Skin Cancer Image Detection and Classification by CNN based Ensemble Learning

Sarah Ali Alshawi, Ghazwan Fouad Kadhim AI Musawi
Imam Ja’far AI-Sadiq University, Mysan

***Abstract*—Melanoma is accounted as a rare skin cancer responsible for a huge mortality rate. However, various imaging tests can be used to detect the metastatic spread of disease with a primary diagnosis or on clinical suspicion. Focus on melanoma detection, irrespective of its unusual occurrence, is that it is often misdiagnosed for other skin malignancies leading to medical negligence. Sometimes melanoma is detected only when the metastasis has entered the bloodstream or lymph nodes. So, effective computational strategies for early detection of melanoma are essential. There are four principal types of skin melanoma with two sub types: Superficial spreading, nodular, lentigo, lentigo maligna, Acral lentiginous, and Subungual melanoma. Amelanotic melanoma, one particular type of melanoma, exists in all kinds of skin tones. Classifications of melanoma with its classes are focused on in this research. The ensemble classifier models, namely Adaboost, random forest, voted ensemble, voted CNN, Boosted SVM, and Boosted GMM, have been used in melanoma classification to address misclassification errors, overfitting issues, and improve accuracy. The results of the ensemble classifier achieve high classification accuracy. However, imbalanced classification is found in all six classes of melanoma. Transfer learning and ensembled transfer learning approaches are implemented to reduce the imbalanced classification issues, and performances are analyzed. Four ML/DL models, six ensembled models, four transfer learning models, and five ensembled transfer learning models are used in this investigation. Implementation of all the 19 classifiers is analyzed using standard performance metrics such as Accuracy, Precision, recall, Mathew’s correlation coefficient, Jaccard Index, F1 measure, and Cohen’s Kappa.**

Keywords—Medical images; skin cancer; machine learning; deep learning; ensemble learning; accuracy

I. Introduction

Humans are largely perceived through their skin. The dermis, epidermis, and subcutaneous layer make up the skin. The skin has the ability to sense its environment and to protect the body’s internal organs and tissues from environmental hazards like bacteria, toxins, and UV radiation. [1]. A variety of internal and external factors can have an impact on the skin. Experimental skin damage, embryogenic infections, chemical exposure, a person’s immunological function, and genetic abnormalities are all factors that influence the emergence of skin diseases. Skin problems have a huge impact on a person’s life and health. People will eventually try home treatments to address their skin conditions. These procedures may have harmful implications if they are not suitable for that skin disease. As skin problems are easily spread from person to person, they must be treated first. Presumptions about a patient’s health are typically based on the doctor’s experience and intuition. It could be dangerous to one’s health if the decision is made incorrectly or delayed [2].

As a consequence, developing efficient strategies for diagnosing and treating skin problems becomes vital and critical. Technological advancement has enabled the design and implementation of a skin monitoring formative days foundational identification of skin issues. There are numerous advancements accessible for pattern-and image- based identification of several skin conditions.

Deep learning is among the disciplines which can help with the practical and exact identification of a variety of skin problems. Image categorization and deep learning can be used to diagnose diseases [3]. Image classification is a basic problem that requires the creation of multiple objective classifications and the development of a training model to acknowledge each subtype. Deep learning-based technologies could be useful for swiftly recognizing clinical information and providing results. Information treatment is essential due to the complexity of skin diseases, the scarcity and misuse of qualified medical professionals, and the urgency associated with an accurate diagnosis. Improvements in photonics and laser-based health care system technology have allowed for much quicker and more accurate diagnosis of skin problems. Even with advances, the price of diagnostic procedures is still prohibitive. Deep learning systems efficiently classify images and data [4,5]. The reliable recognition of anomalies and classification of diseases utilising magnetic MRI, X-ray, PET, CT, and signalling data including EEG, EMG, and ECG has been requested in health diagnostics. Better health care could be provided to patients if diseases were classified more precisely. By automatically identifying data input features, DL approaches can address critical challenges and are adaptable to shifts in the computational complexity [5]. It is expected that learning techniques would be able to discover and start exploring the features in the discovered data patterns with even basic computer modelling, resulting in substantial efficiency gains. As the categorization of skin diseases relies on an image of the affected region, this prompted the researchers to investigate the possibility of using a DL model for classifications. Invasive illness evaluations would be easier and less expensive for doctors and patients to perform with this tool.

II. Related Work

Chaurasia and Pal [1] demonstrated six distinct order frame- works and a multi-model ensemble strategy for predicting skin disorder.

The findings show that differential expression assessment is essential for reducing the dimensionality of data and selecting effective data, thereby increasing the accuracy of prediction and substantially lowering the computing effort. At such a point, the multi-model ensemble methodology employs the predictions of numerous distinct classification models as input. By using the principal organize predictions as highlights, the classification approach minimizes the generation error and obtains more data than if it were trained in isolation. In addition, by utilising classification methods, the intricate relationships between classifiers are discovered, thereby enabling the order strategy in order to make more accurate predictions.

Loganathan, et al. [2] suggested a new DCNN for classifying malignant melanoma (skin cancer). The recommended method comprises pre-processing, enhanced fuzzy clustering for melanoma detection, and enhanced deep convolutional neural networks (E-DCNN) for categorization of dermoscopy images. Enhanced fuzzy clustering is a technique that incorporates modified region grow image segmentation and fuzzy K-means clustering to produce a more precise classification performance than other methodologies suggested by researchers.

Allugunti [3] developed, built, and tested a Convolutional Neural Network (CNN) framework for melanoma detection using a publicly available dataset. The overall accuracy of 88.83% demonstrates the superiority of the proposed method, which would be a two-stage learning platform. This is not unique to DT, RF, or GBT or any other classification algorithm. The proposed technique is based on CNN and can be seen as a powerful means of multiclass categorization.

Kotian & Deepa [4] identify and categorize various diseases using input images. The MATLAB environment serves as the foundation for this project. The photos come from various online sources like Dermnet and DermWeb. The first step is to preprocess the sample images of the four skin diseases. As a second step, a geometric transformation is applied to the vertically-oriented portion of the image. Relying upon it, three types of skin diseases' features are then extracted, and their correlated variables of feature texture and pixels of lesion regions are accumulated via image segmentation.

Verma, et al. [5] proposed a novel method that employs five distinct data mining methods and then develops an ensemble method that integrates all five methods into a single unit. Using descriptive Dermatology data, the researcher examined various data mining techniques to categorize the skin infection, and then apply an ensemble ML technique.

Rea [7] a survey of people with skin diseases was done in Lambeth, London. A stratified specimen of 2180 adults was sent a mail-in questionnaire asking about skin diseases. A subsample of 614 people was questioned at home and their exposed skin was looked at. There were 92 conditions that were found. These were put into 13 groups based on how severe they were for the patient. 22–5% of people were thought to have skin diseases that needed medical care. With a prevalence of 6–1%, eczema was the most widely accepted essential factor. Certain types of skin diseases had different rates of occurrence based on age, gender, and social class. Only 21% of people with a skin disease that should have been treated by a doctor said they had gone to their healthcare professional in the previous six months for a skin problem. Medical treatment and self-medication are considered in relation to the existence of skin infection and certain other factors.

Dildar [11] provides a thorough comprehensive study of DL techniques for skin cancer detection. Research papers from reputable journals on the topic of skin cancer diagnosis were reviewed. To aid comprehension, study results are presented in the form of tools, tables, graphs, methodologies, and frameworks.

A. Gap Analysis

Gap analysis for skin cancer classification using deep learning can be conducted by comparing the current state of research in this field with the desired future state. Some potential gaps that could be identified include:

Lack of standardized datasets: There is a need for standardized datasets that are representative of diverse populations and cover different types and stages of skin cancer [1].

Limited generalizability: Many deep learning models developed for skin cancer classification have been evaluated on small datasets or datasets from a single institution. There is a need for models that can be trained on larger and more diverse datasets and can generalize well to different populations [4, 5, 6].

Limited availability of models in clinical practice: Although deep learning models have shown excellent performance in skin cancer classification, they are not yet widely used in clinical practice. There is a need to develop models that are easy to use, reliable, and can be integrated into clinical workflows [7, 8].

Limited attention to ethical considerations: There is a need for greater attention to ethical considerations in the development and deployment of deep learning models for skin cancer classification, including issues related to bias, privacy, and informed consent [9, 10].

III. Proposed Approach

A. Melanoma Detection using Boosted SVM

Ensembles with an infinite hypothesis are constructed using an infinite ensemble framework. This framework learns all the possible weight combinations for all the possible hypotheses. All the hypotheses are embedded in the kernel of the SVM model. The base classifier is trained by fixing the initial weights, and the error due to misclassification is calculated using the weighted method. Now similar to the boosting algorithm weight of each classifier is adjusted, and the ensemble classifier is computed using the weighted component sum classifier as, where is the weight and is the base classifier. Ensemble majority voting model is shown in Fig. 1.

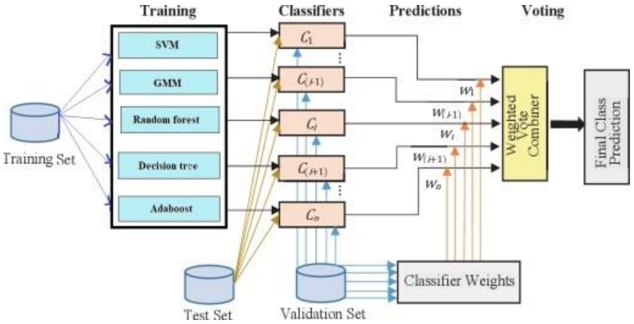

Fig. 1. Ensemble majority voting model.

SVM was developed from the theory of Structural Risk Minimization. In SVM, two essential parameters are to focus on, namely Gaussian width and the regularization parameters [2]. Boosted SVM is applied over the training dataset, and the weights are updated until convergence if the error rate exceeds 0.5. The weighted normal strategy determines the error component (e$_m$). Only the limited weights of misclassified perceptions are taken as the numerator, separated by the total limited weights [3]. Each SVM is prepared to depend upon the dataset because testing and validation errors decrease as data quantity increases. In both the sections mentioned above, the weight update procedure is addressed by taking into account the distance from each group centroid to each misclassified observation. The modified weights for each misclassified observation are assigned concerning the distance from the centroid of the clusters.

$$G_m = \sum_{j=1}^{m} \alpha_j * S_j \tag{1}$$

B. Melanoma Detection using Boosted GMM

Using GMM for Ensemble comes under the standard category of clustering ensemble. It is a model-based Ensemble. A model-based Ensemble assumes that the model's clusters will help optimize the relevance between the data and the underlying model. GMM is a probabilistic model [11] frequently used for density estimation, regression, and classification problems. GMM as a classifier is constructed as discussed in Section III. Then the Gaussian mixture components of each object class are compared with the corresponding class object probability. A threshold is fixed to recognize the object with maximum similarity for the specific object class. To generalize different object components and increase the similarity of objects in each class Adaboost algorithm is applied to create a model-based ensemble GMM framework. Here each component in GMM is considered a weak classifier with low accuracy and high redundancy. Adaboost algorithm combines all the weak classifiers into a robust classifier with effective multiclass object recognition.

C. Melanoma Detection using Ensemble CNN

A sequential voting ensemble could be created using convolution neural networks. Theory and implementation of voted ensembles are similar to Section III.C, with one significant difference. Here instead of using different classifier models, we use three models of CNN, and the highest voted prediction is taken as output. Here for each model, three convolution layers, three batch normalization layers are used. Finally, the dense layer with 256 units and softmax layer are used as the output layer. The Ensemble CNN model is shown in Fig. 2.

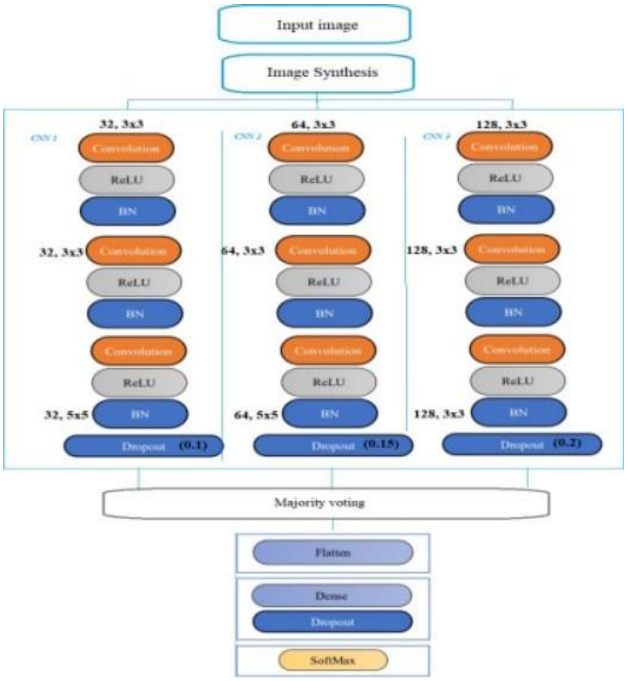

Fig. 2. Proposed CNN model.

Here for the CNN ensemble, three individual models are developed with many ReLu filters. Three models have 32, 64, and 128 filters consecutively. Here the optimizer used is Adam, and the output activation function is softmax. The batch size is 20, and the number of epochs is 30, with 1000 steps for each epoch. An adaptive learning rate is used for each iteration. Dropout for the three models is 0.1, 0.15, and 0.2. Ensemble CNN model is trained using 24000 images, tested using 12000 images, and validated using 12000 images with equally distributed images among each class. During the training of the CNN models, subsamples of the dataset are used. Errors for each subsampled data point are calculated such that if the error is more significant than a threshold, those points are discarded, and training is continued. Also, all three models' features are aligned with highest to lowest priority features. Based on which zero variance features are removed, and features are sent to the next layer.

IV. Results

Improving the accuracy of the classification process and decreasing the misclassification error are two main concerns in ensemble models. This chapter implements six ensemble models for melanoma classification based on the general category aspects. Bagging, boosting, majority voting, infinite ensemble framework, and cluster ensembles are implemented using Random Forest, Adaboost, Majority voting, Boosted SVM, and Boosted GMM classifiers. The Ensemble CNN model is a sequential model developed based on a majority voting ensemble. The ML models used in Section III are SVM, GMM, decision tree, and deepconvnet. These are taken as base models in ensemble models. The basic parameter attributes are tested in Section III, and hyper parameter tuning of the ensemble models is investigated in Section IV. The hyper-parameters which could be fine-tuned for ensemble models are n estimators, learning rate, and m-features. Since deep learning models are analyzed, m features are not fine-tuned for our investigation. And since n- estimators and learning rates are related, these two parameters are considered for hyper-parameter tuning and optimization of the ensemble models. The performance of the classifiers in melanoma prediction is assessed using metrics like accuracy, precision, recall, F1 score, MCC, Jaccard Index, and Kappa. Also, six classes of Melanoma are classified here. In this Stratified 10- fold, cross-validation is used for calculating the performance estimates. Stratification ensures that each class is represented with the same proportions roughly as in the entire data set. Ensemble diversity is used in the datasets to achieve better accuracy and avoid overlapping features in the dataset during clustering. The investigation has carried over 10000 melanoma images generated by GAN. Training of the classifier model has been done with 1000 superficial spreading Melanoma, 1000 Nodular Melanoma, 1000 lentigo melanoma, 1000 acral lentiginous Melanoma, 1000 subungual Melanoma, and 1000 amelanotic Melanoma images. For testing 2000 and validation, 2000 images are used. The entire process has been carried over the python platform. Since CNN models require more datasets to avoid overfitting, it is trained using 24000 images, tested using 12000 images, and validated using 12000 images with images equally distributed among each class.

A. Performance Analysis of Random Forest Classifier

In this investigation, a random forest classifier is implemented for melanoma prediction. Ensemble pruning is done here to reduce the complexity of the network, and the maximum number of features is set to auto. Due to pruning, the random state is fixed to zero. Even though hyperparameter tuning is not necessary for random forest classifiers, the number of estimators varies between 50,100 and 150. The number of estimators (n-estimators) indicates the number of trees or samples required to find the optimum solution. However, it doesn't indicate that the higher the number of trees higher the accuracy. Increasing the number of trees leads to higher computational time. Also, the classifier model's performance will drop for a higher number of estimators. So, choosing the optimum number of estimators is done using the trial-and- error method for our investigation. The minimum sample leaf is restricted to 10, 25, and 50. For each of the six classes of Melanoma, namely superficial spreading Melanoma, Nodular Melanoma, lentigo melanoma, acral lentiginous Melanoma, subungual Melanoma, and amelanotic Melanoma combinations of n estimators and minimum sample leaves were fixed. The performance of the classifier is tested. The random forest classifier's performance is best for melanoma classification with an accuracy of 90.23 for the n estimator minimum leaf combination of 100:50. It is found that for a smaller number of estimators, Random forests suffer from underfitting problems. Performance analysis of random forest classifier is shown in Table I.

B. Performance Analysis of Adaboost Classifier

Adaboost, one of the popular boosting algorithms, is known to reduce outliers and overfitting issues. It also helps in improving the performance and robustness of the classifier. For Adaboost classification, different n estimators for different learning rate combinations are investigated to improve accuracy. N estimators used are 50,100,150 and learning rates are 0.01, 0.001 and 0.0001.

Adaboost classifier performs well with an accuracy of 96.78 for a learning rate of 0.01 with 50 estimators. Since base estimators affect the performance of the Adaboot classifier, svc is used as the base estimator. Performance metrics for melanoma detection using the Adaboost classifier are shown in Table II.

C. Performance Analysis of Ensemble Voted Classifier

Ensemble voted classifier is implemented using the majority voting method. For ensemble networks, pretrained base learners are necessary. GMM and SVM classifiers are implemented using the same settings discussed Decision tree, AdaBoost, and Random Forest classifier's performance vary over the number of estimators, and the learning efficiency differs for each model. So, ensemble majority voted model has been implemented for different estimator learning rate combinations with a fixed threshold of 0.5. Here the combined majority voting is calculated. For each class prediction, if the vote probability is more significant than 0.5, then the majority vote among the classifier output is chosen. If the probability of none of the classifiers is above 0.5, then the model is again restarted for other weights. Performance analysis of the Ensemble voted classifier in Table III shows that the classifier achieves better accuracy of 96.32 for a learning rate of 0.01.

TABLE I. PERFORMANCE ANALYSIS OF RANDOM FOREST CLASSIFIER

Estimators	Learning rate	Precision	Recall	Accuracy	F1-score	MCC	Jaccard Index	Kappa
	10	68.34	69.89	70.4	69.11	48.49	65.66	0.638
50	25	68.57	69.99	74.82	69.27	49.91	69.33	0.712
	50	70.12	71.34	71.27	70.72	51.52	70.49	0.712
	10	73.87	74.21	76.89	74.04	57.34	73.79	0.728
100	25	74.56	75.11	79.45	74.83	68.34	74.99	0.731
	50	86.54	88.31	90.23	87.42	87.31	89	0.771
	10	86.34	88.19	89.87	87.26	86.25	88.17	0.771
150	25	86.36	88.24	89.93	87.29	86.78	88.51	0.764
	50	86.47	88.29	89.97	87.37	86.91	88.81	0.76

TABLE II. PERFORMANCE ANALYSIS OF ADABOOST CLASSIFIER

estimators	Learning rate	Precision	Re-call	Accuracy	F1 score	MCC	Jaccard Index	Kappa
50	0.01	93.21	94.78	96.78	93.99	95.78	93.29	0.836
	0.001	93.2	94.27	96.41	93.73	94.21	93.02	0.827
	0.0001	92.65	93.56	96.17	93.1	92.87	92.13	0.824
100	0.01	93.19	93.21	94.71	93.2	88.34	91.22	0.821
	0.001	93.02	92.16	91.34	92.59	79.48	90.67	0.817
	0.0001	92.89	92.11	92.65	92.5	87.49	89.54	0.802
150	0.01	91.56	91.83	90.47	91.69	77.92	87.48	0.798
	0.001	90.18	91.27	90.65	90.72	72.85	88.01	0.783
	0.0001	89.91	90.42	88.73	90.16	68.79	83.17	0.779

TABLE III. PERFORMANCE ANALYSIS OF ENSEMBLE VOTED CLASSIFIER

Estimators	Learning rate	Precision	Recall	Accuracy	F1-score	MCC	Jaccard Index	Kappa
50	0.01	92.56	93.89	96.32	93.22	93.67	90.45	0.892
	0.001	91.89	92.57	95.48	92.23	91.42	89.93	0.879
	0.0001	90.43	91.27	93.71	90.85	87.17	89.92	0.879
100	0.01	90.12	90.17	92.87	90.64	82.39	89.74	0.872
	0.001	88.67	89.36	91.28	89.01	80.39	87.95	0.865
	0.0001	84.31	86.29	87.91	85.29	78.59	84.77	0.847
150	0.01	83.37	86.17	87.59	84.75	71.44	84.9	0.823
	0.001	81.29	85.99	86.71	83.57	68.53	83.01	0.796
	0.0001	76.15	75.82	82.19	75.98	52.31	79.02	0.747

D. Performance Analysis of Boosted SVM Classifier

From the previous investigation on SVM, it has been proved that the RBF kernel works well on melanoma images for a gamma value of 10. The same parameters are used in the boosted SVM classifier for different n estimators and learning rates. Even though the AdaBoost classifier works well only for 50 n estimators, boosted SVM gives better accuracy of 98.37 with a 0.01 learning rate. MCC is also significantly improved compared to conventional SVM and AdaBoost classifier models. Performance metrics of boosted SVM classifier for melanoma classification are shown in Table IV.

E. Performance Analysis of Boosted GMM Classifier

GMM model works best as a density estimator in clustering, and the EM algorithm is applied for Classification. To maximize the robustness of the GMM classifier AdaBoost algorithm is used here. Boosted GMM classifier is analyzed for different estimators and learning rates, and the best parameter setting for the classifier is fixed. It is found that the boosted GMM classifier works best at 25 estimators for a learning rate of 0.01 to provide an accuracy of 96.17. GMM shows a gradual performance improvement compared to other classifiers. Performance metrics of boosted SVM classifier for melanoma classification are shown in Table V.

TABLE IV. PERFORMANCE ANALYSIS OF BOOSTED SVM CLASSIFIER

Estimators	Learning rate	Precision	Recall	Accuracy	F1-score	MCC	Jaccard Index	Kappa
10	0.01	96.48	92.34	97.16	94.36	89.48	88.28	0.873
	0.001	95.39	92.58	96.9	93.96	89.27	88.15	0.873
	0.0001	92.56	93.26	96.15	92.91	89.1	88.01	0.873
25	0.01	99.78	95.76	98.37	97.73	96.93	89.96	0.886
	0.001	98.89	95.61	97.93	97.22	96.73	89.46	0.881
	0.0001	98.1	94.92	97.61	96.48	96.2	89.31	0.879
50	0.01	97.67	94.36	97	95.99	95.91	89.18	0.871
	0.001	97.41	94.21	96.49	95.78	95.63	89.03	0.868
	0.0001	97.18	94.2	96.12	95.67	95.42	88.97	0.861

TABLE V. PERFORMANCE ANALYSIS OF BOOSTED GMM CLASSIFIER

Estimators	Learning rate	Precision	Recall	Accuracy	F1-score	MCC	Jaccard Index	Kappa
10	0.01	90.18	89.76	95.89	89.97	83.85	84.95	0.723
	0.001	90.04	89.69	95.64	89.86	83.48	84.89	0.798
	0.0001	89.93	89.52	95.39	89.72	89.21	84.73	0.799
25	0.01	90.65	92.87	96.17	91.75	92.9	86.15	0.827
	0.001	90.57	92.74	96.13	91.64	92.84	86.02	0.819
	0.0001	90.43	92.69	96.08	91.55	92.77	85.99	0.811
50	0.01	90.39	92.34	96.01	91.35	92.68	85.81	0.804
	0.001	90.31	92.38	95.98	91.28	92.52	85.71	0.801
	0.0001	90.28	92.1	95.91	91.18	92.41	85.62	0.798

F. *Performance Analysis of Ensemble CNN Classifier*

Ensemble CNN is a sequential voting approach implemented using three different CNN models for variable learning rates. This model provides a more stable melanoma prediction accuracy than other ensemble approaches. Even though the implementation of this Ensemble is similar to the majority voting ensemble, this model performs better due to its distinctive feature extraction. This model extracts low and high- frequency features irrespective of the CNN type unless fixed features from previous models. Also, three different models have different dropout rates and kernel counts, preserving overfitting issues in the CNN model. Model converges earlier without much misclassification error. Performance metrics of the ensemble CNN model, as shown in Table VI, indicate the accuracy of 98.67 for melanoma classification. The network performs best with a learning rate of 0.0001 for 25 estimators, which is relatively less than other ensemble models.

Performance metrics for the classifiers mentioned above for independent classes of Melanoma are shown in Table VII. Based on the classifier performance, it is clear that all the classifiers perform better in classifying superficial spreading Melanoma, Nodular Melanoma. Due to colossal variation and differences in the structural properties in different stages of other types of other types of melanomas, the accuracy of the classifiers is less compared to superficial spreading and nodular Melanoma. Subungual Melanoma is present in nails and nail beds. Properties of this Melanoma sometimes resemble typical characteristics of vitamin deficiency. So, classifiers require extreme robustness to achieve better accuracy. Amelanotic Melanoma is one particular type of Melanoma present in all skin variants. Also, amelanotic Melanoma in certain stages resembles superficial spreading Melanoma and nodular Melanoma.

The ensemble classifiers implemented in this work produce better accuracy than the single classifiers. Adaboost and Boosted SVM classifiers perform better for all five types of Melanomas except amelanotic Melanoma. The other three classifiers, namely boosted GMM, Random Forest, and Ensemble voted classifiers, are performing better in superficial spreading, nodular, and lentigo melanoma classification but is moderate in the other two types despite the best hyperparameter settings, as shown in Table VII. Ensemble CNN models can provide consistent performance for all six types of melanoma classification with slight variation for amelanotic Melanoma.

The convergence plot in Fig. 3 shows the robustness of the ensemble models in melanoma classification. The maximum number of epochs used is 30 to check the training and validation accuracy. Out of six ensemble classifier models used, Ensemble CNN, Adaboost, and Boosted SVM classifiers resulted in better convergence. However, the Ensemble CNN model shows overfitting during validation even though accuracy is higher. Further improvements in the network model and data selection need to be made to avoid overfitting issues. Boosted GMM and Random Forest models are showing underfitting of data points. Ensemble voted model shows the best fit from the 28^{th} epoch.

TABLE VI. PERFORMANCE ANALYSIS OF ENSEMBLE CNN CLASSIFIER

Estimators	Learning Rate	Precision	Recall	Accuracy	F1-score	MCC	Jaccard Index	Kappa
	0.01	90.57	92.74	96.13	91.64	92.84	86.02	0.834
25	0.001	90.65	92.87	96.17	91.75	92.9	86.15	0.839
	0.0001	99.96	99.86	98.67	98.15	97.34	92.71	0.899
	0.01	90.43	92.69	96.08	91.55	92.77	85.99	0.832
50	0.001	90.39	92.34	96.01	91.35	92.68	85.81	0.832
	0.0001	90.31	92.28	95.98	91.28	92.52	85.71	0.832
	0.01	90.28	92.1	95.91	91.18	92.41	85.62	0.832
100	0.001	89.93	89.52	95.39	89.72	83.21	84.73	0.832
	0.0001	90.04	89.69	95.64	89.86	83.48	84.89	0.832

TABLE VII. PERFORMANCE ANALYSIS OF ENSEMBLE CLASSIFIERS

Classifiers	Melanoma classes	Precision	Recall	Accuracy	F1 score	MCC	Jaccard Index	Kappa
	Superficial spreading	99.82	95.83	98.41	97.78	96.94	90	0.886
	Nodular	99.8	95.76	98.41	97.74	96.93	89.96	0.886
Boosted	Lentigo maligna	99.78	95.68	98.37	97.69	96.93	89.96	0.886
SVM	Acral lentignous	99.62	95.68	98.37	97.61	96.91	89.95	0.883
	Subungual	99.58	95.62	98.37	97.56	96.87	89.91	0.881
	Amelanotic	99.58	95.62	98.31	97.56	96.81	89.91	0.88
	Superficial spreading	90.65	92.92	96.25	91.77	92.95	86.46	0.827
	Nodular	90.65	92.92	96.25	91.77	92.95	86.46	0.827
Boosted	Lentigo maligna	90.65	92.9	96.23	91.76	92.95	86.38	0.827
GMM	Acral lentignous	90.62	92.87	96.19	91.73	92.93	86.29	0.819
	Subungual	90.58	92.87	96.17	91.71	92.9	86.15	0.812
	Amelanotic	90.58	92.81	96.17	91.68	92.9	86.15	0.812
	Superficial spreading	86.54	88.31	90.23	87.42	87.31	89	0.771
	Nodular	86.54	88.31	90.23	87.42	87.31	89	0.771
Random	Lentigo maligna	86.54	88.29	90.22	87.41	87.26	89	0.771
forest	Acral lentignous	86.51	88.27	90.2	87.38	87.24	88.97	0.769
	Subungual	86.5	88.25	90.19	87.37	87.21	88.89	0.766
	Amelanotic	86.5	88.25	90.19	87.37	87.21	88.86	0.757
	Superficial spreading	93.54	94.89	96.93	94.21	95.86	93.47	0.836
	Nodular	93.51	94.81	96.84	94.16	95.81	93.41	0.836
Adaboost	Lentigo maligna	93.21	94.78	96.78	93.99	95.79	93.29	0.836
Classifier	Acral lentignous	93.21	94.77	96.78	93.98	95.78	93.31	0.836
	Subungual	93.2	94.77	96.78	93.98	95.78	93.29	0.836
	Amelanotic	93.19	94.77	96.78	93.97	95.78	93.29	0.836
	Superficial spreading	92.59	93.93	96.36	93.26	93.71	90.62	0.892
Ensemble	Nodular	92.56	93.89	96.36	93.22	93.65	90.62	0.892
voting	Lentigo maligna	92.55	93.86	96.32	93.20	93.61	90.57	0.887
classifier	Acral lentignous	92.51	93.82	96.31	93.16	93.56	90.51	0.883
	Subungual	92.49	93.84	96.29	93.16	93.51	90.45	0.881
	Amelanotic	92.49	93.81	96.29	93.15	93.51	90.45	0.881
	Superficial spreading	99.96	99.86	98.67	99.91	97.34	92.71	0.899
Ensemble	Nodular	99.96	95.86	98.67	97.87	97.34	92.7	0.899
CNN	Lentigo maligna	99.94	95.78	98.64	97.82	97.29	92.68	0.897
classifier	Acral lentignous	99.94	95.78	98.64	97.82	97.29	92.68	0.892
	Subungual	99.89	95.78	98.64	97.79	97.29	92.63	0.892
	Amelanotic	99.86	95.71	98.61	97.74	97.27	92.59	0.892

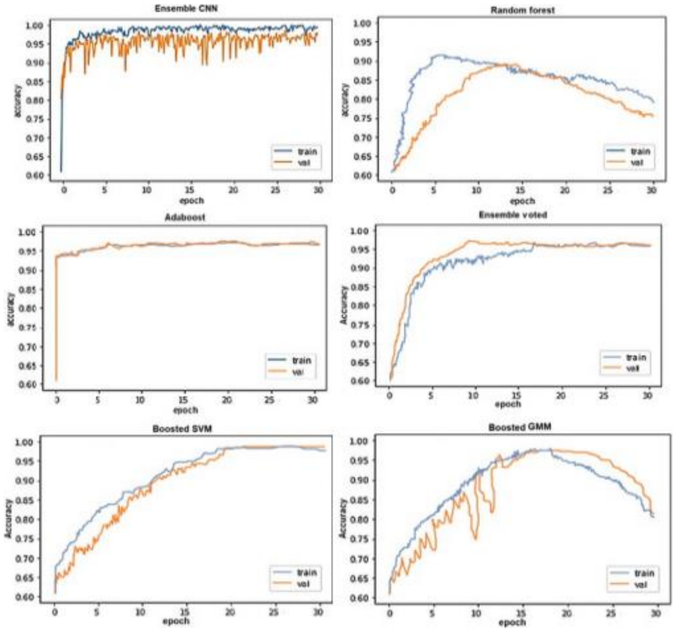

Fig. 3. Accuracy vs. epoch plot for convergence analysis.

V. Conclusion

Ensemble learning models as classifiers for melanoma classification. Bagging, boosting, majority voting, infinite ensemble framework, and cluster ensembles are implemented using Random Forest, Adaboost, Majority voting, Boosted SVM, Boosted GMM classifiers, and Ensemble CNN models.

The performance of the classifiers in melanoma prediction is assessed using metrics like accuracy, precision, recall, F1 score, MCC, Jaccard Index, and Kappa. Also, six classes of Melanoma are classified here. In this Stratified 10-fold, cross-validation is used for calculating the performance estimates. Stratification ensures that each class is roughly represented with the same proportions as in the entire data set. Ensemble diversity is used in the datasets to achieve better accuracy and avoid overlapping of features in the dataset during clustering. Ensemble models can perform well for five classes with consistent accuracy out of six classes. The boosted SVM and Adaboost classifiers have higher performance than boosted GMM, random forest, and Ensemble voted classifiers. Ensemble CNN seems to outperform other ensemble models with an accuracy of 98.67. Though the execution time of ensemble classifiers is high, such a complex network is easier to train, and the network converges ideally. The system's complexity is one point that needs to be considered in the proposed model for further improvement. Also, it was observed that an increase in the number of images in the training dataset increased the size of the feature set, which led to overlapping features.

References

[1] Aggarwal, "Automated skin lesion classification using ensemble of deep neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1925–1936, 2018.

[2] Jha, "A review of deep learning methods for skin lesion classification in dermatology," *Skin Research and Tech- nology*, vol. 27, no. 4, pp. 453–467, 2021.

[3] Esteva, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] Pala, "Classification of skin cancer images using a hybrid deep learning model," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 11, pp. 12 797– 12 806, 2021.

[5] Islam, "Multi-class skin lesion classification using a deep residual network with spatial pyramid pooling," *Computer Methods and Programs in Biomedicine*, vol. 197, pp. 105 742–105 742, 2020.

[6] G. R. Praveena and M. S. Indhumathi, "Skin lesion classification using deep learning: A review and future directions," *Biomedical Signal Processing and Control*, vol. 70, pp. 102 900–102 900, 2021.

[7] Y. Wang, "A novel skin cancer classification approach using deep convolutional neural networks with attention mechanism," *Computerized Medical Imaging andGraph- ics*, vol. 97, pp. 101 996–101 996, 2023.

[8] Wei, "Interpretable classification of skin lesions using attention-based convolutional neural networks," *Journal of Medical Systems*, vol. 46, no. 1, pp. 5–5, 2022.

[9] Lin, "Skin cancer classification using transfer learning and optimized convolutional neural networks," *IEEE Access*, vol. 10, pp. 41 113–41 121, 2022.

[10] Kourou, "Deep learning for skin cancer classification: A comparison study," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2349–2358, 2020.

[11] Codella, "Skin lesion analysis toward melanoma detec- tion: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the Interna- tional Skin Imaging Collaboration (ISIC)," *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 168–172, 2018.