www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2022, VOL. 68, NO. 2, PP. 251-257
Manuscript received May 14, 2021; revised May, 2022. DOI: 10.24425/ijet.2022.139875

An Ensemble of Statistical Metadata and CNN Classification of Class Imbalanced Skin Lesion Data

Sachin Nayak, Shweta Vincent*, Sumathi K, Om Prakash Kumar, and Sameena Pathan

***Abstract*—Skin Cancer is one of the most widely present forms of cancer. The correct classification of skin lesions as malignant or benign is a complex process that has to be undertaken by experienced specialists. Another major issue of the class imbalance of data causes a bias in the results of classification. This article presents a novel approach to the usage of metadata of skin lesions' images to classify them. The usage of techniques addresses the problem of class imbalance to nullify the imbalances. Further, the use of a convolutional neural network (CNN) is proposed to fine-tune the skin lesion data classification. Ultimately, it is proven that an ensemble of statistical metadata analysis and CNN usage would result in the highest accuracy of skin color classification instead of using the two techniques separately.**

***Keywords*—classification; Convolutional Neural Networks; Ensemble Learning; machine learning; metadata**

I. Introduction

SKIN lesions are among the significant causes of death and mental stress in the world [1, 2]. The main reason behind this is the false classification by experts or deprivation of access to an expert's advice. There is a limitation to accurate classification of skin lesions even by experts due to lack of expertise or skin reflections. With the increasing population, there is more pressure on doctors to perform tests faster, which further leads to mistakes. A cost-effective method for diagnosis would help medical testing and reduce stressful situations for patients suffering from various kinds of skin lesions.

Malignant melanoma, if appropriately treated at an early stage, could be cured. But distinguishing malignant melanoma from benign melanoma at an early stage is time-consuming for an expert. In this digital world, having a mobile app/ computer software that could read an image of a skin lesion and with some preliminary data viz., age, the position of a skin lesion on the body, etc., and classify the lesion into a skin disease class, would be beneficial. This could help doctors perform their operations faster. People with smartphones can check if they have any severe skin disease or not. NGOs or other organizations working in underprivileged places can check citizens for any illnesses and provide them the required medical attention.

Another major issue in the efficient classification of skin lesions is the impact of class imbalances. This occurs due to the availability of more images belonging to a particular class A of skin disease and the unavailability of the same number of images in another class B. The resulting drawback is that while training the CNN model, more images of class A are encountered instead of class B. This leads to the misclassification of many instances of class B disease image samples as class A images.

Further, along with image data, there is a plethora of metadata IS available. No considerable effort has been made to mine this metadata. Therefore, an effort must be made to perform a statistical study on this metadata to perform classifications of skin lesions.

In this regard, our article aims at:

- Configuring, training, and testing classification models using the variables selected from statistical study of metadata of skin lesions. Using chosen variables, deep learning classification models are built.
- Using data replication, under-sampling, and class weighting to reduce class imbalance effects in classification models based on metadata.
- Choosing an appropriate CNN model configuration for training and testing to classify skin lesion images.

This article has been divided in the following manner. Section II describes the available literature in our research area. Section III details the statistical methods applied to the metadata of skin lesions' images and the results obtained. Section IV focuses on the issue of class imbalance and techniques to solve it. Section V presents the classification of image data using CNN. Section VI gives the results of using the ensemble model of statistical metadata classification and image classification using CNN. The final section concludes the article and presents the scope for future work.

II. Related Work

Kiran Pai et al. [3] have used VGGNet CNN architecture to predict and classify skin lesions into seven classes. They also developed a website that can predict the three most probable types of skin lesions for a loaded image. MNIST: HAM10000 dataset [15], which contains 10000 images of 7 classes, is used for the experiment. Adams optimizer with an initial learning rate of 0.001 has been used. Agnieszka Mikołajczyk et al. [4], in their research paper, have focused on solving the problem of lack of data and class imbalance using augmentation techniques. Balazs Harangi [5] has used an ensemble of four CNNs, namely GoogLeNet, AlexNet, ResNet, and VGGNet. This improves skin lesions classification accuracy into three classes, viz., melanoma, nevus, and seborrheic keratosis. Enes Ayan et al. [6] have used a CNN architecture and then compared its

Sachin Nayak and Shweta Vincent* are with the Department of Mechatronics Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India (e-mail: sachinunayak369@gmail.com, corresponding author shweta.vincent@manipal.edu).

Sumathi K is with the Department of Mathematics, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India (e-mail: k.sumathi@manipal.edu).

Om Prakash Kumar is with the Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India (e-mail: omprakash.kumar@manipal.edu).

Sameena Pathan is with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India (e-mail: sameena.pathan.k@gmail.com).

performance with and without augmentation. Keras' Image Data Generator API was used for the augmentation of images.

Nils Gessert et al. [7] have compared the ability of oversampling, balanced batches, loss weighting, and diagonal weighting methods in solving class imbalance problems. They have used 224×224×3 images for training and testing.Khalid M. Hosny et al. [8] have presented an article for the automated skin cancer lesions classification. They have used the concept of transfer learning and a pre-trained neural network. Nazia Hameed et al. [9] have showcased the usage of Deep CNN and SVM to classify skin diseases into multiple classes. Five classes of skin diseases have been used, viz., healthy, acne, eczema, benign melanoma, and malignant melanoma. Efficient classification of skin lesions using CNN with a novel regularizer has been performed by Marwan Ali Albahar [10]. An accuracy of 97.49% has been obtained in the classification of skin lesions as benign or malignant.

Achim Hekler et al. [11] have presented an article on the usage of human and artificial intelligence to classify images for skin cancer. A combination of human and artificial intelligence techniques has led to an accuracy of 82.95%, which was higher when using human intelligence separately.H. Kittler et al. [12] have presented a comprehensive article on how dermoscopy increases the accuracy of diagnosis for melanoma instead of using the naked eye for experienced examiners. M. Maragoudakis et al. [13] have showcased using an ensemble of classification techniques comprising of Random Forest with Markov Blanket notion. J. Kawahara et al. [14] have demonstrated the usage of pre-trained CNNs and images of different resolutions to create a novel CNN architecture for the classification of skin lesions.

On a comprehensive note, there are many approaches to the classification of skin lesions. However, there is virtually no effort related to the classification of this dataset using the metadata available. Our research work aims at performing an ensemble of statistical analyses of the metadata. It uses it along with CNN for the classification of images of skin lesions with higher accuracy. The following section of our article presents a theoretical background of our research.

III. STATISTICAL ANALYSIS OF METADATA AND RESULTS OBTAINED

The main motive of this research article is to classify images of skin lesions into different classes of skin diseases. However, with the extensive availability of metadata (data about data) viz., age of the person, sex of the person, and localization of skin lesion, an attempt is being made in our research to build a classification model using this metadata. The propelling factor to do so is to be able to build an application that could be easily usable and understandable by the layman for coarse classification before he approaches a skin specialist for further examination. Other metadata parameters, such as family history, the past occurrence of infection, etc., could also be considered for further research.

The parameters of age, sex, and localization are three different kinds of parameters and need to be handled differently in any programming language. Age is a single value continuous data, whereas localization is a single value categorical data. Sex is taken to be a binary variable. On the other hand, images of the skin lesion are 3-dimensional datasets that comprise of the width and height of the image, and the third dimension corresponds to the RGB values of the image. The following paragraphs describe the methods to handle different forms of data for age, sex, and localization. Using categorical data for machine learning: Models could be built for three types of data viz.,

- categorical input- continuous output
- continuous input-categorical output
- categorical input- categorical output data.

But libraries available in programming languages restrict the ways in which categorical or any other type of data is used. If data involves words, then it is not possible to directly feed that data to a model. Models require data to be in numerical form. Hence, few changes need to be made before using that data. Categorical data can be dealt with using ordinal encoding, one-hot encoding, or learned to embed.

In our research, localization and sex data are encoded using one-hot encoding. Since sex data has only two categories, i.e., male and female, we can encode them with a vector with two dimensions, [0 1] and [1 0], or vice versa. The localization data is taken to have 14 categories (abdomen, back, chest, ear, face, foot, genital folds, lower extremity, hand, palm, neck, scalp, trunk, and upper extremity), which correspond to 14 dimensions in the one-hot encoded vector. One value will be '1' in each vector, and the rest would be '0'. The Kaggle Skin Cancer MNIST: HAM10000 dataset15 is used in this research. Using this dataset, an analysis is performed on the metadata to find any dependencies that exist between variables and skin lesion classes. Normality tests and significance tests are performed on the data. Normality tests are used to check whether a distribution is approximately normal or not normal.

Significance tests are used to check if there exists any significant relationship between variables.

The dataset being used for this research has images of skin lesions with labels. A metadata file has sex data, localization data, and age data of these patients. 10015 instances are available in the dataset, and a brief summary of it is given in Table I. For all tests, a significance level (α) of 5% is being used. Significance level limits type 1 error in the long run. Type 1 error is, we believe and accept that one variable has a significant effect on another variable, but it is just a random effect that we are observing in reality. In this case, we are limiting this error rate to 5%. Therefore, if we reject the null hypothesis based on our observations, then we could be wrong only 5% of the time or less than that.

The various statistical tests performed in this research and their corresponding results are described further.

TABLE I
DESCRIPTION OF THE DATASET USED

Disease	Label	Instances	Abbreviation
Actinic keratoses	0	327	akiec
Basal cell carcinoma	1	514	bcc
Benign keratosis-like lesions	2	1099	bkl
Dermatofibroma	3	115	df
Malignant Melanoma	4	1113	mel
Melanocytic nevi	5	6705	nv
Vascular lesions	6	142	vasc

A. NORMALITY TESTS

Since age data is continuous, the approach towards analysis of age data is different from categorical data like sex and localization data. Statistical tests vary, depending on whether data is normally distributed or not. Therefore, the following normality tests are performed to check if age data for different skin lesion classes are normally distributed or not.

Following this, numerical tests for normality have been performed. There exists a null hypothesis and an alternate hypothesis. Based on the test results, either the null hypothesis would be accepted or rejected. Here the null and alternative hypotheses are as follows:

H0: Age data is normally distributed

H1: Age data is not normally distributed

The significance level (α) is taken as 5%. The p-values are checked for the test results. The p-values are obtained from the distribution of a statistic calculated for input data and are compared with the significance level. If a p-value is less than the required significance level, it signifies that we have enough proof to reject the null hypothesis. Therefore, the null hypothesis is rejected for that data at the required significance level.

Different types of tests like the Shapiro-Wilk test [16], D'Agostino's K-squared test [17], and Anderson-Darling test [18] have been used to measure the normality of a given dataset.

Table II illustrates the results obtained for the Shapiro-Wilk test. The p-values of tests for all skin lesion classes are lower than 0.05(significance level); hence there is enough data proof to reject the null hypothesis. That means age data is not normally distributed for any of the skin lesion classes.

TABLE II
RESULT OF SHAPIRO-WILK TEST

Skin lesion class	Statistic		p-value	H0
Actinic keratoses	0.9615		< 0.0001	Reject
Basal cell carcinoma	0.9258		< 0.0001	Reject
Benign keratosis-like lesions	0.9424		< 0.0001	Reject
Dermatofibroma	0.9690		0.0091	Reject
Malignant Melanoma	0.9665		< 0.0001	Reject
Melanocytic nevi	0.9821		< 0.0001	Reject
Vascular lesions	0.9518		0.0001	Reject

We have enough evidence available to reject the null hypothesis for all skin lesion classes using the D'Agostino's K-squared and Anderson-Darling test. It is concluded that age data is not normally distributed for any of the skin lesion classes. All three tests suggested that the age distributions are not normal. With this conclusion, it is decided to use non-parametric tests for the statistical analysis of data.

B. NON-PARAMETRIC TESTS

Non-parametric tests don't make any assumptions about the underlying distribution of the data. Distributions are defined using a set of parameters, and since no assumption is made regarding the distribution in this test, there are no parameters to define it.

B.1 Kruskal-Wallis test [19]

In this test, the following are considered as the null and alternative hypotheses.

H0: Age distributions for all skin lesion classes are the same.

H1: Age distribution is different of one or more skin lesion classes.

The significance level (α) is considered to be 5%. For a statistic of 2270.9222, the attained p-value is < 0.0001, which is smaller than the α-value. Therefore the alternate hypothesis that age distribution is different for one or more skin lesion classes (H1) is accepted.

This result doesn't tell us which skin lesion classes have a significant difference in their age distributions. But for machine learning purposes, we need to know the differences that exist so that we can decide on whether to use this attribute in our model or not. This requirement leads us to our next step. Now we need to compare the age distributions of all 'pairs' of skin lesion classes. Therefore, in the next section, multiple comparison tests for the age distribution of skin lesion classes are performed in pairs.

C. PAIR-WISE COMPARISON TESTS

C.1 Mann Whitney U test [20]

The following is the null and alternative hypothesis chosen for the Mann-Whitney U test at a significance level (α) of 5%.

H0: Age distributions are equal.

H1: Age distributions are not equal.

Table III gives the results of the test.

TABLE III
RESULTS OF THE MANN WHITNEY U TEST

Lesion 1	Lesion 2	Statistic	p-value	H0
akiec	bcc	79270.0	0.0808	Accept
akiec	bkl	165359.0	0.0244	Reject
akiec	df	8700.0	<0.0001	Reject
akiec	mel	142324.5	<0.0001	Reject
akiec	nv	321411.5	<0.0001	Reject
akiec	vasc	13253.0	<0.0001	Reject
bcc	bkl	248389.0	0.0001	Reject
bcc	df	13834.0	<0.0001	Reject
bcc	mel	216221.5	<0.0001	Reject
bcc	nv	558760.5	<0.0001	Reject
bcc	vasc	20861.0	<0.0001	Reject
bkl	df	34819.5	<0.0001	Reject
bkl	mel	520990.0	<0.0001	Reject
bkl	nv	1394999.5	<0.0001	Reject
bkl	vasc	50046.0	<0.0001	Reject
df	mel	44446.5	<0.0001	Reject
df	nv	289341.0	<0.0001	Reject
df	vasc	8077.0	0.4411	Accept
mel	nv	1864221.0	<0.0001	Reject
mel	vasc	59998.0	<0.0001	Reject
nv	vasc	378023.5	<0.0001	Reject

Table III shows that except for two pairs out of 21 pairs of skin lesion classes' age distributions differ significantly at 5% significance level. We perform a few more pairwise comparison tests on age data to support the obtained result.

C.2 Wilcoxon test [21]

The results of the Wilcoxon test are the same as the results of the Mann-Whitney U test. Only two pairs out of 21 show no significant difference. However, there is a significant difference among age distributions of different skin lesion classes in the Wilcoxon test as opposed to the Mann-Whitney U test.

C.3 Kruskal-Wallis test with Bonferroni's correction

The Kruskal-Wallis test with Bonferroni's correction is performed with a corrected α value of 0.0024. Even from this test, we get results identical to the previous two non-parametric tests, i.e., Wilcoxon and Mann-Whitney U test. Therefore, it is concluded that all non-parametric tests have shown that there is a significant difference among age distributions of skin diseases.

C.4 Z-test [22]

The Z-test results match the results of tests performed under non-parametric tests, i.e., only for two pairs of skin lesions; the null hypothesis has been accepted.

C.5 T-test [23]

The T-test results are the same as the results of the Z-test performed previously. Only for two pairs of skin lesions, the null hypothesis is accepted, and for the rest of the pairs, there exists a significant difference between age distributions.

C.6 Tukey's Honestly Significant Difference (HSD) test [24]

In this test, the null hypothesis for an additional pair of lesions viz., akiec and bkl is accepted. The rest of the results remain the same as the Z and T-tests described above.

D. PARAMETRIC TESTS

Parametric tests are used when the underlying distribution of the data is known. Since a set of parameters defines distribution, these tests are called parametric tests. In this case, the data is assumed to have a normal distribution, and the corresponding parameter of the normal distribution is used.

D.1 ANOVA [25]

After performing the ANOVA test, it is observed that since the p-value is smaller than α, we have enough evidence to reject the null hypothesis. Therefore, there is a significant difference in the age distributions of one or more skin lesion classes.

D.2 ANOVA after Bonferroni's Correction

The ANOVA after Bonferroni's correction method is used after a corrected α value of 0.0024. The result is the same as that of using Bonferroni's method with the Kruskal Wallis test.

Therefore, in conclusion, after analyzing both non-parametric and parametric tests performed in the previous sections, we can see that the results and conclusions have been very similar. Hence, even though the normality tests suggested that age distributions are not normal for any of the seven diseases, we can now say that the population from which samples have been taken is approximately a normal distribution. With the increase in sample size, sample distribution also approaches a normal distribution.

The following section focuses on removing class imbalances in images of the seven classes of skin lesions used for testing.

IV. THE ISSUE OF CLASS IMBALANCE

The Class imbalance problem arises when we have a dataset that has a big difference in the number of instances of different classes. The summary of the dataset being used in this project is given below in Table IV.

TABLE IV
DATA DISTRIBUTION ABOUT SKIN LESION CLASSES

Disease	Label	Instances (original)	After removing unknown values	Abbreviation
Actinic keratoses	0	327	327	akiec
Basal cell carcinoma	1	514	509	bcc
Benign keratosis-like lesions	2	1099	1076	bkl
Dermatofibroma	3	115	115	df
Malignant Melanoma	4	1113	1101	mel
Melanocytic nevi	5	6705	6501	nv
Vascular lesions	6	142	142	vasc

In Table IV, it is observed that there is a big difference between the number of instances of the class Melanocytic nevi and the number of instances of other classes. For simplicity, let us consider only two classes, Melanocytic nevi, and Dermatofibroma. We can observe that Melanocytic nevi have 6590 more instances than Dermatofibroma. This difference is termed as Class imbalance, which causes undesirable effects on classification models trained on this data. Suppose we trained a model on data from the above specified two classes for classifying an instance as either Melanocytic nevi or Dermatofibroma. In that case, the model will have a tendency to classify Dermatofibroma instances as Melanocytic nevi instances.

Let us consider two models,

1. Model 1 classifies 60 out of 100 Dermatofibroma instances into the Melanocytic nevi class and 10 out of 100 Melanocytic nevi instances into the Dermatofibroma class.
2. Model 2 classifies 60 out of 100 Dermatofibroma instances into the Melanocytic nevi class and 60 out of 100 Melanocytic nevi instances into the Dermatofibroma class.

Even though Model 1 classifies a lesser number of instances incorrectly compared to Model 2, Model 2 is better compared to Model 1 as it gives equal importance to both classes. Therefore Model 2 is less biased. If there are a greater number of patients with Dermatofibroma than Melanocytic nevi, then during classification, Model 2 will have better performance compared to Model 1. The problem associated with Model 1 is due to class imbalance.

A. PERFORMANCE MEASURE FOR DATA WITH IMBALANCES

Since the data is highly imbalanced, accuracy does not provide much information about the model's actual performance. Accuracy is a measure of a model's overall performance, and it does not consider different classes present in the data separately. We need to compare the performance of the model for all classes present in the data. A model with good performance for majority class and poor performance for minority class shows high accuracy, but we cannot consider it as a good model as it shows poor performance for minority class, and the model will fail if more number of instances from minority class is fed to it and accuracy drops suddenly. Therefore, some new measures evaluate a model's performance and understand how good a model is. Hence, we use three additional metrics to measure the model viz., Precision, Recall, and F1-score.

B. SKIN LESION CLASSIFICATION AFTER RESOLVING CLASS IMBALANCE EFFECTS

The classification process is performed using the parameters above of Precision, Recall, F1-score, and accuracy on the data concerning the age and localization of skin lesions' images.

B.1 Classification using Age data

Since age distributions of few skin lesion classes overlap, it is impossible to distinguish between a few of the skin lesion class pairs solely using age data. But age data does show good performance while distinguishing between few other pairs of skin lesion classes. Tensorflow library of Python is used to build the deep learning models. While building a machine learning model, weights are updated automatically, but the model requires tuning on the number of layers of neurons, the number of neurons in each layer, and the learning rate.

The configuration of 1-20-20-7 is used where 1 is the number of input neurons followed by two hidden layers of 20 neurons each and an output layer with 7 neurons corresponding to 7 skin lesion classes in the dataset. Since there is an overlap of the age distributions for some skin lesion classes, building a classification model for all classes at once yields poor results. Therefore, training and testing for two skin lesion classes at a time are performed. The model is trained for 100 epochs. 80% of the data is used for training, and 20% is used for testing. Table V illustrates the original data model with class imbalances and its result of classification. A recall value of more than 50% is considered a correct classification.

Table V shows that there are only two class pairs with both recall values greater than or equal to 50%. The mean performance of the models is bad. This is due to the class imbalance effect and the smaller number of samples in some of the skin lesion classes.

TABLE V
ORIGINAL DATA MODEL WITH CLASS IMBALANCES

L1	L2	Number of instances in the training dataset		Precision		Recall		F1 score		Accuracy
		L1	L2	L1	L2	L1	L2	L1	L2	
akiec	bcc	258	410	0.00	0.59	0.00	1.00	0.00	0.74	0.59
akiec	bkl	260	862	0.00	1.00	0.00	0.76	0.00	0.86	0.76
akiec	df	259	94	0.77	0.50	0.99	0.50	0.86	0.09	0.76
akiec	mel	263	879	0.00	0.78	0.00	1.00	0.00	0.87	0.78
akiec	nv	265	5189	0.00	0.95	0.00	1.00	0.00	0.98	0.95
akiec	vasc	264	111	0.73	0.67	0.92	0.32	0.92	0.32	0.72
bcc	bkl	411	857	0.00	0.69	0.00	1.00	0.00	0.82	0.69
bcc	df	405	94	0.83	0.00	0.97	0.00	0.89	0.00	0.81
bcc	mel	395	893	0.00	0.65	0.00	1.00	0.00	0.78	0.65
bcc	nv	416	5184	0.65	0.94	0.12	1.00	0.20	0.97	0.94
bcc	vasc	402	118	0.85	0.60	0.96	0.25	0.90	0.35	0.83
bkl	df	862	90	0.90	0.00	1.00	0.00	0.94	0.00	0.90
bkl	mel	859	882	0.52	0.53	0.55	0.50	0.54	0.52	0.53
bkl	nv	853	5200	0.61	0.88	0.21	0.98	0.31	0.92	0.86
bkl	vasc	865	109	0.88	0.83	1.00	0.15	0.94	0.26	0.88
df	mel	96	876	0.00	0.92	0.00	1.00	0.00	0.96	0.92
df	nv	95	5189	0.00	0.98	0.00	1.00	0.00	0.99	98
df	vasc	94	111	0.62	0.64	0.24	0.90	0.34	0.75	0.63
mel	nv	855	5218	0.63	0.85	0.07	0.99	0.12	0.91	0.84
mel	vasc	872	122	0.92	0.00	1.00	0.00	0.96	0.00	0.92
nv	vasc	5198	108	0.97	0.00	1.00	0.00	0.99	0.00	0.97

After removing the class imbalances using under-sampling, the model's performance is better compared with the model trained on original data. There are 14 class pairs with recall values for both classes greater than or equal to 50%. Many classes are very close to this threshold value. Next, class imbalances are removed using the method of allocating class weights. Weights are assigned based on how many instances are present in each class. There are 13 class pairs with recall values for both classes greater than or equal to 50%.

B.2 Classification using Localization data

The neural network used for classification using localization data has the configuration 14-20-20-7 where 14 is the number of input neurons that correspond to the 14 localization classes, two 20 neurons correspond to two hidden layers, and at the end, 7 stands for 7 output classes corresponding to 7 skin lesion classes. Models based on the localization data also follow the same performance pattern as in age data. However, the localization data is better at classifying skin lesions in terms of recall value. The original imbalanced data has 8 class pairs with recall values for both classes greater than or equal to 50%. The class imbalance is removed, and classification based on localization data is performed. After classification, there are 18 class pairs with recall values for both classes greater than or equal to 50%. On removing class imbalances by weighting, there are 14 class pairs with recall values for both classes great than or equal to 50%.

On combining the results of models based on age data and localization data, 13 class pairs have recall values greater than 50% with the original data, and 17 class pairs have recall values higher than 50% after removing the class imbalance effects.

As it is evident from the above tables, results have improved after combining the two models' predictions for age and localization data. For models considering two classes simultaneously, the number of class pairs having an average F1 score and recall greater than or equal to 50% has increased. If models with all classes are considered, there is not much difference in the original data model results, but for increased data models and weighted class models, the average F1 score and average recall have increased from the previous models.

A summary of the classification results obtained from resolving the class imbalance effects is illustrated in Table VI.

TABLE VI
SUMMARY OF CLASSIFICATION USING METADATA OF AGE AND LOCALIZATION

Metadata	Type of data	Pairs with Recall > 50%
Using Age data	Class imbalanced data	2
	Class balanced data using undersampling	14
	Class balanced data using class weights	13
Using Localization data	Class imbalanced data	8
	Class balanced data using undersampling	18
	Class balanced data using class weights	14

V. CLASSIFICATION OF IMAGE DATA USING CNN

Borders, curves, shapes, colors and color changes, brightness and brightness variation, and many more features make up an image. These features of an image are useful in recognizing or classifying that image. The higher the image's resolution, the better the model's performance as there will be a greater number of features to analyze and use. Bigger datasets consume too much memory space. Greater computational power is required to deal with big datasets. Too much time is also consumed while dealing with bigger datasets. These disadvantages also result in higher costs. Therefore, it is imperative to reduce image size before using them. The tradeoff is a loss of features for the lesser cost of processing the image.

For our research of classification of skin lesions, images are downloaded from Kaggle into Google Colab. All images are of the shape 450×600×3. Image augmentation is used to increase the number of images of the minority classes. The Gaussian filter and Bilateral filter were used to de-noise the images.

The DenseNet201 model imported from Keras in Python was used as the CNN model for our classification. For the output layer, the softmax activation function with categorical cross-entropy loss function was used. Adam optimizer was used to optimize the results. The initial learning rate was 0.001, and it decayed after every 5 steps by 0.5 (half of the present value). Callbacks are used to achieve this. The CNN results for the original class imbalanced data model are shown in Table VII.

The class imbalance is improved by using image augmentation. The results of classification using CNN after image augmentation are shown in Table VIII.

TABLE VII
CNN RESULTS ON THE ORIGINAL DATA MODEL

Lesion	Number of instances	Precision	Recall value	F1 score
akiec	255	0.53	0.49	0.51
bcc	402	0.63	0.67	0.55
bkl	875	0.71	0.51	0.59
df	94	0.57	0.47	0.52
mel	860	0.58	0.53	0.55
nv	5207	0.89	0.94	0.91
vasc	115	0.70	0.80	0.74
Average		0.66	0.63	0.64
Accuracy	0.81			

TABLE VIII
CNN RESULTS ON DATA AFTER IMAGE AUGMENTATION TO REMOVE CLASS IMBALANCES

Lesion	Number of instances	Precision	Recall value	F1 score
akiec	255	0.59	0.56	0.57
bcc	402	0.65	0.71	0.68
bkl	875	0.64	0.63	0.63
df	94	0.63	0.50	0.56
mel	860	0.48	0.63	0.54
nv	5207	0.92	0.88	0.90
vasc	115	0.83	0.83	0.83
Average		0.68	0.68	0.67
Accuracy	0.80			

The recall value is 0.68 and has improved when compared to the original data model.

VI. ENSEMBLE OF STATISTICAL CLASSIFICATION AND IMAGE CLASSIFICATION MODELS

The final step in this research is to combine the statistical metadata model results and the image classification model.

TABLE IX
CLASSIFICATION USING STATISTICAL DATA AND IMAGE DATA USING CNN

Lesion	Number of instances	Precision	Recall value	F1 score
akiec	254	0.60	0.53	0.56
bcc	411	0.64	0.70	0.67
bkl	857	0.59	0.65	0.62
df	92	0.63	0.60	0.62
mel	854	0.53	0.68	0.60
nv	5229	0.92	0.87	0.90
vasc	119	0.75	0.68	0.71
Average		0.67	0.67	0.67
Accuracy	0.87			

Table IX shows and an ensemble of the statistical classification of localization and age data as well as the CNN image data.

As is evident from Table IX, the ensemble model gives the highest accuracy and overall best recall values, precision, and F1-scores.

CONCLUSION AND FUTURE WORK

This article presented a comprehensive analysis of using statistics of metadata of images and conventional image classification methods for skin lesions. The effect of class imbalances was explored, and methods to mitigate the same was implemented. They were combining metadata models with image data models resulted in performance improvement.

However, due to the limitation of the RAM space, images of 175×175×3 were used in this research. However, using images of larger size, i.e., the conventional 224×224×3, would yield better results. Further, there needs to be a mechanism to remove errors caused due to skin reflections.

As time passes, with advancements in technology, mathematics, statistics, machine learning, and complex programming languages, there will be numerous opportunities for us to explore and improve upon the present models.

REFERENCES

[1] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, A. Jemal, "Cancer statistics," *CA: A cancer journal for Clinicians*, 67(1), 7–30, 2017. https://doi.org/10.3322/caac.21395

[2] Z. Apalla, D. Nashan, R. B. Weller, X. Castellsague´, "Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches," *Dermatology and Therapy*, 7(1), 5–19, 2017. http://dx.doi.org/10.1007/s13555-016-0165-y

[3] K. Pai, A. Giridharan, "Convolutional Neural Networks for classifying skin lesions," *Proceedings of TENCON 2019 IEEE conference*, 1794-1796, 2019. https://doi.org/10.1109/TENCON.2019.8929461

[4] A. Mikołajczyk, M. Grochowski, "Data augmentation for improving deep learning in image classification problem," *Proceedings of 2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 117-122, 2018. http://dx.doi.org/10.1109/IIPHDW.2018.8388338

[5] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *Journal of Biomedical Informatics*, 86, 25-32, 2018. http://dx.doi.org/10.1016/j.jbi.2018.08.006

[6] E. Ayan, H. U. Ünver, "Data Augmentation Importance for Classification of Skin Lesions via Deep Learning," *Proceedings of 2018 Electric Electronics, Computer Science, Biomedical Engineering Meeting(EBBT)*, 10-15, 2018. http://dx.doi.org/10.1109/EBBT.2018.8391469

[7] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, A. Schlaefer, "Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting," *IEEE Transactions on Biomedical Engineering*, 1-1, 99, 2019. http://dx.doi.org/10.1109/TBME.2019.2915839

[8] K. M. Hosny, M. A. Kassen, M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with Alex-net," PLOS One, 17-20, 2019. http://dx.doi.org/10.1371/journal.pone.0217293

[9] N. Hameed, A. M. Shabut, M. A. Hossain, "Multi-Class Skin Diseases Classification Using Deep Convolutional Neural Network and Support Vector Machine," *Proceedings of 12th International Conference on Software, Knowledge, Information Management and Applications* (SKIMA), 23-30, 2019 http://dx.doi.org/10.1109/SKIMA.2018.8631525

[10] M. A. Albahar, "Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer," *IEEE Access*, 7, 2019. http://dx.doi.org/10.1109/ACCESS.2019.2906241

[11] A. Hekler, J. S. Utikal, A. H. Enk, "Superior skin cancer classification by the combination of human and artificial intelligence," *European Journal of Cancer*, 120, 114-121, 2019. http://dx.doi.org/10.1016/j.ejca.2019.07.019

[12] H. Kittler, H. Pehamberger, K. Wolff, M. Binder, "Diagnostic accuracy of Dermatoscopy," Lancet Oncology, 3(3), 159-165, 2002. https://doi.org/10.1016/s1470-2045(02)00679-4

[13] M. Maragoudakis, I. Maglogiannis, "Skin lesion diagnosis from images using novel ensemble classification techniques," in *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference* , 1–5, 2010. http://dx.doi.org/10.1109/ITAB.2010.5687620

[14] J. Kawahara, G. Hamarneh, "Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers," *in International Workshop on Machine Learning in Medical Imaging. Springer*, 164–171, 2016. http://dx.doi.org/10.1007/978-3-319-47157-0_20

[15] P. Tschandl, C. Rosendahl, H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, 5, 180161, 2018. http://dx.doi.org/10.1038/sdata.2018.161

[16] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples) ," *Biometrika*, 52(3–4), 591–611, 1965. https://doi.org/10.2307/2333709

[17] B. Ralph, D. E. Cureton, "Test of Normality Against Skewed Alternatives," *Psychological Bulletin*, 78, 262-265, 1972. 2 https://dx.doi.org/10.1037/h0033113

[18] M. A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*. , 69: 730–737, 1974. https://doi.org/10.2307/2286009

[19] W. H. Kruskal, W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of American Statistics Association*, 583–621, 1952. https://doi.org/10.2307/2280779

[20] N. Nachar, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, 4(1), 13-20, 2008. http://dx.doi.org/10.20982/tqmp.04.1.p013

[21] S. W. Scheff, "Chapter 8- Nano-parametric Statistics," *Fundamental Statistical Principles for the Neurobiologist-A Survival Guide*, 157-182, 2016.

[22] K. K. G. Lan, Y. Soo, C. Siu, M. Wang, "The use of Weighted Z-tests in Medical Research," *Journal of Biopharmaceutical Statistics*, 15(4), 625-639, 2005. http://dx.doi.org/10.1081/BIP-200062284

[23] H. Yusop, F. F. Yeng, A. Jumadi, A. Mahadi, M. N. Ali, N. Johari, "The Effectiveness of Excellence Camp: A study on Paired Sample," *Procedia Economics and Finance*, 31, 453-461, 2015. https://doi.org/10.1016/S2212-5671(15)01174-0

[24] H. Abdi, L. J. Williams, "Tukey's Honestly Significant Difference (HSD) Test", *Encyclopedia of Research Design*., 2010.

[25] D. Fraiman, R. Fraiman, "An ANOVA approach for statistical comparisons of brain networks," *Scientific Repository,* 8, 4746, 2018. https://doi.org/10.1038/s41598-018-23152-5