EasyChair Preprint

№ 11697

Comparative Analysis of Various Deep Learning Algorithms for Skin Cancer Disease

Rohan Kukreti, Shivam Chawla, Sarthak Rawat,
Sakshaat Dhiman and Gunjan Chhabra

January 6, 2024

Comparative analysis of various deep learning algorithms for Skin Cancer Disease

Rohan Kukreti,
(Student, Graphic Era Hill University)
rohankukreti07@gmail.com

Shivam Chawla,
(Student, Graphic Era Hill University)
sc2936892@gmail.com

Sarthak Rawat,
(Student, Graphic Era Hill University)
Sarthak.rawat1011@gmail.com

Sakshaat Dhiman,
(Student, Graphic Era Hill University)
sakshaatdhiman999@gmail.com

Dr.Gunjan Chhabra
(Associate Professor,Dept. of Computer Science, Graphic Era Hill University)

Abstract: ***These days, skin illness among people has been a common illness, particularly in millions of individuals are enduring with different sorts of skin based illness. As a rule, these maladies have covered up perils which lead to human not as it were need of self-confidence and mental discouragement but too a chance of skin cancer. Determination of these sorts of illnesses as a rule required therapeutic specialists with high-level disobedient due to a need of visual determination in skin infection pictures. Additionally, manual determination of skin malady is frequently subjective, time-consuming, and required more human exertion. Skin cancer is one of the most hazardous types of cancer impacting millions of lives on a daily basis. Skin cancer originates by uncorrected (DNA) inside skin cells causing genetic mutation in the skin. Skin cancer spreads slowly to body parts and is therefore easier to treat in its early stages; so it's best to catch it in its early stages. The increase in skin cancer, high mortality rates and high medical costs require early detection of its symptoms.***

Given the seriousness of these problems, scientists have developed many types of cancer at an early stage. Symmetry, colour, size, shape, etc. lesion parameters. It is used to identify skin cancers and distinguish between benign tumour's and melanoma. This article provides detailed information about deep learning techniques for early skin detection. Research articles published in reputable journals on the subject of cancer screening were examined. Research findings are presented in the form of tools, graphs, tables, methods and procedures for better understanding

***Index Terms*— Deep learning; deep; machine learning; support vector machine (SVM), convolutional neural networks, random forest classifier, naïve-bayes, skin disease.**

I. Introduction

Skin illness is a common issue in the world, which is affecting millions of people. These illnesses often lead to skin based issues and a higher risk of skin cancer. Detecting these illnesses often requires therapeutic specialists with high-level disobedience due to the need for visual determination in skin infection pictures. Manual determination is subjective, time-consuming, and requires more human effort. Skin cancer is a dangerous type caused by uncorrected DNA in skin cells, causing genetic or mutations. It spreads slowly and is easier to treat in its early stages. Early detection of symptoms is crucial due to the increase in skin cancer, high mortality rates, and high medical costs. Scientists have developed lesion parameters to identify skin cancers and distinguish between benign tumours and melanoma. This article provides detailed information about deep learning techniques for early skin detection.. Skin cancer is a significant public health concern, with increasing frequency rates and the importance of early detection for effective treatment.[1]. The field of skin cancer detection has seen significant advancements driven by innovative technologies and strategies. Skin cancer includes various types, including melanoma, basal cell carcinoma, and squamous cell carcinoma. Early detection and management are crucial for managing the disease, as early-stage analysis leads to higher survival rates and less intrusive treatment options.[2][3] The continuous development of diagnostic tools and strategies is essential to address the growing challenges related to the rising rate of skin cancer. Mechanical advancements, such as artificial intelligence, machine learning, and computer vision, have revolutionized restorative diagnostics, increasing precision and effectiveness through mechanized examination of meatoscopic images, side effect recognition, and risk evaluation. Non-invasive imaging advances and atomic diagnostics also hold promise for improving early skin cancer diagnosis, providing valuable insights into infection movement and personalized treatment strategies. This term paper looks for to fundamentally assess and synthesize the current state of skin cancer location techniques, shedding light on their qualities, impediments, and potential affect on clinical hone.[4] By investigating the crossing point of pharmaceutical and innovation, we point to contribute to the progressing talk on progressing symptomatic approaches, eventually encouraging the early recognizable proof and administration of skin cancer, and in this manner progressing persistent results.[5] To solve the problem, we implemented a approach based on transfer based learning for skin disease classification

II. Dataset, Methodology and Architecture

. ***Dataset*** This research paper explores the use of deep learning technology in automatic skin detection using HAM10000 data, a dataset of 10,000 high-resolution pigmented skin samples. The dataset contains images of various pigmented skin lesions, including melanomas and benign ones, and detailed descriptions. The research uses algorithms like convolutional neural networks (CNN) for image classification and transfer learning techniques to extract features from pre-trained models. The HAM10000 dataset is divided into training, validation, and testing stages to develop and evaluate the different models performance. The case study discusses the results obtained with deep learning models trained on the dataset, including measurements of accuracy, sensitivity, specificity, accuracy, and F1 score, to understand the model's performance in distinguishing benign and malignant skin lesions.[6]

A. Dataset Classification

DNN's play an important role in diagnosis of skin cancer. They consist of a group of nodes connected to each other. Their structure shows resemblance with the human brain in terms of neuron connections. Nodes work together to solve specific types of problems. These are trained for specific tasks; These networks then act as experts in the field they are researching. In this research, these networks are used to classify images and distinguish different types

. of cancer. Introduction to different types of skin diseases in the data classification by Age in Figure2.We searched for different techniques of learning, such as CNN, Decision Tree, Support Vector Machine, Random Forest Generator, Naïve Bayes, KNN and for systems based classification and detection of skin cancer disease. The Fig.1 shows dataset classification by Sex.

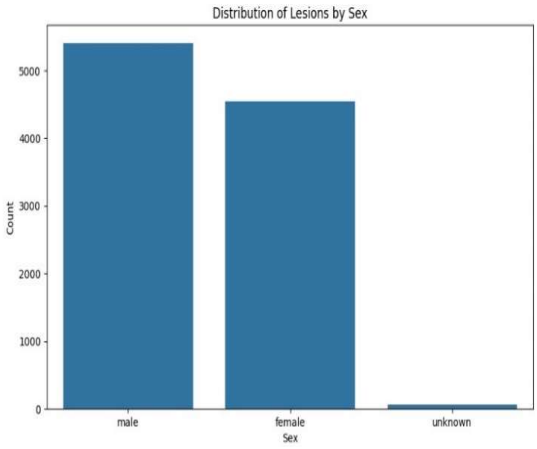

Fig1.Distribution of Lesions by Sex

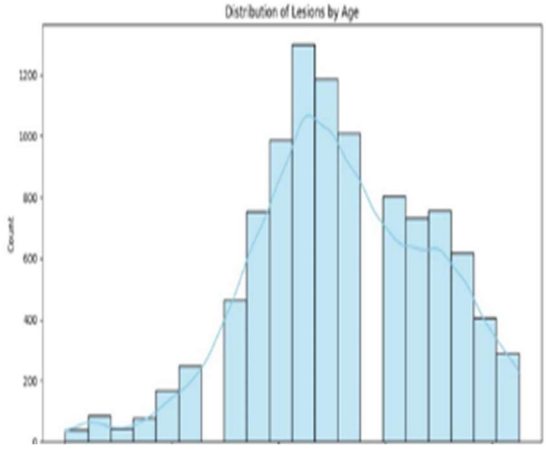

Fig2.Distributions of Lesions by Age

B. Methodology

. DNN's play an important role in diagnosis of skin cancer affected cells. It comprises of a group of layers connected to each other. Their neural network are meant for specific tasks; These networks then act as veterans of field they are researching. In our study, neural networks are used to distinguish images and differentiate various kinds in types of cancer. Introduction to different types of skin diseases in the HAM-10000(Human Against Machine) dataset in Figure 2.

We searched for different techniques of learning, such CNN, Decision Tree, Support Vector Machine, Random Forest Generator, Naïve Bayes, KNN[20] and for skin cancer detection systems. Research related to various models used in classification and identification of skin cancer disease using neural networks is discussed in detail in this section [11].

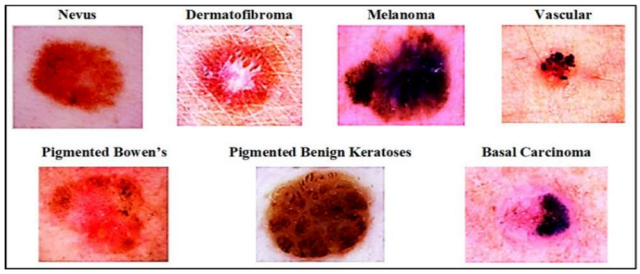

Fig 3.Dataset Image

a) **K-nearest neighbour algorithm**

Nearest neighbour classification is one of the simplest classification techniques in image space. Therefore, when

a test image is considered, the label of the closest point in the subject cluster is assigned based on the position in the image. By default, in KNN the distance between multiple points in an image is usually calculated using the Euclidean measure and a distance is assigned to each pixel. The Euclidean distance between two pixels is called "distance".

After extraction, the extraction is added directly to the product or machine learning tool and divided into two groups. The whole process has two stages. The first is used for training and the second is used for testing of images against the trained dataset[14-16].

Using real method to determine the optimal K value. Additionally, in this study, the T.C(time complexity) of the KNN is O(n^2),.

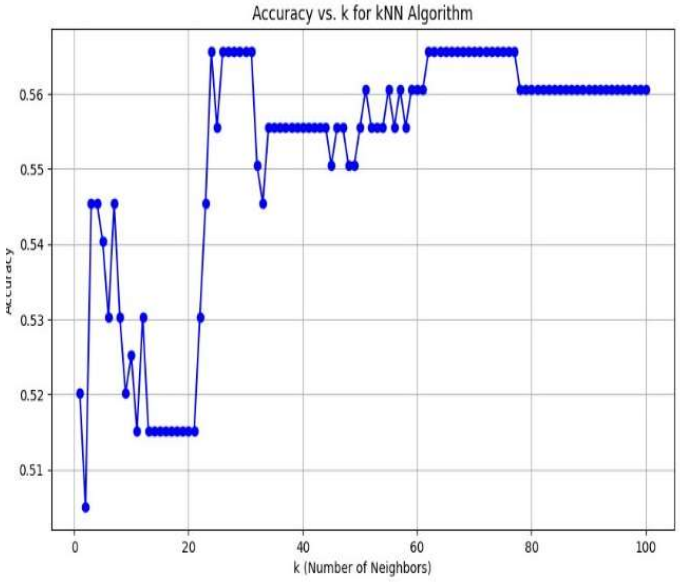

Fig4.KNN Results

. *b)* **Convolutional neural network**: CNN is an important type of deep learning model based on neural networks and is widely used in computer vision. It is used to split the image, combine the input image and create the image view. CNN is an optimal approach for collecting and learning data based on images, from collecting simple features consisting of curves and edges. The hidden layer of CNN consists of a convolutional layer, a nonlinear pooling layer, and an entire connection layer. A CNN may consist of many layers which are connected to each other where output of one layer is input of another layer.Majorly there are three main types of layers in CNN are convolutional layers, layered layers, and full layers. Automated CNN-based deep learning algorithms achieve superior performance in detection, segmentation and classification of medical images.[12][13]

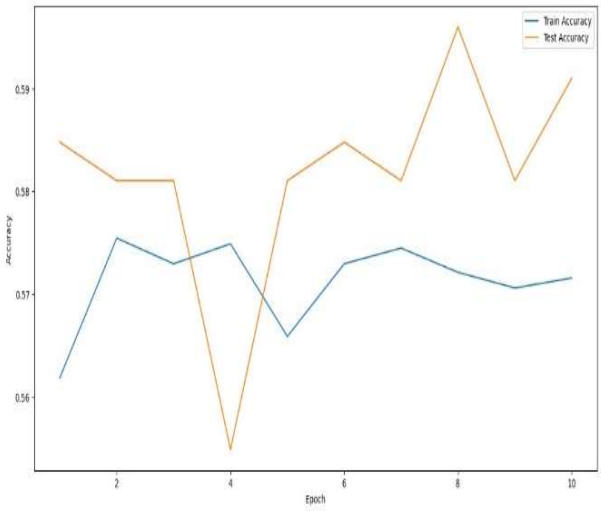

Fig5.CNN accuracy for different eposch values

c) **Decision Tree**

. Decision trees include a distributed monitoringsystem. This thouht comes from the classical model,that consists of roots,leaves and internal-nodes. Similarly, a decision tree consists of nodes that represent circles, while branches represent line sections connecting the nodes.The decision tree originates from Root Node and continues downwards, it usually moves from left to right.Root node is its starting point. The node that does not have any child node is called the "leaf" node.Multiple branches may extend from a leafless node. A node represents the attribute, while a branch represents the level of values. These values serve as breakpoints for a set of values for a particular property.[17][18]

. The integration of input-data into a decision tree model is based on the key characteristics of the given data.. Classification of clusters depends on best suited attributes that classify the data. Data objects are divided into sections based on the importance of these properties. This method is applied recursively to each partitioned subset of the data object. The process is terminated because all data items in the subset now belong to the same category.

. Random Tree :A random tree is a randomtree constructed from a set of possible trees, each with K random features. "Random" in this case means that each tree in the tree has an equal chance of being sampled. Or we could say that the tree has a "normal" distribution. Random trees can be constructed well, and combinations of large trees often produce the correct structure. In recent years, random trees have been intensively studied in machine learning.[19]

d) Random forest classifier:

Created by Leo Bierman ,a random forest is a set of random distributions or regression trees containing samples selected from the training data.

. During induction, random features are selected.Prediction is made by summing all the predictions (usually voting for distribution or average return). Each tree is planted as described in: if there are N problems that consist of the training set, by randomly sampling N, but making changes from the initial information. This model will be used as an

educational tool in tree planting. For the input variable Z, the variable z is chosen such that m< is specified at each node, we select z variables from Z and most optimal value is allocated to the m used. split node The value of z remains constant during the growth of the forest.[21][22]

Every tree grows as much as it can.

. No clipping is used. Random forests often show significant performance improvements compared to single tree classifiers. It produces general errors similar to Adaboost, but is more robust to noise.[23][24]

. Ross Quinlan developed the C4.5 algorithm for decision trees. The decision tree was created with version J48, C4.5 of the open source Java implementation of the WEKA data search tool. This is a standard decision tree algorithm. One of the classification algorithms in data search is decision tree inference. Classification algorithms use inductive learning to build models from predefined data sets. Each file is identified by an attribute or the value of the attribute. A classification can be viewed as a map from a set of criteria to a particular group.

e) **Support vector machine** (SVM) training is used to optimize cost allocation. A key advantage of SVMs is that they provide a common framework within which different machine learning architectures can be built by selecting kernels.[25] Reducing statistics and model risk is the principle used in SVM which minimizes one aspect of the overall error. After extracting the video, the extracted material is sent directly to a classifier, the machine learning tool, to be classified into two different classes.

b) The process has two phases: the training phase and the testing phase.[26]

c) Ø Training phase: Feature samples and class lists of positive and negative images are fed into the classifier for training.

d) Ø Test phase: Enter the unknown test,The information obtained in the training phase

. Will classify the unknown model.SVM classifier used for texture analysis features to predict skin inflammation. Classification is done by comparing each skin image in the test set with the skin image in the training set. SVM has many advantages over various classification models.

. Varying the K parameter from 1 to 10 means we ae distributing testing and training images randomly.

. *f)* **Naive Bayes** (NB),This classification model is based on a very popular conditional probability theorem which is called Bayes' Theorem which is why it is called "Naïve Bayes".Therefore, It works on the assumption that all features are conditionally dependent ,dataset features/labels cannot be learned one by one, this is easier and faster.Bayesian networks often have different sequences from the system. There are links between nodes that represent the relationship between them. The direction of the connection is from the cause If the connection is from node A to B, then node A is called the "parent" of B's "child". Nodes are defined by states, their properties can be continuous as well as discrete.

It works on Bayes Equation,which is stated as :

$$P(X|Y)=P(Y|X)*P(X)/P(Y),$$

Where X is the event to happen in the future,Y is the event that has happened in the past.

III. Result and Discussion

. Convolutional Neural Networks (CNNs) have proven to be especially effective in machine learning models for skin cancer diagnosis when compared to other models because of their ability to extract complex patterns and textures from images, which results in higher classification accuracy. On the other hand, conventional machine learning models such as the Random Forest Classifier provide excellent performance, particularly in terms of interpretability and computational efficiency. Relevant models include Decision Trees, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machines (SVM), each of which has a unique set of benefits based on the dataset and therapeutic needs. Although deep learning models require large datasets and computer power, conventional methods can be more effective with subtle feature extraction while using less resources overall. As a result, selecting the best model requires carefully weighing elements like accuracy requirements,

e) **Accuracy Score:** This metric denotes the proportion of correct predictions out of the overall predictions. It provides a comprehensive evaluation of the model's ability to detect disaster-affected areas from satellite imagery. CNN performs the best out of all the models with accuracy approx..72 %

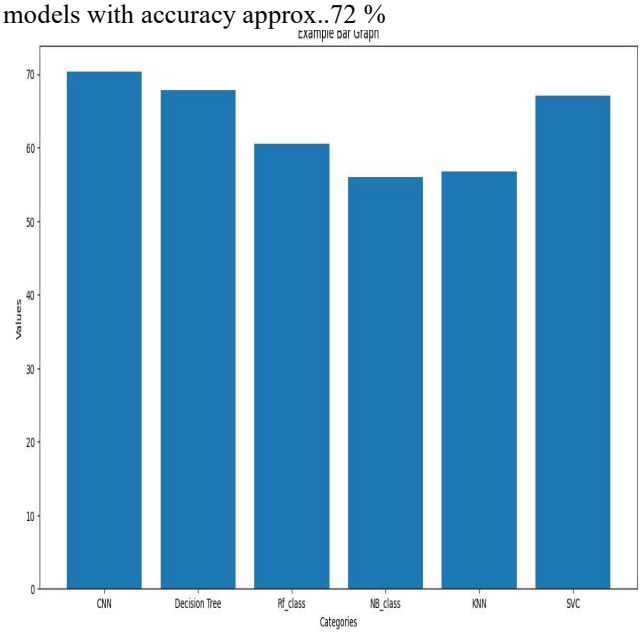

Fig 6 .Comparison of different models accuracy

F1 Score: It is defined as the harmonic mean of precision and recall score, the F1 Score offers a defined estimation of algorithms evaluation based on performance. It considers both the model's precision and recall, thereby making it a strong metric for assessing the effectiveness of

disaster detection. KNN has the highest out of all 66% approx..

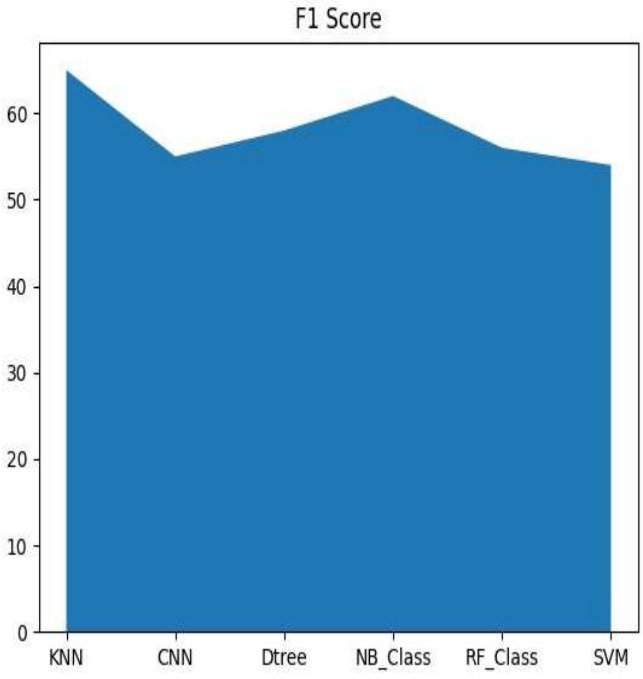

Fig7. F1 Score

f) **Precision Score:** This score can be mathematically defined as the ratio of accurate positive predictions to all predicted positive observations. It concentrates on relevance and precision of the model's positive predictions in disaster situations.

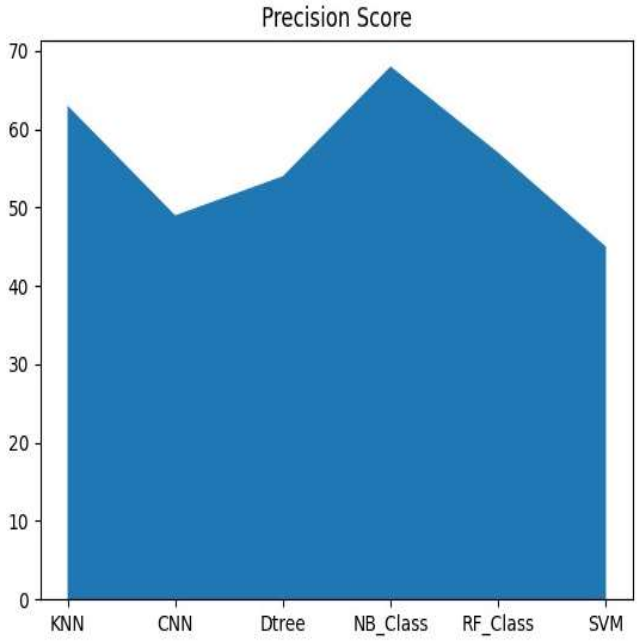

Fig8.Precision Score

g) **Recall Score:** This score indicates the ratio in actual positives accurately recognised by the model. It stresses the algorithms capacity in correctly recognize and classify areas affected by disasters .Decision tree has the highest out of all

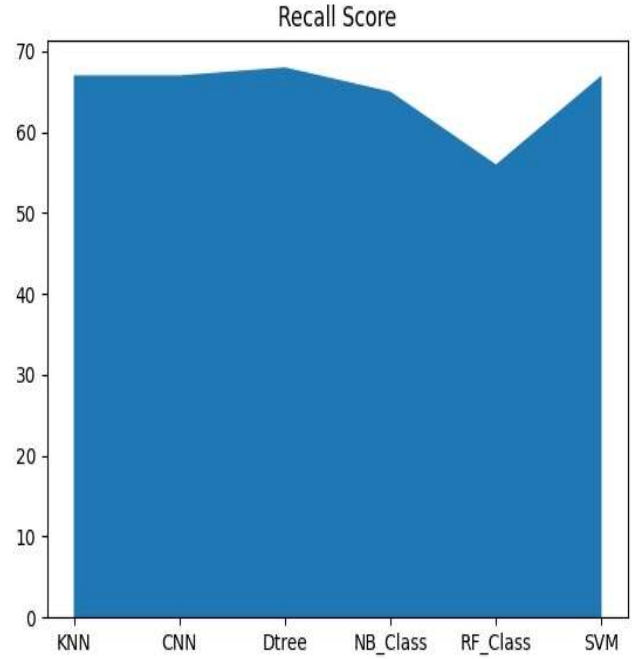

Fig9. Recall Score

h) In addition to these metrics, a confusion matrix was employed to visualize the models' performance in a tabular manner. This matrix shows the number of predictions which could be True Positive, True Negative ,False Positive or False Negative leading to a transparent view of the model's classification abilities.

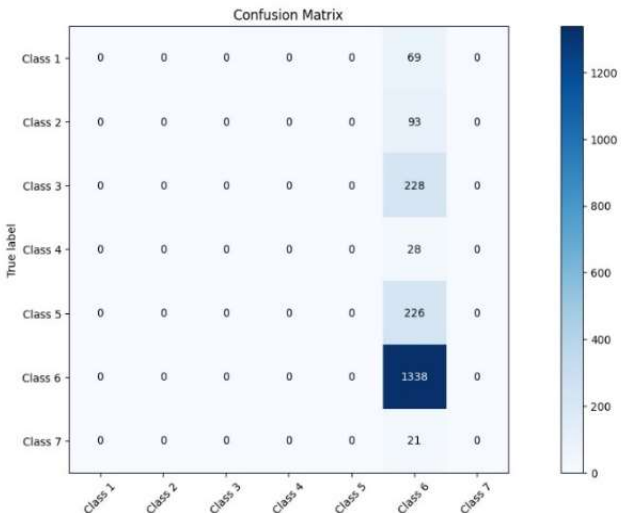

FIg 10.CNN Confusion Matrix

IV. CONCLUSION

i) All things considered, a careful analysis of metrics demonstrated that the Convolutional Neural Network (CNN) was the most effective approach for the given task. This algorithm is suggested for the given problem because of its outstanding performance, as evidenced by its high recall, accuracy, precision, and F1 score. Across several assessment criteria, the CNN model consistently outperformed other algorithms, despite their comparable performance. An extensive analysis of metrics led to the conclusion that the Convolutional Neural Network (CNN) was the best strategy for achieving the stated goal. This approach performs exceptionally well, as demonstrated by its high recall, accuracy, precision, and F1 score, making it recommended for the given problem. CNN consistently outperformed other models in terms of various performance metrcs

REFERENCES

[1] Orchid. 2015. ORCHID researchers provide Nepal earthquake response. Retrieved from: http://www.orchid.ac.uk/presses/orchid-researchers- provide-nepal-earthquake-response/

[2] Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., et al. (2019).“Creating xbd: A dataset for assessing building damage from satellite imagery,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 10–17

[3] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539.

[4] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324.

[6] Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied Logis- tic Regression (3rd ed.). John Wiley Sons. doi:10.1002/9781118548387.

[7] Cover, T., Hart, P. (1967). Nearest Neighbor Pattern Classifi- cation. IEEE Transactions on Information Theory, 13(1), 21-27. doi:10.1109/TIT.1967.1053964.

[8] Mather, P. M., Koch, M. (2011). Computer Processing of Remotely-Sensed Images: An Introduction (4th ed.). John Wiley Sons.

[9] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

[10] Garc´ıa, S., Luengo, J., Herrera, F. (2015). Data Preprocessing in Data Mining. Springer. doi:10.1007/978-3-319-10247-4.

[11] Bergstra, J., et al. (2012). Random Search for Hyper-Parameter Opti-mization. Journal of Machine Learning Research, 13, 281-305.

[12] Cutter, S. L. (Ed.). (2018). Oxford Research Encyclopedia of Natural Hazard Science. Oxford University Press.

[13] Sayeed, A., Marco, A. (Eds.). (2019). Real-time Data Analytics for Large Scale Sensor Data. Elsevier.

[14] Soden, R., Palen, L. (2018). Informating crisis: Expanding critical perspectives in crisis informatics. Proc. ACM Hum.-Comput. Interact., 2(CSCW), 162. doi:10.1145/3274457.

[15] James WD, Berger TG, Elston DM, Odom RB. Andrews' diseases of the skin: clinical dermatology. Saunders Elsevier;2006.

[16] Recognizing neglected skin diseases. World Health Org Glob. 2018 [cited 2020 Aug 26]. Available from: https://www.who.int/neglected_diseases/news/WHO-publishes-pictorial-training-guide-on-neglected-skin-disease/en

[17] Seth D, Cheldize K, Brown D, Freeman EF. Global burden of skin disease: inequities and innovations. PMC. 2018;6(3): 204–10.

[18] Kelbore AG, Owiti P, Reid AJ, Bogino EA, Wondewosen L. Pattern of skin diseases in children attending a dermatology clinic in a referral hospital in Wolaita Sodo, southern Ethiopia. BMC Dermatol. 2019;5(19):1–8.

[19] Nutten S. Atopic dermatitis: global epidemiology and risk factors. Karger AG; 2015.

[20] Tuchayi SM, Makrantonaki E, Ganceviciene R, Dessinioti C, Feldman SR, Zouboulis CC. Acne vulgaris. Nat Pub

[21] Gebauer K. Acne in adolescents. Aust Fam Physician. 2017;46(12):892–5.

[22] Acne vulgaris. Clinical Gate, 2015 [cited 2020 Mar 13]. https://clinicalgate.com/acne-vulgaris/

[23] Silverberg JI. Association between childhood atopic dermatitis, malnutrition, and low bone mineral density:

[24] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, “A

[25] Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction,” Journal of Applied Science and Technology Trends, vol. 1, no. 2, pp. 56–70, 2020.

[26] D. V. Patil and R. S. Bichkar, “A Hybrid Evolutionary Approach To Construct Optimal Decision Trees With Large Data Sets,” in 2006

[27] IEEE International Conference on Industrial Technology, Dec. 2006,

[28] Sethi, K., and Yoo, J. H., Design of multicategory, multifeature split decision trees using perceptron learning. Pattern Recognition, 27(7):939-947, 1994.

[29] Shafer, J. C., Agrawal, R. and Mehta, M. , SPRINT: A Scalable Parallel Classifier for Data Mining, Proc. 22nd Int. Conf. Very Large Databases, T. M

[30] Rissanen, J., Stochastic complexity and statistical inquiry. World Scientific,1989.