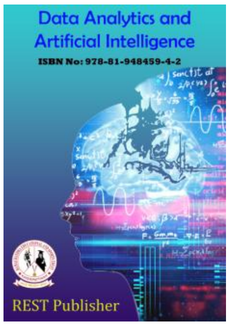
Data Analytics and Artificial Intelligence

Vol: 3(2), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/3/2/31



Skin Disease Prediction Machine Learning Model Using Ensemble Classifier with PCA

***K.Shunmuga Priya, V. Selvi**

Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India.

*Corresponding Author Email: Shunmugapriyaphdcs013@gmail.com

Abstract. *In the medical era, skin disease is considered one of the most common diseases among humans. Skin cancer is the most dangerous type, which can be curable if identified at the initial stage. The severity of skin cancer and the rapid count of affected people make it necessary to introduce an automatic detection scheme. Generally, analyzing and identifying skin disease in a short time is the most complex and challenging task. Several deep learning (DL) and machine learning (ML) are introduced to achieve this. However, the still fulfilling the skin cancer diagnosis is not accomplished completely. To achieve this, we proposed a machine learning model using an ensemble classifier with PCA to predict skin disease with maximum accuracy. The proposed Ensemble classifier is based on similar features and classifies several stages. It is executed by labeling vertebral disorder images according to these statistical features. The performance obtained by the ensemble classifier is compared with Support Vector Machine (SVM) and Resent with several evaluation metrics. The analysis shows that the accuracy attained by the proposed ensemble classifier is 97 % which is far better than the others in terms of classification and accuracy.*

Keywords: *Skin Diseases Diagnosis, Skin cancer, Classification and accuracy, machine learning model, ensemble classifier.*

1. INTRODUCTION

Skin cancer is the most crucial diseases in U.S (United States) and occurs in the skin tissues. In the human body, skin is the important part and is responsible for blocking heat, infections, sunlight, and wounds. There are three layers of skin such as hypodermis, dermis and epidermis [3]. Epidermis is the outmost layer of the skin that generates the skin tone. In the skin, second layer is the dermis which contains hair follicles, sweat glands and rough connective tissues. Finally, the lowest layer is the hypodermis which comprises fat and connective tissues [4]. Skin cancer is the most threatening disease to the skin and is the most common cancer globally. Among the other type of cancers, a minimum of 40% comprises skin cancer, and its growth is predicted to about 9,500 people with skin cancer every day. The interventional treatments such as surgery, chemotherapy, and radiotherapy are used to diagnose and classify the cancers. This study defines the application of modern information systems like image processing mechanisms in successfully processing cancer diagnosis and classification. According to various skin cancers, Melanoma is the 19th most prevalent cancer among the humans. Several new cases around 300,000 were identified in 2018. Melanoma cancer is the 4th most frequent cancer globally, according to the Cancer Cell Organization, with 15000 cases. According to this organization, Melanoma is also the nineth most common cause of cancer death in 2019. Because of the prevalence of many skin lesions, including Carcinoma and Melanoma, a skin cancer analysis is a complicated process. Several noninvasive approaches to minimize unnecessary biopsy for melanoma diagnosis have been proposed. Most approaches have three primary components: segmentation, feature extraction, and classification. In this scenario, several studies were conducted. Suggested a deep learning-based image feature extraction technique for melanoma diagnosis. The scientists used convolutional neural networks (CNN) for feature extraction with transfer learning and a variety of classifiers for final classification, such as AdaBoost, random forest (RF) and k-nearest neighbor (KNN). The ISIC dataset was used to test this strategy, and the results revealed each classifier's accuracy levels. This technique is good but took longer to complete due to the complex configuration. In the last few years, AI models, machine learning, and deep learning techniques have rapidly progressed in the medical field. To enhance skin disease prediction, various ensemble-based [5] ML and artificial neural network-based techniques are extensively employed. Many skin disorders require computer vision in terms of physical symptoms. The use of computer vision aids in the more precise diagnosis of skin disorders. This study uses a combine model of ML and DL to diagnose skin disorders. Three types of machine learning models were utilized for classification: Decision Tree, Adaboost Ensemble Classifier and SVM. VGG16, Alexnet, Resnet50, and Googlenet were used to choose features from four pre-trained deep neural networks. Subsequently, tests were conducted to determine the most effective predicting models.

Machine Learning Models: We investigate the supervised machine learning method's overall performances such as Gaussian Naive Bayes (GNB), Decision Tree Classifier (CART), K-nearest Neighbors classifier (KNN), Linear Discriminant Analysis (LDA) and Logistic Regression (LR). Among these, two algorithms were chosen for consideration. The first approach (LR) is well adapted for binary problems categorization and it is widely biomedical domains. Using the parametric system such as LR approach, the class membership probability is identified among the dataset with two categories. Secondly LDA is widely used for classifying the supervised patterns and generates linear decision boundaries among the classes. The k-nearest Neighbors classifier (KNN) is based on proximity in the following method. This method employs a metric that allocates a distinct category to each item based on its similarity to other data points. Its varieties make it a good choice for studying classification difficulties. Decision tree classifiers which include Classification and Regression Trees (CART) decision-tree technique, are increasingly being employed in biomedical applications. The Gaussian Naive Bayes (GNB) algorithms is another most efficient classifiers based on Bayes' theorem. We use an ensemble learning mechanism for enhancing the accuracy along with using single ML algorithm for image categorization. The soft-voting decision is used to assess the ML algorithm groups and from each group average predictions are identified. E1, E2, and E3 are three separate ensembles or sets of ML algorithms that we define. E1 is a classifier with highest diversity and average predictions from various machine learning techniques (GNB, LR and KNN).E2 averages the predictions from all machine learning techniques which are employed in this study (GNB, KNN, CART, LR and LDA). Finally, E3 compiles the best-performing predictions among the CART, LR and LDA machine learning algorithms.

Problem Statement: As disease variants increase, the day-by-day need for improvements in skin diseases. Lack of accuracy in skin cancer classification and accuracy. Lack of accuracy in Melanoma Skin Cancer classification and accuracy. Existing methods of processing time are high need optimization on execution time. Organization of the paper is as follows: Introduction part is described in the section 1. In section2 related works are discussed. Our proposed mechanism workflow and performance evaluation is discussed in section 3 and finally section 4 contains the conclusion part.

2. RELATED WORK

Xinrong Lu et al. [1] introduced an improved Xception Net mechanism for automatic skin cancer diagnosis from dermos copy samples (images). The improved Xception Net is enhanced with a swish activation function and depth wise separable convolutions. The combined architecture enhances the classification accuracy of the existing original Xception. The proposed Xception deep network architecture is very efficient in enhancing the classification accuracy in CNN (Convolutional Neural Networks). MuhammadArif et al. [2] proposed automatic diagnosing of Nonmelanoma Skin Cancer using a Deep Convolutional Neural Network mechanism. This work aims to detect skin cancer earlier through an intuitive skin lesion classification process by applying a pretrained deep learning network along with transferring learning concepts. In Melanoma diagnosing during its premature stages, the proposed system further extended with g k-means and modified k-means clustering for image segmentation. Next, the feature extraction technology is implemented for extracting the d first-order statistics and Gray Level Co-occurrence Matrix. The Harris Hawks optimization (HHO) method is used for selecting the features from the image. SoleneBechelli et al. [3] proposed the classification of skin cancer from Dermoscopic Images using a combination of Machine Learning and Deep Learning Algorithms. In this work, a comparison experiment is conducted between the ML and DL models. In ML models, linear systems such as linear regression and linear discriminant analysis are very efficient in achieving the highest classification accuracy. At the same time, DL models are very effective in gaining the fine-grained changeability of the dermoscopic images. Finally, from the observation, effective overall performance is achieved without re-training and ResNet50 models. Samir et al. [4] discussed the integration of Machine Learning and Deep learning in predicting skin diseases. In this work, the author introduced an electronic skin disease diagnosis system by the combination of ML and DL models. The combined proposed architecture achieves the feature extraction with four deep learning models and three prominent machine learning classifiers. Finally, the analysis shows the combination of SVM Classifier with deep network model Resnet50 achieves the highest accuracy in the prediction results.

A. Murugan et al. [5] proposed skin cancer diagnosis using machine learning approaches. In this work, skin cancer is detected from the given input skin images. The feature extraction is done by employing GLCM, Moment Invariants, and GLRLM features. Next, the obtained extractions are further classified through the combined mechanism of SVM, Random Forest, Combined SVM+ RF classifier, and Probabilistic Neural Networks. The Combined SVM+ RF classifier has achieved more classification accuracy than the others. Xu et al. [6] proposed a practical approach for detecting Melanoma earlier. They employed a strategy that included feature extraction, classification, image segmentation, and image noise reduction sequentially. The segmentation approach employed is the improved convolutional neural network (CNN) with satin bowerbird optimization (SBO). SBO extracted only the most significant features from the segmented images. Finally, a Support Vector Machine (SVM) was employed based on the acquired features to classify the photos. The proposed method was tested on the American Cancer Society database, and the findings revealed that it was effective. However, the proposed method produced effective results, and the deep learning method learning with the SBO algorithm resulted in a complicated approach. Razmjooy et al. [7] proposed optimized ANN as a diagnostic method for identifying skin cancer. Initially, the smoothing and edge detection were used to eliminate superfluous scales. The approach then segmented the desired area. Mathematical morphology was used to remove the extra information. The model uses an optimized MLP neural Network (ANN) from the World Cup Optimization algorithm to achieve enhanced outcomes. The authors employed the improved ANN to diagnose skin cancer in that study. The Australian Cancer Database (ACD) was used in simulations, and the findings showed that the suggested strategy improved the method's performance. The ANN approach was employed in the /e method, regarded as an old and less accurate method in recent years. Skin disorders are considered the most prevalent than other illnesses. Skin disorders can be caused by a fungal infection, germs, viruses, or allergies, among other things. A skin condition can cause abnormalities in the skin surface or skin tone. Skin infections are long-term, contagious, and can progress to skin cancer in certain situations. Consequently, skin illnesses must always be detected early to avoid their progression and spread. It takes longer to diagnose and treat a skin condition and damages the patient physically and financially. Several researchers studied photos of skin ailments in order to develop a technique for identifying various skin disorders. Various skin disease detection approaches that are mentioned in table 1 were examined in this section.

TABLE 1.

Problem Statement	**Method Used**	**Accuracy**
Skin Diseases Detection [8]	Image processing is taken for diagnosing the disease area, and classification is performed using multi-class SVM	Three kinds of skin diagnoses are done with 100% accuracy.
Skin Diseases Detection [9]	Several machine learning algorithms are employed, such as Logistic Regression, Random Forest, Naive Bayes, Kernel SVM and Convolution Neural Network.	The best result with 96% accuracy is achieved with CNN
Melanoma Detection [10]	The multi-direction three-dimensional color-texture features selection and multilayer back propagation neural network classifier is used for feature extraction.	Enhanced accuracy results
Skin Cancer Classification [11]	Using ResNet50	Accuracy almost 92%
Skin Diseases Classification [12]	Using Deep Neural Networks [LSTM and Mobile NET V2]	Accuracy almost 85.34%
Melanoma Skin Cancer Detection [13]	Using SVM	Accuracy 96.9%
Skin cancer Detection [14]	Deep Learning Network	Good results
Skin cancer Detection [15]	Feature extraction is done with the dermoscopic images and classification by employing a convolution neural network.	Accuracy 89.5%

3. PROPOSED METHODOLOGY

IDS Ensemble classifier & Working Principle Ensemble classifiers is a classifiers collection aimed at analyzing target processes. Ensemble classifier’s estimations are gathered and intended for new sample categorization. It is used for enhancing the machine learning functions through joining multiple models. It improves the classification performance of a particular classification group on a domain. The ensemble classifier behaves depending on the similarity attributes and categorizes various phases. In this work, we proposed an ensemble of classifiers incorporated for labeling vertebral disorder images as normal or abnormal according to their statistical attributes. In conversely to individual classifier, ensemble classifier merges classifier clusters to improve the classifier's performance.

Why executing Ensemble mechanisms;

- Similar training performances starting with classifiers can present varied generalization performances.
- Combining outputs of multiple classifiers results in minimizing danger of selecting poor performing classifiers.
- A single classifier is inefficient when dealing with a huge amount of data.
- When dealing with a small amount of data, ensemble systems can be employed as re sampling strategies.

Ensemble the classifier's prediction workflow;

- The final output is the classifier output with the best performance
- Each classifier's outputs are combined to determine the conclusion
- The final class label is chosen using precise preset rules.

- ➢ The most effective rule permutations are Borda count, knowledge space common behaviors and weighted majority voting
- ➢ The total classifier in the Ensemble is calculated by dividing the speed and accuracy of the classifier.
- ➢ Larger ensembles and over trained categorization necessitate longer prediction training time.

The ensemble learning includes multiple models to improve the performance that contains several methods like; Random subspace: Before running the training algorithms, the subgroup of attributes is picked at random. And then, the outcome of the method is selected by the majority vote. Bagging (Bootstrap Aggregation): A set of models is constructed based on random data. Estimation of final prediction is done by summing of models with averaging. Boosting: It functions based on the voting or averaging of several models. These models are evaluated and created depending on their performance efficiency. We enforce the majority voting principle via linear discriminant with the subspace ensemble in our suggested design. Subspace discriminant: Ensemble subset learning methods perform a key role in low-dimension data. It utilizes a LDA method to handle low-dimension data. Currently, multiple works are grown with the following methods: resampling, weighting, and different sub-spaces. The entire studies aim in ensemble learning's performance enhancements in the aspect of classification and efficiency [16, 17, 18, and 19]. In the random subspace approach (RSM) [20], a random attribute sample is employed to minimize the error rate in developing a model. The primary disadvantage of RSM is incorrect discrimination probability from random sampling. To resolve this issue, the majority voting-VM technique is used. An ensemble is used, and every classifier offers a new or unknown sample. The whole classifier generates all new or unknown samples gathered, and majority vote casting is processed for obtaining the final classification output. In this work, we implemented discriminant learning which implies the subsets along with categorizes the fibrosis levels and the normal case accordingly. The obtained segments are essential for the learning algorithms.

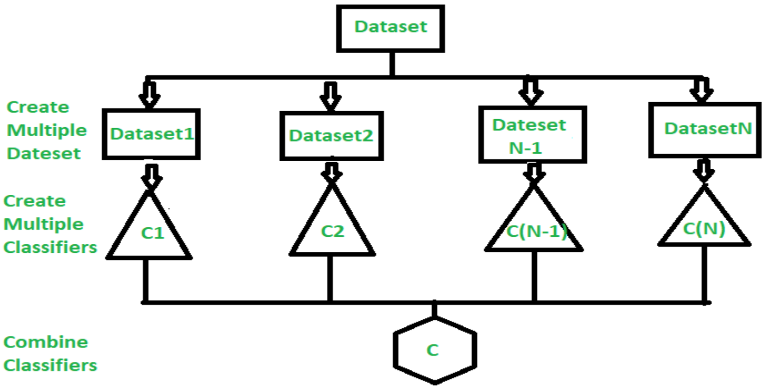


FIGURE 1. Proposed architecture

Figure 1 depicts the distribution of a dataset into several datasets and generates various classifiers correspondingly. Multiple subspaces from the initial dataset are constructed with equal records. These subspaces are the base for developing a base model. Every model is parallelly assessed for generating the training set and is independent to one another. Every model depicts its attribute connections in the form of tree, which brings out normal or abnormal datasets. The output prediction contains multiple evaluation measures that are addressed in the research investigation section. The predictions from all techniques are gathered and assessed to obtain the final predictions. Model which attain the maximum accuracy would considered as the final prediction.

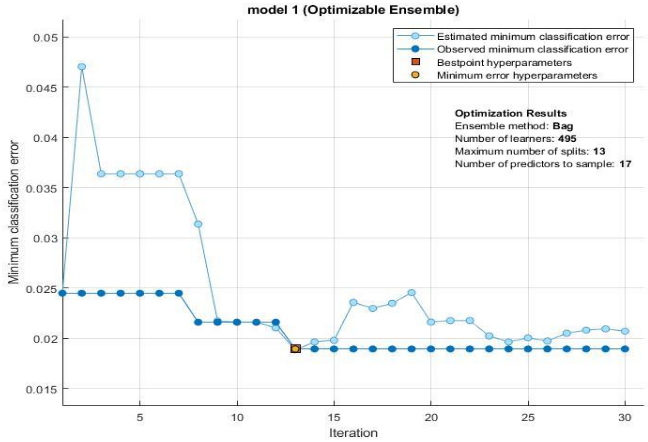


Figure 2 optimizable ensemble

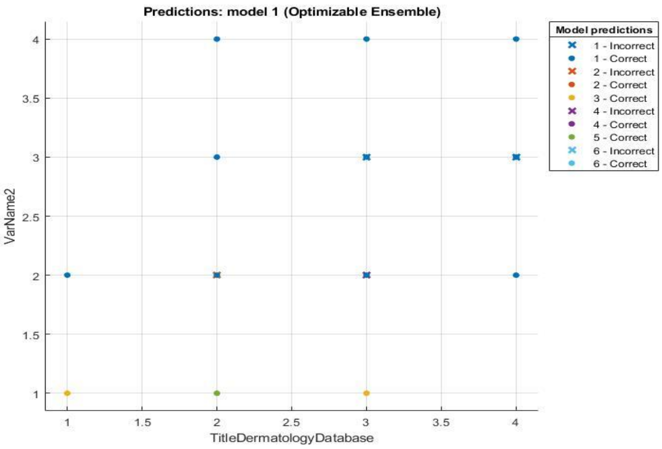


Figure 3 model predictions

Above figure 2 & 3 illustrates the proposed optimized Ensemble with multiple datasets and its prediction model. In figure 2, the x-axis represents the various iterations, and y-axis represents the minimum classification error obtained at each iteration. The graph plots factors like minimum hyper para meter, observed minimum classification error, best point hyper para meter and estimated classification error. Figure 3 shows the correct and incorrect predictions from the respective models. A detailed summary of obtained accuracy under various parameters such as Precision, TP rate, FP rate, ROC area, PRC area, F-Measure and Recall are described the below table1. Table 2 shows the summary of Stratified cross-validation with various parameters and its obtained values.

TABLE 2. Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	**MCC**	ROC Area	PRC Area
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
0.902	0.023	0.887	0.902	0.894	0.873	0.990	0.948
0.972	0.003	0.986	0.972	0.979	0.974	0.994	0.991
0.857	0.016	0.894	0.857	0.875	0.856	0.989	0.938
1.000	0.003	0.981	1.000	0.990	0.989	1.000	1.000
1.000	0.003	0.952	1.000	0.976	0.974	0.999	0.987
0.959	0.007	0.959	0.959	0.959	0.952	0.996	0.980

Time is taken to build the model: 0.21 seconds.

TABLE 3. Stratified cross-validation and Summary

Correctly Classified Instances	351	97.9016 %
Incorrectly Classified Instances	15	4.0984 %
Kappa statistic		0.9487
Mean absolute error		0.0165
Root mean squared error		0.1062
Relative absolute error		6.1909 %
Root relative squared error		29.1055 %
Total Number of Instances		366

Confusion matrix: Observations: Several evaluation criteria are used to assess the performance of our proposed approach. Here, we present the vectors that were used to analyze the data and results. **Confusion Matrix:** A confusion matrix is a real-time report on the classification problem's prediction results. Correct and wrong forecasts are the two categories of predictions. Each class's correct and incorrect predictions are determined by split down using count values. It not only defines the classifier that produced an error but also specifies the error type that was created occurred.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

FIGURE-4. Class Prediction

Definition of the Terms:

- Positive (P): Observation is positive
- Negative (N): Observation is not positive
- True Positive (TP): Observation is positive and is predicted to be positive.
- False Negative (FN): Observation is positive but is predicted negative.
- True Negative (TN): Observation is negative and is predicted to be negative.
- False Positive (FP): Observation is negative but is predicted positive.

Classification Rate/Accuracy: Classification Rate or Accuracy is defined by:
Accuracy =
Recall: The Recall is the ratio of correctly categorized positive cases to the total positive cases. The recognition of correctly identified instances is stated by the high recall values (a limited number of FN). It is described as below:

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

Precision: The result obtained by dividing the total accurately categorized positive cases by the total number of anticipated positive examples is referred to as precision. The value that is marked as positive with high precision is, in fact, positive (a small number of FP). It can be defined as below;

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

F-measure: Precision and Recall are used to calculate F-Measure. As it handles extreme values frequently, it employs Harmonic Mean rather than Arithmetic Mean to calculate F-measure. F-Measure is constantly lower than the Recall or Precision when the three are compared. The F-measure is defined as below:

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

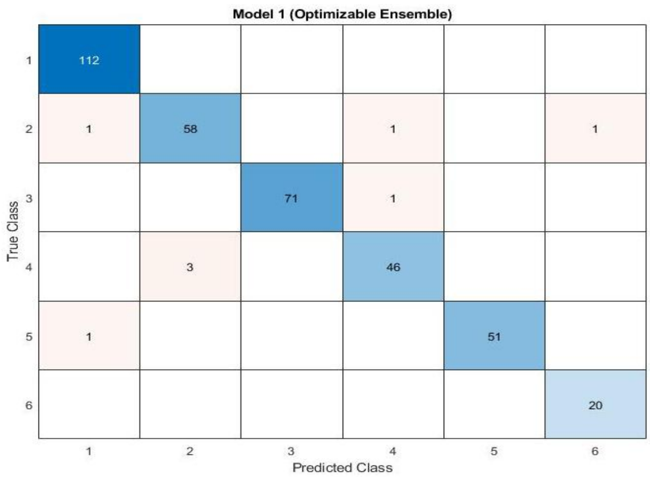


FIGURE 4.

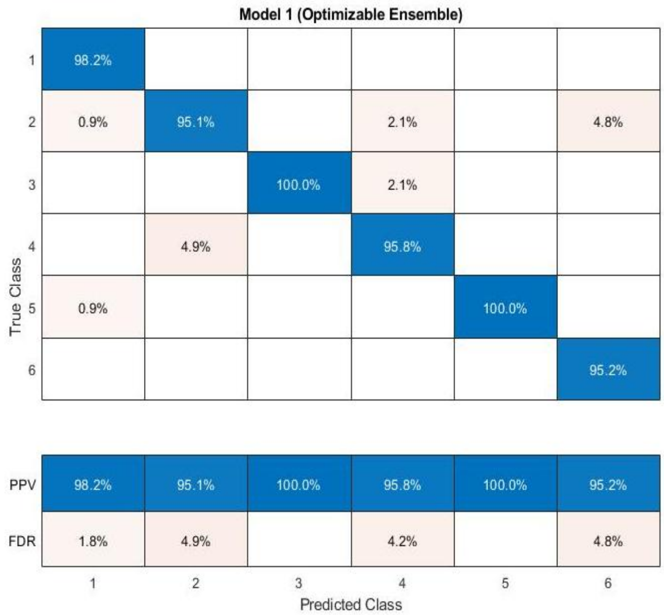


FIGURE 5.

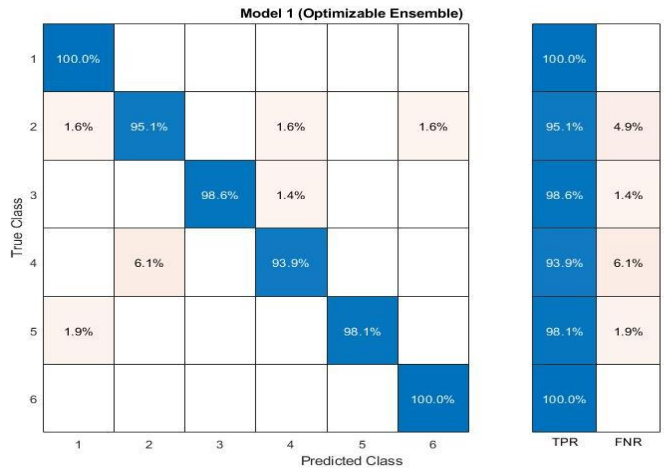


FIGURE 6.

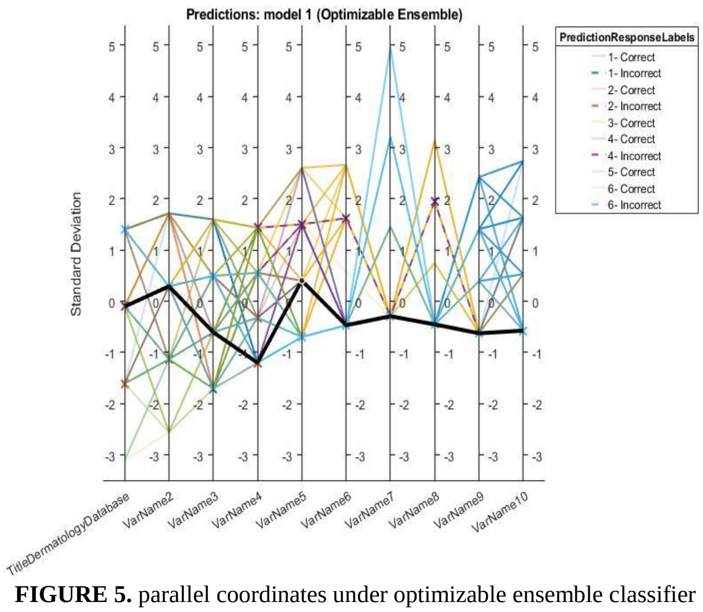


FIGURE 5. parallel coordinates under optimizable ensemble classifier

TABLE 4.

S.No	Classification algorithm	Accuracy performance
1	SVM[14]	96%
2	Ensemble proposed	97 %
3	Resnet[12]	92%

The above figures 4, 5, 6& 7 illustrate the performance of the proposed optimized Ensemble under a confusion matrix with the models. Figure 4, 5, 6 shows the obtained true class and predictive class from the models along with the PPV and FDR. In the confusion matrix, the x-axis denotes the predicted classes, and the y-axis denotes the obtained true classes from the models. Figure 7 shows the obtained standard deviation with multiple datasets and correct and incorrect predictions. The x-axis denotes the various datasets used, and the y-axis denotes the standard deviation of each respective dataset. Table 3 shows the performance analysis of the prediction model between the SVM and Resnet with the proposed Ensemble. The overall accuracy is defined through the evaluation metrics which includes classification accuracy, Recall, precision, and F-measure. The proposed Ensemble achieves 97% accuracy, which is higher than the SVM and Resnet models.

4. CONCLUSION

This paper proposes an ensemble classifier with PCA for detecting skin cancer from the images. Automatic skin disease detection is very favorable for the patients affected with skin cancer. It helps the physician with more accuracy and can cure the diseases at the beginning stage. In this work, we analyzed the existing challenges and complexities in achieving the accurate, correct prediction from the images. Initially, the proposed ensemble classifier labeled the vertebral disorder image, effectively extracting the features through similarity and classification in several stages. The proposed approach applies multiple classifiers for handling huge data and employs resembling strategies to deal with the small amount of data. To analyze the performance of the proposed ensemble classifier, a comparison work is carried out between the SVM and Resent models. The evaluation metrics for comparison are classification accuracy, Recall, precision, and F-measure. The performance attained by the proposed ensemble classifier is more effective, especially in accuracy, than the others.

REFERENCES

[1]. Xinrong Lu and Y. A. FiroozehAbolhasaniZadeh, "Deep Learning-Based Classification for Melanoma Detection Using XceptionNet", Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 2196096, 10 pages https://doi.org/10.1155/2022/2196096

[2]. MuhammadArif ,FelixM. Philip,F.Ajesh,DianaIzdrui , MariaDanielaCraciun ,and OanaGeman,"Automated Detection of Nonmelanoma Skin Cancer Based on Deep Convolutional Neural Network", Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 6952304, 15 pages https://doi.org/10.1155/2022/6952304

[3]. SoleneBechelli and Jerome Delhommelle, "Machine Learning and Deep Learning Algorithms for Skin Cancer Classification from Dermoscopic Images", Bioengineering 2022, 9, 97. https://doi.org/10.3390/bioengineering9030097 https://www.mdpi.com/journal/bioengineering

[4]. Samir Kumar Bandyopadhyay , Payal Bose , AmiyaBhaumik and SandeepPoddar, "Machine Learning and Deep Learning Integration for Skin Diseases Prediction",International Journal of Engineering Trends and Technology Volume 70 Issue 2, 11-18, February, 2022 ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I2P202

[5]. Murugan a , Dr. S. Anu H Nair b , Dr. A. Angelin Peace Preethi c and Dr. K. P. Sanal Kumar d, "Diagnosis of skin cancer using machine learning techniques", Elsevier, Microprocessors and Microsystems 81 (2021) 103727, https://doi.org/10.1016/j.micpro.2020.103727

[6]. Z. Xu, F. R. Sheykhahmad, N. Ghadimi, and N. Razmjooy, "Computer-aided diagnosis of skin cancer based on soft computing techniques," Open Medicine, vol. 15, no. 1, pp. 860–871, 2020

[7]. N. Razmjooy, F. R. Sheykhahmad, and N. Ghadimi, "A hybrid neural network – world Cup optimization algorithm for melanoma detection," Open Medicine, vol. 13, pp. 9–16, 2018.

[8]. AlkolifiAlenezi, N. S. A Method of Skin Disease Detection Using Image Processing and Machine Learning. Procedia Computer Science. 163 (2019) 85–92.

[9]. Bhadula, S., Sharma, S., Juyal, P., &Kulshrestha, C. Machine Learning Algorithms based Skin Disease Detection. International Journal of Innovative Technology and Exploring Engineering. 9(2) (2019) 4044–4049.

[10]. Warsi, F., Khanam, R., Kamya, S., &Suárez-Araujo, C. P. An efficient 3D colour-texture feature and neural network technique for melanoma detection. Informatics in Medicine Unlocked, 17(November 2018) (2019) 100176.

[11]. Gouda, N., &Amudha, J. Skin Cancer Classification using ResNet. 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020. (2020) 536–541.

[12]. Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors. 21(8) (2021) 2852.

[13]. Banasode, P., Patil, M., &Ammanagi, N. A Melanoma Skin Cancer Detection Using Machine Learning Technique: Support Vector Machine. IOP Conference Series: Materials Science and Engineering. 1065(1) (2021) 0–5.

[14]. Kadampur, M. A., & Al Riyaee, S. X. Skin cancer detection: Applying a deep learning-based model-driven architecture in the cloud for classifying dermal cell images. Informatics in Medicine Unlocked, 18(December 2019) (2021) 100282.

[15]. Hasan, M., Barman, S. Das, Islam, S., & Reza, A. W. Skin cancer detection using convolutional neural network. ACM International Conference Proceeding Series: March 2020 (2019) 254–258.

[16]. DR.(Mrs) Selvi V, Mrs.K.Shunmugapriya"Prediction Of Skin Diseases Using Bigdata Techniques" in International Journal of Creative Research Thoughts (IJCRT), Volume-9, Issue-06, June 2021.

[17].DR.(Mrs) Selvi V, Mrs.K.Shunmugapriya "An Effective Pre-Processing Method For Classification Of Skin Disease Using Machine Learning Techniques " DogoRangsang Research Journal UGC Care Group I Journal ISSN : 2347-7180 Vol-12 Issue-07 No. 02 July 2022.