Chapter 2

Literature Review

In this chapter, we begin with a literature review of local pattern-based texture descriptors, highlighting their potential in image retrieval and various other applications within computer vision. This is followed by a discussion on a range of DL frameworks developed for COVID-19 detection utilizing CXR images. The chapter next includes a survey of current CNN based methods for SL classification using dermoscopic images.

2.1 Local pattern based feature descriptors

A key feature in images is texture, and over the course of several decades, an extensive amount of study has been directed towards the fundamental yet difficult topic of texture representation in computer vision and pattern recognition. With outstanding performance, texture representations based on handcrafted schemes, CNNs including bag of words (BoW) have been investigated intensively since 2000 [49].

In a broad array of scientific domains, including pattern recognition, remote sensing, and medical imaging, the texture features have been proven invaluable. Humans can recognise a wide variety of things in the world by using their unique textures, as each object has a unique texture. In computer vision, texture analysis occurs in statistical and structural levels. At the statistical level, local features are calculated in parallel at every point within a texture image. Based on their distributions, a set of statistics is then extracted from these local features. A local feature is determined by the intensities at certain positions around each

point in the image. First-order, second-order, and advanced-order statistics are categorised based on the quantity of points that characterise the local feature [50].

One of the most well-known second-order statistical texture features in the literature is the grey-level co-occurrence matrix (GLCM). It extracts details regarding the frequency of the occurrence of two adjacent pixels in an image. An image can be converted into a matrix that represents the relationships between the pixels in the original image using a GLCM. For a given distance and direction, it computes the mutual occurrence of pixel pairs. The displacement of neighbouring pixels and the orientation between them are the two terms that are used to compute GLCM. Orientations may be diagonal, vertical, or horizontal. It is possible to extract up to fourteen textural features from each matrix. Typically, only four features —energy, entropy, contrast, and homogeneity —are obtained from the input image [51]. The literature encompasses various feature descriptors, such as Histogram of Gradients (HOG)[52], Scale Invariant Feature Transform (SIFT)[53], and Speeded up Robust Features (SURF)[54], among others. An approach for recognizing unique invariant features in images is presented in [53]. Robust matching is demonstrated throughout a wide range of affine distortion, and the features remain constant despite changes in image size and rotation. The SIFT technique is denoted by this approach, as it converts image data into coordinates that remain invariant to scale in relation to local features. The author introduced SURF, an effective interest point identifier and descriptor that is invariant to scale as well as rotation. The concept of the first local pattern-based descriptor for texture analysis, local binary pattern (LBP), was proposed by Ojala et al. in [20]. An LBP operator unites both structural and statistical texture evaluation features. The LBP displays the eight adjacent pixels in binary code and operates on a pixel-by-pixel basis. The extraction of a textural feature is therefore made easier by the LBP, which then compiles all the codes into a histogram. A 3×3 neighbourhood would generate a 256-texture pattern. If I_c and I_n, respectively, stand for the centre and neighbour pixels, where n ϵ [0, 7] denotes the circular neighbours, the binary code produced by the difference between I_c and I_n is given by Equation (2.1)

$$f(a) = \begin{cases} 1, & if\ a \geq 0 \\ 0, & else \end{cases} \tag{2.1}$$

Following the creation of binary code, the pattern values are acquired via Equation

(2.2)

$$LBP = \sum_{n=0}^{7} f(I_n - I_c) \times 2^n \tag{2.2}$$

Because of the fact that LBP is not rotational invariant, it is undesired in some applications. A rotation invariant version is developed in [55]. Ojala et al. [56] discovered that some local binary patterns, referred to as "uniform" represents essential components of the texture in images, and their histogram of occurrence has been shown to be an extremely potent texture characteristic. In order to identify "uniform" patterns, they develop a representation that is both generalized and rotational invariant. This makes it possible to identify unique patterns. Given that the descriptor is, as defined, unaffected by any consistent modification of the grayscale, this method is particularly resilient with respect to variations in the grayscale. Early research has concentrated on texture analysis and classification processing because the LBP approach is primarily utilised to describe texture feature information. Later, a variety of significant LBP improvement techniques are created. An associated completed LBP (CLBP) method for classification of texture is established in [57], along with a completed model of the LBP operator. CLBP–Magnitude (CLBP_M) and CLBP-Sign (CLBP_S) are the two operators that are suggested. Rotation-invariant texture classification can be significantly improved by integrating CLBP_M, CLBP_S, and CLBP-Centre (CLBP_C) features into combined or integrated distributions. A broader interpretation of the LBP descriptor, local ternary patterns (LTP), was presented by Tan and Triggs [58]. In uniform regions, LTP is less susceptible to noise and is more discriminating. LTP is a three-valued code in which grey levels within a specified zone are quantized to 0, values greater than this zone are quantized to +1, and those less than it are quantized to -1. The local derivative pattern (LDP), a new high-order local pattern descriptor for face recognition, was proposed in [59]. Based on local derivative variants, LDP provides a general framework for encoding directional pattern information. In contrast to the first-order local pattern utilised in LBP, the Nth-order LDP is suggested to encode the $(N-1)_{th}$ order orientation changes in the local derivative. This allows for the collection of more precise information. Local ternary co-occurrence patterns (LTCoP), a novel feature extraction approach was presented in [22]. While the normal LDP reflects the co-occurrence of the 1st-order derivatives in a particular orientation, the LTCoP reflects the co-occurrence of similar ternary edges, generated in accordance with the intensity values of the central pixel and adjacent neighbours. The LTCoP was proposed as a result of the concepts of LDP and LTP. With the help of LBP texture operator and the well-known SIFT

descriptor, the author developed a novel interest region descriptor [60]. They refer to it as the CS-LBP descriptor, which stands for centre-symmetric local binary pattern. It is a descriptor with extremely few dimensions. The CS-LBP only produces 16 unique binary patterns, however the LBP generates 256 distinct ones. In earlier local pattern methods, features were extracted as histograms representing each pattern's frequency without considering the patterns' mutual occurrence within the histogram. In [61], the author addressed this limitation by improving the feature extraction process. CSLBP was selected for extracting local patterns from the original image, and GLCM was applied to extract features from the CSLBP pattern map. Inspired by LBP, Murala and Wu introduced a novel technique for retrieval of biomedical image termed as local mesh patterns (LMeP)[22]. The suggested approach captures the association among the adjacent neighbours for a specific referenced pixel in a picture, whereas the standard LBP captures the association between the referenced pixel and neighbours. Local mesh peak valley edge pattern (LMePVEP), a new pattern-oriented feature, is proposed in [62]. The first-order derivative is used to determine the peak or valley edges that connect the neighbours. The primary limitation of LBP, LTP, LTCoP, and LMeP is the "curse of dimensionality", which arises as the feature dimension increases. The only relationships exploited in the LBP, LTP, LTCoP, and CSLBP methods are those between the central and the neighbouring pixels, while the LMeP approach only uses the relationships between the immediate neighbours. The authors suggest a innovative and effective local feature descriptor known as local diagonal extrema pattern (LDEP) [25], in response to the excessive dimensionality of the current techniques. Utilizing 1st-order local diagonal derivatives, this method determines the indexes and values of the local diagonal maxima and minima to benefit from the relationship between any image's centre pixel's diagonal neighbours.

Numerous handcrafted descriptors in the transform domain also have been introduced in recent years. In order to retrieve CT images, Dubey devised the local wavelet pattern (LWP) [21]. Unlike the LBP, which only takes into account the connection between middle pixel and its surrounding pixels, this method first uses local wavelet decomposition to analyse the interaction between the neighbouring pixels before taking into account the relationship with the centre pixel. The study in [63] introduced a novel class of wavelet feature descriptors called Local Neighborhood-Based Wavelet Feature Descriptors (LNWFD) for retrieval of medical image. This method utilized the Triplet Half-Band Filter Bank (THFB), which used three kernels to improve the properties of wavelet

filters. Initially, the THFB was applied for wavelet decomposition at single-level, generating four sub-bands. Then, the relationship between the coefficients derived from wavelet analysis in each sub-band is captured utilizing a 3×3 neighbourhood window, forming the LNWFD pattern. The key innovation of this descriptor lies in analyzing the relationship between wavelet-transformed pixel values instead of their intensity values, providing more detailed local information within the wavelet sub-bands. The method in [64] describes an image retrieval approach built on the YCbCr colour model, utilizing a histogram of canny edge and discrete wavelet transform (DWT). Combining the edge histogram with the DWT enhanced the outcomes of the CBIR framework. Additionally, various wavelets were compared to determine the most suitable function for image retrieval tasks. Statistical modelling regarding the wavelet subbands has often been applied in identifying and retrieving images. However, due to their inherent limitations, conventional wavelets are ineffective for images containing distributed discontinuities, like edges. Shearlets, a newer advancement of wavelets, are more suitable for image characterization in such cases. The authors of this study [65] presented texture retrieval and classification techniques that used linear regression to describe the dependence between consecutive shearlet subbands. The study in [66] presented a method for texture-based image retrieval that employs the Discrete Non-Separable Shearlet Transform (DNST) domain, utilizing a Local Neighborhood Intensity Pattern (LNIP). This method effectively captures discriminative information by analyzing the relationships between a specific coefficient and its neighbouring pixels inside a local window. First, the original picture was disaggregated into various frequency and orientation subbands employing the DNST. Then, inspired by the CS-LBP theory, the DNST domain LNIP descriptor's sign and magnitude patterns were developed based on an innovative computation rule.

While LBP and its variants exhibit commendable performance, they demonstrate limitations in effectively capturing extremely fine details within images. To resolve this matter, for the purposes of indexing and retrieval of biomedical image, a novel feature descriptor for images that relies on local bit-plane decoded patterns (LBDP) is presented [26]. In contrast to existing local feature descriptors, the values of intensity of the local neighbours regarding any referenced pixel were not used directly. Instead, these values were decomposed into local BPs for additional processing. To calculate each image pixel's local BP transformed values, a local BP transformation scheme is put forth. A central pixel's value of intensity is compared to the local BP transformed values for

each BP to create LBDP. Local bit-plane dissimilarity pattern (LBDISP), a local BP dissimilarity pattern-based image descriptor, is presented in [27]. The operator is calculated by determining the dissimilarity map comparing the centre pixel with its neighbouring pixels across each BP. The LBDISP descriptor is subsequently formed by encoding the association of the center pixel with the dissimilarity map. The number of neighbouring pixels exclusively determines the dimension of the descriptor. A BP-based descriptor known as the local bit-plane adjacent neighbourhood dissimilarity pattern (LBPANDP) was presented by the authors in [29]. In light of the dimensionality problem, the authors of this paper only took four MSB planes into account because these planes have more important texture information. In contrast to traditional encoding methods, LBPANDP encodes the binary neighbours' dissimilarity information. The LBPANDP descriptor carries highly discriminative information with significantly fewer feature vector dimensions. Local bit plane-based dissimilarities and adder pattern (LBPDAP), another BP-based feature descriptor, is proposed in [28]. The neighbor-neighbor mutual information on dissimilarity and the center-neighbor information on dissimilarity in each BP are combined by an adder to encode the BPs. Only the four most significant BPs have been accounted for, to restrict the size of feature. The authors in [30], presented a unique descriptor, called as non-subsampled shearlet transform (NSST) local bit-plane neighbour dissimilarity pattern (NSST-LBNDP) for retrieval of biomedical image. In this descriptor, the NSST decomposes the input image initially, and then uses local energy feature computation to introduce non-linearity affecting the NSST coefficients. Decomposing the normalized NSST sub band feature into BP slices allows us to acquire sub band characteristics ranging from extremely fine to coarse.

Paulhac et al. expanded the LBP technique to three dimensions (3D) and compared this with the two-dimensional (2D) approach for analyzing 3D textures. To evaluate the performance of both methods, they carried out classification experiments on three different databases of 3D texture images, each with unique characteristics. The 3D LBP method performed better in each case than the 2D approach [67]. A study introduced 3D rotation-invariant texture descriptors derived from the widely known LBP. In this approach, they enhanced a 3D LBP method by incorporating the region-growing algorithm, utilizing features initially designed for 2D LBPs, such as pixel intensities and intensity differences [68]. Recently, extraction of features from multi-scale Gaussian-filtered images has received more attention. Different scales of discriminative information can be captured by these strategies. Local binary patterns currently in use, capture the

interplay between the central pixel and the pixels directly surrounding it within 2D local regions of a picture at a particular scale. The technique described in [48] uses a multi-resolution Gaussian filter bank to create a 3D plane from a 2D image, encoding the association of the central pixel and its adjacent pixels in five chosen directions named as spherically symmetric 3D-LTP (SS-3D-LTP). Furthermore, they suggest a colour SS-3D-LTP, in which the RGB spaces are viewed as three 3D volume planes. For image retrieval applications, a three-dimensional local ternary co-occurrence pattern (3D-LTCoP) is presented in [46]. The primary concept involves examining the statistical geometry of images obtained with a Gaussian filter bank over various scales. To create a 3D image, five different scales of images are used. It computes first-order derivatives in the neighbourhood using LTCoP, which enables the capture of fluctuations (fine to smooth) at various levels. Five distinct planes are used by 3D-LTCoP to capture multidirectional variation. Ref. [47] proposed two methods for the retrieval of biomedical images: 3D local circular difference wavelet patterns (3D-LCDWP) and 3D local circular difference patterns (3D-LCDP). By employing three planes derived from the original image, a 3D volume is produced for the purpose of calculating local circular difference patterns. RGB components are employed as three planes in colour images, and different resolution Gaussian filter banks are used in grayscale images.

The majority of aforementioned descriptors, however, are based on the same fundamental concept of LBP and only extract the texture image's circular isotropic microstructure, which is insufficient to characterise the details of the texture and does not adequately handle the rotation invariant concerns. To overcome this problem, a novel image descriptor known as Local Directional ZigZag Pattern (LDZP) was developed [31]. Local ZigZag pattern (LZP) uses ZigZag scanning to determine the association of a center pixel with its local neighbouring pixels. Another descriptor, local ZigZag Max histograms of pooling pattern (LZMHPP), is proposed in [32] after the advantages of ZigZag features became popular [31]. This descriptor initially computes the dissimilarity between the centre pixel and its adjacent pixels for every image patch across the entire image, encoding the dissimilarity map using a diverse ZigZag ordering techniques, and then pooling the max histograms to devise the LZMHPP descriptor. The suggested descriptor in [35] used three distinct 3D zigzag patterns in four distinct orientations to express the association between a standard pixel and corresponding surrounding pixels in a 3D plane. Therefore, to represent the association between the reference and its adjacent pixels in a 3D surface, a set of 12 efficient 3D zigzag structures were presented. The author of [36] presented 3 distinct 3D

zigzag patterns in 4 different orientations and suggested a local-oriented zigzag ternary co-occurrence fused pattern (3D-LOZTCoFP). Using the suggested 3D zigzag configuration at radii 1 and 2, they first compute the 3D LTP inside a local 3D region surrounding a reference in 3D-LOZTCoFP. To further improve the descriptor's discriminating ability, the co-occurrence of comparable ternary edges inside the local 3D cube is then calculated.

The literature reveals that most BP-based feature descriptors rely on conventional circular patterns, with limited exploration of 3D sampling patterns, highlighting a significant research gap. Utilizing arbitrary patterns with angular variations could potentially extract more robust and discriminative features. A lack of published work on the establishment of effective sampling patterns is evident from the literature review. To better characterize uniform and non-uniform textures, sampling architectures with a greater number of angular deviations between subsequent sampling locations are expected to perform better. Enhanced biomedical image retrieval outcomes are the consequence of better capture of very coarse to very fine image features.

2.2 Deep learning based approaches on COVID-19 diagnosis using chest X-ray scans

For the detection of inflammation of the lungs, swollen lymph nodes, pneumonia, and additional breathing-related issues, X-ray imaging has been the standard method. The lungs' lining epithelial cells are first impacted by SARS-CoV-2 infections. In this instance, a CXR can be performed to examine the lungs of patient for signs of COVID-19 contamination. CXR is also useful for researching how a disease develops and how it affects the lungs after it has ended. The radiologist employing X-ray scans is supposed to visually evaluate the image and identify any indicators related to the viral infection. This is because previous research has connected abnormalities brought on by the COVID-19 infection to anomalies in chest radiography [69]. When it comes to assessing radiological images, the most significant obstacle that the radiologist face is the visual inspection of minute features. Furthermore, a significant number of chest radiological scans need to be analyzed in a comparatively short duration, which increases the likelihood of incorrect categorizations being made. Because of this, the use of clever techniques that are capable of automatically classifying radiological scans of the chest is

absolutely justified. In the creation of computer based automated systems, DL algorithms have been gaining a lot of favour. DL algorithms, also known as CNNs have the ability to automatically evaluate radiological images and determine whether or not an individual displays positive results for COVID-19. Even while CNN systems perform very well in image categorization, the construction of an automated framework that utilizes of a CNN needs a significant power and a huge dataset in order to get satisfactory results. Transfer learning (TL) is a ML scheme that may be used to address a similar learning issue with a dataset that lacks training samples by transferring the information that pre-trained CNNs gain during their training. TL is useful in avoiding overfiting issues.

The TL technique was applied to a pretrained CNN-based AlexNet architecture in [70]. The Support Vector Machine technique was then used to classify the effective features from all layers of the architecture that had been chosen using the relief feature selection procedure. The study [71] introduces a classification approach by utilizing features extracted from well-known CNN models, including AlexNet, ResNet18, ResNet50, Inceptionv3, Densenet201, Inceptionresnetv2, MobileNetv2, and GoogleNet. Two key contributions are highlighted: the use of Bayesian optimization for selecting hyperparameters of ML algorithms and the implementation of Artificial Neural Network (ANN)-based image segmentation. Initially, lung segmentation is automatically conducted from raw images using ANNs. To optimize classification accuracy, Bayesian optimization is applied to fine-tune the hyperparameters of each ML algorithm. The paper [72] proposes a fine-tuned DenseNet201 model for classifying CXR images. Initially, DenseNet121, DenseNet169, and DenseNet201 models were trained and validated on the same database. Results from experiments revealed that the DenseNet201 model outperformed the other DenseNet variants. Additionally, the DenseNet201 model was tested with different optimizers, and it was observed that RMSprop, Adagrad, and Adamax yielded better performance. The authors in [73] introduced a novel framework for diagnosing COVID-19 infection in the article. They modified the established VGG-16 model by removing its fully connected (FC) layers and replacing them by introducing a new, simplified set of FC layers initialized with random weights. This was applied to the DCNN, which had already learned to recognize discriminative features, shapes, and objects. To preserve these valuable features, the FC head was pre-trained by freezing all layers in the network body, after which all layers were unfrozen for fine-tuning. A relative evaluation of fine-tuned DL models was conducted in [74]to accelerate the identification and categorization of COVID-19 cases from various

pneumonia categories. The models involved in this analysis include MobileNetV2, InceptionV3, ResNet50, NASNetMobile, Xception, VGG16, DenseNet121, and InceptionResNetV2. All models were fine-tuned by substituting the original network head with a set of new layers. The study in [75] presents a DCNN framework for classifying CXR and CT images using TL. The goal is to carry out binary as well as multi-class categorization to recognize pneumonia and COVID-19 cases. Two varieties of TL are employed: the initial one involves fine-tuning the DenseNet201, DenseNet169, DenseNet121, ResNet152, ResNet50, VGG19, and VGG16 models. The second one focuses on deep-tuning models such as AlexNet, LeNet-5, VGG16, and Inception naive v1. Through the combination of standard data augmentation approaches and generative adversarial networks (GANs), the authors of this method have not only solved the issue of data constraints but have also made it possible to extract features more deeply by applying various filter banks, including the Sobel, Gabor filters, and Laplacian of Gaussian (LoG) [41].

A new two-phase method for categorising CXR images is presented in [76]. This method involves training a DCNN to function as a feature extraction tool in the first step and employing Extreme Learning Machines (ELMs) for real-time identification in the second. This work employs the Chimp Optimisation Algorithm to preserve real-time capabilities while enhancing performance and network stability. Due to the scarce quantity of COVID-19 instances, the majority of the studies have an issue with class imbalance. Chamseddine et al. looked at the Synthetic Minority Oversampling Technique (SMOTE) and Weighted Categorical Loss on each dataset independently to resolve the challenge of uneven class distribution. Six state-of-the-art CNNs were trained by transfer learning [77]. Dhere and Sivaswamy trained their suggested architecture [78] using a new conicity-based loss function and a novel multi-scale attention architecture known as Multi-scale Attention Residual Learning (MARL). In the first step, pneumonia cases were distinguished from normal instances using a DenseNet-derived model, and in the second stage, COVID and NCP cases were distinguished using the MARL architecture. The study in [79] presented MAG-SD, a Multiscale Attention Guided deep network with Soft Distance regularization, designed to classify COVID-19 from pneumonia CXR images automatically. To enhance the model's resilience and address the limited training sample, the approach incorporates attention-guided augmentations alongside soft distance regularization. This strategy aims to create meaningful augmentations while minimizing noise. In [80], through TL, a CNN classifier has been utilised to distinguish between the

COVID-19 and normal-healthy image categories. They propose a DenseNet model and employ the idea of early stopping to improve its accuracy.

Most of the existing COVID-19 literature emphasises the utilisation of individual networks for feature extraction, with each network equipped to extract features in a distinctive way. The usefulness of multi-CNN, a concatenation of multiple pre-trained CNN architectures, for the automatic recognition of COVID-19 from CXR pictures is examined in [81]. For the purpose of predicting COVID-19, this method integrates features retrieved from multi-CNN with the Bayesnet classifier and correlation-based feature selection technique. To create the model in [40], a number of pre-trained architectures and their combinations were applied. This approach employs features obtained from networks that have been pre-trained, a feed forward neural network for COVID-19 detection, and a sparse autoencoder for dimensionality reduction. Sharma et al.[82] presented COVDC-Net, a categorization system inspired by Deep Convolutional Networks technique that can recognise SARS-CoV-2 infection based on CXR images. In order to enhance classification accuracy, this method combines two modified pre-trained networks, VGG16 and MobileNetV2 without their classification layers. Both of these two models are consequently fused utilizing the confidence fusion technique. In [83], the author proposed a method for classifying CXR images by blending features from VGG16 and DenseNet models. The approach integrates an attention mechanism to extract deep features. Additionally, ResNet was employed to isolate relevant image data, ensuring accurate and efficient classification. The work in [84] integrates the strengths of various networks, specifically EfficientNetV2 and ResNet, for feature fusion. It incorporates spatial and channel attention units to improve the network's feature extraction capabilities. Additionally, Grad-CAM++ is employed to provide more intuitive feature visualization and improve the DL model's interpretability.

In the realm of biomedical image classification, there have been promising developments as of late, particularly when incorporating a fusion of handcrafted features with DL techniques. A novel categorization framework that combines features from a DCNN and a hand-picked set of features has been proposed in [19]. Three components make up this framework. They use ResNet-50 to train the network in the first module. They build a pool of manually selected features in the second module based on frequency and texture, which are then further reduced using PCA. The authors in [85] have developed an innovative

framework, called HANDEFU, to support feature extraction methodologies for feature engineering that are deep, handcrafted, and fusion-based. The framework includes handcrafted feature extraction algorithms such as GABOR, HOG, and LBP. The framework includes well-known DL approaches like DenseNet121, AlexNet, VGG19, VGG16, ResNet50, InceptionV3, and shallow-fully connected network models, as well as deep-fully connected and shallow-convolutional network models. The FM-HCF-DLF model [86], a unique fusion model developed by Shankar and Perumal using DL characteristics, is presented for the identification and categorization of COVID-19. The model comprises three primary phases: FM for extraction of feature and classification, and preprocessing based on Gaussian filtering. The FM model uses the CNN-based Inception v3 approach and combines handcrafted features with DL features with the aid of LBP. The learning rate scheduler employing Adam Optimizer is applied to enhance the potency of the Inceptionv3 structure even more. Finally, the classification process is completed using MLP. In the research [87], the authors present an innovative fusion model that combines hand-crafted features with DCNNs to categorize ten chest conditions including COVID-19 and normal conditions, making use of CXR images. Initially, the Info-MGAN network is applied to segment the original CXR data, creating lung images corresponding to the ten chest ailments. In the next step, these segmented images of lung are processed through a pipeline that extracts distinctive features utilizing hand-crafted methods like ORB and SURF. The extracted features are then integrated with trained DCNNs. Finally, different ML models are being used as the final layer of the DCNN architectures to classify chest diseases.

The study in [88] introduced a novel meta-heuristic driven fusion framework for diagnosing COVID-19. Initially, the Wiener filtering technique is employed for image preprocessing. Next, feature extraction is performed using a fusion approach that integrates GLCM, LBP, and Gray Level Run Length Matrix. The Salp Swarm Algorithm is then employed for selection of an optimal group of features. Finally, ANN is used to classify and distinguish between infected and healthy individuals. In the study [89], the authors introduced an enhanced hybrid categorization technique for COVID-19 pictures, leveraging the benefits of CNNs for feature extraction and a swarm-based feature selection (FS) method, the Marine Predators Algorithm (MPA), for selecting the most pertinent features. They combined fractional-order calculus, a powerful mathematical tool, with MPA to create an integrated approach for improved feature selection. In the study [90], an extremely effective detecting system was developed using three

distinct CNN models—ResNet101, ResNet50, and InceptionResNetV2—along with X-ray images categorized into three different classes. FS was performed using the Ant Colony Optimization (ACO) and the Particle Swarm Optimization algorithms, and their results were evaluated and compared. Ref. [91] presented an alternative approach to identify COVID-19 instances apart from other normal and abnormal instances by extracting significant features and applying a novel FS method to identify the most relevant features. An enhanced Cuckoo Search optimization technique is proposed to optimize the multiclass classification of COVID-19 instances. With the use of CNN and an enhanced grey-wolf optimizer with genetic algorithm, also known as CXGNet, the work [92] suggests a tri-stage COVID-19 categorisation model based on CXR images.

In the research [93], the authors introduced CoroNet, a DCNN intended to identify COVID-19 infections employing CXR images automatically. The model is built upon the Xception architecture and trained end-to-end on a dataset compiled from two publicly available sources, containing X-ray images showing COVID-19 and additional forms of chest pneumonia. CoroNet employs Xception as the base model, incorporating a dropout layer along with two additional fully connected layers at the final stage. The authors in [94] developed an end-to-end model by customizing the original ResNet18 architecture to categorize and predict COVID-19. While the standard ResNet18 model is designed for colour images, the customized version is particularly well-suited for the grayscale images used in this study for diagnosing infections. In the modified ResNet18, the average pooling layer of the initial architecture is substituted with a Global Average Pooling (GAP) layer, and two additional compression layers are introduced after the GAP layer An enhancement of the ResNet50 architecture is presented in [95] to categorize patients as COVID-19 negative or positive. This enhancement involves appending three additional layers: Convolution, Batch Normalization, and Activation ReLU. These layers are integrated into the ResNet50 framework to enhance precise differentiation and resilient feature extraction. Comprehensive investigations were carried out to assess the efficacy of the architecture. The study in [96] introduced a modified version of MobileNet and a revised ResNet model for categorizing COVID-19 CXR and CT pictures respectively. Specifically, a CNN customizing technique is developed to address the gradient vanishing issue and enhancing classification accuracy through the dynamic integration of features from various network layers. This enhanced version MobileNet is utilized to classify COVID-19, Tuberculosis, various types of pneumonia, along with healthy controls based on CXR images. The study [97] intends to design

a modified DL-based CNN network to enhance the interpretation of CXRs for improved classification of COVID-19 cases. To achieve this, the authors proposed a modified CNN technique that utilizes X-rays for COVID-19 classification by integrating ResNet50V2 and VGG19 architectures, with extra dense layers incorporated into the merged feature extractors.

The majority of DL research focus on feature extraction using single network models, even though each network model has the potential to capture features in its own unique way. It is thought that by combining deep characteristics from several models, the feature might be made more liberal. Consequently, the creation of a trustworthy and precise DL model for COVID-19 diagnosis using CXRs is highly demanding when data is limited, it becomes very difficult to extract a complete set of features and needs attention. Moreover, combining deep features with handcrafted features can enhance the effectiveness of COVID-19 image classification by leveraging the complementary strengths of both approaches, leading to improved accuracy and robustness. The literature review revealed a dearth of published work integrating CNNs with hand-designed feature maps and requires more investigation.

2.3 Deep learning based approaches on skin cancer detection using dermoscopic images

Dermoscopy is a type of imaging employed for examining the specific area of interest in the skin by applying required magnifying and removing any reflection from the surface. Implementing this method becomes beneficial for the prompt diagnosis of SC, which contributes to a reduced mortality rate. Visual inspection of the lesion without the use of magnification may be a laborious, subjective, and imprecise process. A study revealed that the correct identification of lesion class, just through physical examination, requires dermatalogists to process substantial knowledge and expertise, so it is strongly avoided. DL is globally employed for the analysis of biomedical images. An accurate categorization of SLs by DL presents several obstacles. The presence of high-quality samples is a fundamental prerequisite for achieving precise categorizing outcomes when employing DL schemes. While most pre-trained CNN models have undergone training using massive quantity of image samples, the amount of images available for SL examination is limited to a few thousands or less. An additional challenge

arises from the significant degree of similarity among multiple categories and the substantial variability within each class. Moreover, these factors add complexity to the visual assessment procedure.

Numerous efforts have been undertaken to address the complex challenges associated with the classification of SLs.

Zhao et al. [98] suggested a novel SL image categorization framework, called SLA-StyleGAN, incorporating an SL augmentation style-based GAN, to address the intraclass imbalanced datasets issue. It follows the fundamental design principles of DenseNet201 and style-based GANs. To modify the generator and effectively create high-quality SL images, the suggested framework rebuilds the discriminator and reorganises the original generator's style control and noise input structures. They present a novel loss function that can enhance balanced multiclass accuracy by decreasing the distance between samples within the same class and increasing the sample distance across classes. The authors in [99] conducted TL and fine-tuned CNNs by training EfficientNets B0-B7. For EfficientNets B0-B5, all layers were fine-tuned using the pre-trained ImageNet weights. However, for EfficientNets B6-B7, a two-step fine-tuning process was applied. In the first step, only the newly added layers were fine-tuned while the convolutional blocks remained frozen. In the second stage, the endmost four convolutional blocks of the primary representation were unfrozen, while the remaining blocks stayed frozen, and fine-tuning was performed once more. The dataset was artificially augmented using techniques such as rotation, zooming, and both horizontal and vertical flipping. The suggested model in [100] efficiently synthesises high-quality SL images by modifying the configuration of style control and noise input in the initial basic style-based GAN generator and adjusting both the generator and discriminator. Regarding image classification, a TL process is employed to build the classifier on the pre-trained Resnet-50. The research in [101] presented a model for highly accurate SL classification, utilizing TL with the pre-trained GoogleNet model. The initial parameters of the model, obtained from GoogleNet, were adjusted through training. The potency of the model was assessed utilizing the well-known ISIC 2019 dataset, which included various SLs. To boost the generalization of the model, several augmentation techniques were applied, including vertical and horizontal shifts, flips, and rotations. The research in [102] introduced a new deep TL model based on MobileNetV2 for melanoma classification. An assessment of the efficacy of the model was conducted utilizing the ISIC 2020 dataset. Several data augmentation methods were employed,

namely, rescaling, width shifts, rotations, shear transformations, horizontal flips, and channel shifts, among others.

To increase the accuracy of SL classification, an explainable artificial intelligence based approach is suggested in [103]. They employ the transfer learning algorithm ResNet-18 for extraction of feature and the image augmentation techniques of flipping, rotation, and cropping. In order to produce visual explanations that align with pre-existing beliefs and general best practices for explanations, the predictions are analyzed additionally employing the Local Interpretable Model-Agnostic Explanations method. In order to classify SC, Pacheco et al. in [104] address the challenge of merging images with metadata information using DL models. They suggest a unique technique called the Metadata Processing Block, which enhances the most pertinent features during the categorization pipeline to leverage metadata to support data categorization. They used data augmentation with typical image processing operations, like vertical and horizontal flips, changes in brightness, contrast, and saturation, scaling, and the addition of random noise. While DCNNs have achieved significant progress in numerous image classification applications, the shortage of sufficient data for training, intra-class variation, inter-class similarity, and the insufficient emphasis on semantically significant regions of lesions make accurate SL classification a persistent challenge. The authors [105] suggest an attention residual learning CNN (ARL-CNN) technique, which consists of several ARL blocks to address these problems. To enhance its capacity for discriminative representation, every ARL block combines unique attention-learning techniques with residual learning. They used online data augmentation techniques, comprising random rotation, zoom, and flips (horizontal and vertical), to expand the training samples. In order to effectively represent discriminative information from many image perspectives using a sensible weighing method, this research suggests a multi-view filtered transfer learning network. In addition, this approach evaluates the importance of each source image, which can provide valuable insights by excluding unfavourable examples from the source domain [106]. Insufficient training data, melanoma and nevus similarities, and insufficient robustness continue to pose challenges to the proper categorization of SLs. Three long attention networks (LANet) are combined to form the multi-scale long attention network (MSLANet), which the authors propose as a solution to the problems. By using a lengthy attention mechanism, each LANet can integrate context knowledge and enhance its ability to represent discriminatively. Both feature-level and instance-level multi-scale information can be used concurrently by MSLANet. Furthermore, they suggest

a depth data augmentation technique, which when trained, can enhance the model's capacity for generalisation [107].

The work in [108] presents a new CAD system that combines DL features using mutual information measures with handcrafted characteristics related to the medical algorithm ABCD rule. Shape, colour, and texture are examples of handcrafted characteristics that are utilised to represent the ABCD rule. A CNN architecture is then used to extract DL features. As a fusion rule, MI measurement is used to extract the most crucial data from both kinds of characteristics. SMOTE is used as data augmentation method. In Ref. [109], the authors presented a new intelligent system that integrates multiple neural networks across two classification stages. The first stage includes five classifiers: a perceptron combined with colour LBPs, a perceptron integrated with colour HOGs, a GAN paired with the ABCD rule, ResNet, and AlexNet. These classifiers were chosen experimentally, considering melanoma characteristics like texture, shape, colour, etc. At the second level, a single classifier identifies whether the lesion is melanoma relying on the modified weights learned from the decisions of the first-stage classifiers. Data augmentation involved four 90-degree rotations. To increase the efficacy of convolutional neural networks (ConvNets),Ref.[110] introduces a cascaded ensemble network that combines ConvNet with handcrafted features through a multi-layer perceptron. The proposed model extracts non-handcrafted features from images using ConvNet, while simultaneously incorporating colour moments and texture features as handcrafted attributes. The colour images are represented using the RGB channels, and four statistical moments— standard deviation, mean, skewness, and kurtosis are calculated for every channel. Additionally, five GLCM features, namely contrast, correlation, energy, homogeneity, and dissimilarity, are derived. To address the class imbalance, SMOTE is applied to the minority classes for data augmentation. In the study [111], a vigorous SC identification model was suggested, built on feature fusion. Features were manually extracted using LBP, while Inception V3 was employed for automatic feature extraction. Additionally, the Adam optimizer was utilized to adjust the learning rate. During the final step, an LSTM network was applied to the fused features to classify SC as malignant or benign. Various data augmentation methods, like rotation, flipping, and shearing, were performed to expand the volume of training data.

Another StyleGAN based data augmentation technique is adopted to pro-

duce high-quality images to address the issue of the dermoscopy image dataset's unequal and sparse distribution [112]. Three CNN architecture viz. InceptionV3, ResNet50, and VGG16BN are considered through transfer learning. In order to finally make choices using numerous blocks, they employ the block to combine multiple CNNs. Compared to the classic fusion technique, the decision fusion method is more stable and robust, and it can solve the generalisation capability of individual CNN models. In the study [113], the authors introduced an SL classification system that leverages DL techniques and collective intelligence by utilizing multiple CNNs. The CNNs selected for this system, based on their performance, include GoogLeNet, AlexNet, GoogLeNet-Places365, Xception, MobileNet-V2, ResNet-50, InceptionResNet-V2, ResNet-101, and DenseNet201. They evaluated the performance of every network to generate a weight matrix, where the elements represent the weights corresponding to the NNs and the lesion classes. Using this matrix, another decision matrix was created to form a multi-network ensemble configuration that consolidates the decisions of every network into a decision fusion unit. This unit is then responsible for making a final, more accurate prediction by considering the weighted outputs of each network. Data augmentation methods like shearing, rotations, mirroring, etc. were applied to improve the model's effectiveness. The research [114] employed the DL method to identify the two main categories of tumours, malignant and benign forms, employing the ISIC2018 dataset. Initially, the images were enhanced and refined using ESRGAN. Several TL models, including InceptionV3, ResNet50, and Inception ResNet, were then fine-tuned. Alongside experimenting with various models, the novel contribution of this study lies in the application of ESRGAN as a preprocessing step. Data augmentation strategies like rotation, reflection, brightness adjustment, etc. were also employed. In the work [115], the authors introduced a novel CNN architecture named Densenet Residual Network, which utilizes contextual data. The principal focus of this model is to surpass other DL algorithms while reducing computational parameters. By integrating the residual concept into the Densenet structure, the Densenet Residual Network achieves improved accuracy in recognition tasks. In this architecture, the nodes' outputs are integrated with their respective inputs. For data augmentation, Gaussian noise was applied, along with techniques such as flipping, shifting, zooming, and rotation at angles ranging from 2 to 12 degrees.

The author in [116] proposed a DNN model by modifying EfficientNet-B4 and EfficientV2-M baseline models. They have included a flatten layer followed by two dense layers of different neuron size. A modified Resnet-50 architecture is

developed in [117] by adding an extra FC layer to the existing network followed by fine-tuning. GANs are used as data augmentation technique. The authors in [118] developed a new convolutional network with dense connections called DenseSFNet-45 by incorporating their developed structural component, the SE-Fire block, into the dense block of DenseNet. SF block enhances DenseNet's representational power. Using DenseSFNet as a foundation, they introduced an innovative two-stage framework that first segments SLs and then classifies them to achieve accurate SL classification. Additionally, they applied rotation techniques (90°, 180°, and 270°) to the melanoma class. The study [119] introduced an enhanced capsule network named FixCaps to classify dermoscopic images. FixCaps features an expanded receptive field relative to conventional CapsNets by incorporating a substantial, high-efficiency kernel in the initial convolution layer, having a kernel size of 31×31, significantly larger than the commonly used 9×9. To minimize the decrease in spatial information caused due to pooling and convolution, the convolutional block attention module was implemented. Additionally, group convolution was employed in the capsule layer to prevent model underfitting. To further augment the training set, translation and other approaches were applied to enhance the quantity of samples. In [120], the author presented a SL classification method utilizing MobileNet. They proposed a modified version of MobileNet specifically adapted for classifying SLs. The final five layers of the original MobileNet were substituted with a Dropout layer and an FC layer incorporating a Softmax activation function. This modification allowed the FC layer to achieve more accurate predictions for each class compared to the original five layers of MobileNet. The complete quantity of parameters in the modified model was decreased from 4,253,864 to 3,236,039, resulting in lower computational time. Data augmentation techniques included rotating images at various angles, zooming to new scales, width, and height shifts, and both horizontal and vertical flips. In [121], the authors introduced a weighted average ensemble learning architecture to categorize 7 SLs. The ensemble was built using five DNN models: ResNeXt, SeResNeXt, ResNet, Xception, and DenseNet. Few layers were appended to these pre-existing networks. Initially, the dense layers of the pre-trained networks were detached, and a GAP layer was introduced. Following this, batch normalization, dropout, and dense layers were included. The initial dense layer contained 512 neurons, while the concluding dense layer had 7 neurons representing the seven categories in the classifier, using softmax activation. For data augmentation, rotation, flipping, shearing, and zooming techniques were employed.

Ref. [45] proposes a new approach to multiclass SL classification utilising an extreme learning machine and the best DL feature fusion. The suggested approach consists of the following steps: contrast enhancement; transfer learning for DL feature extraction; hybrid whale optimisation and entropy mutual information technique for best feature selection; modified canonical correlation-based approach for fusing selected features; and, lastly, extreme learning machine-based classification. The accuracy and computational efficacy of the system are enhanced by the feature selection stage. Afza et al. [122] suggested a hierarchical structure built on DL and 2D superpixels. They start by combining images enhanced at both local and global levels to enhance the contrast of the original images. The subsequent stage uses a tri-step superpixel lesion segmentation method to segment SLs using the whole set of improved images. ResNet-50 model, which uses transfer learning is used to learn new features. An enhanced grasshopper optimisation method is employed to further optimise the extracted features. The suggested approach in [123] has three main stages. Phase I uses the tiny YOLOv2 model, which uses the squeeze net and open neural network as its foundation, to localise several categories of SLs. The features are obtained from the squeeze net's depthconcat7 layer and supplied as an input to the tiny YOLOv2. Phase II involves segmentation using a 13-layer 3D-semantic segmentation model. Later on in Phase III, the ResNet-18 network is utilized for extracting deep features, and the ACO approach is employed to choose optimised features. The authors in [124] presented a DL and optimal FS framework for multiclass SL categorization. Three pre-trained DL networks were fine-tuned and trained employing the TL approach. During fine-tuning, several additional layers were added or removed to decrease the quantity of parameters, and hyperparameters were selected employing a genetic algorithm (GA) in lieu of manual tuning. The goal of using GA for selection of hyperparameter was to enhance learning performance. The deep features extracted were combined employing a new serial correlation-based method, which reduced the feature vector length while retaining minimal redundant information. To further improve FS, the authors presented an enhanced anti-Lion optimization algorithm to address this redundancy.

The literature indicates that most DL-based approaches for SL classification rely on pre-trained networks or basic TL models. Modifying existing network architectures can enhance performance, and combining original and modified models in an ensemble can further improve results. Although some efforts to modify current pre-trained CNN designs have showed promising results in SL classification application, these approaches have not been fully utilized and require further research.