ISBN No. : 978-93-5768-638-9

PROCEEDING OF NATIONAL CONFERENCE

ON

MACHINE LEARNING, DEEP LEARNING AND IoT

(NCMLDLIOT-2023)

JANUARY 19-20, 2023

EDITION - I

EDITED BY

Dr. Babita Majhi

Dr. Pushpalata Pujari

SPONSORED BY

SCIENCE AND ENGINEERING RESEARCH BOARD

Department of Science and Technology, Government of India
New Delhi

Chhattisgarh Council of Science and Technology

Raipur (C.G.)

ORGANIZED BY

Department of Computer Science and Information Technology
School of Studies of Mathematical & Computational Sciences

GURU GHASIDAS VISHWAVIDYALAYA

(A Central University)

Bilaspur – 495009 (Chhattisgarh)

[4] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, “Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism,” *Appl. Sci. (Basel)*, vol. 10, no. 17, p. 5841, 2020.

[5] Z. Ni, A. C. Yuksel, X. Ni, M. I. Mandel, and L. Xie, “Confused or not confused?: Disentangling brain activity from EEG data using Bidirectional LSTM Recurrent Neural Networks,” *ACM BCB*, vol. 2017, pp. 241–246, 2017.

[6] I. Stancin, M. Cifrek, and A. Jovic, “A review of EEG signal features and their application in driver drowsiness detection systems,” *Sensors (Basel)*, vol. 21, no. 11, p. 3786, 2021.

[7] A. Al-Nafjan and M. Aldayel, “Predict students’ attention in online learning using EEG data,” *Sustainability*, vol. 14, no. 11, p. 6553, 2022.

ISBN No. : 978-93-5768-638-9

IDENTIFICATION OF SKIN CANCER USING CONVOLUTIONAL NEURAL NETWORK

Hema Vastrakar,
Dept. of CSIT,
Guru Ghasidas Vishwavidyalaya, Bilaspur(C.G.)
kajuvastrakar7595@gmail.com

Akhilesh Kumar Shrivas
Dept. of CSIT
Guru Ghasidas Vishwavidyalaya, Bilaspur(C.G.)
proffshrivas@gmail.com

ABSTRACT

This paper investigates the effectiveness of CNN in deep neural architectures. Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available datasets of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations acquired and stored by different modalities. Deep convolutional neural network (DCNN) models have been widely explored for skin disease diagnosis and some of them have achieved diagnostic outcomes comparable to or even superior to those of dermatologists. This paper proposes a strategy for the classification of skin lesions on imbalanced datasets. First, various DCNNs are trained on different and imbalanced datasets to verify the models with moderate complexity outperform the high complex models. Second, regularization Dropout and Drop Block are added to reduce overfitting and modified the data set. Augmentation strategy is proposed to deal with the defects of sample underrepresentation in the dataset. Our study shows that this method can achieve a high classification performance at low cost and less computational resources and inference time.

Index Terms—Skin lesion classification, dermoscopy, medical image analysis, deep learning, class imbalanced, Convolutional neural network (CNN).

1. INTRODUCTION

Skin cancer is one of the most common malignancies in the world with significantly increased incidence over the past decade. Skin cancer is typically diagnosed based on dermatologists' visual inspection, with the support of dermoscopic imaging and confirmation of skin biopsy[1]. However, owing to the shortage of the dermatologic workforce and the lack of pathology lab facilities in rural clinics, patients in rural communities do not have access to prompt detection of skin cancer, leading to increased morbidities and melanoma mortalities [2]. Artificial

ISBN No. : 978-93-5768-638-9

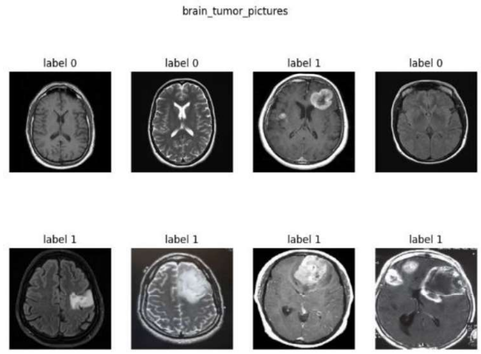

Fig. 1. Brain tumor MRI dataset with label

B. Preprocessing

The preprocessing of the dataset is the most important step in the study of Images and Predictive, as inaccurate data or statistics might impair the function of a distinctively sophisticated classifier. The original dataset must be split into training and validation sets, with training data having the greatest number possible to ensure that the model learns well enough to validate the data[hari]. In this research work also, the dataset has been split into two classes for training, and validation. 30% of MRI images are set aside for validation, and roughly 70% of the total data is set aside for training. These training and validation datasets are divided into tumor data and non-tumor data. The validation data is primarily used to verify and evaluate the trained data; since it was not included in the training process, it provides an unbiased analysis of the suggested model.

a) ***Image resizing and augmentation:***

To obtain the best results, it is suggested to resize the MRI dataset images to have identical height and width. . In this proposed work, the MRI images are resized in width, height, and channel (224 × 224 × 3), which uses 224 × 224 in its input layer.

To learn the Deep Learning model in the best possible way and to prevent over fitting of problems, it must be trained on a large dataset. Moreover, compared to the extensive training dataset, the performance of the DL model can be greatly enhanced. Making a new image out of an existing one is not the goal of image augmentation. It is entirely the acquisition of images of the first image taken from various perspectives. Data Augmentation (DA) is the technique to create an artificial dataset by modifying the original dataset applying scrolling, rotating, enlarging, etc.It is also known as the process of creating multiple image copies of the original image with different angles [19] .The augmented image shown in Fig. 2.

ISBN No. : 978-93-5768-638-9

II. RELATED WORK

Many researchers developed robust classifiers using a variety of machine-learning techniques for the classification of phishing websites. The results of these studies are crucial sources of motivation for future studies.

Ammar Odeh et al. [3] presented a comprehensive review of conventional ML techniques which are significant for detection of malicious attacks on websites. They discussed various challenges to Machine learning based phishing detection techniques. On websites with captcha information, large amounts of image data, and ML algorithms, inefficiency has been found. The literature widely studies overfitting, low precision, and hyper-tuning of ML techniques. The small size of datasets to train the ML techniques is another challenge as identified in this research.

B. Alotaibi and M. Alotaibi [4] proposed a new phishing detection technique based on using two machine learning algorithms, namely, the LightGBM classifier and the AdaBoost classifier, to detect web phishing attacks. This phishing detection method can classify phishing sites in real-time and produce better results than the ones produced by the existing methods.

Ojewumi et al. [5] presented the performance of a machine learning-based classification model for phishing detection using a lightweight Google Chrome extension. PhishNet was developed to detect phishing sites on the web. They used three machine learning techniques namely, K-Nearest Neighbor, Support Vector Machine, and Random Forest as classifiers for classification of phishing attacks. The result given proved that Random Forest performed better than the other two machine learning tools. Consequently, our study has offered a better solution to the issue of phishing attacks on web pages.

Butt et al. [6] proposed machine and deep learning techniques for classifying emails as phish a and not phish. Python programming's NLP language and regular expression are used to obtain the chosen features. These are grouped in an appropriate organizer and divided into multiple classifiers. The classification of email attacks using SVM, NB, and LSTM classifiers where LSTM classifiers achieved highest accuracy.

Y. Su. [7] implemented an LSTM-based phishing detection approach, which tackles the problem that existing machine learning techniques have difficulty extracting relevant features from the data. In this study, the training procedure for the model is optimized using the LSTM deep learning approach in combination with RNN's special characteristics.

ISBN No. 978-93-5768-638-9

Table 2 Confusion Matrix

Actual/Predicted	Normal	Attack
Normal	True Negative (TN)	False Positive (FP)
Attack	False Negative (FN)	True positive (TP)

The performance measures on which the IDS system is evaluated are[16]:

Detection Rate/Recall/Sensitivity: The rate at which model can detect the attack or normal class out of all the samples of the attack or normal class.

$$recall = \frac{TP}{TP + FN}$$

False Alarm Rate: Ratio of normal instances which are incorrectly classified as attack out of total number of normal instances.

$$FAR = \frac{FP}{FP + TN}$$

Precision: The ratio of number of correctly predicted attack class out of all the samples which are attacks.

$$Precision = \frac{TP}{TP + FP}$$

F1-score: It is Harmonic Mean of the precision and recall.

$$F1 - score = 2\frac{Precision \ \times Recall}{Precision + Recall}$$

Sensitivity: It is opposite of recall.

$$Sensitivity = \frac{TN}{TN + FP}$$

V. SIMULATION STUDY

The simulation is carried out in CPU-backed environment -AMD Ryzen 5 3450U processor having 8-GB RAM, 64-bit operating system using Python 3 language.

The pre-processed data is balanced using Random over-sampling, which increases the data to 4,60,645. The data is split into train-test data. The model is trained different batches of train data. In this experiment, the whole train data is shuffled, oversampled using random over sampling and then the SVM classifier with Radial Basis Function kernel is trained in five

ISBN No. 978-93-5768-638-9