ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

CANCER AI

[1]A.Rayudu, [2]B.Ashwini, [3]B.Asritha, [4]A.Umashankar, [5]Mrs.Kalpana,

[1,2,3,4] U.G.Scholor, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.

[5]Assistant Professor, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.

Abstract—

Because few people know the warning symptoms of skin cancer and how to avoid getting it, it ranks high among the most hazardous illnesses. The shockingly high mortality toll from skin cancer has led some to term it a "fourth burden disease" on a global scale. As a result, cancer cannot progress beyond its first stages unless it is detected early. Using ML and IR methods, we identify and categorize multi-label skin cancer and use the best practices in this research. Nevertheless, preprocessing techniques help eliminate superfluous and irrelevant features from the label encoder, and standard features are used to scale the input variance unit and standardize the range of functionality. In addition, the HAM10000 metadata dataset was used to evaluate each classifier's performance using a variety of machine learning approaches. The seven distinct skin cancer types included in the HAM10000 metadata dataset were the subjects of the experimental study. According to the findings, SVM, DT, and GNB were the machine learning algorithms that outperformed the other classifiers in terms of accuracy. Skin cancer, ML, DT, GNB, Multi-class, Support Vector Machines

I. INTRODUCTION

An uncommon characteristic of this illness in the interim, skin cancer is frequent and affects many nations; it occurs in people, animals, and plants. Skin cancer is becoming more common, which is a major problem that affects people all over the globe. Worldwide, skin cancer ranks as the fourth leading cause of mortality. Its primary victims are the young and the old, yet it may impact anybody at any time [1]. Early detection and treatment with a specific kind of surgery are possible for this condition. Melanoma, basal cells, and squamous cells are only a few of its many varieties [2]. Melanoma is the most unexpected kind of cancer. Hair follicles, skin cells, and mucous membranes are all affected. The development of skin cancer is possible in almost all cases of skin injury. It develops when normal skin cells undergo a mutation and become cancerous, a condition that may strike almost anybody. Carcinomas are the medical terms for skin cancers, which may develop anywhere on the body.

High doses of ultraviolet (UV) radiation from the sun are known to cause skin cancer, according to studies [3]. Even though not everyone is aware of it, skin cancer affects a large number of individuals worldwide. The sun's harmful ultraviolet (UV) rays cause skin cancer. One kind of energy that leaves the sun and makes it to Earth is ultraviolet light, or UV rays. It comes in several forms and may affect any part of the skin. Although it can't be completely avoided, there are measures you may do to lessen the likelihood of developing unstable skin cancer cells. There is no denying the gravity and potential risk of skin cancer [4]. Being aware of the risk and taking measures to protect yourself against cancer waves are crucial. The illness has a high prevalence in the US [5]. The skin cancer foundation reports that in 2012, over 63,000 new cases of melanoma were diagnosed, making it the most dangerous form of skin cancer, while over millions of new cases of non-melanoma skin cancers (NMSC) were also diagnosed. When

ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

skin cells start to proliferate uncontrollably, it's called skin cancer. This is known as non-melanoma skin cancer (NMSC) when it occurs on the outermost layers of the skin, and melanoma when it occurs deeper in the dermis [6].

II. RELATED WORK

Researchers and scientists have been working on machine learning strategies to regulate the automated categorization of skin cancer using many classes and levels over the past twenty years. The authors Nazia Hameed et al. urged the use of deep learning and machine learning to improve methods for treating a wide variety of ailments, including skin cancer, handicap, and global distress. Classification algorithm performance has been able to identify improved outcomes for various skin lesions [7]. In order to control the average feature extraction from anecdotal pictures, A. Murugan et al. examined the implications of skin color cleansing and segmentation in human disorders. When compared to other methods, the combined outputs of SVM and RF show superior accuracy [8].
In their study on skin cancer, Carolina Magalhaes et al. suggested using machine learning to observe prospective infrared thermography approaches. According to research [9], using thermal parameter-based ensemble learning for skin cancer detection yields superior prediction accuracy when it comes to simplifying confusion matrices. Melanoma and other skin cell abnormalities may be caused by persistent glitches in deoxyribonucleic acid transmission, according to research by Mehwish Dildar et al. Color, form, symmetry, and other physical characteristics may help in the early diagnosis of cancer, which in turn helps regulate the severity of symptoms and medical therapy. Even though a lot of people have tried using different machine learning and deep learning approaches to identify skin cancer, the results haven't been great [10]. In an effort to provide light on the human condition, Yuheng Wang et al. examined cancer diagnosis tools that use polarization deep learning to create illustrations of the disease and its symptoms. On the other hand, when it comes to skin cancer, there are two types of lesions: malignant and benign [11]. According to Rashmi Patil et.al., melanoma is one of the most dangerous cancers to identify using machine learning approaches for patient categorization, which may be a real pain for patients. To solve the loss function and text processing melanoma tumor thickness classification issue, the suggested method used a CNN model [12]. The doctors' data is more accurate thanks to deep learning techniques, according to Ravi Manne et.al., which make it easy to separate skin cancers using convolution neural networks. Reducing misclassification of pictures and improving accuracy via weaknesses in deep learning approaches was achieved by the CNN model that was examined [13].

III. METHODOLOGY

With the use of machine learning algorithms as SGD Classifier, Logistic Regression, GNB, SVM, DT, and KNN, the proposed method seeks to effectively categorize data with multiple labels. Preprocessing, categorization, and assessment of performance are the three primary processes. Here is the suggested system framework, as shown in Figure 1.

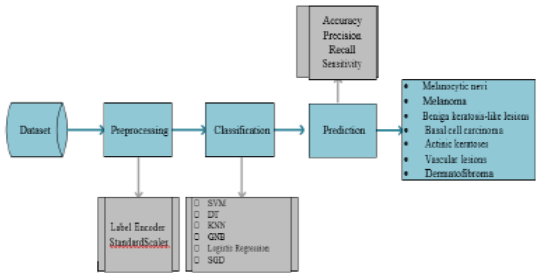


Fig 1: Proposed methodology to Skin cancer classification

A. Preprocessing

In order to get the most out of various machine learning algorithms, it's important that the data be clean. Each feature's coefficient uses a unique data preparation approach. In order for the machine learning approach to read the label data, the label encoder characteristics were used to transform the label data into numeric form. Also, by scaling to unit variance, we institutionalized the range of process in the input data and established using a conventional scalar component [13]. For multi-label data, we used several classification approaches, including SGD Classifier, Logistic Regression, KNN, GNB, and DT (Support Vector Machines). Classification with several labels SVM When it comes to classification

ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

and regression issues, SVM is a supervised technique that recycles. In order to solve issues involving binary classification, the SVM applies the same technique. It is possible to break down the multi-classification issue into its component parts. An often-used technique for applying multi-classification to the issue statement is the One vs. All (OVA) approach. The One vs. All method uses a hyperplane to partition the classes, with one half assigned to a single class point and the other half to all other points. According to Figure 2, the Greenline is the best option for maximizing the distance between the green point and every other point [14].

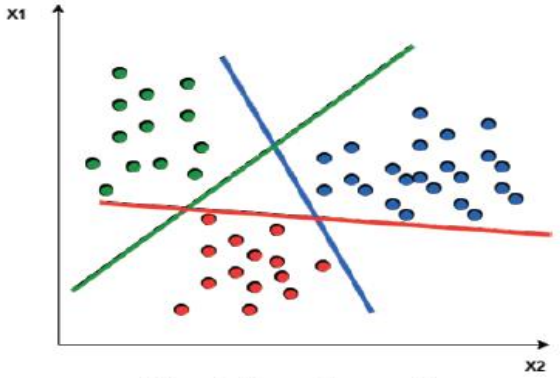
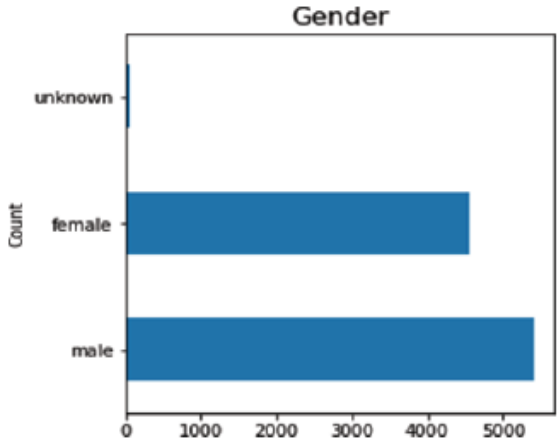


Figure 2: One vs All approach

Multi labels classification DT

When it comes to multi-label classification, the DT classifier is the way to go. It takes a tree-like structure and uses traits and features to formulate several queries. Every internal node in the root partitions the data into distinct records according to several attributes. According to [15], the data is divided into distinct categories by the leaves of the tree. Classification with several labels KNN K-Nearest Neighbors (KNN) is a classification method used in supervised machine learning. The data structure is not a determinant of the KNN algorithm's performance. By using the formula for the geometric mean, we can get the distance between the two feature vectors [16]. Classification with several labels GNB Bye's theorem is the basis of NB, a probabilistic ML approach used for classification. For training data, several functions are used to determine whether the data transit follows a normal or gaussian distribution. To use GNB, we must compute X's mean and variance in addition to substituting the normal distribution's probability thickness [17].Classification with several labels Regression using Logistic Classification is facilitated by LR, a supervised ML method. Using one-vs-rest, the LR method may be used to multiclass classification problems, which entails modifying the loss functions and converting the distribution of probabilities to a multinomial distribution [18]. Classification with several labels SGD's Filter SGD provides a simple and effective way to train linear classifiers and regressors using convex loss functions, including logistic regression and support vector machines. While SGD has been there for quite some time on the machine learning association, it is only very recently that it has garnered substantial interest within the context of large-scale learning [19]. C. Catalog Because it is easily accessible in the Kaggle repository, we repurposed the HAM10000 metadata data for this article. Data is kept in various ways and gathered from diverse populations. Researchers may observe on huge data solving ML problems using datasets, which are open-source machine learning databases. Machine learning models were trained and tested on two types of data from the dataset, with the former monitoring the real-time prediction of the characteristics provided by the latter [20]. Figure 3 displays the numerical distribution of many variables, including gender, age, cell type, and location. It is clear from the data's behavior that improper data grouping will lead to inefficient results when using ML algorithms on the data. Eighty percent of the dataset was used for training, while twenty percent was put to use for testing.



ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

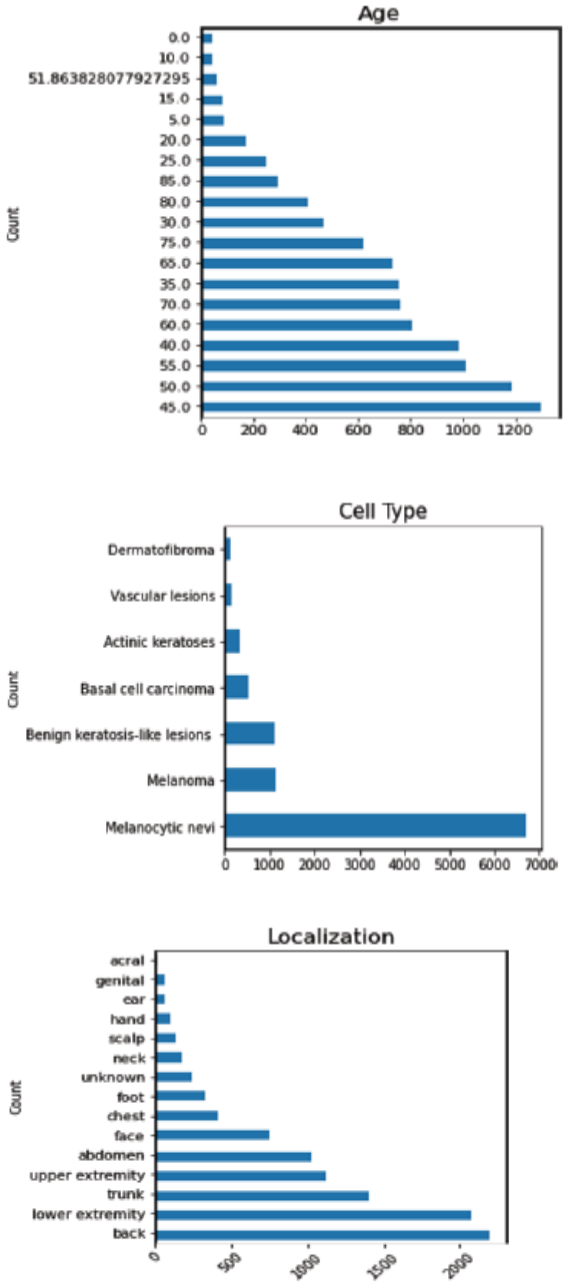


Fig 3: Distribution of different attributes values from the dataset

IV. RESULT AND DISCUSSION

Here you may find a variety of categorization models along with an analysis of each one. For the purpose of evaluating each classifier, the HAM10000 metadata dataset was subjected to six distinct machine learning approaches: support vector machine (SVM), decision tree (DT), k-nearest neighbor (GNB), logistic regression (SGD), and DT. Classifiers are only applied after the data has been normalized and standardized. In this experiment, several machine learning classifiers are trained and evaluated using the data. Training made use of 80% of the data, while testing made use of the remaining 20%. From Table 1 to Table 6, you can see the outcomes of several machine learning methods. What follows is a more in-depth explanation of label mapping:

0: Melanocytic Nevi (NV)
1: Melanoma (MEL)
2: Benign Keratosis-Like Lesions (BKL)
3: Basal Cell Carcinoma (BCC)
4: Intraepithelial Carcinoma / Bowen's Disease (AKIEC)
5: Vascular Lesions (VASC)
6: Dermatofibroma (DF)

Table 1: Multi labels classification SVM

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	96
2	1.00	1.00	1.00	228
3	1.00	1.00	1.00	37
4	1.00	1.00	1.00	1327
5	1.00	1.00	1.00	222
6	1.00	1.00	1.00	32
accuracy			1.00	2003
macro avg	1.00	1.00	1.00	2003
weighted avg	1.00	1.00	1.00	2003

Table 2: Multi labels classification DT

ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	96
2	1.00	1.00	1.00	228
3	1.00	1.00	1.00	37
4	1.00	1.00	1.00	1327
5	1.00	1.00	1.00	222
6	1.00	1.00	1.00	32
accuracy			1.00	2003
macro avg	1.00	1.00	1.00	2003
weighted avg	1.00	1.00	1.00	2003

Table 3: Multi labels classification KNN

	precision	recall	f1-score	support
0	0.85	0.82	0.83	61
1	0.86	0.85	0.86	96
2	0.91	0.99	0.95	228
3	0.94	0.43	0.59	37
4	0.96	0.99	0.98	1327
5	0.95	0.89	0.92	222
6	1.00	0.06	0.12	32
accuracy			0.95	2003
macro avg	0.92	0.72	0.75	2003
weighted avg	0.95	0.95	0.94	2003

Table 4: Multi labels classification GNB

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	96
2	1.00	1.00	1.00	228
3	1.00	1.00	1.00	37
4	1.00	1.00	1.00	1327
5	1.00	1.00	1.00	222
6	1.00	1.00	1.00	32
accuracy			1.00	2003
macro avg	1.00	1.00	1.00	2003
weighted avg	1.00	1.00	1.00	2003

Table 5: Multi labels classification Logistic Regression

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	96
2	0.97	1.00	0.98	228
3	1.00	0.38	0.55	37
4	1.00	1.00	1.00	1327
5	0.93	1.00	0.97	222
6	1.00	1.00	1.00	32
accuracy			0.99	2003
macro avg	0.99	0.91	0.93	2003
weighted avg	0.99	0.99	0.99	2003

Table 6: Multi labels classification SGD Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	0.00	0.00	0.00	96
2	0.69	0.36	0.47	228
3	0.00	0.00	0.00	37
4	0.93	0.98	0.96	1327
5	0.43	0.74	0.54	222
6	0.00	0.00	0.00	32
accuracy			0.80	2003
macro avg	0.44	0.44	0.42	2003
weighted avg	0.77	0.80	0.78	2003

A. Confusion Matrix

Figure 4 shows the confusion matrix for skin cancer prediction results across seven classes. The confusion matrix is a useful tool for visualizing the results of machine learning algorithms.In addition, the majority of datasets include data that is imbalanced and contain more irrelevant examples of data from other classes. By using confusion matrix techniques, the model can accurately predict all characteristics, resulting in the greatest possible accuracy score. Results inside the 100% range for accuracy, precision, recall, and F1-score were better for the SVM, DT, and GNB, as shown in the confusion matrix.

ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

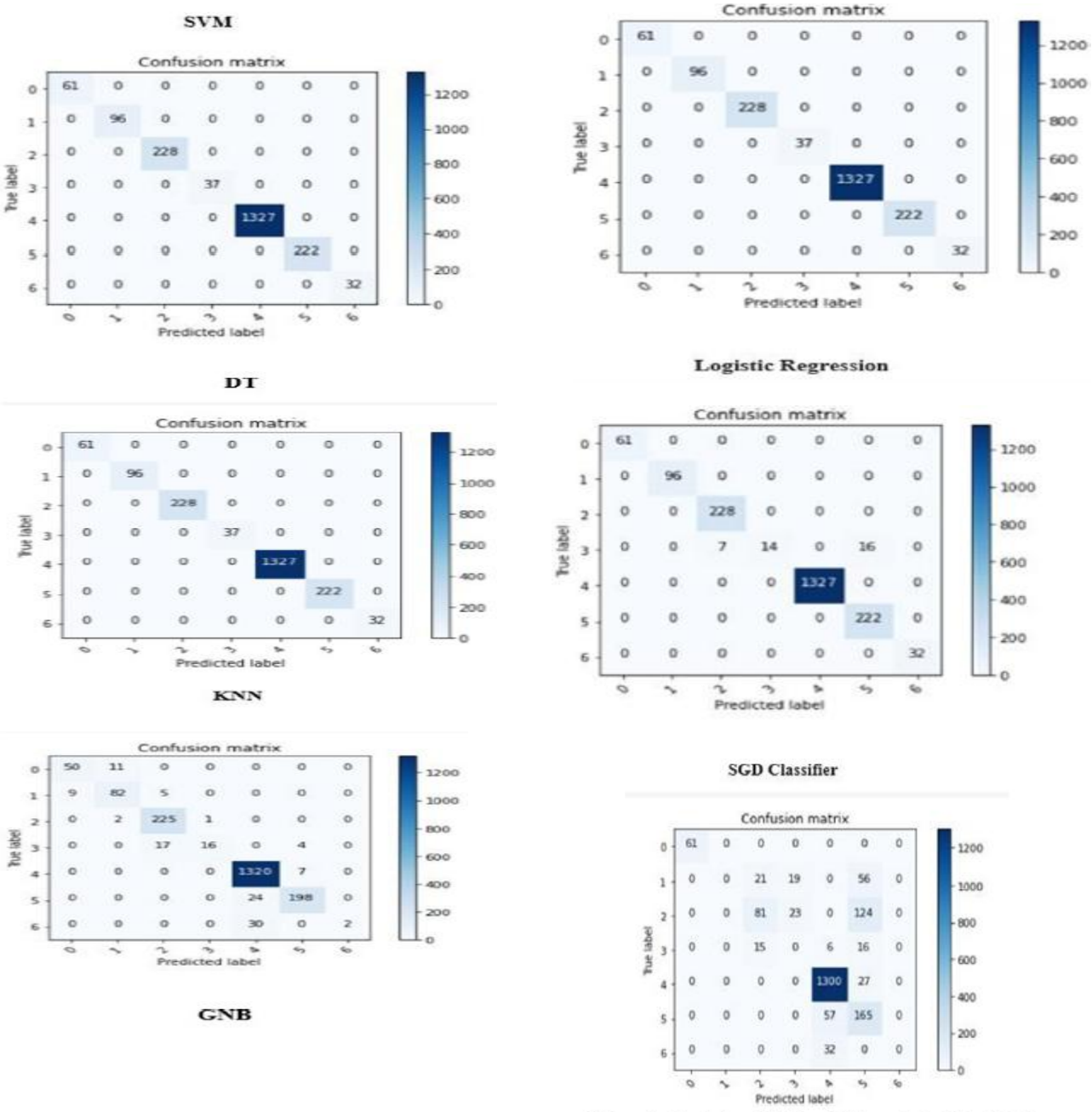


Figure 4: Confusion Matrix of different classifier (SVM, DT, KNN, GNB, Logistic Regression, SGD)

B. Overall Performance

Table 7 and Figure 5 reveal that, according to the results of the examination of all classifiers, the SVM, DT, and GNB achieved the best accuracy.

ISSN: 2456-1134 www.isjcresm.com
Vol-10 Issue-01 Mar 2025

Table 7: Comparison of all ML classifier

Machine Learning Classifier	Mean of Machine Learning Classifiers				
	Acc	Prec	Rec	F1-sco	Support
SVM	100%	100%	100%	100%	2003
DT	100%	100%	100%	100%	2003
KNN	95%	92%	72%	75%	2003
GNB	100%	100%	100%	100%	2003
LR	99%	99%	91%	93%	2003
SGD	80%	44%	44%	42%	2003

With perfect accuracy, precision, recall, and F1-Score, the SVM, DT, and GNB work well. With a recall of 91% and an F1-score of 93%, logistic regression has attained an accuracy of 99%. A 75% F1-score, 92% recall, 92% precision, and 95% accuracy were all attained by the KNN classifier. Using 80% accuracy, 44% precision, 44% recall, and 42% F1-score, the SGD classifier has achieved the minimal performance accuracy. According to Table 7 and Figure 5, the best performing algorithms among the aforementioned machine learning approaches were the gaussian naive bayes, decision tree, and support vector machine classifiers, which obtained the greatest accuracy among all classifiers.

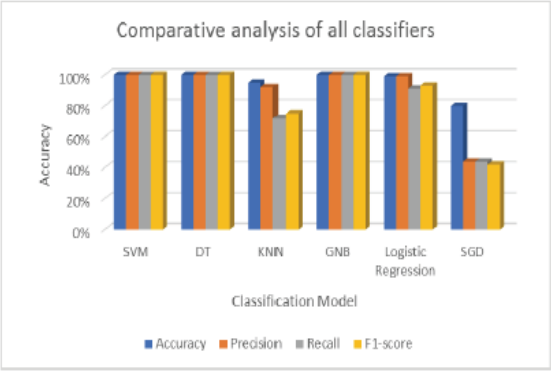


Figure 5: Comparison of all classifiers

V. CONCLUSION

Sadly, skin cancer is a yearly epidemic that affects a large number of people. Worldwide data shows that 5.4 million new cases of melanoma are recorded annually in the US, with a diagnosis rate of 53.3%. In this work, we compared seven different label datasets using the skin cancer HAM10000 metadata dataset and several machine learning approaches. When compared to other classifiers, SVM, DT, and GNB perform at the highest level of 100% accuracy. In comparison, KNN obtained 95% accuracy, Logistic regression 99% accuracy, and SGD reached 80% accuracy. We want to apply other categorization methods to evaluate this dataset's performance in the future.

REFERENCES

[1] A. Salam, F. Ullah, M. Imad, and M. A. Hassan, "Diagnosing of Dermoscopic Images using Machine Learning approaches for Melanoma Detection," in 2020 IEEE 23rd International Multitopic Conference (INMIC), 2020: IEEE, pp. 1-5.

[2] K. Das et al., "Machine Learning and Its Application in Skin Cancer", International Journal of Environmental Research and Public Health, vol. 18, no. 24, p. 13409, 2021. Available: 10.3390/ijerph182413409.

[3] E. Jana, R. Subban and S. Saraswathi, "Research on Skin Cancer Cell Detection Using Image Processing", 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2017. Available: 10.1109/iccic.2017.8524554.

[4] P. Dubal, S. Bhatt, C. Joglekar and S. Patil, "Skin cancer detection and classification", 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017. Available: 10.1109/iceei.2017.8312419.

[5] I. A. OZKAN and M. KOKLU, “Skin Lesion Classification using Machine Learning Algorithms”, Int J Intell Syst Appl Eng, vol. 5, no. 4, pp. 285–289, Dec. 2017.

[6] D. Wen et al., "Characteristics of publicly available skin cancer image datasets: a systematic review", The Lancet Digital Health, vol. 4, no. 1, pp. e64-e74, 2022. Available: 10.1016/s2589-7500(21)00252-1.

[7] N. Hameed, A. Shabut, M. Ghosh and M. Hossain, "Multiclass multi-level classification algorithm for skin lesions classification using

machine learning techniques", Expert Systems with Applications, vol. 141, p. 112961, 2020. Available: 10.1016/j.eswa.2019.112961.
[8] A. Murugan, S. Nair, A. Preethi and K. Kumar, "Diagnosis of skin cancer using machine learning techniques", Microprocessors and Microsystems, vol. 81, p. 103727, 2021. Available: 10.1016/j.micpro.2020.103727.
[9] C. Magalhaes, J. Tavares, J. Mendes and R. Vardasca, "Comparison of machine learning strategies for infrared thermography of skin cancer", Biomedical Signal Processing and Control, vol. 69, p. 102872, 2021. Available: 10.1016/j.bspc.2021.102872.
[10] M. Dildar et al., "Skin Cancer Detection: A Review Using Deep Learning Techniques", International Journal of Environmental Research and Public Health, vol. 18, no. 10, p. 5479, 2021. Available: 10.3390/ijerph18105479 [Accessed 17 August 2022].
[11] Y. Wang et al., "Deep learning enhances polarization speckle for in vivo skin cancer detection", Optics & Laser Technology, vol. 140, p. 107006, 2021. Available: 10.1016/j.optlastec.2021.107006 [Accessed 17 August 2022].
[12] R. Patil and S. Bellary, "Machine learning approach in melanoma cancer stage detection", Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 3285-3293, 2022. Available: 10.1016/j.jksuci.2020.09.002 [Accessed 17 August 2022].
[13] S. I. Ullah, A. Salam, W. Ullah, and M. Imad, "COVID-19 Lung Image Classification Based on Logistic Regression and Support Vector Machine," in European, Asian, Middle Eastern, North African Conference on Management & Information Systems, 2021: Springer, pp. 13-23.
[14] M. Imad, N. Khan, F. Ullah, M. A. Hassan, and A. Hussain, "COVID-19 classification based on Chest X-Ray images using machine learning techniques," Journal of Computer Science
and Technology Studies, vol. 2, no. 2, pp. 01-11, 2020.
[15] A. Hussain, M. Imad, A. Khan and B. Ullah, "Multi-class Classification for the Identification of COVID-19 in X-Ray Images Using Customized Efficient Neural Network", AI and IoT for Sustainable Development in Emerging Countries, pp. 473-486, 2022. Available: 10.1007/978-3-030-90618-4_23
[16] M. Imad, A. Hussain, M. Hassan, Z. Butt and N. Sahar, "IoT Based Machine Learning and Deep Learning Platform for COVID-19 Prevention and Control: A Systematic Review", AI and IoT for Sustainable Development in Emerging Countries, pp. 523-536, 2022. Available: 10.1007/978-3-030-90618-4_26
[17] M. Imad, F. Ullah, and M. A. Hassan, "Pakistani Currency Recognition to Assist Blind Person Based on Convolutional Neural Network," Journal of Computer Science and Technology Studies, vol. 2, no. 2, pp. 12-19, 2020.
[18] M. Imad, S. I. Ullah, A. Salam, W. U. Khan, F. Ullah, and M. A. Hassan, "Automatic Detection of Bullet in Human Body Based on X-Ray Images Using Machine Learning Techniques," International Journal of Computer Science and Information Security (IJCSIS), vol. 18, no. 6, 2020.
[19] M. Imad, S. I. Ullah, A. Salam, W. U. Khan, F. Ullah, and M. A. Hassan, "Automatic Detection of Bullet in Human Body Based on X-Ray Images Using Machine Learning Techniques," International Journal of Computer Science and Information Security (IJCSIS), vol. 18, no. 6, 2020.
[20] "Skin Cancer MNIST: HAM10000", Kaggle.com, 2022. [Online].Available: https://www.kaggle.com/datasets/kmader/skin-cancer-mnistham10000.