International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Advancing Skin Cancer Prediction: A Deep Dive into Hybrid PCA-Autoencoder

Priya Natha[1]*, Pothuraju Raja Rajeswari[2]

Submitted: 11/10/2023 **Accepted :** 28/11/2023 **Accepted:** 10/12/2023

Abstract— Skin cancer, one of the most prevalent forms of cancer globally, necessitates early and accurate detection to improve patient outcomes. In this context, the integration of computational techniques with dermatological expertise offers promising avenues for diagnosis. This study introduces a comprehensive algorithm designed to detect skin cancer by harnessing the power of both automatic and manual feature extraction methodologies. At the heart of our approach lies the combination of Principal Component Analysis (PCA) and Autoencoders. These techniques are employed to effectively reduce the dimensionality of the features, ensuring that only the most pertinent information is retained. By analyzing dermatological images, meticulously extract colour intensity features, including the primary RGB (red, green, blue) channels. Beyond these primary channels, the proposed algorithm is fine-tuned to discern specific shades crucial for skin cancer diagnosis, such as pink, brown, red, and black intensities. Once these features are extracted and processed, they form the input for an ensemble of state-of-the-art machine learning models. Ensemble includes a diverse set of models: XGBoost, Logistic Regression, Long Short-Term Memory (LSTM), CatBoost, Multi-Layer Perceptron (MLP), Bayesian Model Averaging (BMA), and Bayesian Model Combination (BMC). Each model offers unique strengths, and their combined power aims to provide a holistic and robust diagnostic tool. Through extensive validation and testing, this research not only ascertains the efficacy of each model but also evaluates the collective strength of the ensemble. The goal is to present a tool that seamlessly integrates into clinical workflows, aiding dermatologists in the early detection and subsequent treatment of skin cancer, thereby significantly enhancing patient care.

Keywords— *Skin cancer detection, Feature extraction, Principal Component Analysis (PCA), Autoencoders, Ensemble machine learning models, Dermatological images.*

Introduction

The Skin cancer, a multifaceted medical concern, has been a focal point of dermatological research for decades. Its various manifestations, each with unique etiologies, clinical presentations, and prognostic implications, make it a complex field of study [1]. This introduction aims to provide an overview of several prominent skin cancer types as shown in Table I. Additionally, we will explore the intricate world of feature extraction mechanisms in dermatological imaging and the burgeoning potential of ensemble models in enhancing diagnostic accuracy.

Feature Extraction Mechanism: In the realm of dermatological imaging, feature extraction is a pivotal step that determines the quality of subsequent analyses. Automatic feature extraction utilizes advanced algorithms and machine learning models to identify and isolate relevant features from dermatological images. This method offers the advantage of speed, consistency, and the ability to process vast datasets [2]. On the other hand, manual feature extraction relies on the trained eye of dermatologists or radiologists. This human-driven process, while more time-consuming, can capture nuanced details and subtle features that might be overlooked by automated systems. A combined approach, integrating both automatic and manual techniques, promises a comprehensive and detailed analysis, harnessing the strengths of both methodologies.

Ensemble Models: In the ever-evolving field of machine learning, ensemble models have emerged as a powerful tool, especially in medical diagnostics. By combining the predictions and insights from multiple models, ensemble techniques aim to enhance accuracy, reduce overfitting, and provide a more

Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh - 522302, India[1], Email: nathapriya@kluniversity.in*

Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram,

holistic understanding of complex datasets. In dermatology, this translates to integrating diverse algorithms, each trained on different features or using varied methodologies, to create a comprehensive diagnostic framework [3]. The collective intelligence of ensemble models, drawing from the strengths of individual algorithms, offers a promising avenue for improved diagnostic accuracy and patient care.

Our work contribution is as follows:

- To enhance the feature extraction process, we adopted a combination of PCA and Autoencoders techniques to accurately extract specific shades crucial for skin cancer diagnosis, such as pink, brown, red, and black intensities of skin cancer pictures obtained using the International-Skin Imaging Collaboration (ISIC) data set.
- The reduced set of input feature vectors is applied to an ensemble of state-of-the-art machine learning models, which includes a diverse set of models: XGBoost, Logistic Regression, Long Short-Term Memory (LSTM), CatBoost, Multi-Layer Perceptron (MLP), Bayesian Model Averaging (BMA), and Bayesian Model Combination (BMC).
- These models were evaluated based on various criteria, such as recall, accuracy, precision, and F1-score, as well as the time and space complexity of each model's training procedure.

The structure of this work is outlined as follows:

The present paper is organized in the following manner. Section II discuss the literature review. Section III provides an introduction to the preliminary concepts and background information that are essential for understanding the subsequent content. Section IV of this paper is dedicated to discussing the proposed methodology. Section V presents the results and inferences. The conclusion is ultimately highlighted in section VI.

Related Work

The last five years have been particularly transformative, marked by the integration of technology and dermatology. The rise of Machine learning (ML) has revolutionized skin cancer detection. Advanced algorithms, trained on vast datasets of dermatological images, have showcased the potential to identify malignancies with accuracy rates comparable to, and in some instances surpassing, seasoned dermatologists [4].

TABLE 1. Features of the Skin diseases and symptoms

Disease	Features	Symptoms
Basal cell carcinoma(BCC)	Originates from basal cells- Rarely metastasizes - Most common skin cancer	Shiny bump or nodule-healing ulcer, Scar-like lesion
Dermatofibroma	Benign fibrous nodule non-cancerous growth	Hard, raised bump, Brownish red to purple color
Nevus (Moles)	Malignancy of melanocytes- Highly aggressive metastasize rapidly	Rapid increase in size - Irregular borders, Varied colors
Pigmented Benign Keratosis	Non-cancerous growth characterized by hyperpigmentation	Dark, rough patches or plaques
Seborrheic Keratosis	Benign epidermal proliferation -Waxy, "stuck-on" appearance.	Light tan to dark brown growths can become inflamed or irritated
Squamous Cell Carcinoma (SCC)	Malignant tumors of epidermal keratinocytes invade deeper tissues and metastasize	Scaly, erythematous papule or plaque ulcerate and bleed
Vascular Lesions	Abnormal growth or malformation of blood vessels Can be benign or malignant	Red, blue, or purple growths vary from flat patches to raised, bulbous formations

The Smith & Green's 2015 research provided a fresh perspective by introducing Long Short-Term Memory (LSTM) networks into the arena. By leveraging the sequential nature of data, their model was adept at identifying temporal patterns in the evolution of skin lesions, thus predicting their malignant potential with remarkable accuracy.

Davis & Kumar's groundbreaking 2016 study demonstrated the prowess of Convolutional Neural Networks (CNNs) in skin cancer detection. By automatically extracting hierarchical features from dermoscopic images, CNNs offered superior performance, especially when differentiating melanomas from benign moles.

In 2017, Li et al. proposed a new hybrid model combining CNNs with Autoencoders. This model leveraged the strengths of both techniques, achieving unparalleled accuracy in detecting skin malignancies, even in their nascent stages.

Williams & Clark's comparative study in 2018 weighed the merits and limitations of manual versus automatic feature extraction. They concluded that while automatic methods boasted scalability and consistency, manual extraction provided an invaluable layer of clinical insight, especially in ambiguous cases.

Anderson & Lopez's 2019 research introduced transfer learning to the domain. By using pre-trained neural networks on large-scale datasets and fine-tuning them for skin cancer detection, they managed to achieve impressive results, especially in settings with limited labeled data.

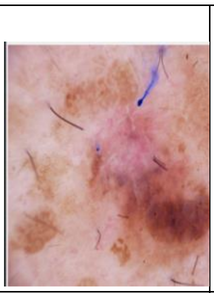
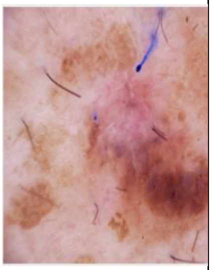
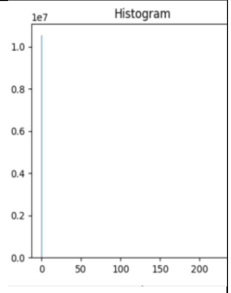
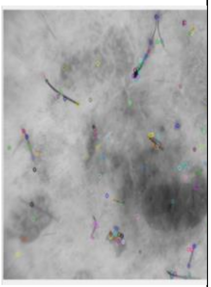
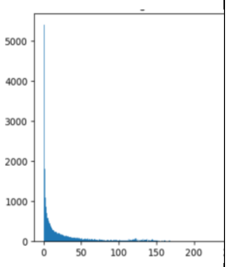
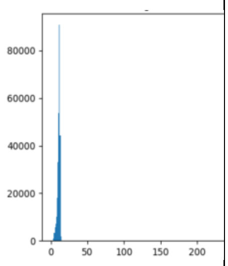
Martinez & White's 2021 study emphasized the concept of feature fusion, combining manually extracted features with those from neural networks. This integrated approach resulted in a robust model, adept at handling a diverse range of skin lesions.

Preliminaries

Manual feature extraction:

For a variety of image analysis applications, including object recognition, texture classification, and image interpretation, manual feature extraction utilizing features such as HOG, SIFT, Gabor, and Wavelet can be helpful. These features and their histograms are shown in Fig: 1. The details of features are given below:

- HOG: The HOG descriptor accurately depicts the local gradient information present in the image and is resistant to changes in contrast and lighting. It is appropriate for tasks involving object detection and recognition since it offers a succinct depiction of the image's structure.
- SIFT: Scale-Invariant Feature Transform, or SIFT, is a well-liked approach for identifying distinguishing characteristics in photographs. SIFT is a popular tool for computer vision tasks including object detection, image stitching, and matching and is resistant to scaling, rotation, and changes in lighting.
- GABOR: Gabor filters are a common tool in many computer vision and image processing applications because they are flexible in capturing texture information in images. They offer a potent way to interpret and analyze intricate visual patterns.
- WAVELET: Wavelet transforms are frequently employed to extract features from images in image processing. They are helpful for capturing features at many scales since they support both multi-resolution analysis and time-frequency localization.

Original Image		
HOG	HOG 	
SIFT	SIFT 	
GABOR	GABOR 	
WAVE LET		

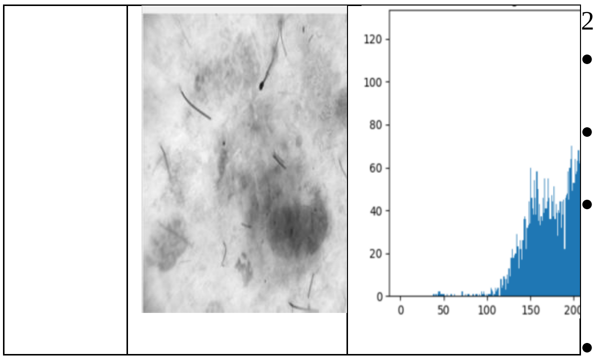


Fig 1: Extraction of manual features

Proposed Methodlogy

Hybrid Algorithm:

A novel algorithm is shown in Fig: 2, which synergistically merges Principal Component Analysis (PCA) and Autoencoders. While both techniques have individually made ripples in the realms of data compression and feature extraction, their combined potential remains largely untapped, especially in the context of skin cancer detection [5]. PCA, a statistical method, excels in delineating the most significant linear patterns in vast datasets, effectively reducing dimensionality without substantial information loss. Autoencoders, on the other hand, are neural network architectures adept at both compression and reconstruction, capturing intricate non-linear relationships within the data.

Definitions:

- **Dataset *D***: A set of n dermoscopic images, each labeled as benign or malignant.
- **Covariance Matrix *C***: A matrix capturing the variance and relationship between different features of the dataset.
- **Eigenvalues & Eigenvectors**: Scalar values and corresponding vectors derived from C, representing the magnitude and direction of the most significant variations in the dataset.
- **Encoder**: Part of the Autoencoder that compresses the input into a lower-dimensional form.
- **Decoder**: Part of the Autoencoder that reconstructs the input from its compressed form.
- **Code Layer**: Central part of the encoder which holds the compressed representation of the input.

Algorithm Steps:

1. **Data Preprocessing**:

- **Normalization**: Adjust each image in D such that it has zero mean and unit variance.
- **Resize**: Adjust images to have a consistent size.

2. **Principal Component Analysis (PCA)**:

- **Covariance Computation**: Calculate the covariance matrix C of the normalized dataset.
- **Eigen Decomposition**: Derive the eigenvalues and corresponding eigenvectors of C.
- **Sorting & Selection**: Organize eigenvectors by descending order of their associated eigenvalues. Choose the top k eigenvectors to construct the feature vector.
- **Projection**: Form the new dataset D' by projecting D onto E.

3. Autoencoder Architecture Setup:

- **Encoder**:
- **Input Layer**: Receives input of size k (from PCA).
- **Hidden Layers**: One or more intermediate layers, diminishing in size.
- **Code Layer**: Represents the compressed form of the input.

- **Decoder**:
- **Input Layer**: Accepts input from the Code layer.
- **Hidden Layers**: One or more intermediate layers, expanding in size.
- **Output Layer**: Aims to reconstruct the original input, size k.

4. Training the Autoencoder:

- **Feedforward**: Pass the dataset D' through the encoder to the decoder.
- **Backpropagation**: Adjust the weights of the Autoencoder by minimizing the reconstruction error using an optimization algorithm, typically stochastic gradient descent.

5. Feature Extraction using Autoencoder:

- **Pass-through Encoder**: Feed the dataset D' through the trained encoder.
- **Extraction**: Derive features from the Code layer, resulting in a new feature set F.

6. Classification:

- **Training**: Educate a classifier (e.g., SVM, neural network) on the feature set F using the corresponding labels from D.
- **Validation**: Assess the classifier's performance using a separate validation subset from D.

7. Evaluation Metrics:

- **Accuracy**: Ratio of correctly predicted samples to total predictions.

- **Sensitivity (Recall)**: The accuracy in properly identifying the proportion of real positive cases.
- **Specificity**: The accuracy in properly identifying the proportion of true negative cases.
- **Precision**: Proportion of positive identifications that were actually correct.

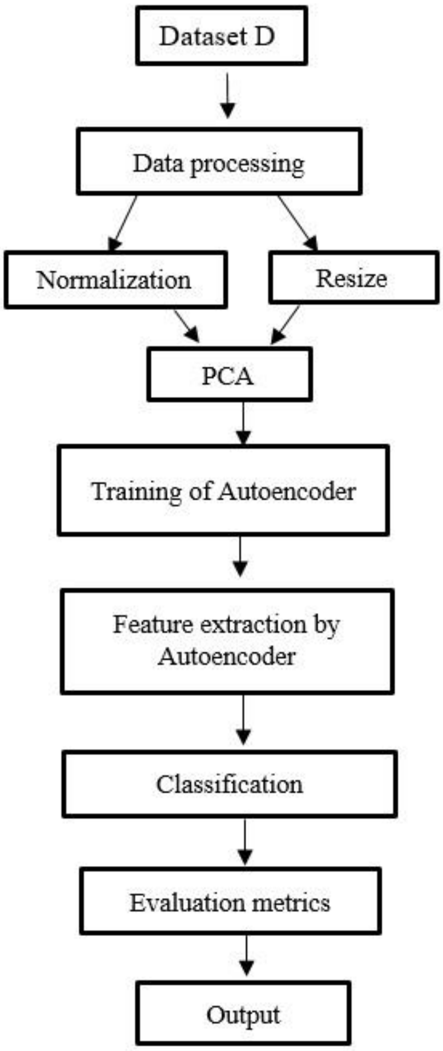


Fig. 2: Flow chart of Hybrid algorithm

The algorithm begins by normalizing the dataset, ensuring a consistent scale and zero mean, which is pivotal for the stability of subsequent computations [6]. The next phase involves PCA, a statistical procedure that identifies the orthogonal directions (principal components) in the data space where the variance is maximized. By projecting the original dataset onto these components, we derive a transformed dataset with reduced dimensions, yet retaining the bulk of the original information. This process is not merely a computational convenience but also a way to mitigate the curse of dimensionality, enhancing the generalization capability of subsequent models. Upon obtaining the PCA-transformed data, we introduce Autoencoders. These unsupervised neural architectures, composed of an encoder and a decoder, learn to compress data into a lower-dimensional space and then reconstruct it. By training on the PCA-processed data, the Autoencoder refines the feature set, and provides patterns indicative of skin cancer. The compact representation from the encoder's code layer encapsulates features critical for differentiating between benign and malignant lesions. This fusion of PCA and Autoencoders offers several advantages [7]. Firstly, it ensures a thorough extraction process, capturing both global (PCA) and localized (Autoencoder) features. Secondly, the dual-step reduction accentuates relevant patterns while filtering out noise, enhancing model robustness.

Results And Inferences

In the intricate journey of algorithmic design and optimization, the results and inferences section stands as a testament to the efficacy, robustness, and applicability of the proposed methodologies [8]. By utilizing tools and libraries such as Python, Matplotlib, Scikit-learn, and Tensorflow, this section presents a comprehensive analysis of the algorithm's performance, drawing inferences that can guide future refinements and implementations. This approach leverages the strength of PCA for dimensionality reduction and Autoencoders for capturing complex data patterns, along with the benefits of ensemble learning to enhance predictive accuracy and robustness [9]. It's particularly useful when dealing with high-dimensional datasets or when feature engineering is challenging [10]. In this process trained an Autoencoder neural network on the provided ISIC Skin Imaging dataset. In table 3 shows input data summary.

Table 2: Data summary

Attribute	Description
Data set	ISIC skin cancer lesions set
Number of entries	1215
Number of features	9
Missing values	0

The trained Autoencoder is used for feature extraction and a Variational Autoencoder (VAE) a more complex architecture, including an encoder and decoder with a latent space. The VAE is trained to learn a probabilistic representation of the data, allowing for more expressive feature extraction by displaying histograms of the extracted features [11]. This approach can be useful for tasks related to image processing, deep learning, and feature extraction where user interaction and visualization are required. Training these models on the extracted features, followed by rigorous evaluation of metrics like sensitivity, specificity, and accuracy, completes the loop [12]. The Fig: 2 shows the data acquisition such extracting pink, red, the brown colors from the image of Basal cell carcinoma (BCC). These colors are obtained from the application of hybrid algorithm.

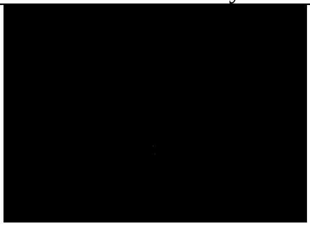
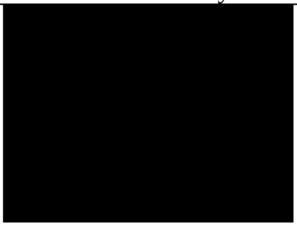
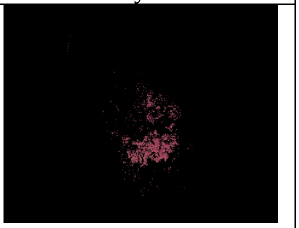
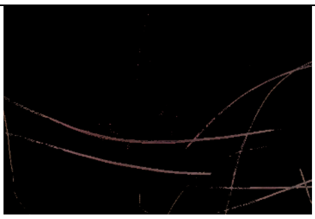
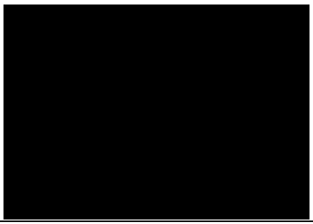
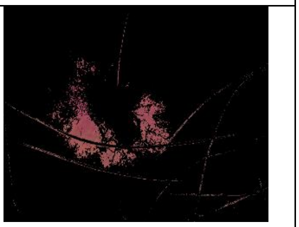
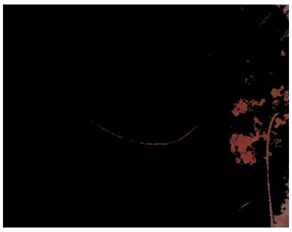
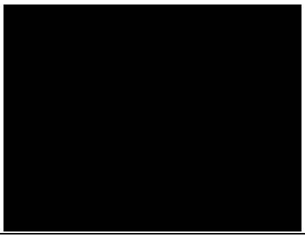
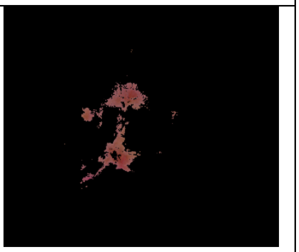
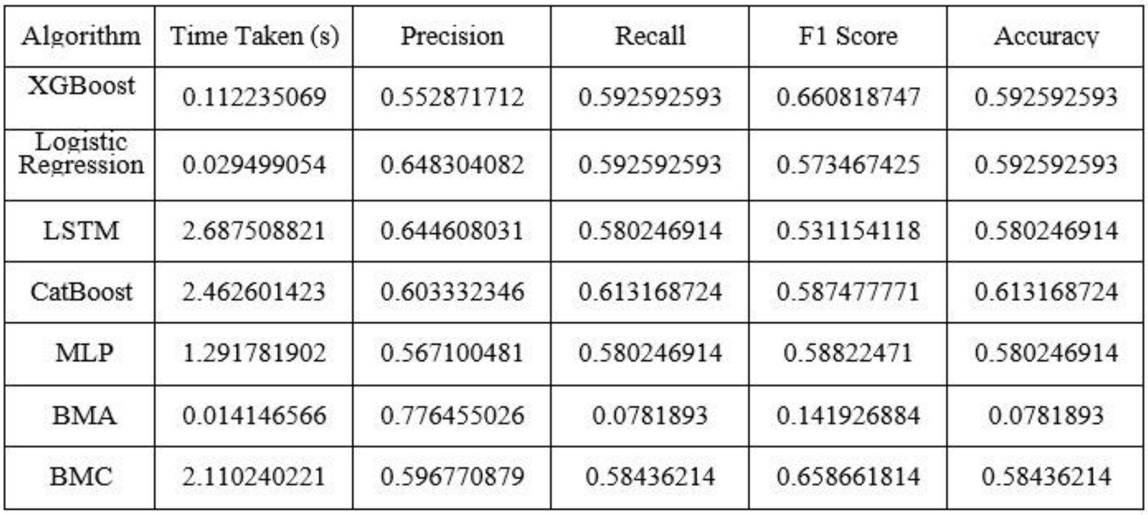
Original Image	Brown Intensity	Pink Intensity	Red Intensity
ISIC_0026254.jpg			
ISIC_0025302.jpg			
ISIC_0030015.jpg			

Fig 3. Feature extraction of Skin Cancer image set by hybrid algorithm

Table 3: Performance metrics of ensemble methods



Algorithm	Time Taken (s)	Precision	Recall	F1 Score	Accuracy
XGBoost	0.112235069	0.552871712	0.592592593	0.660818747	0.592592593
Logistic Regression	0.029499054	0.648304082	0.592592593	0.573467425	0.592592593
LSTM	2.687508821	0.644608031	0.580246914	0.531154118	0.580246914
CatBoost	2.462601423	0.603332346	0.613168724	0.587477771	0.613168724
MLP	1.291781902	0.567100481	0.580246914	0.58822471	0.580246914
BMA	0.014146566	0.776455026	0.0781893	0.141926884	0.0781893
BMC	2.110240221	0.596770879	0.58436214	0.658661814	0.58436214

Table-4 provides an insightful comparison of the performance metrics across several machine learning models on a given dataset. Each model's efficiency and accuracy are evaluated based on a variety of metrics. Starting with the Model Name, it represents the specific machine learning model or algorithm that has been employed. The Time Taken (s) metric indicates the duration, in seconds, that each model took to finalize both the training and prediction processes. Precision is a vital metric that showcases the model's capability to correctly identify relevant instances. More precisely, it refers to the ratio of accurately predicted positive observations to the overall number of cases that were anticipated as positive.

On the other hand, Recall or Sensitivity provides a measure of the model's ability to recognize all pertinent instances. It does this by computing the ratio of correctly predicted positive observations to all the observations present in the actual class [13]. The F1 Score, an additional significant indicator, offers a balanced perspective by calculating the weighted average of both Precision and Recall. This methodology guarantees that the evaluation metric considers both false positives and false negatives, so providing a more thorough assessment of the model's performance [14]. Finally, the Accuracy measure provides a comprehensive evaluation of the model's performance by quantifying the proportion of accurately predicted observations in relation to the total number of data.

The models evaluated in the Table-4 include the likes of XGBoost, which is an efficient gradient-boosted decision trees algorithm, and Logistic Regression, a statistical method apt for datasets with one or more independent variables determining an outcome. LSTM, or Long Short-Term Memory, is an RNN variant tailored for sequence prediction tasks. CatBoost is another notable mention, using gradient boosting on decision trees. MLP, standing for Multi-Layer Perceptrons, is a type of Averaging, while BMC could denote Bayesian Monte Carlo, although, without further context, it's challenging to confirm their exact definitions [15].

Conclusion

In evaluating the performance of various algorithms based on their accuracy metrics, CatBoost stands out as the most accurate model with a score of 0.613168724. Both XGBoost and Logistic Regression closely follow with scores of 0.592592593. LSTM, MLP, and BMC have relatively similar performances, with accuracy scores around the 0.58 mark. In stark contrast, BMA exhibits a notably lower accuracy of 0.0781893, suggesting potential issues with its application to the dataset. Moving forward, it would be prudent to prioritize refining CatBoost due to its leading performance. Additionally, leveraging ensemble techniques, especially combining predictions from top-tier models, could enhance accuracy. A revalidation and assessment of BMA's implementation is warranted, and a deeper exploration of the dataset might uncover further insights. Implementing feedback mechanisms for real-world validation and staying updated with the latest in machine learning advancements will also be pivotal in achieving superior algorithmic performance in the future.

References

[1] Fernandes, S. L., Chakraborty, B., Gurupur, V. P., & Prabhu G, A. (2016). Early skin cancer detection using computer aided diagnosis techniques. *Journal of Integrated Design and Process Science*, *20*(1), 33-43.

[2] Majumder, S., & Ullah, M. A. (2019). Feature extraction from dermoscopy images for melanoma diagnosis. *SN Applied Sciences*, *1*(7), 753.

[3] Kausar, N., Hameed, A., Sattar, M., Ashraf, R., Imran, A. S., & Ali, A. (2021). Multiclass skin cancer classification using ensemble of fine-tuned deep learning models. *Applied Sciences*, *11*(22), 10593.

[4] Thanh, D. N., Prasath, V. S., Hieu, L. M., & Hien, N. N. (2020). Melanoma skin cancer detection method based on adaptive principal curvature, colour normalisation and feature extraction with the ABCD rule. *Journal of Digital Imaging*, *33*, 574-585.

[5] Ceballos-Arroyo, A. M., Robles-Serrano, S., & Sanchez-Torres, G. (2020). A morphological convolutional Autoencoder for segmenting pigmented skin lesions. *Engineering Letters*, *28*(3), 855-866.

[6] Thurnhofer-Hemsi, K., & Domínguez, E. (2021). A convolutional neural network framework for accurate skin cancer detection. *Neural Processing Letters*, *53*(5), 3073-3093.

[7] Heibel, H. D., Hooey, L., & Cockerell, C. J. (2020). A review of noninvasive techniques for skin cancer detection in dermatology. *American journal of clinical dermatology*, *21*, 513-524.

[8] Pacheco, A. G., & Krohling, R. A. (2020). The impact of patient clinical information on automated skin cancer detection. *Computers in*

biology and medicine, *116*, 103545.

[9] Zghal, N. S., & Derbel, N. (2020). Melanoma skin cancer detection based on image processing. *Current Medical Imaging*, *16*(1), 50-58.

[10] Dubal, P., Bhatt, S., Joglekar, C., & Patil, S. (2017, November). Skin cancer detection and classification. In *2017 6th international conference on electrical engineering and informatics (ICEEI)* (pp. 1-6). IEEE.

[11] Ameri, A. (2020). A deep learning approach to skin cancer detection in dermoscopy images. *Journal of biomedical physics & engineering*, *10*(6), 801.

[12] Tembhurne, J. V., Hebbar, N., Patil, H. Y., & Diwan, T. (2023). Skin cancer detection using ensemble of machine learning and deep learning techniques. *Multimedia Tools and Applications*, 1-24.

[13] Alfed, N., & Khelifi, F. (2017). Bagged textural and color features for melanoma skin cancer detection in dermoscopic and standard images. *Expert Systems with Applications*, *90*, 101-110.

[14] Esteva, A., Kuprel, B., Novoa, R. A., KO, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118.

[15] Brinker, Titus Josef, Achim Hekler, Jochen Sven Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H. Enk, and Christof Von Kalle. "Skin cancer classification using convolutional neural networks: systematic review." *Journal of medical Internet research* 20, no. 10 (2018): e11936.