854	0.0654	0.0531	0.0753
Gradient boosting	0.0587	0.0687	0.0782	0.0854	0.0912	0.0741

kernel is Gaussian (which produces a Gaussian bell curve at each data point). The density map of the attributes is shown in Fig. 4.

We have used the Python code to calculate the prediction on the skin disease dataset to calculate the mean value, standard deviations, and accuracy of the six different classification techniques. The values obtained by the six classifiers are shown in Table 2.

The accuracy of any classification method is calculated using the confusion matrix and is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

In another term, it can be represented as:

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

where TP is true positives; TN, true negatives; FP, false positives; and FN, false negatives.

From Table 2, it is clear that the highest accuracy obtained is 98.64% in the case of Bernoulli naïve Bayesian classification algorithm and the least accuracy is 58.67% for radius neighbors classification algorithm with a radius of 10 ( $r = 10$ ). The box and whisker plot of six classifier methods are shown in Fig. 5.

Now, we have applied three ensemble methods boosting, AdaBoost, and gradient boosting to calculate the root mean square error values of six classification methods. This is the square root of the mean of the squared errors. RMSE indicates how close the predicted values are to the actual values; hence, a lower RMSE value signifies that the model performance is better. RMSE value is calculated by the formula:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

The RMSE values for each classification algorithms using ensemble techniques are shown in Table 3.

**Table 4** Output of evaluating algorithms on the scaled dataset

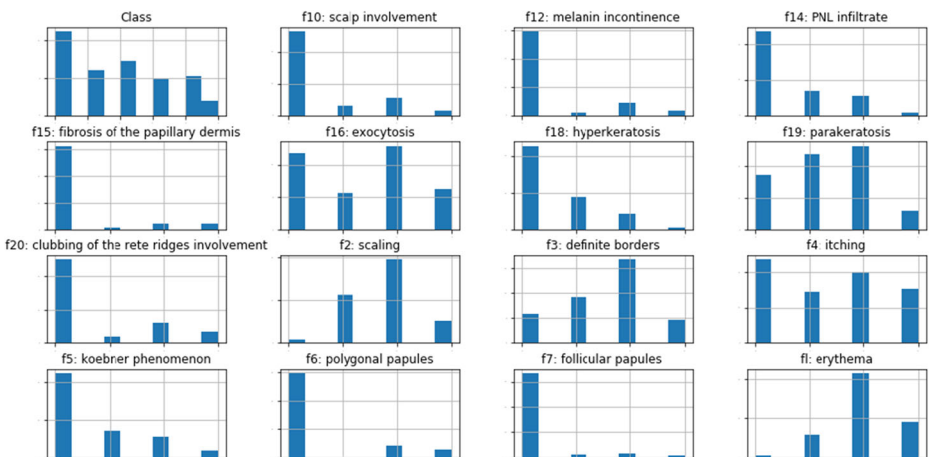
Ensemble methods	Accuracy (%)						Time (in seconds)
	PAC	LDA	RNC	BNB	NB	ETC	
Bagging	97.35	95.23	94.87	96.56	95.76	96.32	0.06
AdaBoost	98.21	96.56	95.23	97.58	96.45	97.56	0.06
Gradient boosting	96.62	98.57	99.46	94.63	97.98	96.39	0.05

**Table 5** Output of ensemble method

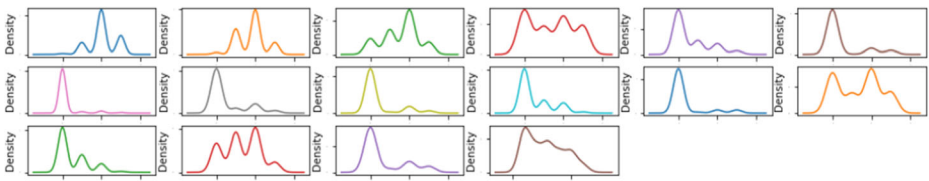
Method	Accuracy	Confusion matrix	Values				
Bagging	97.35%	[[24 0 0 0 0 0] [ 0 10 0 0 0 0] [ 0 0 11 0 0 0] [ 0 0 0 14 0 0] [ 0 0 0 0 11 0] [ 1 0 0 0 0 3]]	Precision	Recall	f1-score	Support	
			1	0.96	1.00	0.98	24
			2	1.00	1.00	1.00	10
			3	1.00	1.00	1.00	11
			4	1.00	1.00	1.00	14
			5	1.00	1.00	1.00	11
			6	1.00	0.75	0.86	4
			Avg/total	0.99	0.99	0.99	74
AdaBoost	98.21%	[[24 0 0 0 0 0] [ 0 0 9 0 0 1] [ 0 0 11 0 0 0] [ 0 0 14 0 0 0] [ 1 0 10 0 0 0] [ 0 0 0 0 0 4]]	1	0.96	1.00	0.98	24
			2	0.00	0.00	0.00	10
			3	0.25	1.00	0.40	11
			4	0.00	0.00	0.00	14
			5	0.00	0.00	0.00	11
			6	0.80	1.00	0.89	4
			Avg/total	0.39	0.53	0.43	74
			Gradient boosting	99.46%	[[24 0 0 0 0 0] [ 0 9 0 0 0 1] [ 0 0 11 0 0 0] [ 0 1 0 13 0 0] [ 0 0 0 0 11 0] [ 0 0 0 0 0 4]]	1	1.00
2	0.90	0.90				0.90	10
3	1.00	1.00				1.00	11
4	1.00	0.93				0.96	14
5	1.00	1.00				1.00	11
6	0.80	1.00				0.89	4
Avg/total	0.98	0.97				0.97	74

The accuracy of three ensemble methods using six classification machine learning algorithms is shown in Table 4. It is clear that the accuracy found in Table 4 is higher than the accuracy found in Table 2. The accuracy of RNC in gradient boosting ensemble method has a very high increase as compared with individual RNC accuracy of 58.67%.

Now, we ensemble all the six classification algorithms using the bagging classifier, AdaBoost classifier, and gradient boosting classifier. The results obtained after applying these three ensemble methods are shown in Table 5.



**Fig. 6** Visualization of skin disease dataset



**Fig. 7** Density map of skin disease dataset

## Experiment 2

In the new obtained dataset, after feature selection using the feature importance method, we obtained 10 clinical features (f1, f3, f2, f7, f6, f10, f4, f5, f8, and f9) and 5 histopathological features (f14, f16, f20, f12, and f15). This shows that clinical features are more important than histopathological features in the accuracy of disease prediction. We first visualize the distribution of values as shown in Figs. 6 and 7. Figure 6 depicts the distribution of values of skin disease used in our study containing 366 instances and 16 attributes. Each feature shows the distribution of frequency among the 4 values (0 to 3) except the feature f35 (which is targeting variable class). Figure 7 depicts the density map of the features of the dataset.

We have again used the Python code to calculate the prediction on the skin disease dataset obtained from the feature selection method to calculate the mean value, standard deviations, and accuracy of the six different classification techniques. The value obtained from the six classifiers is shown in Table 6.

From Table 6, it is clear that the highest accuracy obtained is 97.29% in the case of Bernoulli naïve Bayesian classification algorithm and the least accuracy is 81.08% in the radius neighbors classification algorithm with a radius of 4 ( $r = 4$ ). The box and whisker plot of six classifier methods are shown in Fig. 8.

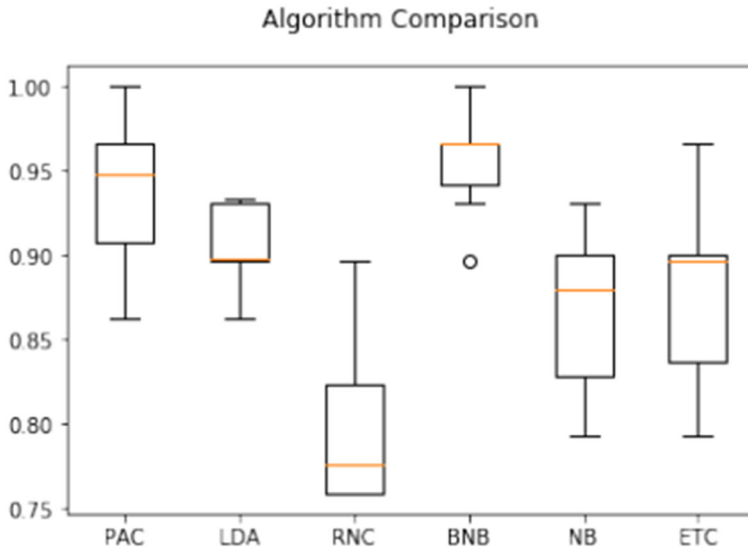
The root mean square error (RMSE) values obtained by a dataset from feature selection method are shown in Table 7.

The accuracy of three ensemble methods using six classification machine learning algorithms on a reduced dataset obtained after features selection is shown in Table 8.

Here, we observe that the accuracy of ensembles methods is increased in comparison to without ensemble data mining techniques. It is also clear that the accuracy obtained from experiment 2 (with feature selection on the reduced dataset) is higher than the accuracy obtained from experiment 1 (using three ensemble techniques on six classifiers on whole skin disease dataset). Now, we ensemble all the six techniques using the bagging classifier, AdaBoost classifier, and gradient boosting classifier. The results obtained after applying these three ensemble methods are shown in Table 9.

**Table 6** Mean and standard deviation of classifiers

Algorithms	Mean value	Standard deviation	Accuracy
PAC	0.9383	0.0399	94.59%
LDA	0.9074	0.0222	91.49%
RNC	0.7975	0.0484	81.08%
BNB	0.9555	0.0267	97.29%
NB	0.8696	0.0461	90.54%
ETC	0.8800	0.0584	94.59%



**Fig. 8** Accuracy of different classifier algorithms

The accuracy obtained in the feature selection method has improved over the first experiment which considered the full spectrum of data. The possible causes are due to the inclusion of irrelevant features, which does not play an important role in the prediction of skin diseases.

ROC curve represents a plot of the true positive rate (sensitivity) in the function of the false positive rate (100-specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal). The ROC curve of the three ensemble techniques bagging, AdaBoost, and gradient boosting is shown in Fig. 9.

## Discussion

In this paper, we have conducted two experiments differently one using three ensemble methods bagging, AdaBoost, and gradient boosting on six different classification algorithms PAC, LDA, RNC, BNB, NB, and ETC, and second using the feature selection method. This experiment is done on the UCI skin disease dataset which contains 366 instances and 34 attributes. The skin dataset consists of six classes of skin disease: C1, psoriasis; C2, seborrheic dermatitis; C3, lichen planus; C4, pityriasis rosea; C5, chronic dermatitis; C6, pityriasis rubra.

**Table 7** RMSA values for ensemble methods

Ensemble methods	Root mean square error (RMSE)					
	PAC	LDA	RNC	BNB	NB	ETC
Bagging	0.0521	0.0658	0.0425	0.0584	0.0754	0.0621
AdaBoost	0.0542	0.0536	0.0732	0.0532	0.0469	0.0611
Gradient boosting	0.0498	0.0539	0.0612	0.0733	0.0836	0.0699

**Table 8** Output of accuracy on the reduced dataset

Ensemble methods	Accuracy (%)						Time (in seconds)
	PAC	LDA	RNC	BNB	NB	ETC	
Bagging	98.56	97.25	95.42	96.89	96.21	97.32	0.05
AdaBoost	99.25	96.98	96.32	97.89	97.65	98.63	0.04
Gradient boosting	97.38	98.98	99.68	95.36	98.66	97.65	0.04

We have analyzed the mean, standard deviation, root mean square error, and accuracy of six different machine learning algorithms to obtain the highest accuracy of 98.64% in the case of Bernoulli naïve Bayesian classification algorithm. The performance demonstrated by the ensemble data mining techniques for skin disease prediction lies in input variable choice and classification method selection. A root mean square error is calculated for each ensemble methods. After applying ensembles method, we get 99.46% accuracy in the case of the gradient boosting classifier method.

In the second experiment, we first choose the important attributes using the feature importance method to obtain the most crucial fifteen features and then reduce the dataset as a subset of the original dataset. The reduced dataset contains 15 attributes and 366 patient records. The same ensemble methods and machine learning algorithms are applied and we get different results; the highest accuracy achieved in this experiment is 99.68% in the case of the gradient boosting classifier method.

A comparison of experiment one and two are shown in Table 10 (FS stands for feature selection).

**Table 9** Output of evaluating ensemble method

Method	Accuracy	Confusion matrix	Values				
Bagging	98.56%	[[24 0 0 0 0]]	Precision	Recall		f1-scor	Support
		[ 0 10 0 0 0]]	1	0.92	1.00	0.96	24
		[ 0 0 11 0 0]]	2	0.91	1.00	0.95	10
		[ 0 1 0 13 0]]	3	1.00	1.00	1.00	11
		[ 1 0 0 0 10]]	4	1.00	0.93	0.96	14
		[ 1 0 0 0 0 3]]	5	1.00	0.91	0.95	11
			6	1.00	0.75	0.86	4
	Avg/total	0.96	0.96	0.96	74		
AdaBoost	99.25%	[[23 1 0 0 0]]	1	0.96	0.96	0.96	24
		[ 0 10 0 0 0]]	2	0.36	1.00	0.53	10
		[ 0 0 11 0 0]]	3	1.00	1.00	1.00	11
		[ 0 14 0 0 0]]	4	0.00	0.00	0.00	14
		[ 0 0 0 0 11 0]]	5	1.00	1.00	1.00	11
		[ 1 3 0 0 0]]	6	0.00	0.00	0.00	4
			Avg/total	0.66	0.94	0.68	74
Gradient boosting	99.68%	[[23 1 0 0 0]]	1	0.96	0.96	0.96	24
		[ 0 9 0 1 0 0]]	2	0.75	0.90	0.82	10
		[ 0 0 11 0 0 0]]	3	1.00	1.00	1.00	11
		[ 0 2 0 12 0 0]]	4	0.92	0.86	0.89	14
		[ 0 0 0 0 11 0]]	5	1.00	1.00	1.00	11
		[ 1 0 0 0 0 3]]	6	1.00	0.75	0.86	4
			Avg/total	0.94	0.93	0.93	74

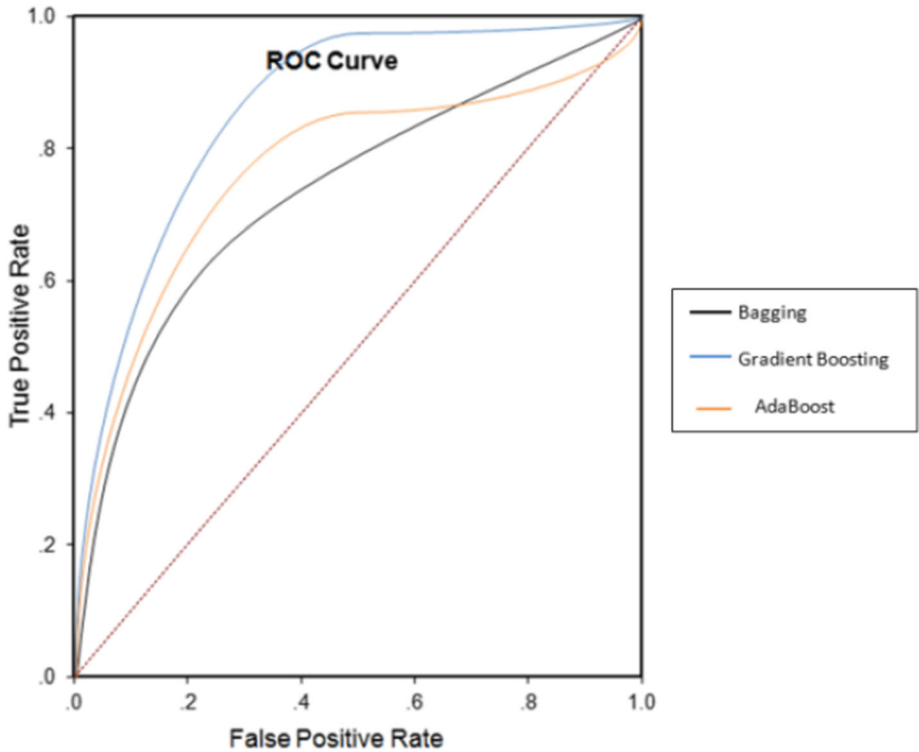


Fig. 9 Receiver operating characteristic (ROC) curve

In this comparison, we execute six classifiers using three different ensemble algorithms without feature selection (all 34 features used) and we calculated the corresponding accuracy of classifiers. The accuracies obtained with all features used and after feature selection in experiment 2 are shown in Table 10. The highest accuracy of each classifier is also presented in the same row in order to make comparison better.

The comparison from Table 10 shows the efficiency of the feature selection method on the reduced dataset is better than without the feature selection in each case, and we get higher accuracy.

Table 10 Comparison of accuracies in experiment 1 and experiment 2

Algorithms	Bagging			AdaBoost			Gradient boosting		
	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
	Without FS	With FS	Difference	Without FS	With FS	Difference	Without FS	With FS	Difference
PAC	97.35	98.56	1.21	98.21	99.25	1.04	96.62	97.38	0.76
LDA	95.23	97.25	2.02	96.56	96.98	0.42	98.57	98.98	0.41
RNC	94.87	95.42	0.55	95.23	96.32	1.09	99.46	99.68	0.22
BNB	96.56	96.89	0.33	97.58	97.89	0.31	94.63	95.36	0.73
NB	95.76	96.21	0.45	96.45	97.65	1.2	97.98	98.66	0.68
ETC	96.32	97.32	1	97.56	98.63	1.07	96.39	97.65	1.26



**Table 11** A few investigations which have dealt with skin disease mining

Author	Year	Method	Classification accuracy
Polat and Güneş [3]	2009	C4.5 and one-against-all	96.71%
Ubeyli [20]	2009	CNN	97.77%
Chang and Chen [5]	2009	Decision tree	80.33%
		Neural network	92.62%
Ubeyli and Dogdu [7]	2010	K-mean clustering	94.22%
Lekka and Mikhailov [21]	2010	Evolving fuzzy classification	97.55%
Xie and Wang [22]	2011	IFSFS and SVM	98.61%
Cataloluk and Kesler [23]	2012	Basic-KNN	94.4%
		Weighted KNN	96.36%
Olatunji and Arif [24]	2013	ANN	78.09%
		ELM	98.36%
Ravichandran et al. [25]	2014	FELM	93.00%
Amarathunga et al. [26]	2015	AdaBoost	85% for eczema
		Bayes net	95% for impetigo
		J48,	85% for melanoma
		MLP (naïve Bayes)	
Parikh et al. [27]	2015	ANN	97.17%
		SVM	94.04%
Maghooli et al. [28]	2016	CART	93.69%
Parvin and Jafar [29]	2017	Multi-SVM	97.4%
		KNN	90%
		Naïve Bayesian	55%
Zhou et al. [30]	2017	ANN	92.4%
Idoko et al. [31]	2018	CART	94.84%
		ARFCMC	75.96%
		ANFIS	95.50%
		AEC	97.32%
		FNN	98.37%
Zhang et al. [19]	2018	ANN	96.8%

To illustrate the success of our approach, the results obtained in this study were compared with other results given in the literature. In order to compare the efficiency of the proposed dermatological classification, we used a large number of technical studies using the same information but using different classification techniques and then developing a multi-model ensemble method. According to these studies, the same partitions of the above test datasets were followed. To illustrate this, the classification efficiency is compared with previous studies. This is shown in Table 11.

Although the predicted values after applying feature selection in experiment 2 gives better results than in experiment 1, it does not guarantee that it will perform in all cases, if we use more instances. The new feature selection process is used on only these 366 instances, so whether the selection of features is accurate or not will depend upon more instances, which can be used in a future study.

## Conclusion

Machine learning techniques play an important role in the diagnosis of diseases in the biotechnology field. Knowledge obtained using machine learning techniques can be used to develop expert systems, which provide help in predicting various types of disease. This paper

describes different data mining techniques for skin disease prediction. Six machine learning classification techniques PAC, LDA, RNC, BNB, NB, and ETC are used to classify the prediction of skin disease and three ensemble techniques bagging, AdaBoost, and gradient boosting classifiers are applied to improve the accuracy obtained by machine learning algorithms. A feature selection method is also applied on skin disease dataset and obtains more accurate accuracy of 99.68% in the case of gradient boosting ensemble method applied on RNC. We got the highest accuracy in the literature available on the skin disease dataset.

This study achieved higher accuracy as compared with previous research done in this field. This study can be useful to correctly predict the skin disease. Hence, we conclude that the proposed feature selection model can be of efficient use in the detection of erythematous-squamous diseases with improvement in speed and accuracy.

## Compliance with Ethical Standards

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Research Involving Human Participants and/or Animals** This paper does not contain any studies with human participants or animals performed by any of the authors.

**Abbreviations** PAC, passive aggressive classifier; LDA, linear discriminant analysis; RNC, radius neighbors classifier; BNB, Bemoulli naïve Bayesian; NB, Gaussian naïve Bayesian; ETC, extra tree classifier; FS, feature selection

## References

1. Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Chronic kidney disease: a predictive model using decision tree. *International Journal of Engineering Research and Technology*, *11*(11), 1781–1794.
2. Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119–126.
3. Polat, K., & Güneş, S. (2009). A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, *36*(2), 1587–1592.
4. Immagulate, I., & Vijaya, M. S. (2015). Categorization of non-melanoma skin lesion diseases using support vector machine and its variants. *International Journal of Medical Imaging*, *3*(2), 34–40.
5. Chang, C. L., & Chen, C. H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, *36*(2), 4035–4041.
6. Ramya, G., & Rajeshkumar, J. (2015). A novel method for segmentation of skin lesions from digital images. *International Research Journal of Engineering and Technology*, *2*(8), 1544–1547.
7. Übeyli, E. D., & Dođdu, E. (2010). Automatic detection of erythematous-squamous diseases using k-means clustering. *Journal of Medical Systems*, *34*(2), 179–184.
8. Güvenir, H. A., Demiröz, G., & İltter, N. (1998). Learning differential diagnosis of erythematous-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, *13*(3), 147–165.
9. Ahmed, K., Jesmin, T., & Rahman, M. Z. (2013). Early prevention and detection of skin cancer risk using data mining. *International Journal of Computer Applications*, *62*(4), 1–6.
10. Fernando, Z. T., Trivedi, P., & Patni, A. (2013). DOCAID: predictive healthcare analytics using naive Bayes classification. In *Second Student Research Symposium (SRS), International Conference on Advances in Computing, Communications and Informatics (ICACCT'13)*, 1–5.
11. Jaleel, J. A., Salim, S., & Aswin, R. B. (2012). Artificial neural network based detection of skin cancer. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *1*(3), 200–205.
12. Theodoraki, E. M., Katsaragakis, S., Koukouvinos, C., & Parpoula, C. (2010). Innovative data mining approaches for outcome prediction of trauma patients. *Journal of Biomedical Science and Engineering*, *3*(08), 791–798.

13. Sharma, D. K., & Hota, H. S. (2013). Data mining techniques for prediction of different categories of dermatology diseases. *Journal of Management Information and Decision Sciences*, 16(2), 103.
14. Rambhajani, M., Deepanker, W., & Pathak, N. (2015). Classification of dermatology diseases through Bayes net and best first search. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(5), 275–86.
15. Bakpo, F. S., & Kabari, L. G. (2011). Diagnosing skin diseases using an artificial neural network. In *Artificial Neural Networks-Methodological Advances and Biomedical Applications*, Suzuki K (ed.), Intech. Available from: <http://www.intechopen.com/articles/show/title/diagnosing-skin-diseases-using-an-artificial-neural-network>.
16. Manjusha, K. K., Sankaranarayanan, K., & Seenaa, P. (2014). Prediction of different dermatological conditions using naive Bayesian classification. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1), 864–868.
17. Yadav, D. C., & Pal, S. (2019). To generate an ensemble model for women thyroid prediction using data mining techniques. *Asian Pacific Journal of Cancer Prevention*, 20(4), 1275–1281.
18. Tuba, E., Ribic, I., Capor-Hrosik, R., & Tuba, M. (2017). Support vector machine optimized by elephant herding algorithm for erythematous-squamous diseases detection. *Procedia Computer Science*, 122, 916–923.
19. Zhang, X., Wang, S., Liu, J., & Tao, C. (2018). Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. *BMC Medical Informatics and Decision Making*, 18(2), 59.
20. Übeyli, E. D. (2009). Combined neural networks for diagnosis of erythematous-squamous diseases. *Expert Systems with Applications*, 36(3), 5107–5112.
21. Lekkass, S., & Mikhailov, L. (2010). Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. *Artificial Intelligence in Medicine*, 50(2), 117–126.
22. Xie, J., & Wang, C. (2011). Using support vector machines with a novel hybrid feature selection method for diagnosis of erythematous-squamous diseases. *Expert Systems with Applications*, 38(5), 5809–5815.
23. Cataloluk, H., & Kesler, M. (2012). A diagnostic software tool for skin diseases with basic and weighted K-NN in International Symposium on Innovations in Intelligent Systems and Applications, IEEE (2012), 1–4.
24. Olatunji, S. O., & Arif, H. (2013). Identification of erythematous-squamous skin diseases using extreme learning machine and artificial neural network. *ICTACT Journal of Softw Computing*, 4(1), 627–632.
25. Ravichandran, K. S., Narayanamurthy, B., Ganapathy, G., Ravalli, S., & Sindhura, J. (2014). An efficient approach to an automatic detection of erythematous-squamous diseases. *Neural Computing and Applications*, 25(1), 105–114.
26. Amarathunga, A. A. L. C., Ellawala, E. P. W. C., Abeysekara, G. N., & Amalraj, C. R. J. (2015). Expert system for diagnosis of skin diseases. *International Journal of Scientific & Technology Research*, 4(01), 174–178.
27. Parikh, K. S., Shah, T. P., Kota, R., & Vora, R. (2015). Diagnosing common skin diseases using soft computing techniques. *International Journal of Bio-Science and Bio-Technology*, 7(6), 275–286.
28. Maghooli, K., Langarizadeh, M., Shahmoradi, L., Habibi-koolae, M., Jebraeily, M., & Bouraghi, H. (2016). Differential diagnosis of erythematous-squamous diseases using classification and regression tree. *Acta Informatica Medica*, 24(5), 338.
29. Pravin, S. R., & Jafar, O. A. M. (2017). Prediction of skin disease using data mining techniques. *IJARCCCE*, 6(7), 313–318.
30. Zhou, H., Xie, F., Jiang, Z., Liu, J., Wang, S., & Zhu, C. (2017). Multi-classification of skin diseases for dermoscopy images using deep learning. In *2017 IEEE International Conference on Imaging Systems and Techniques (IST)* (pp. 1–5). IEEE.
31. Idoko, J. B., Arslan, M., & Abiyev, R. (2018). Fuzzy neural system application to differential diagnosis of Erythematous-squamous diseases. *Cyprus Journal of Medical Sciences*, 3(2), 90–97.