Deep Transfer Learning-Based Approach for Glucose Transporter-1 (GLUT1) Expression Assessment

Maisun Mohamed Al Zorgani[1] · Hassan Ugail[1] · Klaus Pors[2] · Abdullahi Magaji Dauda[2]

Abstract

Glucose transporter-1 (GLUT-1) expression level is a biomarker of tumour hypoxia condition in immunohistochemistry (IHC)-stained images. Thus, the GLUT-1 scoring is a routine procedure currently employed for predicting tumour hypoxia markers in clinical practice. However, visual assessment of GLUT-1 scores is subjective and consequently prone to inter-pathologist variability. Therefore, this study proposes an automated method for assessing GLUT-1 scores in IHC colorectal carcinoma images. For this purpose, we leverage deep transfer learning methodologies for evaluating the performance of six different pre-trained convolutional neural network (CNN) architectures: AlexNet, VGG16, GoogleNet, ResNet50, DenseNet-201 and ShuffleNet. The target CNNs are fine-tuned as classifiers or adapted as feature extractors with support vector machine (SVM) to classify GLUT-1 scores in IHC images. Our experimental results show that the winning model is the trained SVM classifier on the extracted deep features fusion *Feat-Concat* from DenseNet201, ResNet50 and GoogLeNet extractors. It yields the highest prediction accuracy of 98.86%, thus outperforming the other classifiers on our dataset. We also conclude, from comparing the methodologies, that the off-the-shelf feature extraction is better than the fine-tuning model in terms of time and resources required for training.

Keywords GLUT-1 scoring · Deep transfer learning · Colorectal cancer · IHC image analysis · Tumour hypoxia

Introduction

It is known that most solid tumour tissues have an abnormal vasculature resulting in poor delivery of oxygen to localised regions. Therefore, these regions are characterised by low oxygen concentration, known as tumour hypoxia [1]. Hypoxia is currently receiving attention as the centre for the hallmarks of cancer; this is because of its major characteristics of chemotherapy and radiotherapy resistance and a major prognostic factor [2]. Oxygen gradient in the order of diffusion and consumption plays a critical role in the dynamics of the tumour microenvironment, and this creates room for the development of hypoxic regions in almost all solid tumours [3]. The hypoxic regions in all solid tumours vary in size and extent due to the different gradients and supply as well as distribution of oxygen in these regions [4]. The need for oxygen in solid tumours differs; some tumours that are irregular and disorganised require a reduced oxygen gradient, whilst others require an increase in oxygen demand which arises as a result of tumour metabolism [5]. The imbalance between oxygen demand and supply in solid tumours makes hypoxia a unique hallmark of cancer as it creates room for tumours to more adverse features that include epithelial–mesenchymal transition (EMT) and mesenchymal–epithelial transition that produces cancer stem cell (CSC) niche, resistance to chemotherapy and radiotherapy resistance, poor clinical prognosis and immune damping as well as increase genomic instability and increases cells abilities to evade apoptosis [4, 6]. Hypoxia and angiogenesis work hand in hand for all solid tumours to be fully established and therefore are linked to aggressiveness and metastasis of solid tumours, hence among

✉ Maisun Mohamed Al Zorgani
   Maisunmalzorgani@gmail.com

   Hassan Ugail
   h.ugail@bradford.ac.uk

   Klaus Pors
   K.Pors1@bradford.ac.uk

   Abdullahi Magaji Dauda
   A.D.Magaji@bradford.ac.uk; dramdauda@gmail.com

1   Faculty of Engineering and Informatics, School of Media, Design and Technology, University of Bradford, Richmond Road, Bradford BD7 1DP, UK

2   Institute of Cancer Therapeutics, University of Bradford, Richmond Road, Bradford BD7 1DP, UK

the most relevant micro-environmental factors [5]. Tumour hypoxia leads to the production and activation of hypoxia-inducible factor (HIF) family members, including HIF-1 and HIF-2; these HIFs' proteins later activate more genes that promote hypoxia such as the GLUT-1 and carbonic anhydrase-9 (CAIX) [7] which are both well-known membrane-bound biomarkers that induce and increase the hypoxic condition in solid [8]. Thus, the GLUT-1 as a cell surface receptor is one of the genes associated with tumour hypoxia in different forms of cancer including breast, prostate and colorectal and hence classed as a hypoxic stem cell marker cancer [9]. The semi-quantitative scoring system is employed for estimating the amount of GLUT-1 expression. Several studies [9–12] have employed the semi-quantitative scoring system to assess the level of hypoxia marker using the GLUT-1 protein expression. The clinical significance of the different scores of tumour hypoxia therapy is aimed at understanding how much hypoxia is developed and spread across solid tumours and cells. The deeper the hypoxic environment, the more difficult it gets for chemotherapy and radiotherapy to reach tumour sites. Hence, there is a need for the use of hypoxia-activated pro-drugs (HAPs) as targeted therapy. For the apparent progress in cancer therapies, the visual examination still is the standard way of measuring tumour hypoxia markers in clinical practice, which is affected by inter-observer variation [13–15]. In addition, commercial software such as QuPath imaging software requires an expert pathologist to quantify the intensity of GLUT1 ImmunoStaining in tumour cells. Therefore, one of the steps of tumour therapy development would be promoted by a robust automated scoring method to facilitate clinical tests, improve the diagnosis quality, avoid the variation among histopathologists and reduce the time and costs of diagnoses. This, in turn, results in better assessment outcomes and improved patient experience [15].

Traditional machine learning-based methods rely heavily on extracting some specific visual features manually from images. So, such hand-crafted features can only deal with some low-level information about images. Furthermore, the relevant domain knowledge is necessary to select useful features, which can be greatly muddled by subjective extractor bias. In contrast, deep learning-based methods can extract high-level abstract features from images automatically in a standardised way. CNNs are one form of deep learning that can extract the image's hierarchical features at multiple layers, i.e. the features at each layer are computed from the representations of the previous layer. Such hierarchical features can be learned gradually from low-level to high-level through a deep architecture [15]. Thus, multi-level abstraction enables deep learning networks to be well suited for discovering the complex structures within high-dimensional data, such as whole slide images [16]. In addition, such CNNs have the ability to learn complex mapping functions directly from input data, without the help of human-crafted features [16].

Although deep learning algorithms achieved state-of-the-art results in the digital histopathology domain, they have some unique challenges in their implementation. One of the major challenges is the large number of annotated images needed for training deep CNNs that may not be available, especially in the digital histopathology domain [16]. In contrast, training deep CNNs with limited training data leads to over-fitting and a poor generation of features on data. Over-fitting is critical when the data contain high image appearance variance, which is usually common in digital histopathological images. Furthermore, training deep CNN from scratch requires high computational costs and extensive memory resources and time. Thus, such approaches have practical limitations in the digital histopathology field [14]. In recent years, there has been a debate that the methodology of deep transfer learning could tackle the aforementioned challenges more effectively. Deep transfer learning is leveraged as a helpful tool to overcome the data scarcity problem in the digital histopathology field. Therefore, this work has adopted the most common deep transfer learning to present an automated method for scoring the tumour hypoxia marker using the GLUT-1 protein expression. Towards the end, this work proposes alternative tumour hypoxia scoring evaluation using deep learning technology to objectively predict those patients that require targeted therapy HAPs instead of chemotherapy or radiotherapy.

The rest of the paper is organised as follows: the related works are briefly discussed in the second section. The proposed methodology is explained in the third section. The carried out experiments and their results are presented in the fourth section. Lastly, the fifth section concludes the paper.

Related Work

The traditional IHC scoring methods rely on hand-craft features. This is due to the lack of IHC images in clinical practice. So IHC images need the high cost of antibodies, autostainer machine equipment and the complex laboratory process. To our knowledge, no work has been done for the automatic GLUT1 scoring using deep learning techniques. Therefore, this section reviews some of the works that are similar to our work. These works have participated in the contest human epidermal growth factor receptor 2 (HER2) scoring challenge on invasive breast cancer images. It was organised by the University of Warwick, the University of Nottingham and the Academic–Industrial Collaboration for Digital Pathology. It aimed to advance automated methods for HER2 scoring on IHC-stained images. HER2 expression is used as a predictor of invasive breast cancer progression in clinical practice. More information on the HER2 Challenge dataset can be found in the paper published by Qaiser et al. [17].

Since the HER2 scoring challenge contest, several studies [18–21] have been carried out by using traditional machine learning methods for HER2 scoring. Cordeiro et al. [18] have utilised and compared the SVM, K-nearest neighbours (KNN), multi-layer perceptron classifier (MLP) and decision tree classifiers for HER2 scoring according to image patch level and patient level with colour and texture features. Mukundan [19] has employed uniform local binary pattern (ULBP), characteristic curves, entropy and energy features with logistic regression and SVM classifier to score HER2-stained tissue samples. Tewary et al. [20] have utilised colour space-based membrane extraction followed by an SVM classifier for HER2 scoring. Chang et al. [21] have employed the colour channel to extract the morphology, texture and intensity features, and then they were utilised for training the SVM classifier. But these methods depend on extracting visual features manually from images which in turn requires an expert pathologist.

The others [17, 22–25] investigated deep learning methods. This is due to deep learning methods providing impressive results in many applications of histopathological image analysis. Starting from [17], Qaiser et al. have presented the winning teams of the HER2 challenge contest that employed CNN-based methods. The winning teams leveraged deep transfer learning methodologies to address the lack of annotated training images in the HER2 challenge dataset. They have fine-tuned the latest pre-trained CNNs at the time as classifiers for HER2 scoring. For instance, the MUCS team [17] submitted three versions, the AlexNet model was adapted to MUCS-1 and MUCS-2, and the GoogLeNet model was adapted to MUCS-3. MUCS-1 network is fine-tuned to classify four output classes that correspond to HER2 scores from 0 to 3 +. MUCS-2 and MUCS-3 had an additional output class for the background. The background class contained the regions with texture having only a weak appearance of nuclei (without blueish or brownish colour). MTB NLP team [17] trained the modified architectures of AlexNet and VGG-16 and then employed a random forest classifier to produce the final class probabilities for each score. In a similar manner, the VISILAB team [17] fine-tuned the GoogLeNet model, and the FSUJena team [17] fine-tuned the AlexNet model to four HER2 scores. Whilst in [22], the authors have used direct feeding to the AlexNet model for HER2 scoring. In [23], the authors have proposed Her2Net architecture with LSTM recurrent network for segmenting and labelling the HER2-stained tissue samples. In [24], the authors have applied super-pixel-based tissue region segmentation to extract colour and texture features and then followed by SVM to distinguish epithelial and stromal regions, which are scored using a modified UNet model. In [25], the authors utilised deep transfer learning with fine-tuned five pre-trained CNNs by fully connected dense layers for 3 classes and then presented a collective voting scheme

for HER2 scoring. They have adopted VGG16, VGG19, ResNet50, MobileNetV2 and NASNet-Mobile architectures for image-based and patch-based labelling.

In this work, we have explored the common deep transfer learning methodologies to score GLUT-1 protein expression in a colorectal cancer tissue microarray (TMA) of IHC images. We have investigated and compared the predictive performance of six different pre-trained CNNs architectures: AlexNet, VGG16, GoogLnet, ResNet50, DenseNet-201 and ShuffleNet. These architectures were fine-tuned as classifiers or adapted as extractors to get the transferable off-the-shelf features on our IHC image dataset.

Materials and Methods

In this section, we first present the acquisition approach of dataset images and then outline the essential steps to be followed in dataset image preparation. Next, we show the appropriate choice of architecture and methodology for our application. Finally, the proposed methodology for the discrimination of the GLUT1 scoring is quantitatively explained.

Image Dataset Acquisition

The images were obtained from the institute of cancer therapeutics, the University of Bradford, and the necessary ethical approval has been obtained. The images were prepared by immunohistochemistry (IHC) staining for ALDH7A1 and GLUT-1 on HT-29 tissue microarray (TMA) of human clinical specimens of colorectal cancer adenocarcinoma. The IHC is a widely used technique in pathology. It is now used in all aspects of modern research to identify specific antigens within a tissue section from formalin-fixed paraffin-embedded (FFPE) tissue, e.g. in tissue microarrays (TMAs) and 3D dimensional spheroids grown from cells. The method utilises an antigen-specific antibody interaction and detection using a light microscope [26]. The TMA clinical sample slide number G063 (Biomax. us) carries 150 cores of clinical sample on the whole side in which 100 are colorectal cancer (CRC) tissues, and 50 were either malignant, adjacent tissue to the cancer tissue or normal tissues; these give a total number of 50 cases of colorectal cancer in each whole TMA slide. A whole slide IHC analysis was done to analyse ALDH7A1 protein expression in CRC clinical tissues. The clinical samples were collected between July and August 2019 from thirty-three male patients and thirteen female patients. So, these clinical samples whole slide comprises both male and female colorectal cancer patients and different age distributions from the highest age distribution of 82-year-old male patient sample with grade 1 and stage IIB CRC, whilst the least clinical sample on the whole slide was from a 33-year-old female patient sample with grade 1 and staged IIB CRC (CO1505). The IHC images were scanned using Aperio Digital
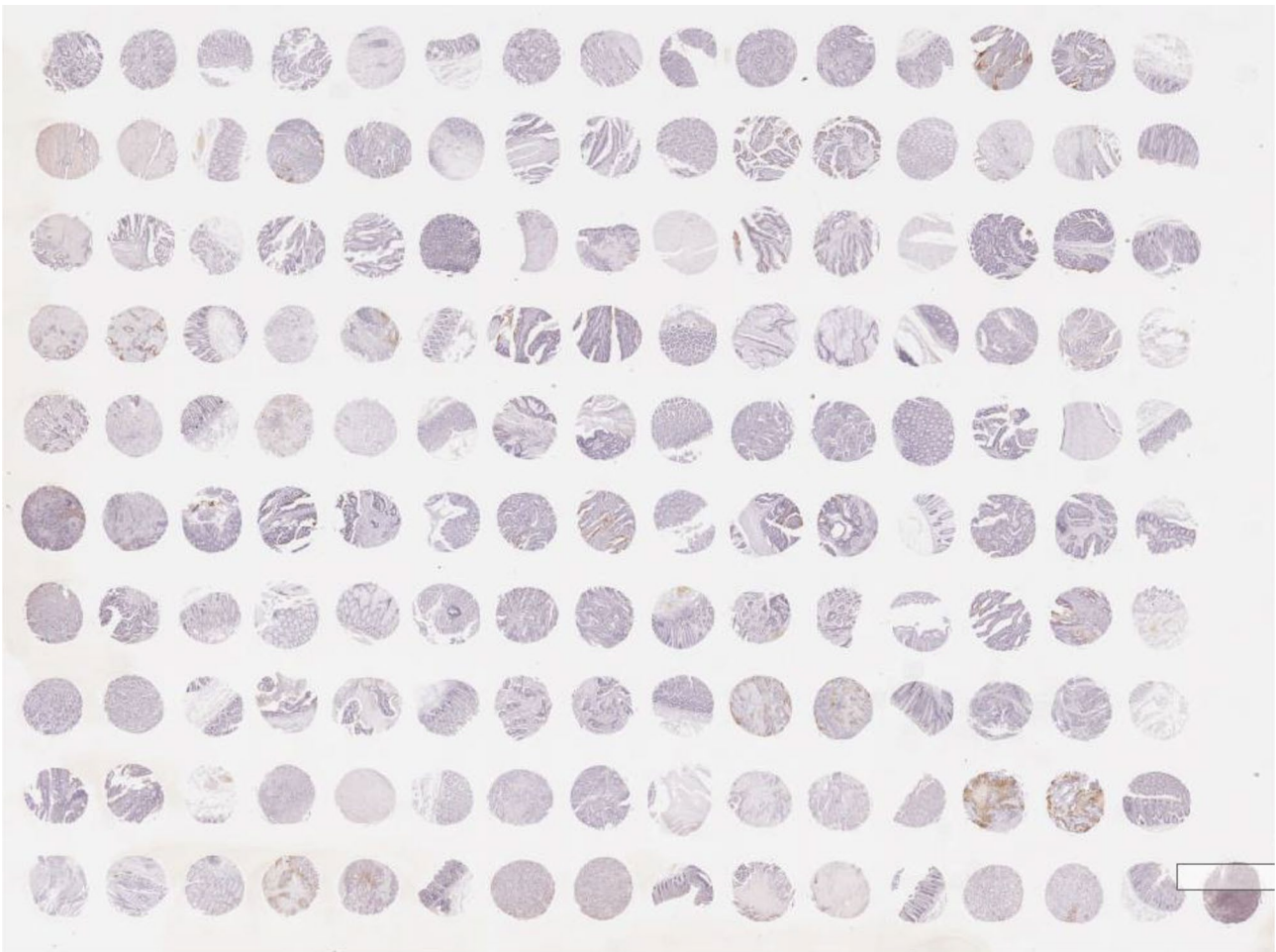
**Fig. 1** The G063 clinical sample of 150 tumour cores of Aperio Leica software

Pathology Slide Scanners (Aperio AT2) and then captured at 0.5 um/pixel and 200 μm diameter. The whole cores and examples of GLUT-1 expression of IHC colon adenocarcinoma images were shown in Figs. 1 and 2, respectively.

Annotation of Dataset Images

Two trained histopathologists scored the TMA cores according to the proportion of Glut-1 and ALDH7A1 staining in the entire cores. They have quantified the different scores using QuPath Imaging Software with version 0.2.6. Such software provides percentage (%) mean expression of protein measured as intensity per pixel. So they have analysed all the semi-quantitative scores per expression of GLUT-1 and ALDH7A1 at different stages of CRC TMA using the QuPath Imaging Software version 0.2.6. The TMA cores were semi-quantitatively scored: score 0 (no staining with 0%), score 1 (light staining with 0–5%), score 2 (medium

Fig. 2 Examples of GLUT-1 expression at different stages of IHC colon adenocarcinoma images; magnification of the upper panel is ×5 (0.1 um/pixel), and the lower panel is ×20 (0.5 um/pixel)
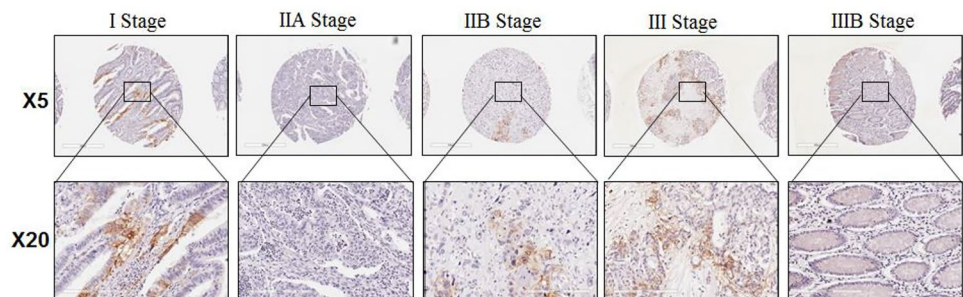
staining with 5–15%), score-3 (heavy staining with15–30%) and score 4 (intense staining with > 30%) and a score given for whole section. The settings used were TMA de-arrayer (optional) at TMA core diameter of 1.2 mm, intensity parameters threshold of 0.1 and Max background intensity of 2. The intensity threshold parameters are under score compartment; cytoplasm, DAB OD mean, threshold $1+0.2$, threshold $1+0.4$ and threshold $1+0.6$. And all of these were ran at single threshold. The examples of our dataset images are shown in Fig. 2.

Derivation of Dataset Images

The dataset images were obtained by hand-picking the regions of interest from TMA cores that contained the most representative samples from each class. The regions were selected at a low resolution and mapped to the highest resolution (0.5 um/pixel) to generate the patches. We derived approximately ten to twelve images for each TMA core to balance the number of images among the classes. A total of 1750 IHC images were extracted at 0.5 um/pixel. Each class contains 350 images, where we derived dataset images with size 512×512 pixels and then stored them in jpeg compression format.

Colour Normalisation

For this purpose, we utilised a colour de-convolution method described in [27] for highlighting the brown-stain areas (Diaminobenzidine, DAB) of reactive membranes in the IHC image.

Partition of Dataset Images

The dataset images were divided randomly into 80% training set (1400 images) and 20% testing set (350 images); i.e. each class was trained with 280 images and tested with 70 images. The testing set does not utilise for training our proposed architectures.

Augmentation of Dataset Images

Several studies [16, 28] have observed that the limited amount of training data is one of the major challenges in deep learning. In this regard, data augmentation is one of the possible solutions to create additional artificial training images through some transformations for learning deep features from the images and thus increasing the deep network performance. In this study, we augmented the training images of our dataset by rotated them with angles of 90, 180 and 270 degrees and then flipped them in the horizontal and vertical direction. This is to enlarge the training images size without affecting the quality of input images [29] and avoid the problems of over-fitting and poorly generation of features [30].

Standard Performance Evaluation Metrics

The performance of different classifiers is evaluated according to standard metrics formulated for multi-class classification as depicted in Eqs. 1, 2, 3 and 4:

- Accuracy metric: this criterion is used to measure a classifier's ability for predicting actual classes correctly. It is formulated as

$$Acc = \frac{1}{L} \sum_{i=1}^{L} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}, \tag{1}$$

- Recall metric: it is also known as sensitivity; this criterion is used to measure a classifier's ability for predicting each individual class correctly. It is formulated as

$$R_i = \frac{tp_i}{tp_i + fn_i}, \tag{2}$$

- Precision metric: this criterion is used to measure a classifier's ability for predicting relevant instances for each individual class. It is formulated as

$$P_i = \frac{tp_i}{tp_i + fp_i}, \tag{3}$$

- F1-score metric: this criterion is a harmonic mean of the recall and precision metrics. It is formulated as

$$F_1 score = \frac{2 \times P_i \times R_i}{P_i + R_i}, \tag{4}$$

where tpi is the number of true positives (i.e. the correctly classification for ith class), fpi is the number of false positives (i.e. wrongly classification for ith class), fni is the number of false negatives (i.e. missed classification for ith class), tni the is the number of true negatives (i.e. correctly classification not belong to ith class), and L is the number of classes.

In this study, standard performance metrics were extracted from the confusion matrix [31]. Thus the 5×5 confusion matrix for classifying five-class that utilised in this work is shown in Fig. 3. So the accuracy, recall and precision could be calculated using the following equations:

$$ACC = \frac{\sum_{i=0}^{L-1} tp_i}{\sum_{i=0}^{L-1} tp_i + \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} E_{ij}}, \forall i \neq j, i = j = \{0, 1, 2, 3, 4\} \tag{5}$$

$$R_i = \frac{tp_i}{tp_i + \sum_{j=0, \forall i \neq j}^{L-1} E_{ij}}, i = \{0, 1, 2, 3, 4\} \tag{6}$$

$$P_i = \frac{tp_i}{tp_i + \sum_{j=0,\forall i \neq j}^{L-1} E_{ji}}, i = \{0, 1, 2, 3, 4\} \tag{7}$$

From Fig. 3, the term E_{ij} is the error classification to the actual class i according to predicted class j. By substituting into Eqs. 1, 2 and 3, we find that $tp_i + fp_i = \sum_{i=0}^{L-1} tp_i$, $fn_i = \sum_{j=0,\forall i \neq j}^{L-1} E_{ij}$ and $fp_i = \sum_{j=0,\forall i \neq j}^{L-1} E_{ji}$.

Choice of Deep Transfer Learning Methodology

In this section, we introduce the concept of two common methodologies.

Fine-Tuning CNN

In this methodology, the target pre-trained CNN network layers can be replaced by new ones and then retrained on the new dataset. It is possible to fine-tune all the network layers or fix the earlier layers and only fine-tune the last layers of the network [32]. The weights in transferred layers are frozen and initialised from a source pre-trained network, whilst the weights in new layers are trained using a back-propagation algorithm [33]. In the back-propagation algorithm, the weights in the new layers are adjusted by back-propagating the new task's errors into the transfer layers during training. It is faster than constructing a new CNN network, i.e. a pre-trained CNN on millions of images could be taken to retrain for new classification using only hundreds of images [33]. Furthermore, it gains convergence faster than learning from scratch so that it can solve the convergence problem [34].

Fig. 3 A confusion matrix for the five-class classification

Off-the-Shelf Feature Extraction

Such methodology adapts the pre-trained CNNs to extract the deep features which then are utilised to train a separate classifier for prediction without consuming time and effort for training. These extractors are characterised by the generalisation property which enables the deep features to be transferred to other applications [35]. Therefore, the generalisability property is particularly useful when there is not enough dataset for training the CNN from scratch.

Architecture of the Proposed Models

Choice of the appropriate architecture for a specific application is an essential step. In this work, we have investigated six pre-trained CNNs trained on the ImageNet dataset [36]:

- *AlexNet architecture* [13] was the winner of the ImageNet 2012 challenge that popularised CNNs. It contains five convolutional layers followed by rectification, three max-pooling layers and three fully connected layers.
- *VGG architecture* [37] was proposed by the Visual Geometry Group at the University of Oxford. It has a similar architecture to AlexNet with more convolutional layers. In VGG architecture, the convolutional layers use only 3×3 filters, as well the pooling layers employ only 2×2 filters. VGG16 model has thirteen convolutional layers followed by rectification, five max-pooling layers and three fully connected layers.
- *GoogLeNet architecture* [32] contains twenty-two convolutional layers with nine inception blocks and a fully connected layer. The inception module consists of four parallel convolution kernels that process the same input, and the extracted different features are then concatenated at the end. Each inception module has six convolutional layers followed by rectification.
- *Residual architecture* [38] uses skip connections to reduce the effect of the vanishing gradient problem significantly. ResNet-50 has 50 convolutional layers followed by rectification and a fully connected layer. It contains 16 residual blocks with the skip layer element-wise addition layer to enable the network to pass the features from lower to higher levels to acquire more complicated features.
- *DenseNet architecture* [39] utilises the skip connections from each layer to the succeeding layers that promote reusing the features through the entirety of the network. DenseNet-201 contains 201 convolutional layers and a fully connected layer. It has dense blocks that concatenate outputs from all the previous layers as its input.
- *ShuffleNet architecture* [40] contains grouped convolution, channel shuffling and element-wise addition of two inputs. So, it combines the characteristics of residual and

dense blocks. It utilises the channel shuffle operation to overcome the consequences of using group convolutions.

In the AlexNet and VGG-16 models, the feature extraction part is regarded from the input layer to the last max-pooling layer, whereas the classification part is three fully connected layers. Whilst in the other models, from the input layer to the global average pooling layer is the feature extraction part, and the fully connected layer is the classification part. The different blocks for proposed architectures are shown in Fig. 4.

Proposed Methodology

This study has investigated the common deep transfer learning methodologies: fine-tuning of pre-trained CNNs for a new classification and off-the-shelf features extraction from pre-trained models. In the first approach, we initialise the target network with pre-trained weights on the source task and then partially re-training them on the target task. Whilst the second approach extracts the deep features from the source task without re-training the network and then uses them to train a third-party classifier.

Fine-Tuning Target CNN Classifiers

In this experiment, we compare the performance of six different pre-trained models: AlexNet, VGG16, GoogLeNet, ResNet-50, DenseNet-201 and ShuffleNet. We carried out our experiments by fine-tuning these target models as classifiers. We fine-tuned the fully connected layer of these models by replacing the last three layers with new layers for five classes; i.e. the new layers

are mapped to our dataset. We specifically replace the fully connected, softmax and classification layers with three corresponding new layers. Subsequently, the weights in transferred layers were preserved, whilst the weights in new layers were updated continuously using the back-propagation algorithm.

Tuning Target CNN Feature Extractors

In this experiment, we have compared the performance of SVM classifiers that were trained on the extracted deep features from a specific layer in the proposed networks as follows: global average pooling "avg-pool" layer for GoogLeNet, ResNet-50, DenseNet-201 and ShuffleNet, whilst "f7" fully connected layer for AlexNet and VGG16. The SVM classifiers were tuned and trained on the extracted deep features using the training set and then evaluated using the test set. In the sequel, the extracted deep feature vectors from fully connected layer fc7 of AlexNet and VGG-16 will be referred to as Feat-Alex and Feat-Vgg16, respectively. As the same, the extracted deep feature vectors from the global average pooling layer of GoogLeNet, ResNet-50, DenseNet and SuffleNet will be referred to as Feat-GoogLe, Feat-Res50, Feat-Dense and Feat-Suffle, respectively.

Experiments and Results

Three experiments are carried out using our dataset images. The experiments are implemented in MATLAB R2020a on a desktop computer with a 3.60-GHz Intel® CPU, Dual-Core-i7-7700, 32 GB RAM and NVIDIA GeForce GTX 1070
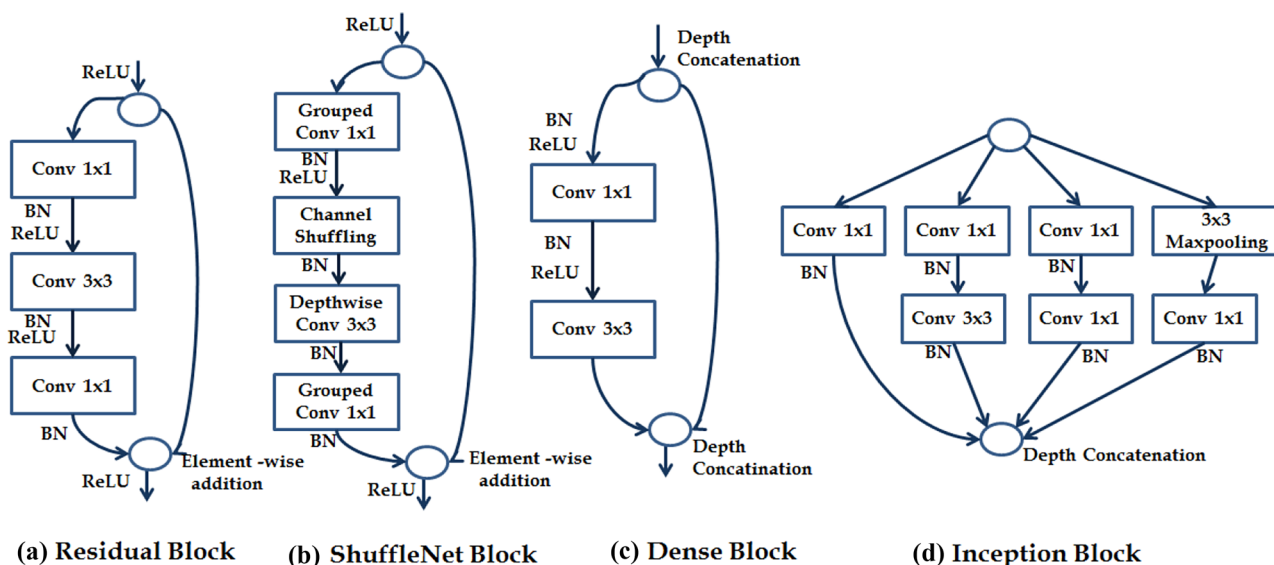

Fig. 4 Block structure for different proposed architectures, where "Conv", "ReLU" and "BN" represents the convolution layer, rectified linear unit and batch normalisation layer, respectively

GPU. In our experiments, the target CNN models were fed with the augmented IHC images, which were resized according to their input layer size. AlexNet architecture requires input images of size $227 \times 227 \times 3$, whilst other architectures require input images of size $224 \times 224 \times 3$, where 3 is the number of colour channels. Now, the dataset images are ready to employ according to the used approach. In the fine-tuning classifier, augmented data is fed to the target networks for training. Whereas in feature extraction approach, augmented data is used to train target SVM classifiers. This will be further explained in the following sections. Figure 5 illustrates the flow chart diagram of how these approaches can be carried out.

Evaluation of Tuned Classifiers

In this section, we discuss the setup of the tuned classifiers and the analysis of obtained results.

Tuned Classifier Setup

The tuned classifier setup is as follows:

- *For transfer learning*, to learn the new layers faster than the transferred layers, set the parameter values of

the fully connected layer as learn rate factors (LRFs) to a large value, whereas initial learning rate (ILR) is to a small value. Therefore, weight learn rate factor, bias learn rate factor and initial learning rate values were set to 20, 20 and 0.0001, respectively. Then the target networks were retrained by augmented training data.

- *For network training*, target networks were run for 90 epochs. The training was done using stochastic gradient descent (SGD) with momentum set to 0.90 with a batch size of 128. The learning rate was initially set to 0.0001 as a starting point and was decremented after each update. The programme validated target networks every three iterations; 3 iterations per epoch were selected as the maximum number. The different models were trained and tested with the same training and testing sets for a fair comparison. Figure 6 shows the training process of the different fine-tuned classifiers.

Analysis of the Results Obtained

The analysis of the results obtained is as follows:

- *Training progress analysis*: the analysis of the obtained results in Fig. 6 indicates that the fine-tuning models are

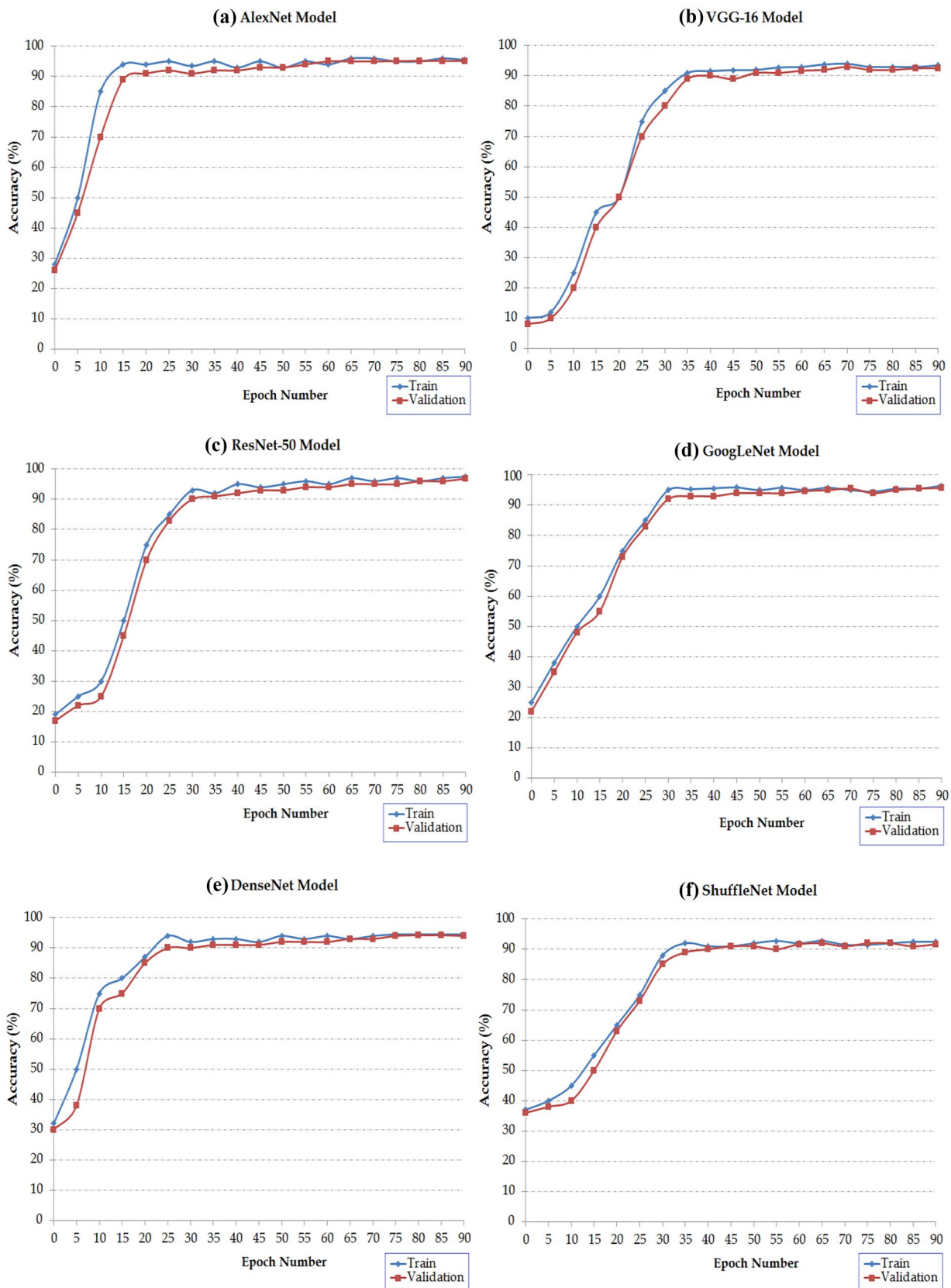**Fig. 5** Flowchart diagram for the proposed methodologies
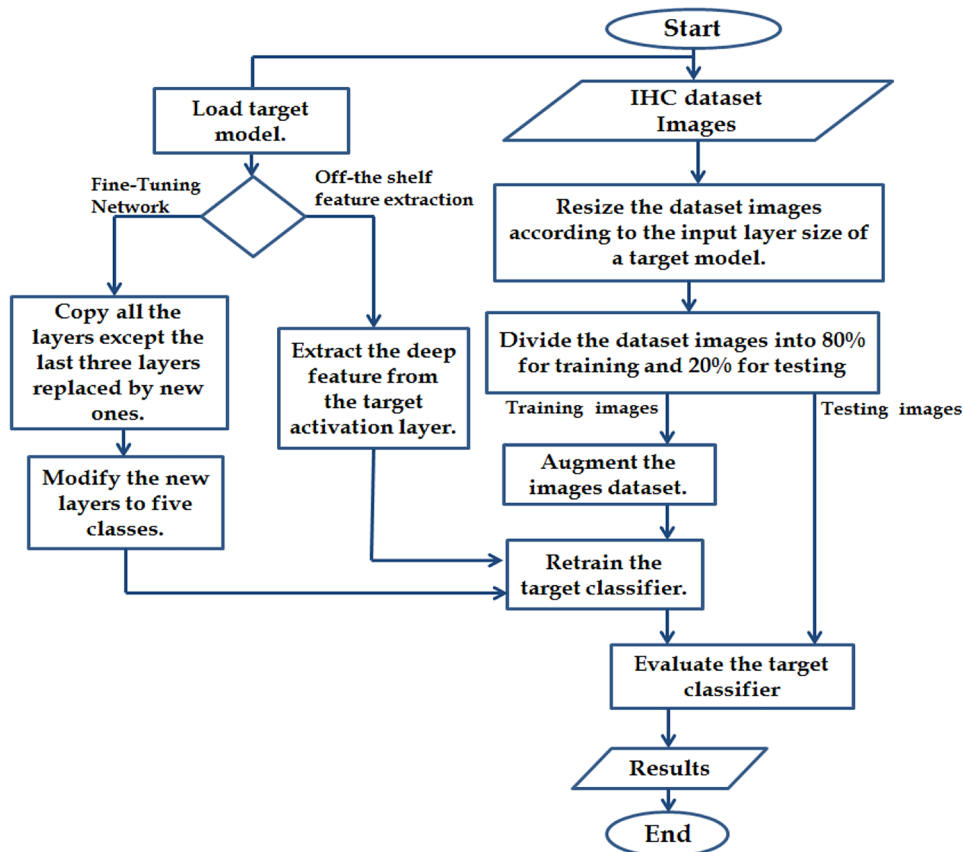
Fig. 6 Visualisation of the training progress for different target networks
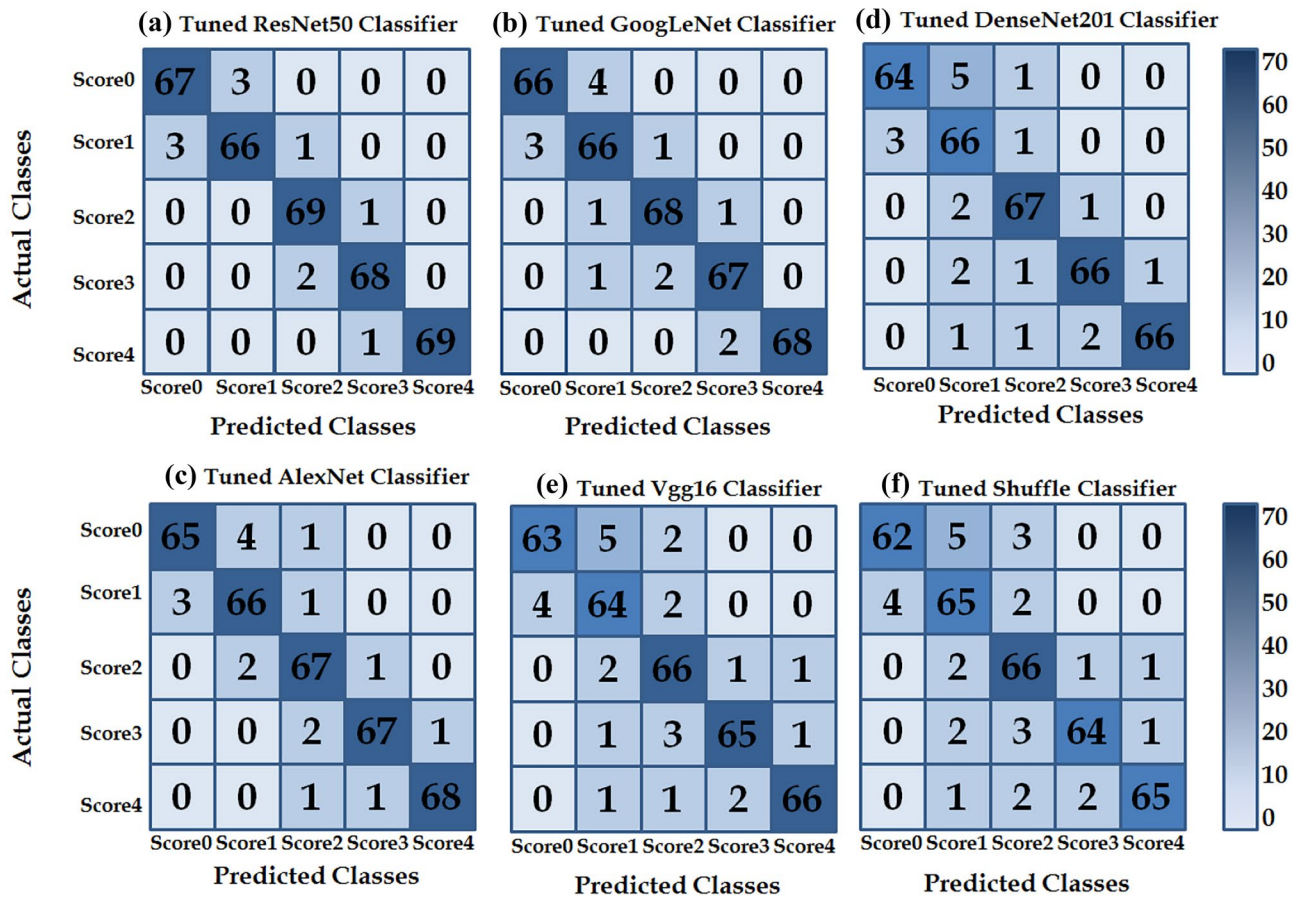
Fig. 7 Confusion matrices for various fine-tuned CNN evaluations

rapidly gaining convergence during the training stage. Where the convergence approximately occurred in the first twenty epochs. This is also shown by Tajbakhsh et al. [31] indicating that the fine-tuning CNN approach gains convergence faster. From the 60th epoch, the accuracy rate was approximately steady for all target networks; i.e. the values of accuracy rates rose slowly. This is the main reason for ending the training at the 90th epoch.

- *Timing training analysis*: from the analysis of timing training of the six target networks, we found that AlexNet is the fastest; it took 45 min to train. For the same training data, VGG16, GoogLeNet, ResNet50, DenseNet and ShuffleNet took 90, 270, 120, 300 and 200 min, respectively.

- *Performance analysis*: 5×5 confusion matrices were used to represent the obtained prediction results of the cancerous pathological samples. Five scores represented the percentage of GLUT1 staining in tumour cells. Here, the *X*-axis represents the actual values, and the *Y*-axis represents the predicted values. These 5×5 confusion matrices are shown in Fig. 7. Statistical performance measurement results of different fine-tuned classifiers were summarised in Tables 1

and 2. Accuracy, recall, precision and F1-score metrics were computed by using Eqs. 5, 6, 7 and 4, respectively. As can be seen from Table 1, the higher successful classification was achieved by the tuned ResNet50 and tuned GoogLeNet with accuracy rates that were 96.86% and 95.17%, respectively. Followed by tuned AlexNet, DenseNet, VGG16 and Shuffle models with accuracy rates were 95.14%, 94.00%, 92.57% and 91.17%, respectively. In the same way, from Table 2, we find that the tuned ResNet-50 model has obtained the highest F1-score values than the other models.

Table 1 Accuracy metric for different tuned classifiers

Fine-tuned model	Accuracy (%)
Tuned AlexNet	95.14
Tuned VGG16	92.57
Tuned GoogLeNet	95.71
Tuned ResNet-50	96.86
Tuned DenseNet-201	94.00
Tuned ShuffleNet	91.71

Table 2 Standard metrics for various fine-tuned classifiers

Proposed architecture	Metrics	IHC Score				
		Score0	Score1	Score2	Score3	Score4
AlexNet	P_i	0.93	0.94	0.96	0.96	0.97
	R_i	0.96	0.92	0.93	0.97	0.99
	$F_1 score$	0.94	0.93	0.94	0.96	0.98
VGG16	P_i	0.90	0.91	0.94	0.93	0.94
	R_i	0.94	0.88	0.89	0.96	0.97
	$F_1 score$	0.92	0.89	0.91	0.94	0.95
GoogleNet	P_i	0.94	0.94	0.97	0.96	0.97
	R_i	0.96	0.92	0.96	0.96	1.00
	$F_1 score$	0.95	0.93	0.96	0.96	0.98
ResNet50	P_i	0.96	0.94	0.99	0.97	0. 99
	R_i	0.96	0.96	0.96	0.97	1.00
	$F_1 score$	0.96	0.95	0.97	0.97	0.99
DenseNet	P_i	0.91	0.94	0.96	0.94	0.94
	R_i	0.96	0.87	0.94	0.96	0.99
	$F_1 score$	0.93	0.90	0.95	0.95	0.96
ShuffleNet	P_i	0.89	0.93	0.94	0.91	0.93
	R_i	0.94	0.87	0.87	0.96	0.97
	$F_1 score$	0.91	0.90	0.90	0.93	0.95

Evaluation of the SVM Classifiers

In this methodology, the experiment was carried out by using the target SVM classifiers that have been trained on the extracted deep feature vectors from the activation layer of the six proposed extractors. The extracted deep feature vectors are illustrated in Table 3. The flow diagram of the steps involved in the experiment is illustrated in Fig. 5. As in the previous section, we discuss the setup of the SVM classifiers and then the analysis of obtained results.

SVM Classifier Setup

The target SVM classifiers were set up by combining them with an error-correcting output codes (ECOC) function as described in [41]. ECOC function is commonly used for modelling a multi-class classification problem. It divides a K class problem into $\frac{K(K-1)}{2}$ binary learners, then assigns a one-versus-one coding design to determine the classes for

such binary learning [41]. Setting SVM classifiers with the ECOC framework could improve classification accuracy, even compared to other multi-class models.

Analysis of the SVM Classifiers' Performance

By the analysis of timing training, we recorded that classifier training time on *Feat-Alex*, *Feat-Vgg16*, *Feat-GoogLe*, *Feat-Res50*, *Feat-Dense* and *Feat-Suffle* took 29, 35, 49, 44, 58 and 32 min, respectively.

The obtained results of the experiment were represented in the 5×5 confusion matrices as shown in Fig. 8. Statistical performance measurement results of SVM classifiers were reported in Tables 4 and 5. The performance results were computed by using Eqs. 5, 6, 7 and 4. It can be noted that the trained SVM classifier on the *Feat-Dense* vector has obtained the highest value than the other vectors in terms of classification accuracy (97.43%), recall, precision and F1-score values.

Table 3 Feature vectors for different activation layers

Feature Vector	Activation Layer	Feature
Feat-Alex	"fc7"	$1 \times 1 \times 4096$
Feat-Vgg16	"fc7"	$1 \times 1 \times 4096$
Feat-GoogLe	"pool5-7×7-s$_1$"	$1 \times 1 \times 1024$
Feat-Res50	"avg-pool"	$1 \times 1 \times 2048$
Feat-Dense	"avg-pool"	$1 \times 1 \times 1920$
Feat-Suffle	"node-200"	$1 \times 1 \times 544$

Evaluation of the Proposed Feature Selection Approach

In the previous experiment, Feat-Dense201, Feat-Res50 and Feat-GoogLe vectors won the three best accuracies, which is why they were chosen in this experiment. Several studies [42–44] have employed the feature selection approach to improve classification accuracy and reduce over-fitting and training time. The goal of this experiment
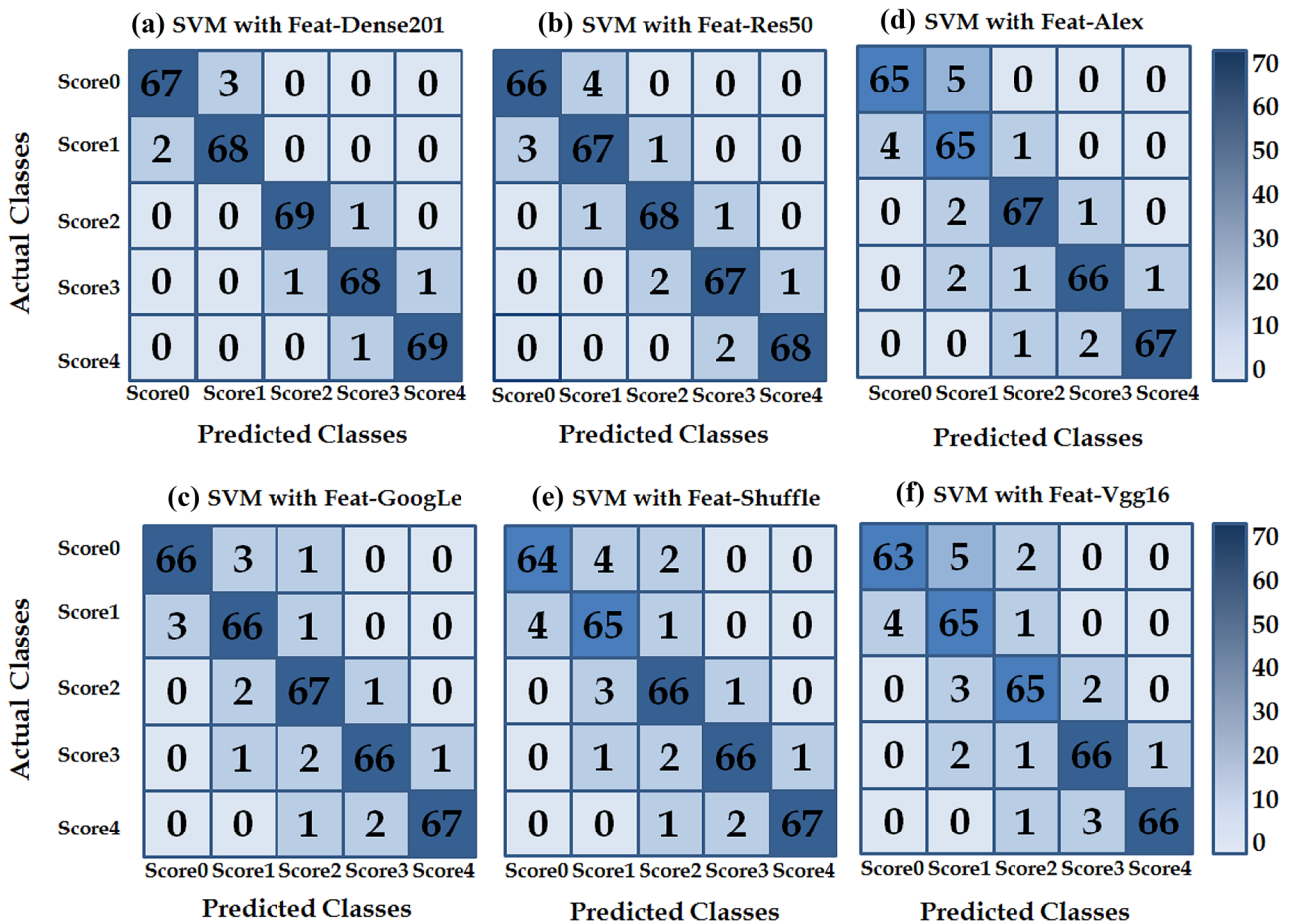
Fig. 8 Confusion matrices for SVM classifiers trained on different feature vectors

was to determine the features that yield better performance than the others. We concatenate these superior deep features and select the best features for representing our application to improve the performance. We employed a sparse support vector machine (SSVM) classifier to select the best deep features into a Feat-Concat vector. Feat-Concat vector has been utilised for training other SVM with ECOC classifier. The performance metric of the SVM classifier that trained on the Feat-Concat vector is given in Table 6, and Fig. 9 illustrates a 5×5 confusion matrix. It can be seen from Table 6 that the SVM classifier achieved the highest accuracy rate of 98.86%. It performed the classification with better accuracy than the other classifiers. The SSVM [41] solves the optimisation problem by minimising the following equation:

$$SSVM = \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(b + zh(x_i)\right)\right] + \lambda|z|, \qquad (8)$$

where n is the number of input images, $x_i = \left(x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{n,d}\right)$ is a vector of the ith feature, d is

the number of feature, and y_i is the class label and belonged to $\{+1, -1\}$, for $i = 1, \ldots, n$, where $y_i = +1$ indicates the ith sample is in class 1 (e.g. has cancer) and $y_i = -1$ indicates the ith sample is in class 2 (e.g. does not have cancer). z is a hyper-plane parameter, $\left[1 - y_i\left(b + zh(x_i)\right)\right]$ is the convex hinge loss function, the scalar b is denoted as the bias, $\lambda|z|$ is the L_1-norm, and $\lambda > 0$ is the tuning parameter controlling the trade-off between minimising the hyper-plane coefficients and the classification error.

Table 4 Accuracy metrics for SVM Classifiers

Trained SVM classifier with	Accuracy (%)
Feat-Alex	93.70
Feat-Vgg16	92.86
Feat-GoogLe	94.85
Feat-Res50	96.00
Feat-Dense	97.43
Feat-Shuffle	94.28

Table 5 Standard metrics for SVM classifiers

Feature vector	Metrics	IHC Score				
		Score0	Score1	Score2	Score3	Score4
Feat-Alex	P_i	0.93	0.93	0.96	0.94	0.96
	R_i	0.94	0.88	0.96	0.96	0.99
	F_1score	0.93	0.90	0.96	0.95	0.97
Feat-VGG16	P_i	0.90	0.93	0.93	0.94	0.94
	R_i	0.94	0.87	0.93	0.93	0.99
	F_1score	0.92	0.90	0.93	0.93	0.96
Feat-GoogLe	P_i	0.94	0.94	0.96	0.94	0.96
	R_i	0.96	0.92	0.93	0.96	0.99
	F_1score	0.95	0.93	0.94	0.95	0.97
Feat-Res50	P_i	0.94	0.96	0.97	0.96	0.97
	R_i	0.96	0.93	0.96	0.96	0.99
	F_1score	0.95	0.94	0.96	0.96	0.98
Feat-Dense	P_i	0.96	0.97	0.99	0.97	0.99
	R_i	0.97	0.96	0.99	0.97	0.99
	F_1score	0.96	0.96	0.99	0.97	0.99
Feat-Shuffle	P_i	0.91	0.93	0.94	0.94	0.96
	R_i	0.94	0.89	0.92	0.96	0.99
	F_1score	0.92	0.91	0.93	0.95	0.97

Table 6 Performance metric for SVM classifier with *Feat-Concat* vector

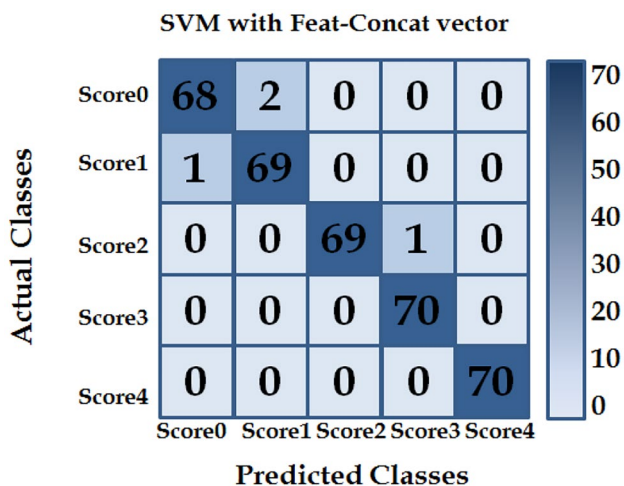| Metrics	IHC score				
	Score-0	Score-1	Score-2	Score-3	Score-4
Pi	0.97	0.99	0.99	1.00	1.00
Ri	0.99	0.97	1.00	0.99	1.00
F1score	0.98	0.98	0.99	0.99	1.00

Fig. 9 Confusion matrix for SVM Classifier with *Feat-Concat* vector

Discussion and Conclusion

In the discussion part, we will provide an analytical comparison among the methodologies used in the three experiments as follows:

- Fine-tuning ResNet-50 architecture is more suitable for our application than the other proposed architectures.
- The transferable deep features of DenseNet architecture are learned more than the deep features of other proposed architectures. Therefore the performance of the SVM classifier that is trained on the Feat-Dense vector has outperformed the other vectors on our dataset images.
- By comparing the time taken to train classifiers, we find that extracting and training SVM classifiers are much less than fine-tuning the CNN model as a classifier. Therefore, off-the-shelf feature extraction methodology is better than fine-tuning CNN methodology in terms of performance, training time and resources required for training.
- Although several studies [14, 16, 28, 35] have compared deep transfer learning methodologies in the histopathological field, there is no consensus about whether one is better than the other. However, this work showed that using off-the-shelf features methodology yielded performances slightly superior to using fine-tuning methodology but with the advantage of not having to re-train the network.
- The feature selection approach is proposed in this work to improve the performance of the SVM classifier for scoring tumour hypoxia markers as it yields optimal per-

formance. Therefore, our contribution is to introduce the *Feat-Concat* vector to train the SVM predictor.

- Our study does have some limitations. Firstly, the proposed framework was implemented during COVID-19, so the resources were restricted. Secondly, the used dataset images were prepared in the same laboratory and under the same conditions (i.e. dataset images acquired with the same digital scanners and staining techniques). Finally, choosing the appropriate patches was one of the difficulties that we faced in carrying out the experiments.

In conclusion, this paper proposed a deep learning-based framework to automatically assess the GLUT1 scoring as a biomarker of tumour hypoxia in IHC images. It helps to avoid inter-observer disagreement between pathologists and improves diagnostic performance. Three experiments were carried out using two common deep transfer learning methodologies with six various CNN architectures for classifying tumour hypoxia markers scores. From the results obtained in our experiments, it is observed that deep transfer learning approaches can significantly improve classification accuracy. Therefore, it is the best strategy in case of scarcity of dataset images, as is the case with histopathological images. This may be considered an initial step towards developing a reliable computer-assisted diagnosis tool for GLUT-1 scoring of digitised colorectal cancer histology slides. The future indications include the extension of our dataset and the inclusion of IHC images of different cancers from various laboratories. Also, state-of-the-art pre-trained models need to be included in future work. Finally, it will be targeted to apply deep transfer learning to develop a comprehensive computer-assisted diagnosis tool for GLUT-1 scoring of diverse cancerous tissue slides.

Data Availability The image dataset can be made available.

Code Availability Not applicable.

References

1. Brown, J. M., Wilson, W.R. Exploiting tumour hypoxia in cancer treatment, Nature Reviews Cancer, vol.4(6), pp.437,(2004).

2. Mirabello, V., Cortezon-Tamarit, F. and Pascu, S.I., 2018. Oxygen sensing, hypoxia tracing and in vivo imaging with functional metalloprobes for the early detection of non-communicable diseases, Frontiers in chemistry, vol.6, p.27, (2018).

3. Lepp{\"a}nen, J., Helminen, O., Huhta, H., Kauppila, J.H., Isohookana, J., Haapasaari, K.M., Karihtala, P., Parkkila, S., Saarnio, J., Lehenkari, P.P. and Karttunen,T.J., 2018. Toll-like receptors 2, 4 and 9 and hypoxia markers HIF-1alpha and CAIX in pancreatic intraepithelial neoplasia, Wiley Online Library, Apmis, vol.126(11), pp.852–863, (2018).

4. Bhandari, V., Hoey, C., Liu, L.Y., Lalonde, E., Ray, J., Livingstone, J., Lesurf, R., Shiah, Y.J., Vujcic, T., Huang, X. and Espiritu, S.M. Molecular landmarks of tumor hypoxia across cancer types, Nature Publishing Group, Nature genetics, vol.51(2), pp.308-318, (2019).

5. Godet, I., Shin, Y.J., Ju, J.A., Ye, I.C., Wang, G. and Gilkes, D.M. Fate-mapping post-hypoxic tumor cells reveals a ROS-resistant phenotype that promotes metastasis, Nature communications, Nature Publishing Group, Nature communications, vol.10(1), pp.1-18, (2019).

6. Zhao, S., Yu, W., Ukon, N., Tan, C., Nishijima, K.I., Shimizu, Y., Higashikawa, K., Shiga, T., Yamashita, H., Tamaki, N. and Kuge, Y., 2019. Elimination of tumor hypoxia by eribulin demonstrated by 18 F-FMISO hypoxia imaging in human tumor xenograft models, Springer, EJNMMI research, vol.9(1), pp.1–10, (2019).

7. Meier, V., Guscetti, F., Roos, M., Ohlerth, S., Pruschy, M. and Rohrer Bley, C. Hypoxia-related marker GLUT-1, CAIX, proliferative index and microvessel density in canine oral malignant neoplasia, Public Library of Science San Francisco, CA USA, PloS one, 11(2), p.e0149993, (2016).

8. Huizing, F.J., Hoeben, B.A., Franssen, G.M., Boerman, O.C., Heskamp, S. and Bussink, J. Quantitative imaging of the hypoxia-related marker CAIX in head and neck squamous cell carcinoma xenograft models. ACS Publications, Molecular pharmaceutics, 16(2), pp.701-708, (2018).

9. Raleigh, J.A., Chou, S-C., Bono, E.L., Thrall, D.E., Varia, M.A. Semiquantitative immunohistochemical analysis for hypoxia in human tumors, Elsevier International Journal of Radiation Oncology* Biology* Physics, vol.49(2), pp. 569–574, (2001).

10. Manu, V., Hein, T.A., Boruah, D., Srinivas, V. Serous ovarian tumors: Immunohistochemical profiling as an aid to grading and understanding tumorigenesis, Medical Journal Armed Forces India,(2018).

11. Albertella, M.R., Loadman, P.M., Jones, P.H., Phillips, R.M., Rampling, R., et al. Hypoxia-selective targeting by the bioreductive prodrug AQ4N in patients with solid tumors: results of a phase I study, Clinical cancer research,vol.14(4), pp.1096-1104, (2008).

12. Sullivan, C.AW., Chung, G.G. Biomarker validation: in situ analysis of protein expression using semiquantitative immunohistochemistry-based techniques, Clinical colorectal cancer, vol.7(3), pp.172–177, (2008).

13. Krizhevsky, A., Sutskever, I., Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks, In Proc. 25th International Conference on Neural Information Processing Systems, NIPS'12 Current Associates Inc., USA, pp.1097–1105,(2012).

14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., et al. A survey on deep learning in medical image analysis, Elsevier journal of medical image analysis, vol.42, pp. 60-88, (2017).

15. Stathonikos, N., Veta, M., Huisman, A., van Diest, P.J. Going fully digital: Perspective of a Dutch academic pathology lab, J. of pathol. inform., vol.4(1), pp.15, (2013)

16. Bayramoglu, N., Heikkil{\"a}, J. Transfer learning for cell nuclei classification in histopathology images, In: European Conference on Computer Vision, Springer, pp.532–539, (2016).

17. Qaiser, T., Mukherjee, A., Reddy Pb, C., Munugoti, S.D., Tallam, V., Pitk{\"a}aho, T., Lehtim{\"a}ki, T., et al. Her 2 challenge contest: a

detailed assessment of automated her 2 scoring algorithms in whole slide images of breast cancer tissues, Histopathology, vol.72(2), pp.227–238, (2018).

18. Cordeiro C.Q., Ioshii S.O., Alves J.H., Oliveira L.F. et al: An Automatic Patch-based Approach for HER-2 Scoring in Immunohistochemical Breast Cancer Images Using Color Features. arXiv preprint, (2018)

19. Mukundan R: Analysis of Image Feature Characteristics for Automated Scoring of HER2 in Histology Slides. Journal of Imaging pp. 5–35, (2019).

20. Tewary S., Arun I., Ahmed R., Chatterjee S., Mukhopadhyay S., et al: AutoIHC-Analyzer: computer-assisted microscopy for automated membrane extraction/scoring in HER2 molecular markers. Journal of Microscopy 281:pp. 87-96, (2021).

21. Chang, C.-Y., Huang Y.-C., Ko C.-C. Automatic analysis of her-2/neu immunohistochemistry in breast cancer, in: 2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications, IEEE, pp. 297–300, (2012).

22. Pitkäaho, T., Lehtimäki, T.M., McDonald, J. and Naughton, T.J.: Classifying HER2 breast cancer cell samples using deep learning. In Proc. Irish Mach. Vis. Image Process. Conf, pp. 1–104, (2016).

23. Saha, M. and Chakraborty, C.: Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Transactions on Image Processing, 27(5), pp.2189-2200,(2018).

24. Khameneh, F.D., Razavi, S. and Kamasak, M., Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Computers in biology and medicine, 110, pp.164-174, (2019).

25. Tewary, S. and Mukhopadhyay, S., HER2 molecular marker scoring using transfer learning and decision level fusion. Journal of Digital Imaging, 34, pp.667-677,(2021).

26. Drew, C.P., Shieh, W.-J. Immunohistochemistry, In: Current Laboratory Techniques in Rabies Diagnosis, Research and Prevention, Elsevier, vol.2, pp.109--115, (2015).

27. Ruifrok, A.C., Johnston, D.A., et al. Quantification of histochemical staining by color deconvolution, Analytical and quantitative cytology and histology, vol.23(4), pp. 291--299, (2001)

28. Mormont, R., Geurts, P., Mar{\'e}e, R. Comparison of deep transfer learning strategies for digital pathology, In Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.2262–2271, (2018).

29. Miko{\l}ajczyk, A., Grochowski, M. Data augmentation for improving deep learning in image classification problem, In 2018 international interdisciplinary PhD workshop (IIPhDW), pp. 117–122, (2018).

30. Shorten, C., Khoshgoftaar, T.M. A survey on image data augmentation for deep learning, Springer, J. Big Data, vol.6, pp. 60 (2019).

31. Sokolova, M., Lapalme, G. A systematic analysis of performance measures for classification tasks, Elsevier Journal of Information Processing \& Management, vol.45(4), pp. 466–475, (2009).

32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. et al. Going deeper with convolutions, In: Proc. IEEE conference on computer vision and pattern recognition, pp. 1–9, (2015).

33. Yosinski, J., Clune, J., Bengio, Y., Lipson, H., How transferable are features in deep neural networks?, In: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol.2, pp.3320–3328, (2014).

34. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B. et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?, journal of IEEE transactions on medical imaging, vol.35(5), pp. 1299–1312, (2016).

35. Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P. Understanding the mechanisms of deep transfer learning for medical images, In: Deep Learning and Data Labeling for Medical Applications, Springer, pp. 188–196, (2016).

36. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L. Imagenet: A large-scale hierarchical image database, In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248--255,(2009).

37. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition International Conference on Learning Representations (ICLR), 2015.

38. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, J. Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition, PP.770–778, (2016).

39. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Densely connected convolutional networks, In Proc.of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708 (2017).

40. Zhang, X., Zhou, X., Lin, M., Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices, In Proce. of the IEEE conference on computer vision and pattern recognition, pp. 6848--6856 (2018).

41. Escalera, S., Pujol, O., Radeva, P.On the Decoding Process in Ternary Error-Correcting Output Codes, IEEE transactions on pattern analysis and machine intelligence, vol.32(1), pp. 120-134, (2010).

42. Jammal, M., Canu, S.,Abdallah, M. R., Sparse Support Vector Machines via Mixed Integer Programming, In International Conference on Machine Learning, Optimization, and Data Science, Springer, pp. 572—585,(2020).

43. Yao, L., Zeng, F., Li, D.-H., Chen, Z.-G. Sparse Support Vector Machine with L p Penalty for Feature Selection, Journal of Computer Science and Technology, Springer, vol(1)32,pp. 68—77, (2017).

44. Kahya, M. A., Al-Hayani, W., Algamal, Z. Y. Classification of breast cancer histopathology images based on adaptive sparse support vector machine, Journal of Applied Mathematics and Bioinformatics, vol.(1)7, pp.49,(2017).