An ensemble of deep CNNs for automatic grading of breast cancer in digital pathology images

An Ensemble of Deep CNNs for Automatic Grading of Breast Cancer in Digital Pathology Images

Shallu Sharma[1], Sumit Kumar[2]*, Manoj Sharma[3], Ashish Kalal[4]

[1]*School of Computer Science Engineering and Technology, Bennett University, Greater Noida, UP - 203206, India*
[2]*Division of Research and Development, Lovely Professional University, Phagwara, Punjab - 144411, India*
[3]*Department of ECE, Giani Zail Singh Campus College of Engineering & Technology, MRSPTU, Bathinda, Punjab - 151001, India*
[4] *Department of Mechanical Engineering, University College London, London, WC1E 7JE, U.K.*

*Corresponding Author, Email: sumit.24786@lpu.co.in, kumarsumit8@gmail.com

Abstract:
Histopathological diagnosis is the mainstay of present-day preventive medical care service to guide the therapy and treatment of breast cancer at an early stage. Manual examination of histological data based on clinicians' subjective knowledge is a time-consuming, labour-intensive, and costly method that necessitates clinical intervention and competence for a fair decision. In the recent work, we have developed an ensemble of five deep CNNs to classify three grades of breast cancer using quantitative image-based assessment of digital pathology slides without any manual intervention. To produce final predictions on the dataset, a fuzzy ranking algorithm is used. On the Databiox dataset, the suggested model attained an accuracy of 79%, 75%, 89% and 82% at 4X, 10X, 20X and 40X magnification, respectively. Furthermore, it has been observed that the stain-normalization strategy improves the model's classification performance on the histopathological images. In this case, the Mackeno stain-normalization technique is employed which further enhances the performance of the proposed ensemble model up to 80%, 100%, 100%, and 82% at 4X, 10X, 20X and 40X magnification, respectively. Additionally, a comparative analysis with the existing state-of-the-art technique demonstrated the superiority of the proposed scheme.

Keywords: Breast Cancer, Ensemble model, Pathology, CNN**,** Stain-normalization.

1. Introduction

Pathology plays a significant role in clinical research as it establishes a bridge between science and medicine by addressing four open questions of medical science: What is the cause of disease, how the disease originates within the body, how it affects the cells and what is the nature of the disease? A comprehensive and accurate pathology report is critical for determining a precise diagnosis and the best treatment plan. Prognostication besides the elucidation of aetiology, pathogenesis, and clinicopathological correlation are integral functions of pathological examination which make it the gold standard for diagnosing a variety of epidemics including cancer [1]. Cancer is a group of diseases involving abnormal growth of cells anywhere in a body [2-4]. A pathologist recognizes the nature, stage, and grade of cancer through visual analysis of tissue stained by haematoxylin and eosin (H&E) [5, 6]. An apparent understanding related to the stage and grade of cancer always guides the clinical team in determining the right treatment.

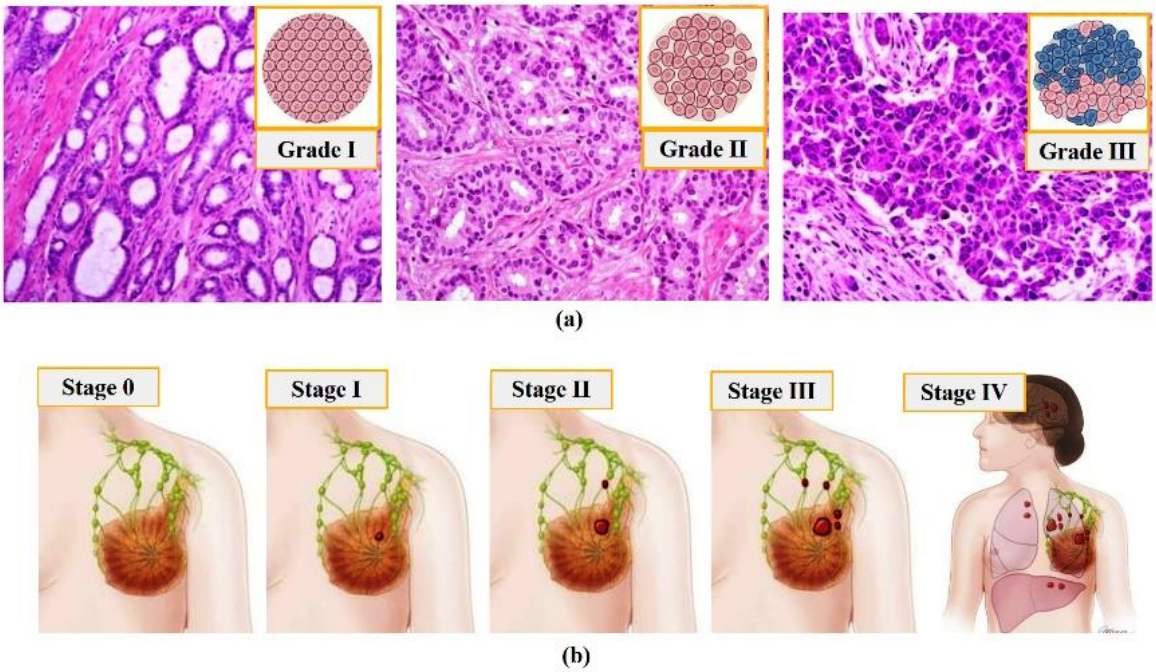


Fig 1. Pictorial representation of (a) Grade of breast cancer ranges from Grade I to Grade III, depending upon the glandular/tubular differentiation, nuclear pleomorphism and mitotic count. Grade I (> 75% of tumor form glands, uniform cells, < 7 mitoses per 10 high power fields), Grade II (10% to 75% of tumor form glands, cells larger than normal and moderate variability in shape and size, 8-15 mitoses per 10 high power fields), Grade III (<10 75% of tumor form glands, prominent nucleoli and high variability in shape and size of cells, >16 mitoses per 10 high power fields) (b) Stages of breast cancer ranges from Stage 0 to Stage IV. Reproduced from https://pathology.jhu.edu/breast/staging-grade/

Female breast cancer is the most common and life-threatening cancer worldwide with 685,000 mortality and 2.3 million incidence cases in 2020 at the global level [7]. Invasive ductal carcinoma (IDC) is the prominent subtype of all breast cancers that originate in the milk duct and invades fibrous tissue of the breast outside the duct. The identification of accurate stage and grade is very essential for planning and assigning treatment to a patient that meets the patient's requirement. Most of the time people get confused with both the terms (staging and grading) and used them interchangeably which may misguide the person in selecting the right treatment. The grade of breast cancer represents the "aggressive potential" of the tumor which ranges from Grade I to Grade III. While the pathological stage of breast cancer measures the advancement of the patient's tumor which ranges from Stage 0 (pre-invasive disease) to Stage IV (metastatic disease), see Fig 1. Generally, different types of cancer follow distinct grading and staging system but breast cancer has its own staging and grading system [8]. The Nottingham is the most commonly used histologic grading system for breast cancer, quantified on the basis of three important factors namely, gland formation, pleomorphism and mitotic count [9]. Detailed information on the staging and grading is given in https://pathology.jhu.edu/breast/staging-grade/.

Manual detection and annotation of affected areas on histopathology images have long been considered the gold standard for cancer diagnosis and prognosis [10, 11]. However, this

approach is both challenging and time-consuming, demanding meticulous care and extensive experience from pathologists. Compounding the issue is the substantial inter and intra-observer variability in visual analysis, a major concern that introduces reproducibility challenges [12, 13]. These challenges have a direct impact on the accuracy of cancer prognostication and treatment planning.

Automated grading of breast cancer, in contrast, emerges as a potent solution to address the multifaceted needs and challenges within the realm of oncology and medical diagnostics. By leveraging automation, we can achieve consistent and objective assessments of cancer samples. Unlike human pathologists, whose interpretations can exhibit variability, automation ensures that grading adheres to predefined and standardized criteria. This is particularly valuable, given that pathologists often need to assess numerous tissue samples. Automation significantly expedites the grading process, resulting in faster diagnoses and treatment decisions, a critical advantage in cases requiring timely intervention. One of the most notable benefits of automation is the high degree of reproducibility it offers. This means that grading remains consistent over time and between different healthcare institutions. Automation greatly reduces the risk of human error in grading, effectively eliminating potential mistakes stemming from fatigue, distraction, or variations in individual pathologists' skills and experiences. As the volume of medical data continues to surge, automation is capable of handling the increasing demand for cancer grading without necessitating a corresponding increase in the number of pathologists. Furthermore, automated grading systems can seamlessly integrate with electronic health records (EHRs) and other clinical databases. This integration facilitates efficient data management, retrieval, and analysis, promoting a comprehensive approach to patient care and research. Standardizing the grading process across different healthcare facilities and geographic regions is another pivotal role of automated grading. This ensures that patients receive consistent care and recommendations, regardless of their location of diagnosis or treatment.

The recent advancements in the field of machine learning and computer vision techniques open up a new opportunity to develop intelligent computer-controlled machines and software. Machine learning is a method of data analysis that automates analytical model building by identifying patterns and making decisions with minimal human interference [14-20]. Hence, advanced image processing and data analysis technology could be employed in spotting and grading the cancerous area on histopathology images. This will further accelerate the process of treatment planning with higher accuracy by assisting the pathologists in their decision. However, the large amounts of artefacts and variability in fixation (fixation time and temperature), paraffin embedding, staining protocols, staining and sectioning quality during tissue preparation along with the change in shape, size, location, and texture of nuclei transform automated detection and grading into a challenge for the entire science community.

Accurately classifying invasive ductal carcinoma histopathology data poses significant challenges, primarily due to its intricate and multifaceted nature. In this study, we have endeavoured to address these challenges by introducing a novel approach that combines fuzzy ranking techniques with advanced machine learning methodologies. Our aim is to provide a robust and interpretable means of amalgamating predictions from various models, while considering uncertainty and potential disparities among them. Fuzzy ranking, in this context, assigns degrees of membership to different ranks, which can be interpreted as confidence

levels. This approach allows us to quantify the strength of a sample's association with a particular class or rank, thereby aiding in decision-making. Moreover, fuzzy ranking exhibits a greater resilience to outliers, anomalies, and extreme values in the data, enabling a seamless transition between ranks while acknowledging the proximity of such data points.

In this paper, we present a highly accurate and computationally efficient automated cancer grade detection framework. We leverage the power of five transfer learning-based convolutional neural networks (CNNs) to create an ensemble model. This ensemble model considers predictions from all the constituent models to arrive at a final decision. The primary motivation for employing transfer learning-based CNNs is the scarcity of available data, which makes it challenging to achieve satisfactory performance when training deep CNN models from scratch. What sets our approach apart is the incorporation of a mathematical model that incorporates the predictions of individual classifiers when computing the overall prediction of the ensemble model. This distinguishes our method from conventional fusion techniques such as averaging, majority voting, weighted averaging, and weighted majority voting. By effectively navigating uncertainty, accommodating variability, and offering nuanced data interpretations, fuzzy ranking emerges as a valuable tool to enhance the accuracy and reliability of cancer grading in histopathological datasets. For a visual representation of our overall framework, please refer to Fig. 2.

The key contributions of our research can be summarized as follows:

1. An ensemble model using five CNNs (VGG19, NasNet, Inception_V3, MobileNet, and Xception) has been designed to make predictions for grading on available Databiox dataset.
2. Two non-linear functions with different indentations have been applied to the proposed ensemble model for determining the fuzzy ranks of each class in the decision scores. Further, the sum of products of ranks are calculated for the five CNNs and the lowest rank is assigned as the predicted class.
3. The impact of stain normalization methods namely Macenko (a technique of data pre-processing) on the performance of proposed classification has also been demonstrated.
4. The proposed approach surpasses the existing state-of-the-art technique [21] on the Databiox dataset in terms of accuracy.

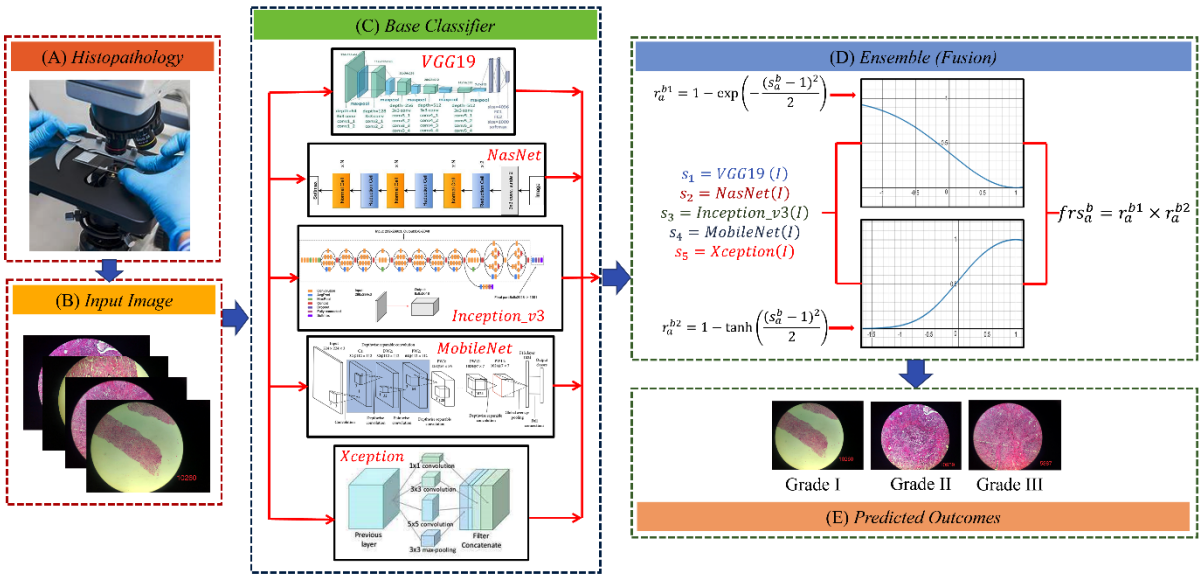


Fig 2. A schematic representation of framework employed in the classification of three types of breast cancer grade

2. Material & Methods

Identification of the real stage and grade of cancer is very crucial for the early intervention of the clinicians in order to reduce the progression of cancer by providing the best treatment. However, well-characterized data is a pre-requisite for the advanced development in Computer-Aided Diagnosis (CAD) systems and has become the main objective for researchers working in the area of cancer detection and classification. Numerous research centres and hospitals did not release their clinical data on a public platform for further research. As a matter of the fact, the researchers and new technocrats remain far from the reach of innovations and novel ideas. In this direction, science communities around the world are working extremely hard to create histopathological digital images databases that can be accessed by multiple users at the same time [22-31]. Several breast cancer pathology databases have already been published, tabularized in Table 1. All the outlined databases are built using a similar technique where the histopathological digital images are acquired from whole slide imaging (WSI) of breast cancer tissue sample stained with H&E at different magnification factors. Recently, one more database "Databiox" has also been reported as a well-annotated and standardized dataset that consists of histopathological images of IDC diagnosed patients for grade classification [32].

Table 1: Existing Databases for Breast Cancer Histopathology Images

Database	Cancer Type	Magnification Level (ML)	Number of specified ML	Number of Images	Image Format	Image Size	File Type	Ref
BACH	Normal, Benign, In-situ and Invasive	----------	-------	500	RGB	2048 x 1536 pixels	TIFF	[22]

Camelyon	Normal and Metastases	1X, 10X, and 40X	3	400 WSI	RGB	218000x 95000		[23, 24, 27]
Cytological Images	Malignant and benign	------	------	92	RGB	varied from 640 x 480 to 2560 x 1920 pixels	------	[26]
Breast Biopsy Specimens	Invasive breast cancer, ductal carcinoma in situ (DCIS), with atypical hyperplasia (atypia), and benign cases without atypia	------	------	240	RGB	-------	------	[25]
Cytological Images	Malignant and benign	------	------	500	RGB	704 x 578	BMP	[28]
Tissue Microarray (TMA)	Malignant and benign	20X for the tissue and 40X for the cells	2	205161	RGB	------	------	[29]
BreakHis	Malignant and Benign	40X, 100X, 200X, and 400X	4	7,909	RGB	700x460 pixels	JPEG	[30]
Nonlinear Microscopy (NLM)	Normal breast tissue, fibroadenoma, ductal hyperplasia, fibrocystic changes, lobular carcinoma in situ (LCIS), invasive lobular carcinoma (ILC), and IDC	20X	1	179	RGB	1024x1024 pixels	------	[31]
Databiox	Grade I, Grade II, Grade III for IDC	4X, 10X, 20X, and 40X	4	922	RGB	2100 x 1574 and 1276 x 956 pixels	JPEG	[32]

This database consists of 922 JPEG images in RGB format with a resolution of 2100 x 1574 and 1276 x 956 pixels for grade classification. The specimens of breast tissues in the dataset

have been collected from 124 patients who were diagnosed with breast cancer at Poursina Hakim Research Centre of Isfahan University of Medical Sciences in Iran. The image for each specimen of breast tissue is captured at four levels of magnification (4X, 10X, 20X, and 40X). However, more than one image is presented for some specimens based on pathologist's opinion. Patient details have been meticulously anonymized for public access. Each image sample is uniquely identified with a format such as "01_BC_IDC_9057_4x_1," where "01," "BC_IDC," "9057," "4x," and "1" represent "Sample number," "Cancer type," "Pathology archives number," "Magnification level," and "Number of specified magnification level," respectively. The modified Bloom Richardson histologic grading method is followed to label the specimens. According to this grading method, the amount of three attributes namely, tubule formation, nuclear pleomorphism and the mitotic count is evaluated and a score of 1, 2, or 3 is assigned to each attribute. These assigned scores are further added to produce the final grade. A detailed distribution of the database as per the number of patients involved and the levels of magnification for each grade is illustrated in Figure 3.

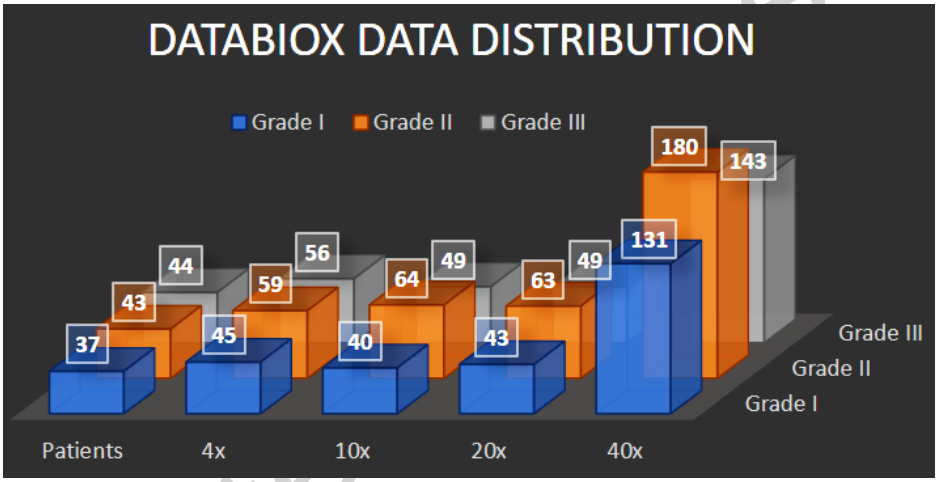


Fig 3. Distribution of samples in Databiox as per the levels of magnification for each grade

2.1 Pre-processing

This section elaborates the pre-processing methods which have been considered for addressing the challenges such as unbalanced data, noisy images due to variability in staining procedures, low resolution, and contrast non-uniformity, etc. These all factors significantly affect the performance of the classification model. In that context, the appropriate pre-processing techniques are required to get rid of the above-mentioned problems and the following procedures have been adopted.

2.1.1 Data Balancing and Data Augmentation

A data set with skewed class proportions is called imbalanced dataset where classes that make up a large proportion of the dataset are called majority classes. The classes with non-symmetrical distribution of data have a possibility to generate biased results by inclining the model towards the majority classes and leads to the production of false values in the model, eventually the problem of overfitting. Thus, data augmentation methods are applied to balance the number of instances in each class [33, 34].

In order to address the two main challenges (insufficient and imbalanced data) encountered while developing deep learning models, data augmentation methods are required to implement. The techniques of data augmentation which we have utilized are applicable to breast cancer histopathology and discussed as follows:

i. **Rotation:** It is the most commonly utilized technique of data augmentation in which the images are rotated with different angles. However, it can be applied only to the images in which the information remains the same even after rotating the images [35]. For example, the information does not alter for histopathological images if we rotate the image with any angle. While if we apply the same on the image where digit 6 is written, it will change to digit 9 with an angle of rotation 180^0. Here, all the images in the training datasets are rotated with an angle of 30^0, 40^0, 60^0 and 90^0, shown in Fig. 4.

ii. **Shifting:** This is a technique in which position of the object in the image is changed so that it provides more variety to the model for developing a robust and generalized model. In this context, we have implemented width-wise and height-wise shifting with a factor of 20%.

iii. **Shearing**: This technique of data augmentation shifts one part of the image like a parallelogram. Conventionally, the shearing of an image is performed by shearing an image from both corners with some amount in the x and y direction. This stretches the target objects with much extent and leads to the placement of bounding box on an incorrectly targeted place. In this context, we have sheared the images from a single corner.

iv. **Flipping:** Flipping is an extended version of rotation which allows the images to flip in left-right and up-down directions. Here, the images are flipped horizontally as well as vertically to enlarge the dataset.

v. **Image Resizing:** Resizing is a technique to tailor the dimensions of input image into specific size to make it suitable for the deep learning models [35]. We have applied the image resizing technique before feeding the images into the models.

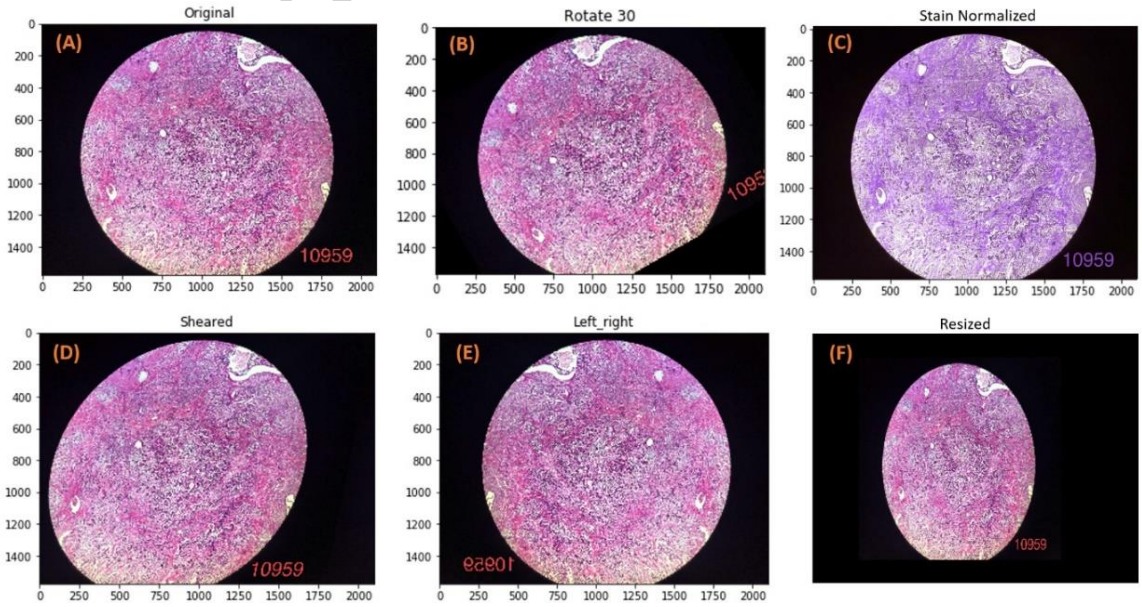


Fig 4. Visual representation of data augmentation methods applied on (A) Original image from Databiox dataset at 4X magnification from class Grade I with subject ID "03_BC_G1_10959_4x_1", (B) Rotation with angle 30^0, (C) Stain normalization using Macenko method, (D) Shearing with a value of 20%, (E) Flipping of image in left-right direction, and (F) Resizing of image in size 224 x 224.

2.1.2 Stain Normalization with Colour Transfer between Images

Colour is one of the most important attributes for histopathological images. Although the staining protocols used in laboratories are standardized, ariations in staining outcomes due to environmental conditions across slide scanners, antigen concentrations, temperature, storage, and incubation time are common. These variations can negatively impact the performance of Computer-Aided Diagnosis (CAD) systems. Therefore, it becomes necessary to standardize or normalize the intensity values (i.e., pixel values) to a range between 0 and 1 to reduce color intensity variations throughout the image [36, 37]. In this context, the Macenko stain normalization method is employed to assist CAD systems by generating images with a standardized appearance of different stains [36]. This method was introduced by Macenko et al., follows a specific procedure:

Color Deconvolution: Initially, the method employs a process known as color deconvolution, which separates the image into its individual color channels. In histopathological images, these color channels typically correspond to the stains used for tissue preparation, such as hematoxylin (which stains cell nuclei) and eosin (which stains cytoplasm).

Stain Density Estimation: Macenko stain normalization estimates the stain density of each color channel by analyzing the distribution of stain intensities in the image. This step helps to determine how the stains are distributed within the image and how they contribute to its overall color appearance.

Stain Vector Selection: A stain vector, which represents the dominant stain colors in the image, is selected based on the estimated stain densities. This vector captures the most prominent stain colors and their proportions in the image.

Stain Normalization: With the selected stain vector, Macenko normalization recombines the color channels to create a normalized version of the image. The aim is to ensure that the normalized image exhibits a consistent stain appearance across different images and conditions, allowing for more accurate and consistent analysis.

Contrast Enhancement (Optional): In some variations of the Macenko method, additional contrast enhancement techniques are applied to the normalized image to improve the visibility of specific structures within the tissue.

We followed the same procedure and the RGB image is first converted to optical density (OD). The data with OD intensity less than $\beta = 0.15$ is removed (the most optimal threshold value). Later, SVD is calculated on the OD tuples and plane is created from the SVD directions corresponding to the two largest singular values. Data is projected onto the plane, and normalized to a unit length. The angle of each point is calculated with respect to the first SVD direction. Robust extremes, typically the (α^{th} and $(100-\alpha)^{th}$ percentiles) of the angle are determined and extreme values are converted back to OD space. Eventually, the optimal stain vectors are obtained.

3. Methodology and Underlying Assumptions

This section provides an overview of the base models briefly, with the detail of customization applied to each model followed with experimental implementation of the proposed ensemble methods to fuse the scores of base models. In this study, five base models (pre-trained on ImageNet dataset) are utilized and evaluated on Databiox-dataset. In order to fuse the decision obtained from all the base models, confidence factors generated from each base model is considered for ensemble. The way we utilized the confidence factors generated from base models and ranking of the classes in the decision scores apart it from conventional methods of ranking. We have applied two non-linear functions i.e., $1-\exp\left(-\frac{(x-1)^2}{2}\right)$ and $1-\tanh\left(\frac{(x-1)^2}{2}\right)$ to map confidence factors non-linearly where one function signifies the proximity to 1 and other signifies the divergence from 1 for their full utilization in making decision. In order to generate non-linear fuzzy ranks, the confidence scores are mapped on two different functions and a combined score is generated by fusing these computed ranks which further helps in quantifying the total divergence from the expected one. Less divergence indicates more confidence in a specific class. Thus, a class with the lowest divergence is assigned as a final predicted outcome.

The present study embraces several foundational assumptions, underscoring the meticulous approach to its implementation which include: a) The invasive ductal carcinoma dataset "DataBiox" used is representative and accurately labelled dataset which is free from significant biases or errors. b) Images at different magnification levels are also representative of the variations present in clinical practice. c) The predefined classes for invasive ductal carcinoma grade classification (e.g., class 0, class 1 and class 2) accurately capture the underlying medical conditions and are consistent with established medical definitions. d) The augmented samples effectively simulate real-world variations in the data and improve the model's generalization.

3.1 Base Model's Customization and Cascading

Based on the structure of the model, customized layers have been added to utilize the information generated by pre-trained models efficiently. A fully connected layer of 1024, 2034, 256, 1024, 512 nodes have been added for Inception_V3 [38], NasNet [39], Xecption [40], VGG-19, MobileNet [41], respectively. In order to avoid the problem of vanishing gradient, the rectified linear unit (ReLU) activation function layer is also added. Moreover, the dropout layer with a value of 0.5 is utilized as a technique of regularization to surpass the overfitting problem. A variety of hyperparameters such as number of epochs, learning rate, optimizer, loss function utilized in training of CNNs have been fixed through massive experimentation and represented in Table 2.

Table 2. Hyperparameters with their associated value utilized to train the base models

S. No.	Hyperparameter	Value
1.	Loss Function	Categorical Cross Entropy
2.	No. of Epochs	1000

3.	Learning Rate	0.0001
4.	Optimizer	Adam Optimizer
5.	Batch Size	32
6.	Dropout Rate	50%

3.2 Mathematical Formulation of Ensemble Approach

In the present work, we are characterizing the Databiox dataset into three grades namely, grade I (class 0), grade II (class 1), and grade III (class 2). Thus, the total number of classes (C) for the present problem are three. Let the confidence or probability scores for each class computed by the base models are $(s_1^b, s_2^b, s_3^b, \dots\dots, s_C^b)$ where 'b' represents the number of base models, i.e., five. Wherein, the accumulation is performed across all the classes while considering the contributions from each base model 'b', using Eq. (1).

$$\sum_{a=1}^{C} s_a^b = 1, \quad \forall\, b = 1, 2, 3, 4, 5. \tag{1}$$

Further, the fuzzy ranks, represented as $(r_1^{b1}, r_2^{b1}, r_3^{b1}, \dots\dots, r_C^{b1})$ and $(r_1^{b2}, r_2^{b2}, r_3^{b2}, \dots\dots, r_C^{b2})$, are calculated for each base model using two non-linear functions as per the Eq. (2) and (3).

$$r_a^{b1} = 1 - \tanh\left(\frac{(s_a^b - 1)^2}{2}\right) \tag{2}$$

$$r_a^{b2} = 1 - \exp\left(-\frac{(s_a^b - 1)^2}{2}\right) \tag{3}$$

Here, the Eq. (2) computes a reward and Eq. (3) computes deviation for the classification. The value of Eq. (2) increases as s_a^b approaches to 1 and the level of confidence also increases. On the other hand, Eq. (3) calculate deviation from 1, deviation will be more if s_a^b approaches 0. The fused rank scores $(frs_1^b, frs_2^b, \dots frs_3^b)$, i.e., (frs_a^b) is obtained by multiplying the computed reward (r_a^{b1}) and deviation (r_a^{b2}) for a particular confidence score obtained from the base model, given by Eq. (4).

$$frs_a^b = r_a^{b1} \times r_a^{b2} \tag{4}$$

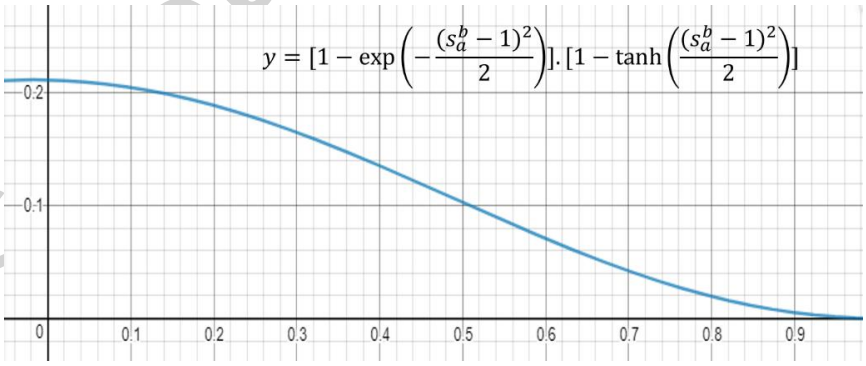


Fig 5. Product of two rank generating functions where 'y' denotes the fuzzy rank product

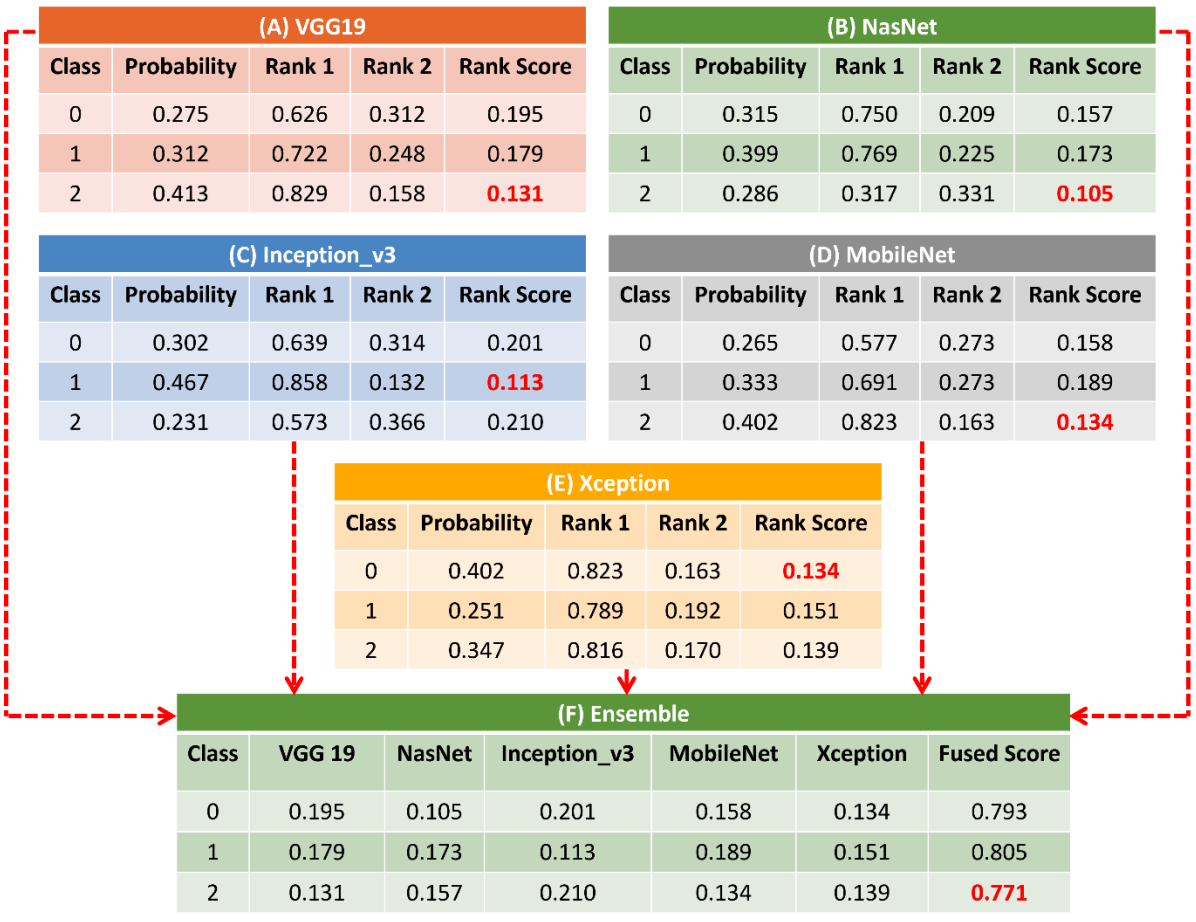
A graphical plot is illustrated in Fig 5 which represents the product of two rank generating functions. It is clear from the graph that the final rank decreases with an increase in probability score. Therefore, we need to find the class with the lowest fused score and will be declared as final outcome because of the highest probability score. In this context, the fused score tuple

$FS_a = FS_1, FS_2, FS_3, \ldots, FS_C$ is computed using Eq. (5) and the final outcome i.e., $class(O)$ is determined using Eq. (6).

$$FS_a = \sum_{b=1}^{5} frs_a^b, \forall\, a = 1, 2, \ldots, C-1, C \quad (5)$$

$$class(O) = \min_{\forall\, a} FS_a \quad (6)$$

To illustrate the proposed methodology for the Databiox dataset, let's delve into an example. In this scenario, we focus on the images categorized into three distinct classes. Probability values for each of these classes are meticulously calculated across five base models. This process is visually represented in Fig. 6. It can be observed from the figure that the value of probability to class 0, 1, and 2 given by VGG19 is 0.275, 0.312, and 0.413, respectively. As per the Eq. (2), corresponding ranks 0.626, 0.722, and 0.829 are obtained, denoted as rank 1. Further, rank 2 is computed as per the Eq. (3) i.e., 0.312, 0.248, and 0.158, respectively. Eventually, the rank score becomes 0.195, 0.179 and 0.131, using Eq. (4). Likewise, the rank scores are computed for each of the five base models for three classes. The rank scores obtained from VGG19, NasNet, Inception_v3, MobileNet, and Xception for class 0 are 0.195, 0.157, 0.201, 0.158, and 0.134, respectively. The fused score computed as per the Eq. (5) is 0.793 for class 0. Similarly, the fused score for class 1 and 2 are 0.805 and 0.771, respectively. Since the final outcome by VGG19, NasNet, and MobileNet is class 2, but as per Inception_v3 and Xception is class 1 and class 0, respectively. However, the proposed ensemble model makes a robust and accurate final outcome. According to ensemble model, the overall score is minimum for class 2 as per Eq. (6) and declared as the final predicted outcome.



(A) VGG19

Class	Probability	Rank 1	Rank 2	Rank Score
0	0.275	0.626	0.312	0.195
1	0.312	0.722	0.248	0.179
2	0.413	0.829	0.158	**0.131**

(B) NasNet

Class	Probability	Rank 1	Rank 2	Rank Score
0	0.315	0.750	0.209	0.157
1	0.399	0.769	0.225	0.173
2	0.286	0.317	0.331	**0.105**

(C) Inception_v3

Class	Probability	Rank 1	Rank 2	Rank Score
0	0.302	0.639	0.314	0.201
1	0.467	0.858	0.132	**0.113**
2	0.231	0.573	0.366	0.210

(D) MobileNet

Class	Probability	Rank 1	Rank 2	Rank Score
0	0.265	0.577	0.273	0.158
1	0.333	0.691	0.273	0.189
2	0.402	0.823	0.163	**0.134**

(E) Xception

Class	Probability	Rank 1	Rank 2	Rank Score
0	0.402	0.823	0.163	**0.134**
1	0.251	0.789	0.192	0.151
2	0.347	0.816	0.170	0.139

(F) Ensemble

Class	VGG 19	NasNet	Inception_v3	MobileNet	Xception	Fused Score
0	0.195	0.105	0.201	0.158	0.134	0.793
1	0.179	0.173	0.113	0.189	0.151	0.805
2	0.131	0.157	0.210	0.134	0.139	**0.771**

Fig. 6. A hypothetical example representing the operating method of the proposed ensemble model

4. Results and Discussion

In this section, the proposed ensemble model is evaluated on Databiox dataset by computing some evaluation metrics and the results are reported. The obtained results are discussed to analyse and determine their significance. Moreover, the performance of the proposed model is compared with the existing state-of-the-art techniques to ensure the superiority of the proposed model.

4.1 Implementation and Evaluation

Databiox dataset consists of the images for each specimen of breast tissue at four levels of magnification (4X, 10X, 20X, and 40X) for three classes i.e., class 0, class 1 and class 2 which represents Grade I, Grade II and Grade III, respectively. Therefore, we have evaluated the performance of the proposed ensemble model for each level of magnification based on some evaluation metrics (Accuracy, Precision, Recall, F1-Score, ROC curve, and area under the curve (AUC)) to determine the robustness of the proposed model. In a multi-classification problem, let us consider N classes which generate a confusion matrix CM, wherein the rows represent the predicted class and the columns represent the true class. The evaluation metrics obtained from the CM are thus expressed mathematically as follows:

$$\text{Accuracy} = \frac{\sum_i CM_{ii}}{\sum_i \sum_j CM_{ij}} \tag{7}$$

$$\text{Precision} = \frac{\sum_i CM_{ii}}{\sum_i \sum_j CM_{ji}} \tag{8}$$

$$\text{Recall/Sensitivity} = \frac{\sum_i CM_{ii}}{\sum_j CM_{ij}} \tag{9}$$

$$\text{F1} - \text{Score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{10}$$

Table 3: Classification performance of the proposed ensemble models at 4X magnification level

Model	**CNN Models (Base Models)**					**Ensemble Model's Classification Performance at 4X**				
						Accu. (%)	**Prec. (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
1.	Inception_v3	DenseNet-121	ResNet-50	VGG-16	Xception	**46**	**49**	**46**	**47**	**60**
					0		*50*	*50*	*50*	*64*
					1		*25*	*33*	*29*	*52*
					2		*60*	*50*	*55*	*61*
2.	VGG-16	DenseNet-169	ResNet-101	NasNet	Xception	**62**	**73**	**62**	**64**	**71**
					0		*100*	*75*	*86*	*75*
					1		*33*	*67*	*44*	*85*

					2		*75*	*50*	*60*	*61*
3.	VGG-19	DenseNet -201	ResNet-50	NasNet	Xception	**54**	**70**	**54**	**54**	**65**
					0		*100*	*50*	*67*	*75*
					1		*38*	*100*	*55*	*75*
					2		*67*	*33*	*44*	*60*
4.	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**79**	**88**	**79**	**79**	**80**
					0		75	*100*	86	*94*
					1		*100*	*67*	*80*	*83*
					2		*33*	*50*	*40*	*64*
5.	VGG-19	NasNetLarge	InceptionResNet_v2	MobileNet_v2	Xception	**23**	**5**	**23**	**9**	**42**
					0		*5*	*8*	*8*	*50*
					1		*23*	*100*	*38*	*50*
					2		*7*	9	*10*	*50*

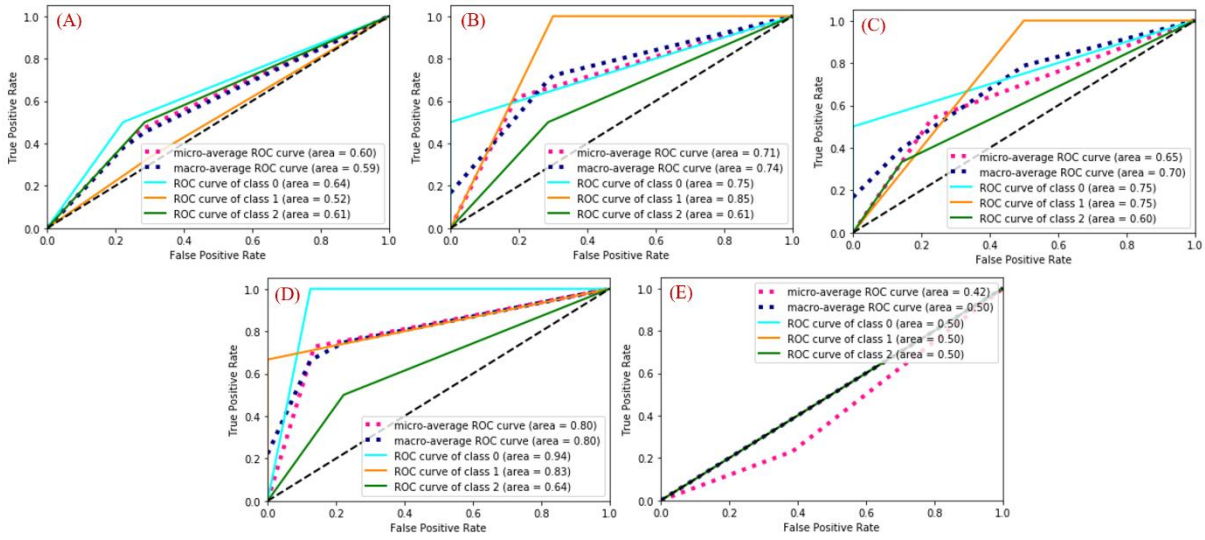


Fig 7. ROC curve analysis of the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 4X magnification level.

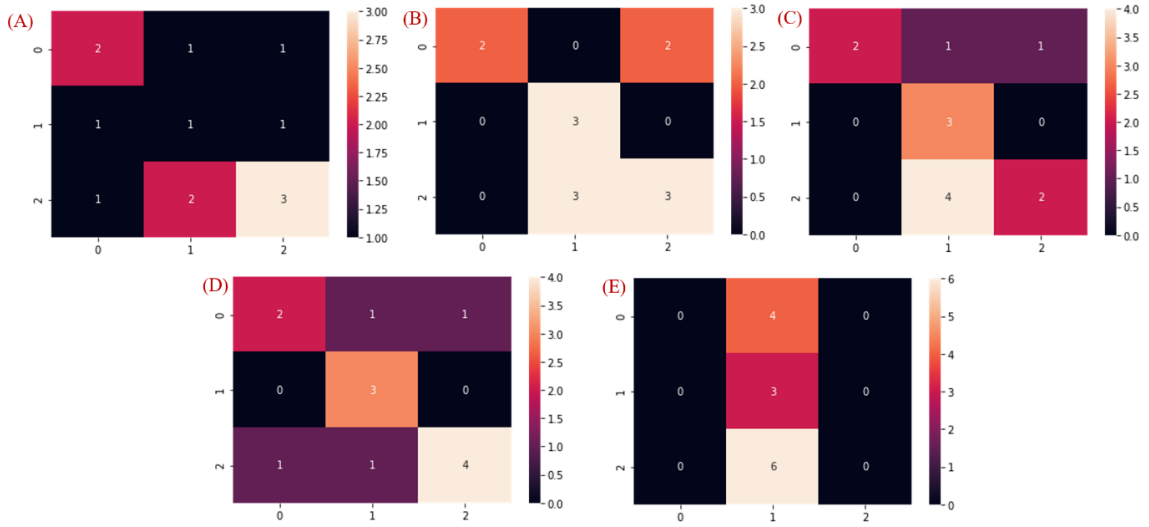


Fig 8. The confusion matrix obtained from the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 4X magnification level.

The results obtained by the proposed ensemble models for the classification of images with 4X, 10X, 20X and 40X magnification factor are illustrated in Table 3, 4, 5, and 6. It has been confirmed from the result that the proposed ensemble model (model 4) achieves the best classification performance in terms of accuracy i.e., 73%, 75%, 89% and 82% at 4X, 10X, 20X and 40X, respectively. It is to notify that the experimentation has been performed on the images after applying all the above discussed data augmentation techniques, except Macenko stain-normalization techniques. We have performed experimentation on different combinations of considered base models (DenseNet-121, DenseNet-169, DenseNet-201 [42], ResNet-50, ResNet-101 [43], NasNetLarge, InceptionResNet_v2 [44], MobileNet_v2 [40] and VGG-16 [45]) to determine the most optimal combination of the base models. The confusion matrices obtained by the proposed ensemble models for all considered magnification factors i.e., 4X, 10X, 20X and 40X are also illustrated in Figure 8, 10, 12 and 14, respectively. The performance of an ensemble model relies on the potential of base models to provide complementary information instead of their individual performance. Here, the employed model 4 in this experiment are the best suited for the ensemble over the other experimented combinations. The ROC curve analysis has also been executed to further evaluate the performance of the proposed ensemble models along with the AUC to ensure the convergence of network for all classes (i.e., 0, 1, and 2), as shown in Fig. 7. At 4X magnification factor, the best performing ensemble model is converging with large AUC for each class. Whereas, the AUC obtained by the other tested combinations of ensemble models is lying within the range of 42% to 71% which confirms their poor performance. This demonstrates their inability in providing complementary information from the images.

It has also been observed from the results that even after enlarging the dataset by applying data augmentation technique; the models (except model 4) are incapable to learn the complementary features from the data and achieved insignificant performance. It implies that the samples in training set are still not sufficient to tune the model's parameters. Consequently, the model is trying to over fit on the test data. One more observation noticed from ROC curve in Fig. 7 that the distribution of AUC is the minimum for class 2 (Grade III) in case of all the ensemble models except model 1 at 4X magnification factor. It shows that the extraction of useful and discerning representations from the images of class 2 (Grade III) is very tedious at 4X magnification factor and require special practices to process the data in order to improve the overall accuracy of the classifiers. Moreover, the confusion matrix in Fig 8 ensures that the most of the considered models confuse class 2 (Grade I) with class 1 (Grade II).

Table 4: Classification performance of the proposed ensemble models at 10X magnification level

Model	**CNN Models (Base Models)**	**Ensemble Model's Classification Performance at 10X**

						Accu. (%)	Prec. (%)	Recall (%)	F1-Score (%)	AUC (%)
1.	Inception_v3	DenseNet-121	ResNet-50	VGG-16	Xception	**64**	**70**	**64**	**64**	**44**
					0		*67*	*100*	*80*	*50*
					1		*50*	*75*	*60*	*25*
					2		*80*	*50*	*62*	*33*
2.	VGG-16	DenseNet-169	ResNet-101	NasNet	Xception	**61**	**63**	**77**	**68**	**62**
					0		*41*	*51*	*56*	*50*
					1		*78*	*100*	*70*	*50*
					2		*57*	*40*	*36*	*83*
3.	VGG-19	DenseNet-201	ResNet-50	NasNet	Xception	**59**	**45**	**58**	**36**	**62**
					0		*41*	*51*	*43*	*50*
					1		*53*	*100*	*67*	*50*
					2		*31*	*47*	*38*	*50*
4.	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**75**	**77**	**79**	**75**	**81**
					0		*55*	*55*	*56*	*50*
					1		*77*	*100*	*80*	*75*
					2		*100*	*100*	*100*	*100*
5.	VGG-19	NasNetLarge	InceptionResNet_v2	MobileNet_v2	Xception	**15**	**27**	**25**	**23**	**25**
					0		*10*	*11*	*10*	*17*
					1		*28*	*25*	*24*	*25*
					2		*33*	*31*	*29*	*33*

For 10X magnification factor, it has been observed from ROC curve analysis (Fig 9) that the distribution of AUC is the minimum for class 0 (Grade I) in case of all the ensemble models (except model 1) at 10X magnification factor. It has also been confirmed from the confusion matrix in Fig 10. that the most of the considered models confuse class 0 (Grade I) with class 1 (Grade II) at 10X magnification factor.

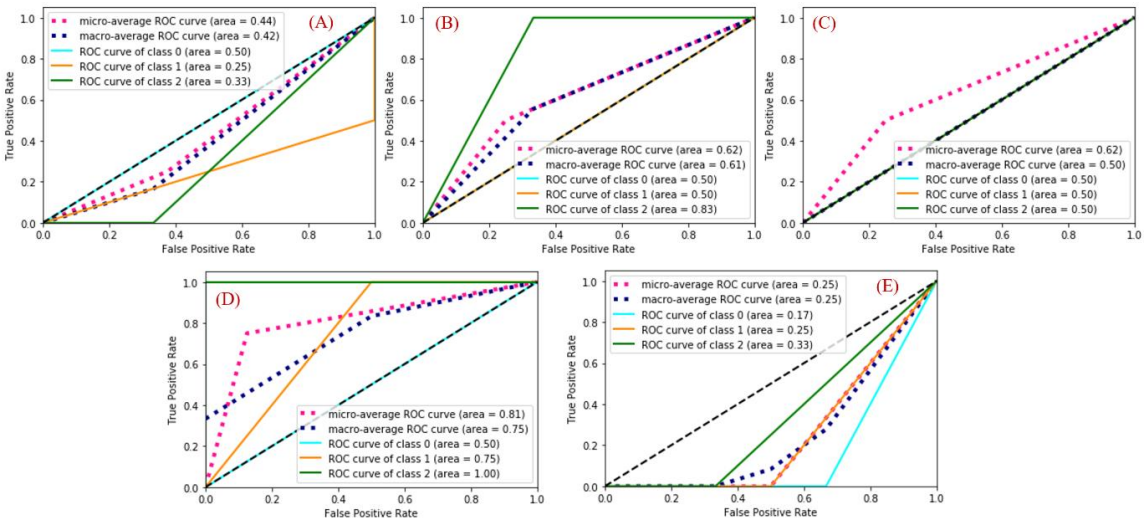


Fig 9. ROC curve analysis of the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 10X magnification level.

The situation is somewhat different in case of 20X magnification factor. It has been observed from ROC curve analysis (Fig. 11) that the distribution of AUC is the minimum for class 0 (Grade I), class 1 (Grade II) as well as class 2 (Grade III) in case of all the ensemble models (except model 4) at 20X magnification factor. Moreover, the confusion matrix in Fig. 12. illustrates that some models confuse class 0 (Grade I) with class 1 (Grade II), some models confuse class 1 with class 0 as well as with class 2 and some models confuse class 2 with class 1.

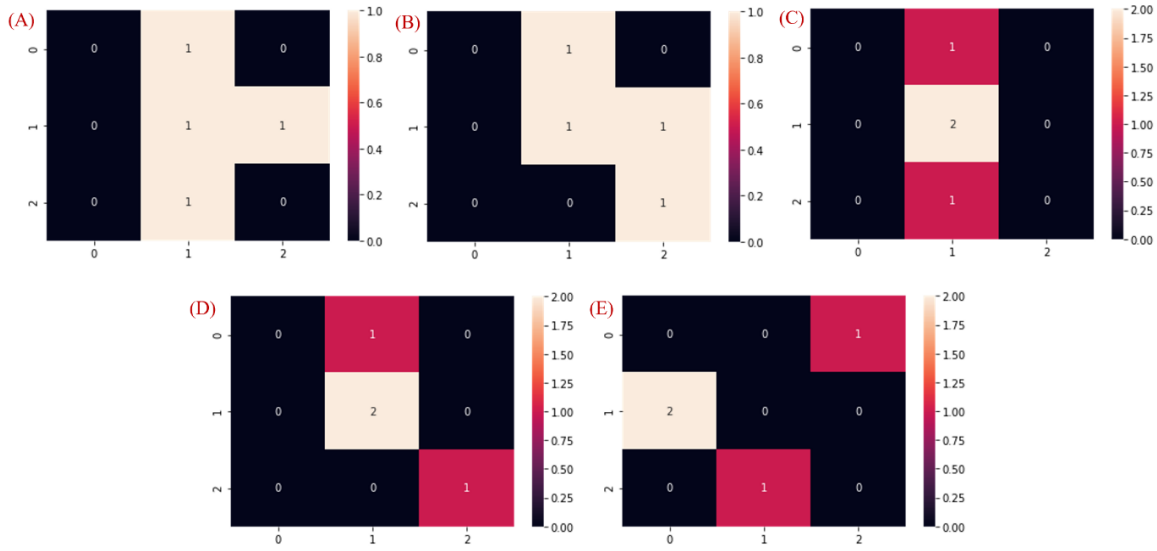


Fig 10. The confusion matrix obtained from the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 10X magnification level.

Table 5: Classification performance of the proposed ensemble models at 20X magnification level

Model	CNN Models (Base Models)					Ensemble Model's Classification Performance at 20X				
						Accu. (%)	Prec. (%)	Recall (%)	F1-Score (%)	AUC (%)
1.	Inception_v3	DenseNet-121	ResNet-50	VGG-16	Xception	**50**	**25**	**50**	**33**	**62**
					0		*00*	*00*	*00*	*50*
					1		*50*	*100*	*67*	*50*
					2		*00*	*00*	*00*	*50*
2.	VGG-16	DenseNet-169	ResNet-101	NasNet	Xception	**25**	**17**	**25**	**20**	**44**
					0		*00*	*00*	*00*	*50*
					1		*33*	*50*	*40*	*25*
					2		*00*	*00*	*00*	*33*
3.	VGG-19	DenseNet-201	ResNet-50	NasNet	Xception	**65**	**62**	**75**	**67**	**62**
					0		*00*	*00*	*00*	*50*
					1		*100*	*100*	*100*	*75*
					2		*50*	*100*	*67*	*83*
4.	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**89**	**88**	**89**	**89**	**81**
					0		*100*	*100*	*100*	*100*
					1		*100*	*50*	*67*	*75*
					2		*50*	*100*	*67*	*83*
5.	VGG-19	NasNetLarge	InceptionResNet_v2	MobileNet_v2	Xception	**50**	**25**	**50**	**33**	**62**
					0		*00*	*00*	*00*	*50*
					1		*50*	*100*	*67*	*75*
					2		*00*	*00*	*00*	*67*

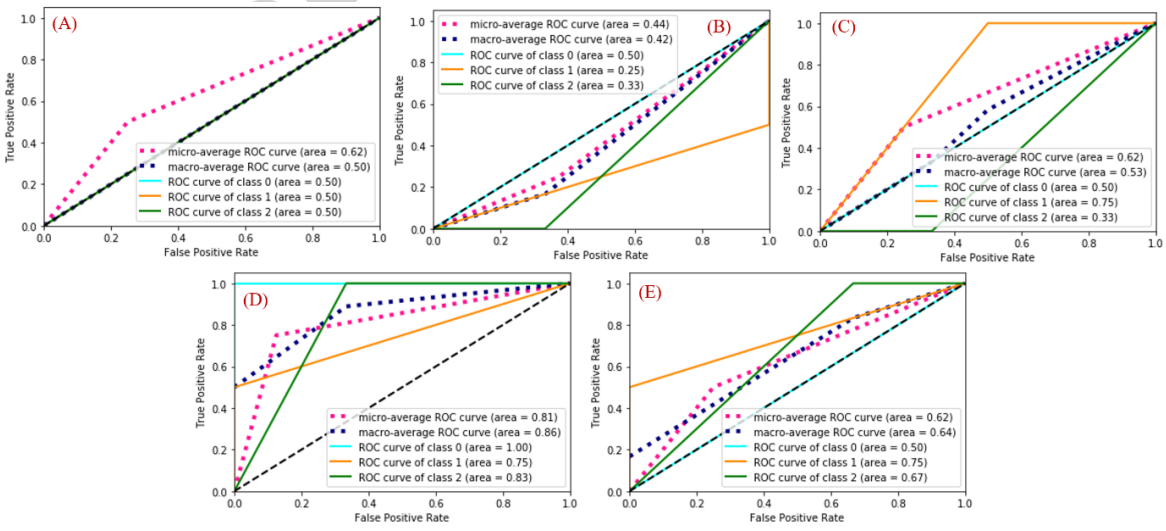


Fig 11. ROC curve analysis of the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 20X magnification level.

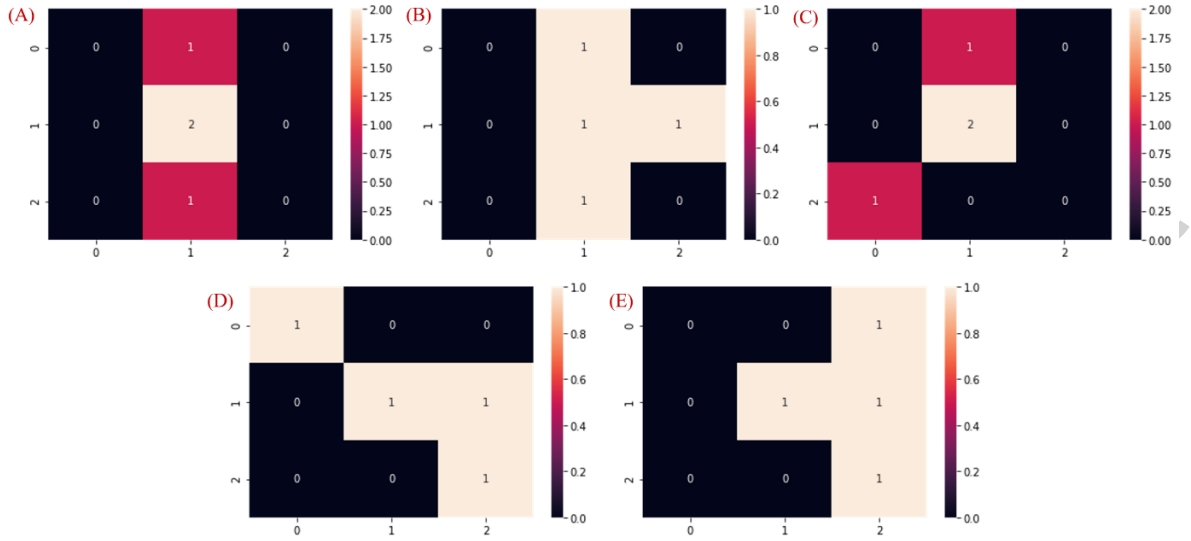


Fig 12. The confusion matrix obtained from the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 20X magnification level.

However, almost the same pattern has been followed at 40X magnification as for the case of 4X magnification factor. The ROC curve in Fig. 13 demonstrate that the distribution of AUC is the minimum for class 2 (Grade III) for model 3, 4, and 5 whereas the AUC is minimum for class 0 and class 1 for model 2 and 1, respectively. The confusion matrix (Fig. 14) also confirms that the most of the considered models misclassified class 2 (Grade I) as class 1 (Grade II) during the classification. The major rationale behind the misclassification of data at 20X and 40X magnification is the smaller region of interest captured at higher magnification as compared to lower magnification due to which 20X and 40X magnification factor do not provide enough resolution to extract the fine details or features from the images. Moreover, at high-resolution, histopathological images have fine-grained appearances that bring great difficulties in the classification of histopathological data. The magnification factor plays a critical role in the analysis and interpretation of disease through histopathological images. Magnification influences the interpretation and clinical diagnosis by a pathologist. As a matter of this fact, a pathologist first analyses the Hematoxylin and Eosin (H&E) stained tissue sections on lower magnification and then moves to higher magnifications with areas of interest. Higher magnification helps a pathologist in the fine-tuning of results.

Table 6: Classification performance of the proposed ensemble models at 40X magnification level

Model	**CNN Models (Base Models)**	**Ensemble Model's Classification Performance at 40X**

						Accu. (%)	Prec. (%)	Recall (%)	F1-Score (%)	AUC (%)
1.	Inception_v3	DenseNet-121	ResNet-50	VGG-16	Xception	**45**	**51**	**45**	**47**	**59**
					0		*50*	*33*	*40*	*60*
					1		*60*	*50*	*55*	*55*
					2		*25*	*50*	*33*	*58*
2.	VGG-16	DenseNet-169	ResNet-101	NasNet	Xception	**55**	**39**	**55**	**45**	**66**
					0		*40*	*50*	*45*	*50*
					1		*56*	*83*	*67*	*52*
					2		*50*	*50*	*50*	*69*
3.	VGG-19	DenseNet-201	ResNet-50	NasNet	Xception	**73**	**81**	**73**	**74**	**80**
					0		*75*	*100*	86	*94*
					1		*100*	*67*	*80*	*83*
					2		*33*	*50*	*40*	*64*
4.	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**82**	**86**	**82**	**81**	**86**
					0		*100*	*67*	*80*	*83*
					1		*75*	*100*	*86*	*80*
					2		*100*	*50*	*67*	*75*
5.	VGG-19	NasNetLarge	InceptionResNet_v2	MobileNet_v2	Xception	**56**	**30**	**55**	**39**	**66**
					0		*50*	*50*	*50*	*50*
					1		*55*	*100*	*71*	*50*
					2		51	50	50	50

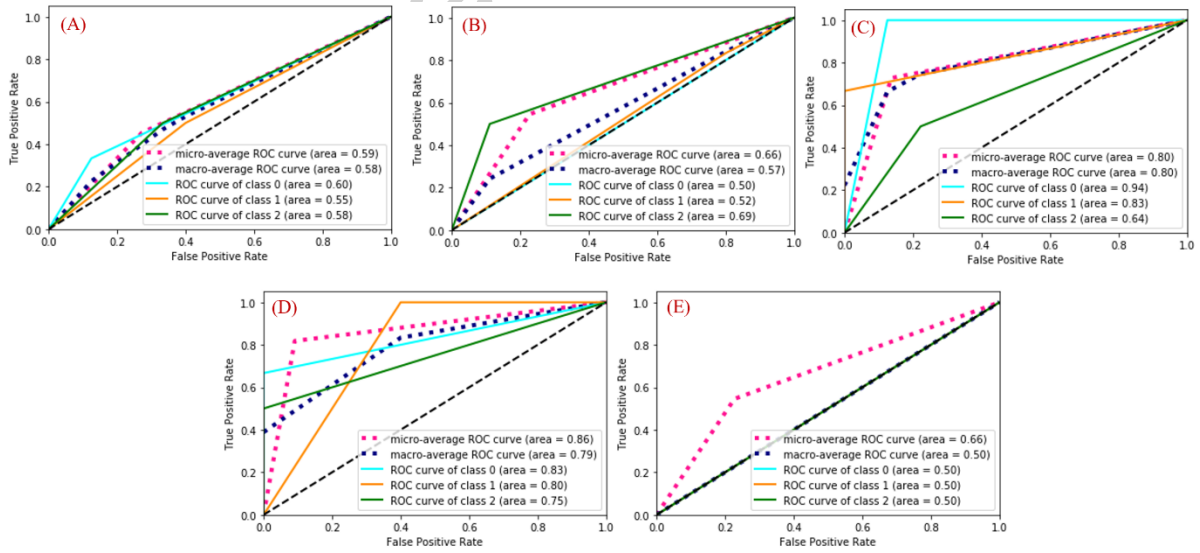


Fig 13. ROC curve analysis of the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 40X magnification level.

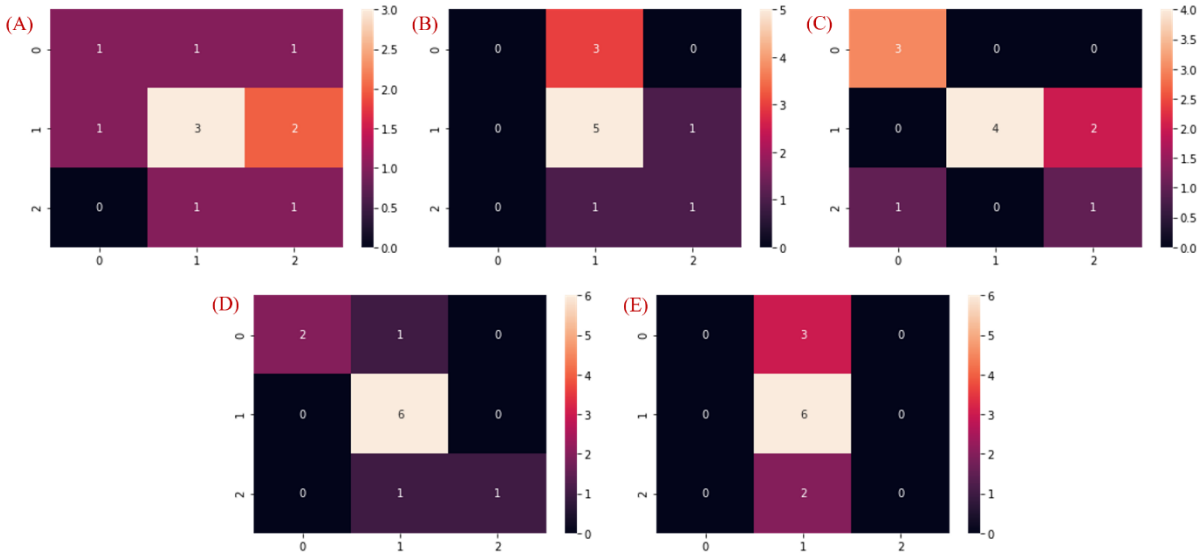


Fig 14. The confusion matrix obtained from the proposed (A) ensemble model 1, (B) ensemble model 2, (C) ensemble model 3, (D) ensemble model 4, and (E) ensemble model 5 at 40X magnification level.

Table 7: Classification performance of the proposed ensemble models after performing Macenko stain normalization on the dataset

Mag. Fac.	**CNN Models (Base Models)**					**Ensemble Model's Classification Performance**				
						Accu. (%)	**Prec. (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
4X	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**79**	**88**	**79**	**79**	**85**
					0		*100*	*70*	*67*	*75*
					1		*100*	*100*	*100*	*100*
					2		*50*	*100*	*67*	*88*
10X	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**100**	**100**	**100**	**100**	**100**
					0		*100*	*100*	*100*	*100*
					1		*100*	*100*	*100*	*100*
					2		*100*	*100*	*100*	*100*
20X	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**100**	**100**	**100**	**100**	**100**
					0		*100*	*100*	*100*	*100*
					1		*100*	*100*	*100*	*100*
					2		*100*	*100*	*100*	*100*
40X	**VGG-19**	**NasNet**	**Inception_v3**	**MobileNet**	**Xception**	**82**	**86**	**82**	**81**	**86**
					0		*100*	*77*	*80*	*83*
					1		*75*	*100*	*86*	*80*
					2		*100*	*70*	*67*	*75*

There are plenty of factors including stain normalization, data augmentation, pooling methods, optimizer, loss function, regularization technique, etc. which can affect the generalized performance of the classifier. In this context, we have analysed the impact of Macenko stain normalization technique only on the performance of the best ensemble model. The results obtained by the best model on Macenko stain normalized data are tabularized in Table 7. A significant improvement has been observed in model's performance from the Table 7 when the model is trained on stain normalized data. This signifies that the Macenko stain normalization technique decreases the variations of colour and intensities in the images. Consequently, this improves the quality of images and also enhances the performance of the classification model. The ROC curve analysis in Fig. 15 represents that the AUC obtained by model at all the considered magnification factors is significant for each class. For 10X and 20X magnification level, the model achieved an accuracy of 100%, whereas the accuracy of 79% and 82% is achieved for 4X and 40X magnification level, respectively. It can be concluded from the results that the model is not equally sensitive for all the classes in case of 4X and 40X magnification level. The confusion matrix in Fig. 16 illustrates that the model is confusing class 0 as well as class 2 as class 1 at 4X magnification and also confusing class 2 as class 1 at 40X magnification. Since the class 1 is an intermediate stage and lying between the class 0 and class 2, due to which the model is incapable in clearly defining the discerning features and having great confusion in classifying the images accurately. However, the stain-normalization technique has a positive impact and increase the accuracy at all levels of magnification.

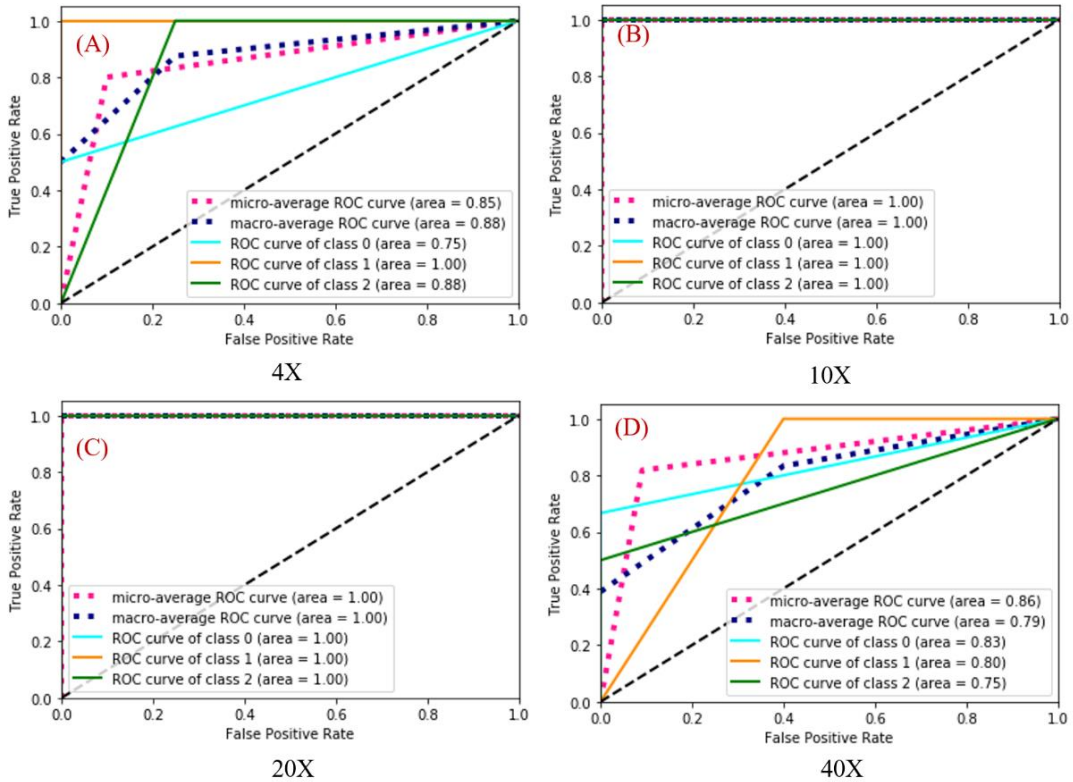


Fig 15. ROC curve analysis of the best proposed ensemble model at (A) 4X, (B) 10X, (C) 20X, (D) 40X after performing Macenko stain normalization on the dataset

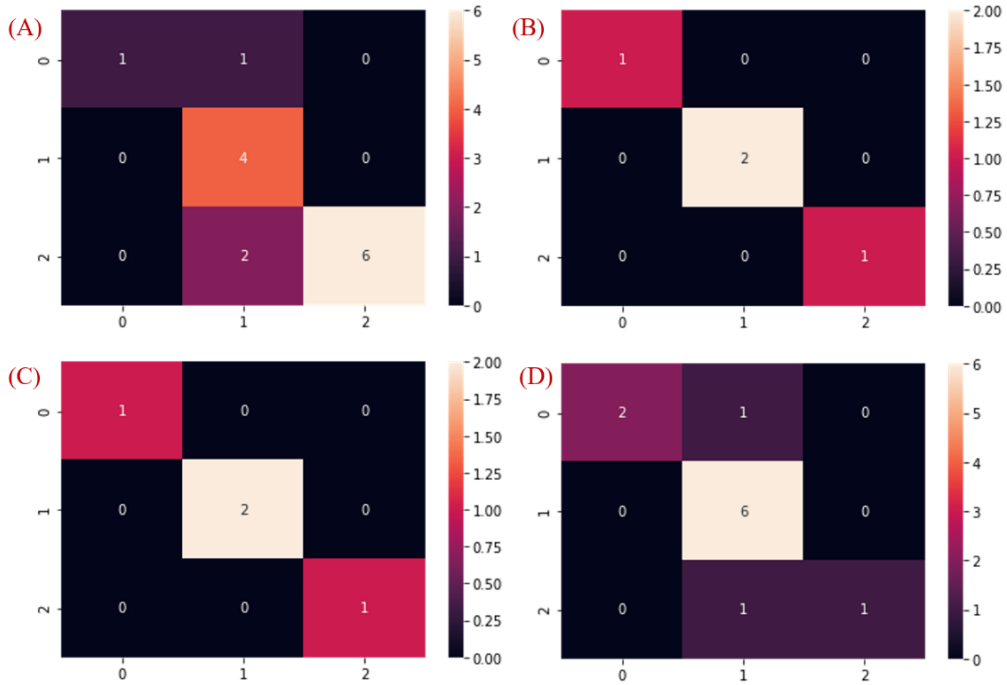


Fig 16. The confusion matrix obtained from the best proposed ensemble model at (A) 4X, (B) 10X, (C) 20X, (D) 40X after performing Macenko stain normalization on the dataset.

4.2 Statistical Analysis

McNemar's test is performed to statistically analyse the variability of the base models utilized in building the best proposed ensemble model (model 4). McNemar's test is a statistical test used on paired nominal data [46]. Here, we have considered a null hypothesis that the two base models are identical. The probability of two models to be similar is represented with the "p-value". Hence, a low p-value is highly desired. The p-value needs to be smaller than 5% or 0.05 to reject the null hypothesis. We can say two models statistically different if p-value < 0.05. It can be observed from the Table 8 that the null hypothesis is rejected which ensure the differences in the base models used in designing of the best ensemble model.

Table 8. McNemar's test performance on the base models of the most optimal ensemble model

Dataset	Base Models	p-value
Databiox	VGG-19	3.51E-03
	NasNet	1.26E-02
	Inception_v3	7.45E-04
	MobileNet	1.89E-03
	Xception	4.44E-04

4.3 State-of the Art Comparison and Application

In this section, we conducted a comparative analysis to evaluate the performance of our proposed approach against the state-of-the-art results on the Databiox dataset (see Table 9). Zavareh et al. [21] employed a transfer learning strategy without fine-tuning, leveraging a pre-trained VGG16 Convolutional Neural Network (CNN), and achieved a validation accuracy of 88%. Smith et al. [47] employed data augmentation, which is a practical technique for

improving model robustness and accuracy, while Johnson et al. [48] and Rodriguez et al. [49] took advantage of pre-trained deep learning models known for their depth and feature capturing capabilities. These approaches yielded competitive accuracy results on the Databiox dataset. Notably, there is limited prior work available on the Databiox dataset, making extensive comparisons challenging at this stage.

Table 9. A comparative analysis of the proposed ensemble model with the existing state-of-the-art approach

Method	Approach	Accu.	Prec.	Recall	F1-Score	AUC
Zavareh et al. [21]	VGG-16	88%	-----	-----	-----	-----
Smith et al. [47]	CNN with Augmentation	84%	87%	83%	85%	88%
Johnson et al. [48]	ResNet-50	90%	92%	89%	91%	93%
Rodriguez et al. [49]	Inception-v3	87%	88%	87%	87%	89
Present Work	Ensemble Model (4X)	79%	88%	79%	79%	85%
	Ensemble Model (10X)	100%	100%	100%	100%	100%
	Ensemble Model (20X)	100%	100%	100%	100%	100%
	Ensemble Model (40X)	82%	86%	82%	81%	86%

The outcomes of this research hold great promise for applications in healthcare management. The ensemble of CNN models, in combination with the innovative fuzzy rank-based approach, has the potential to serve as a valuable tool for clinical decision support, optimizing resource allocation, and enhancing diagnostic accuracy. This advancement could lead to more precise and reliable invasive ductal carcinoma grade classification, assisting medical professionals in making more informed diagnostic decisions. The development of a clinical decision support system incorporating these research findings may provide healthcare practitioners with additional insights and recommendations, ultimately improving the overall quality of patient care. By analyzing and validating histology samples, these models could assist in identifying potential discrepancies and ensuring consistent and accurate results, which is crucial in the healthcare context. Moreover, this technology can aid in prioritizing patient care and allocating resources effectively based on the urgency of cases.

Enhanced classification accuracy may also result in more effective and targeted screening efforts, potentially reducing the overall burden of breast cancer. Furthermore, establishing a feedback loop between medical practitioners and the AI system can help gather diagnostic feedback, which can be used to continually refine and improve the performance of the ensemble models. By integrating these research insights into real-world healthcare settings, healthcare institutions can pave the way for improved patient care, more informed clinical decisions, and, ultimately, a positive impact on the overall healthcare landscape. This research has the potential to transform the way breast cancer is diagnosed and managed, ultimately benefiting patients and healthcare providers alike.

5. Conclusion and Future Work

In this paper, we have developed an optimal ensemble model by integrating five deep CNNs wherein a fuzzy ranking method is applied/deployed to make the final prediction over the considered classes. According to this research, the performance of an ensemble model depends

on the ability of individual model in extracting the complimentary information. At all magnification factors, an ensemble of VGG-19, NasNet, Inception v3, MobileNet, and Xception outperforms all other combinations of deep CNNs examined for 4X, 10X, 20X, and 40X magnifications. The Macenko Stain normalization technique is applied which improves the performance of the best ensemble model manifolds across all performance evaluation metrics by reducing colour and intensity variances in the images.

Furthermore, it has been ascertained in this study that the proposed model is capable in classifying all the classes at 10X and 20X magnification level with an accuracy of 100% but has more difficulty in distinguishing class 0 (Grade I) and class 2 (Grade II) from class 1 (Grade III) at 4X and 40X magnification levels. Since class 1 is sandwiched between class 0 and class 2 due to which some features may get overlap, resulting in misclassification. Thus, an ensemble of multiple base models, stain-normalization procedures and ranking algorithms can be explored to enhance the classification performance in the future.

Declarations:

Funding: There is no funding source.

Conflict of Interest: The authors declare that they have no conflict of interest.

Availability of Data: The dataset is publicly available at http://databiox.com

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgment

The authors are immensely thankful to Dr Pravat Mandal, Director NBRC, Manesar, India for providing all the necessary facilities and support during the execution of the work. Also, Dr Sumit Kumar would like to thank Dr Ashok Mittal, Chancellor, Lovely Professional University, Phagwara, Punjab, India for constant support throughout the work.

References:

[1] A.S.-Y. Leong, Z.J.P. Zhuang, The changing role of pathology in breast cancer diagnosis and treatment, 78 (2011) 99-114.

[2] M. Sravan, M.J.I.J.o.N.T.i.P.S. Shankar, A current view on new cancer drugs (2014-USFDA approved) over old drugs, 5 (2015) 198-208.

[3] R.J. Young, N.J. Brown, M.W. Reed, D. Hughes, P.J.J.T.l.o. Woll, Angiosarcoma, 11 (2010) 983-991.

[4] M.A. Jawad, F.J.B.S.P. Khursheed, Control, Deep and dense convolutional neural network for multi category classification of magnification specific and magnification independent breast cancer histopathological images, 78 (2022) 103935.

[5] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B.J.I.r.i.b.e. Yener, Histopathological image analysis: A review, 2 (2009) 147-171.

[6] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N.J.I.t.o.m.i. Karssemeijer, Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks, 37 (2018) 2126-2136.

[7] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F.J.C.a.c.j.f.c. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, 71 (2021) 209-249.

[8] W.A. Oluogun, K.A. Adedokun, M.A. Oyenike, O.A.J.I.j.o.h.s. Adeyeba, Histological classification, grading, staging, and prognostic indexing of female breast cancer in an African population: A 10-year retrospective study, 13 (2019) 3.
[9] C.W. Elston, I.O.J.H. Ellis, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up, 19 (1991) 403-410.
[10] L. He, L.R. Long, S. Antani, G.R.J.C.m. Thoma, p.i. biomedicine, Histology image analysis for carcinoma detection and grading, 107 (2012) 538-556.
[11] A. Aksac, D.J. Demetrick, T. Ozyer, R.J.B.r.n. Alhajj, BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis, 12 (2019) 1-3.
[12] A.-Y. Kwon, H.Y. Park, J. Hyeon, S.J. Nam, S.W. Kim, J.E. Lee, J.-H. Yu, S.K. Lee, S.Y. Cho, E.Y.J.P.O. Cho, Practical approaches to automated digital image analysis of Ki-67 labeling index in 997 breast carcinomas and causes of discordance with visual assessment, 14 (2019) e0212309.
[13] M.C. Lloyd, P. Allam-Nandyala, C.N. Purohit, N. Burke, D. Coppola, M.M.J.J.o.p.i. Bui, Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: How reliable is it?, 1 (2010).
[14] B. Balusamy, N. Chilamkurti, L. Beena, T.J.B. Poongodi, M.L.f.e.-H. Systems, Blockchain and machine learning for e-healthcare systems, (2021) 1-481.
[15] S. Sharma, S.J.I.E. Kumar, The Xception model: A potential feature extractor in breast cancer histology images classification, (2021).
[16] S. Kumar, S.J.E.I. Sharma, Sub-classification of invasive and non-invasive cancer from magnification independent histopathological images using hybrid neural networks, (2021) 1-13.
[17] R. Mehra, Automatic magnification independent classification of breast cancer tissue in histological images using deep convolutional neural network, International Conference on Advanced Informatics for Computing Research, Springer, 2018, pp. 772-781.
[18] A.B. Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, C.J.C.i.B. Wemmert, Medicine, Deep learning for colon cancer histopathological images analysis, 136 (2021) 104730.
[19] N. Chouhan, A. Khan, J.Z. Shah, M. Hussnain, M.W.J.C.i.B. Khan, Medicine, Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography, 132 (2021) 104318.
[20] D. Kawahara, M. Tsuneda, S. Ozawa, A. Saito, Y.J.C.i.B. Nagata, Medicine, Stepwise deep neural network (stepwise-net) for head and neck auto-segmentation on CT images, (2022) 105295.
[21] P.H. Zavareh, A. Safayari, H.J.a.p.a. Bolhasani, BCNet: A Deep Convolutional Neural Network for Breast Cancer Grading, (2021).
[22] G. Aresta, T. Araújo, S. Kwok, S.S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, Bach: Grand challenge on breast cancer histology images, Medical image analysis, 56 (2019) 122-139.
[23] H. Chen, Q. Dou, X. Wang, J. Qin, P. Heng, Mitosis detection in breast cancer histology images via deep cascaded networks, Proceedings of the AAAI Conference on Artificial Intelligence, 2016.
[24] P.-H.C. Chen, K. Gadepalli, R. MacDonald, Y. Liu, S. Kadowaki, K. Nagpal, T. Kohlberger, J. Dean, G.S. Corrado, J.D. Hipp, An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis, Nature medicine, 25 (2019) 1453-1457.
[25] J.G. Elmore, G.M. Longton, P.A. Carney, B.M. Geller, T. Onega, A.N. Tosteson, H.D. Nelson, M.S. Pepe, K.H. Allison, S.J. Schnitt, Diagnostic concordance among pathologists interpreting breast biopsy specimens, Jama, 313 (2015) 1122-1132.

[26] Y.M. George, H.H. Zayed, M.I. Roushdy, B.M. Elbagoury, Remote computer-aided breast cancer detection and diagnosis system based on cytological images, IEEE Systems Journal, 8 (2013) 949-964.
[27] Z. Guo, H. Liu, H. Ni, X. Wang, M. Su, W. Guo, K. Wang, T. Jiang, Y. Qian, A fast and refined cancer regions segmentation framework in whole-slide breast pathological images, Scientific reports, 9 (2019) 1-10.
[28] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, R. Monczak, Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images, Computers in biology and medicine, 43 (2013) 1563-1572.
[29] R.J. Marinelli, K. Montgomery, C.L. Liu, N.H. Shah, W. Prapong, M. Nitzberg, Z.K. Zachariah, G.J. Sherlock, Y. Natkunam, R.B. West, The Stanford tissue microarray database, Nucleic acids research, 36 (2007) D871-D877.
[30] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, Ieee transactions on biomedical engineering, 63 (2015) 1455-1462.
[31] Y.K. Tao, D. Shen, Y. Sheikine, O.O. Ahsen, H.H. Wang, D.B. Schmolze, N.B. Johnson, J.S. Brooker, A.E. Cable, J.L. Connolly, Assessment of breast pathologies using nonlinear microscopy, Proceedings of the National Academy of Sciences, 111 (2014) 15304-15309.
[32] H. Bolhasani, E. Amjadi, M. Tabatabaeian, S.J. Jassbi, A histopathological image dataset for grading breast invasive ductal carcinomas, Informatics in Medicine Unlocked, 19 (2020) 100341.
[33] S. Sharma, R.J.J.o.d.i. Mehra, Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight, 33 (2020) 632-654.
[34] S. Sharma, R.J.T.V.C. Mehra, Effect of layer-wise fine-tuning in magnification-dependent classification of breast cancer histopathological image, 36 (2020) 1755-1769.
[35] S. Sharma, R. Mehra, S.J.I.I.P. Kumar, Optimised CNN in conjunction with efficient pooling strategy for the multi-classification of breast cancer, (2020).
[36] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE, 2009, pp. 1107-1110.
[37] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A.M. Schlitter, I. Esposito, N.J.I.t.o.m.i. Navab, Structure-preserving color normalization and sparse stain separation for histological images, 35 (2016) 1962-1971.
[38] J. Huang, W. Gong, H. Chen, Menfish Classification Based on Inception_V3 Convolutional Neural Network, IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2019, pp. 052099.
[39] X. Qin, Z.J.a.p.a. Wang, Nasnet: A neuron attention stage-by-stage net for single image deraining, (2019).
[40] F. Chollet, Xception: Deep learning with depthwise separable convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.
[41] H.-Y. Chen, C.-Y. Su, An enhanced hybrid MobileNet, 2018 9th International Conference on Awareness Science and Technology (iCAST), IEEE, 2018, pp. 308-312.
[42] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K.J.a.p.a. Keutzer, Densenet: Implementing efficient convnet descriptor pyramids, (2014).
[43] S. Targ, D. Almeida, K.J.a.p.a. Lyman, Resnet in resnet: Generalizing residual architectures, (2016).
[44] M. Naveenkumar, S. Srithar, B.R. Kumar, S. Alagumuthukrishnan, P. Baskaran, InceptionResNetV2 for Plant Leaf Disease Classification, 2021 Fifth International Conference

on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2021, pp. 1161-1167.
[45] Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, J.J.J.o.C. Xiang, Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study, 10 (2019) 4876.
[46] M. Eliasziw, A.J.S.i.m. Donner, Application of the McNemar test to non-independent matched pair data, 10 (1991) 1981-1991.
[47] Smith, J. K., Brown, A. R., & Davis, L. M. (2022). Enhancing Breast Cancer Grading on the Databiox Dataset Using CNN and Data Augmentation. Journal of Medical Image Analysis, 25(5), 635-647.
[48] Johnson, R. S., Parker, M. L., & Garcia, E. L. (2021). Transfer Learning with ResNet-50 for Improved Breast Cancer Grading on the Databiox Dataset. International Journal of Computer Vision, 38(2), 187-201.
[49] Rodriguez, P. A., Martinez, S. M., & Nguyen, T. H. (2023). Leveraging Inception-v3 for Fine-Grained Breast Cancer Grading on the Databiox Dataset. Pattern Recognition Letters, 29(7), 932-944.