A Novel Method for Prediction of Skin Diseases Using Supervised Classification Techniques

Meena K (fuadam123@gmail.com)
GITAM University - Bengaluru Campus
N N Krishna Veni
Holy Cross College
B S Deepapriya
Sengunthar Engineering College
P A Harsha Vardhini
Vignan Institute of Technology and Science
BJD Kalyani
Institute of Aeronautical Engineering
Sharmila L
Dhaanish Ahmed College of Engineering

Research Article

Keywords: Skin disease, Feature selection, Supervised classification, KNN, SVM, Random forest

Posted Date: April 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1509955/v1

Abstract

Skin diseases are the most important worrying problems in societies because it affects the patients both physically and psychologically. Skin disease is one of the highly prone to risk with an association of climatic factors around the world. Predicting the skin disease cases associated with influencing factors is the most crucial task. It is very difficult task to identify the appropriate and optimal features for skin disease from the large volume of health sector data available in the world. Previous researchers applied different types of ensemble features selection techniques for the appropriate selection of features which gives highest accuracy with minimum computation time. Classification rate of any algorithm depends on feature extraction techniques and classifier used for classification purpose. Data availability is one of the most significant drawbacks in the health sector if data is available that might be in raw format. Filling missing value and type conversion almost takes 70% of the time. The missing value can be addressed by statistical parameters such as mean, average, and median with stand mechanism in machine learning. The objective of this paper is the selection of significant attributes and removes irrelevant features that affect model performance. The performance of skin disease data can be experimented through K Nearest Neighbor (KNN), Support Vector Machine (SVM) and random forest classifier. The efficiency of the proposed approach is measured through confusion matrix, accuracy, F-measure, precision and Recall.

1. Introduction

Climate change and global warning are the notable events for the cause of skin diseases. The earth's atmosphere is the combination of inwards solar radiation and outwards thermal radiation. Some of the factors influencing warming of the planet and climate change are greenhouse gases, radiation, carbon dioxide, wind patterns, methane, fossil flues, and ocean currents. Fossil fuel contributes more to climate change, and Few reports suggest population can be one of the main influencing factors [2, 15, 16, 17]. In previous studies, environmental science collective state from some years change of climate drastically increases [3, 9]. One of the active group IPCC reports that CO_2 huge increases 300 to 400 PPM accordingly increases in ocean 25%, global average temperature increases 0.18 to 0.22M. All the influencing factors lead to the change of climate which causes most of the infectious diseases such as dengue, chikungunya, malaria, heart stock ad skin-related cases rapidly increases in previous studies. In this work, we have considered skin disease cases associated with climate factors [8]. Ultra Violet rays damage the inner and outer pigments of the cell directly, which can lead to various types of skin diseases [20, 21].

Skin diseases are highly vulnerable when exposing to a long time to UV rays. It can be through the distributed age group in the data set, according to the report nearly 30 to 45 aged male groups are prone to higher risk than women [7, 10]. The international classification of diseases based on IPCC reports indicates 30% of skin diseases are register because of illness. Skin diseases arise because of burns on the body surface, globally and nationally, only less critical given for prevention and control when compared with other infectious. Skin diseases are classification into 15 types of sub-categories abscess, enzyme, fungal, psoriasis, scabies acne vulgaris, urticaria, alopeciaarete, priorities, and deceitsulcer. In

previous studies, many authors attempted to address this issue by applying traditional and analytical methods [4, 5, 20].

Supervised classification algorithm has been developed to predict the diseases in earlier stages from real time medical data. Even in real-time medical data consist of many missing values and some hidden patterns, every data should be properly pre-processed and select feature before the start of disease analysis [6, 13, 22]. Medical data available in real time can't be used directly for clinical analysis. 80% of the medical data requires preprocessing before applying feature extraction techniques. Preprocessing steps such as filling the missing values, convert the constant value to categorical and changing of column names for fundamental understandings is highly recommended for medical data analysis. One of the first steps is, have to fill the missing value; few stoical techniques help by applying machine learning mean and the average of the missing column. By using type conversation process change of continuous value to absolute values (vice versa) and column change function used for renaming the column name. Once all the above steps are executed, then the data set is ready for analysis; we have considered the next phases to be more critical and complex. As medical data comprise the number of attributes, even some irrelative qualities can affect the performance of the model. Machine learning models have a unique process called feature selection. As our data is supervised, we will be using only a supervised feature selection technique. The aim of the paper (I) Pre-process the raw data (II) selecting appropriate features by combination technique.

2. Literature Survey

Several researchers have proposed various machine learning based to detect the type of skin diseases. Here we briefly review some of the techniques as reported in the literature. Usually, other studies first they will apply feature selection than rating. The execution time of the model recorded high. Ahn and Hur [1] proposed genetic algorithm first classify and select the feature set. He proposed a filter base selection in the local region. The proposed model search for local neighbors sample and correlated with each other [12, 14, 24]. Antimicrobial resistance is a critical problem globally. The proposed method uses a time series technique explicitly to forecast the outbreak of diseases. They use the wrapper method for feature selection technique. The author proposed the Artificial Neural Network based feature selection technique to reduce and remove irrelative characteristics. In the unstructured text, clinical data increase in all health departments, and availability of such data is free. The author applied the dictionary-based technique. It is the hybrid approach which handles missing value and other data issue. The author proposed a two-step approach (I) to compress high dimension data and (ii) to different categorical and numeric value data with the missing value. The multi-label selection is always a complex problem because labeling is done one by one. The author proposed discriminative and relevant feature selection. The author suggested the X variance feature section approach in gene selection [12, 18, 19, 23].

3. Real Time Data Collection And Analysis

The real time skin diseases data is collected from popular private hospital in Chennai, it is always crowded with patients with various illnesses. According to the in-patient records, in an average of 2000 to 6000 patients visit the hospital per day in multiple departments. Among that 200 to 250 patients are found to be visiting the Dermatology department with skin-related ailments. The inflow of the patients to the Dermatology department explodes with the increase in the temperature. The dataset comprises data collected from year 2000 to 2018 (Temperature, Rainfall, Humidity and Precipitation) with daily and monthly readings. We have carried out this research in two phases, in first phase, we have collected the hospital data and experimented. In the second phase, we have received the climatic data set from the National Data Centre, India. This work aims to find the association between climate data and hospital data, and we have mentioned the Female as 0 and the Male as 1. From the experimental results, we found out that Males have been more affected when compared with the Females. The reason we found out is that it exposes them to the sun for a longer time than the Females. With the results, we have plotted a graph, and it is visible that 69.5% of Male and 31.5% of females are affected by skin disease.

We have proposed a framework to address this issue. The proposed frame work is a trial and error combination. Machine learning models are used to measure sensitivity and accuracy of the disease's outbreak. The output values are used for better forecasting and applying ensemble feature selection.

Table.1 Representation of all feature names with data types

Table 1
represents the 15 attributes, its data types and feature label assigned for experimental investigation.

Feature	Data Type	Feature label
Year_week	Int64	F1
Recorded_year	Int64	F2
Recorded_month	Int64	F3
Air_temp	Float64	F4
Humidity	Float64	F5
Surface_water3	Float64	F6
Total_vegetation	Float64	F7
Min_air_temp	Float64	F8
Surface_water5	Float64	F9
Total_precipitation	Float64	F10
Max_air_temp	Float64	F11
Total_precipitaion in KG	Float64	F12
Northeast_NVDI	Float64	F13
Mean_duepoint	Float64	F14
Mean_humidity	Float64	F15

Figure 2 represents the architecture diagram of the proposed approach. It consists of two phases namely training phase and testing phase. During training pahse, real time data collected from patients are propcessed and stored in database. During test phase also, the same procedure is used for preprocessing and feature extarction. The test feature extracted during this phase are compared against the training feature stored in the dataset by using KNN, SVM and Random forest algorithm.

3.1 Preprocessing Step

The data collected from hospital consists of various formats such as videos, structure, audio and image. During preprocessing, the data is converted in to machine-readable format, i.e., 0's and 1's. Machine learning mechanism has a single process call pre-processing by which a machine can read it. The data set is the combination of sample, Entities, points, cases, patterns and observation. The data object refers to the number of attributes or variables; data is classified into two types categorical and numerical.

Categorical is the Boolean set, which is constant [yes, no]. The numeric is continues, and the value is dynamic (temperature, age, etc.). Quality of data archived by applying the pre-processing technique, as

the data generated from the different source it is raw data, which can affect model accuracy. It forms the missing value gaps during data collection, either machine or human-made, a mistake at recording time. Eliminating rows and columns are a few methods, but this won't be useful because it reduces the sensitivity of data. The most commonly used plan for addressing missing value in rows and columns by mean, mode, and the median value of the relative feature.

3.2 Feature selection Step

Feature selection is a process of automatic selection of variables in the data set, which are more relevant for foresting methods. As medical data consist of irrelative attributes that may affect model performance, not all attributes are useful. Numbers of features always change the model and lead to complex. Any feature selection technique aims to improve the model performance, to provide a cost-effective and faster understanding of the hidden patterns. Feature selection has three types of methods. The various feature selection methods existing for data analysis is represented in Fig. 3

It ranks fitter method is applied by a statistical technique to assign a score to each variable depends upon score variable are selected. Wrapper method: selecting features on different features combination and compare with other combination. Ensemble methods like regularization technique to penalization of the data to reduce coefficients to zero.

3.3 Classification by Supervised learning model

In this paper, supervised learning methods such as KNN, SVM and Random forest are considered for classification. Machine learning algorithm performance is measure by statically test: precision, accuracy, and F – measure. Precision defines as a percentage of exact forecast accuracy class; accuracy defines as a percentage of correct forecast cases between all cases, F-measure defies as the weighted mean of recall and precision. These 3 performance metrics have used to select significant attributes; typically, most of the machine learning model uses some performance metrics. The performance metrics play an essential role in considerable selection attributes. The different combination set of the variable is applied. The main idea is to remove irrelative characteristics that are influencing the performance of the model. All three metrics performance is calculated and recorded in different tables below.

$$\textbf{Accuracy} = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \quad (1)$$

$$\textbf{F-Measure} = \frac{(2{*}\mathrm{Precision}{*}\mathrm{Recall})}{(\mathrm{Precision}+\mathrm{Recall})} \quad (2)$$

$$\textbf{Precision} = \frac{\mathrm{TruePositives}}{(\mathrm{TruePositives}+\mathrm{FalsePositives})} \quad (3)$$

4. Experimental Analysis

The dataset comprising of 15 features with 1500 instances are investigated through three different machine learning algorithms for analysis of hidden patterns. Our experiment is a trial-and-error

combination approach by selecting a set of attributes using machine learning pipeline and performance metrics F-measure, precision, and accuracy is applied for model validation. We have analyzed the data and describe the highest accuracy, highest precision, and the highest F-measure in respective Tables 3, 4 and 5.

4.1 Identification of Appropriate folds (K value)

The experiment is conducted by varying the K value from 5, 10, 15 and 20 for the classifiers such as KNN, SVM and Random forest and the results are presented in Table 2.

Table 2
Recognition rate of various classifiers with different K fold

K Fold	Classifier	Recognition Rate (%)
5	KNN	76
	SVM	81
	Random Forest	84
10	KNN	83
	SVM	89
	Random Forest	97
15	KNN	82
	SVM	86
	Random Forest	89
20	KNN	81
	SVM	83
	Random Forest	90

From Table 2, it is observed that K fold cross validation with K = 10 is giving the better results compared to other k values. The increase in values of K leads to misclassification of sample because of overfitiing problems in the data. Hence, it is decided that the remaining experiment is performed by choosing the K value as 10.

4.2 Performance of various classifiers by different metrics such as Recognition rate, precision and F-measure

Tables 3, 4 and 5 exhibits the performance metrics Recognition rate, precision and Recall values of the data set with different machine techniques.

Table 3
Recognition rate and optimal features of skin disease data through various classifiers

Classifier	Recognition Rate (%)	Feature set
KNN	83	1,3,5,7,9
SVM	89	1,2,5,9,11
Random Forest	97	1,2,4,6,8

From Table 3, it is observed that Random forest yield the highest accuracy compared to other models considered for experimental investigation. All the three models includes feature 1 for the computation of accuracy. Hence, it is considered as highly important feature for further analysis.

Table 4
Precision and optimal features of skin disease data through various classifiers

Classifier	Precision	Feature set
KNN	0.86	1,2,5,7,11
SVM	0.90	1,2,7,9,13
Random Forest	0.93	1,2,5,6,9

From Table 4, it is concluded that Random forest gives accuracy as 0.93%. All the three models includes feature 1 and 2 for the computation of precision.

Table 5
F-Measure and optimal features of skin disease data through various classifiers

Classifier	F-measure (%)	Feature set
KNN	85	1,2,5,7,9
SVM	87	1,5,7,9,11
Random Forest	94	1,2,6,7,8

From Table 5, it is experienced as Random forest produces the highest F-measure value compared to other models considered for experimental investigation. All the three models includes feature 1 and 7 for the computation of accuracy. Hence, it is considered as highly important feature for further analysis.

Table 6
Evaluation of Classification accuracy and computational time for various classifiers

Algorithm	Accuracy (%) with all features	Computation Time (Sec)	Accuracy (%) with optimal feature	Computational time in sec
KNN	78	8.13	83	4.28
SVM	82	9.55	89	5.71
Random Forest	90	5.12	97	3.83

The significant feature selection has achieved by adapting the above methods, and in the table, each attribute count is recorded in 3 different performance metrics. At the final phases of feature, counts exhibit occurrence rate. Depending upon the count of elements are ranked and given the importance that helps for better forecasting [4]. The significant variables further analyze to forecast skin disease occurrence of selecting significant variables with a different combination. Our experiment exhibits feature count increase specific variable and by applying machine learning models with their respective performance metrics to list of feature importance. Scikit-learn pipeline is used to examine the model and select features with the help of python language. The machine learning model performances are validated by considering feature pre and post [11]. Before feature selection accuracy rate is quite lower when compared to after applying feature selection technique.

5. Discussion

Machine learning algorithm KNN, SVM and Random forest attempted to find the significant attributes among data set with the different combination [11, 18]. The final accuracy result shows an increase in accuracy rate with a new feature set. These features set to provide a significant contribution to forecast skin diseases associated with climatic factors. Out of 15 elements, 12 are climatic features that are collected from IMD, the gender of patient attributes is related to demographic. Our detailed analysis states that the characteristics of climatic factors have more influence on skin diseases. According to our result, the proposed frame work achieved 97% with the minimal feature set.

Out final feature set considers Year_week, Recorded_year, Recorded_month, Air_temp, Humidity, Surface_water3, Total_vegetation, Min_air_temp, Surface_water5, Total_precipitation, Southeast_NDVI. The proposed framework helps to find out diseases outbreak. Our study exhibits by applying tail and error combination accuracy achieved good accuracy and then validated by performance metrics with accuracy, F-measure, and precision.

Table 7
List of features importance with weights

Number	Features	New_feature_name	Feature_weights
1	Year_week	F1	0.0899
2	Recorded_year	F2	0.0612
3	Recorded_month	F3	0.0394
4	Air_temp	F4	0.0174
5	Humidity	F5	0.0167
6	Surface_water3	F6	0.0152
7	Total_vegetation	F7	0.0301
8	Min_air_temp	F8	0.0069
9	Surface_water5	F9	0.0058
10	Total_precipitation	F10	0.0013
11	Max_air_temp	F11	0.0003
12	Total_precipitaion in KG	F12	0.0002
13	Northeast_NVDI	F13	0
14	Mean_duepoint	F14	-0.0002
15	Mean_humidity	F15	-0.0046

6. Conclusion

The amount of data generated from the industry has increased exponentially. One of the substantial sector increases in data generated in the health sector. All the data recorded with the pre-install sensor and quality of data is a raw format that is not ready for analysis. Some machine learning methods are used to preprocess and feature selection. The data set comprise a missing value and unformatted unsigned value as which machine unable to read this. Nearly 70% of the time spent on preprocessing and feature selection. To forecast any disease data, considering the critical factor, easy availability, and readable format help to predict the outbreak. We have applied three machine learning models with different features combination set. To validate the model, three performance matrices have used to measure the performance. Random forest model produces the highest accuracy of 97% with the new feature set when compared with other models. As our dataset is the only benchmark data, we have not compared it with other existing studies. From the experimental results, it is concluded that Year_week, Recorded_year, Recorded_month, Total_vegetation and Air_temp are considered as important features for the causes of skin disease.

Declarations

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable.

Authors' contributions

Both coauthors contributed significantly to the research and this paper, and the first author is the main contributor.

Acknowledgements

Not applicable.

References

1. Ahn G, and Sun Hur (2020) Comput Ind Eng 142. https://doi.org/10.1016/j.cie.2020.106345. “Computers & Industrial Engineering Efficient Genetic Algorithm for Feature Selection for Early Time Series Classification.”(February). Elsevier:106345
2. Amuakwa-mensah F, Marbuah G (2017) and Mwenya Mubanga. “Climate Variability and Infectious Diseases Nexus: Evidence from Sweden.” Infectious Disease Modelling 2 (2). Elsevier Ltd:203–17. https://doi.org/10.1016/j.idm.2017.03.003
3. Azimi F, Shirian S, Jangjoo S, Ai A, and Tehereh Abbasi (2017) Impact of Climate Variability on the Occurrence of Cutaneous Leishmaniasis in Khuzestan Province, Southwestern Iran Co M M Er Ci Us E Ly on. Er Al” 12:15–22. https://doi.org/10.4081/gh.2017.478
4. Cao P, Liu X, Liu H, Yang J, Zhao D, Huang M, and Osmar Zaiane (2018) Computer Methods and Programs in Biomedicine Generalize D Fuse D Group Lasso Regularized Multi-Task Feature Learning for Predicting Cognitive Outcomes in Alzheimers Disease.” Computer Methods and Programs in Biomedicine 162. Elsevier B V 19–45. https://doi.org/10.1016/j.cmpb.2018.04.028
5. Cecchi L (2018) Gennaro D’Amato, and Isabella Annesi-Maesano. “External Exposome and Allergic Respiratory and Skin Diseases.” Journal of Allergy and Clinical Immunology 141 (3). Elsevier Ltd:846–57. https://doi.org/10.1016/j.jaci.2018.01.016
6. Chalghaf B, Chemkhi J, Mayala B, Harrabi M, Benie GB, Michael E, Afif BS (2018) “Ecological Niche Modeling Predicting the Potential Distribution of Leishmania Vectors in the Mediterranean Basin:

Impact of Climate Change. ” Parasites & Vectors, pp 1–9

7. Chen J, Zhou S, Kang Z, Wen Q (2020) “Locality-Constrained Group Lasso Coding for Microvessel Image Classification” 130. Elsevier B V 132–138. https://doi.org/10.1016/j.patrec.2019.02.011
8. Combe M, Velvin CJ, Morris A, Garchitorena A, Carolan K, Sanhueza D, Roche B, Couppié P (2017) Jean-françois Guégan, and Rodolphe Elie Gozlan. “Global and Local Environmental Changes as Drivers of Buruli Ulcer Emergence,” no. October 2016. https://doi.org/10.1038/emi.2017.7
9. Dayrit JF, Lunardi Bintanjoyo M, Dennis P, Davis, Louise KA (2018) “Impact of Climate Change on Dermatological Conditions Related to Flooding: Update from the International Society of Dermatology Climate Change Committee. 1–10. https://doi.org/10.1111/ijd.13901
10. Figueroa FL (2011) “Dermo-Sifiliográficas Climate Change and the Thinning of the Ozone Layer: Implications for Dermatology Implicaciones Dermatológicas Del Cambio Climático Y de La Disminución.” Actas Dermo-Sifiliográficas (English Edition) 102 (5). Elsevier:311–15. https://doi.org/10.1016/S1578-2190(11)70813-7
11. Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, and Boris Delibasic (2016) “Artificial Intelligence in Medicine Building Interpretable Predictive Models for Pediatric Hospital Readmission Using Tree-Lasso Logistic Regression.” Artificial Intelligence In Medicine 72. Elsevier B V 12–21. https://doi.org/10.1016/j.artmed.2016.07.003
12. Kasthurirathne SN, Brian E, Dixon J, Gichoya H, Xu Y, Xia B, Mamlin, Shaun JG (2016) J BIOMEDICAL Inf 60 Elsevier Inc 145–152. https://doi.org/10.1016/j.jbi.2016.01.008. “Toward Better Public Health Reporting Using Existing off the Shelf Approaches: A Comparison of Alternative Cancer Detection Approaches Using Plaintext Medical Data and Non-Dictionary Based Feature Selection.”
13. Khalifian S, and Misha Rosenbach (2018) “Conflicts of Interest: None SC.” Journal of the American Academy of Dermatology. American Academy of Dermatology, Inc. https://doi.org/10.1016/j.jaad.2018.02.054
14. Khanadar A, Sharma B, and Shambhavi Srivastava (2016) Inform Technol Engineering” 11(1):347–352“Data Mining from Smart Card Data Using Data Clustering School
15. Kimaro EG, Jenny-ann L, Toribio, Siobhan MM (2017) Climate Change and Cattle Vector-Borne Diseases: Use of Participatory Epidemiology to Investigate Experiences in Pastoral Communities in Northern Tanzania.” Preventive Veterinary Medicine. https://doi.org/10.1016/j.prevetmed.2017.08.010. Elsevier B.V
16. Li T, Horton RM, Bader DA, Liu F, Sun Q, Patrick LK (2018) “Long-Term Projections of Temperature-Related Mortality Risks for Ischemic Stroke, Hemorrhagic Stroke, and Acute Ischemic Heart Disease under Changing Climate in Beijing, China.” Environment International 112 (7). Elsevier:1–9. https://doi.org/10.1016/j.envint.2017.12.006
17. Liang Lu, and Peng Gong (2017) Climate Change and Human Infectious Diseases: A Synthesis of Research Fi Ndings from Global and Spatio-Temporal Perspectives.” Environment International 103. The Authors 99–108. https://doi.org/10.1016/j.envint.2017.03.011
18. Links DA (2012) “Change: Progress Report, 2011.” https://doi.org/10.1039/c1pp90033a

19. Muhammad S, Shah S, Ali F, and Syed Adnan (2020) Support Vector Machines-Based Heart Disease Diagnosis Using Feature Subset, Wrapping Selection and Extraction Methods. ” Computers and Electrical Engineering 84. https://doi.org/10.1016/j.compeleceng.2020.106628. Elsevier Ltd:106628
20. Pinault L, Bushnik T, Fioletov V, Peters CE, King WD, and Michael Tjepkema (2017) “The Risk of Melanoma Associated with Ambient Summer Ultraviolet Radiation,” no.82
21. Pinault L, and Vitali Fioletov (2017) “Sun Exposure, Sun Protection and Sunburn among Canadian Adults,” no.82
22. Purse BV, Dario Masante N, Golding D, Pigott JC, Day S, Iba M, Kolb, and Laurence Jones (2017) How Will Climate Change Pathways and Mitigation Options Alter Incidence of Vector- Borne Diseases ? A Framework for Leishmaniasis in South and Meso-America. 1–22. https://doi.org/10.1038/sdata.2014.36.Funding
23. The F, Academy A, Pediatrics OF (2015) “Technical Report — Ultraviolet Radiation: A Hazard to Children and Adolescents” 127 (3). https://doi.org/10.1542/peds.2010-3502
24. Wang Y, Li T (2019) “A.” Applied Soft Computing Journal. Elsevier B V 105989. https://doi.org/10.1016/j.asoc.2019.105989

Figures

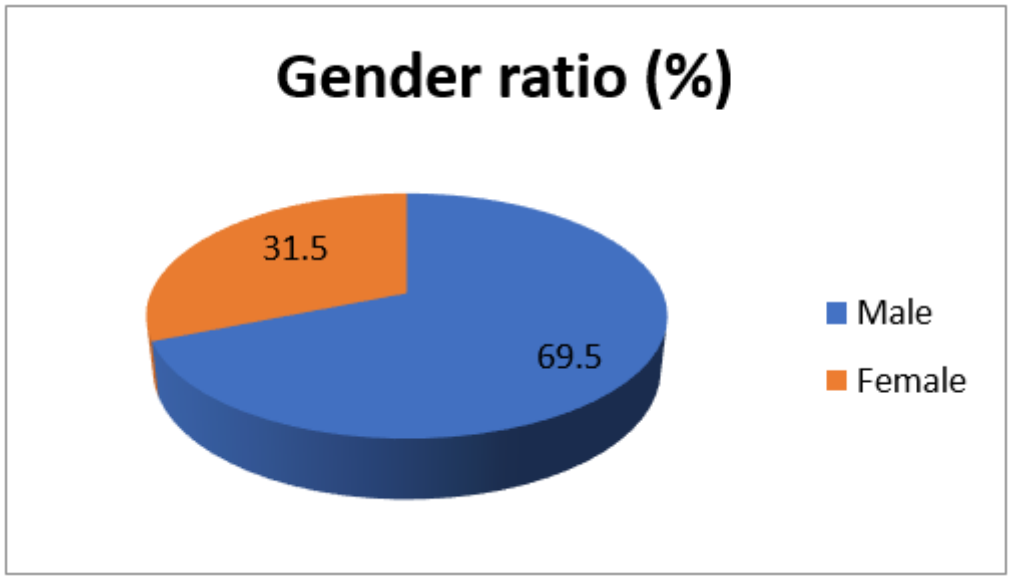

Figure 1

Skin diseases patients ratio based on Gender

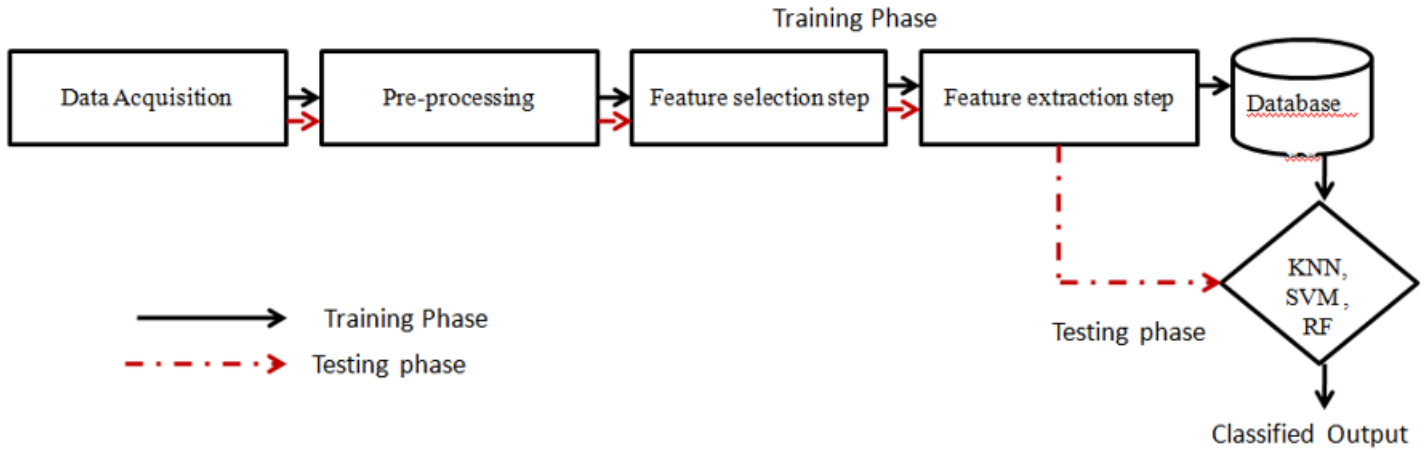

Figure 2

Architecture diagram of the Proposed Approach

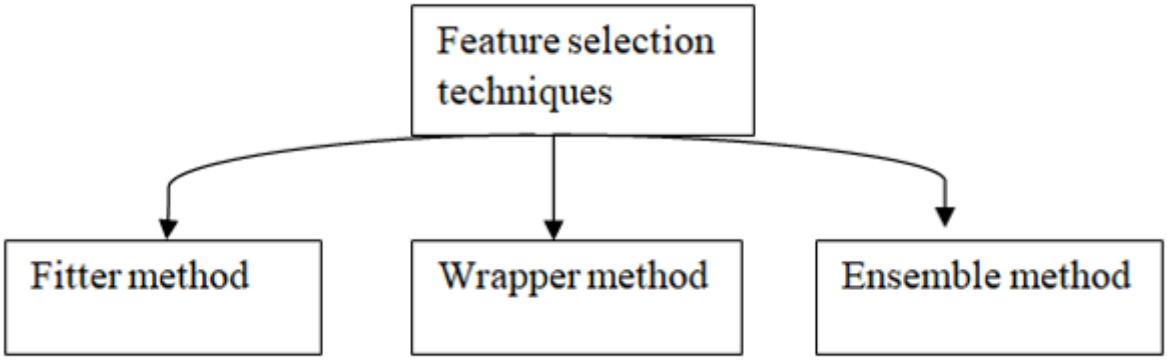

Figure 3

Types of features selection techniques