doi:10.20944/preprints202111.0246.v1

Article

Cancer Diagnosis of Microscopic Biopsy Images using Social Spider Optimization tuned Neural Network

Prasanalakshmi Balaji[1]*, Kumarappan Chidambaram[2]

[1] Department of Computer Science, Center for Artificial Intelligence, King Khalid University, Abha 62529, Saudi Arabia. Email: prengaraj@kku.edu.sa
[2] Department of Pharmacology, School of Pharmacy, King Khalid University, Abha 62529, Saudi Arabia. Email: kumarappan@kku.edu.sa
Correspondence: Prasanalakshmi Balaji*(prengaraj@kku.edu.sa)

Abstract: One of the most dangerous diseases that threaten people is Cancer. Cancer if diagnosed in earlier stages can be eradicated with its life threatening consequences. In addition, accuracy in prediction plays a major role. Hence, developing a reliable model that contributes much towards the medical community in early diagnosis of Biopsy images with perfect accuracy come to the scenario. The article aims towards development of better predictive models using multi-variate data and high-resolution diagnostic tools in clinical cancer research. This paper proposes the social spider optimization (SSO) algorithm tuned neural network to classify microscopic biopsy images of cancer. The significance of the proposed model relies on the effective tuning of the weights of the NN classifier by the SSO algorithm. The performance of the proposed strategy is analyzed with the performance metrics, such as accuracy, sensitivity, specificity, and MCC measures, and are attained to be 95.9181%, 94.2515%, 97.125%, and 97.68% respectively, which shows the effectiveness of the proposed method in effective cancer disease diagnosis.

Keywords: biopsy, cancer diagnosis, predictive models, neural network, optimization.

1. Introduction

In 2019, the cancer burden is estimated 1.7 million new cases and 0.6 million deaths in the United States [3]. As the incidence rate of cancer and its mortality have risen sharply, prolonging survival and reducing local recurrence depends on laparoscopy surgery, robotic surgery, tumor adjuvant therapy, and other new technologies [4]. There exist multiple options in treating cancer [5]. This led to the increase in the effectiveness of cancer treatment [6]. However, in spite of new technologies and developments that can be seen in the research areas towards cancer prognosis, prediction and diagnosis, due to uncertainties in diagnostic prediction expected curative results could not be obtained. Thus, patient-specific optimal treatment, like precision medicine based on preliminary diagnosis could be adopted if an accurate prognosis could be made. Here is where the role of prediction accuracy plays an important role. Perfect prediction accuracy could assist the doctors in planning treatments that could reduce the stress in patients. Preliminary clinical observations when combined with traditional TNM staging approach based on tumor size (T), spread to nearby lymph nodes (N) and spread to other parts of body (M) , but again the erroneous predictions of prognosis continue to create a bottleneck for the clinicians [7]. Even though many researches are carried out towards development of Machine learning based models, improvement in prognosis accuracy stays a critical challenge to the clinicians. Technical advancements in statistics, Machine learning and its collaboration with Medical imaging based on computer programs and software has brought improvements using multi-factor analysis, regression analysis like linear regression, polynomial regression, Cox regression . The empirical predictions were overcome by the recent researches carried out on prognosis and diagnosis using machine learning and deep learning models. These methods currently play an important role in improving the accuracy of cancer susceptibility, recurrence, and survival predictions [3]. The digital pathology field has grown dramatically over recent years, largely due to technological advancements in image processing and machine learning algorithms, and increases in computational power. As part of this field, many methods have been proposed for automatic histopathological image analysis and classification [2].

This paper introduces an Artificial intelligence (AI) model to diagnose cancer using the microscopic biopsy images. Initially, the input images taken under consideration containing a balanced set of classes are fed as input to the neural network. The neural network taken under consideration is a basic neural network. The two classes taken under consideration for classification of biopsy images are Benign and Malignant. The images undergo various layers of abstraction to get themselves classified. The input images are pre-processed through Region of Interest (ROI) segmentation method. The pre-processed image of RGB form is converted into binary image to make it subject to watershed segmentation, where the nucleus based morphological features are extracted. Then the features, such as area, perimeter, diameter, and compactness from the nucleus based morphological features are extracted. The extracted features develop as a feature vector that acts as the input to the proposed NN classifier, the weights of which are optimally tuned using the SSO algorithm. This assists in enhancing the performance of the NN classifier in diagnosing the input image as benign and malignant.

The contribution and the key topics discussed in the study are;

The proposed model designs an automated model to diagnose the microscopic biopsy images as benign and malignant.

The optimal tuning of the weights of the NN classifier by the SSO algorithm plays a vital role in the enhancement of the performance of the proposed classification model.

The proposed model is analyzed with the conventional methods in terms of performance measures, such as accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC) to validate the effectiveness of the proposed strategy.

The rest of the paper is organized as: section 2 presents the survey of the recent strategies of cancer diagnosis with the challenges associated with them. Section 3 describes the proposed NN model in the diagnosis of cancer. Section 4 deliberates the outcomes of the proposed model, and section 5 concludes the paper.

2. Literature survey

This section deliberates the existing methods of cancer diagnosis and the challenges associated with the existing models that act as the motivation for the development of the proposed model.

2.1 Related works

The literature reviews of the existing methods of cancer diagnosis are stated as: Irfan Ullah Khan *et al.* [1] developed a study for the early analysis of cervical cancer with the aid of reduced risk set of feature and three ensemble classification strategies, such as extreme Gradient Boosting (XGBoost), Random Forest (RF) and AdaBoost, along with Firefly algorithm for hyper parameter tuning process. The model achieved poor sensitivity measure, which is considered as the major drawback of the method. Subrata Bhattacharjee *et al.* [2] examined the stained microscopic biopsy images to perform image treatment and extract the important features to be fed as input to the support vector machine (SVM) classifier for the diagnosis of cancer disease. However, the model cannot make exact predictions as errors in one class affects the accuracy in diagnosis. Ghulam Murtaza *et al.* [5] created a consistent and more precise model that used minimum resources with the aid of transfer learning based convolution neural network (CNN) model. However, it is a tedious task to gather more number of images of all types of cancers with proper labels.

2.2 Challenges

The challenges associated with the existing methods of cancer diagnosis are

The DNN based strategies, particularly the CNN based methods has solved the problem of handcrafted extraction of features. However, when this model is trained from scratch, it need a more annotated images and requires very high resources [5].

In breast biopsy, the samples of breast cancer is taken and preserved into microscopic slides for manual evaluation. The microscopic investigation is carried out by expert pathologist, and the final conclusion is made after the agreement of more than two pathologists for enhanced diagnosis. However, it may need increase time for diagnosis and there may be a disagreement of opinion among two pathologists [5].

3. Proposed method of cancer diagnosis

An optimization tuned NN classifier is proposed in this research to classify the cancer as benign or malignant. In the initial step, the microscopic biopsy images of the patients are subjected to pre-processing for the removal artifacts present in the raw image data. In order to isolate the required object from the input image data, watershed segmentation of image is performed after the conversion of RGB image into binary image. The object is then analyzed using the proposed NN model with the provision of the features of the image as the training samples. The features, such as area, perimeter, diameter, and compactness are extracted from the image, with which the feature vector is developed through feature concatenation. The trained NN model using the feature vector developed with the image features is tested with the test data, in such a way to classify the cancer disease. Sample biopsy images are found in Figure 1.

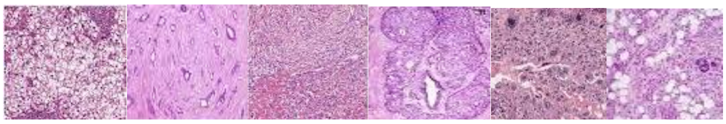


Figure 1 Sample Biospy images

The schematic representation of proposed model is depicted in figure 2.

3.1 Image pre-processing

The pre-processing step is the initial steps of the proposed cancer disease diagnosis module, with the objective of getting rid of the artifacts that are present in the images taken from the microscopic biopsy images. The pre-processing step is necessary in such a way to enhance the prediction accuracy of the proposed NN classifier. The raw microscopic biopsy images are normalized using the pre-processing step through ROI segmentation with the extraction of Stroma, nuclei and lumen in such a way to subject it for the successive steps of the diagnosis process [8].

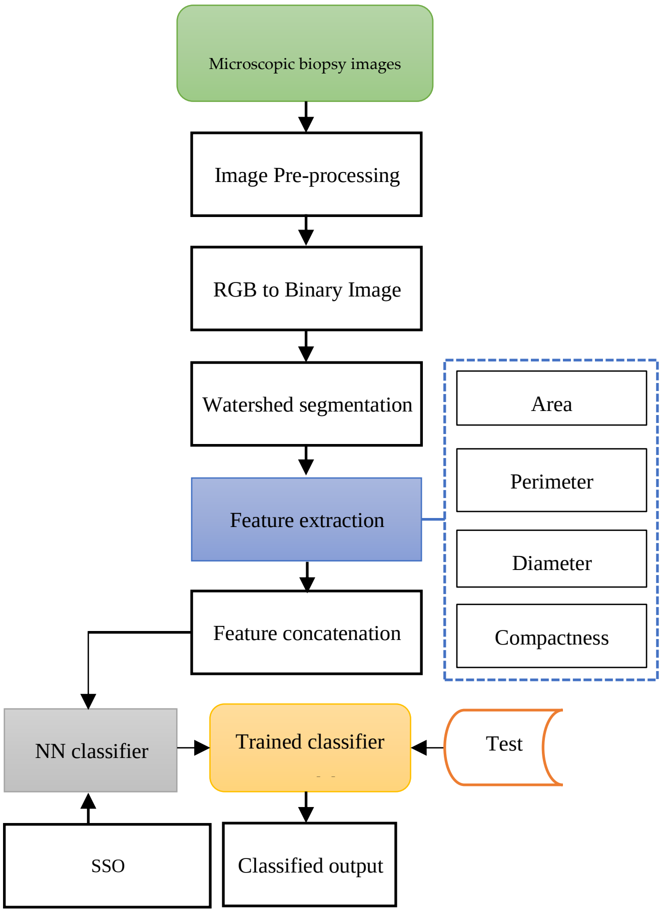


Figure 2. Schematic diagram of proposed cancer disease diagnosis model

3.2 Watershed segmentation

The image in the form of RGB is converted into binary form in such a way to describe the color of the image with more recognizable comparisons, like brightness, color, and vitality [8]. The watershed segmentation is applied to the binary image to execute the process of object segmentation. The watershed segmentation process assists in the extraction of nucleus based morphological features, from which the significant features are extracted.

3.3. Feature extraction and concatenation

The significant features, such as area a_f, perimeter p_f, diameter d_f, and compactness c_f are extracted from the nucleus based morphological features of the input microscopic biopsy images. The extracted features are then concatenated to form the feature vector that acts as the input to the proposed NN classifier of cancer diagnosis. The feature vector thus generated is represented as,

$$F = \{a_f, p_f, d_f c_f\} \qquad (1)$$

Thus, the concatenated feature acts as the input to the proposed SSO-based NN classifier in such a way to execute the cancer diagnosis process.

3.4 Proposed social spider optimization tuned neural network classifier in cancer diagnosis

The NN is a mathematical model designed on the basis of biological neural networks with interconnected gathering of artificial neurons and processes the data by a connection strategy for computation. NNs have come out as an area of unusual opportunity for analysis, advancement and application in the past few years to different types of real world problems. The NNs are the tremendously equivalent computing systems comprising of number of interconnections and simple processors. The principle of operation executed by human is the basic tactic for the development of NNs and as the purpose of neurons in humans. The NN possess the neuron layers in such a way to process the data acting as input. The NN consists of input that is multiplied with the weights to represent the flow of data. The mathematical functions involves in evaluating the weights that in turn outputs the activation function of the neuron. The output of the NN can be adjusted as expected with the optimal tuning of weights using the proposed SSO algorithm. The architecture of the NN is depicted in figure 3.

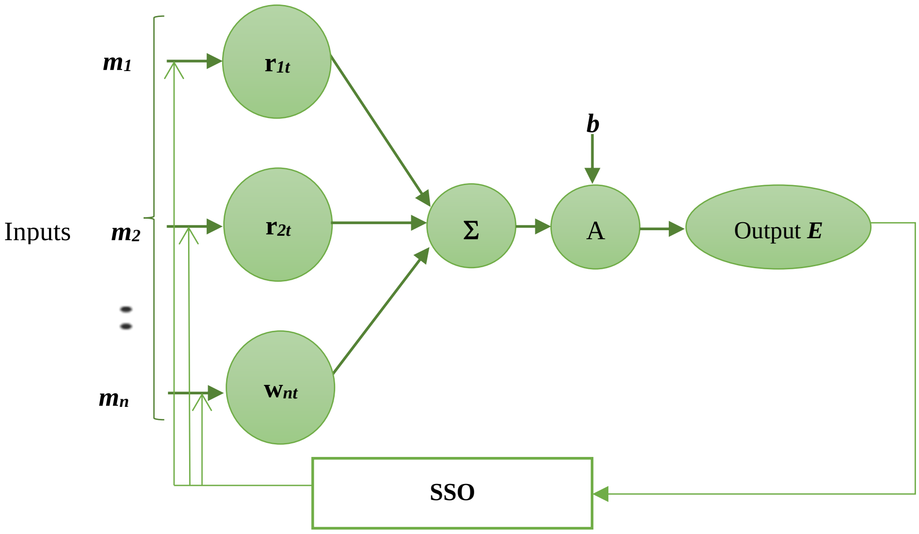


Figure 3. Structure of Proposed network

The output of the ANN can be mathematically expressed as,

$$E = A\left(\sum_{g=1}^{n} r_g m_g\right) + b \qquad (2)$$

Where, A represents the activation function, m_g is the input, and r_g indicates the weight values. The evaluation of position of all possible positions in the proposed system is performed using the NN classifier. The significance of using NN [9] is that the classifier performs the evaluation in an automatic manner without the requirement for human to check the process. The evaluation accuracy of the method relies on proper usage of the proposed NN with noteworthy weight values using the proposed SSO algorithm.

3.5 Social spider optimization algorithm in update of NN weights

The vibration sensitive social characteristic is derived from the spiders, which is one of the most different species among all kinds of creatures and they make in use of variety of techniques for foraging. However, the model of sensing the vibrations is the most effectual method of foraging in social spiders. They are highly sensitive to vibrations and the vibrations assist them in catching the prey. When the observed vibrations of the prey are within a defined frequency, the social spider catches the prey. One of the special features of the social spider is its ability to sense the difference among the vibrations of other social spider and the vibration produced by prey. The vibration generated by other social spiders helps in obtaining a clear view about the web that effectively reduces the loss of information. The social characteristics can be devised as a cooperative movement of the entire social spiders in reaching the position of the prey. The vibrations form the prey and other social spiders are received and analyzed in such a way to find the optimal position of the prey [10].

The search space of the optimization problem is initially framed as the hyper-dimensional web of the social spider. All the positions in the web are considered as a feasible solution, and the web acts as a medium of transmission of vibrations. Each social spider occupies a position on the web, and the fitness of the solution depends on the objective of finding the position of prey. The vibrations contain information of a social spider, which can be used by other social spider on the web. The standard expression of social spider is generalized as,

$$B_v^q = \left(L - B_v^{q-1}\right) \times rand \qquad (3)$$

where, B_v^q is the position of v^{th} spider at q^{th} iteration. The above equation can be remodeled as,

$$B_v^{q+1} = \left(L - B_v^q\right) \times rand \qquad (4)$$

$$B_v^{q+1} = \left(L - 0.5B_v^q - 0.5B_v^q\right) \times rand \qquad 5)$$

$$B_v^{q+1} = \left(L - 0.5B_v^q - 0.5\left[\left(L - B_v^{q-1}\right) \times rand\right]\right) \times rand \qquad (6)$$

$$B_v^{q+1} = rand\left(L - 0.5B_v^q - 0.5L \times rand + 0.5B_v^{q-1} \times rand\right) \qquad (7)$$

$$B_v^{q+1} = rand\left(L\left[1 - 0.5rand\right] - 0.5B_v^q + 0.5B_v^{q-1} \times rand\right) \qquad (8)$$

where, B_v^{q+1} is the position of the social spider in $(q+1)^{th}$ iteration, L is the lower bound of the search space, and $rand$ is the random floating point number ranging between 0 and 1. The above equation is the standard equation of the SSO algorithm that involve in the optimal tuning of the weights of NN classifier. Algorithm 1 deliberates the pseudocode of the proposed SSO optimization algorithm.

Algorithm 1. Pseudo code of SSO algorithm

```
Initialize the population of social spiders
Initialize the target vibration
Initialize the parameters, $rand$ and $L$
Evaluate the fitness measure for all social spiders
For all social spiders,
{
Calculate the vibration intensity
Select the strongest vibration among all
{
    $n_{int} > t\arg et_{int}$
    Store position of foraging spider $v$ as the best solution
}
While $q < q_{max}$
 {
 Update the position of foraging spider as per equation (8)
 }
End For
Update the parameters, $rand$ and $L$
Evaluate fitness for all social spiders
    Sort the positions as per fitness measure (accuracy)
$q = q + 1$
  }
```

```
End For
Return $B_v^{q+1}$
```

4. Results and discussions:

The outcomes of the proposed cancer disease diagnosis module and the comparative study for proving the performance of the proposed NN classifier in disease cancer disease diagnosis is discussed in this section.

4.1 Experimental setup

The experimentation is done in MATLAB tool installed in Windows 10 OS and 64-bit operating system with 16GB RAM. The dataset considered includes biopsy images of 58 Hematoxylin and Eosin (H&E) strain microscopic images of the breast tissues(UCSB center for bio-image Informatics)- 31 benign and 27 malignant images. The images are augmented to increase the count of samples. Augmentations include in general standard techniques include rotation, shearing, zooming, cropping, flipping, and fill.

4.3 Evaluation metrics:

The effectiveness of the proposed cancer diagnosis method is tested using the following measures.

a) Accuracy: The accuracy of the system is evaluated as the rate of closeness to the obtained quantity to the real quantity. It is expressed mathematically as,

$$Accuracy = \frac{True\ \ positive + True\,negative}{real\ \ positive + real\ \ negative} \tag{9}$$

b) Sensitivity: The sensitivity is termed as the proportion of true positive values to the number of real positive cases. It is expressed mathematically as,

$$Senitivity = \left(\frac{True\ \ positive}{no\ of\ real\ positivecases} \right) \tag{10}$$

c) Specificity: It is characterized as proportion of true negatives to the count of real negative cases and formulated as,

$$Specificity = \left(\frac{True\ negative}{no\ of\ real\,negative\,cases} \right) \tag{11}$$

d) Matthew's correlation coefficient: The measure of MCC can be calculated using the expression,

$$MCC = \frac{(True\ \ positive \times TrueN\ \ negative) - (False\ \ psoitive \times False\ negative)}{\sqrt{(True\ \ positive + False\ \ positive) \times (True\ \ positive + False\ negative) \times (True\ negative + False\ \ positive) \times (True\ negative + False\ negative)}} \tag{12}$$

4.4 Comparative analysis of methods involved in diagnosis of cancer

The methods considered for comparison are the genetic algorithm (GA) [11], Particle swarm optimization (PSO) [12], Firefly algorithm (FF) [13].

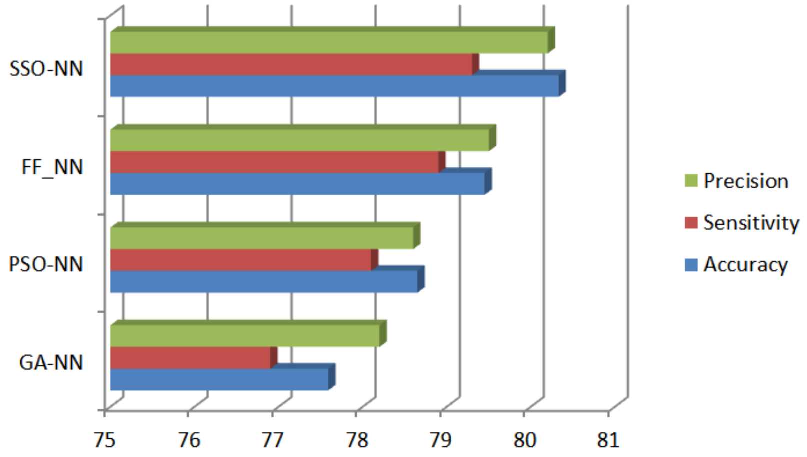


Figure 4. Performance Comparison of models

The GA-NN, PSO-NN, FF-NN and the proposed method has shown a considerable increase in the Accuracy in the specified order.The results does not show much variation in the accuracy, since the accuracy of medical images and their thresholds are very important.

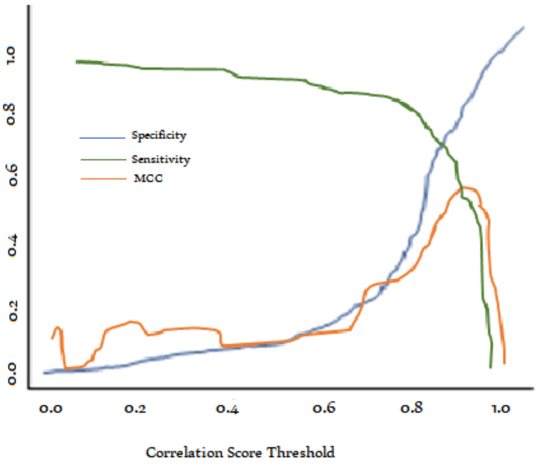
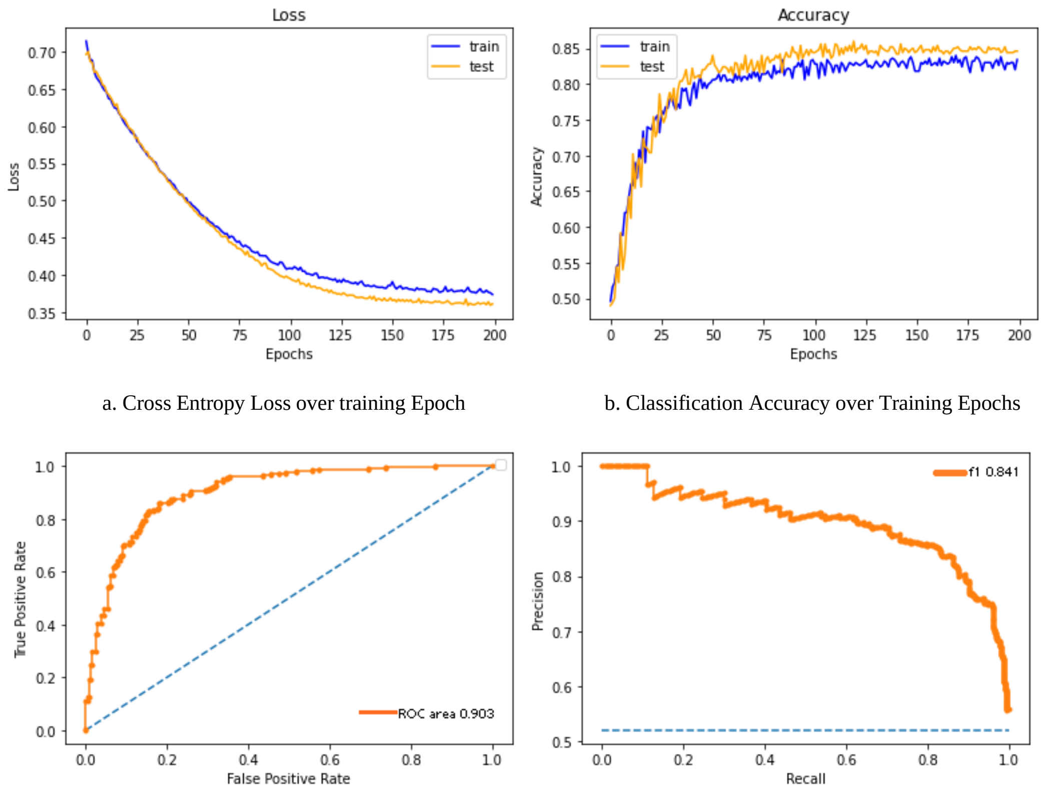


Figure 5. MCC, Sensitivity, Specificity vs Threshold

In Figure 5 we show the MCC vs. threshold on the same plot as sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)) for the selected data. We see that the peak of the MCC occurs where the specificity is somewhat greater than the sensitivity. A user might move the classification threshold somewhat lower or higher depending on whether it is more important to retrieve all or practically all true positives, or whether it is rather more important to ensure that the positive results are not contaminated with false positives.



a. Cross Entropy Loss over training Epoch

b. Classification Accuracy over Training Epochs

c. ROC plot

d. Precision-Recall Plot

Figure 6. Performance analysis of Proposed method

5. Conclusion

This paper proposes an Artificial intelligence (AI) based classification model for the diagnosis of cancer using microscopic biopsy image. The input image is initially subjected to pre-processing for the removal of artefacts present in the image. The pre-processed image is subjected to watershed segmentation for the extraction of nucleus based morphological features. Then the features that include area, perimeter, diameter, and compactness from the nucleus based morphological features are extracted. The extracted features acts as the input to the proposed NN classifier that involve in the diagnosis of cancer disease. The weights of the NN classifier are optimally tuned using the SSO algorithm that enhances the performance of the proposed model. The effectiveness of the proposed strategy is analyzed with the performance indices, namely accuracy, sensitivity, specificity, and MCC measures. The measures of accuracy, sensitivity and MCC are 95.9181%, 94.2515% and 97.68% respectively, which shows the effectiveness of the proposed method in effective disease classification. In future, the weights of the NN classifier will be optimally tuned by hybrid optimization algorithms to further increase the accuracy of the proposed model of cancer disease diagnosis.

Author Contributions

Conceptualization, Prasanalakshmi B; Data curation, Prasanalakshmi B; Formal analysis, Prasanalakshmi B; Funding acquisition, Kumarappan Chidambaram; Investigation, Prasanalakshmi B; Methodology, Prasanalakshmi B; Project administration, Kumarappan Chidambaram; Resources, Kumarappan Chidambaram; Software, Prasanalakshmi B; Supervision, Prasanalakshmi B.

Funding:

The authors are thankful to King Khalid University (KKU), Saudi Arabia for providing financial support through the Small Research Program (Grant Approval Number: 275/1442)

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khan, Irfan Ullah, Nida Aslam, Rawan Alshehri, Seham Alzahrani, Manal Alghamdi, Atheer Almalki, and Maryam Balabeed. "Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization." Scientific Programming 2021 (2021).
2. Bhattacharjee, Subrata, Hyeon-Gyun Park, Cho-Hee Kim, Deekshitha Prakash, Nuwan Madusanka, Jae-Hong So, Nam-Hoon Cho, and Heung-Kook Choi. "Quantitative analysis of benign and malignant tumors in histopathology: Predicting prostate cancer grading using SVM." Applied Sciences 9, no. 15 (2019): 2969.
3. Alix-Panabières, Catherine, and Klaus Pantel. "Circulating tumor cells: liquid biopsy of cancer." Clinical chemistry 59, no. 1 (2013): 110-118.
4. Fowler, Jackson E., Steven A. Bigler, Derek miles, and Denis a. Yalkut. "Predictors of first repeat biopsy cancer detection with suspected local stage prostate cancer." The Journal of urology 163, no. 3 (2000): 813-818.
5. Murtaza, Ghulam, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Ghulam Raza, and Nor Aniza Azmi. "Breast cancer classification using digital biopsy histopathology images through transfer learning." In Journal of Physics: Conference Series, vol. 1339, no. 1, p. 012035. IOP Publishing, 2019.
6. Kasivisvanathan, Veeru, Antti S. Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A. Mynderse, Markku H. Vaarala, Alberto Briganti et al. "MRI-targeted or standard biopsy for prostate-cancer diagnosis." New England Journal of Medicine 378, no. 19 (2018): 1767-1777.
7. Iizuka, O., Kanavati, F., Kato, K. *et al*. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci Rep* **10,** 1504 (2020). https://doi.org/10.1038/s41598-020-58467-9
8. Eichler, Klaus, Susanne Hempel, Jennifer Wilby, Lindsey Myers, Lucas M. Bachmann, and Jos Kleijnen. "Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: a systematic review." The Journal of urology 175, no. 5 (2006): 1605-1612.
9. Stephan, Carsten, Henning Cammann, Axel Semjonow, Eleftherios P. Diamandis, Leon FA Wymenga, Michael Lein, Pranav Sinha, Stefan A. Loening, and Klaus Jung. "Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies." Clinical chemistry 48, no. 8 (2002): 1279-1287.
10. James, J. Q., and Victor OK Li. "A social spider algorithm for global optimization." Applied soft computing 30 (2015): 614-627.
11. Mirjalili, Seyedali. "Genetic algorithm." In Evolutionary algorithms and neural networks, pp. 43-55. Springer, Cham, 2019.
12. Poli, Riccardo, James Kennedy, and Tim Blackwell. "Particle swarm optimization." Swarm intelligence 1, no. 1 (2007): 33-57.
13. Arora, Sankalap, and Satvir Singh. "The firefly optimization algorithm: convergence analysis and parameter selection." International Journal of Computer Applications 69, no. 3 (2013).