Malignancy Identification From Cytology Images Using Deep Optimal Features

A thesis submitted in partial fulfilment of the requirement for the

Degree of Master of Computer Application

Of

Jadavpur University

By

OINDRILA GHOSH

Registration Number: 149902 of 2019-2020

Examination Roll Number: MCA226041

Under the Guidance of

Prof. Nibaran Das

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

June, 2022

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled "**Malignancy Identification From Cytology Images Using Deep Optimal Features**" has been satisfactorily completed by **OINDRILA GHOSH** (University Registration No.:149902 of 2019-2020, Exam Roll No.:MCA226041). It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of **Master of Computer Application in Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata 700032**.

Prof. Nibaran Das (Thesis Supervisor)
Professor
Department of Computer Science and Engineering,
Jadavpur University, Kolkata-700032.

Countersigned

Prof. Anupam Sinha
Head, Computer Science and Engineering,
Jadavpur University, Kolkata-700032.

Prof. Chandan Mazumdar
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata-700032.

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

<u>CERTIFICATE OF APPROVAL</u>

This is to certify that the thesis entitled "**Malignancy Identification From Cytology Images Using Deep Optimal Features**" is a bonafide record of work carried out by OINDRILA GHOSH in partial fulfilment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of 2019 to 2022. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

Signature of Examiner
Date:

Signature of Supervisor
Date:

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis entitled "**Malignancy Identification From Cytology Images Using Deep Optimal Features**" contains literature survey and original research work by the undersigned candidate, as part of her Degree of Master of Computer Application.
All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.
I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: OINDRILA GHOSH
University Registration No. : 149902 of 2019-2020
Examination Roll No. : MCA226041

Thesis Title: Malignancy Identification From Cytology Images Using Deep Optimal Features

Signature:
____

Date:
____

ACKNOWLEDGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my thesis guide, **Prof. Nibaran Das**, Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis.

I am also grateful to the **Center for Microprocessor Applications for Training Education and Research Lab, Computer Science and Engineering department, Jadavpur University** for giving me the proper laboratory facilities as and when required.

I am thankful to all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least; I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

OINDRILA GHOSH
University Registration No. : 149902 of 2019-2020
Examination Roll No. : MCA226041
Master of Computer Application
Department of Computer Science and Engineering
Jadavpur University

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION:

1.1 WHAT IS CANCER?

In recent years, there has been a startling increase in the number of people diagnosed with cancer [2]. Typically, cancer is caused by the aberrant multiplication of cells at their origin. Despite the fact that genetic factors are the most common causes of cancer, there are a number of other factors that might also hasten the onset of the disease. Plastics, heavy metals, radiation, toxic chemicals, eating fast food and processed foods, and simply living can all cause cells to divide incorrectly. Apart from this factor like Excessive smoking and tobacco consumption, UV, X-ray, and Gamma rays from the sun and other sources, Infection with various viruses, bacteria, and germs etc. can be the cause of cancer.

According to the World Health Organization, 4.6 million people die from cancer each year. Cancer caused 2.8 million deaths in 2020. Breast Cancer claimed the lives of 437000 people. If it is detected at an earlier stage, the mortality rate can be decreased.

The objective of this thesis is to identify the optimal number of feature subsets from breast cytology images using a specific optimization technique that yields a high recognition accuracy.

The experiment for the proposed work is tested on breast cytology images. Breast cancer affects the majority of women between the ages of 40 and 60. Women who carry the BRCA1 and BRCA2 genes have a significantly increased risk of breast cancer.

1.2 WHAT IS CYTOLOGY?

The human body is made up of several cells, each of which has a nucleus and cytoplasm. The genetic imprint, or DNA, is found in the nucleus, and it is this DNA that is mutated when a disease is acquired. The change can be seen in the form of size, colour, texture, and nature of the nucleus and cytoplasm, and malignant and pre-malignant diseases can be detected by observing these parameters under a microscope. Cytology, often stated as cytopathology is a discipline of medicine that deals with the examination of cells in order to discover the existence of any aberrant cellularity.

- **Utility of Cytology**

  To detect lumps or masses, imaging techniques such as magnetic resonance imaging (MRI), X-ray (plain film and computed tomography [CT], ultrasound (US), optical imaging, and others are used. However, these imaging techniques are unable to evaluate images at the cellular level. As a result, cytology is commonly employed to discover abnormalities in cell structure. Expert cyto-technologists use cytology to identify cancer by studying cell samples obtained through biopsy or cytology. When compared to biopsy, cytology has more number of advantages.

- **WHEN IS CYTOLOGY USED?**

  Cytology tests can also be used for following purposes:

  - Infectious diseases are diagnosed.
  - Inflammatory conditions are diagnosed.
  - Thyroid lesions are examined.
  - Diseases affecting specific body cavities, such as the pleural cavity (the space between two thin membranes that line and enclose your lungs), are identified.

- **TESTS COMMONLY USED IN CYTOLOGY:** In cytology, there are two types of tests that are often used to diagnose malignancy.

  - **Screening Test:** This test is usually performed before the disease manifests itself, that is, when symptoms are not yet visible, and is advised for people who are in the high-risk category. Regular screening tests can aid in the early detection of the condition, allowing it to be treated with the best potential outcome.
  - **Diagnostic Test:** This test is frequently performed when symptoms first appear and is suggested for individuals who are diagnosed with a serious illness.

Both manual and automatic CAD (Computer Assisted Design) [4] based systems can be used to conduct the tests. A manual test is usually carried out under human supervision, and it entails human labour, time, and tiredness. Examining hundreds of slides per day is a time-consuming operation that necessitates the use of skilled cyto-technologists, which are currently in short supply.

As a result, systems based on automatic CAD are being developed that can process hundreds of specimens per minute while requiring little human participation. Nonetheless, it saves the cyto-technologist's time and efforts by supporting them technically in a variety of ways.

- **ISSUES WITH CYTOLOGY:** Despite the various benefits of cytology, there are a few difficulties that have yet to be tackled.

  - Neoplastic lesion to an exact anatomic location cannot be localized
  - Pre-invasive and invasive cancer cannot be distinguished.
  - Reactive from dysplastic cannot be distinguished.
  - Tumor type may not be determined.
  - The specimen's cellular structure is influenced by the specimen collector's experience, as well as the technique and instruments utilised..
  - With FNA, there are more false negative cases.

1.3 WHAT IS OPTIMAL FEATURE SELECTION?

Optimization is a method that maximizes the accuracy or minimizes the error of a model and finds the optimal solution, under certain constrains, according to the requirement.

When creating a predictive model, feature selection is the process of minimising the number of input variables to lower the computational cost of modelling and to increase the model's performance.

- **TYPES OF OPTIMAL FEATURE SELECTION:**There are numerous forms of feature selection algorithms, broadly divided into three categories: wrapper based, filter based and embedded techniques.

In filter-based method, a single statistical measure is selected that best fits the data and provides a score for each feature column and the columns are then sorted by their feature accuracy and returned in that order.

Filter and wrapper methods are combined in embedded methods. Algorithms with built-in feature selection methods are used to implement it.

Wrapper based approaches take a subset of features and utilise them to train a model. The features to be added or removed from the subset are determined based on the inferences obtained from the prior model. The problem can be reduced to a simple search problem. These approaches are frequently quite time-consuming to compute. Forward feature selection (a iterative strategy), recursive feature elimination (a greedy optimization algorithm, and other wrapper approaches are common examples.

1.4 IMPORTANCE OF OPTIMAL FEATURE SELECTION:

When creating a predictive model, feature selection is the process of minimising the number of input variables to lower the computational cost of modelling and to increase the model's performance.

Recent advances in applied science and related technologies have resulted in a massive evolution and generation of data with a huge number of features, potentially increasing the problem's processing cost. These increases in feature dimension may have a negative impact on the classification process' overall performance. Overfitting, which is computationally expensive and difficult to comprehend, is a common result. Due to the complicated nature of data, the number of characteristics in many real-world applications, such as biological data analysis, text processing, signal processing, data science, and microarray data analysis, tends to increase. Due to the availability of such high-dimensional information in relation to the short sample size, it leads to a typical problem known as the course of dimensionality.

The main benefit of the dimensionality reduction method is that it reduces the number of redundant, irrelevant, and noisy features while preserving the required information. It improves the model's performance while also lowering the problem's computing cost. Text mining, micro array data analysis, picture retrieval, face recognition, protein categorization, digit recognition, and intrusion detection are some of the uses of the dimensionality reduction method. Feature reduction, or FS, is the most extensively used method for reducing the problem's size. To separate out the redundant, superfluous, and noisy information, each input feature is turned into its matching new feature in feature reduction. The FS technique is more effective in terms of classification accuracy if related features are used. It employs an exhaustive search algorithm that examines every conceivable solution space. With an increase in the number of features, this search strategy increases the computational time of FS issues significantly. This stimulates the reduction of feature size in order to improve the FS approach's classification accuracy. Alternatively, it reduces the FS approach's overall computational cost.

Wrapper-based techniques use learning algorithms as a black box to select the most appropriate optimal subset of relevant features based on the classification model's accuracy. Based on the relevance of the feature, it determines the best suitable optimal features.The key advantages of the wrapper-based FS approach over the filter-based FS approach are that (i) it considers the classifier hypothesis, which may handle feature dependencies more effectively, and (ii) it determines optimal features via an induction algorithm.

1.5 CONTRIBUTION:

The problem of feature selection can be framed as an optimization problem. When there are few input features, all feasible combinations may be analysed, and the best subset can be determined decisively. A stochastic optimization approach can be used to explore the search space and find an effective subset of features when there are a large number of input features.

This thesis presents the Artificial Electric Field Algorithm for optimal feature selection to enhance the accuracy from cytology images dataset. The work flows in several steps:

- extracting feature from the last layer where loss is minimum
- classification algorithms is used to determine the initial accuracy
- Artificial Electric Field Algorithm is used to find out the optimal subset of features
- Classification algorithm is again applied on the optimal subset to check whether the accuracy has been enhanced.

CHAPTER 2

LITERATURE SURVEY

Numerous works have been reported on the classification of medical images utilizing the optimal feature selection. Optimization algorithms like GA [9], DE [10], PSO [27] and ACO [12] are popularly used for optimal features selection. Few works that have been studied during the journey of the work are stated below.

- Das et al. published a paper [3] on feature reduction using a neural fuzzy model. A Linguistic Neuro-Fuzzy with Feature Extraction (LNF-FE) model was used to analyze medical data for disease classification in this paper. In the Neuro-Fuzzy (NF) model, Feature Extraction (FE) algorithms are utilised to extract only those features (a smaller feature set) that significantly contribute to the network. With a range of benchmark biomedical data obtained from diverse fields, the usefulness of this suggested model has been evaluated and validated. For proof of correctness, the acquired results were evaluated using statistical techniques such as Friedman and Holm-Bonferroni.
- In paper [6] on optimized feature selection using Jaya Algorithm by Das et

al., a novel feature selection (FS) approach based on Jaya optimization algorithm (FSJaya) along with supervised machine learning techniques has been selected to select the optimal features.The method used a search methodology to identify the most appropriate features by updating the worst features in order to lower the feature space's dimensions.The performance was evaluated on Pima dataset and achieved accuracy 79%. The Friedman and Holm test were used to validate the suggested approach's statistical significance.

- Das et al.[7] presented a teaching-learning-based optimization (TLBO) based on the feature selection (FS) approach called FSTLBO for finding the optimal features in their paper [7] titled 'Optimal Selection of Features by Teaching-learning-based Optimization Algorithm for Classification.' This method replaced weak features with strong ones.

- In the paper [8], Peng et al. presented a novel feature selection strategy for dealing with biological data classification difficulties involving high dimensionality.The method suggested in this paper combines filter and wrapper approaches into a sequential search procedure to improve the classification performance of the features selected. The method proposed in this paper is highlighted in such a way that a pre-selection step was added and hence the efficiency of searching feature subsets was improved with better classification results.

- A new hybrid genetic algorithm (HGA) for feature selection (FS), called HGAFS is presented in the paper [9] by Kabir et al. HGAFS contains a novel local search operation to tune the search in the FS process in a fine way, that was designed and implemented in HGA. The local search method is based on the distinct and informative nature of input features, as determined by their correlation information. The goal is to direct the search process so that less correlated (different) information such as general and special properties of a dataset can be used to adapt the newly created offspring. HGAFS was tested on 11 real-world classification datasets which have dimensions that varies from 8 to 7129 and the performance was compared with ten existing well-known FS algorithm and it was found HGAFS produced better performances with essential features with better accuracy.

- In the paper [10] by Qin et al., he presented the self-adaptive DE (SaDE) algorithm, in which both trial vector generation strategies and their associated control parameter values are gradually self-adapted by learning from previous experiences in creating promising solutions. Consequently, a more suitable generation strategy was determined along with parameter settings. The approach was tested on 26 bound-constrained numerical optimization problems and the results were satisfactory.

- In the papers [11] and [12], Ant Colony Optimization which is inspired by real-life observations of ants searching for the shortest routes to sources of food is presented for the improvement of text categorization performance which is of utmost important in feature selection.The proposed approach is simple to build and has a low computing complexity due to the usage of a basic classifier. The feature selection using AOC is tested the Reuters-21578 dataset and the performance is compared to genetic algorithm, information gain, and CHI. The proposed algorithm's superiority was demonstrated by simulation results on this dataset.

- The paper [13], Too et al. tackled the feature selection problem using a multiple inertia weight technique and innovative co-evolution binary particle swarm optimization (MIWS-CBPSO). Ten benchmark datasets from the UCI machine learning repository are used to validate the proposed technique. Four contemporary and popular feature selection approaches, namely BPSO, genetic algorithm (GA), binary gravitational search algorithm (BGSA), and competitive binary grey wolf optimizer (CBGWO), are used in a performance comparison to analyse the effectiveness of the method.

- Mitra et al.[16] proposed a superpixel based cytology image segmentation approach by applying various morphological and clustering algorithms like anisotropic diffusion, DBscan, Fuzzy C-means, etc. After that the features were extracted from the segmented nuclei and finally they achieved an accuracy 91% on SVM classifier.

- In the paper [17], Artificial Bee Colony (ABC) Algorithm has been implemented in CT Scan images of cervical cancer. ABC and SVM (Gaussian Kernel) has been used for feature selection and classification and the accuracy came out to be 99%

CHAPTER 3

PROPOSED METHODOLOGY

This thesis presents the AEFA approach for optimal feature selection to enhance the accuracy from cytology images dataset extracted by using ResNet-18 [19] CNN model. The work flows in several steps:

- extracting feature from the last layer
- classification algorithms is used to determine the initial accuracy
- Artificial Electric Field Algorithm is used to find out the optimal subset of features
- Classification algorithm is again applied on the optimal subset to check whether the accuracy has been enhanced.

A block diagram for the proposed methodology is given in Figure 3.1

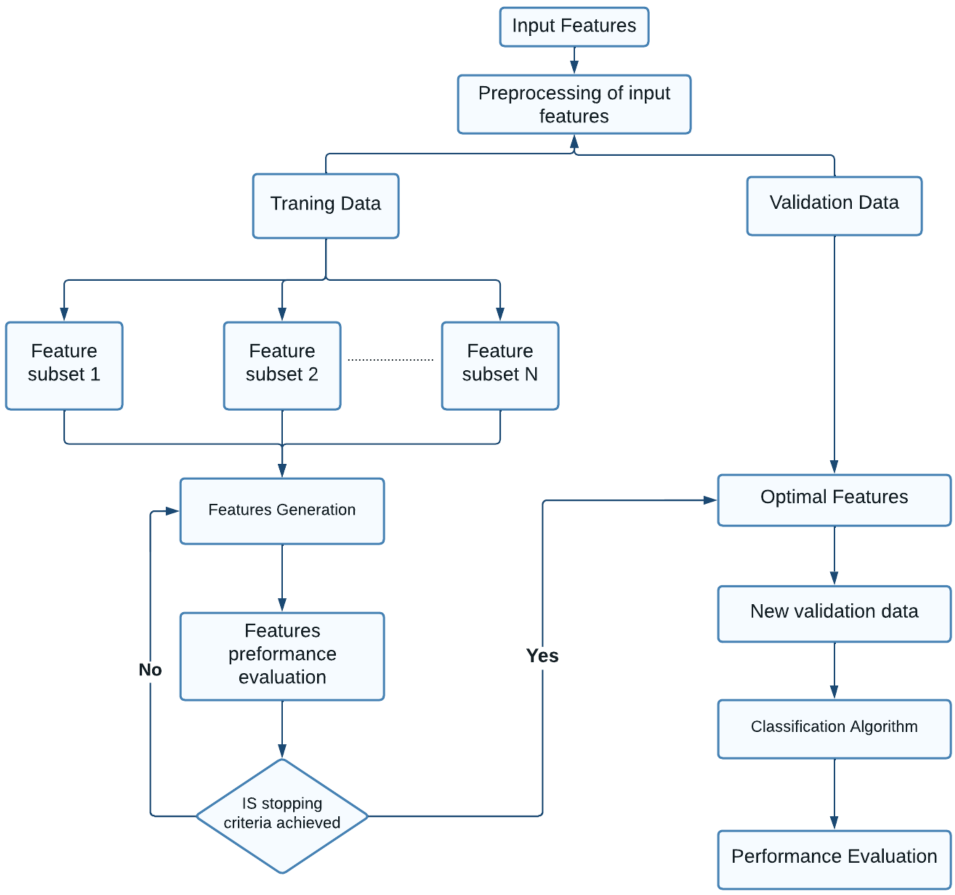


Figure 3.1: A Block Diagram of the Proposed Methodology

3.1 DATASET:

The study used FNAC-based breast cytology pictures obtained from the pathology centre ”Theism Medical Diagnostics Centre, Dumdum, West Bengal.” A total of 212 cytology pictures were taken, with 118 malignant and 94 normal samples. The photographs were acquired with a 5-megapixel resolution using a 40x optical zoom Olympus microscope.

The description of the dataset is given in Table 3.1

Some examples of cytology image samples are given in Figure 3.2

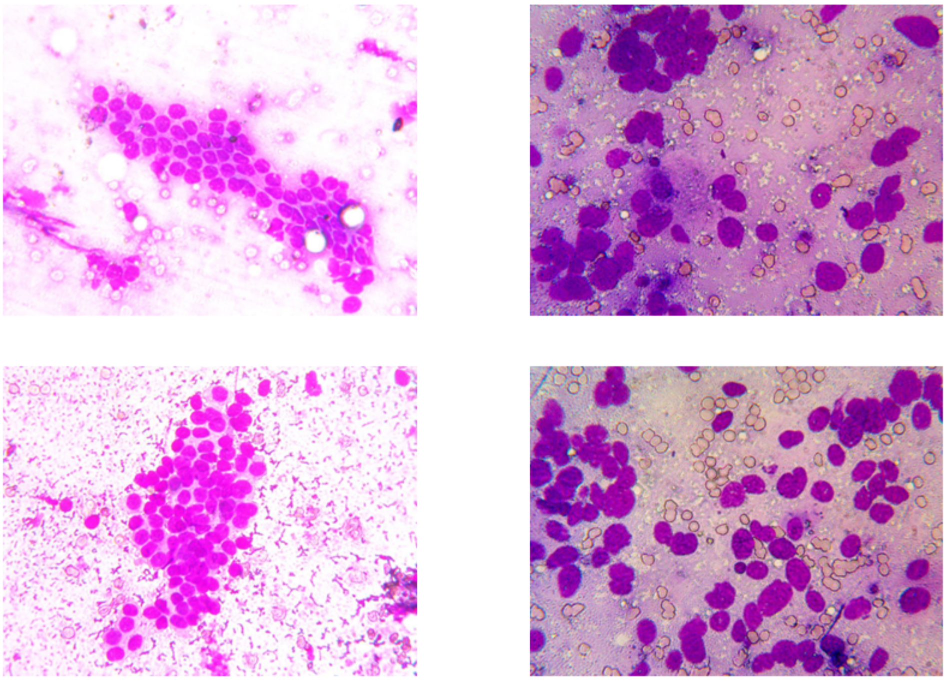
Table 3.1: Description of the Dataset

	Train	Test	Validation
Benign	77	09	08
Malignant	79	20	19

Brief description of RESNET-18, A popular Convolutional Neural Network, Architecture The Convolutional Neural Network (CNN) model used for extracting the features is RestNet-18 [19] .

The ResNet18 design has 72 levels and 18 deep layers. The architecture of this network was designed to allow for the efficient operation of a high number of convolutional layers. When neural networks are trained using back propagation, they use gradient descent to find the smallest weights by descending the loss function. Due to the presence of numerous layers, repeated multiplication causes the gradient to get smaller and smaller, eventually ”vanishing”, causing network performance to be saturated or even degraded.

ResNet’s central concept is the usage of jumping connections, also known as shortcut connections or identity connections. These connections work essentially by jumping across one or more levels, creating shortcuts between them. The goal of establishing these shortcut links was to tackle the problem of deep networks’ disappearing gradient. These shortcut connections fix the vanishing gradient problem by reusing the previous layer’s activations. Initially, these identity mappings do little more than bypass the connections, resulting in the utilisation of preceding layer activations. The network is compressed as a result of omitting the connection; as a result, the network learns faster.Following the compression of the connections, the layers are expanded to allow the residual part of the network to train and explore greater feature space. Because of its complicated layered architecture and the fact that the layers receive input from several layers and output to



(a) (b)

Figure 3.2: Examples of Cytology Image Samples [16], (a) Benign Samples, (b) Malignant Samples

numerous layers, the network is classified as a DAG network. For the analysis of medical pictures, residual networks and their modifications have been widely used. The Architecture of Resnet18 is shown in Figure 3.3

3.2 TRAINING PROCESS:

The 212 segmented images (94 benign and 118 malignant samples) that were previously segmented 14 using a superpixel-based technique are randomly separated into train, test, and validation sets in the ratio 3:2:1.Using training data, the validation set is used to pick an effective deep learning model. Depending on the

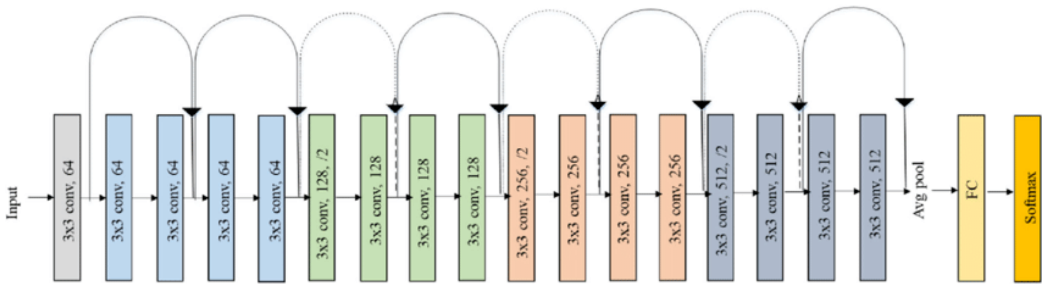


Figure 3.3: Architecture of ResNet18 [1]

validation set, the best model will be chosen. It is used to calculate a tuned model estimate.

To begin, images from the trained set are randomly flipped horizontally, and the flipped images are randomly cropped (horizontal and vertical). As a result, an image with a resolution of 960 x 1280 is arbitrarily cropped into the ResNet18 dimension of 224 x 224 and rotated at 30 degree angles. The input images must be scaled in these dimensions in order for pytorch to work. Neural networks were used to train these randomly cropped images. The clipped portions will be modified randomly at each epoch during training because the transformation processes are done at runtime. The model is trained in NVIDIA 510 and the PyTorch environment on a 6Gb GPU. The best is saved when there is the least amount of validation loss. The characteristics are derived from the best-trained ResNet-18 model's last layer.

3.3 ARTIFICIAL ELECTRIC FIELD ALGORITHM:

The artificial electric field algorithm [14] is based on the Coulomb's law of electrostatic force which states that the electrostatic (attraction or repulsion) force between two charge particles is directly proportional to the product of their charges and inversely proportional to the square of the distance between their positions.

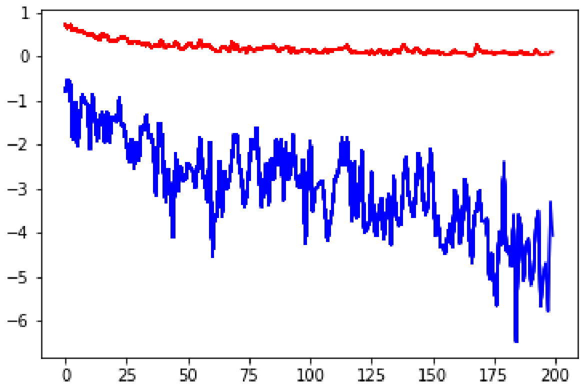


Figure 3.4: Plot-ResNet18 (Training error)

Agents are treated as charge particles in this method, and their strength is determined by their charges. An electrostatic force attracts or repels these particles, and objects move in the search space as a result of this force. As a result, charges are used as a direct method of communication via electrostatic force, and the position of the charge corresponds to a problem solution.Here, the charges are considered as the fitness function of the solution and the population . Only the attraction electrostatic force is taken into account in the algorithm, so that a charge particle with the highest charge (”best”) attracts all other particles with lower charges and moves slowly across the search space. As a result, the Artificial Electric Field ALgorithm can be seen of as a closed system of charges that obeys Coulomb’s law of electrostatic force and the laws of motion as follows:

- **Coulomb’s First Law of Electrostatic**: It states that like charges repel each other and unlike charges attract each other.
- **Coulomb’s Second Law of Electrostatic**: It states that The force of attraction or repulsion between like and unlike charge particles is proportional to the product of their charges and inversely proportional to the square of the distance between their centres.

- **Law of Motion**: It states that any charge's current velocity is equal to the total of its prior velocity's fractions plus the changes in velocity. The force exerted on the system divided by the mass of the particle equals variation in the velocity or acceleration of any charge.

The Physics of the AEFA algorithm can now be defined using the above laws. Let $Y_i = (Y_i^1, Y_i^2,...., Y_i^d)$ be the position of ith particle in the d-dimensional search space for i = 1, 2, 3,..., N, where Y_i^d is the position of ith particle in the dth dimension. The position of the global best fitness achieved by all charged particles, as well as the personal best fitness history of each particle, are used in the AEFA.The following Eq. gives the position of the best fitness value acquired by any particle i at any time t

$$Y(t+1) = \begin{cases} \text{Y(t)} & \text{if } f(Y_i(t)) < f(Y_i(t+1)) \\ \text{Y(t+1)} & \text{if } f(Y_i(t)) > f(Y_i(t+1)) \end{cases}$$

The Coulomb's constant K(t) can be derived using the following equation as a function of iteration and maximum iteration

$$K_t = K_0 * exp(-\alpha \frac{i}{maxiter})$$

where K_0 is the initial value and α is the parameter. The current iteration is i and the maximum number of iterations is maxiter. The Coulomb's constant is set to a high value at the start of the algorithm for exploration, and then gradually decreases iteration by iteration to manage the method's search accuracy.

The fitness functions calculate the charge of the particles, assuming that the charge of each particle is equal. One example of a good charge function is the equation below. The charge function can take various forms to suit the requirements, with some being better than others for specific types of optimization issues.

The most basic criterion for selecting a charge function is that the charge of the best particle should have the biggest value (normalised), i.e. qbest = 1, while all other particles of lower fitness should have a smaller value in the range [0,1]. The choice of charge is determined by whether the purpose is to maximise or minimise the fitness. It should be monotonically growing or decreasing for maximisation or minimization problems, respectively, to ensure the biggest charge and therefore greater force for the best fitness value.

$$q_i(t) = exp(\frac{fit_{y_i}(t) - worst(t)}{best(t) - worst(t)})$$

$$q_i(t) = \frac{q_i(t)}{\sum_{i=1}^{N} q_i(t)}$$

where f_{p_i} is the fitness of the i^th particle at time t.

For maximization problem:

$$best(t) = max(f_j(t)), j \in (1, 2, 3, ..., N)$$

$$worst(t) = min(f_j(t)), j \in (1, 2, 3, ..., N)$$

For minimization problem:

$$best(t) = min(f_j(t)), j \in (1, 2, 3, ..., N)$$

$$worst(t) = max(f_j(t)), j \in (1, 2, 3, ..., N)$$

The force exerted on the charge i from charge j at any time t is now defined as follows:

$$F_{ij}^d = K(t)\frac{q_i(t) * q_j(t)(Y_j^d(t) - Y_i^d(t))}{r_{ij}(t) + \epsilon}$$

where $q_i(t)$ and $q_j(t)$ are the charges of i_{th} and j_{th} particle at any time t, K_t is the Coulomb's constant at any time t, ϵ is a small positive constant $r_{ij}(t)$ is the euclidean distance between the two particles i and j given by

$r_{ij}(t) = ||Y_i(t), Y_j(t)||$

The total electric force exerted on the ith particle by all other particles in a d-dimensional search space at any moment t is represented by the equation below.

$F_i(t) = \sum_{j=1}^{N} rand() F_{ij}^d(t)$

rand() returns a uniform random number in the range [0, 1], which is utilised to give the algorithm its stochastic nature. F_i is the resulting force acting on the i_{th} charge particle, and N is the number of particles in the search area.

The following equation gives the electric field of the i_{th} particle at any time t and in the d_{th} dimension

$e_i^d(t) = \frac{F_i^d(t)}{q_i(t)}$

As a result, the acceleration of the i_{th} particle at any time t and in the d_{th} dimension is given by the following equation according Newton's second law of motion

$a_i(t) = \frac{q_i(t) e_i^d(t)}{m_i(t)}$

where $m_i(t)$ is the unit mass of the $i_t h$ particle at any time t.

The position of the particle and the velocity are updated as follows:

$v_i^d(t+1) = rand() * v_i^d(t) + a_i^d(t)$

Table 3.2: Different Parameters used for AEFA

Parameter	Description
d	Number of dimensions
n	Number of particles
maxiter	Maximum number of iterations
Y_i	Position of the i_th particle
$v_i(t)$	Velcity of the i_th particle at time t
$f(Y_i(t))$	Fitness of the i_th particle at time t
K	Coulomb's constant
best	Best fitness
worst	Worst fitness
q_i	Charge of the i_th particle
r_ij	Euclidean distance between the two vectors Y_i and Y_j
$F_i(t)$	Force exerted by the i_th particle at time t
$e_i(t)$	Electric field of the i_th particle at time t
$a_i(t)$	Acceleration of the i_th particle at time t

$Y_i^d(t+1) = Y_i(t) + v_i^d(t+1)$

The list of parameters and their meaning for the Artificial Electric Field Algorithm is given in Table 3.2. The flowchart of the Artificial Electric Field Algorithm is given in Figure 3.5.

3.4 SELECTION FEATURES USING AEFA

The challenge of selecting the best appropriate optimal subset of the feature is NP-hard and requires exhaustive search. With an increase in the number of characteristics, this search method continuously increases the number of solutions in the solution space. It is critical to choose the best value from among a variety of options, which is a difficult process. It improves the performance of feature selection difficulties at the end. The selection of optimal features is a three-step process: (1) A set of subsets of features is generated; (2) The fitness of these subsets of features are compared and evaluated among themselves to find the optimal

Algorithm 1 Algorithm of the Artificial Electric Field Algorithm

The values of dimensions, number of instances and maximum number of iteration are initialized
The position of the particles $(Y_1(t), Y_2(t), Y_3(t), ..., Y_n(t))$ are initialized randomly in the search range $[Y_{min}, Y_{max}]$
The velocities are initialized to a random value.
The fitness values $(f_1(t)), f_2(t)), f_3(t), ..., f_n(t))$ of $(Y_1(t), Y_2(t), Y_3(t), ..., Y_n(t))$ are evaluated
$iteration\ t \leftarrow 0$
while Stopping criterion is not achieved **do**
  Coulomb's constant K, best fitness and worst fitness are calculated
  The total force $F_i(t)$, the total electric field $e_i(t)$ and acceleration $a_i(t)$ is calculated
  The velocity $v_i(t+1)$ and position $Y_i(t+1)$ of each particle are updated.
  if fitness of the updated position is greater than the fitness of its previous value **then**
    the updated position is considered for next iteration
  else
    The previous position is taken into consideration
  end if
end while

subset of features or until the termination criteria are satisfied; and (3) The result is computed solely using optimal features. A new strategy of selecting features by AEFA is proposed to overcome the aforementioned issue by selecting the best acceptable subset of characteristics for classification. It speeds up and improves the accuracy of classification for prediction. The algorithm's goal is to choose the most appropriate feature set C from the initial feature set D, such that $C <= D$.

Initially, a random binary population of N particles $S = [S_1, S_2, S_3, ..., S_N]^T$ with dimension d denoting the total number of features or dimensions of the problem is generated. $S_i = [S_{i,1}, S_{i,2}, S_{i,3}, ..., S_{i,d}]$, where i = 1, 2,..., N is the sample size and $S_{i,d}$ represents the i_th particle position of the d_th feature of the population in the d dimensional feature space. Create a subset of characteristics for each instance of the population, using $S_{i,d}$ = 1 for analysis. Every feature in the popu-

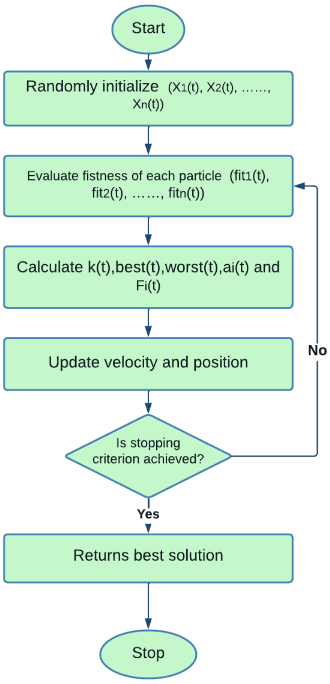


Figure 3.5: Flowchart of AEFA

lation is represented as a binary form (either 1 or 0). The bits 1 and 0 reflect the features that are selected or discarded for the classification process, accordingly.

For fitness (accuracy) computation, the randomly picked features are considered by a classification method, the Support Vector Machine (SVM) Radial Basis Function (RBF) kernel. The goal of the algorithm is to select the most appropriate subset of optimal features in order to maximise the algorithm's fitness.

$f_{S_i} = accuracy\ of\ the\ individual\ S_i^d$ in the d-dimensional feature space.

The charge for the i_{th} particle in d-dimensional feature space is calculated as follows

$q_i = exp(\frac{f_{s_i}) - worst(f)}{best(f) - worst(f)})$

$q_i = \frac{q_i}{\sum_{i=1}^{N} q_i}$

where f_{s_i} is the fitness of the i^th particle at time t.

For maximization problem:

$best(f) = max(f_j), j \in (1, 2, 3, ..., N)$

$worst(f) = min(f_j), j \in (1, 2, 3, ..., N)$

The Coulomb's constant K can be derived using the following equation as a function of iteration and maximum iteration

$K = K_0 * exp(-\delta \frac{iteration}{maxiteration})$

where S_0 is the initial value and δ is the parameter. The current iteration is iteration and the maximum number of iterations is maxiteration. The Coulomb's constant is initially set to a high value for exploration and then gradually reduced to control the search accuracy.

On generation g the force acting on the charged particles i and j is indicated by $F^g_{i,j}$.

$F^g_{i,j} = K \frac{q^g_i * q^g_j (S^g_{j,d} - S^g_{i,d})}{r^g_{i,j}) + \epsilon}$

where $q^g_i(t)$ and $q^g_j(t)$ are the charges of i_{th} and j_{th} particle at iteration g, K is the Coulomb's constant, ϵ is a small positive constant $r_{i,j}$ is the euclidean distance between the two vectors S_i and S_j given by

$r_{i,j} =$ square root of sum of the squares of the difference between corresponding element of the two vectors.

The total electric force exerted on the i_{th} particle by all other particles in a d-dimensional search space at generation g is represented by the equation below.

$$F_{i,d}^{g} = \sum_{j=1}^{N} rand() F_{i,j}^{g}$$

$F_{i,d}^{g}$, is the total electric force acting on the i_{th} particle in a d-dimensional feature space. Here, rand() is a uniformly distributed random number in the range [0–1] that gives the method its stochastic nature, and N is the total number of particles in the search space.

$e_{i,d}^{g}$ gives the electric field of the i_{th} particle at generation g in d-dimensional feature space.

$$e_{i,d}^{g} = \frac{F_{i,d}^{g}}{q_{i}^{g}}$$

As a result, the acceleration of the i_{th} particle at generation g in d-dimensional feature space is given by the following equation according Newton's second law of motion

$$a_{i,d}^{g} = K * e_{i,d}^{g}$$

The position of the particle and the velocity for the next generation g+1 are updated as follows. Here rand() is a random number generated uniformly in the interval of 0 and 1.

$$v_{i,d}^{g+1} = rand() * v_{i,d}^{g} + a_{i,d}^{g}$$

$$S_{i,d}^{g+1} = S_{i,d}^{g} + v_{i,d}^{g+1}$$

Now, the local maxima is calculated for each element and the position with higher value of fitness is considered for next generation.

By determining the probability of the new individuals probability$\left(S_{i,d}^{g+1}\right)$, the transfer function is employed to convert the continuous feature into its corresponding binary feature. The following equation would be used to update the elements in each population. If and only if the new individuals' probability is less than or equal to a random value, it is set to 0, otherwise it is set to 1. The rand() function is used to produce random numbers that are dispersed between 0 and 1, where probability is the probability.

$$\text{probability}\left(S_{i,d}^{g+1}\right) = 1/\left(1 + exp\left(-v_{i,d}^{g+1}\right)\right)$$

$$S_{i,d}^{g+1} = \begin{cases} 1 & \text{if } probability\left(S_{i,d}^{g+1}\right) > \text{rand()} \\ 0 & \text{otherwise} \end{cases}$$

Individuals $S_{i,d}^{g+1}$'s new fitness is assessed only on the basis of their former fitness. The freshly computed fitness value is compared to the old fitness value and updated. For the next generation, the best individuals with maximal fitness values are considered. For the next iteration, the best individuals' fitness value are taken into account.

$$S_{i,d}^{g+1} = \begin{cases} S_{i,d}^{g} & \text{if } f(S_{i,d}^{g}) < f(S_{i,d}^{g+1}) \\ S_{i,d}^{g+1} & \text{otherwise} \end{cases}$$

The list of parameters and their description for the proposed methodology is given in Table 3.3. The flowchart of the proposed method is given in Figure 3.6.

Algorithm 2 Algorithm of the feature selection technique by AEFA

$D \leftarrow 512, N \leftarrow 25, maxiteration \leftarrow 10$
The binary feature subsets $(S_1, S_2, S_3, ..., S_N)$ are initialized randomly in the search range
The velocities are initialized to a random value.
The fitness values $f_{S_1}, f_{S_2}, f_{S_3}, ..., f_{S_n}$ of $(S_1, S_2, S_3, ..., S_N)$ are evaluated
$iteration\ g \leftarrow 0$
while $g < 10$ **do**
    Coulomb's constant S, best fitness and worst fitness are calculated
    Fitness f_{s_i} is calculated for the particular iteration
    The total force $F^g_{i,d}$, the total electric field $e^g_{i,d}$ and acceleration $a^g_{i,d}$ are calculated
    The velocity $v^{g+1}_{i,d}$ and position $S^{g+1}_{i,d}$ of each particle are updated.
    The probability of each element of each subset is calculated
    if $Probability > random\ value$ **then**
        The value of that element of that subset is set to 1
    else
        The value of that element of that subset is set to 0
    end if
    if fitness of the updated subset is greater than the fitness of its previous value **then**
        the updated subset is considered for next iteration
    else
        The previous vector of that particular subset is taken into consideration
    end if
end while

Table 3.3: List of different parameters used in the proposed methodology:

Parameter	**Description**
D	Number of dimensions
N	Number of subsets
maxiter	Maximum number of iterations
S_i	Binary vector of $i_t h$ subset
$v_i(t)$	Velcity of the $i_t h$ subset
$f(S_i)$	Fitness of the $i_t h$ subset
K	Coulomb's constant
best	Best fitness
worst	Worst fitness
q_i	Charge of the $i_t h$ subset
$r_i j$	Euclidean distance between the two vectors S_i and S_j
$F_i(t)$	Force exerted by the $i_t h$ subset
$e_i(t)$	Electric field of the $i_t h$ subset
$a_i(t)$	Acceleration of the $i_t h$ subset

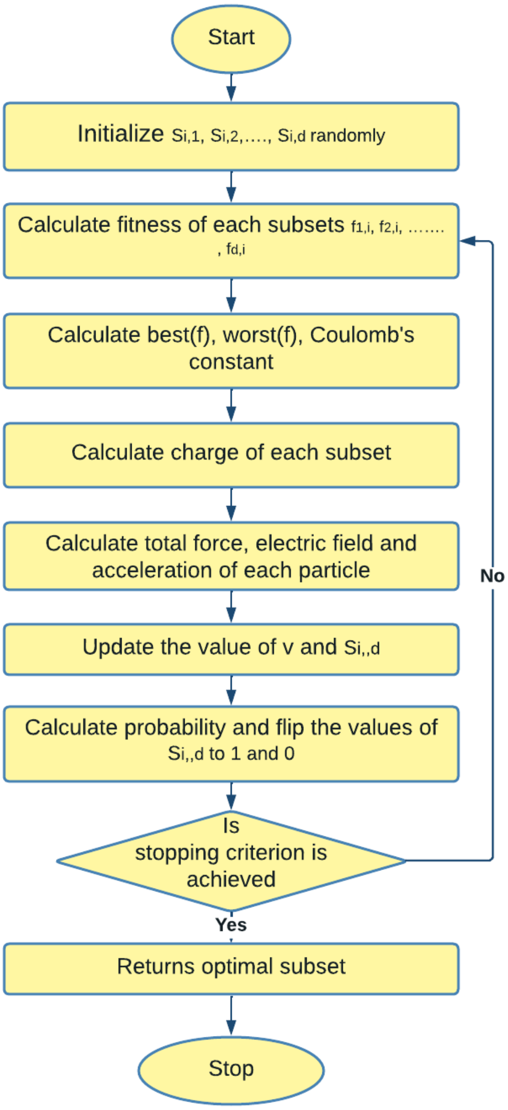


Figure 3.6: Flowchart for Feature Selection by AEFA

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

The breast cytology image dataset has been randomly divided into three halves: 75% for training data, 13% for test data, and 12% for validation for fitness (accuracy) evaluation. Random horizontal flipping and cropping are performed on images from the trained set (horizontal and vertical) to increase the size of the training set. As a result, an image with a resolution of 960 x 1280 is arbitrarily cropped into the ResNet18 dimension of 224 x 224 and rotated at 30-degree angle. The model is trained in NVIDIA 510 and the PyTorch environment on a 6 GB GPU. The best is saved when there is the least amount of validation loss. The features are extracted using the ResNet18 deep CNN model from the model's last layer, and the results (accuracy, precision, recall, f1 score) are tested using the Support Vector Machine (SVM) on Radial Basis Function (RBF) kernel. The maximum number of iterations and population size is set to 10 and 25, respectively. The stopping criterion for the experiment is also set to 10 iterations.

The recall, precision, and F1-Score for the model are calculated by using the following formulae, where TP = True Positive, FN = False Negative, FP = False Positive, TN = True Negative.

Table 4.1: Performances of SVM classifier with and without using optimal features selection

Performance Matrices	Without Feature Selection	With Feature Selection
Number of features	512	91
Accuracy(%)	0.62	**0.85**
F1 Score	0.61	0.83
Precision	0.62	0.81
Recall	0.64	0.82

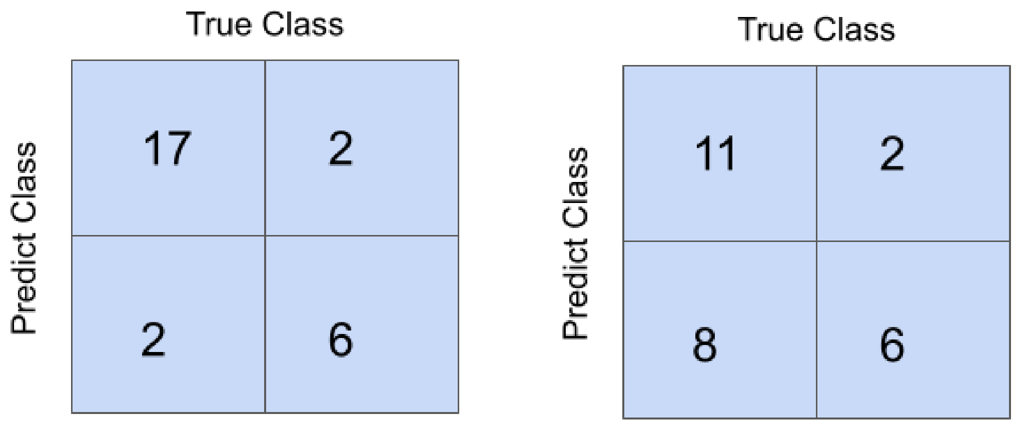


Figure 4.1: Confusion matrices of classification result using SVM. Left: After Optimal feature selection, Right: Before features selection

$$Recall = \frac{TP}{(TP+FN)}$$
$$Precision = \frac{TP}{(TP+FP)}$$
$$F1Score = \frac{(2*Precision*Recall)}{(Precision+Recall)}$$

The comparison of the performances of SVM classifier with and without using optimal feature selection is given in Table 4.1

The confusion matrices before and after feature selection are demonstrated in Figure 4.1.

The comparative analysis of the performance evaluation before and after optimal feature selection is given in Figure 4.2

The experimental results came out to be satisfactorily good as compared to the other optimization algorithms used for feature selection. When the features (512

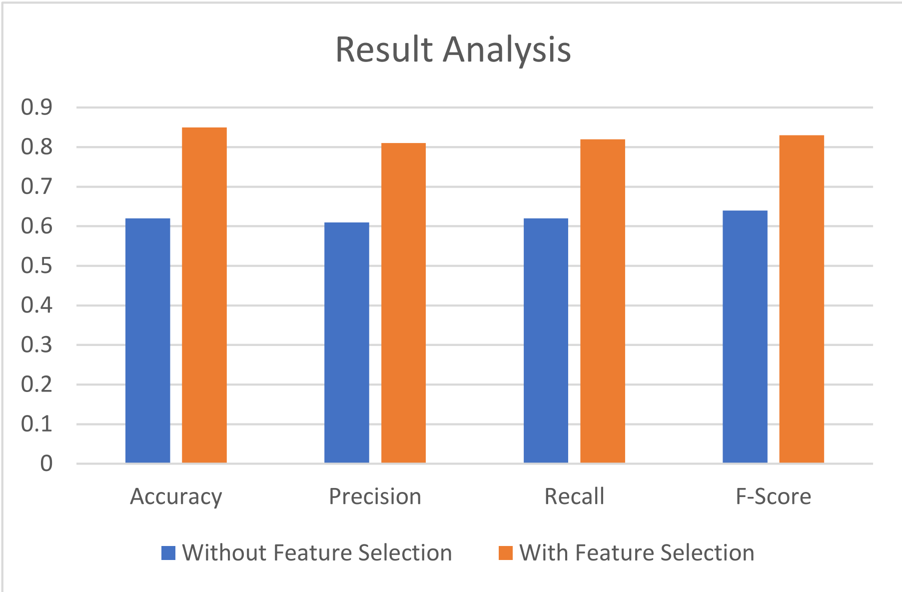


Figure 4.2: Comparison of different matrices obtained with and without feature selection

features) were in unfiltered form, the accuracy, precision, recall, and F1 score tested against the validation set were 62%, 62%, 64% and 61% respectively. The results came out better after the optimal subset was selected using the Artificial Electric Field Algorithm. The number of features selected in the optimal subset is 91 and the accuracy, precision, recall, and F1 score tested against the validation set were 85%, 71%, 82% and 83% respectively, which is better than the performance evaluation of the unfiltered feature set.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis, FSAEFA, a new feature selection method based on AEFA metaheuristic optimization, is proposed for cytology image classification. The objective of this study was to successfully determine the optimal subset of features using the AEFA method. The experimental results demonstrate that the proposed model FSAEFA significantly outperforms the competition. Compared to the majority of cases, the proposed model achieves superior computational performance and greater precision. Additionally, it chooses the optimal number of features. The test and validation precision of the breast cytology images are satisfactory.

In the future, a number of innovative and efficient metaheuristic algorithms could be combined with the wrapper-based FS method to select optimal features from a variety of classification models. These FS models have a wide range of potential applications, including disease prediction, weather forecasting, and text data analysis.

REFERENCES

[1] TY - JOUR AU - Ramzan, Farheen AU - Khan, Muhammad Usman AU - Rehmat, Asim AU - Iqbal, Sajid AU - Saba, Tanzila AU - Rehman, Amjad AU - Mehmood, Zahid PY - 2019/12/18 SP - T1 - A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks VL - 44 DO - 10.1007/s10916-019-1475-2 JO - Journal of Medical Systems ER -

[2] Silverstein, A., Silverstein, V. B., & Nunn, L. S. (2006). Cancer: conquering a deadly disease. Twenty-first century books.

[3] Himansu Das, Bighnaraj Naik, H.S. Behera, Shalini Jaiswal, Priyanka Mahato, Minakhi Rout, Biomedical data analysis using neuro-fuzzy model with post-feature reduction, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part A, 2022, Pages 2540-2550, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.01.007. (https://www.sciencedirect.com/science/article/pii/S1319157819311656)

[4] Bandler, J. W. (1969). Optimization methods for computer-aided design. IEEE Transactions on Microwave Theory and Techniques, 17(8), 533-552.

[5] Lever, J. V., Trott, P. A., & Webb, A. J. (1985). Fine needle aspiration cytology. Journal of clinical pathology, 38(1), 1-11.

[6] Himansu Das, Bighnaraj Naik, H.S. Behera, A Jaya algorithm based wrapper method for optimal feature selection in supervised classification, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part B, 2022, Pages 3851-3863, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.05.002. (https://www.sciencedirect.com/science/article/pii/S1319157820303670)

[7] Das, H., Chakraborty, S., Acharya, B., & Sahoo, A. K. (2020). Optimal selection of features using teaching-learning-based optimization algorithm for classification. In Applied Intelligent Decision Making in Machine Learning (pp. 213-227). CRC Press.

[8] Yonghong Peng, Zhiqing Wu, Jianmin Jiang, A novel feature selection approach for biomedical data classification, Journal of Biomedical Informatics, Volume 43, Issue 1, 2010, Pages 15-23, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2009.07.008. (https://www.sciencedirect.com/science/article/pii/S1532046409001014)

[9] Md. Monirul Kabir, Md. Shahjahan, Kazuyuki Murase, A new local search based hybrid genetic algorithm for feature selection, Neurocomputing, Volume 74, Issue 17, 2011, Pages 2914-2928, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2011.03.034. (https://www.sciencedirect.com/science/article/pii/S0925231211002748)

[10] Qin, A. K., Huang, V. L., & Suganthan, P. N. (2008). Differential evolution algorithm with strategy adaptation for global numerical optimization. IEEE transactions on Evolutionary Computation, 13(2), 398-417.

[11] M. Dorigo, M. Birattari and T. Stutzle, ”Ant colony optimization,” in IEEE Computational Intelligence Magazine, vol. 1, no. 4, pp. 28-39, Nov. 2006, doi: 10.1109/MCI.2006.329691.

[12] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, Text feature selection using ant colony optimization, Expert Systems with Applications, Volume 36, Issue 3, Part 2, 2009, Pages 6843-6853, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2008.08.022. (https://www.sciencedirect.com/science/article/pii/S0957417408005459)

[13] Too, J., Abdullah, A. R., & Mohd Saad, N. (2019, May). A new co-evolution binary particle swarm optimization with multiple inertia weight strategy for feature selection. In Informatics (Vol. 6, No. 2, p. 21). MDPI.

[14] Anita, Anupam Yadav, AEFA: Artificial electric field algorithm for global optimization, Swarm and Evolutionary Computation, Volume 48, 2019, Pages 93-108, ISSN 2210-6502, https://doi.org/10.1016/j.swevo.2019.03.013. (https://www.sciencedirect.com/science/article/pii/S2210650218305030)

[15] Shirude, S. B., & Kolhe, S. R. (2020). Improved hybrid approach of filtering using classified library resources in recommender system. In Intelligent Computing Paradigm: Recent Trends (pp. 1-10). Springer, Singapore.

[16] Mitra, S., Dey, S., Das, N., Chakrabarty, S., Nasipuri, M., & Naskar, M. K. (2020). Identification of malignancy from cytological images based on superpixel and convolutional neural networks. In Intelligent Computing Paradigm: Recent Trends (pp. 103-122). Springer, Singapore.

[17] Agrawal, V., & Chandra, S. (2015, August). Feature selection using Artificial Bee Colony algorithm for medical image classification. In 2015 Eighth

International Conference on Contemporary Computing (IC3) (pp. 171-176). IEEE.

[18] da Silva, S. F., Brandoli, B., Eler, D. M., Neto, J. B., & Traina, A. J. (2010, October). Silhouette-based feature selection for classification of medical images. In 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 315-320). IEEE.

[19] Ayyachamy, S., Alex, V., Khened, M., & Krishnamurthi, G. (2019, March). Medical image retrieval using Resnet-18. In Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications (Vol. 10954, pp. 233-241). SPIE.

[20] Aghdam, M. H., & Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research, 5.

[21] Khushaba, R. N., Al-Ani, A., & Al-Jumaily, A. (2011). Feature subset selection using differential evolution and a statistical repair mechanism. Expert Systems with Applications, 38(9), 11515-11526.

[22] Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iranian journal of basic medical sciences, 19(5), 476.

[23] Mitra, S., Das, N., Dey, S., Chakraborty, S., Nasipuri, M., & Naskar, M. K. (2021). Cytology Image Analysis Techniques Toward Automation: Systematically Revisited. ACM Computing Surveys (CSUR), 54(3), 1-41.

[24] Afza, F., Khan, M. A., Sharif, M., & Rehman, A. (2019). Microscopic skin laceration segmentation and classification: A framework of statistical nor-

mal distribution and optimal feature selection. Microscopy research and technique, 82(9), 1471-1488.

[25] Poonguzhali, S., Deepalakshmi, B., & Ravindran, G. (2007, February). Optimal feature selection and automatic classification of abnormal masses in ultrasound liver images. In 2007 International Conference on Signal Processing, Communications and Networking (pp. 503-506). IEEE.

[26] Narayanan, B. N., Hardie, R. C., Kebede, T. M., & Sprague, M. J. (2019). Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities. Pattern Analysis and Applications, 22(2), 559-571.

[27] Kharrat, A., Gasmi, K., Messaoud, M. B., Benamrane, N., & Abid, M. (2011). Medical image classification using an optimal feature extraction algorithm and a supervised classifier technique. International Journal of Software Science and Computational Intelligence (IJSSCI), 3(2), 19-33.

[28] Niwas, S. I., Palanisamy, P., & Sujathan, K. (2010, October). Wavelet based feature extraction method for breast cancer cytology images. In 2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA) (pp. 686-690). IEEE.

[29] Lakshmi, G. K., & Krishnaveni, K. (2014, February). Multiple feature extraction from cervical cytology images by Gaussian mixture model. In 2014 World Congress on Computing and Communication Technologies (pp. 309-311). IEEE.

[30] Saha, M., Mukherjee, R., & Chakraborty, C. (2016). Computer-aided diagnosis of breast cancer using cytological images: A systematic review. Tissue and Cell, 48(5), 461-474.

[31] Basak, H., Kundu, R., Chakraborty, S., & Das, N. (2021). Cervical cytology classification using PCA and GWO enhanced deep features selection. SN Computer Science, 2(5), 1-17.

[32] Jelen, L., Fevens, T., & Krzyzak, A. (2008). Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. International Journal of Applied Mathematics and Computer Science, 18(1), 75.

[33] Filipczuk, P., Fevens, T., Krzyżak, A., & Monczak, R. (2013). Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. IEEE transactions on medical imaging, 32(12), 2169-2178.

[34] Dey, P. (2021). The emerging role of deep learning in cytology. Cytopathology, 32(2), 154-160.

[35] Manna, A., Kundu, R., Kaplun, D., Sinitca, A., & Sarkar, R. (2021). A fuzzy rank-based ensemble of CNN models for classification of cervical cytology. Scientific Reports, 11(1), 1-18.

[36] Sornapudi, S., Brown, G. T., Xue, Z., Long, R., Allen, L., & Antani, S. (2019). Comparing deep learning models for multi-cell classification in liquid-based cervical cytology image. In AMIA Annual Symposium Proceedings (Vol. 2019, p. 820). American Medical Informatics Association.