Received: November 25, 2020. Revised: January 28, 2021.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Edge Boost Curve Transform and Modified ReliefF Algorithm for Communicable and Non Communicable Disease Detection Using Pathology Images

Shiva Sumanth Reddy[1]* **Nandini Channegowda[1]**

[1]*Department of Computer science and Engineering,*
Dayananda Sagara Academy of Technology and Management, Bangalore, India
* Corresponding author's Email: sumanthdsatm@gmail.com

Abstract: In this paper, a five phase model is proposed for early detection of communicable and non-communicable diseases like Haemoprotozoan and breast cancer using pathology images. At first, color normalization technique is utilized to improve the visual quality of the collected histology images. Next, edge boost curve transform is employed to segment nuclei and non-nuclei cells from the enhanced images. The developed segmentation methodology delivers good results in overlapped database. Further, the segmented image is converted into one dimensional vectors and then modified reliefF algorithm is applied to choose the active feature vectors to achieve better classification. Finally, deep neural network is accomplished to classify the Haemoprotozoan images as anaplasmosis, babesiosis and theileriosis, and breast images as malignant or benign. From the experimental result, the proposed model; modified reliefF-deep neural network obtained maximum classification accuracy of 97.6% in Haemoprotozoan disease detection and 95.94% in breast cancer detection, which are better related to other comparative techniques like Random Forest, Multi Support Vector Machine and K-Nearest Neighbor.

Keywords: Breast cancer detection, Canny edge detection, Circular hough transform, Color normalization, Deep neural network, Haemoprotozoan disease detection, Modified reliefF algorithm.

1. Introduction

In recent times, breast cancer has a higher mortality and morbidity among women according to the world cancer report. In India, breast cancer is the 2nd largest chronic disease, approximately 300,000 people get affected each year [1, 2], so early detection is essential to diminish the mortality rate of breast cancer (non-communicable disease). Additionally, Haemoprotozoan disease (communicable disease) is very common in tropical and sub-tropical regions, which causes economic losses to the livestock industry [3]. Haemoprotozoan disease is mainly transmitted by blood transfusion and occasionally through ixodid tick. The two most important Haemoprotozoan diseases transmitted of cattle are theileriosis and babesiosis, which are caused by Theileria spp and Babesia spp. The rickettsial disease caused by Anaplasma spp is named as anaplasmosis [4, 5]. Though, Haemoprotozoan tick not only transmit the diseases to the animals and also causes hide damage, anaemia and tick paralysis [6]. The Haemoparasitaemic animals are emaciated with poor reproductive and productive performances, anaemic and reduced working capacity in bullocks [7, 8]. So, early diagnosis and an effective treatment are compulsory to prevent the animals from death that improves the production ratio of a country [9]. Recently, histopathological image analysis is an effective imaging modality technique for cancer diagnosis and recognition. Histopathological image analysis assists clinicians in diagnosing the tumor and its sub-types, where the two basic types of tasks in the pathology image analysis are image segmentation and classification [10]. In this paper, a deep learning based model is proposed to perform pathology cell segmentation and classification for early diagnosis of Haemoprotozoan disease and breast cancer.

Initially, the Haemoprotozoan disease related to pathology images is collected from a real time database and breast cancer pathology images is collected from BreaKHis dataset. Next, a color normalization technique is used to improve the visibility level of the collected images by altering the range of pixel intensity values. Then, the nuclei and non-nuclei cell segmentation is performed using edge boost curve transform. In this technique, canny edge detection is applied to obtain edge images and it is fed to circular Hough transform to redefine the images as circles, and ellipses for better cell segmentation. Further, the cell regions are precisely separated from the segmented images based on the radius and center location of each cell and then the separated cells are resized as 32×32. The obtained 2D pathology image is converted into 1D vectors, and then modified reliefF algorithm is applied to select the active feature vectors from the total vectors. Modified ReliefF algorithm reduces the "curse of dimensionality" problem that results in better disease classification. The obtained features are fed to Deep Neural Network (DNN) classifier to classify Haemoprotozoan images as anaplasmosis, babesiosis and theileriosis, and breast images as malignant or benign. In the experimental section, the proposed modified reliefF-DNN model performance is validated by means of accuracy, balanced accuracy, sensitivity, specificity and f-score.

This research article is prepared as follows: In section 2, a few recent research papers on the topic "pathology image segmentation and classification" are surveyed. The detailed explanation about the proposed modified reliefF-DNN model is given in the Section 3. The experimental analysis of the proposed modified reliefF-DNN model is stated in the Section 4. Conclusion of the present research is given in the Section 5.

2. Literature survey

C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, [11] developed a hybrid Convolutional Neural Network (CNN) model for breast cancer histopathology image classification. The hybrid CNN model contains a local model and a hybrid model branch, where the developed model has strong representation ability by merging two branch information and local voting. Additionally, the redundant channels were removed from the hybrid CNN model by including squeeze excitation pruning block in the embedding layer. This procedure decreases the overfitting problem and also helps in delivering a higher classification accuracy. The simulation result showed that the developed hybrid CNN model outperformed the existing models in breast cancer histopathology image classification. Y. Xu, Z. Jia, L.B. Wang, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang, [12] developed leveraging deep CNN activation features to perform visualization, segmentation and classification in the large scale tissue histopathology images. In this study, ImageNet was utilized to transfer the extracted features from trained image databases to histopathology images. By visualizing the neuron components in the hidden layers, the properties of CNN features were explored. However, CNN is a region based pixel labeling, so it cannot explicit the higher level dependency between the points on the object boundaries to preserve the overall smoothness. In addition, CNN model is highly expensive in real time applications, because it needs computing hardware like neuromorphic chips and graphics processing units.

A. Chakravarty, and J. Sivaswamy, [13] developed a Recurrent Neural Network (RNN) based solution named as RACE-net for bio-medical image segmentation. In this literature, the developed RACE-net model performance was validated on three segmentation tasks like left atrium in cardiac MRI scans, cell nuclei in histopathology images, and optic cup and disc in fundus retinal images. The experimental results showed that the RACE-net model achieved better segmentation performance compared to existing U-net model. Hence, the developed RACE-net model mitigate the vanishing gradients concerns, so it cannot incorporate with high level features to achieve better classification accuracy. Further, X. Li, Y. Wang, Q. Tang, Z. Fan, and J. Yu, [14] developed a dual U-Net structure to segment the overlapped glioma nuclei from the histology images. The developed dual U-Net structure use both region and boundary information to enhance the segmentation accuracy of glioma nuclei. A new regression methodology was used to predict the distance map in order to refine the segmentation and the final segmentation was achieved using the fusion layers. The dual U-Net structure overcomes the issues faced by the researchers in the existing studies like touching or overlapping nuclei, irregular shapes, and intra or inter color variations. Hence, dual U-Net structure achieved a good performance in glioma cases, since the accuracy of touching nuclei with serious deformations are less which leads to over-segmentation problem.

A. Albayrak, and G. Bilgin, [15] developed a two phase segmentation method to segment the cell structures from the histology images. Initially, a simple linear iterative clustering method was applied to segment the super-pixels from the images and then

a global clustering methodology was used to cluster the same super-pixels that contains cell nuclei. The simple linear iterative clustering method was effective in eliminating the image artifacts and smoothening the local variance of the neighborhood pixels. The experimental results showed that the developed two phase segmentation method achieved better histopathological cell segmentation performance by means off-measure, true positive rate, precision, computation time and true negative rate. The performance of the developed method completely depends on the quality of pre-computed boundary maps. H. Jiang, S. Li, W. Liu, H. Zheng, J. Liu, and Y. Zhang, [16] developed a Geometric Feature Spectrum Extreme-Net (GFS Extreme-Net) model for cell detection. The developed model showed a promising and broader application potential in microscopic image analysis. Hence, the developed GFS Extreme-Net model consumes more time for labeling, and also it is very difficult to identify the specific extreme points that reflects the best geometric features of a target. Additionally, H. Li, X. Zhao, A. Su, H. Zhang, J. Liu, and G. Gu, [17] developed a weight map on the basis of distance transformation weight and class weight to improve the ability of loss function in U-Net for effectively learning the cell border feature. The experimental results showed that the developed model achieved better performance in white blood cell segmentation on the ALL-IDB1 database. However, the developed model is not suitable to solve the segmentation problem on small medical data sample that is a major problem in this literature. P. Alirezazadeh, B. Hejrati, A. Monsef-Esfahani, and A. Fathi, [18] developed a new unsupervised system for histopathological breast cancer detection. Initially, correlation metric was used to overcome the mismatch between the test and trained feature values into a domain invariant space. Then, an adaptation approach was developed based on representation learning to improve the detection rate of malignant images from the benign images. Finally, classification was carried out using decision tree, random forest, nearest neighbor, SVM and Quadratic Linear Analysis (QLA). In that, QLA attained better classification accuracy of 88.50% on BreaKHis database. Major issue with the adaptation approach was the registration of multiplexed images, because the physical displacements were occurred easily during the sequential image of the similar individual. In order to address the aforementioned problems, modified reliefF-DNN model is proposed to improve the histopathological cell segmentation and classification performance in communicable and non-communicable diseases.

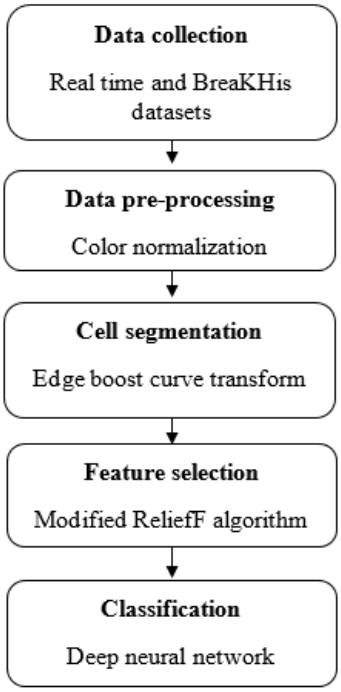


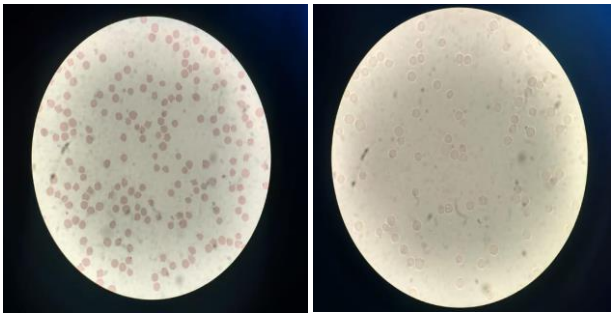
Figure. 1 Workflow of proposed system

3. Methodology

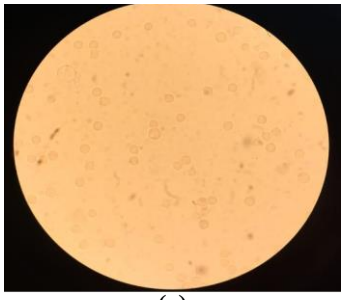
The proposed system includes five phases such as **data collection:** real time and BreaKHis datasets, **data pre-processing:** color normalization, **cell segmentation:** edge boost curve transform, **feature selection:** modified reliefF algorithm, and **classification**: DNN. The work flow of the proposed system is graphically indicated in Fig. 1.

3.1 Data collection and pre-processing

In this research study, real time and BreaKHis datasets are used for experimental investigation. The real time dataset comprises of 98 pathology images, 11 anaplasmosis images, 60 babesiosis images, and 27 theileriosis images. At the border of the cell, a rink link occurrence will be there in anaplasmosis images, and a cell with two dual structure is called as babesiosis images. In addition, a cell with circular big dot or rod like structure is called as theileriosis images. The graphical depiction of anaplasmosis, babesiosis and theileriosis images are indicated in Fig. 2. BreaKHis dataset comprises of 7909 image samples with two major classes such as malignant and benign. The malignant subset consists of 5429 samples, and the benign subset consists of 2440 samples, and it is graphically stated in Fig. 3. After data collection, color normalization technique is undertaken for enhancing the visible level of the collected pathology images [19]. General formula of color normalization technique is defined in Eq. (1).

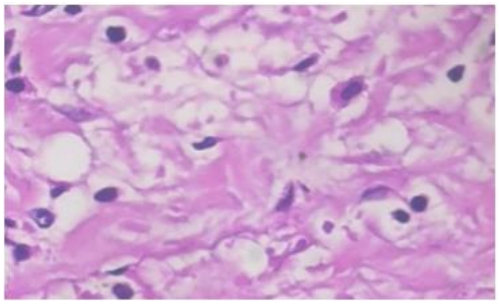


(a) (b)

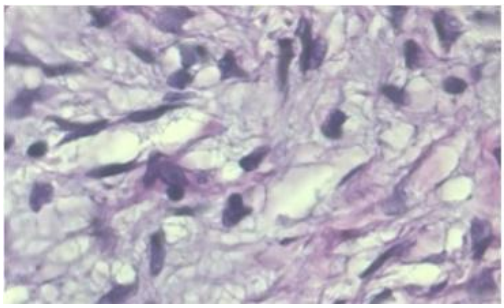


(c)

Figure. 2 Collected haemoprotozoan images: (a) anaplasmosis, (b) babesiosis, and (c) theileriosis

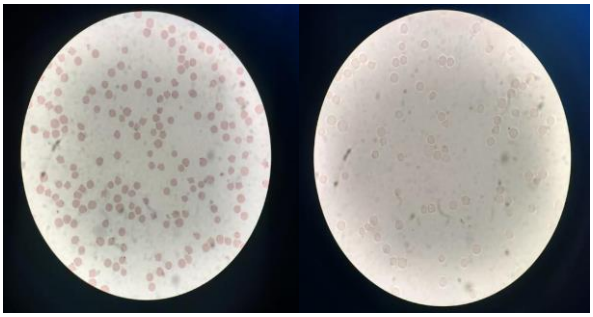


(a)

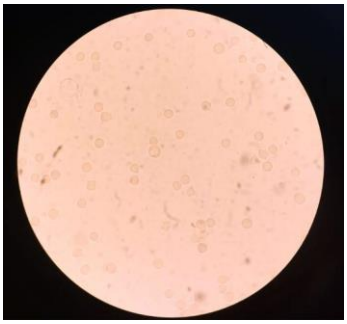


(b)

Figure. 3 Sample breast images: (a) malignant class and (b) benign class

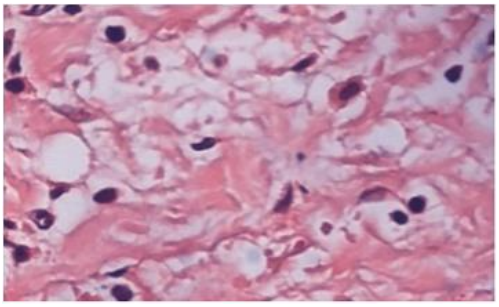


(a) (b)

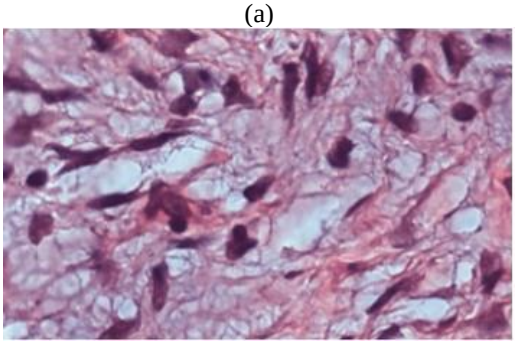


(c)

Figure. 4 Normalized images: (a) anaplasmosis, (b) babesiosis, and (c) theileriosis



(a)



(b)

Figure. 5 Normalized breast images: (a) malignant class and (b) benign class

$$I_{norm} = (I - Min) \times \frac{newMax - newMin}{Max - Min} + newMin \quad (1)$$

Where, I is indicated as collected pathology images, I_{norm} is denoted as normalized images, and $Max - Min$ is specified as minimum and maximum range of image pixel intensity value that ranges between 0 to 255. The graphical representation of normalized anaplasmosis, babesiosis and theileriosis pathology images are indicated in Fig. 4. Hence, the normalized breast images are indicated in Fig. 5.

3.2 Cell segmentation

After improving the visibility level of images, edge boost curve transform is applied to segment the nuclei and non-nuclei cells. In this technique, canny edge detector is an effective edge detection operator, which is used to detect the extensive range of edges in the enhanced histology images [20]. Steps involved in canny edge detector are given as follows:

Step 1: Initially, Gaussian filter [21] is used to remove noise from the enhanced histology images.

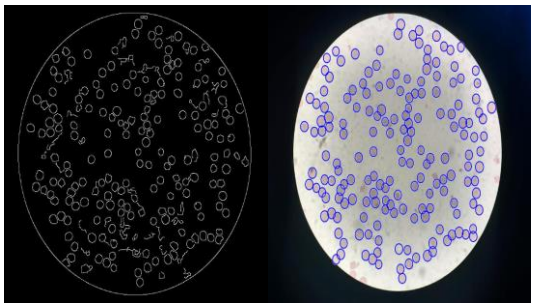
Step 2: Then, sobel operator is used to identify the image gradients for highlighting the nuclei and non-nuclei cells.

Step 3: Next, suppress the image pixels that are not at the maximum (non-maximum suppression).

Step 4: Hysteresis is applied to track the residual image pixels that are not suppressed. Further, the double thresholding technique utilizes 2 thresholds T1 and T2 for classifying the gradients into 3 groups.

- Gradients<T1 is a non-edge point.
- Gradients>T2 is an edge point.
- Or-else, the decision is taken based on the existing edge paths and direction of the point. The output image of the canny edge detector is fed to circular Hough transform to segment the cell regions.

Circular Hough transform is utilized to locate the regular curve in the output images of canny edge detector. This circular Hough Transform re-defines the images as circles, ellipses and expressions with powers of three and above. In this transformation technique, circle candidates are generated by voting in the Hough parameter space and then select local maxima in the accumulator matrix [22]. The output image of canny edge detector and circular Hough transform is represented in Figs. 6 and 7. By using bounding box, cell regions are separated based on center location and radius of every cell. Next, the cell size is fixed as 32×32, and the respective two dimensional histology image is converted into one dimensional vector.



(a) (b)

Figure. 6 Segmented haemoprotozoan images: (a) canny edge detection and (b) circular hough transform

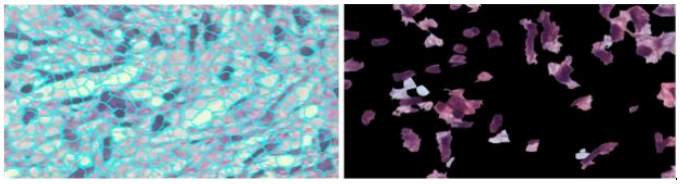


Figure. 7 Segmented breast images: (a) canny edge detection and (b) circular hough transform

3.3 Feature selection

After obtaining the one dimensional feature vectors x, modified reliefF algorithm is applied to select the optimal or relevant feature vectors for better classification [23]. Generally, reliefF algorithm is an extension of relief algorithm, where the conventional algorithm can able to deal with numerical and nominal attributes. But it is ineffective in unstructured or incomplete data and also it is limited to binary class issues. The reliefF algorithm resolves the aforementioned problems and effectively deals with noisy and incomplete data. As similar to relief algorithm, the reliefF randomly chooses the instances r_i and then search for k-nearest neighbors from the different classes is named as nearest miss M_i instances and the k-nearest neighbors searched from similar classes is named as nearest hit H_i instances. Generally, Manhattan distance is used to identify the nearest miss and hit instances. In modified reliefF algorithm, Chebyshev distance is used instead of Manhattan distance to identify the nearest miss and hit instances. Major benefit of Chebyshev distance is it needs only limited time to decide the distances between the instances. Although, Chebyshev distance uses only limited number of features to represent the data that is enough to attain precise neighbourhood selection and better prediction and also it completely reduces the "curse of dimensionality" problem.

In modified reliefF algorithm, the searched nearest miss M_i and nearest hit H_i instances updates the quality estimation $W[x]$ for all attributes x [24], as indicated in the Eqs. (2) - (4).

$$W[x] = \frac{\bar{M}+\bar{H}}{q} \tag{2}$$

Where,

$$\bar{H} = -\sum_{i=1}^{k} D\,(x, r_i\,, H_i)/k \tag{3}$$

$$\bar{M} = \sum_{C \neq cl(r_i)} \left[\left(\frac{P(C)}{1-P(cl(r_i))}\right) \sum_{i=1}^{k} D\,(x, r_i\,, M_i(C))\right]/k \tag{4}$$

Where, $q = 25$ is represented as user defined parameter, D is represented as Manhattan distance between the selected instances r_i, $C = 3$ is indicated as total classes (Anaplasmosis, Babesiosis, and Theileriosis), $cl(r_i)$ is stated as class of i^{th} sample, and $P(C)$ is represented as previous class. After applying reliefF algorithm, actual features x is 3072 and the selected features $W[x]$ is 922. Finally, the selected features $W[x]$ are fed to DNN classifier.

3.4 Classification

After selecting the optimal feature vectors $W[x]$, histopathological image classification is performed by utilizing stacked auto-encoder. It is an unsupervised deep learning algorithm, where the number of input nodes are lower than the number of hidden nodes. The number of output nodes in auto-encoder is equal to the number of input nodes. During pathology image classification, the possibilities of missing value is low in stacked auto-encoder. Initially, it assigns a classification score $f(W[x])$ for the optimal features during prediction time. The function f includes a sequence of layers for computation that is mathematically defined in Eq. (5).

$$Z_{ij} = I_i P_{ij};\ Z_j = \sum_{ij} Z_{ij} + h_j;\ O_j = g(Z_j) \quad (5)$$

Where, I_i is represented as input layer, P_{ij} is indicated as model parameter, O_j is indicated as output layer, h_j is denoted as hidden layer and $g(Z_j)$ is stated as a mapping or pooling function. The layer wise relevance propagation in auto-encoder decomposes $f(W[x])$ into relevance attribute l_i that plays a vital role in classification decision, which is mathematically defined in Eq. (6).

$$f(W[x]) = \sum_i l_i \quad (6)$$

where $l_i = \sum_j \frac{zij}{\sum_i zij}$

If $l_i < 0$, it is a neutral or negative evidence, and if $l_i > 0$, it is a positive evidence that supports classification decision. In auto-encoder, the hidden layers are trained on the input data for learning the primary features. All the weight and bias parameters are learned during the pre-training process to reduce the cost function, as mathematically defined in Eq. (7).

$$cost = \frac{1}{2n}\sum_{i=1}^{n}\left(\hat{I}_i - I_i\right)^2 + \beta \sum_{j=1}^{m} KL\left(p|\widehat{p_j}\right) + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\theta_{ij}^2 \quad (7)$$

Where, p is indicated as sparsity parameter, β is stated as weight of sparsity penalty, θ is indicated as weight of hidden layers, λ is indicated as weight delay, KL is represented as Kullback-Leibler divergence function, $\widehat{p_j}$ is represented as probability of firing activity, n is represented as the number of input nodes and m is indicated as hidden nodes. The parameter settings of auto-encoder is given as follows; input layer is 1, output layer is 1, hidden layer is 125 and 250, and learning rate is 0.1. Generally, the deep learning techniques like stacked auto-encoder requires more number of images to achieve better classification. Here, the experiment is carried out with and without augmentation, because the collected database contains minimum number of images.

4. Experimental results

In this research, the proposed modified reliefF-DNN model is simulated using MATLAB (2018a) environment with the system requirements; **RAM:** 16 GB, processor: Intel core i7, and **Operating System**: windows 10 (64 bit). In this scenario, the performance of modified reliefF-DNN model is analysed by means of sensitivity, specificity, accuracy, balanced accuracy, and f-score on real time and BreaKHis dataset. In histopathological medical diagnosis, specificity is defined as the test to correctly identify the regions without disease (true negative rate). Sensitivity is defined as the test to correctly identify the regions with disease (true positive rate). Further, accuracy is the most important performance measure that utilized in medical diagnosis, where it is the ratio of correctly predicted observations from the total observations. Specificity, sensitivity, and accuracy are mathematically defined in the Eqs. (8)-(10).

$$Specificity = \frac{TN}{FP+TN} \times 100 \quad (8)$$

$$Sensitivity = \frac{TP}{FN+TP} \times 100 \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (10)$$

F-score is determined as the harmonic mean of model's recall and precision, and the balanced accuracy is defined as the harmonic mean of model's sensitivity and specificity. The mathematical expressions of f-score and balanced accuracy are defined in the Eqs. (11) and (12).

$$F - score = \frac{2TP}{2TP+FP+FN} \times 100 \quad (11)$$

Table 1. Performance analysis of modified reliefF-DNN model without augmentation in light of sensitivity, specificity and accuracy

Without Augmentation				
Feature selection	**Classifier**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Without feature selection	MSVM	90.80	92.50	94.45
	Random forest	83.47	88.20	86.20
	KNN	89.67	95.50	82.80
	DNN	92.53	96.13	91
Mutual information	MSVM	88.93	88.20	96.20
	Random forest	85.60	90.90	89.20
	KNN	84.40	97.20	77.40
	DNN	92.27	98.40	97.40
Correlation based feature selection	MSVM	96.20	96.40	98.25
	Random forest	84.53	88.80	89.60
	KNN	93	95.60	93.80
	DNN	95.73	97.20	97
Infinite	MSVM	63.63	67.95	78.80
	Random forest	66.10	74.75	74.30
	KNN	56.67	77.90	54.60
	DNN	64.77	80.87	78.86
ReliefF	MSVM	93.47	97.80	92.01
	Random forest	87.87	90.80	91.80
	KNN	91.33	96.70	98.60
	DNN	96.73	98.80	98.80
Modified ReliefF	MSVM	94.49	98.50	92.74
	Random forest	89.90	90.88	92.91
	KNN	91.87	96.98	98.87
	DNN	**97.90**	**98.98**	**98.92**

Table 2. Performance analysis of modified reliefF-DNN model without augmentation by means of balanced accuracy and f-score

Without Augmentation			
Feature selection	**Classifier**	**Balanced accuracy (%)**	**F-score (%)**
Without feature selection	MSVM	96.25	93.43
	Random forest	87.20	82.33
	KNN	89.15	86.12
	DNN	92.60	90.68
Mutual Information	MSVM	97.70	96.81
	Random forest	90.05	86
	KNN	87.30	84.30
	DNN	98.80	98.47
Correlation based feature selection	MSVM	98.20	96.89
	Random forest	89.20	84.89
	KNN	94.70	92.62
	DNN	98.10	97.67
Infinite	MSVM	73.38	64.65
	Random forest	74.53	66.15
	KNN	66.25	54.48
	DNN	74.23	67.25
ReliefF	MSVM	93.90	97.13
	Random forest	91.30	87.52
	KNN	91.65	96.23
	DNN	99	98.60
Modified ReliefF	MSVM	94	97.18
	Random forest	91.80	87.73
	KNN	92.35	96.80
	DNN	**99.08**	**98.80**

$$Balanced\ accuracy = \frac{Sensitivity+Specificity}{2} \times 100 \quad (12)$$

Where, True Positive is denoted as TP, False Positive is indicated as FP, True Negative is denoted as TN, and False Negative is represented as FN.

4.1 Analysis on haemoprotozoan disease

In this section, the performance of modified reliefF-DNN model is analysed without augmentation on a real time database. Here, the performance analysis is carried-out with different feature selection techniques (mutual information, correlation based feature selection, infinite algorithm and reliefF algorithm) and classification techniques (Multi Support Vector Machine (MSVM), random forest, K-Nearest Neighbor (KNN) and DNN). The undertaken database contains 98 pathology images (11 anaplasmosis images, 60 babesiosis images, and 27 theileriosis images) in that 80% of the images are used for training and 20% of the images are used for testing. By inspecting Table 1, the performance analysis is done with different feature selection and classification techniques by means of accuracy, sensitivity and specificity. Compared to other combinations, modified reliefF-DNN model achieved maximum accuracy of 97.90%, sensitivity of 98.98%, and specificity of 98.92%.

In Table 2, the performance evaluation is done without augmentation by means of balanced accuracy and f-score. By investigating Table 2, the modified reliefF with DNN model achieved a maximum balanced accuracy of 99.08% and f-score of 98.80%. The deep learning algorithm eliminates the need for data labeling and has the ability to deliver high quality results compared to other machine learning algorithms.

In Table 3, the performance evaluation is carried out with augmentation by means of sensitivity, accuracy and specificity. By inspecting Table 3, the undertaken models attained better classification performance with augmentation compared to with-out augmentation. As similar to the Tables 1 and 2, the combination (modified reliefF-DNN) achieved a significant performance in Haemoprotozoan disease detection related to other combinations (dissimilar feature selection and classification techniques). In this section, modified reliefF-DNN model attained maximum classification accuracy of 97.6%, sensitivity of 98.92% and specificity of 98.70% in Haemoprotozoan disease detection. Modified ReliefF algorithm effectively detects the statistical

Table 3. Performance analysis of modified reliefF-DNN model with augmentation by means of sensitivity, accuracy, and specificity

With Augmentation				
Feature selection	**Classifier**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Without feature selection	MSVM	90.73	91.15	94.50
	Random forest	90.30	92.85	93.10
	KNN	79.57	89.90	68.90
	DNN	93	96.33	86.40
Mutual information	MSVM	89	91.45	92
	Random forest	87.97	90.60	92.20
	KNN	74.37	91.90	59.70
	DNN	93.63	97.82	83.10
Correlation based feature selection	MSVM	93.77	95.25	95.30
	Random forest	91.10	92.80	94.80
	KNN	84.97	92.85	76.20
	DNN	95.60	97.53	91.60
Infinite	MSVM	90.93	91.80	97.20
	Random forest	86.13	87.90	91.80
	KNN	86.60	93.60	84.20
	DNN	92.40	96.30	94.20
ReliefF	MSVM	92	90	93.93
	Random forest	92.02	91.29	93.22
	KNN	91	90	88
	DNN	95.02	90	92.09
Modified ReliefF	MSVM	94.27	92.90	97.50
	Random forest	92.13	94.10	95.10
	KNN	92.90	95.85	90
	DNN	**97.60**	**98.92**	**98.70**

Table 4. Performance analysis of modified reliefF-DNN model with augmentation in terms of balanced accuracy and f-score

With Augmentation			
Feature selection	**Classifier**	**Balanced accuracy (%)**	**F-score (%)**
Without feature selection	MSVM	92.83	89.19
	Random forest	92.98	89.84
	KNN	79.40	72.69
	DNN	89.40	87.45
Mutual Information	MSVM	91.73	88.03
	Random forest	91.40	87.53
	KNN	75.80	67.63
	DNN	86.53	85.21
Correlation based feature selection	MSVM	95.28	93.19
	Random forest	93.80	90.74
	KNN	84.53	79.81
	DNN	92.60	91.47
Infinite	MSVM	95.90	92.70
	Random forest	89.85	85.10
	KNN	88.90	85.21
	DNN	95.25	93.36
ReliefF	MSVM	96.20	93.36
	Random forest	94.60	92.07
	KNN	92.93	90.77
	DNN	96.82	97.26
Modified ReliefF	MSVM	97	94.96
	Random forest	95.90	94.77
	KNN	93	94
	DNN	**97.80**	**98.16**

interactions from the histopathological images, so it can able to select the relevant feature subsets from the higher dimensional extracted features. This process completely reduces the "curse of dimensionality" problem that results in better classification.

In Table 4, the modified reliefF-DNN model with augmentation achieved maximum balanced accuracy of 97.80% and f-score value of 98.16%. In this research study, modified reliefF algorithm plays a vital role in Haemoprotozoan disease detection, where the effect of modified reliefF feature selection is given in the Tables 1, 2, 3, and 4. The proposed modified reliefF-DNN model includes two major benefits like cost effective related to other machine learning algorithms, and assists clinicians in early diagnosis of Haemoprotozoan disease.

4.2 Analysis on breast cancer

In this section, the classification performance of the proposed modified reliefF-DNN model is validated with dissimilar classification approaches such as MSVM, random forest and KNN, and also the effectiveness of the proposed modified reliefF-DNN model is analysed with and without augmentation. In Table 5, the performance validation of the proposed modified reliefF-DNN model is done in light of accuracy, sensitivity, and specificity. From the

Table 5. Performance analysis of modified reliefF-DNN model with and without augmentation in terms of sensitivity, specificity and accuracy

Cell separation	**Classifier**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
Without Augmentation	MSVM	79.09	88.43	83.04
	Random forest	76.63	80.90	79
	KNN	82.98	84.90	83
	DNN	**91**	**90.52**	**94**
With Augmentation	MSVM	78.90	89.92	84
	Random forest	80.90	84.22	82.18
	KNN	85.12	88.36	86.9
	DNN	**92.90**	**94.39**	**95.94**

Table 6. Comparative study of proposed and existing work

Methodology	Classification accuracy (%)
QLA[18]	86.6
Modified reliefF-DNN	**95.94**

inspection, the classification accuracy of proposed modified reliefF-DNN model is 95.94%, which is higher compared to other classifiers. In this scenario, the proposed model almost showed 1.93% to 13.81% improvement in accuracy compared to other classifiers. In addition, the sensitivity and specificity of the proposed modified reliefF-DNN model are superior related to other comparative classifiers.

Table 6 represents the comparative study of proposed and existing works. P. Alirezazadeh, B. Hejrati, A. Monsef-Esfahani, and A. Fathi, [18] developed a system for histopathological breast cancer image classification. Initially, correlation metric was used to reduce the mismatch between the test and trained feature values. Then, an adaptation method was utilized for enhancing the detection rate of benign and malignant pathology images. Finally, QLA classifier was used to classify malignant and benign images. In this developed work, an extensive experiment was performed on BreaKHis database, and the developed system achieved 86.6% of classification accuracy. Compared to this existing work, the proposed modified reliefF-DNN model achieved better performance in breast cancer detection.

5. Conclusion

In this research, modified reliefF-DNN model is proposed for early detection of communicable and non-communicable diseases like Haemoprotozoan disease and breast cancer. The modified reliefF-DNN model includes three major phases; segmentation, feature selection, and classification for disease detection. In the segmentation phase, edge boost curve transform is used for nuclei and non-nuclei cell segmentation. Next, modified reliefF algorithm and DNN classifier are used to select the optimal feature vectors and to classify the segmented images. Related to the comparative models like MSVM, KNN, and random forest, modified reliefF-DNN model achieved a maximum sensitivity of 98.92%, specificity of 98.70%, f-score of 97.26%, classification accuracy of 97.60%, and balanced accuracy of 96.82% in Haemoprotozoan disease detection. Similarly, modified reliefF-DNN model achieved maximum sensitivity of 92.90%, specificity of 94.39% and accuracy of 95.94% in breast cancer detection. In the future work, a hybrid clustering algorithm is included in modified reliefF-DNN model to improve the performance of histopathological cell segmentation and classification in both communicable and non-communicable diseases.

I	Collected pathology images
I_{norm}	Normalized images
q	User defined parameter
D	Manhattan distance
C	Total classes
$P(C)$	Prior class
P_{ij}	Model parameter
$g(Z_j)$	Mapping or pooling function
p	Sparsity parameter
β	Weight of sparsity penalty
θ	Weight of hidden layers
λ	Weight delay
KL	Kullback-Leibler divergence function
$\widehat{p_j}$	Probability of firing activity

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] T. Wan, J. Cao, J. Chen, and Z. Qin, "Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features", *Neurocomputing*, Vol. 229, pp. 34-44, 2017.

[2] B. Gecer, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks", *Pattern recognition*, Vol. 84, pp .345-356, 2018.

[3] M. A. Bary, M. Z. Ali, S. Chowdhury, A. Mannan, M. N. E. Azam, M. M. Moula, Z. A. Bhuiyan, M. T. W. Shaon, and M. A. Hossain, "Prevalence and molecular identification of haemoprotozoan diseases of cattle in Bangladesh", *Advances in Animal and*

Veterinary Sciences, Vol. 6, No. 4, pp. 176-182, 2018.

[4] G. Patra, S. Ghosh, D. Mohanta, S. Kumar Borthakur, P. Behera, S. Chakraborty, A. Debbarma, and S. Mahata, "Prevalence of haemoprotozoa in goat population of West Bengal, India", *Biological Rhythm Research*, Vol. 50, No. 6, pp. 866-875, 2019.

[5] K. Jayalakshmi, M. Sasikala, M. Veeraselvam, M. Venkatesan, S. Yogeshpriya, P. K. Ramkumar, P. Selvaraj, and M. K. Vijayasarathi, "Prevalence of haemoprotozoan diseases in cattle of Cauvery delta region of Tamil Nadu", *Journal of Parasitic Diseases*, Vol. 43, No. 2, pp. 308-312, 2019.

[6] D. R. Prameela, V. V. Rao, V. Chengalvarayulu, P. Venkateswara, T. V. Rao, and A. Karthik, "Prevalence of Haemoprotozoan infections in Chittoor District of Andhra Pradesh", *Journal of entomology and zoology studies*, 2020.

[7] S. Ghosh, G. Patra, S. Kumar Borthakur, P. Behera, T. C. Tolenkhomba, A. Deka, R. Kumar Khare, and P. Biswas, "Prevalence of haemoprotozoa in cattle of Mizoram, India", *Biological Rhythm Research*, Vol. 51, No. 1, pp. 76-87, 2020.

[8] S. B. Swami, J. S. Patel, S. H. Talekar, B. Kumar, V. L. Parmar, A. K. Bilwal, and B. R. Patel, "Prevalence of Haemoprotozoan Infection in Gir Cattle in and around Junagadh, Gujarat", *The Indian Journal of Veterinary Sciences and Biotechnology*, Vol. 15, No. 2, pp. 46-48, 2019.

[9] K. J. Ananda, and J. Adeppa, "Prevalence of Haemoprotozoan infections in bovines of Shimoga region of Karnataka state", *Journal of Parasitic Diseases*, Vol. 40, No. 3, pp. 890-892, 2016.

[10] P. Mohapatra, B. Panda, and S. Swain, "Enhancing histopathological breast cancer image classification using deep learning", *Int. J Innov. Technol. Explor. Eng*., Vol. 8, pp. 2024-2032, 2019.

[11] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact CNNs", *BMC Medical Informatics and Decision Making*, Vol. 19, No. 1, pp. 198, 2019.

[12] Y. Xu, Z. Jia, L. B. Wang, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features", *BMC bioinformatics*, Vol. 18, No. 1, pp. 1-17, 2017.

[13] A. Chakravarty and J. Sivaswamy, "RACE-net: a recurrent neural network for biomedical image segmentation", *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 3, pp. 1151-1162, 2018.

[14] X. Li, Y. Wang, Q. Tang, Z. Fan, and J. Yu, "Dual U-Net for the Segmentation of Overlapping Glioma Nuclei", *IEEE Access*, Vol. 7, pp. 84040-84052, 2019.

[15] A. Albayrak and G. Bilgin, "Automatic cell segmentation in histopathological images via two-staged superpixel-based algorithms", *Medical & Biological Engineering & Computing*, Vol. 57, No. 3, pp. 653-665, 2019.

[16] H. Jiang, S. Li, W. Liu, H. Zheng, J. Liu, and Y. Zhang, "Geometry-Aware Cell Detection with Deep Learning", *Msystems*, Vol. 5, No. 1, 2020.

[17] H. Li, X. Zhao, A. Su, H. Zhang, J. Liu, and G. Gu, "Color space transformation and multi-class weighted loss for adhesive white blood cell segmentation", *IEEE Access*, Vol. 8, pp. 24808-24818, 2020.

[18] P. Alirezazadeh, B. Hejrati, A. Monsef-Esfahani, and A. Fathi, "Representation learning-based unsupervised domain adaptation for classification of breast cancer histopathology images", *Biocybernetics and Biomedical Engineering*, Vol. 38, No. 3, pp. 671-683, 2018.

[19] K. M. Koo, and E. Y. Cha, "Image recognition performance enhancements using image normalization", *Human-centric Computing and Information Sciences*, Vol. 7, No. 1, pp. 1-11, 2017.

[20] R. Biswas and J. Sil, "An improved canny edge detection algorithm based on type-2 fuzzy sets", *Procedia Technology*, Vol. 4, pp. 820-824, 2012.

[21] G. Deng and L. W. Cahill, "An adaptive Gaussian filter for noise reduction and edge detection", In: *Proc. of IEEE Conf. Record Nuclear Science Symposium and Medical Imaging Conf.*, pp. 1615-1619, 1993.

[22] S. J. K. Pedersen, "Circular hough transform", *Aalborg University, Vision, Graphics, and Interactive Systems*, Vol. 123, No. 6, 2007.

[23] I. Sangaiah and A. V. A. Kumar, "Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction", *Cluster Computing*, Vol. 22, No. 3, pp. 6899-6906, 2019.

[24] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review", *Journal of Biomedical Informatics*, Vol. 85, pp. 189-203, 2018.