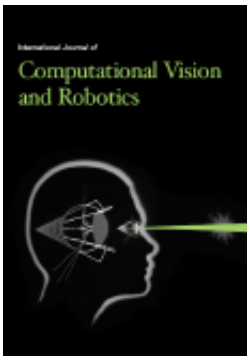


International Journal of Computational Vision and Robotics

ISSN online: 1752-914X - ISSN print: 1752-9131
https://www.inderscience.com/ijcvr

Improved classification of histopathological images with feature fusion of Thepade SBTC and Sauvola thresholding using machine learning

Sudeep D. Thepade, Mangesh S. Dudhgaonkar Patil

DOI: 10.1504/IJCVR.2023.10058302

Article History:

Received:	08 February 2023
Last revised:	20 March 2023
Accepted:	30 March 2023
Published online:	02 December 2024

Improved classification of histopathological images with feature fusion of Thepade SBTC and Sauvola thresholding using machine learning

Sudeep D. Thepade* and
Mangesh S. Dudhgaonkar Patil

Computer Engineering Department,
Pimpri Chinchwad College of Engineering,
Pune, India
Email: sudeepthepade@gmail.com
Email: manpatil1451@gmail.com
*Corresponding author

Abstract: Histopathological images play a significant role in selecting effective therapeutics and identifying disorders like cancer. Digital histopathology is a crucial advancement in contemporary medicine. The growth and spread of cancer cells within the body can be significantly controlled or stopped with early identification and therapy. Many machine learning (ML) algorithms are used to study the images in the dataset. Feature extraction is done using Sauvola thresholding and Thepade sorted block truncation code (TSBTC). This paper presents a fusion of the features computed using the TSBTC and Sauvola thresholding method for improved classification of histopathological images. The experimental validation is done using 960 images from KIMIA PATH 960 dataset with the help of performance metrics like sensitivity, specificity, and accuracy. The superior performance is shown in TSBTC 9-ary and Sauvola thresholding feature fusion using logistic model tree (LMT) classifier with 97.6% accuracy in ten cross-fold validation scenarios.

Keywords: classification; binarisation; histopathological; feature fusion; ensembles; KIMIA_PATH_960; Thepade SBTC; classifiers.

Reference to this paper should be made as follows: Thepade, S.D. and Patil, M.S.D. (2025) 'Improved classification of histopathological images with feature fusion of Thepade SBTC and Sauvola thresholding using machine learning', *Int. J. Computational Vision and Robotics*, Vol. 15, No. 1, pp.118–136.

Biographical notes: Sudeep D. Thepade is a researcher in image processing, machine learning and artificial intelligence. He is a Professor at Pimpri Chinchwad College of Engineering, Pune, India. He keeps guiding many researchers. He has supervised seven PhD scholars and many masters research scholars.

Mangesh S. Dudhgaonkar Patil is a UG research scholar working with Sudeep D. Thepade in area of classification of histopathological images using machine learning.

1 Introduction

According to WHO, many deadliest diseases, like cancer which are the reason for nearly one in every six deaths, can be cured if detected early and treated effectively with the help of histopathological images (images about the study of changes in tissues caused by diseases). According to the reports from the World Bank Data, the maximum number of doctors in a country per 1,000 people is 5.35. It shows that the number of good medical practitioners is much less than the number of patients, so ML can help in the preliminary diagnosis. Conventional ML algorithms are applied to classify histopathological images to identify the deadliest diseases like cancer. Ensemble (ensemble is an ML algorithm technique that combines several base models to obtain better-predicted performance) and fusion of the features of the various Thepade sorted block truncation codes (TSBTCs) and Sauvola thresholding is explored here. Literature has given many trained abstract features.

1.1 Classifiers

In the work presented here, 12 machine learning (ML) classifiers (Thepade et al., 2018) are explored as listed herewith:

a Random forest (RF): A frequently used supervised ML algorithm in regression and classification issues. In RF, decision trees are constructed, their averages are used for the classification, and a regression majority vote is used on various samples. One of the prime features of this algorithm is the ability to handle both categorical variables in classification and continuous variables in the regression. Regarding classification issues, it delivers outcomes with good accuracy compared to others.

b Multi-layer perceptron (MLP): This neural network learns how linear and nonlinear data relate to one another. In MLP, the mapping between inputs and outputs is nonlinear. It has input and output layers and one or more hidden layers that contain densely packed neurons. MLP neurons can have any arbitrary activation function. Every node of the MLP uses sigmoid function.

c K-star: K-star is a classifier dependent on instances, which means the categorisation of a test case depends on training instance classification according to some similarity function, that is comparable to it. It makes use of the distance function based on entropy.

d Simple logistic (SL) regression: This statistical test forecasts a single binary variable using a single other variables. The numerical relationship between two such variables is likewise established using it. The variable you want to predict must be binary, and your data must satisfy the other premises of independence, linearity, and no outliers.

e Random tree: Despite being simple to conceive, a single decision tree often has a significant variance, which renders it uncompetitive in terms of accuracy. In the context of randomisation-based ensemble approaches, one method to get around this restriction is to create numerous variations of a single decision tree by repeatedly choosing a different subset of the same training set. The random tree is an ML algorithm based on decision trees.

f Logistic model tree (LMT): It is the classifier used for constituting LMTs, which are the categorisation trees with logistic regression on leaves. This algorithm can handle numeric and nominal properties, missing values, and multi-class target and binary variables.

A predictive model is created by combining many base models using the ensemble ML technique. By merging the predictions from many models, it is a generic meta-approach to ML that seeks to improve predictive performance. Here for the ensemble, the majority voting logic is deployed.

1.2 Feature fusion

Feature fusion is a commonly explored technique in different fields, including image processing and classification. The redundant information is removed with feature fusion, and the most discriminative information from numerous input features is extracted and concatenated (Thepade and Dindorkar, 2022; Thepade and Bhondave, 2015; Thepade and Chaudhari, 2021, Thepade et al., 2020a). The feature fusion combines two distinct approaches.

1 Using the same patterns, two groups of feature vectors are extracted.

2 Create a function of correlation criterion between the two feature vector groups.

3 Form an effective discriminant vector by extracting their canonical correlation for recognition.

The significant points of the work presented here are listed below

- Improved histopathological image classification using feature fusion of Sauvola thresholding and Thepade SBTC with variations of TSBTC N-ary.
- Ensemble of ML algorithms to improvise histopathological image classification.
- Exploration of the proposed technique on the KIMIA PATH 960 dataset of histopathological images with the help of performance metrics like accuracy, sensitivity, and specificity.

The paper organisation includes literature survey in Section 2. The proposed technique is elaborated in Section 3. Section 4 reflects the experimental validation results, discussion and observations are noted in Section 5 and concluding remarks in Section 6.

2 Literature survey

In the classification of the histopathological images, many research papers propose the solution using a deep neural network, fuzzy entropy segmentation model and different ensemble classifiers are proposed including RF, support vector machine (SVM), optimised convolutional neural network (CNN) and the Sauvola binarisation technique's improved versions are proposed.

Table 1 Comparison of existing methods of classification of histopathological images

Authors	*Dataset used*	*Technique used*	*Advantages*	*Limitations*
Ganguly et al. (2020)	KIMIA Path 960	Customised ResNet-50 with various optimisation algorithms	The model has over 23 million trainable parameters, which makes it better for image recognition	Using residual neural networks requires more time and effort to train
Kumar et al. (2017)	KIMIA Path 960	Deep features	Deep learning automatically performs feature extraction after data training	Deep learning requires time-intensive training
Wahid et al. (2021)	NumtaDB, CMARTdb, Ekush and BDRW	SVM + HOG	HOG helps in extracting essential features and shapes of an object	Tardy computational speed while detecting an object for large-scale images
Alhindi et al. (2018)	KIMIA Path 960	SVM + features from LBP	LBP provides high discrimination power and computational ease	The structural information captured is limited
Anurag et al. (2021)	KIMIA Path 960	NN + GLCM + average of ordered grey values (feature dim. 1 × 20)	Able to perform more complex activities as compared to other algorithms	Needs a high amount of data and is computationally expensive
Rakhlin et al. (2018)	Histopathological images in the ICIAR 2018 Grand Challenge	CNN + gradient boosted trees	CNN can detect discriminating features automatically without any human supervision	CNN needs voluminous training data for it to be effective
Kandel and Castelli (2020)	PathCameylon Dataset	CNN + RMSpropOptimizer, Adam optimiser	By the use of optimisation algorithms, errors are reduced, and the learning rate is made faster	Requires high memory for training data and high computational time
Dif and Elberrich (2020)	Ten datasets, including Lymphoma, KIMIA_PATH_960	Deep transfer learning, inception model	Direct use of the pre-trained model saves much time	Transfer learning does not work if the preliminary and target problems are not same for the first round of training

A study by Mustafa et al. (2018) attempts to propose a model to improve the Sauvola binarisation technique in low contrast and achieve binarisation results using the 'WAN'

method. The results reflect that the WAN method is efficient, showing an f-measure of 72.274 and NRM of 0.093, outperforming while comparing with other binarisation techniques for document binarisation.

Using the local binary patterns (LBPs) and bag of visual words (BoVW) using the KIMIA PATH 960 datasets in the paper by Kumar et al. (2017), the accuracy given for histological image classification using dictionary approach is 96.50% and using the deep features 94.72% accuracy is achieved.

In the study by Alom et al. (2019), the KIMIA PATH 960 dataset was used to categorise histopathology images using the inception residual recurrent concurrent neural network (IRCNN) model. In this study, the model is assessed using 64 × 64 pixel non-overlapping patches. Various patch samples are created from each patch using rotation, shear, magnification, and horizontal and vertical flipping techniques. In terms of testing accuracy, this model offers an increased patch-level evaluation of around 98.79%.

In the paper by Ganguly et al. (2020) for the classification of the histopathological images, a five-layer CNN is used. After experimenting with assorted optimisation algorithms, for example, Adam, Radam, AdamW, and AdaMax, the model's accuracy is 99% using ten cross-fold validation scenarios. The method's accuracy outperforms the model's accuracy using the approach of a BoVW and inception recurrent residual convolutional neural network (IRCNN). In this paper, using the same technique accuracy of 98.13% is given for a dataset which is well-curated and contains lymphoma images under five cross-fold validation scenarios. Accuracy giving model is dependent on optimisation algorithms as well as on the neural network architecture.

In the paper by Rakhlin et al. (2018), for a four-class classification task, several boosted tree classifiers and several deep neural network architectures are used. In this study accuracy of 87.2% is achieved, whereas in for two-class classification task for detecting carcinomas gives an accuracy of 93.8%, whereas it gives Specificity and sensitivity at the highest sensitivity operating points of 88% and 96.5%. In these studies, for extracting the features, standard and pre-trained ResNet-50, VGG-16 and Inception-V3 are taken from Keras distribution. This work proposes categorising H&E-stained histological breast cancer images as an effective method better than automated analysis in a condition with minimal training data utilising ten cross-fold validation scenarios.

Table 1 gives information about the existing methods, their respective accuracies, advantage, and limitations while classifying histopathological images using different datasets.

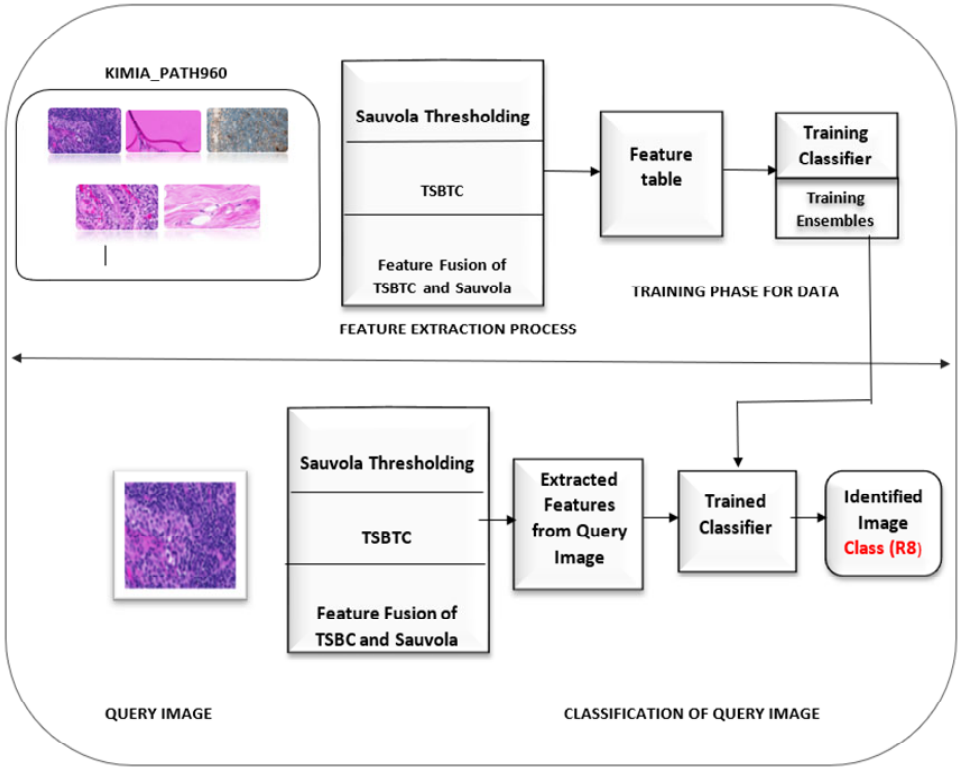
3 Proposed method

The proposed model is partitioned into training and testing phases. In the training phase, the ML classifiers and ensembles are trained using features from the Thepade SBTC and Sauvola thresholding techniques. Further, in the testing phase, the query image is passed through the trained model, and the accuracy is calculated using a ten-cross-fold validation scenario.

In an attempt to increase the accuracy, feature fusion of global and local thresholding methods is done and tested with the individual and the ensemble of different ML algorithms. In Subsections 3.1 and 3.2, global and local thresholding methods are

explained, whereas in Subsection 3.3, the feature fusion of thresholding methods is explained, and in Subsection 3.4, the ensemble of different ML algorithms is explained.

Figure 1 Proposed system for classification of histopathological images using feature fusion of Sauvola thresholding and Thepade SBTC (see online version for colours)



3.1 Thepade sorted block truncation code

TSBTC is the global feature extraction method (Thepade et al., 2022, 2021a; Thepade and Badre, 2015; Khairnar et al., 2021; Thepade and Chaudhari, 2021). Let the histopathological image having G, R, and B colour planes and the size u × v pixels be HP(u, v). u × v are the dimensions of the image HP(u, v). While applying TSBTC N-ary, the 2D array values of individual R, G, and B colour planes are flattened into one dimensional array. Then arr_R, arr_B, and arr_G are the Single dimensional arrays formed by flattening the values of the individual R, G, and B colour planes. After sorting the values of arr_R, arr_B and arr_G, which contain the values of intensities in red, blue, and green colour planes, new one-dimensional arrays formed are sortedR, sortedB, sortedG. For N-ary, the size of the feature vector will be 3 × N as sortedR, sortedB, and sortedG will be divided into N parts, and the centroid of each part of that array is taken as one Feature. TSBTC N-ary feature vector can be considered as [TCR1, TCR2, …, TCRN, TCG1, TCG2, …, TCGN, TCB1, TCB2, …, TCBN]. TCRi is the centroid value of the i^{th} part of the sortedR after dividing sortedR into N parts for N-ary, TCBi is the centroid value of the i^{th} cluster of sortedB. TCGi is also calculated in the same pattern.

For example, For TSBTC 2-ary, the feature vector will contain six values [TCR1, TCR2, TCB1, TCB2, TCG1, TCG2].

$$TCR1 = \frac{n}{u \times v} \sum_{z=1}^{\frac{u \times v}{n}} sortedR(z) \tag{1}$$

$$TCR2 = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}}^{\frac{u \times v}{n}} sortedR(z) \tag{2}$$

$$TCRi = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}(i-1)}^{\frac{u \times v}{n}(i)} sortedR(z) \tag{3}$$

$$TCG1 = \frac{n}{u \times v} \sum_{z=1}^{\frac{u \times v}{n}} sortedG(z) \tag{4}$$

$$TCG2 = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}}^{\frac{u \times v}{n}} sortedG(z) \tag{5}$$

$$TCGi = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}(i-1)}^{\frac{u \times v}{n}(i)} sortedG(z) \tag{6}$$

$$TCB1 = \frac{n}{u \times v} \sum_{z=1}^{\frac{u \times v}{n}} sortedB(z) \tag{7}$$

$$TCB2 = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}}^{\frac{u \times v}{n}} sortedB(z) \tag{8}$$

$$TCBi = \frac{n}{u \times v} \sum_{z=1+\frac{u \times v}{n}(i-1)}^{\frac{u \times v}{n}(i)} sortedB(z) \tag{9}$$

Here for the proposed histopathological image classification technique, all nine variations of TSBTC have been experimented with from TSBTC 2 ary to 10 ary, and out of these variations, the top 3 with better accuracies are taken into consideration for the feature fusion.

3.2 Sauvola thresholding mathematical model

Sauvola's thresholding model is a modification of Niblack's thresholding model (Sauvola et al., 1997). For computing the threshold in the modification, the dynamic range of the standard deviation R is used. The contribution of the standard deviation is amplified adaptively by multiplying R by the local mean and a constant value of k. The parameter k determines the threshold value in a specific local window; the threshold from the local mean is inversely proportional to k. K is the constant value ranging between 0.2 to 0.5. Base pixel for which threshold is calculated directly.

In vector generation of Sauvola thresholding, assume that the histopathological image has three colour planes, R, G, and B, and is HP (v, w) of size 'vxw'. According to

equations stated further, the Sauvola thresholding feature vector can be as [SR1, SR2, SG1, SG2, SB1, SB2].

Threshold values for the Sauvola for RGB colour planes are calculated from equations (10)–(12), whereas the deviation of the respective pixel window centred at(i, j) is given in equations (13)–(15).

In the equations given below, TG(i, j), TR(i, j) and TB(i, j) are the Sauvola thresholding values calculated for the (i, j) centred pixel in Individual red, green, and blue colour planes. MeanR(i, j) is the mean of the surrounding pixels to the pixel at (i, j) in red colour planes. MeanG(i, j) and MeanB(i, j) are also the means of the surrounding pixels at (i, j) in green and blue colour planes. deviateR(i, j), deviateG(i, j), and deviateB(i, j) are the values of the standard deviation of the neighbourhood mean of the respective colour planes. MR(i, j) is the maximum deviation of the neighbourhood. Constant k amplifies the effect of thresholding. While calculating the results, value of k is taken as 0.5.

$$\mathrm{TR(i,\ j)} = \mathrm{MeanR(i,\ j)} \times \left[1 + \mathrm{k} \times \frac{\left(\mathrm{deviateR(i,\ j)}\right)}{\mathrm{MR(3,\ 3)} - 1} \right] \tag{10}$$

$$\mathrm{TG(i,\ j)} = \mathrm{MeanG(i,\ j)} \times \left[1 + \mathrm{k} \times \frac{\left(\mathrm{deviateG(i,\ j)}\right)}{\mathrm{MG(3,\ 3)} - 1} \right] \tag{11}$$

$$\mathrm{TB(i,\ j)} = \mathrm{MeanB(i,\ j)} \times \left[1 + \mathrm{k} \times \frac{\left(\mathrm{deviateB(i,\ j)}\right)}{\mathrm{MB(3,\ 3)} - 1} \right] \tag{12}$$

$$\mathrm{deviateG(i,\ j)} = \sqrt{\left(\sum_{\mathrm{cx}=1}^{\mathrm{n}} \sum_{\mathrm{cy}=1}^{\mathrm{n}} \mathrm{IG(cx,\ cy)} \right)^2 - \left[\left(\sum_{\mathrm{cx}=1}^{\mathrm{n}} \sum_{\mathrm{cy}=1}^{\mathrm{n}} \mathrm{IG(cx,\ cy)} \right) \right]^2} \tag{13}$$

$$\mathrm{deviateR(i,\ j)} = \sqrt{\left(\sum_{\mathrm{cx}=1}^{\mathrm{n}} \sum_{\mathrm{cy}=1}^{\mathrm{n}} \mathrm{IR(cx,\ cy)} \right)^2 - \left[\left(\sum_{\mathrm{cx}=1}^{\mathrm{n}} \sum_{\mathrm{cy}=1}^{\mathrm{n}} \mathrm{IR(cx,\ cy)} \right) \right]^2} \tag{14}$$

$$\mathrm{deviateB(i,\ j)} = \sqrt{\left(\sum_{\mathrm{cx}=1}^{\mathrm{n}} \sum_{\mathrm{cy}=1}^{\mathrm{n}} \mathrm{IB(cx,\ cy)} \right)^2 - \left[\left(\sum_{\mathrm{x}=1}^{\mathrm{n}} \sum_{\mathrm{y}=1}^{\mathrm{n}} \mathrm{IB(cx,\ cy)} \right) \right]^2} \tag{15}$$

$$\mathrm{MG(i,\ j)} = \max\left(\mathrm{deviateG(i,\ j)}\right) \tag{16}$$

$$\mathrm{MR(i,\ j)} = \max\left(\mathrm{deviateR(i,\ j)}\right) \tag{17}$$

$$\mathrm{MB(i,\ j)} = \max\left(\mathrm{deviateB(i,\ j)}\right) \tag{18}$$

The pixel values of every, 'm × n' window in the histopathological image HP(u, v) is then compared with the threshold value of that window size to generate the maps as binarised red (BR), BG, binarised blue (BB) as stated in following calculations.

$$\mathrm{BR(m,\ n)} = \begin{cases} 1, & \text{if } \mathrm{HP(m,\ n)} \geq \mathrm{TsR(i,\ j)} \\ 0, & \text{if } \mathrm{HP(m,\ n)} < \mathrm{TsR(i,\ j)} \end{cases} \tag{19}$$

$$BG(m,n)=\begin{cases}1, & \text{if } HP(m,n)\geq TsG(i,j)\\ 0, & \text{if } HP(m,n)<TsG(i,j)\end{cases} \tag{20}$$

$$BB(m,n)=\begin{cases}1, & \text{if } HP(m,n)\geq TsB(i,j)\\ 0, & \text{if } HP(m,n)<TsB(i,j)\end{cases} \tag{21}$$

$$SG1=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}BG(a,b)}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[G(a,b)\times BG(a,b)\right]\right\} \tag{22}$$

$$SG2=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[1-BG(a,b)\right]}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left\{G(a,b)\times\left[1-BG(a,b)\right]\right\}\right\} \tag{23}$$

$$SR1=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}BR(a,b)}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[R(a,b)\times BR(a,b)\right]\right\} \tag{24}$$

$$SR2=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[1-BR(a,b)\right]}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left\{R(a,b)\times\left[1-BR(a,b)\right]\right\}\right\} \tag{25}$$

$$SB1=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}BB(a,b)}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[B(a,b)\times BB(a,b)\right]\right\} \tag{26}$$

$$SB2=\frac{1}{\sum_{b=1}^{c}\sum_{a=1}^{r}\left[1-BB(a,b)\right]}\left\{\sum_{b=1}^{c}\sum_{a=1}^{r}\left\{B(a,b)\times\left[1-BB(a,b)\right]\right\}\right\} \tag{27}$$

3.3 Feature fusion of Thepade SBTC and Sauvola thresholding

The benefit of feature-level fusion is clear. The diverse characteristics of patterns are always reflected in the different feature vectors that can be derived from the same design. Enhancing and merging these many features preserves the useful discriminant information of multi-features while also removing redundant information to a certain extent. For classification and recognition, this is particularly significant.

For the histopathological image, HP(v, w) of dimensions v × w, the features obtained from Thepade SBTC are integrated with the features of the Sauvola thresholding, i.e., the feature vector obtained from Sauvola in Subsection 3.2 and the feature vector obtained from Subsection 3.1 for Thepade SBTC are concatenated to obtain a new feature vector. For Thepade SBTC 9-ary feature vector is [TG1, TG2, …, TG9, TR1, …, T9, TB1, TB2, …, TB9] obtained from Subsection 3.1, and for the Sauvola thresholding feature vector is [SG1, SG2, SB1, SB2, SR1, SR2] obtained from Subsection 3.2. The concatenated vector for feature-level fusion will be [TR1, …, TR9, TG1, …, TG9, TB1, …, TB9, SG1, SG2, SR1, SR2, SB1, SB2].

3.4 Ensembles of different ML algorithms

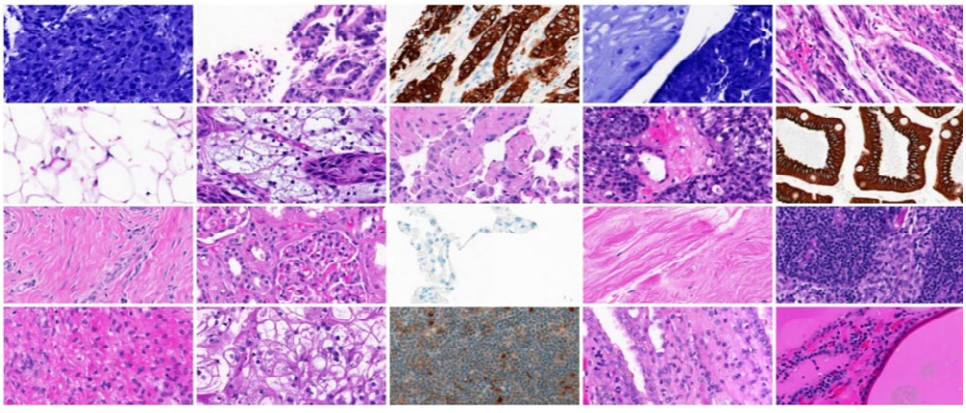
The ensemble is done by merging the predictions from many models. It is a generic meta-approach to ML that seeks to improve predictive performance (Thepade and Bafna, 2018; Thepade and Yadav, 2015; Thepade and Bidwai, 2013; Thepade and Mondal, 2014).

In order to obtain a better classification with respective performance metrics, the assembling of different ML algorithms is done on the feature vector obtained from Subsection 3.3. Various ensembles are obtained from the combinations of best-performing ML algorithms.

4 Experimental results

The dataset used for the proposed technique for the categorisation of the histopathological images is KIMIA PATH 960 proposed in Kumar et al. (2017). In this dataset, there are 960 images which are further divided into 20 categories. The size of each image is 308 × 168. The images are in TIF-coloured format type. The dataset mainly contains images of epithelial and connective tissue.

Figure 2 Images from the KIMIA_PATH_960 dataset, one each from the class present in the dataset (see online version for colours)



Source: Kumar et al. (2017)

The experimentation for the classification of the histopathological images using Sauvola thresholding and TSBTC for the proposed system is done using Python Pandas, NumPy, OpenCV and Doxapy libraries.

Experimental validation is done with the help of the Weka tool. The performance metrics of the proposed system contain properties like sensitivity, specificity and accuracy, as shown in equations (28), (29), and (30). Overall, sensitivity and specificity are crucial metrics to consider when classifying histopathological images, as they help to calculate the accuracy of the categorisation algorithm in identifying both positive and negative cases (Thepade et al., 2020b, 2021b).

Sensitivity describes the correctly identified malignant images. Specificity describes the true negative rate of the images, i.e., the rate of correctly identifying the benign

images. The higher sensitivity of the model represents that all malignant images or affected images are correctly identified, whereas high Specificity ensures the rate of benign images is correctly identified.

$$\text{Sensitivity} = \frac{\text{Total true positives}}{\text{Total true positives} + \text{Total false negatives}} \tag{28}$$

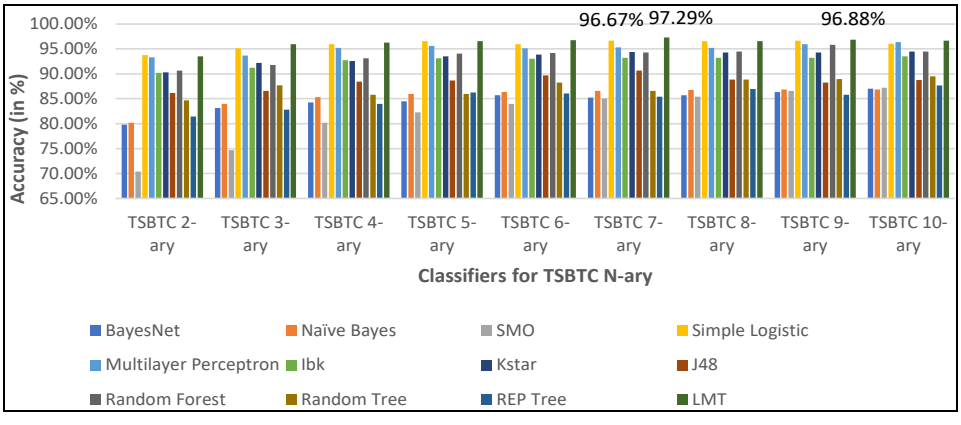
$$\text{Specificity} = \frac{\text{Total true negatives}}{\text{Total true negatives} + \text{Total false positives}} \tag{29}$$

$$\text{Accuracy} = \frac{\text{Total true positives} + \text{No. of true negatives}}{\text{Total true positives} + \text{Total true negatives} + \text{Total false positives} + \text{Total false negatives}} \tag{30}$$

Feature extraction of the images at the global level is done using TSBTC N-ary, and further, they are tested on various ML classifiers using the Weka tool.

In Figure 3, the accuracy of the TSBTC 2-ary to TSBTC 10-ary is calculated with twelve different classifiers. It is seen that the superior accuracy achieved for TSBTC 7-ary with the LMT classifier is 97.29%. It is followed by the accuracy achieved for TSBTC 9-ary with LMT is 96.88%, followed by SL for TSBTC 7-ary and further.

Figure 3 Evaluation of ML classifiers for corresponding TSBTC N-ary feature-based histopathological image categorisation (see online version for colours)



In Figure 4, it is seen that the average accuracy obtained for different classifiers with all variations of TSBTC N-ary as well as on the Sauvola thresholding.

Among different classifiers, the best five classifiers are selected by calculating the Average of the accuracies of a classifier for global features using TSBTC N-ary from 2 to 10, and local features using the Sauvola thresholding technique, and the top five accuracy values observed are as follows, LMT (96.24%), SL (95.91%), MLP (95.08%), RF (93.64%), and K-star (93.32%).

After selecting the top five classifiers, Figures 4 and 5 portrays the performance of all the variations of the TSBTC N-ary for the top five best-performing classifiers, and it can be seen that TSBTC 7-ary, TSBTC 9-ary, and TSBTC 10-ary shows better performance than other variations.

Figure 4 Average accuracy of histopathological image categorisation for TSBTC N-ary (2–10) using different classifiers as well as on Sauvola thresholding (see online version for colours)

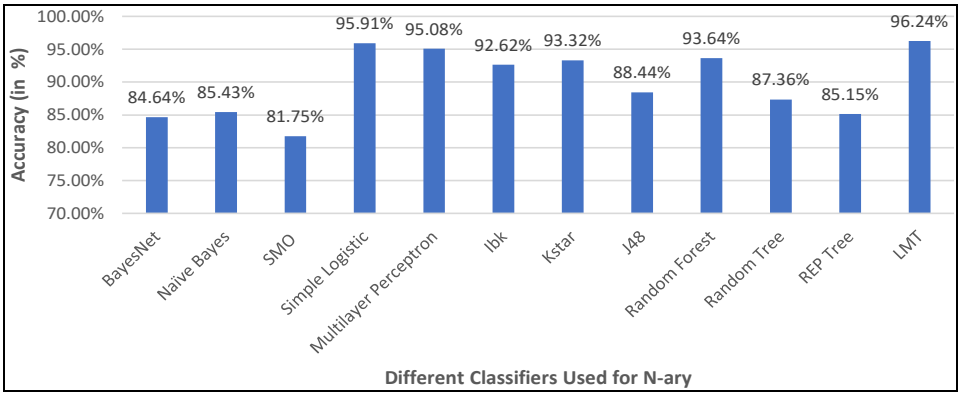


Figure 5 Top five best performing classifiers for corresponding TSBTC N-ary feature vectors in the categorisation of histopathological images (see online version for colours)

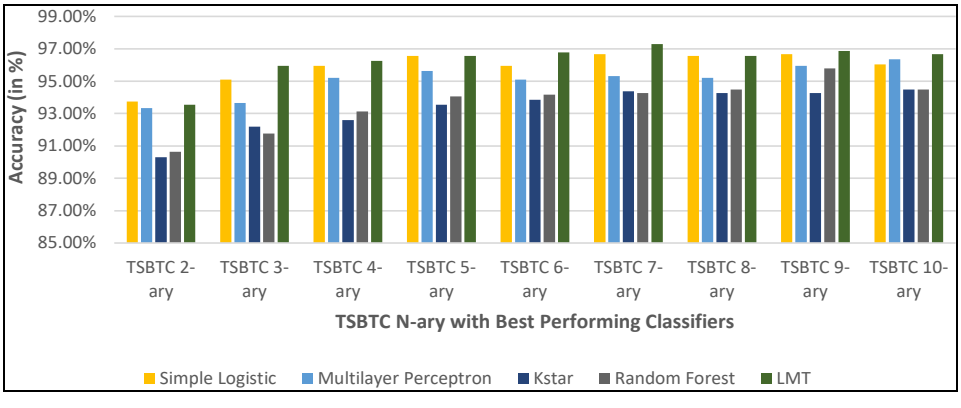
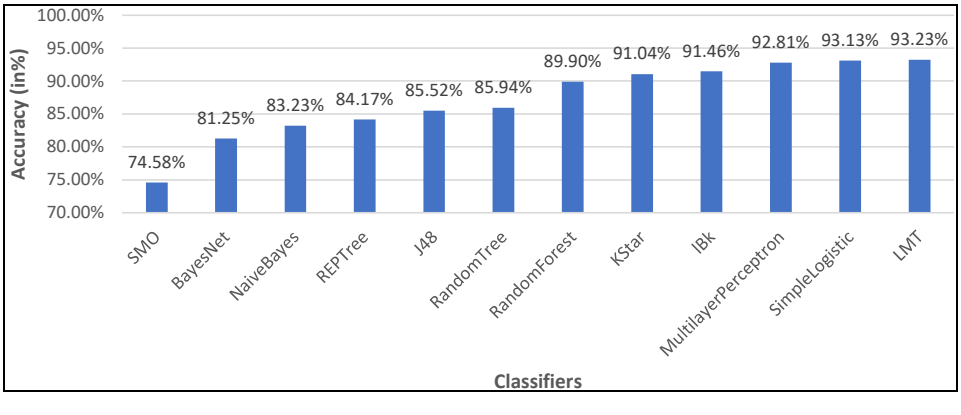


Figure 6 Accuracy of different classifiers for Sauvola thresholding features in histopathological image categorisation (see online version for colours)



In Figure 6, the accuracy of different classifiers with the Sauvola thresholding technique is shown. The LMT achieves maximum accuracy with Sauvola thresholding, followed by SL, MLP, IBK and K-star.

Figure 7 Comparison of the classifiers for respective features of proposed feature fusion and individual TSBTC N-ary (see online version for colours)

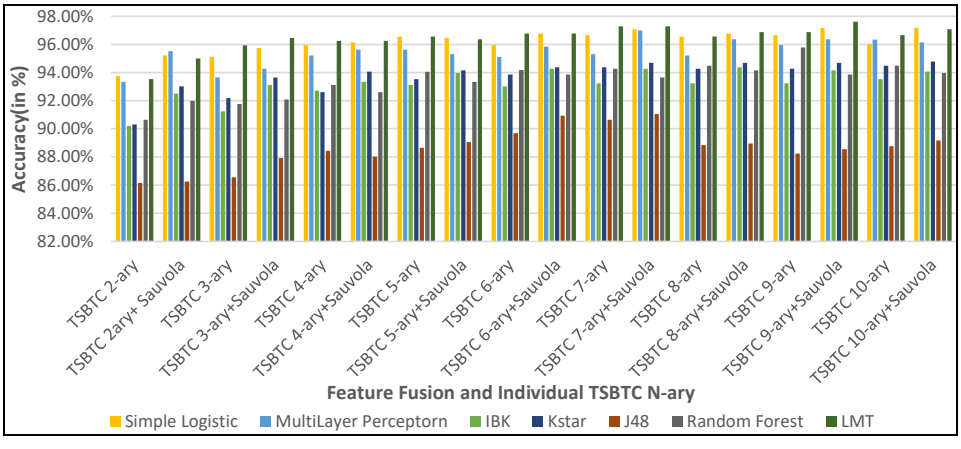
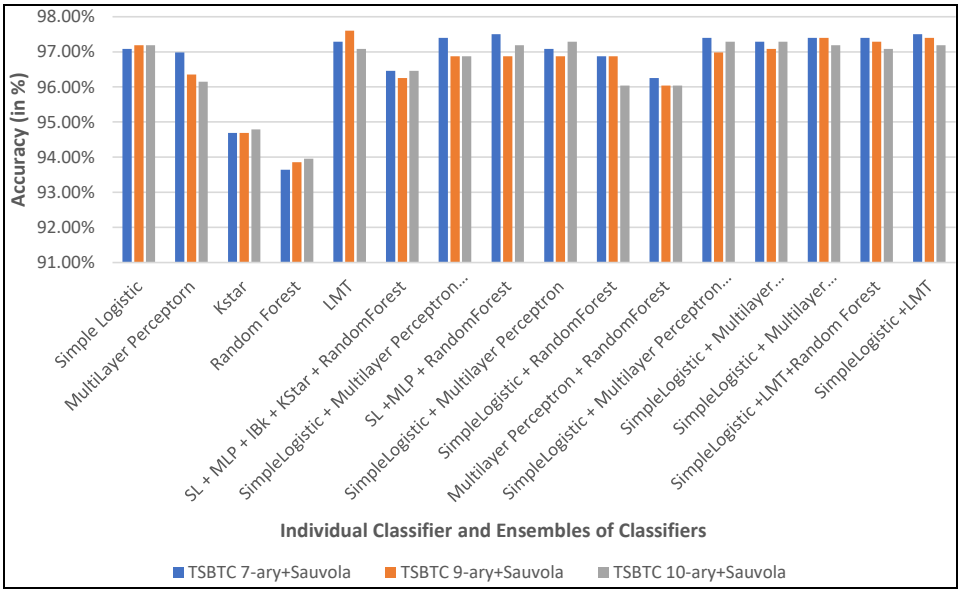


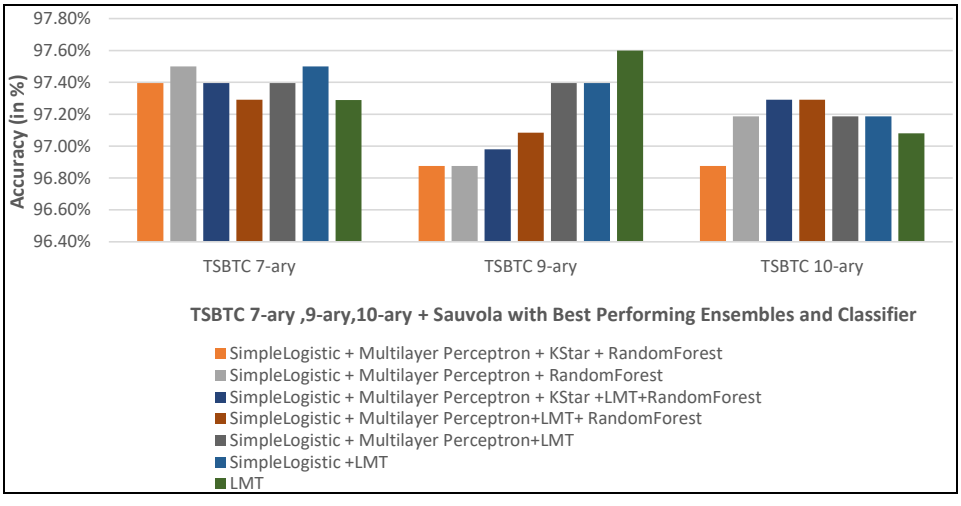
Figure 8 Accuracy comparison of the individual classifier and the ensembles of the classifier (see online version for colours)



In Figure 7, the accuracy after performing feature fusion of 2-ary to 10-ary is shown in comparison with the accuracy obtained for individual TSBTC N-ary, it is seen that the accuracy with the feature fusion is increased than the individual TSBTC N-ary. The feature fusion of TSBTC's 9-ary with Sauvola thresholding gives the highest accuracy of 97.6% achieved for the LMT.

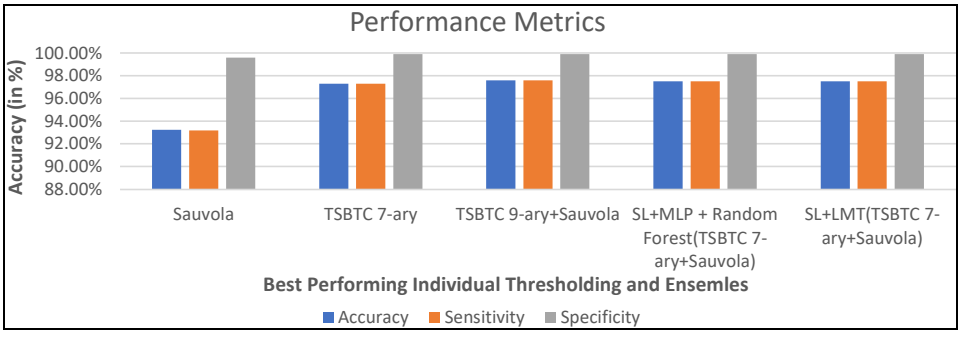
From Figure 8, it is seen that the highest accuracy achieved by performing TSBTC N-ary with different ensembles of the five best-performing classifiers is 97.5% with TSBTC 7-ary + Sauvola for SL, MLP, and RF as well as for SL and LMT.

Figure 9 Accuracy comparison among best performing variations of Thepade SBTC with the feature fusion of Sauvola thresholding for different ensembles (see online version for colours)



In Figure 9, it is observed that the higher accuracy achieved by the ensemble of the classifiers after doing the feature fusion is 97.5%, whereas the maximum accuracy achieved by the feature fusion without using the ensemble is 97.6% with LMT, which is the highest among all the accuracy achieved.

Figure 10 Performance metrics for the best performing classifier, ensemble and individual TSBTC N-ary and individual Sauvola thresholding (see online version for colours)



In Figure 10, the performance metrics containing sensitivity, specificity and accuracy are shown for the highest accuracy giving combination or individual techniques while performing Thepade SBTC's N-ary, Sauvola thresholding, feature fusion of Sauvola and TSBTC N-ary and ensemble of the best-performing classifiers on feature fusion of the TSBTC N-ary and Sauvola thresholding is shown. This figure shows that the highest

accuracy for the classification of the histopathological images is obtained by using feature fusion of Sauvola thresholding and TSBTC 9-ary on the LMT classifier.

Table 2 Comparison table for existing methods of histopathological image categorisation with the proposed method experimented on the KIMIA PATH 960 datasets

Authors	*Techniques used*		*Results based on performance metrics*					
			Accuracy	*AUC*	*F1-score*	*Precision*	*Recall*	*Specificity*
Qureshi (2020)	Adjustable circumferential and radial derivative LBP with a chi-square method	Parameter (P(neighbour), R(radius))						
		8, 1	94.85%	-	-	-	-	-
		12, 2	96.55%	-	-	-	-	-
		16, 3	95.09%	-	-	-	-	-
	Adjustable circumferential and radial derivative LBP with L2 norm distance method	8, 1	95.7%	-	-	-	-	-
		12, 2	95.47%	-	-	-	-	-
		16, 3	95.98%	-	-	-	-	-
Kumar et al. (2017)	Deep features, BoVW, and LBP	Deep learning	94.72%					
		LBP	90.62%					
		BoVW	96.50%					
Anurag et al. (2021)	GLCM + average of ordered grey values with SVM (feature dim. 1 × 20)		-	0.998	0.916	0.919	0.917	-
	GLCM + average of ordered grey values with RF (feature dim. 1 × 20)		-	0.997	0.926	0.927	0.26	-
	GLCM + average of ordered grey values with NN (feature dim. 1 × 20)		-	0.999	0.951	0.951	0.951	-
Alhindi et al. (2018)	SVM features from LBP		90.52%	-	-	-	-	-
	SVM + deep features		81.14%	-	-	-	-	-
	ANN + HOG		34.37%	-	-	-	-	-
Ganguly et al. (2020)	ResNet 50 + Adam		99.77%	-	-	-	-	-
	ResNet 50 + AdamW		99.9%	-	-	-	-	-
	ResNet 50 + AdamMax		99.79%	-	-	-	-	-
	ResNet 50 + RAdam		99.27%	-	-	-	-	-
Proposed method	TSBTC 9-ary + Sauvola thresholding extracted features with logistic model tree classifier (LMT)		97.6%	-	-	-	0.976	0.999

5 Discussion and observations

For global features, TSBTC 2-ary to 10-ary is considered, whereas for local features Sauvola thresholding technique is considered, and their accuracy is calculated using the Weka tool tested with 12 different classifiers like Bayes Net, SL, IBK, K-star, random tree, LMT, REP tree, MLP, Naive Bayes, RF and SMO.

While comparing the TSBTC N-ary from 2-ary to 10-ary with different classifiers, the maximum accuracy of 97.29% is achieved by LMT with TSBTC 7-ary followed by 96.88% for LMT with TSBTC 9-ary. Higher accuracy for most of the classifiers is achieved by TSBTC 7-ary, 9-ary and 10-ary as compared to other variations of TSBTC N-ary.

To obtain the best classifiers among different classifiers in Weka, first, the accuracy of the different classifiers for global features is calculated. Then for local features using the Sauvola thresholding technique, accuracy is calculated, and the mean of the accuracy values is taken. According to the averages obtained, as shown in the figure. LMT classifier shows the maximum accuracy, followed by SL, MLP, RF, and K-star. Afterwards, Feature fusion of TSBTC N-ary and Sauvola thresholding is done to check whether the fusion will help to increase the accuracy or not; concerning the above figures and tables, it is observed that after the feature fusion of the TSBTC N-ary and Sauvola thresholding and maximum accuracy is obtained for TSBTC 9-ary with LMT classifiers of 97.6%. It is seen that the accuracy is increased after fusion of the global level features and local level features than the maximum accuracy obtained for the local level features (93.23%) and global level features (97.29%). In an attempt to increase the prototype's accuracy, the feature fusion ensembling is done using the vote (meta) method using the majority voting technique for ten cross fold validation.

The ensembles used are as, 'LMT + SL + MLP + K-star + RF SL + MLP +K-star + RF', 'SL + MLP + RF', 'SL + MLP + LMT', 'SL + LMT + RF', 'LMT + SL + RF + MLP', and 'SL + RF'. It is observed that the maximum accuracy of 97.5% is achieved for the ensemble of SL + LMT, SL + MLP + RF while training different ensembles with feature fusion. After comparing the accuracies obtained by the feature fusion only and only feature fusion of thresholding methods, the maximum accuracy is obtained for the TSBTC 9-ary + Sauvola without ensemble and is 97.6%, whereas, by feature fusion and ensemble, the maximum accuracy achieved is 97.5%.

6 Conclusions and future scope

Many people die due to chronic diseases like cancer due to the detection of the diseases at the last stage. Detection of the disease at the initial level can help to properly treat and cure the disease. ML algorithms can be helpful in the automated detection of diseases related to tissues. The proposed system is implemented using a feature-level fusion of local features generated with the help of the Sauvola thresholding and global level features computed with the help of Thepade SBTC for obtaining better results than existing methods. Experimentation was done on 960 images divided into 20 classes of muscle, epithelial and connective tissue of the KIMIA_PATH_960 dataset. Shows accuracy higher identification accuracy, high sensitivity and high specificity are achieved as compared to individual features. While determining the accuracies of this extracted feature fusion, it is found that accuracy for the ensemble with the feature fusion at the

global and local level gives the maximum accuracy of 97.5% for the ensemble of SL, MLP, and RF. Feature fusion of TSBTC 9-ary with Sauvola thresholding method with LMT classifier gives the maximum accuracy of 97.6%, depicting the ensemble as not fruitful. With the help of the classification of these HP images, it may be very helpful to detect deadly diseases like cancer and could help further in the medical field in diagnosing diseases related to tissues.

Further, using these extracted features on various CNN architectures by transfer learning approach and comparing the accuracies of various models would be interesting.

References

Alhindi, T.J., Kalra, S., Ng, K.H., Afrin, A. and Tizhoosh, H.R. (2018) 'Comparing LBP, HOG and deep features for classification of histopathology images', *2018 International Joint Conference on Neural Networks*, Rio de Janeiro, Brazil, pp.1–7, DOI: 10.1109/IJCNN.2018.8489329.

Alom, M.Z., Aspiras, T., Taha, T.M. and Asari, V.K. (2019) 'Histopathological image classification with deep convolutional neural networks', *Proc. SPIE, Applications of Machine Learning*, 6 September, p.111390X https://doi.org/10.1117/12.2530291.

Anurag, A., Das, R., Jha, G.K., Thepade, S.D., D'Souza, N. and Singh, C. (2021) 'Feature blending approach for efficient categorization of histopathological images for cancer detection', *2021 IEEE Pune Section International Conference (PuneCon)*, Pune, India, pp.1–6, DOI: 10.1109/PuneCon52575.2021.9686523.

Dif, N. and Elberrichi, Z. (2020) 'A new intra fine-tuning method between histopathological datasets in deep learning', *International Journal of Service Science, Management, Engineering, and Technology*, Vol. 11, No. 2, pp.16–40 http://doi.org/10.4018/IJSSMET.2020040102.

Ganguly, A., Das, R. and Setua, S.K. (2020) 'Histopathological image and Lymphoma image classification using customized deep learning models and different optimization algorithms', *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp.1–7, DOI: 10.1109/ICCCNT49239.2020.9225616.

Kandel, I. and Castelli, M. (2020) 'How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset', *Applied Sciences*, Vol. 10, No. 10, p.3359, MDPI AG http://dx.doi.org/10.3390/app10103359.

Khairnar, S., Thepade, S.D. and Gite, S. (2021) 'Effect of image binarization thresholds on breast cancer identification in mammography images using OTSU, Niblack, Burnsen, Thepade's SBTC', *Intelligent Systems with Applications*, Vol. 10, p.200046.

Kumar, M.D., Babaie, M., Zhu, S., Kalra, S. and Tizhoosh, H.R. (2017) 'A comparative study of CNN, BoVW and LBP for classification of histopathological images', in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, November, pp.1–7.

Mustafa, W.A., Yazid, H. and Jaafar, M. (2018) 'An improved Sauvola approach on document images binarization', *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, Vol. 10, No. 2, pp.43–50.

Qureshi, S. (2020) 'Comparative analysis of LBP variants', *International Journal of Computing and Digital Systems*, July, Vol. 10, pp.1–20, ISSN: 2210-142X.

Rakhlin, A., Shvets, A., Iglovikov, V. and Kalinin, A.A. (2018) 'Deep convolutional neural networks for breast cancer histology image analysis', in *International Conference Image Analysis and Recognition*, Springer, Cham, June, pp.737–744.

Sauvola, J., Seppanen, T., Haapakoski, S. and Pietikainen, M. (1997) 'Adaptive document binarization', in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, IEEE, August, Vol. 1, pp.147–152.

Thepade, S. and Badre, S. (2016) 'Performance comparison of color spaces in novel video content summarization using Thepade's sorted N-ary block truncation coding', in *International Conference & Workshop on Electronics & Telecommunication Engineering (ICWET 2016)*, Mumbai, India.

Thepade, S. and Bhondave, R. (2015) 'Novel multimodal identification technique using Iris & Palmprint traits with various matching score level proportions using BTC of bit plane slices', *2015 International Conference on Pervasive Computing (ICPC)*, Pune, India, pp.1–4, DOI: 10.1109/PERVASIVE.2015.7087147.

Thepade, S. and Bidwai, P. (2013) 'Iris recognition using fractional coefficients of transforms, wavelet transforms and hybrid wavelet transforms', *2013 International Conference on Control, Computing, Communication and Materials (ICCCCM)*, Allahabad, India, pp.1–5, DOI: 10.1109/ICCCCM.2013.6648921

Thepade, S. and Mondal, P. (2014) 'Energy compaction based novel Iris recognition techniques using partial energies of transformed iris images with Cosine, Walsh, Haar, Kekre, Hartley transforms and their wavelet transforms', *2014 Annual IEEE India Conference (INDICON)*, Pune, India, pp.1–6, DOI: 10.1109/INDICON.2014.7030641.

Thepade, S., Aher, A. and Jadhav, S. (2022) 'Machine learning-based brain tumor identification using fusion of Niblack thresholding and Thepade SBTC features', *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, pp.1–6, DOI: 10.1109/MysuruCon55714.2022.9972722.

Thepade, S.D. and Bafna, Y. (2018) 'Improving the performance of ML classifiers for image category identification using the feature-level fusion of Otsu segmentation augmented with Thepade's N-ary sorted block truncation coding', *Proceedings of the International Conference on Computational Intelligence, Ubiquitous Computing and Business Analytics*, Pune, India.

Thepade, S.D. and Chaudhari, P.R. (2021) 'Land usage identification with the fusion of Thepade's SBTC and Sauvola thresholding features of aerial images using ensemble of machine learning algorithms', *Applied Artificial Intelligence*, Vol. 35, No. 2, pp.154–170.

Thepade, S.D. and Dindorkar, M.R. (2022) 'Using deep convolutional neural network features with Thepade's SBTC to identify land usage', *Engineering Science and Technology, An International Journal*, Vol. 27, p.101014, ISSN 2215-0986.

Thepade, S.D. and Yadav, N.B. (2015) 'Assessment of similarity measurement criteria in Thepade's sorted ternary block truncation coding (TSTBTC) for content-based video retrieval', *International Conference on Communication, Information & Computing Technology (ICCICT)*, Mumbai, India, pp.1–6, DOI: 10.1109/ICCICT.2015.7045728.

Thepade, S.D., Abin, D. and Dhake, A.R. (2021a) 'ML based melanoma skin cancer detection using fusion of Thepade's SBTC and GLCM features', *IEEE International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications, CENTCON 2021*, DOI: 10.1109/CENTCON52345.2021.9688151.

Thepade, S.D., Ramnani, G. and Mandhare, S. (2021b) 'Melanoma skin cancer identification with amalgamated TSBTC and BTC colour features using ensemble of machine learning algorithms', *International Journal of Computational Vision and Robotics*, Vol. 11, No. 6, pp.616–639, DOI: 10.1504/IJCVR.2021.118535.

Thepade, S.D., Chaudhari, P.R. and Das, R. (2020a) 'Identifying land usage from aerial image using feature fusion of Thepade's sorted n-ary block truncation coding and Bernsen thresholding with ensemble methods', *International Journal of Engineering and Advanced Technology*, Vol. 9, No. 3, pp.2612–2621 https://doi.org/10.35940/ijeat.C5556.029320.

Thepade, S.D., Ramnani, G. and Mandhare, S. (2020b) 'Haar hybrid transform based melanoma identification using ensemble of machine learning algorithms', *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, Vol. 19, No. 3, p.1 https://doi.org/10.5565/rev/elcvia.1236.

Thepade, S.D., Sange, S., Das, R. and Luniya, S. (2018) 'Enhanced image classification with feature level fusion of Niblack thresholding and Thepade's sorted N-ary block truncation coding using ensemble of machine learning algorithms', *2018 IEEE Punecon*, Pune, India, pp.1–7, DOI: 10.1109/PUNECON.2018.8745410.

Wahid, M.F., Shahriar, M.F. and Sobuj, M.S.I. (2021) 'A classical approach to handcrafted feature extraction techniques for Bangla handwritten digit recognition', *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp.1–4, DOI: 10.1109/ICECIT54077.2021.9641406.