Alexandria Engineering Journal 104 (2024) 359–370

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

A multi-fusion approach to classify pharyngitis, tonsillitis and oral cancer with iterative relief feature weighting

M. Swathi, Rajeshkannan Regunathan *

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India

ARTICLE INFO

Keywords:
Tonsillitis
Cancer(Oral)
Pharyngitis
Bag of Visual Words(BOVW)
Histogram of Oriented Gradients(HOG)
Gabor filters
Machine Learning
Transfer Learning

ABSTRACT

Bacterial pharyngitis and tonsillitis can lead to severe complications if untreated, while oral cancer poses risks of spreading if not detected early. Classifying pharyngitis, oral cancer, and tonsillitis is essential for timely and accurate medical interventions. Early diagnosis enables effective treatment planning, potentially saving lives and minimizing complications. Classification aids in tailoring specific medical approaches, contributing to better patient outcomes and healthcare management. The global COVID-19 pandemic has prompted a renewed interest in telemedicine for managing these conditions. Existing techniques, like Bag of Visual Words and individual pre-trained models, fall short in achieving optimal classification. To address this, we propose GHBResIncep, a framework that combines deep features from SE-ResNet, Inception-v4 and EfficientNetV2 with shallow features from improved BOVW, HOG, and Gabor filters. Employing an iterative relief feature weighting algorithm enhances feature selection. Our proposed model achieved highest classification accuracy of 95.58 %.The multi-fused features are fed into ML classifiers like XGBoost, showcasing significant improvements in various classification parameters compared to prior works.

1. Introduction

The imperative to mitigate risks associated with the susceptibility of sick patients to doctors has spurred the adoption of remote healthcare diagnosis and treatment. The COVID-19 pandemic further emphasizes the necessity of remote medicine for individuals exhibiting respiratory symptoms. Leveraging the widespread usage of smartphones, capturing images of the mouth or throat can serve as input to software or processing system for classifying pharyngitis, tonsillitis and oral cancer.

Tonsillitis and pharyngitis share similar symptoms like sore throat, difficulty swallowing, and tonsils that are inflamed [15]. However, oral cancer presents distinct symptoms such as ulcers of oral nature, chewing or swallowing difficulties, abnormal white or red patches in the mouth, and a mass or hardening in the cheek [18]. Despite symptom overlap, these conditions have different causes and risk factors. Advanced feature extraction methods like Bag of Visual Words, Histogram of Oriented Gradients (HOG), Gabor Filters, and feature fusion are essential to tackle this complexity.

The Bag of Visual Words is a feature extraction technique commonly used in computer vision for image classification and object recognition (Sultani and Ban, 2021). It involves creating a visual vocabulary by clustering local image descriptors, and then representing an image as a histogram of occurrences of visual words from the vocabulary [19,20,21]. HOG is a feature extraction method used to capture local gradient information from an image. It divides the image into small cells and computes histograms of gradient orientations within each cell. HOG is particularly popular for pedestrian detection and object recognition tasks. Gabor filters are filters of linear nature used for analysis of texture in image processing. They are defined by a sinusoidal wave of plane nature modulated by a Gaussian envelope. Gabor filters are effective in capturing texture information at various scales and orientations, making them suitable for various computer vision applications.

SE-ResNet is an extension of the ResNet architecture that incorporates Squeeze-and-Excitation blocks [10]. These blocks adaptively recalibrate the feature maps by learning channel-wise scaling factors, enhancing the representational power of the network and improving accuracy on various tasks. Inception-V4 is a deep convolutional neural network architecture that builds upon the Inception family of models [13,12]. It utilizes a combination of different convolutional filters and pooling operations at multiple scales to efficiently capture features from the input data. Inception-V4 achieves state-of-the-art performance on various image recognition tasks due to its efficient and effective design.

* Corresponding author.
E-mail addresses: swathi.m2020@vitstudent.ac.in (M. Swathi), rajeshkannan.r@vit.ac.in (R. Regunathan).

https://doi.org/10.1016/j.aej.2024.06.027
Received 7 May 2024; Received in revised form 3 June 2024; Accepted 11 June 2024
Available online 2 July 2024

Table 1
Cancer (Oral), Pharyngitis, Healthy and Tonsillitis Dataset Distribution.

Class	Data	Number of Training Set (%)	Number of Test Set(%)	Number of Validation Set(%)
Pharyngitis	Raw Data	104 (80)	13(10)	13(10)
Oral Cancer		69 (80)	8(10)	8(10)
Tonsillitis		80(80)	10(10)	10(10)
Healthy		160(80)	20(10)	20(10)
Pharyngitis	Basic	416 (80)	52(10)	52(10)
Oral Cancer	Augmentation	276(80)	32(10)	32(10)
Tonsillitis		320(80)	40(10)	40(10)
Healthy		640(80)	80(10)	80(10)
Pharyngitis	CycleGAN-based	750(80)	94(10)	94(10)
Oral Cancer	Augmentation	750(80)	94(10)	94(10)
Tonsillitis		750(80)	94(10)	94(10)
Healthy		750(80)	94(10)	94(10)

The below paragraph addresses the need for using deep learning models like SE-ResNet and Inception-V4.

SE-ResNet and Inception-V4, arises from the pursuit of overcoming challenges like vanishing gradients and optimizing model efficiency. ResNet introduces the concept of residual learning, enabling the training of extremely deep neural networks without degradation in performance. On the other hand, Inception-V4, an evolution of the Inception architecture, emphasizes improved feature extraction through diverse receptive fields.These models address the demand for enhanced accuracy and efficiency in image classification tasks by leveraging advanced architectural innovations. ResNet's skip connections mitigate the vanishing gradient issue, promoting smoother and more effective training. Inception-V4, with its inception modules, excels in capturing intricate features across different scales. Their adoption underscores the continuous quest for more robust and efficient deep learning architectures in the field of image classification.

The fusion of deep and shallow features enhances model performance by combining high-level abstract representations with specific local information(Parvin et al., 2022). This approach improves robustness, reduces overfitting, and provides better discrimination among classes [17]. The complementary nature of deep and shallow features contributes to a more flexible and adaptable model design, resulting in overall improved classification capabilities. The shallow features are extracted from improved BOVW, HOG and Gabor filters. Our proposed method encompasses the multi-fusing of features from SE-ResNet, Inception-V4, improved BOVW, HOG and Gabor filters.

2. Related work

Pharyngitis is a prevalent acute upper respiratory tract infection affecting approximately 11 million patients in the United State [5]. Streptococcus pyogenes is the primary bacterial cause of pharyngitis, accounting for about 20–30 % of sore throat cases [16,23]. Sankaranarayanan et al. (2005) demonstrated pragmatic features of pharyngitis in Cycle GAN-generated synthetic images. In developing countries, oral cancer predominantly affects men, contributing to two-thirds of global deaths (145,500) and new cases (274,300) in 2002 [1,4]. Tonsillitis is associated with the risk of pneumonia, and it has been calculated that around 58 % of impacted children seek medical attention promptly (Observatory and (GHO) Data, 2015;Phensadsaeng and Kosin, 2017).

In a recent study conducted by M Swathi and colleagues in 2023 (Swathi et al., 2023), they proposed an enhanced BOVW approach to tackle the challenges of oral cancer, pharyngitis and tonsillitis image classification. They fused the features of improved BOVW and Inception-V3 and got an accuracy of 88.3 % and AUC of 0.9288.In another work by Ma et al. in 2016, they employed HOG features and SVM to detect grape leaves, showing robustness to lighting and environmental variations but struggling with grape leaf detection under incorrect postures [26]. Earlier, numerous researchers both domestically and internationally utilized shape features, including eccentricity, second moment, contrast, rectangularity, entropy, roundness,aspect ratio, as well as various texture features, to identify weeds or crops [3]. For instance, Ishak et al. [9] applied features of Gabor and gradient distribution for weed classification.In a work by Chen Y et al. (2020), they developed corn and weed seedling detecting methods using multiple feature fusion and support vector machine [14],. They used HOG and Gabor filters as one of the feature extraction techniques [15].

In a research conducted by Farjana Parvin and her team in 2022 (Parvin et al., 2022), they devised a method to address the issue of information loss by combining deep and shallow features and also similar work in [11]. The merged feature vector was then utilized to train a support vector machine classifier. The experimentation was performed on a publicly accessible dataset. Their novel framework demonstrated remarkable performance with an accuracy of 92.48 % (accompanied by a precision of 93.64 %, recall of 94.55 %, and f1-score of 93.97 %) in the accurate detection of brain tumors.The paper by [6] introduces a novel method for 3D human pose recovery from silhouettes, overcoming challenges in high-dimensional image features. It utilizes multiview locality-sensitive sparse coding, incorporating a local similarity term for stability and improving sparse coding with multiview data. The research by [8] employs multitask manifold deep learning for accurate face-pose estimation, integrating multimodal information to enhance performance. The work by [25] combines user click data and visual features to improve image retrieval rankings by leveraging both user interaction and visual content. The work by [24] develops a stratified deep learning model for fine-grained image recognition, emphasizing the prediction of click features for enhanced understanding of subtle visual details. Hong et al., [7]propose a deep autoencoder-based approach for human pose recovery, utilizing multimodal information to improve accuracy and efficiency in 3D pose recovery from diverse data sources. Our proposed method encompasses the multi-fusing of features from SE-ResNet, Inception-V4, improved BOVW, HOG and Gabor filters.

3. Data acquisition

We have adopted the same data set of pharyngitis,tonsillitis, oral cancer and healthy images that were used for our earlier work by (Swathi et al., 2023).We have applied GAN and Cycle GAN techniques ([2]; Lu and Diagle, 2020; [22])to augment the raw data set. Pharyngitis and non-pharyngitis images were sourced from [23](DOI: https://doi.org/10.17632/8ynyhnj2kz). Oral cancer images were obtained from an open-source link (https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images. Tonsillitis images, collected from Bangkok's hospital in Thailand (Phensadsaeng et al., 2017) included some from open sources. The distribution of number of images in each class of images is as shown in below Table 1. Each input image is of dimension 3x256x256.

The collected raw data, initially imbalanced, underwent augmentation using basic GAN transformations. These transformations included left-right flips, width-height translation (5 % to +5 %), random rotation (10°), zooming (0–20 %), and random brightness change (10 %). Subsequently, Cycle GAN technique was applied to further increase the samples for each class.

4. Proposed methodology of GHBResIncep

4.1. Overview of methodology

The proposed idea encompasses the preparation of deep features extracted from pre-trained models SE-ResNet, Inception-V4 after fine tuning on the dataset being considered and shallow features extracted

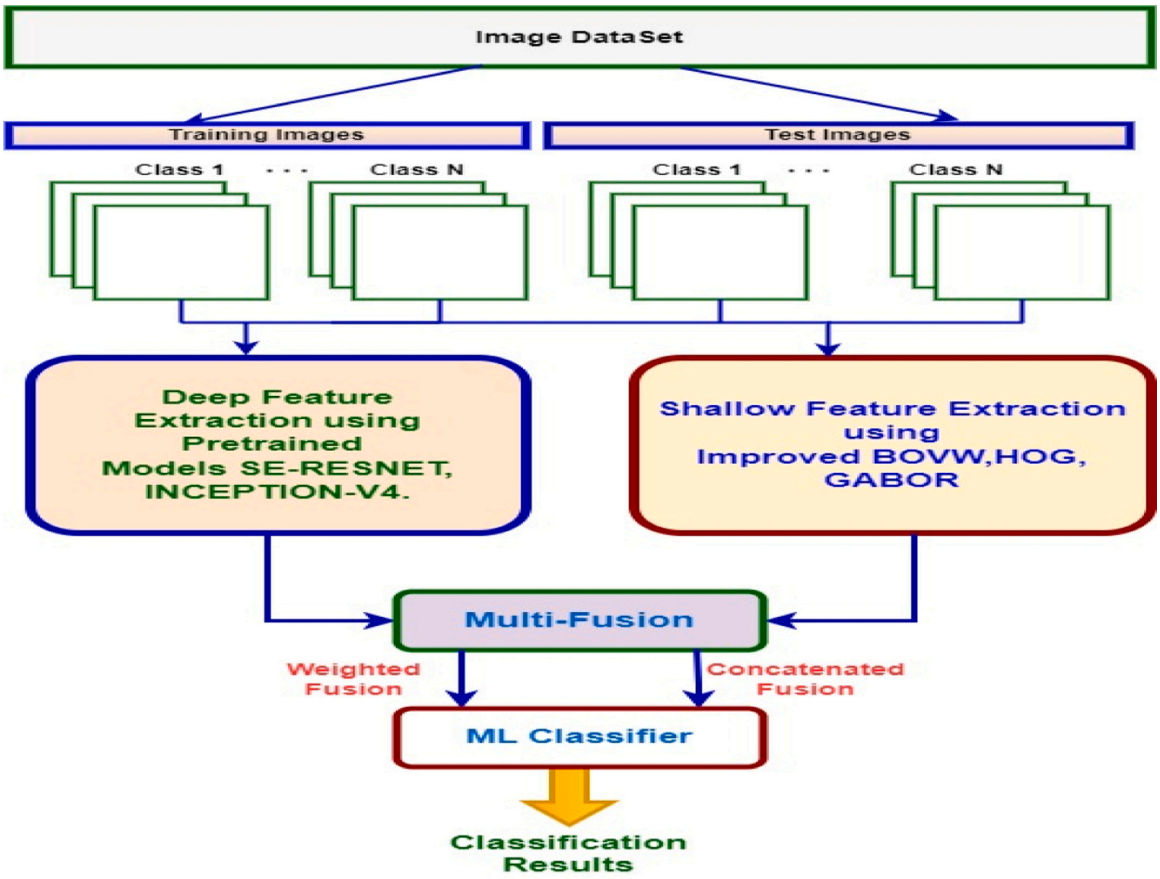


Fig. 1. The overall architecture of proposed framework GHBResIncep.

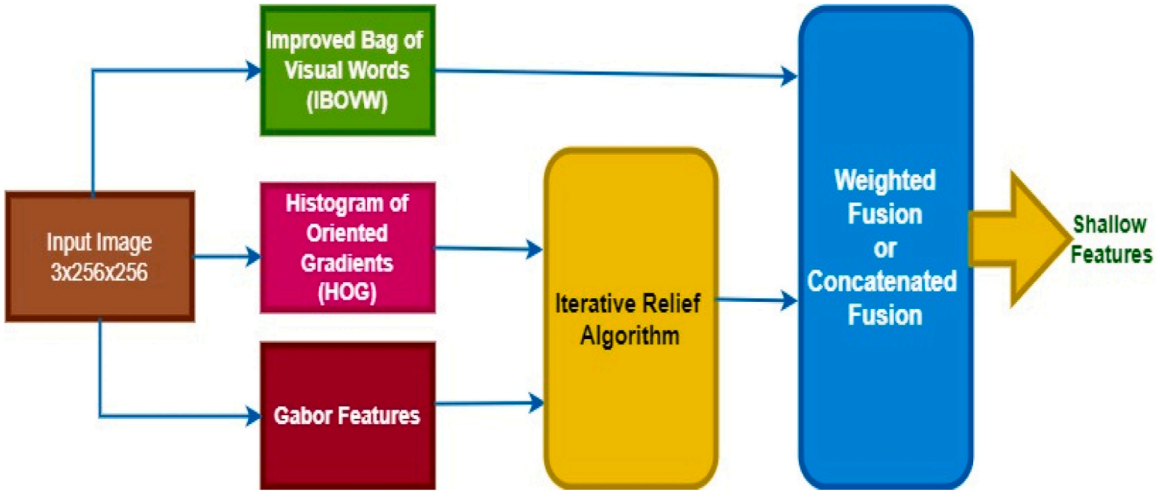


Fig. 2. Shallow Feature Extraction.

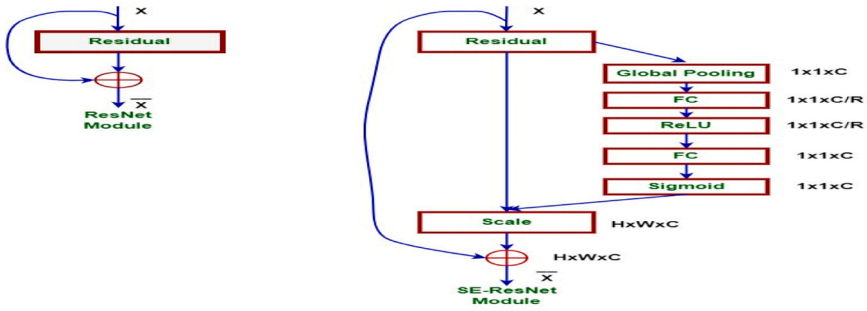


Fig. 3. Architecture of SE-ResNet.

Table 2
Summary of SE-ResNet Trailing Layer (Classifier Layer).

Layer (type)	Output Shape	Param #
Flatten-304	[-1, 2048]	0
Linear-305	[-1, 1536]	3147,264
fc_intermediate-306	[-1, 1536]	0
Linear-307	[-1, 128]	196,736
ReLU-308	[-1, 128]	0
Linear-309	[-1, 32]	4128
ReLU-310	[-1, 32]	0
Dropout-311	[-1, 32]	0
Linear-312	[-1, 4]	132
custom_fc-313	[-1, 4]	0

from improved Bag of Visual Words, Histogram of Oriented Gradients (HOG) and Gabor features. These both deep and shallow features are fused in two kinds of ways. The first way is weighted fusion and second way is concatenated fusion. Next, these fused features are given to ML classifier XGBoost to get classification results. The proposed framework GHBResIncep is combination of techniques Gabor, HOG, Bag of Visual Words and pretrained models SE-ResNet and Inception-V4. The main architecture of the proposed framework GHBResIncep is as given in Fig. 1. The main function of algorithm is given in Algorithm 1.

4.2. Shallow feature extraction

Shallow feature extraction in computer vision holds importance due to its computational efficiency, interpretability, and robustness to small datasets. It offers a more interpretable representation of basic visual patterns like edges and textures, benefiting real-time and resource-constrained applications. Additionally, shallow features can complement deep features, enhancing overall performance. The shallow feature extraction involves initially the retrieval of features from improved BOVW, HOG and Gabor techniques. As the number of features from both HOG and Gabor is large, we are sending them to iterative Relief algorithm and extracting best features out of them. The complete architecture of shallow feature extraction is as given in Fig. 2.

4.2.1. Bag of visual words

As part of Bag of Visual Words, we extract SIFT features (descriptors) from images and then we cluster descriptors of all training images using k-means clustering. Then we initialize histogram of every image to k zeros. Next, we create histogram of an image after incrementing i^{th} value of histogram if i^{th} centroid is nearest to a descriptor of image. We repeat this process for every descriptor of an image. In our earlier work by M Swathi et al. [21], we have selected 4 nearest means in the increasing direction of distances to each descriptor of image and let those indices be I1, I2, I3 and I4. We assign 0.6 to histogram value at index named I1, 0.2 to histogram value at index I2, 0.1 to histogram value at index I3, and 0.1to histogram value at index I4.

4.3. Proposed methodology of improved bag of visual words

In versions of various techniques of BOVW, the SIFT technique extracts features after converting color image into gray scale image. But in this paper, we are proposing an improved version of BOVW such that the Blue, Green and Red components are extracted from image and then we perform SIFT technique on each of these color component. Here, we are strongly believing that RGB scale will preserve information of the image compare to grey scale. The descriptors for each of Red, Green and Blue

Table 3
Description of Customization of Trailing layer (Classifier Layer) of Inception-V4.

Layer (type)	Output Shape	Param #
Dropout-786	[-1, 1536]	0
Linear-787	[-1, 128]	196,736
ReLU-788	[-1, 128]	0
Linear-789	[-1, 32]	4128
ReLU-790	[-1, 32]	0
Dropout-791	[-1, 32]	0
Linear-792	[-1, 4]	132
custom_fc-793	[-1, 4]	0

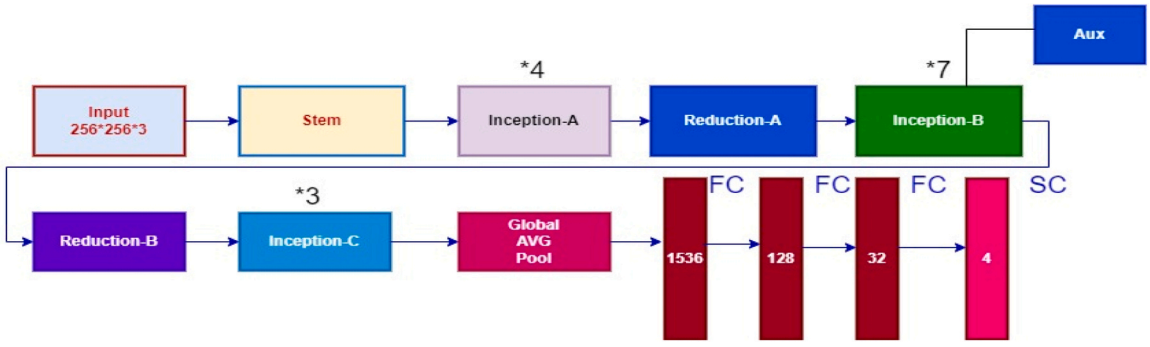


Fig. 4. The architecture of Inception-V4.

are clustered using k-means clustering with k centroids. In this work, we have chosen k=512 which is found to be optimal after rigorous trials. Next, we generated histograms for each color channel and we concatenated those features which will be final representative features of Improved BOVW. The choice of centroids in generating histograms is taken as same as that of our earlier work by M Swathi et al. [15]. Next, we concatenated the histogram features of Red, Green and Blue channel so that resulting number of features will be 3×512=1536.The algorithm is given in Algorithm 2.

Algorithm 1. :***Multi_fusion(Rd): Input***: Rd:Raw datasets of Pharyngitis,Oral Cancer,Tonsillitis and Healthy images.

Output: *Classification result parameters such as Accuracy, Precision, F1 score, Recall.*

1. *Imgdset* ←
   Data set after preprocessing the raw data set Rd using GAN and Cycle GAN techniques.
2. *Divide the dataset Imgdset into Train and Validation and Test data sets with*
   a. *percentages of* 80%, 10% *and* 10% *respectively and let they be Trset, Vset and Tset and*
   b. *make them global to be accessible to each fusion technique.*
3. *Extract features for each of fusion technique as given below.*
4. featbovwtrain ← IBOVW(Trset)
5. featbovwval ← IBOVW($Vset$)
6. featbovwtest ← IBOVW($Tset$)
7. *fea*thogtrain ← HOG(Trset)
8. *fea*thogval ← HOG($Vset$)
9. *featogtest*← HOG($Tset$)
10. *fea*tgabortrain ← Gabor(Trset)
11. *fea*tgaborval ← Gabor($Vset$)
12. *featgabortest*← Gabor($Tset$)
13. $netfeathoggabortrain \leftarrow ReliefF(feathogtrain, featgabortrain)$
14. $netfeathoggaborval \leftarrow ReliefF(feathogval, featgaborval)$
15. $netfeathoggabortest \leftarrow ReliefF(feathogtest, featgabortest)$
16. $featresnettrain, featresnetval, featresnettest \leftarrow SEResNet(Trset, Vset, Tset)$
17. $featinceptrain, featincepval, featinceptest \leftarrow InceptionV4()$
    i. $fusedfeaturestrain \leftarrow$
    Weighted sum of $featbovwtrain, netfeathoggabortrain, featresnettrain, featinceptrain$
    according to various combinations of weights.
18. $fusedfeaturesval \leftarrow$
    Weighted sum of $featbovwval, netfeathoggaborval, featresnetval, featincepval$
    according to various combinations of weights
19. $fusedfeaturestest \leftarrow$
    Weighted sum of $featbovwtest, netfeathoggabortest, featresnettest, featincepval$
    according to various combinations of weights
20. $fusedconcattrain \leftarrow$
    Concatenated featutes of $featbovwtrain, netfeathog, gabortrain, featresnettrain, featinceptrain$
21. $fusedconcatval \leftarrow$
    Concatenated featutes of $featbovwval, netfeathoggaborval, featresnetval, featincepval$
22. $fusedconcattest \leftarrow$
    Concatenated featutes of $featbovwtest, netfeathoggabortest, featresnettest, featinceptest$
23. *Results* ← *Classicication results after giving fused and concat features to XGBoost.*
24. *Print Results*

Algorithm 2. :IBOVW(Dset)

1. $Extract\ the\ SIFT\ descriptors\ of\ Dset\ for\ each\ of\ Red, Green\ and\ Blue\ channel\ and\ let\ these\ be\ DesRed, DesGreen\ and\ DesBlue$
2. $Cluster\ the\ DesRed, DesGreen\ and\ DesBlue\ using\ k-means\ clustering\ to\ form\ visual\ words\ or\ centroids\ VisRed, VisGreen\ and\ VisBlue\ respectively.$
3. $Generate\ histograms\ by\ adding\ 0.6\ to\ nearest\ centroid, 0.2\ to\ next\ nearest\ centroid\ and\ 0.1\ to\ next\ centroid\ and\ 0.1\ to\ next\ centroid\ for\ each\ color\ channel.$
4. $Concatenate\ the\ histogram\ features\ of\ Red, Green\ and\ Blue\ channels\ to\ form\ bovwfeature.$
5. $return\ bovwfeature.$

Algorithm 3. :HOG(Dset)

1. $Extract\ HOG\ features\ by\ passing\ the\ parameters\ such\ as\ orientations = 12, pixelspercell = (16,16), cellsperblock = (3,3) to\ hog\ function\ in\ python\ and\ images\ in\ Dset.$
2. $Reshape\ obtained\ features\ in\ previous\ step\ after\ reshaping\ to\ 3x16x441.$
3. $Now, Apply\ two\ dimensional\ \max pooling\ size\ of\ (1,6) with\ a\ padding\ of\ (0,3)\ to\ get\ intermediate\ features$
4. $Reshape\ the\ features\ obtained\ in\ previous\ step\ to\ (1,3552) as\ Hogfeature.$
5. $return\ hogfeature.$

Algorithm 4. :Gabor(Dset)

1. $Create\ Gabor\ filter\ bank\ with\ kernel\ size = 31\ and\ numangles = 5\ and\ numscales = 4.$
2. $Now\ Extract\ Gabor\ features\ by\ passing\ the\ input\ images\ of\ Dset\ through\ filters\ obtained\ in\ previous\ step.$
3. $Next, Apply\ 2D\ \max pooling\ with\ kernel\ size\ of\ (6,10)\ to\ convert\ number\ of\ features\ from\ 3x256x256\ into\ 3150\ number\ of\ features.$
4. $Reshape\ the\ features\ obtained\ in\ previous\ step\ to\ (1,3150)\ as\ gaborfeature.$
5. $return\ gaborfeature.$

Algorithm 5. :SEResNet(Trset,Vset,Tset)

1. $Import\ the\ pretrained\ model\ SEResNet$
2. $Train\ the\ model\ SEResNet\ on\ Trset\ using\ Vset.$
3. $Extract\ features\ featresnettrain, featresnetval\ and\ featresnettest\ after\ giving\ input\ image\ to\ the\ model\ from\ Trset, Vset\ and\ Tset\ from\ the\ layer\ whose\ number\ of\ out\ features\ are\ 1536.$
4. $return\ featresnettrain, featresnetval\ and\ featresnettest.$

Algorithm 6. :InceptionV4(Trset,Vset,Tset)

1. *Import the pretrained model InceptionV4*
2. *Train the model SEResNet on Trset using Vset.*
3. *Extract features featinceptrain, featincepval and featinceptest after giving input image to the model from Trset, Vset and Tset from the layer whose number of out features are* 1536.
4. *return featinceptrain, featincepval and featinceptest.*

4.4. Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) feature, a prominent descriptor for target detection, was initially introduced in 2005. The core principle underlying HOG is to leverage the gradient or directional density of edges to characterize the local contour of the object within the image. This approach exhibits notable resilience against variations in illumination and background conditions under real-world settings [13].

Convolution is used to compute the gradient of pixel intensities in both horizontal and vertical directions, yielding Gx and Gy values for each pixel. Edge strength and direction are indicated by the gradient's amplitude g and angle θ. The equations for the same are as given below in Eqs. (1) and (2).

$$g = \sqrt{{G_x}^2 + {G_y}^2} \tag{1}$$

$$\theta = \arctan \frac{G_y}{G_x} \tag{2}$$

During the process of extracting HOG features, an image of size 256×256 is partitioned into smaller regions called cells, with each cell having a size of 16×16 pixels. As a result, the image is divided into a total of 256 such cells. Additionally, the cells are further organized into blocks, with each block comprising 3×3 cells. With selection of 12 orientations, the number of features for an image will be 21168 which will be reshaped to 3x16x441size. Next, these will be sent to a two dimensional max pooling with kernel size of (1, 6) with a padding of (0, 3) so that the resulting number of features will be 3552. These features concatenated with Gabor features are given to iterative relief feature weighting algorithm which is given in Section 4.6.The algorithm is given in Algorithm 3.

4.5. Gabor features

We have chosen Gabor features over CNN due to Gabor's effectiveness in capturing texture and edge information. Gabor filters excel in tasks where texture patterns are crucial. In scenarios with limited labeled data, Gabor features, being handcrafted, might be preferred over CNNs, which demand substantial labeled samples. Gabor features offer interpretability, important when understanding feature contribution is key. CNNs, being complex, lack transparency in this regard. The choice hinges on data nature, labeled samples availability, and interpretability needs. So we go for blend of deep and shallow features.

Two-dimensional Gabor filters, with varying frequencies and orientations, can effectively extract valuable features from images in the discrete domain and is given by following Eqs. (3) and (4).

$$G_c[i,j] = B \quad e^{\frac{-(i^2+j^2)}{2\sigma^2}} \cos(2\pi f(i\cos\theta + j\sin\theta)) \tag{3}$$

$$G_s[i,j] = C \quad e^{\frac{-(i^2+j^2)}{2\sigma^2}} \sin(2\pi f(i\cos\theta + j\sin\theta)) \tag{4}$$

Where in, f defines the frequency being looked for inside the texture and θ states the orientation of texture in some direction. We change the support of basis and size of region of image by adjusting σ.

The Gabor filter is renowned for its capacity to capture superior local features, encompassing both frequency and spatial domains. In this research, the Gabor filter was applied to the input image with 4 different scales and 5 distinct directions, resulting in the extraction of 20 sub-images. The output feature obtained from the Gabor filter yields a

Table 4
System Requirements.

Parameter	Details
Platform	Google Colab
CPU	GPU
Language	Python 3.7
PyTorch	2.10.0
Keras	2.9.1
Python Packages	Numpy, Sci-kit Learn, LightGBM, Matplotlib

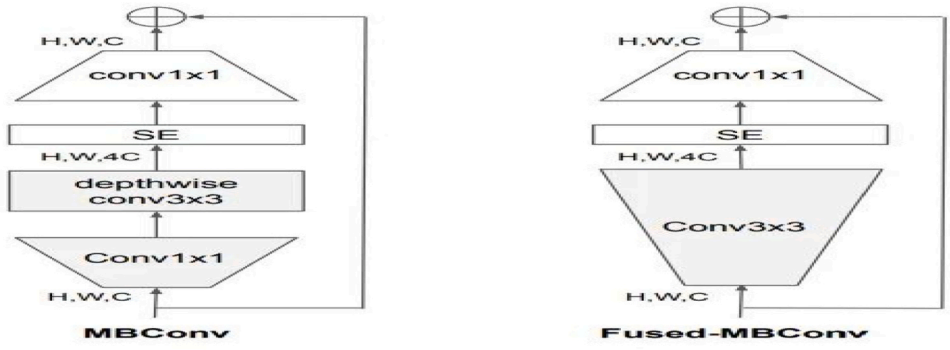


Fig. 5. Connections inside MBConv and Fused-MBConv blocks.

Table 5
Performance of Classification of Tonsillitis, Oral Cancer, Pharyngitis and Healthy.

Model	Accuracy(%)	Precision(%)	Recall(%)	F-1 Score(%)
Only Traditional BOVW	77.39	77.39	77.39	77.39
Only Improved BOVW[Swathi.et al.,2023]	81.12	81.12	81.12	81.12
Only Improved BOVW [Proposed work]	82.21	82.21	82.21	82.21
Only HOG	81.14	81.14	81.14	81.14
Only Gabor	82.13	82.13	82.13	82.13
HOG+Gabor	83.26	83.26	83.26	83.26
HOG+Gabor+ReliefF	87.46	87.46	87.46	87.46
Only InceptionResNetV2	92.13	92.11	92.11	92.11
Only Inception-V4	92.45	92.45	92.45	92.45
Only MobileNetV2	92.56	92.56	92.56	92.56
Only EfficientNetB0	92.61	92.61	92.61	92.61
Only SE-ResNet	93.26	93.26	93.26	93.26
Only EfficientNetV2	94.15	94.15	94.15	94.15
(IBOVW=0.13,HOG=0.13, Gabor=0.13,SE-ResNet=0.3, Inception-V4=0.3)	95.11	95.11	95.11	95.11
(IBOVW=0.2,HOG=0.2,Gabor=0.2 SE-ResNet=0.2, Inception-V4=0.2)	95.27	95.27	95.27	95.27
(IBOVW=0.06,HOG=0.06,Gabor=0.06 SE-ResNet=0.4,Inception-V4=0.4)	94.91	94.91	94.91	94.91
(IBOVW=0,HOG=0,Gabor=0 SE-ResNet=0.5,Inception-V4=0.5)	94.34	94.37	94.34	94.31
(IBOVW=0.1,HOG=0.1,Gabor=0.1 SE-ResNet=0.2,Inception-V4=0.25, EfficientNetV2=0.25)	95.50	95.50	95.50	95.50
(IBOVW=0.1,HOG=0.1,Gabor=0.1 SE-ResNet=0.2,Inception-V4=0.2, EfficientNetV2=0.3)	95.78	95.78	95.78	95.76
Concatenation of All 6 models	95.43	95.42	95.41	95.43

shape of 3x256x256, representing the spatial distribution of the filtered responses.

Subsequently, these extracted features are subjected to 2D max pooling with a kernel size of (6, 10) and no padding. Max pooling involves selecting the maximum value within each pooling region to down sample the feature maps and reduce their spatial dimensions while retaining the most salient information. This will result in 3150 features which will be concatenated with HOG features and to be given to iterative relief algorithm to extract best features. The algorithm is given in Algorithm 4.

4.6. Best Feature extraction from HOG and gabor using ReliefF

As number of features extracted from input images through HOG and Gabor filters is very large we have applied Iterative Relief feature weighting algorithm to extract best features. The algorithm is given in reference (Robnik-Šikonja et al.,2003). The total number of features from HOG and Gabor is 3552+3150=6702. With help of ReliefF algorithm, we are selecting best 1536 . features out of 6702. In this algorithm, W is weight vector and m is number of iterations. $R_i$is a random instance and H_j is the hit instances from training or test data set whichever is under consideration. The algorithm is outlined in Algorithm 7.

Algorithm 7. :ReliefF(features1,features2)

1. $Set\ weights\ W[A] = 0.0, and\ merge\ the\ features1\ and\ features2\ into\ new\ set\ of\ instances.$
2. $for\ l = 1\ to\ m\ do\ begin$
3. $select\ an\ instance\ R_l.$
4. $Calculate\ k\ nearest\ hits\ H_j.$
5. $for\ every\ class\ C \neq class(R_l)\ do$
6. $find\ nearest\ k\ misses\ M_j(C)\ from\ class\ C$
7. $for\ B = 1\ to\ a\ do$
8. $W[B] := W[B] - \sum_{j=1}^{k} \frac{diff(B,R_i,H_j)}{(m.k)} + \sum_{C \neq class(R_i)} \left[\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff(B, R_i, M_j(C)) \right] / (m.k);$
9. $return\ best features\ according\ to\ weight\ vector\ W.$

In the ReliefF algorithm, m determines the number of iterations, impacting the balance between accuracy and computational cost. A higher m improves feature weight estimation but increases computational complexity. k, representing the number of neighbors, influences the stability of feature importance estimation. Larger k values offer stability but might be sensitive to noise. The optimal k and m are dataset-dependent, requiring experimentation. Tuning involves assessing their impact on feature selection.

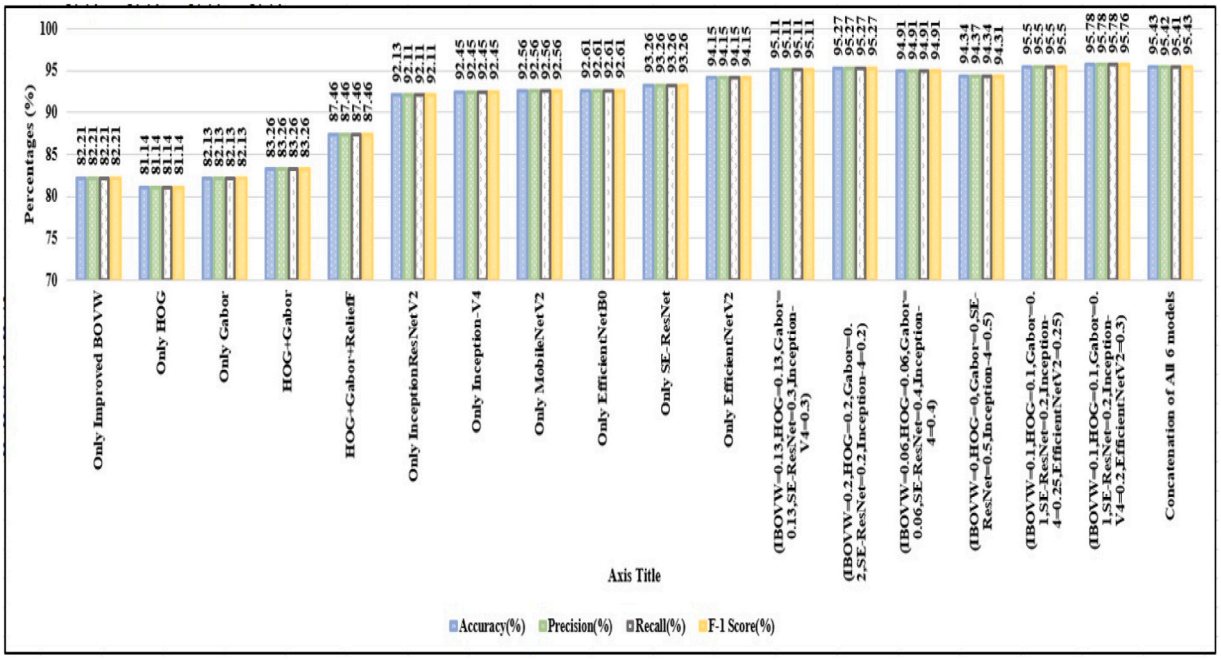
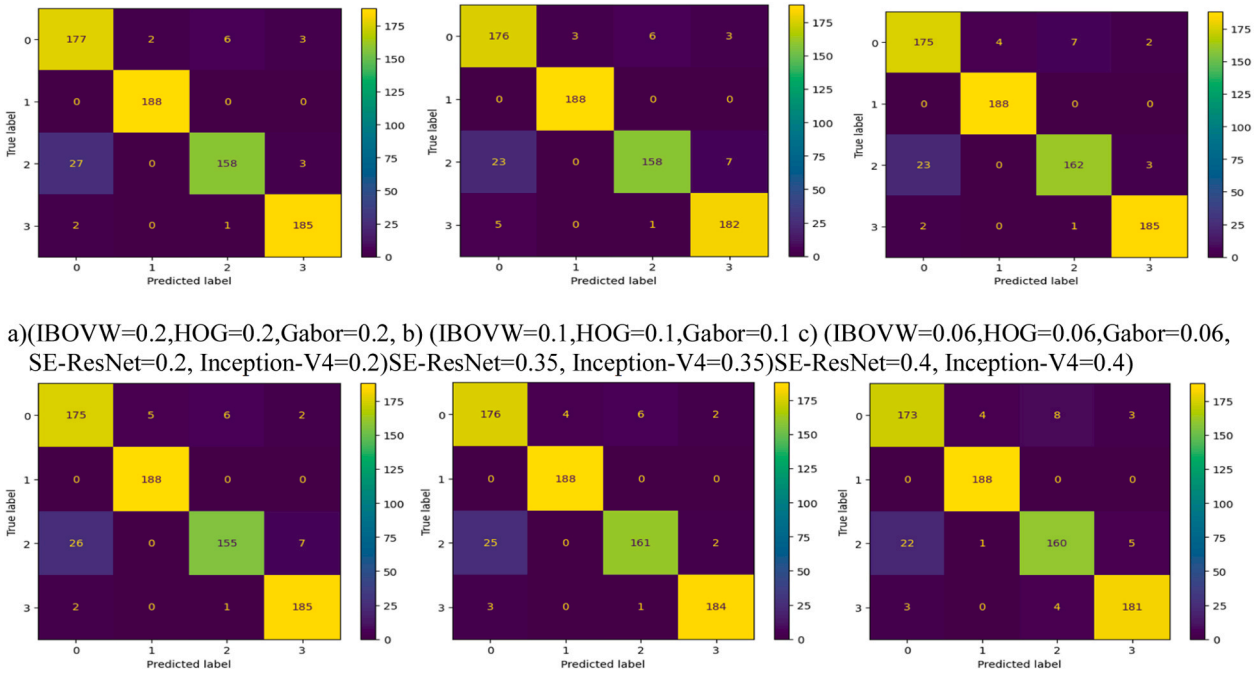


Fig. 6. Bar chart showing Accuracy, Precision, Recall and F1-Score for various models.



a)(IBOVW=0.2,HOG=0.2,Gabor=0.2, b) (IBOVW=0.1,HOG=0.1,Gabor=0.1 c) (IBOVW=0.06,HOG=0.06,Gabor=0.06, SE-ResNet=0.2, Inception-V4=0.2)SE-ResNet=0.35, Inception-V4=0.35)SE-ResNet=0.4, Inception-V4=0.4)

Fig. 7. Confusion matrices of selected 6 combinations of multi-fusion.

4.7. Deep feature extraction

Before jumping to incorporate latest models, we tested the accuracy of the basic models like InceptionResNetV2, MobileNetV2, EfficientNetB0. As the performance of these models is not upto the mark, we moved to incorporate models like SE-ResNet, Inception-V4 and EfficientNetV2.

Pre-trained models like SE-ResNet, and Inception-V4 are preferred over training from scratch due to their learned rich features from large datasets. This accelerates adaptation to new tasks, crucial for limited data scenarios, and reduces the risk of overfitting. The use of pre-trained models also enhances computational efficiency, saving time and resources compared to training from the ground up. The pre-trained models were customized to tune with our proposed work and is explained in Sections 4.7.1 and 4.7.2.

Applying models like EfficientNetV2 is crucial for multi-class image classification due to their advanced architecture that balances model size, computational efficiency, and accuracy. EfficientNetV2's innovative design optimizes resource utilization, enabling effective feature extraction from images of varying complexities. This results in improved classification performance, making it a valuable choice for tasks where model efficiency and accuracy are paramount. This model is explained in Section 4.7.3.

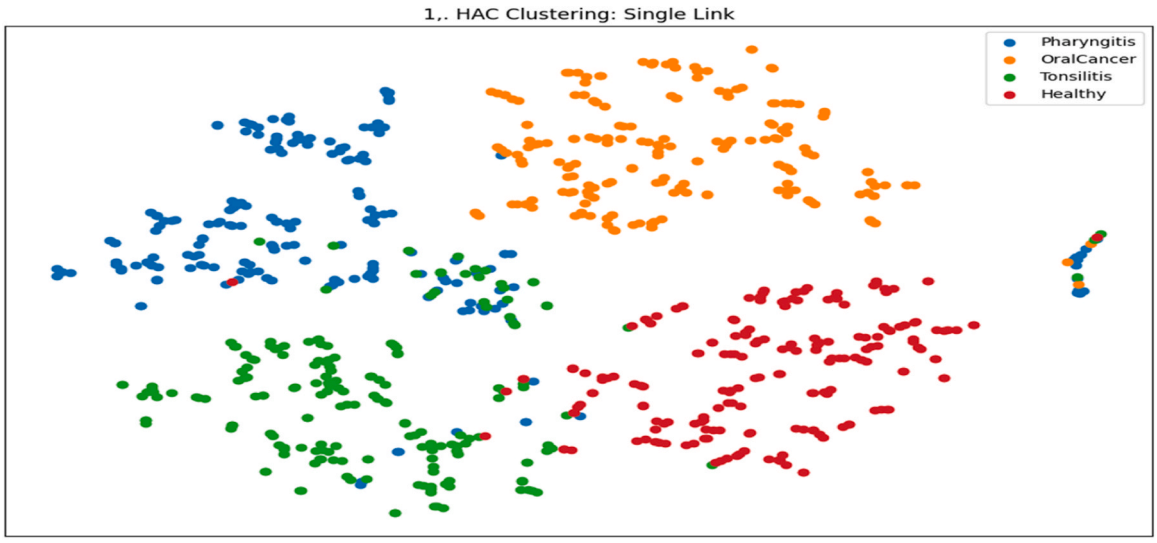


Fig. 8. t-SNE graph for best result of multi-fused weights of (IBOVW=0.1, HOG=0.1, Gabor=0.1, SE-ResNet=0.2, Inception-V4=0.2, EfficientNetV2=0.3). The t-SNE graph for this scenario is shown in Fig. 7.It is also observed that ReliefF feature weighting improved by 4 % when compared to that of HOG and Gabor taken alone.

4.7.1. SE-ResNet

The internal workings of SE-ResNet can be delineated as follows:

1. Residual Blocks: Similar to the original ResNet, SE-ResNet comprises residual blocks that encompass skip connections. These connections enable the seamless flow of input information to the output of each block, effectively addressing the issue of vanishing gradients during training.

2. Squeeze-and-Excitation Block: The pivotal innovation in SE-ResNet lies in the SE block. Within this block, each channel of the feature map undergoes a squeeze operation, wherein spatial information along that channel is aggregated. This aggregation is achieved through global average pooling, computing the channel-wise average of the feature map, and thereby generating a channel descriptor representing the significance of each channel.

3. Excitation Operation: Subsequent to the squeeze operation, the SE block executes an excitation operation utilizing two fully connected layers. These layers serve to uncover the inter-channel relationships and compute channel-wise scaling factors. These learned scaling factors denote the level of importance that should be assigned to each channel.

4. Rescaling and Fusion: The computed scaling factors are then employed to rescale the feature map via channel-wise multiplication. This rescaling step amplifies the contribution of pertinent channels while suppressing less relevant ones, empowering the model to concentrate on discriminative features.

5. Integration with Residual Blocks: The SE block is seamlessly integrated into the conventional residual blocks of the ResNet architecture. The scaled feature maps are combined with the original feature maps through element-wise addition, culminating in the generation of the final output of the block. By incorporating the SE block, SE-ResNet achieves dynamic adaptability in assigning importance to each channel across different layers, endowing the model with the ability to focus more on relevant features and disregard less informative ones. The architecture of SE-ResNet is given in Fig. 3.

After loading pretrained SE-ResNet we have to fine tune the model on our custom data set for a few iterations. As the number of out features is 2048 in SE-ResNet, we have passed these 2048 features into fully connected linear layer whose out features is of number 1536. These 1536 features are passed to dense connected layer whose out features are 128 and these again down sampled to 32 and 4 where 4 is number of classes. After fine tuning with last sigmoid layer we will take the features from layer where 1536 out features are coming so that number of features are same as that of BOVW, HOG and Gabor. The customized architectures of trailing layer of SE-ResNet is as given below in Table 2. The algorithm is given in Algorithm 5.

4.7.2. Inception-V4

Inception-V4 is an advanced version of the Inception architecture, characterized by Inception blocks with parallel convolutional layers, factorized convolutions, reduction blocks for down sampling, and auxiliary classifiers for regularization.

It employs batch normalization and ReLU activation, and a stem network for initial feature extraction. This combination enables Inception-V4 to efficiently capture multi-scale features, leading to enhanced performance in image classification and object recognition tasks. The use of factorized convolutions and auxiliary classifiers contributes to faster training and improved generalization, making Inception-V4 a widely adopted deep learning architecture. The generalized architecture of Inception-V4 is given in Fig. 3.The algorithm is given in Algorithm 6.

After loading pertained Inception-V4 we have to fine tune the model on our custom data set for a few iterations. As the number of out features is 1536 in InceptionV4 we have passed these 1536 features to dense connected layer whose out features are 128 and these again down sampled to 32 and 4 where 4 is number of classes. After fine tuning with last sigmoid layer we will take the features from layer where 1536 out features are coming so that number of features are same as that of BOVW, HOG and Gabor. The customized architectures of trailing layer of Inception-V4 is as given below in Table 3.

4.7.3. EfficientNetV2

EfficientNetV2 features a compound scaling method that optimally scales model dimensions, resolution, and depth to balance performance. It introduces a new technique called "Residual Swish" and utilizes inverted residuals with linear bottleneck. The architecture incorporates mobile inverted bottleneck blocks and utilizes Squeeze-and-Excitation (SE) modules for enhanced feature representation. EfficientNetV2 achieves superior efficiency and accuracy by adaptively scaling network parameters based on a compound coefficient.

This architecture enables effective learning across a broad spectrum of tasks, making it versatile for various image classification challenges. The connections inside MBConv and Fused-MBConv blocks are shown in Fig. 5 (https://towardsdatascience.com/efficientnetv2-faster-smaller-and-higher-accuracy-than-vision-transformers-98e23587bf04). MBConv(Mobile Inverted Bottleneck Convolution) blocks in the

EfficientNetV2 architecture play a crucial role by combining inverted residual connections with depth wise separable convolutions. This design enhances model efficiency and computational performance. The introduction of Fused-MBConv further optimizes operations, consolidating the benefits of MBConv blocks to achieve a well-balanced and efficient neural network structure for diverse image classification tasks.

5. Experimental results

We have used Google Colab Pro for running the experiment. With regard to improved BOVW technique that we have selected number of centroids k=512 so that it amounts to 1536 features for 3 color channels. Table 4 shows the system requirements to run the experiment. The values of various experimental parameters with regard to classification are as given in Table 5. The bar chart for these results has been shown in Fig. 6.The confusion matrices for each of six combinations of multi-fused weights are shown from Fig. 7(a) to Fig. 7(f) with {0: 'Healthy', 1: 'Oral Cancer', 2: 'Pharyngitis', 3: 'Tonsillitis'}.The models SE-ResNet, Inception-V4 and EfficientNetV2 have taken around 20 minutes for 30 epochs with transfer learning and the improved BOVW took 2 hours for extraction of features and HOG for 1 hour and Gabor for 1 hour.

The hyper-parameters of the experiment are accuracy, precision, recall and F-1 Score.SE-ResNet and Inception-V4 employ cross-entropy loss, optimized using stochastic gradient descent during training. The process involves fine-tuning model parameters through back-propagation for classification tasks. Key training details, such as batch size, learning rate, and regularization techniques, are tuned for convergence and preventing overfitting.

From the table of results, it is observed that highest accuracy of 95.78, precision of 95.78 and recall of 95.78 and F1-Score of 95.76 were achieved under weighted combination of (IBOVW=0.1,HOG=0.1, Gabor=0.1, SE-ResNet=0.2,Inception-V4=0.2 and EfficientNetV2 =0.3). The weights should be chosen in such a way that slightly greater importance to deep features compared to shallow features and is confirmed based on rigorous trails.

6. Conclusion

Due to limited size of real world data sets, we have applied GAN and CycleGAN techniques to augment the data set. To get the benefit of early diagnosis in classifying oral cancer, pharyngitis, tonsillitis, the enhanced multi-fused techniques proved that multi-fusion will extract salient features of disease specific characteristics in a diversified manner. The improved version of bag of visual words techniques proved that extraction of features from each of color component is beneficial compared to our earlier work. The usage of iterative relief algorithm retrieved best features out of HOG and Gabor and proved beneficial. The combination of multi-fusion weights of 0.3 for EfficientNetV2, 0.2 for Inception-V4, 0.2 for SE-ResNet, 0.1 for Improved BOVW, 0.1 for HOG and 0.1 for Gabor attained the best accuracy, precision, recall and F1-score. It is also observed that ReliefF feature weighting improved by 4 % when compared to that of HOG and Gabor taken alone. Hence it is proved that as each technique is special in its own extraction of features, the multi-fusion of features attained best results in classification of oral cancer, tonsillitis, pharyngitis. The future works can include applying vision transformers and improve the results of performance within less amount of time compared to our work.

CRediT authorship contribution statement

Rajeshkannan Regunathan: Supervision. **Swathi M:** Writing – original draft.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest that could potentially influence the objectivity, integrity, or impartiality of the research findings presented in this work.

References

[1] R.O. Alabi, O. Youssef, M. Pirinen, M. Elmusrati, A.A. Mäkitie, I. Leivo, A. Almangush, Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future-A systematic review, Artif. Intell. Med 115 (2021) 102060 https:// doi.org/10.1016/j.artmed.2021.102060.

[2] T. Carneiro, R.V.M. Da NoBrega, T. Nepomuceno, G.B. Bian, V.H.C. De Albuquerque, P.P.R. Filho, Performance analysis of Google colaboratory as a tool for accelerating deep learning applications, IEEE Access 6 (2018) 61677–61685, https://doi.org/10. 1109/ACCESS.2018.2874767.

[3] Y. Chen, Z. Wu, B. Zhao, C. Fan, S. Shi, Weed and corn seedling detection in field based on multi feature fusion and support vector machine, Sensors 21 (1) (2020) 212.

[4] J. Ferlay, D.M. Parkin, P. Pisani, GLOBOCAN 2002: cancer incidence, mortality and prevalence worldwide, IARC Press, Lyon, 2004.

[5] Global Health Observatory (GHO) Data. WHO. Care Seeking for Pneumonia, WHO, accessed on Jun. 15, 2015. [Online]. Available: http://www.who.int/gho/child_health/prevention/pneu monia_text/en/.

[6] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Ind. Electron. 62 (6) (2014) 3742–3751.

[7] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670, https://doi.org/10.1109/TIP.2015.2487860.

[8] C. Hong, J. Yu, J. Zhang, X. Jin, K.-H. Lee, Multimodal Face-Pose Estimation With Multitask Manifold Deep Learning, IEEE Trans. Ind. Inform. 15 (7) (2019) 3952–3961, https://doi.org/10.1109/TII.2018.2884211.

[9] A.J. Ishak, A. Hussain, M.M. Mustafa, Weed image classification using Gabor wavelet and gradient field distribution, Comput. Electron. Agric. 66 (1) (2009) 53–61.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.

[11] L. Lu, B.J. Daigle, Prognostic analysis of histopathological images using pre-trained convolutional neural networks: application to hepatocellular carcinoma, PeerJ 8 (2020) e8668 https://doi. org/10.7717/peerj.8668.

[12] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, G. JamalipourSoufi, Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning, Med. Image Anal. (2020) https://doi.org/10. 1016/j.media.2020.101794.

[13] Nakasi R., Mwebaze E., Zawedde A., Tusubira J., Akera B., Maiga G. (2020) A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models.

[14] F. Parvin, M. Al Mamun, Feature Fusion Based Effective Brain Tumor Detection Approach Using MRI. In 2022. 25th International Conference on Computer and Information Technology (ICCIT), IEEE, 2022, pp. 611–616.

[15] P. Phensadsaeng, K. Chamnongthai, The Design and Implementation of an Automatic Tonsillitis Monitoring and Detection System, IEEE Access 2699665 (2017), https://doi.org/10.1109/ACCESS.2017.

[16] A. Rao, B. Berg, T. Quezada, R. Fader, K. Walker, S. Tang, U. Cowen, D. Duncan, J. Sickler, Diagnosis and antibiotic treatment of group a streptococcal pharyngitis in children in a primary care setting: impact of point-of-care polymerase chain reaction, BMC Pedia 19 (2019) 24, https://doi.org/10.1186/s12887-019-1393-y.

[17] M. Robnik-Šikonja, I. Kononenko, Mach. Learn. 53 (2003) 23, https://doi.org/10.1023/A:1025667309714.

[18] R. Sankaranarayanan, K. Ramadas, G. Thomas, R. Muwonge, S. Thara, B. Mathew, B. Rajan, Trivandrum Oral Cancer Screening Study Group, Effect of screening on oral cancer mortality in Kerala, India: a cluster-randomised controlled trial, Lancet (Lond., Engl.) 365 (9475) (2005) 1927–1933, https://doi.org/10. 1016/S0140-6736(05)66658-5.

[19] C. Sitaula, S. Aryal, New bag of deep visual words-based features to classify chest x-ray images for COVID-19 diagnosis, Health InfSciSyst 9 (24) (2021) https://doi.org/10.1007/s13755-021- 00152-w.

[20] Z. Sultani, B.N. Dhannoon, Modified bag of visual words model for image classification article's information abstract. AlNahrain J. Sci. 24 (2021) 78–86, https://doi.org/10.22401/ ANJS.24.2.11.

[21] M. Swathi, R. Regunathan, A novel feature fusion-based approach for detecting pharyngitis, oral cancer, and tonsillitis using improved bag of visual words, Soft Comput. (2023) 1–13.

[22] T.K. Yoo, J.Y. Choi, H.K. Kim, CycleGAN-based deep learning technique for artifact reduction in fundus photography, Graefes Arch. Clin. Exp. Ophthalmol. (2020) https://doi.org/10.1007/s00417-020- 04709-5.

[23] T.K. Yoo, J.Y. Choi, Y. Jang, E. Oh, I.H. Ryu, Toward automated severe pharyngitis detection with smartphone camera using deep learning networks, Comput. Biol. Med 125 (2020) 103980 https://doi. org/10.1016/j.compbiomed.2020.103980.

[24] J. Yu, M. Tan, H. Zhang, Y. Rui, D. Tao, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2022) 563–578, https://doi.org/10.1109/TPAMI.2019.2932058.

[25] J. Yu, D. Tao, M. Wang, Y. Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2015) 767–779, https://doi.org/10.1109/TCYB.2014.2336697.

[26] M. Yuan, F. Quan, Y. Mei, Detection of wine grape leaves based on HOG, Comput. Eng. Appl. 52 (15) (2016) 158–161.

M. Swathi completed her M.Tech from G.Pulla Reddy Engg College from Kurnool, Andhra Pradesh. She has been pursuing research as a full-time research scholar in the department of CSE, Vellore Institute of Technology, Vellore in December 2020. She has published 3 Conferences in Springer Proceedings. Her research interests include Machine Learning, Image Processing, and Deep Learning.

Rajeshkannann Regunathan has received his MTech in Computer Science and Engineering from SASTRA University, Tanjore, and Ph.D in Computer Science and Engineering from Vellore Institute of Technology, Vellore, India. He works as an Associate Professor at VIT, Vellore, India. His area of interest is cloud computing, artificial intelligence, natural language processing and data science. He has 15+ years of teaching experience. He has published more than 35 papers in international journals and conferences. He is a member of CSI, IEEE(WIE).