

Mahipal Singh Choudhry

A study of feature Optimization Methods for Lung Cancer Detection

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SIGNAL PROCESSING AND DIGITAL DESIGN

Submitted by:

Jyoti Yadav

2K20/SPD/07

Under the supervision of

Prof. Mahipal Singh Choudhry



**DEPARTMENT OF ELECTRONICS AND
COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY, 2022

**DEPARTMENT OF ELECTRONICS AND
COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I **JYOTI YADAV** student of M.Tech (Signal Processing and Digital design), hereby declare that the project Dissertation titled “**A study of feature Optimization Methods for Lung Cancer Detection**” which is submitted by me to the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi

Date: 31/05/2022

Jyoti

Jyoti Yadav

**DEPARTMENT OF ELECTRONICS AND
COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

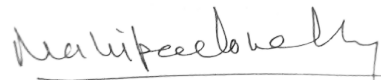
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Report titled “**A study of feature Optimization Methods for Lung Cancer Detection**” which is submitted by **Jyoti Yadav**, Roll No., 2K20/SPD/07 of Electronics and Communication Department, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 31/05/2022



Prof. Mahipal Singh Choudhry

SUPERVISOR

ACKNOWLEDGEMENT

A successful project can never be prepared by the efforts of the person to whom the project is assigned, but it also demands the help and guardianship of people who helped in the completion of the project.

I would like to thank all those people who have helped me in this research and inspired me during my study.

With a profound sense of gratitude, I thank Prof., my Research Supervisor, for his encouragement, support, patience, and his guidance in this research work.

Furthermore, I would also like to thank Prof. S. Indu and Prof. D. R. Bhaskar, who permitted me to use all required equipment and the necessary format to complete the report.

I take immense delight in extending my acknowledgment to my family and friends who have helped me throughout this research work.

Jyoti

NAME

ABSTRACT

In this project, Lung cancer remains an extremely important disease in the world that causes deaths. Early Diagnosis can prevent large amounts of deaths. Classifiers play an important role in detecting lung cancer by means of a machine learning set of rules in addition to CAD-based image processing techniques. For the classifier's accuracy, there is the need for a good feature collection of images. Features of an image can help to find all relevant information for identifying disease. Features are the important parameter for finding results. Mostly, features are extracted from feature extraction techniques like GLCM or some datasets already have features of lung cancer images by using some techniques. For different models of classifier, dimension, storage, speed, time and performance create an impactful effect on the results because we have large amount features of the images. An optimized method like the feature selection technique is the one solution that leads to finding relevant features from datasets containing features or features extracted from feature extraction techniques. The lung cancer database has 32 case records with 57 unique characteristics. Hong and Young compiled this database, which was indexed in the University of California Irvine repository. Take out medical information and X-ray information, for example, are among the experimental materials. The data described three categories of problematic lung malignancies, each with an integer value ranging from 0 to 3. A new strategy for identifying effective aspects of lung cancer is proposed in our work in Matlab 2022a. It employs a Genetic Algorithm. Using a simplified 8-feature SVM classifier and four-feature KNN, 100% accurateness is achieved. The new method is compared to the existing Hyper-Heuristic method for the feature selection. Through the maximum level of precision, the projected technique performs better. As a result, the proposed approach is recommended for determining an effective disease symptom.

CONTENTS

Candidate's Declaration	ii
Certificate	iii
Acknowledgement	iv
Abstract	v
Contents	vi-vii
List of Figures	viii
List of Table	ix
CHAPTER 1 INTRODUCTION	1
1.1 Background	1-2
1.2 Layout of the whole process	3
1.2.1 Data collection	3
1.2.1.1 Dataset including features	3
1.2.1.2 Dataset without including features	4
1.2.2 Image processing	4
1.2.3 Image segmentation	4
1.2.4 Feature extraction	5
1.2.5 Feature selection	5
1.2.6 Classification	5
1.3 Objective	6
1.4 Motivation	6
1.5 Report organization	6
CHAPTER 2 LITERATURE REVIEW	7-10
CHAPTER 3 PROPOSED METHODOLOGY	11
3.1 Theoretical foundations and key competencies	11

3.1.1 Machine learning	12
a) Artificial intelligence	13
b) Data mining	13
c) Optimisation	14
d) Generalisation	14
e) Approaches	14
f) Self-learning	14
g) Feature learning	15
3.1.2 Feature selection	16
3.1.3 Genetic algorithm	16-17
3.1.4 Model	18-22
3.2 Matlab implementation	23
3.2.1 Toolbox used	23
3.2.2 Model functions used in Matlab	23
3.3 The implemented structure for lung cancer detection procedure	24
3.4 Performance parameters	25
CHAPTER 4 SIMULATION RESULTS	25
4.1 Code snippet	25
4.2 Results	26-30
4.3 Performance of proposed work	31
4.4 Comparison analysis	32
CHAPTER 5 CONCLUSION	33
REFERENCES	34-36

LIST OF FIGURES

Fig. No.	Title	Page No.
Fig.1	The Architecture of the CAD System	5
Fig.2	The Architecture when the dataset has features of images	14
Fig.3	The Machine learning is a subfield of AI	15
Fig.4	The flow chart of the Genetic Algorithm	17
Fig.5	The Hyperplane in SVM classifier	19
Fig.6	The Supported vectors	20
Fig.7	The Holistic View of KNN	21
Fig.8	The K-NN Classification Algorithm	22
Fig.9	The framework of the Proposed method	24

LIST OF TABLES

Table. No.	Title	Page No.
Table.1	The Image segmentation techniques	07
Table.2	The Summary Table for the whole Survey	10
Table.3	Matlab functions	15
Table.4	The Comparison of the existing algorithm with the proposed work in terms of used different datasets, selected features, accuracy, and error rate	17

CHAPTER 1

INTRODUCTION

1.1 Background

Lung cancer is one of the commonly found cancers, with a survival rate of roughly 15%. The most important determinant in lung cancer survival is early detection. Lung cancer symptoms (features) do not develop until malignancy has spread to other body parts. To improve the survival rate, accurate early identification of lung cancer is required. It is necessary to characterize efficient features and delete redundant features among all characteristics for reliable detection. The difficulty of selecting informative characteristics from a huge number of options is known as feature selection.

In recent years, cancer has been one of the deadly diseases. Lung cancer is a disease that primarily affects men. However, the occurrence of breast cancer in emerging countries stays lesser than in Western countries, and the disease is on the rise across the region. Cancer mortality rates could be decreased significantly in the long run if the disease is discovered early. Cancer is the world's another-biggest reason for death, 9.6 million people died in 2018. According to World Health Organization figures [1], cancer is responsible for around 1 in every 6 deaths worldwide. In light of this growing worry, a new technology capable of reliably diagnosing and detecting cancer is required to prevent cancer at an early stage. It is critical to have a technology that can detect cancer at an early stage and is very sensitive. Methods like classification and data mining are useful for categorizing data, especially in the medical area. Computing and machine learning methods can greatly aid physicians in diagnosing and forecasting diseases at an early stage by establishing expert systems. One of the most important study fields in the field of medical diagnosis is Computer-Aided Diagnosis (CAD), as well as health proficient organizations and apparatuses [2]. Feature Selection is one of the most significant steps to produce optimized results to detect diseases like breast cancer, and lung cancer. In Popular domains of machine learning, pattern recognition, statistics, and data mining, feature selection helps in reducing dimensional complexity by choosing a subdivision of related attributes from the real attributes based on a few conditions or extracted characteristics using the feature extraction technique.

The following are a few good examples of feature selection techniques: thresholding technique, PCA, Chi-square, IG, forward selection, and relief, ACO, etc. It enhances the classifier's predictability, understandability, scalability, and generalization power, among other things. It also helps with knowledge discovery by reducing the memory requirement and computed difficulty, providing an earlier and added price-effective prototype [3]. It also provides new insights into which aspects are the most important or instructive. Feature selection, on the other hand, is a multi-stage process that is frequently expensive. Even the ideal model parameters for the entire collected attributes may be modified before the best design constraints for a particular subset of attributes can be determined. An attribute array can n-dimensional array in machine learning which signifies the attributes by calculating values across each trial. The attribute space is a term used to describe the space in which these array vectors exist. Most of the mentioned techniques can be used to minimize the dimension of the attribute space. Feature selection is a subset of feature extraction, which is a broader field. Feature extraction decreases the dimensionality of the data by transforming a novel attribute into a different attribute in the matrices by resetting the axes [4], whereas feature selection converts the novel attributes space into a subspace without alteration. Principal Component Analysis (PCA), Factor Analysis (FA), and Linear Discriminant Analysis (LDA) are instances of normal element extraction drawing near. Information Gain, Relief, Chi Squares, Fisher Score, and Lasso are some examples of feature selection approaches. Feature extraction is further generic compared to feature selection, and transformation paths might deliver superior inequitable capability. However, the main problem with the altered separate space is that there might not be a realistic sense of good explanation [5].

1.2 Layout of the whole process

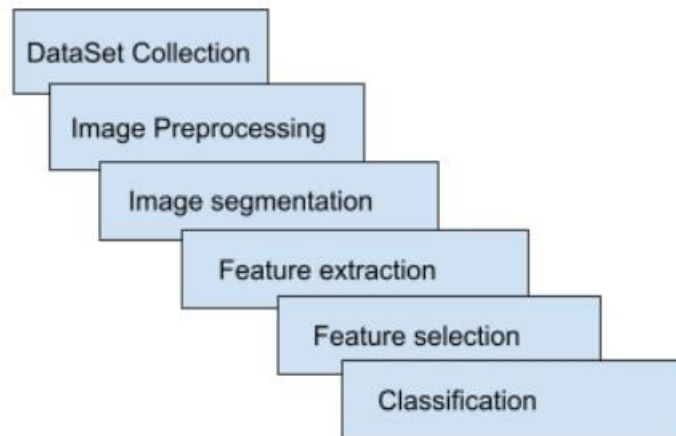


Fig. 1. The architecture of the CAD system

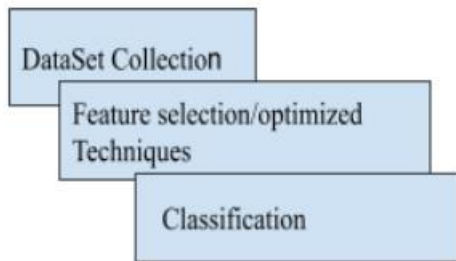


Fig. 2. Architecture when the dataset has features of images

Figure 1, consists of a whole process of the CAD system to detect a disease from classification. We will see a whole overview of the system. There are six steps in the process. Figure 2, This is a subset of the CAD system but there are three steps in which data already consists of features of images and the next step is to select relevant features for the classification so we use some feature selection techniques and optimized techniques for the classifier.

1.2.1 *Data collection*

1.2.1.1 *Dataset including features*

The lung cancer database has 32 case records with 57 unique characteristics. Hong and Young compiled this database, which was indexed in the University of California Irvine repository. Extracted from clinical data and X-ray data, for example, are among the experimental materials. The data described three categories of problematic lung malignancies, each with an integer value ranging from 0 to 3.

1.2.1.2 Dataset without including features

Here data is found which includes images of a particular disease and it can be CT or Diacom type images. The most common dataset used was the LIDC-IDC dataset which includes four experienced thoracic radiologists contributing lesion annotations. There are 1,018 low-dose lung CTs in the LIDC-IDRI database from 1010 lung patients. It is a global source for the advancement, training, and valuation of CAD system approaches for lung cancer discovery and analysis that are accessible via the internet. The purpose of this procedure was to identify nodes in every CT examination as completely as possible without necessitating forced consensus. Modalities "CT" with image size 127 * 127 pixels are included in the data.

1.2.2. Image pre-processing

A designed system will not be able to perform CT pictures directly. So completely processed pictures are needed for the implementation. These methods help to remove the corrupted or noisy photos and get proper pictures. This boosts the overall classification's effectiveness which ensures accurate results. Several techniques are being recorded such as Wiener filtering, Gaussian and Gabor filtering, Median filtering, and CLAHE.

1.2.3. Image Segmentation

The method of dividing a single photo into several chunks is known as image segmentation. The goal is to determine edges in a photo. The procedure of judging the objects in photos turns out to be easier as the pictorial complication gets vanished through this method. The multiple segmentation approaches used, as well as the number of sample pictures used, are listed in table1.

Table 1: Image segmentation techniques [6]

S.No.	Technique	Number of samples
1.	Watershed Transform Marked-Controlled watershed transform Thresholding & Marker Controlled	14 CT Images
2.	Morphological Operations	216 CT Images (128-train & 88- test)
3.	Region of Interest	33 CT Images (25-train & 8-test)
4.	Thresholding & Region Growing	60 CT Images
5.	Otsu Thresholding	1000 CT Images

1.2.4. Feature Extraction

The exercise of decreasing the number of assets essential to describe a huge quantity of statistics is referred to as function extraction. The massive number of variable quantities concerned is one of the most difficult components of carrying out massive statistics analysis. A big wide variety of variables calls for a whole lot of memory and processing strength, and it'd cause a typing algorithm to overfit training times and fail to generalize to new information. characteristic extraction is a huge term that pertains to techniques for combining variables to keep away from those issues at the same time as nonetheless correctly describing the facts. Many device mastering professionals agree that properly-optimized characteristic extraction is crucial to growing strong models. constructed units of application-structured capabilities, often built through an expert, can improve consequences. function engineering is an instance of this type of procedure. widespread dimensionality discount techniques also are employed, consisting of Structural Co-prevalence Matrix (SCM), PCA, and ROI. The maximum not unusual approach used is GLCM. The distribution of co-occurring pixel values (grayscale values, or shades) at a given offset is defined by way of grey-level co-incidence matrices (GLCMs). it is a texture evaluation technique it is hired in a variety of settings, which include medical photograph evaluation [8].

1.2.5. Feature Selection

Following feature extraction, feature selection is carried out mostly. This technique is made from a seek method for signifying novel function subcategories and an assessment degree for scoring the specific function subcategories. A mostly simple algorithm is to being assessed every possible subgroup of attributes and pick only with the bottom errors free. That is a complete examination of the space, and all but the smallest function units are computationally communicable. Many techniques such as PCA, thresholding, CCSA, and ACO are examples of feature selection techniques that help to produce good results.

1.2.6. Classification

Photo categorization becomes a fundamental duty that entails interpreting a photo in the wholeness. The main intent takes place to recognize the desired object by providing particular labeling. Categorization is the utmost frequently pre-owned method to label photos in that only one object is observable and estimated. On the other hand, Target findings require the completion of both classification and positioning tasks and are being applied to study further

real scenarios in which an image may contain a large number of objects. The aim of the test gets done to see if a lung nodule is malicious or benign. The various categorization methods as well as the classification methodologies used in machine learning like SVM, KNN, ANN, Fuzzy Particle Swarm Optimization (FPSO) and CNN, Decision Tree. The record frequently is being used methods like SVM and KNN.

1.3 Objective

- To propose an optimized method for selecting useful features and using classifiers to detect the lung cancer by measuring accuracy
- To give a review of past research and try to cover the most useful data for feature selection to identify cancer at an early stage.
- To compare the optimized method with the accuracy measure and error rate.

1.4 Motivation

For early detection, the most commonly used method is feature extraction and classification. It includes image pre-processing, image segmentation, feature extraction, and classification. Generally, the theory says that as we increase the number of features, it leads to good results. In any case, nowadays, this approach isn't delivering improved results. This is the essential motivation to present element determination techniques. And it helps to reduce time, cost, and storage and produce the best results for accuracy.

1.5 Report organization

This report has been organized into five chapters. Chapter 1, starts with the introduction of the need for feature selection. It also includes the motivation and objective behind the report. Chapter 2 deals with the literature study of various feature selection methods. Chapter 3 deals with the description Genetic algorithm and a complete study of the proposed method. In chapter 4, all the simulation results of the proposed methods. Chapter 5 arrangements are the conclusion and future scope of the proposed method.

CHAPTER 2

LITERATURE REVIEW

In 2021, Joana et al (2021) [9] proposed a complete approach for EGFR mutation status classification utilizing radiomic characteristics collected as of ROI spanning the complete lung where the nodule is being placed, reversing typical approaches that had previously focused solely on the nodule. The findings revealed that evidence from additional widespread sections of the lung cancer nodule allowed for a further thorough lung cancer characterization using this unique holistic approach. Linear SVM, Elastic Net, and Logistic Regression were shown to be the most reliable system for predicting EGFR transmutation position, and the findings encourage usage by others in future radiogenomics investigations. Also, they advocate using approaches that minimize the attributes figures before training the system, such as PCA. They are found to enhance analytical results. The best-performing classifiers were Linear Support Vector Machine, Elastic Net, and Logistic Regression paired with the Principal Component Analysis, a feature selection approach executed with 70 percent of the modification in the attribute subgroups, with AUC values are being ranged from 0.725 to 0.737.

Negar Maleki et al (2021) [22] proposed a genetic algorithm with KNN to classify lung cancer patients' risks into three levels: low, medium, and high. The classification accuracy improves dramatically when the kNN approach is combined with a feature selection algorithm. Six characteristics had been mentioned previously. As previously stated, the GA algorithm selected six features [2, 7, 10, 16, 17, 19], with the cost function value convergence. 912.5702 seconds are being taken to run the program in the fourth iteration. k was likewise greatest at 6. Future works could include comparing the results of alternative Machine learning classification algorithms for feature selection or population-based metaheuristics to those produced by the suggested approach.

Waseem et al (2020) [11] proposed a system for classifying automated lung cancer. From CT scan pictures, the suggested approaches allow a capable and exact detection of the nodule area of the lungs. The proposed scheme's performance was shown to be good in the experimental study. The suggested method's automatic nodule detection accuracy is 97.34 percent for the LIDC dataset and 96.55 percent for the ELCAP dataset. The LIDC dataset can detect nodules with a diameter of 2.2038 while the ELCAP dataset can detect nodules with a diameter of 2.2173. This plan could be expanded in the future to include 3D assessment and categorization of lung nodule cancer utilizing 3D slices. Because lung cancer detection and classification using CT scans are not currently employed in clinical practice, it will be necessary to design a method for a clinical application that will be useful in the automatic CAD structure for lung nodule cancer diagnosis at an initial phase.

Yasemin et al (2020) [10] suggested a diagnosis framework that has a lot of promise for lung cancer patients, and it could aid doctors by presenting initial findings and diagnoses. In her research, she found that the k-NN, NB, and DT processes worked well for classification jobs. (1) Z-score for normalizing methods, (2) principal component analysis (PCA) for

dimensionality reduction methods, and (3) information gain for feature selection methods were the greatest effective pre-processing approaches in terms of the performance of classification algorithms on LCDs. She conducted research into the use of machine learning approaches to improve lung cancer classification accuracy. 32*56 sized numerical data from the University of California, Irvine's Machine Learning Repository website. Instead of using classification algorithms directly, the accuracy of the classification system was improved in this work by using effective pre-processing approaches. Nine datasets were created using pre-processing approaches to achieve this improvement, and six machine-learning classification methods were applied. According to the findings, the k-nearest neighbor's approach outperforms random forest, naive Bayes, logistic regression, decision tree, and support vector machines in terms of accuracy. On the lung cancer dataset, the performance of pre-processing approaches was evaluated. Z-score (83 percent accuracy) for normalizing methods, principal component analysis (87 percent accuracy) for dimensionality reduction methods, and information gain (71 percent accuracy) for feature selection were the most successful pre-processing approaches.

Md. Siraj et al (2019) [12] proposed a model that produces significantly better results than the average Data mining was used in this study to aid in the diagnosis of cancer. To that end, an ensemble method was utilized to extract the most important characteristics in cancer detection and classify the cancer pattern based on the most influential features. Set union operation was used to assess the advantages of Principal Component Analysis (PCA), Pearson Correlation Coefficient (PCC), and Chi-Square (Chi2) for feature extraction. The major goal of this study is to investigate the impact of the dataset's most common features on the final result by lowering the features using the proposed hybrid feature selection method's set union operation. proposed a model that produces significantly better results than the average Data mining was used in this study to aid in the diagnosis of cancer. To that end, an ensemble method was utilized to extract the most important characteristics in cancer detection and classify the cancer pattern based on the most influential features. Set association activity was utilized to evaluate the benefits of Principal Component Analysis (PCA), Pearson Correlation Coefficient (PCC), and Chi-Square (Chi2) for including extraction. The major goal of this study is to investigate the impact of the dataset's most common features on the final result by lowering the features using the proposed hybrid feature selection method's set union operation

Sannasi et al (2019) [13] proposed a methodology to increase classification efficiency by selecting acceptable extracted features, and the results reveal that the PNN with CCSA-based feature selection performs better than the PNN without CCSA feature selection.

Lakshmanaprabu et al. (2019) [24] developed a hybrid method combining an optimum deep neural network (ODNN) and a linear discriminate analysis (LDA) to categorize lung nodules as malignancy or benign. The ODNN was originally utilized in their research to extract key features from computed tomography (CT) lung images. The features were then reduced in dimensionality using LDA. Finally, the ODNN was optimized using a modified gravitational search strategy. Their algorithm's sensitivity, specificity, and accuracy were measured at 96.2 percent, 94.2 percent, and 94.56 percent, respectively. Alirezaei et al. (2019) used four bi-objective meta-heuristic algorithms to find the fewest number of attributes with the highest classification accuracy rate (multi-objective firefly (MOFA), multi-objective imperialist competitive algorithm (MOICA), non-dominated sorting genetic algorithm (NSGA-II), and multi-objective particle swarm optimization (MOPSO)). They used several

pre-processing procedures first because of the importance of data quality. Then, as a classifier, SVM was utilized. MOFA was the most accurate of the above meta-heuristics, with 95.12 percent accuracy.

Lin et al (2019) [14], proposed a logistic regression version primarily based on CT examinations to become aware of the EGFR mutation position. Their model incorporated radiometric characteristics, and the AUC was 0.748, with 74.2 percent accuracy, 70.6 percent specificity, and 78.6 percent sensitivity in our study. They also developed a prediction system based on medical characteristics, with an AUC of 0.645. They will have a higher diagnostic value if radiomic features can be paired with medical aspects. Their findings indicated that radiomic characteristics can accurately predict EGFR mutation status. The radiomics technique for determining EGFR transmutation position can be improved and used in clinics in the future. The software AK (Version V3.2.0.R) was used to find abrasion attributes and select features. At that time, to predict EGFR mutation status, a logistic regression model was created. Finally, the receiver operating characteristic (ROC) curvatures were stayed to calculate the analytical presentation of characteristic characteristics. As a result, radiomic characteristics having a higher diagnostic value were eliminated.

Moumita et al (2018) [15] proposed an automated decision-making method for detecting and classifying nodules in lung CT scan pictures. To detect lung nodules, an adaptive/iterative thresholding technique, together with filtering and morphological operations, is utilized instead of classical thresholding. The features of the nodules are extracted to determine if they are benign or cancerous. Three types of characteristics are employed to categorize in this paper: geometric, texture, and intensity-based features. The redundant features are removed using the ranker searched feature selection approach, which minimizes computational complexity. Then, using KNN and SVM classifiers, classification is done. In comparison to other classifiers, the KNN and SVM with polynomial classifiers produce higher results. The proposed nodule detection method also has taken lesser time to detect nodules in the lungs. As a result, this approach can aid in automatic lung cancer diagnosis decision-making. The proposed technique performs better in terms of run time, taking 14 to 18 seconds for each scan.

Mitra Montazeri et al (2015) [21] was developed, which was based on a new heuristic called Hyper-Heuristic. By adding exploration, this approach may successfully search the solution space and appropriate exploitation of lower iteration (200). The planned strategy could strike a balance between manipulation and protection. The feature selection problem is being investigated. Exploitation was carried out using exploiter heuristics in this approach. Explorer heuristics were used to conduct the exploration. In other words, exploiter heuristics boost the candidate's quality. exploration heuristics dwell on random solutions to provide a better potential solution at each phase increase in exploration and disruption (not necessarily produce a better candidate solution). The proposed method compared five of the most used machine learning methods, their findings revealed that the recommended strategy performed better. Using a reduced 11 feature set, it achieved an accuracy of 80.63 percent.

Chen et al. (2013) [23] introduced a fuzzy system for Parkinson's disease (PD) diagnosis based on KNN (FkNN). They also employed principal component analysis to discover the most discriminating features on which to build the best FkNN model. They tested their system against the SVM algorithm, in addition, discovered that their proposed way outperformed it. Their FkNN had the best classification accuracy of 96.07 percent.

Table 2: Summary Table for the whole Survey

Year	Ref. No.	Author Name	Technique	No. of features extracted	Selected features from extracted	Dataset contains features with/without feature extraction technique	Findings
2021	9	Joana et al	Pairwise correlation, PCA, QR, No filter	Highest	Lowest	NSCLC-Radiogenomics Dataset	After extracting 1311 features, PCA 70% reduces features up to 8 with AUC 0.737± 0.018 by SVM (linear kernel)
2020	10	Yasemin Gultepe et al	Chi-square, information gain, forward selection, and relief.	56	28	The LCDS includes records of 56 features	The best accuracy value was found for the k-NN algorithm with IG, namely, an accuracy value of 0.71. Processing time is reduced and classification accuracy is good.
2020	11	Waseem Abbas et al	PCA	12	Reduced	LIDC & ELCAP	The suggested method's automatic nodule detection accuracy is 97.34% for the LIDC dataset and 96.55% for the ELCAP. LIDC gives better results with nodule size 2.203 cm
2019	12	Md. Sirajum et al	Combining PCA, PCC, and Chi2	—	Reduced	Wisconsin Lung Cancer Dataset	Individual classifier results were voted on by majority vote to achieve the best feasible result. The model's performance was evaluated using k-fold cross-validation (10 and 5 folds) with a train (80% and 70%) test (20% and 30%) split (AEEMCD).
2019	13	Sannasi et al	CCSA	13	6	lung CT images	The computation metrics employed include specificity, sensitivity, positive and negative predictive values, and accuracy. The results show that feature selection based on CCSA effectively achieves a 90% accuracy rate.
2019	14	Lin et al	ANOVA	396	9	lung CT images	After six steps in feature selection technique, it could able reduce 396 to 9 randomic features by using ANOVA technique with AUC 0.748
2018	15	Moumita et al	Threshold Technique	17	12	LIDC-IDRI	The proposed nodule detection also consumes less time to detect the nodules present in the lung region and take 14 to 18 scan per scan

CHAPTER 3

PROPOSED METHODOLOGY

The dataset with features was used in this study. The Genetic Algorithm was used in this section, along with classifiers like KNN and SVM. Let's talk about the project work in detail. Then we show how GA can help the kNN and SVM approach become more accurate.

3.1 Theoretical foundations and key competencies

3.1.1 Machine learning

AI (ML) is an area of examination connected with understanding and creating how-to "learn." H. The most effective method to utilize information is to work on the exhibition of different undertakings. Computerized reasoning is related to it. AI calculations fabricate models in light of preparing information and settling on forecasts and choices without unequivocal programming. AI calculations are utilized in an assortment of utilizations where it is troublesome or difficult to foster customary calculations to play out the necessary errands, for example, medication, email separating, discourse acknowledgment, and PC vision.

AI is firmly connected with PC insights, which center around making forecasts utilizing PCs. All the considered things, factual learning isn't all AI. The field of AI benefits from numerical improvement research since it gives devices, hypotheses, and application spaces. Information mining is a comparative field of study zeroed in on unaided learning for exploratory information investigation. Information and brain networks are utilized in some AI executions to reproduce how the natural cerebrum functions. AI is otherwise called prescient investigation when used to settle business challenges.

(a) Artificial Intelligence

AI emerged from the quest for man-made consciousness as a logical pursuit. A few scholastics were keen on causing machines to gain from information at the beginning of computer-based intelligence as a scholarly discipline. They attempted different representative techniques as well as what was then alluded to as "brain organizations," which were to a great extent perceptrons and different models that were in this manner found to be rehashes of summed up direct models of insights. Probabilistic thinking was likewise utilized, especially in clinical determination programming.

In any case, as the accentuation of intelligent, information-based approaches has grown, a break has arisen between simulated intelligence and AI. Hypothetical and reasonable information get-together and portrayal of issues tormented probabilistic frameworks. Master frameworks had assumed control over artificial intelligence by 1980. Hypothetical and common information about social occasions and portrayal issues tormented probabilistic frameworks. Master frameworks had overwhelmed simulated intelligence by 1980, and insights had become undesirable. Work on representative/information-based learning went on inside simulated intelligence, prompting inductive rationale programming, however, the more measurable line of examination, in design acknowledgment and data recovery, was present outside the discipline of man-made intelligence appropriate. Around a similar time, simulated intelligence and software engineering had deserted brain network research. Analysts from different disciplines, like Hopfield, Rumelhart, and Hinton, proceeded with this way as "connectionism" beyond the man-made intelligence/CS field. Their greatest advancement happened during the 1980s when they rehashed backpropagation.

AI (ML), which was laid out as a different field during the 1990s, started to create. The objective of the field moved from man-made brainpower to reasonable issues that could be addressed. It dismissed its concentration from the emblematic methodologies it had gotten from man-made intelligence and toward measurements, fluffy rationale, and likelihood hypothesis philosophies and models. The qualification between AI and man-made reasoning is as often as possible misjudged. ML learns and predicts in light of aloof perceptions, while simulated intelligence alludes to a specialist that collaborates with the climate to learn and make moves that increment its possibilities of achieving its goals. Many destinations guarantee that AI is a subfield of computer-based intelligence in 2020.

Others contend that not all AI will be computer-based intelligence and that just an "astute subset" of AI ought to be considered computer-based intelligence.

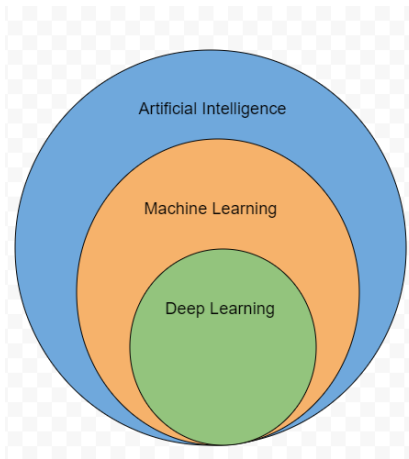


Fig. 3. Machine learning is a subfield of AI.

(b) Data mining

While AI and information mining both utilize comparative strategies and have huge cross-over, AI centers around expectation in light of realized characteristics gained from preparing information, though information mining centers around the finding of (already) obscure properties in the information (this is the investigation step of information disclosure in data sets). AI utilizes information mining techniques as "unaided learning" or as a pre-handling step to further develop student exactness, while information mining utilizes a few AI strategies with different points. The hidden suspicions they work with creating a ton of turmoil between these two scholarly networks (which regularly have particular meetings and diaries, with ECML PKDD being a striking special case). The capacity to recreate realized information is ordinarily estimated in AI, though the essential point of information disclosure and information mining (KDD) is the finding of already unseen information. When contrasted with existing information, a clueless (solo) procedure will effectively beat other regulated techniques, be that as it may, administered strategies can't be utilized in a commonplace KDD task because of the absence of preparing information.

(c) Optimization

Many learning issues are expressed as minimization of some misfortune work on a preparation set of occurrences, which relates AI to improvement. The distinction between the model's expectations and the real issue occurrences is communicated by

misfortune capacities (for instance, in arrangement, one needs to appoint a name to cases, and models are prepared to anticipate the pre-allocated marks of a bunch of models accurately).

(d) Generalisation

The objective of speculation recognizes advancement from AI: while enhancement procedures can limit a misfortune on a preparation set, AI is worried about diminishing misfortune on concealed examples. Describing the speculation of different learning calculations, especially profound learning calculations, is a hot region in momentum research.

(e) Approaches

Contingent upon the kind of the "sign" or "input" accessible to the learning framework, AI frameworks are by and large sorted into three significant classifications:

Regulated learning: A "instructor" gives the PC test inputs and wanted yields, determined to become familiar with a common principle that guides contributions to yields.

Unaided learning is the point at which the learning calculation isn't given names and is left to reveal structure in the information all alone. Solo learning can be an objective all by itself (tracking down secret examples in information) or a necessary evil (tracking down secret examples in information) (including learning).

Support learning happens when a PC program associates with a powerful climate to accomplish a particular objective (like driving a vehicle or playing a game against a rival). The product gets input as impetuses as it explores its issue space, which it endeavors.

to maximize.

(f) Self-learning

In 1982, self-advancing as an AI worldview was proposed, combined with the crossbar versatile exhibit, a brain network equipped for self-learning (CAA). There are no outside remunerations or educator ideas in this sort of learning. The CAA self-learning calculation figures the two activities and feelings (sentiments) in outcome situations in a crossbar style. The connection between acumen and feeling drives the

framework. Oneself learning calculation refreshes a memory network $W = ||w(a,s)||$ so the accompanying AI methodology is executed in every cycle.

It is a framework with solitary info, circumstances, and a solitary result, activity (or conduct). The environmental factors give no extra support or counsel. The feeling toward the outcome situation fills in as a backpropagated esteem (auxiliary support). The CAA exists in two conditions: one is the conduct climate, wherein it acts, and the other is the hereditary climate, where it obtains first feelings about conditions that it will insight into the social world for the solitary time. The CAA learns an objective looking for conduct in a climate that incorporates both alluring and undesirable circumstances after procuring the genome (species) vector from the hereditary climate.

environment.

(g) Feature learning

A few learning calculations are pointed toward finding better portrayals of the preparation inputs. Head part investigation and group examination are two notable models. Include learning calculations, otherwise called portrayal learning calculations, attempt to moderate the data in their contribution while likewise changing it in a helpful style, generally as a pre-handling venture before directing grouping or forecasts. This strategy takes into account the remaking of contributions from obscure information producing appropriation while keeping away from setups that are outlandish for that dispersion. This wipes out the requirement for manual component designing and allows a machine to learn and utilize highlights to satisfy a given undertaking.

The learning of highlights may be administered or unaided. Highlights are shown involving marked input information in administered including learning. Fake brain organizations, multi-facet perceptrons, and directed word reference learning are a few models. Highlights are learned with unlabeled info information in solo element learning. Word reference learning, free part examination, autoencoders, lattice factorization, and different grouping strategies are models.

Under the imperative that the learned portrayal is low-layered, complex learning calculations look to achieve so. Meager coding procedures attempt to do such while

remembering that the learned portrayal is scanty, and that implies the numerical model has a ton of zeros. The objective of multilinear subspace learning methods is to advance low-layered portrayals for multi-faceted information directly from tensor portrayals, instead of twisting them into higher-layered vectors.

Profound learning techniques uncover many layers of portrayal, or an ordered progression of elements, with more significant level, more dynamic highlights indicated (or created) as far as lower-level qualities. An astute PC, it has been proposed, is one that learns a portrayal that unravels the fundamental wellsprings of variety that make sense of the noticed information.

AI errands like grouping regularly request input that is hypothetically and computationally advantageous to investigate, which rouses include learning. True information, for example, photographs, video, and tangible information have opposed endeavors to algorithmically characterize specific characteristics. Another choice is to look at such elements or portrayals instead of relying upon express techniques.

3.1.2 Feature Selection

The strategy of element determination is pivotal in information mining. A system that makes sense of trait determination, variable choice, and variable subset determination is included in the determination. It alludes to the most common way of choosing a subset of important data (factors, indicators) for use in model development. An element choice calculation is a blend of a quest philosophy and a technique for producing new component subsets. It has an assessment measure that doles out a score to the different element subsets. The set hypothesis is utilized to get the last capabilities. More specifically, the association interaction is utilized to separate normal highlights from include sets found utilizing individual element choice techniques. Just the normal highlights from the malignant growth datasets are acquired after this interaction, and the capabilities that are erased are respected to be less powerful because they are not normal in all capabilities after directing individual element determination systems.

3.1.3 Genetic Algorithm

GA is a type of heuristic search. It can be used to find the best answer in spaces that are too large to be thoroughly examined. This algorithm is based on natural selection, the mechanism that drives biological evolution, and it can solve both limited and unconstrained optimization

problems. Natural sciences, mathematics, computer science, finance and economics, industry, management, and engineering are just a few of the fields where it might be used. It can mimic the kNN and SVM algorithm's characteristic determination approach. A genetic algorithm is divided into five stages:

1. The starting population
2. Function of fitness
3. Choosing
4. Transition
5. Variation

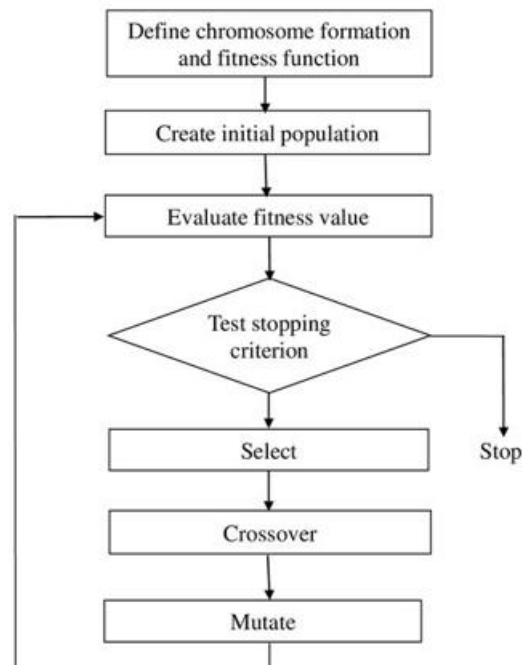


Fig. 4. The flow chart of Genetic Algorithm.

The GA method is an iterative method that involves a population connecting with a look space to find answers to a problem using a limited set of images termed the genome, which is gathered in a chromosome (solution). The basic GA remains the same as before: an underlying population of chromosomes is created haphazardly or heuristically. The chromosomes in the population are encoded and assessed using a fitness function that depicts the streamlining issue in the search space for each developmental advance (generation).

Chromosomes are chosen for their fitness to shape another population (the next generation). There are several options here, one of the simplest of which is the fitness proportionate decision, in which chromosomes are picked based on their relative fitness. This ensures that a picked individual is present the normal amount of times about its population performance. As a result, high-fitness chromosomes have a better chance of reproducing and transmitting new individuals to the population than low-fitness chromosomes.

Crossover and mutation are hereditary processes that introduce new chromosomes into the population. The crossover procedure involves two chosen individuals (parents) transferring bits of their genomes to generate two new chromosomes (offspring). In the meantime, the change activity forestalls untimely association to neighboring optima by arbitrarily assessing new concentrations in the pursuit space; it is carried out by flipping bits at random, with a low likelihood. GA is a stochastic iterative method with no guarantee of finding the best solution. Furthermore, the stopping condition could be specified as a maximum number of generations or the desired fitness value in several regions.

3.1.4 Model

Machine learning entails developing a model that is trained on a set of training data and then using that data to generate predictions. For machine learning systems, various types of models have been utilized and investigated.

(a) Training Model

Machine learning algorithms, in general, require a large amount of reliable data to make correct predictions. Machine learning engineers must aim and acquire a broad and representative sample of data while training a machine learning model. A corpus of text, a collection of photographs, sensor data, and data collected from individual customers of service are all examples of data from the training set. When training a machine learning model, keep an eye out for overfitting. Skewed or unwanted predictions might occur from trained models based on biased or non-evaluated data. Bias models may produce poor outcomes, exacerbating the negative effects on society or goals. Algorithmic bias can occur when data is not properly prepared for training. Machine learning ethics is becoming a subject of study, with many machine learning engineering teams incorporating it.

(b) SVM

A discriminative classifier is a Support Vector Machine (SVM). The essential notion of this classifier is decision planes, which establish decision limits. This classifier is described by a separating hyperplane. This approach returns an ideal hyperplane for supervised learning, which categorizes fresh cases. The hyperplane in two-dimensional space is a line that divides a plane into two sections by a line, with each class on each side. This classifier can handle several continuous and categorical variables and can perform regression and classification tasks. For categorical data, a dummy variable is produced with case values of 0 or 1. A regularisation parameter in this classifier prevents overfitting. It employs the kernel trick. Over-fitting the model selection criterion can be quite damaging to kernel models.

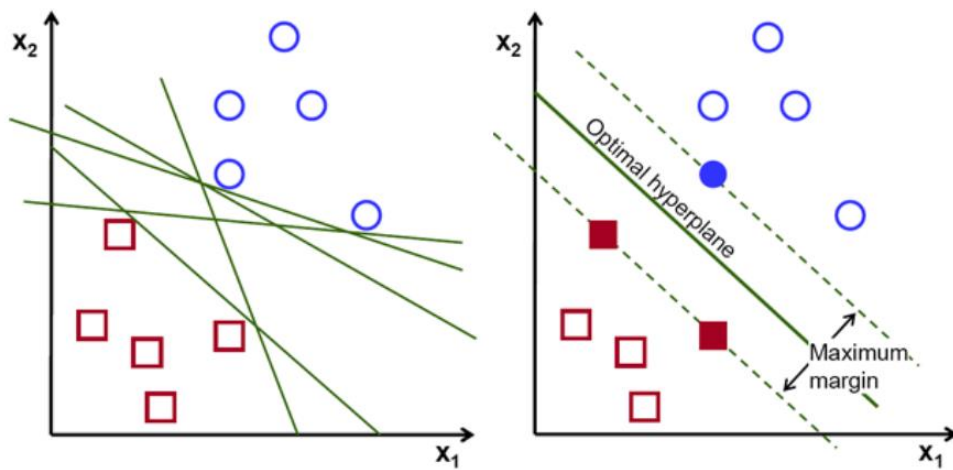


Fig. 5. Hyperplanes in SVM Classification.

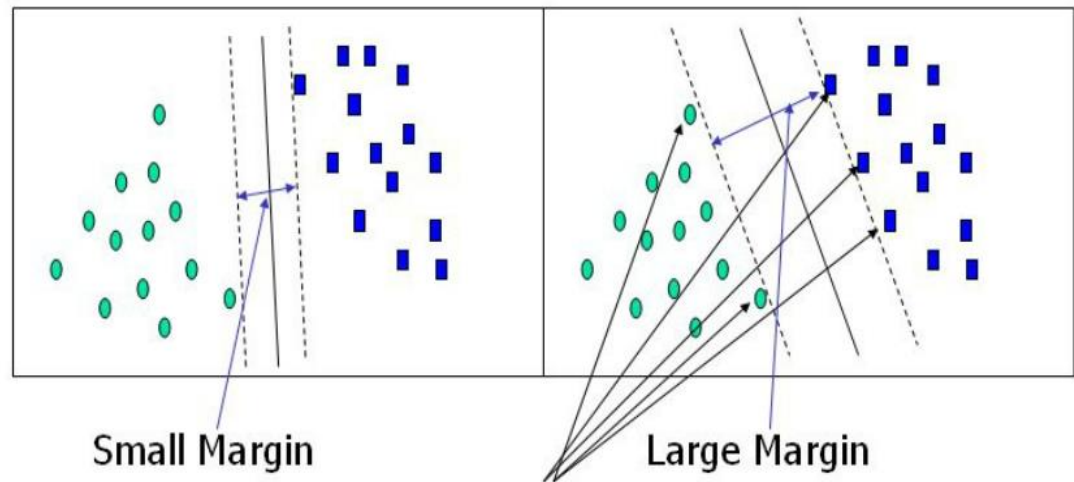


Fig. 6. Supported vectors.

(c) KNN

K-Nearest Neighbors (KNN) is a calculation that holds generally accessible models and orders new cases utilizing a similitude activity [18]. KNN can be utilized to take care of grouping and relapse foreseeing issues. With expanding K, the limit becomes smoother. On various K-esteems, the preparation blunder rate and the approval mistake rate are two boundaries to get to. Since it is hearty to uproarious preparation information, viable assuming the preparation information is tremendous, it has a critical hunt issue to find closest neighbors, and it additionally requires information capacity, KNN is exceptionally valuable in the field of order procedures.

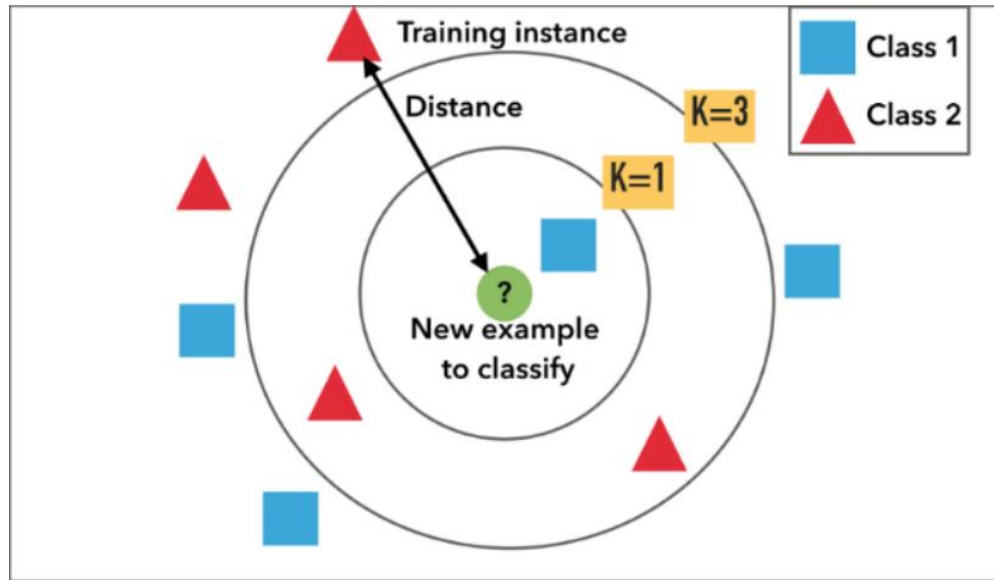


Fig. 7. Holistic View of KNN

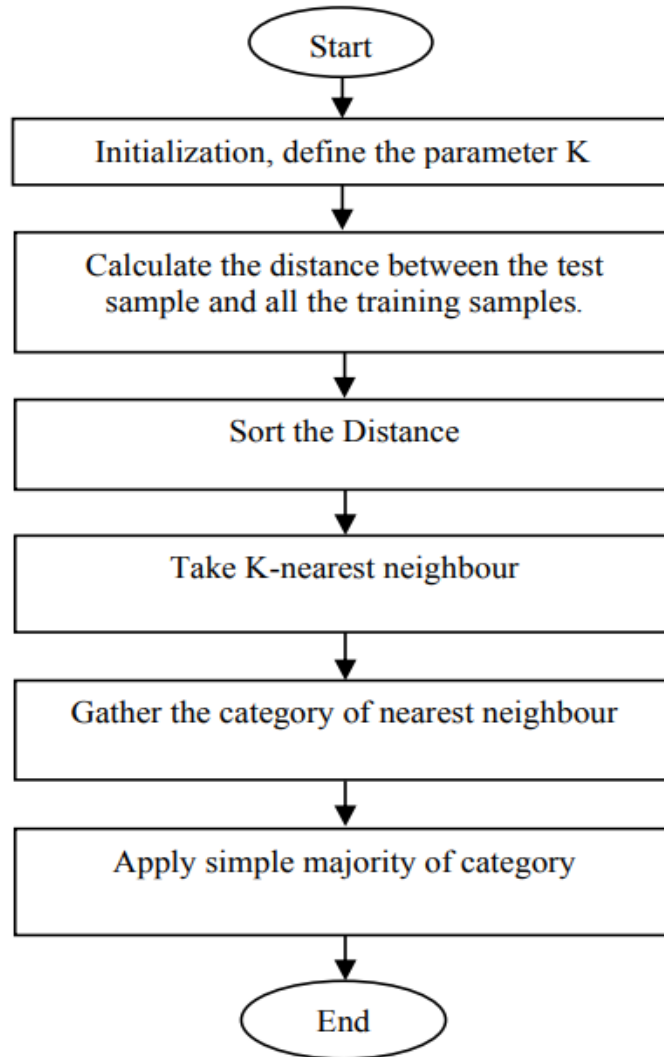


Fig. 8. K-NN Classification Algorithm

3.2 Matlab Implementation

MATLAB is a logical registering framework with an intelligent programming climate. Numerous specialized spaces depend on MATLAB for information examination, critical thinking, trial and error, and calculation creation. MATLAB-based discipline-explicit programming, organized into libraries of capacities called tool stash, is added every now and again utilized. MATLAB is broadly utilized in specialized training as the establishment for computational research facility work; more than 1000 course readings use it as an instructing device. The Mathworks of Natick, Massachusetts, created MATLAB. The fundamental purposes behind MATLAB's prosperity, as per the specialized figuring local area, are without a doubt the instinctive, exact, and clear documentation, its utilization of intricate networks as that of the default information type object, the force of the made administrators, easily utilized realistic plans, and its straightforward and agreeable programming model, which considers simple language expansion. The power of the computational models whereupon tasks are created can be added to this.

3.2.1 Toolboxes used

- Statistics and Machine learning
- Image processing
- Curve fitting

3.2.2 Model Functions used in Matlab

Table 3: Table for Matlab functions

S.No.	Function	Syntax	Description
1.	fitcknn	Md1=fitcknn(--, Name, Value)	Deploy the k-nearest neighbor classifier.
2.	fitcecoc	Md1=fitcecoc(--, Name, Value)	Multiclass models should be fitted to support vector machines (SVM or other classifiers).

3.3 The implemented structure for lung cancer detection procedure.

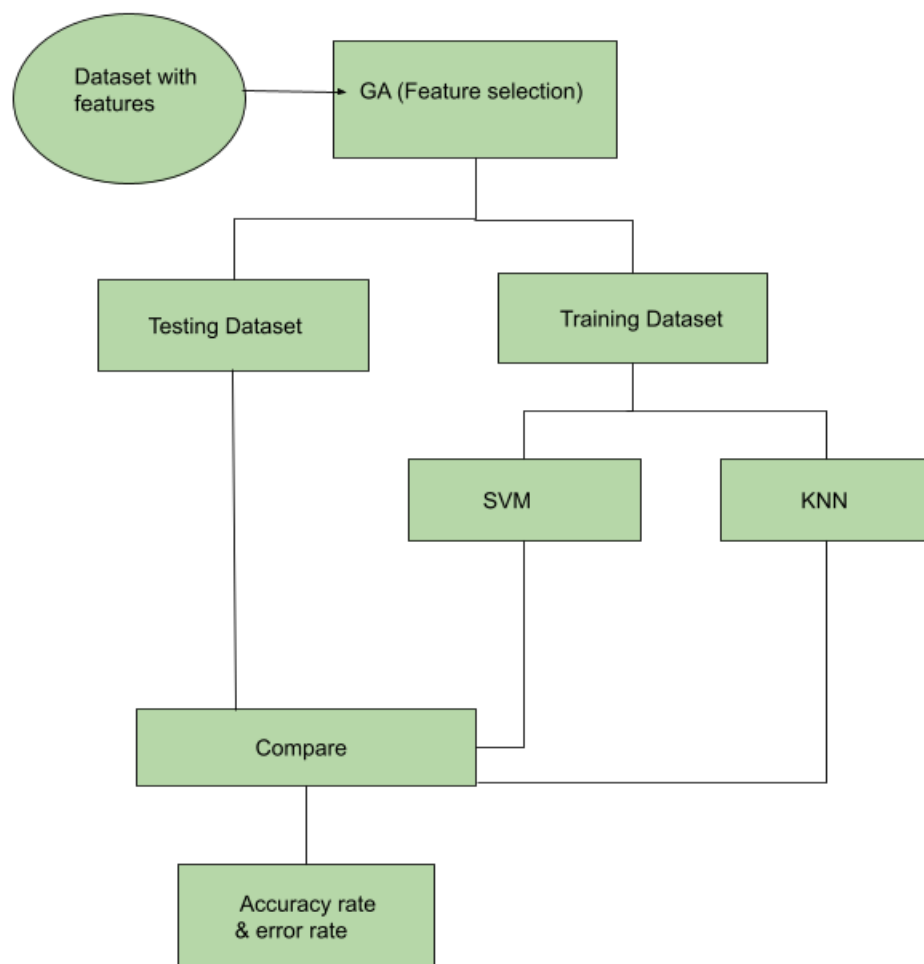


Fig. 9. The framework of the Proposed method

CHAPTER 4

SIMULATION RESULTS

4.1. CODE SNIPPET

```
main.m x +
1   clc
2   clear
3   close all
4   format compact
5   ML='SVM'; % 'SVM' or 'KNN'
6   %% load and divide dataset:
7   rng('default');
8   ds=readmatrix('ds.csv');
9   holdout=0.2; % 20% for test and 80% for training
10  NS=size(ds,1);
11  NF=size(ds,2)-1;
12  idx=randperm(NS, round(holdout * NS));
13  testData=ds(idx,2:end);
14  testLabels=ds(idx,1);
15  ds(idx,:)=[];
16  trainData=ds(:,2:end);
17  trainLabels=ds(:,1);
18  %% GA + SVM
19  % Genetic Parameter Settings:
20  maxGeneration=400;
21  popsize=20;
22  % initial population:
23  for i=1:popsize
24      pop{i}=randsrc(1,NF,[0,1]);
25  end
26  % Optimization:
27  for iter=1:maxGeneration
```

Fig. 10. Snippet of code

Name	Value	Size	Class
best	1×56 double	1×56	double
best_cost	33.3333	1×1	double
cost	1×20 double	1×20	double
ds	26×57 double	26×57	double
fitness	1×20 double	1×20	double
holdout	0.2000	1×1	double
i	20	1×1	double
idx	1×20 double	1×20	double
iter	1	1×1	double
maxGe...	400	1×1	double

Fig. 11. Workspace

4.2. RESULTS

4.2.1 GA+KNN

```

Command Window
-----
Generation #400  best cost=5.2632  No. of Features=12
=====
bestFeatures =
    2    3    4    5    7   17   19   20   29   32   37   46
accuracy =
    94.7368
MSE =
    0.0526
RMSE =
    0.2294

```

Fig. 12. Command window which is showing best features using genetic algorithm and KNN classifier.

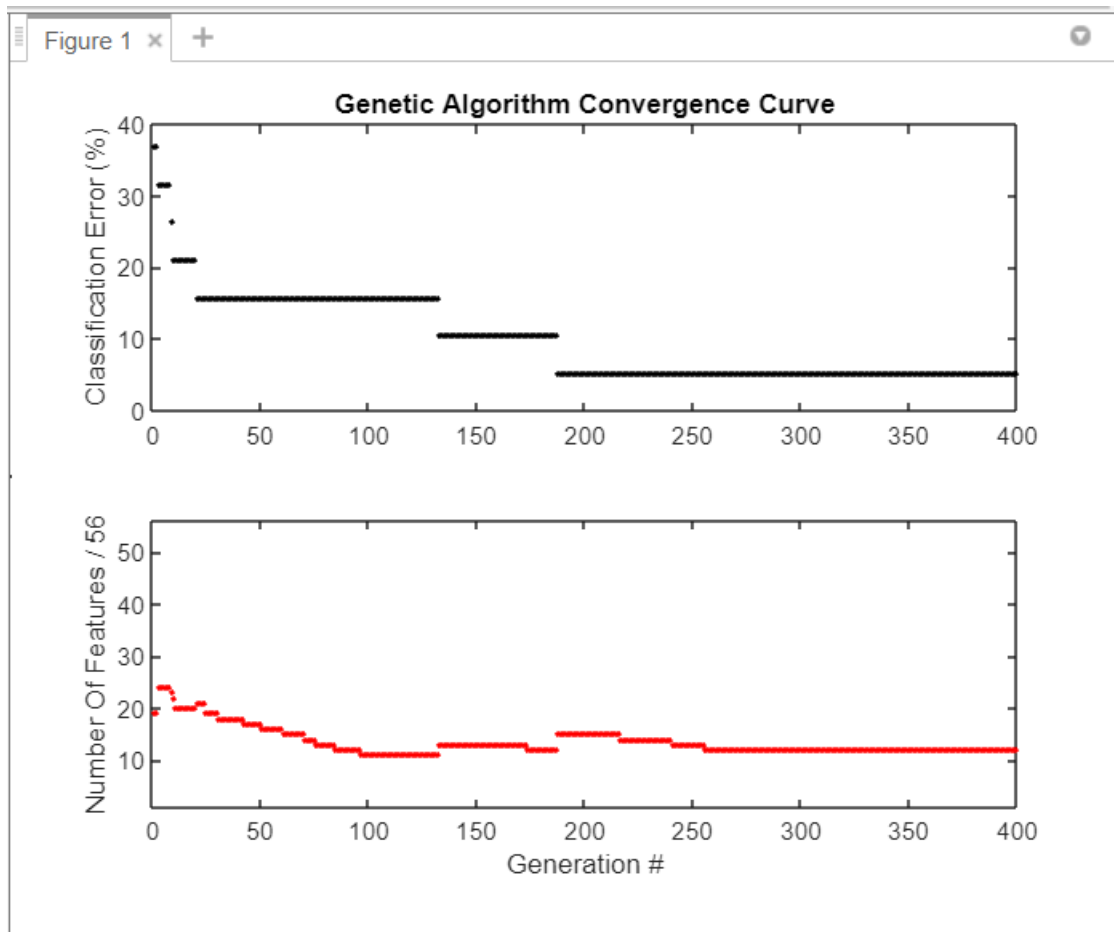


Fig. 13. Plotted Graphs for no. of features reduced in 400 generations and for classification error.

4.2.2 GA+SVM

```

Command Window
Generation #399 best cost=15.3846 No. of Features=6
Generation #400 best cost=15.3846 No. of Features=6
=====
bestFeatures =
    6    33    40    41    45    54
accuracy =
    84.6154
MSE =
    0.1538
RMSE =
    0.3922

```

Fig. 14. Command window which is showing best features using genetic algorithm and KNN classifier.

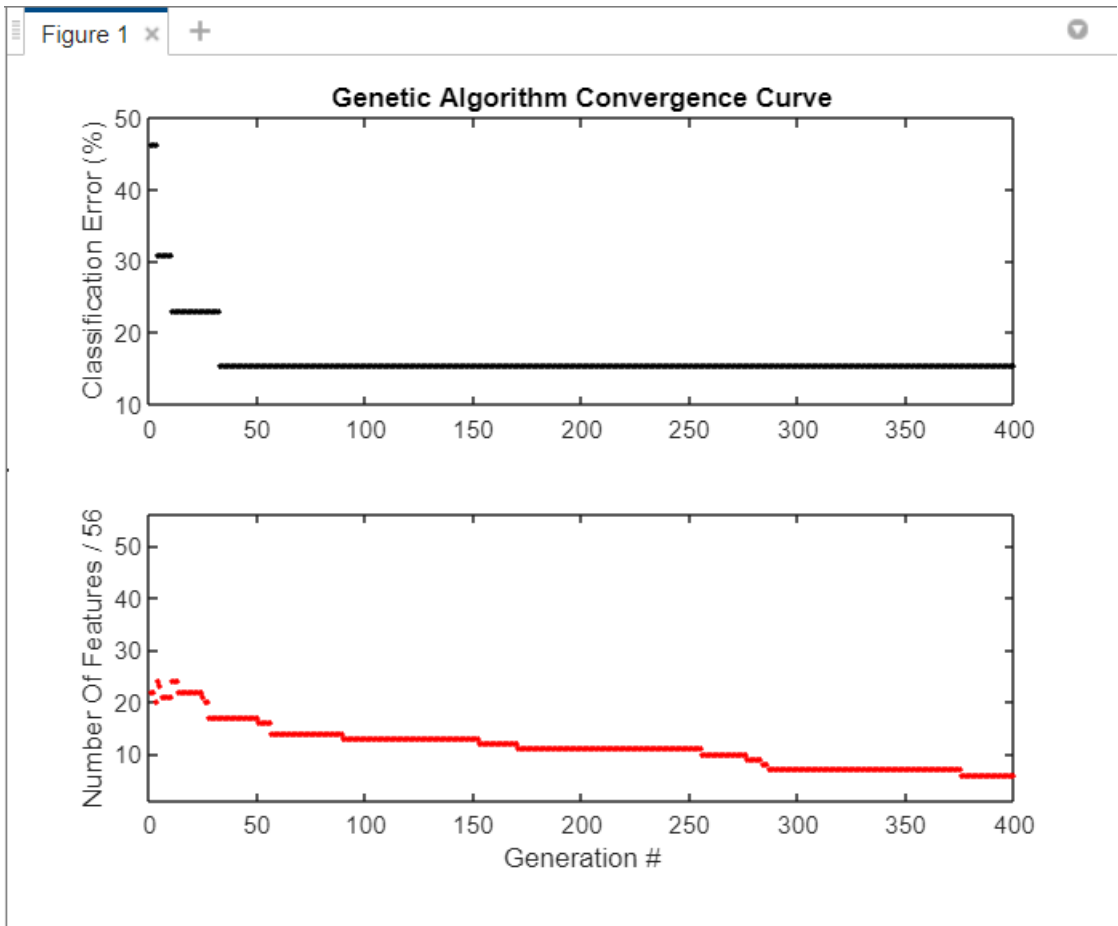


Fig. 15. Plotted Graphs for no. of features reduced in 400 generations and for classification error.

4.2.3 Cuckoo search Algorithm

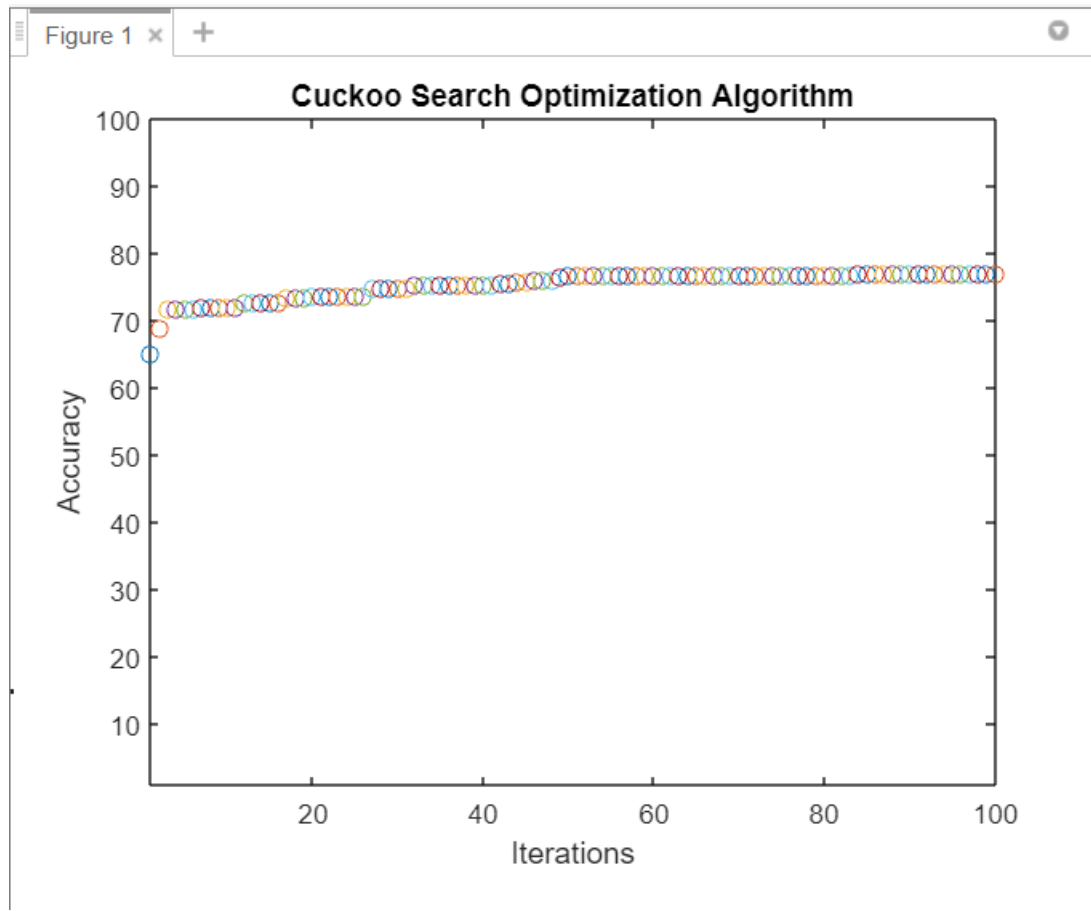


Fig. 16. Plot for Accuracy and best features selected using Cuckoo-Search Algorithm

4.2.4 IG+KNN

```
'KNN'  
  
Accuracy_Before_IF =  
34.1667  
  
Accuracy_After_IF =  
58.3333  
  
reduced_features =  
60  
  
best_features =  
2 7 10 13 14 19 20 23 29 37 44 46 49 52 57 59 63 75 78 79
```

Fig. 17. Plot for Accuracy and best features selected using Information Gain Algorithm with KNN

4.3 Performance of Proposed Work

4.3.1 Accuracy

Accuracy is one of the exhibition measures with a few implications in various regions. In the characterization strategies, in any case, exactness is characterized as a factual proportion of how well a paired grouping test accurately recognizes or avoids a condition. That is, the exactness is the extent of genuine outcomes (both genuine up-sides and genuine negatives) among the all-out number of cases analyzed in the investigation. Eq. (1) is utilized to measure the quantify binary accuracy:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

where, TP = True positive, FP = False positive, TN = True negative, FN = False-negative. Other execution measures are "responsiveness" and "explicitness", likewise referred to in insights as an order work, which are broadly utilized in medication and bioinformatics studies.

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (3)$$

		Predicted Class	
		TP	FP
True Class	TP	TP	FP
	FN	FN	TN

Fig. 16. Confusion Matrix

4.3.2 RMSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (4)$$

RMSE is an indicator used to determine the error rate between the measured values and the values estimated by a model. Therefore, the closer RMSE approaches zero the more it indicates that the predictive power of the model has increased.

4.4 Comparison analysis

Table 4: Comparison of the existing algorithm with the proposed work in terms of used different datasets, selected features, accuracy, and error rate

Technique	Dataset	Features	Selected features	Accuracy rate (%)	Error rate (%)
IG + KNN [10]	Dataset1	56	28	71	65
IG + KNN [10]	Dataset3	80	20	58.33	70
Hyper-Heuristic Algorithm +KNN [21]	Dataset1	57	11	80.63	-
Cuckoo Search Algorithm	Dataset2	56	Reduced	75.8	-
Chi-square [41]	Dataset2	23	10	92	-
Anova [42]	Dataset2	23	15	70	-
Proposed Method GA+SVM	Dataset1	56	6	84.6	22.9
GA+KNN	Dataset1	56	4	94.7	39.2

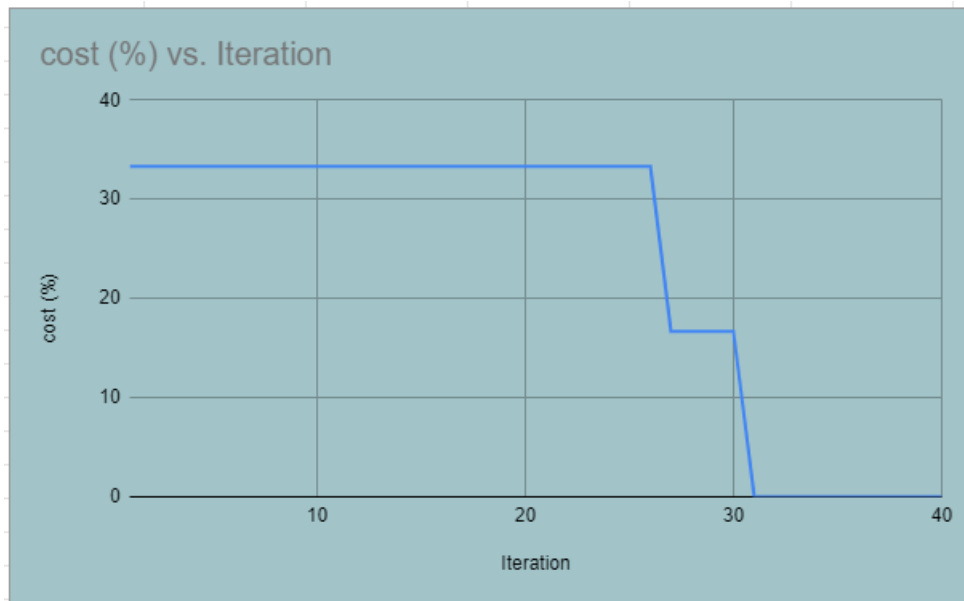


Fig. 17. The best cost function (for SVM is 0)

CHAPTER 5

CONCLUSION

Image processing encompasses a wide range of application areas that require a unique approach to describing raw images utilizing features that not only minimize dimensionality but also storage difficulties. This paper combines the most widely used feature selection techniques from a variety of domains and human biological models in one place. Furthermore, systems with larger databases and a wide range of processing features require selection algorithms that aid in the performance of machine learning algorithms. Every feature selection approach aims to reduce the number of features, and the proposed method does just that. In this method, KNN, and SVM machine learning algorithm is used with a Genetic algorithm to predict lung cancer. The best 4 and 6 out of 57 features are selected by using GA in the KNN and SVM respectively. The best results produced by KNN in terms of features selection, high accuracy i.e., 94.7%, zero error rate, and the reduced cost function value at the starting generation (approx., 100). The run time of this program which is done on Matlab 2022a is 180secs.

There is a trade-off between generation and cost, selected features and generation, generation and time. Some existing methods compare it with the proposed work. It found that other methods are produced with less generation but their costs and period are maximum. In the proposed work, there is more generation but accuracy, time, and costs are getting reduced.

REFERENCES

- [1] World Health Organization website: <https://www.who.int/en/news-room/fact-sheets/detail/cancer>
- [2] Z. Pang, D. Zhu, D. Chen, L. Li, and Y. Shao, "A computer-aided diagnosis system for dynamic contrast-enhanced MRI images based on level set segmentation and relief feature selection," *Computational and mathematical methods in medicine*, vol. 2015, no. 2015, 2015.
- [3] M. Gutkin, R. Shamir, and G. Dror, "SlimPLS: A method for feature selection in gene expression-based disease classification," *PLoS ONE*, vol. 4, no. 7, p. e6416, Jul. 2009
- [4] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel unsupervised feature filtering of biological data," *Bioinf.*, vol. 22, no. 14, pp. e507–e513, Jul. 2006
- [4] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel unsupervised feature filtering of biological data," *Bioinf.*, vol. 22, no. 14, pp. e507–e513, Jul. 2006
- [5] P. Krizek, "Feature selection: Stability, algorithms, and evaluation," Ph.D. dissertation, Dept. Cybern., Faculty of Elect. Eng., Czech Technical Univ., Praha, Czech Republic, 2008.
- [6] Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021, March). Prediction and classification of lung cancer using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1099, No. 1, p. 012059). IOP Publishing.
- [7] A. A. Raweh, M. Nassef and A. Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," in *IEEE Access*, vol. 6, pp. 15212-15223, 2018, DOI: 10.1109/ACCESS.2018.2812734.
- [8] https://en.wikipedia.org/wiki/Co-occurrence_matrix
- [9] Morgado, J, Pereira, T., Silva, F., Freitas, C., Negrão, E., de Lima, B. F & Oliveira, H. P. (2021). Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. *Applied Sciences*, 11(7), 3273.
- [10] Gultepe, Y. A. S. E. M. İ. N. (2021). Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms. *CMC-COMPUTERS MATERIALS & CONTINUA*, 67(2).
- [11] W. Abbas, K. B. Khan, M. Aqeel, M. A. Azam, M. H. Ghouri and F. H. Jaskani, "Lungs Nodule Cancer Detection Using Statistical Techniques," 2020 IEEE 23rd International Multitopic Conference (INMIC), 2020, pp. 1-6, DOI: 10.1109/INMIC50486.2020.9318181.
- [12] M. S. Munir Prince, A. Hasan and F. M. Shah, "An Efficient Ensemble Method for Cancer Detection," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-6, DOI: 10.1109/ICASERT.2019.8934817.
- [13] Sannasi Chakravarthy, S. R., & Rajaguru, H. (2019). Lung cancer detection using a probabilistic neural network with the modified crow-search algorithm. *Asian Pacific journal of cancer prevention: APJCP*, 20(7), 2159.
- [14] J. Liu, L. Liu, Y. Ma, K. Xue, Z. Zhou, and M. Zhang, "In Non-Small Cell Lung Cancer, Can Radiomic Features Predict EGFR Mutations?," 2019 IEEE International Conference

- on Mechatronics and Automation (ICMA), 2019, pp. 2180-2184, DOI: 10.1109/ICMA.2019.8815966.
- [15] M. Mukherjee and P. K. Biswal, "Segmentation of lungs nodules by iterative thresholding method and classification with Reduced Features," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 450-455, DOI: 10.1109/ICICCT.2018.8473287.
- [16] M. Kurkure and A. Thakare, "Lung cancer detection using Genetic approach," 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), 2016, pp. 1-5, DOI: 10.1109/ICCUBEA.2016.7860007.
- [17] Senthil, S., & Shubha, B. A. (2019). Improving the performance of lung cancer detection at an earlier stage and prediction of reoccurrence using the neural networks and ant lion optimizer. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2), 6378-6391.
- [18] Delzell, D. A., Magnuson, S., Peter, T., Smith, M., & Smith, B. J. (2019). Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Frontiers in oncology*, 1393.
- [19] Senthil Kumar, K., Venkatalakshmi, K., & Karthikeyan, K. (2019). Lung cancer detection using image segmentation using various evolutionary algorithms. *Computational and mathematical methods in medicine*, 2019.
- [20] Prince, M. S. M., Hasan, A., & Shah, F. M. (2019, May). An efficient ensemble method for cancer detection. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (pp. 1-6). IEEE.
- [21] Montazeri, M., Baghshah, M. S., & Enhesari, A. (2015). Hyper-Heuristic algorithm for finding efficient features in the diagnosis of lung cancer disease. *arXiv preprint arXiv:1512.04652*.
- [22] M. K. Sohrabi and A. Tajik, "Multi-objective feature selection for Warfarin dose prediction," *Comput. Biol. Chem.*, vol. 69, pp. 126–133, Aug. 2017, DOI: 10.1016/j.compbiolchem.2017.06.002.
- [23] A. Sahoo and S. Chandra, "Multi-objective grey wolf optimizer for improved cervix lesion classification," *Appl. Soft Comput.*, vol. 52, pp. 64–80, Mar. 2017, DOI: 10.1016/j.asoc.2016.12.022.
- [24] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, H. Alhussian, M. G. Ragab, and A. Alqushaibi, "Binary multi-objective grey wolf optimizer for feature selection in classification," *IEEE Access*, vol. 8, pp. 106247–106263, 2020.
- [25] H. Liu and Z. Zhao, "Manipulating data and dimension reduction methods: Feature selection," in *Computational Complexity: Theory, Techniques, and Applications*. New York, NY, USA: Springer, 2012, pp. 1790–1800.
- [26] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever-evolving frontier in data mining," in *Proc. Feature Selection Data Mining*, 2010, pp. 4–13.
- [27] L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina, "A preference-based evolutionary algorithm for multi-objective optimization," *Evol. Comput.*, vol. 17, no. 3, pp. 411–436, Sep. 2009.
- [28] A. Abraham and L. Jain, "Evolutionary multiobjective optimization," in *Evolutionary Multiobjective Optimization*. London, U.K.: Springer, 2005, pp. 1–6.

- [29] L. T. Bui, Ed., *Multi-Objective Optimization in Computational Intelligence: Theory and Practice*. Hershey, PA, USA: IGI Global. 2008.
- [30] Y. Zhai, Y.-S. Ong, and I. W. Tsang, “The emerging’ big dimensionality,” *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 14–26, Aug. 2014.
- [31] W. Siedlecki and J. Sklansky, “A note on genetic algorithms for largescale feature selection,” in *Handbook of Pattern Recognition and Computer Vision*. Singapore: World Scientific, 1993, pp. 88–107.
- [32] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and G. Prisma, “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” *Ann. Internal Med.*, vol. 151, no. 4, pp. 264–269, 2009.
- [33] Y. Zhang, D.-W. Gong, X.-Z. Gao, T. Tian, and X.-Y. Sun, “Binary differential evolution with self-learning for multi-objective feature selection,” *Inf. Sci.*, vol. 507, pp. 67–85, Jan. 2020.
- [34] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, “A multi-objective approach for profit-driven feature selection in credit scoring,” *Decis. Support Syst.*, vol. 120, pp. 106–117, May 2019, DOI: 10.1016/j.dss.2019.03.011.
- [35] J. González, J. Ortega, M. Damas, P. Martín-Smith, and J. Q. Gan, “A new multi-objective wrapper method for feature selection—Accuracy and stability analysis for BCI,” *Neurocomputing*, vol. 333, pp. 407–418, Mar. 2019, DOI: 10.1016/j.neucom.2019.01.017.
- [36] A. Sharma and R. Rani, “C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods,” *Comput. Methods Programs Biomed.*, vol. 178, pp. 219–235, Sep. 2019, DOI: 10.1016/j.cmpb.2019.06.029.
- [37] H. E. Kiziloğlu, A. Deniz, T. Dokeroglu, and A. Cosar, “Novel multiobjective TLBO algorithms for the feature subset selection problem,” *Neurocomputing*, vol. 306, pp. 94–107, Sep. 2018, DOI: 10.1016/j.neucom.2018.04.020.
- [38] M. Amoozegar and B. Minaei-Bidgoli, “Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism,” *Expert Syst. Appl.*, vol. 113, pp. 499–514, Dec. 2018, DOI: 10.1016/j.eswa.2018.07.013.
- [39] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay, “Pareto front feature selection based on artificial bee colony optimization,” *Inf. Sci.*, vol. 422, pp. 462–479, Jan. 2018, DOI: 10.1016/j.ins.2017.09.028.
- [40] M. Dashtban, M. Balafar, and P. Suravajhala, “Gene selection for tumor classification using a novel bio-inspired multi-objective approach,” *Genomics*, vol. 110, no. 1, pp. 10–17, Jan. 2018, DOI: 10.1016/j.ygeno.2017.07.010.
- [41] Almutiri, T., & Saeed, F. (2019, December). Chi-square and support vector machine with recursive feature elimination for gene expression data classification. In *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)* (pp. 1-6). IEEE.
- [42] Fatimaezzahra, M., Mohamed, S., & Abdelaziz, E. (2017). ‘A combined cuckoo search algorithm and genetic algorithm for parameter optimization in computer vision. *Int. J. Appl. Eng. Res*, 51, 12940-12954.