GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM

B. Tech. Major Project Report

BY

KAPPETA SRINIVASA REDDY (14115042)

KAVETI ANIRUDH REDDY (14115046)

TAGGI KISHORE (14115089)

DEPARTMENT OF COMPUTER SC. & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

RAIPUR, CG (INDIA)

MAY, 2018

GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM

Major Project Report

Submitted in partial fulfillment of the

requirements for the award of the degree

Of

Bachelor of Technology

In

COMPUTER SCIENCE AND ENGINEERING

BY

KAPPETA SRINIVASA REDDY (14115042)

KAVETI ANIRUDH REDDY (14115046)

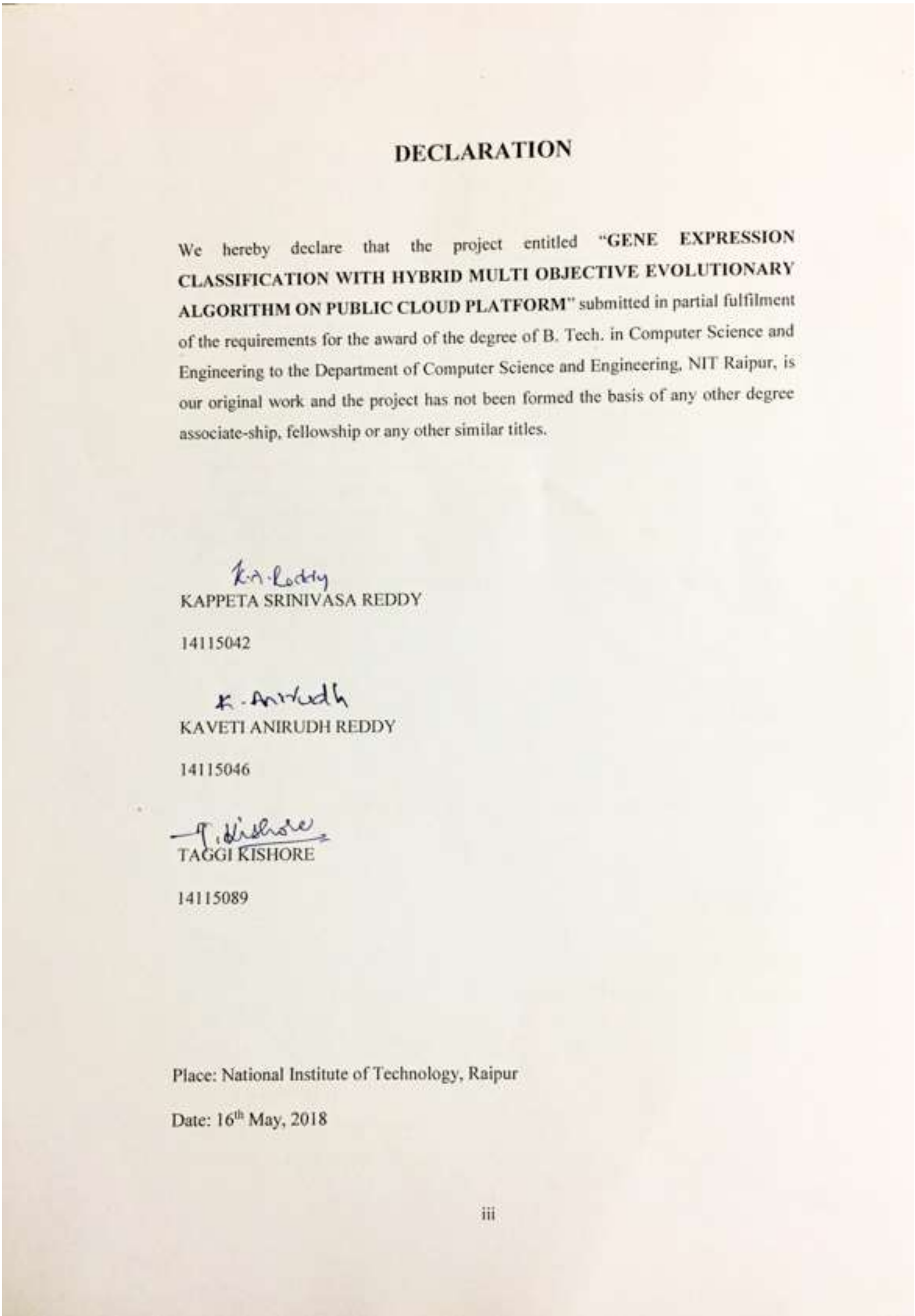
TAGGI KISHORE (14115089)

DEPARTMENT OF COMPUTER SC. & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

RAIPUR, CG (INDIA)

MAY, 2018



DECLARATION

We hereby declare that the project entitled **"GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM"** submitted in partial fulfilment of the requirements for the award of the degree of B. Tech. in Computer Science and Engineering to the Department of Computer Science and Engineering, NIT Raipur, is our original work and the project has not been formed the basis of any other degree associate-ship, fellowship or any other similar titles.

K.A.Reddy
KAPPETA SRINIVASA REDDY

14115042

K. Anirudh
KAVETI ANIRUDH REDDY

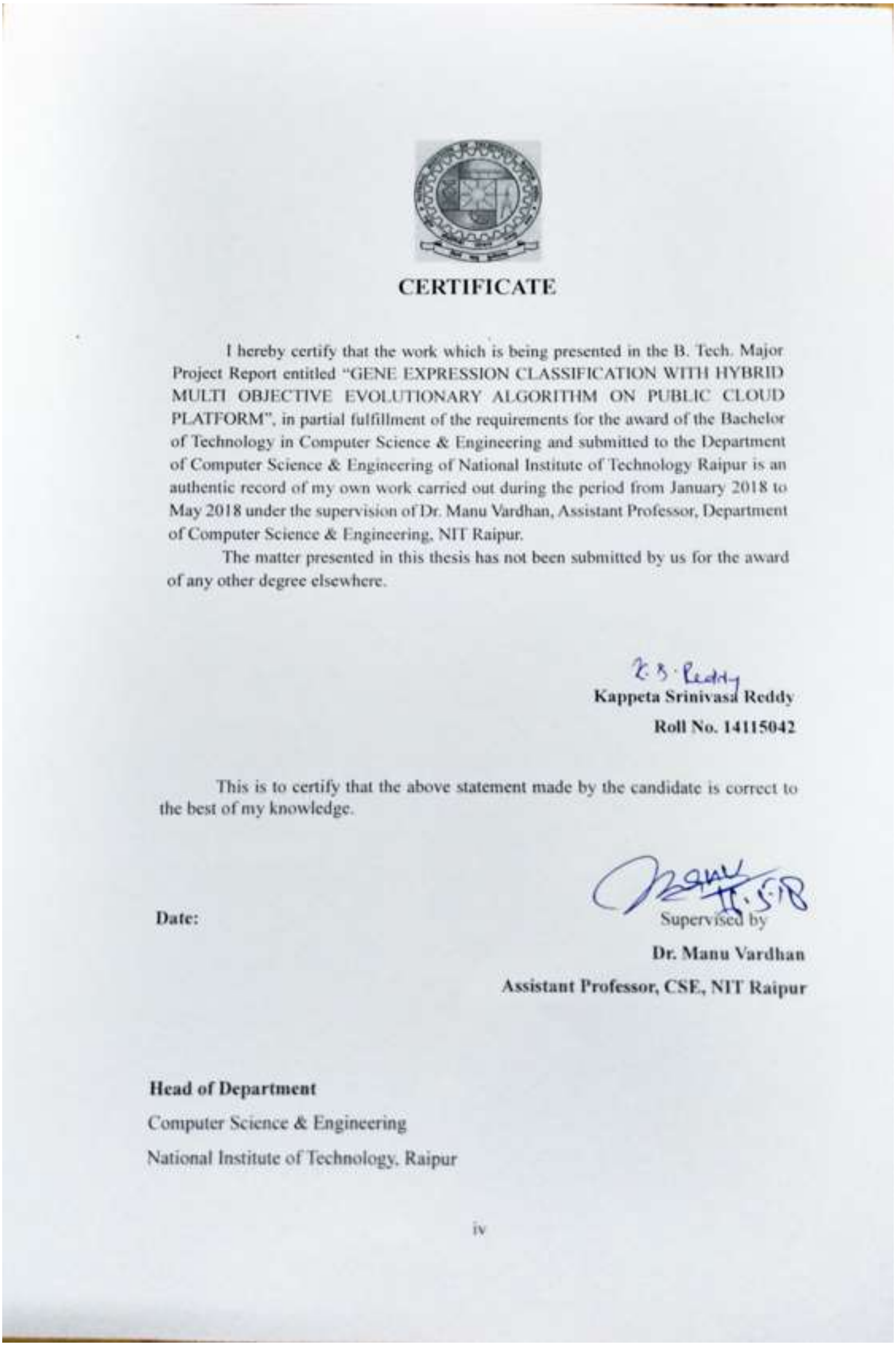
14115046

T. Kishore
TAGGI KISHORE

14115089

Place: National Institute of Technology, Raipur

Date: 16th May, 2018



CERTIFICATE

I hereby certify that the work which is being presented in the B. Tech. Major Project Report entitled "GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM", in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Science & Engineering and submitted to the Department of Computer Science & Engineering of National Institute of Technology Raipur is an authentic record of my own work carried out during the period from January 2018 to May 2018 under the supervision of Dr. Manu Vardhan, Assistant Professor, Department of Computer Science & Engineering, NIT Raipur.

The matter presented in this thesis has not been submitted by us for the award of any other degree elsewhere.

K.S. Reddy
Kappeta Srinivasa Reddy
Roll No. 14115042

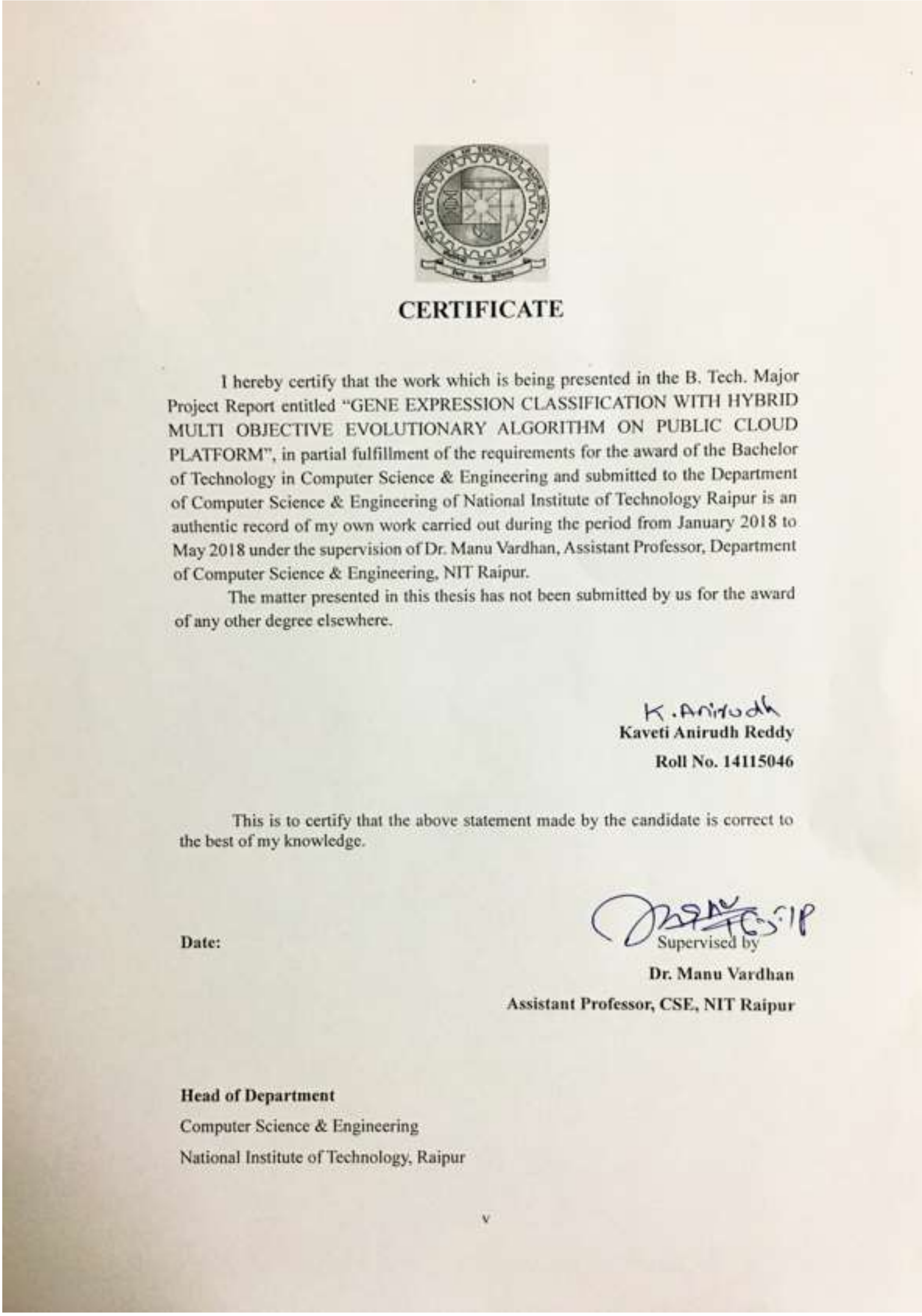
This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

11.5.18

Date:

Supervised by
Dr. Manu Vardhan
Assistant Professor, CSE, NIT Raipur

Head of Department
Computer Science & Engineering
National Institute of Technology, Raipur



CERTIFICATE

I hereby certify that the work which is being presented in the B. Tech. Major Project Report entitled "GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM", in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Science & Engineering and submitted to the Department of Computer Science & Engineering of National Institute of Technology Raipur is an authentic record of my own work carried out during the period from January 2018 to May 2018 under the supervision of Dr. Manu Vardhan, Assistant Professor, Department of Computer Science & Engineering, NIT Raipur.

The matter presented in this thesis has not been submitted by us for the award of any other degree elsewhere.

K.Anirudh
Kaveti Anirudh Reddy
Roll No. 14115046

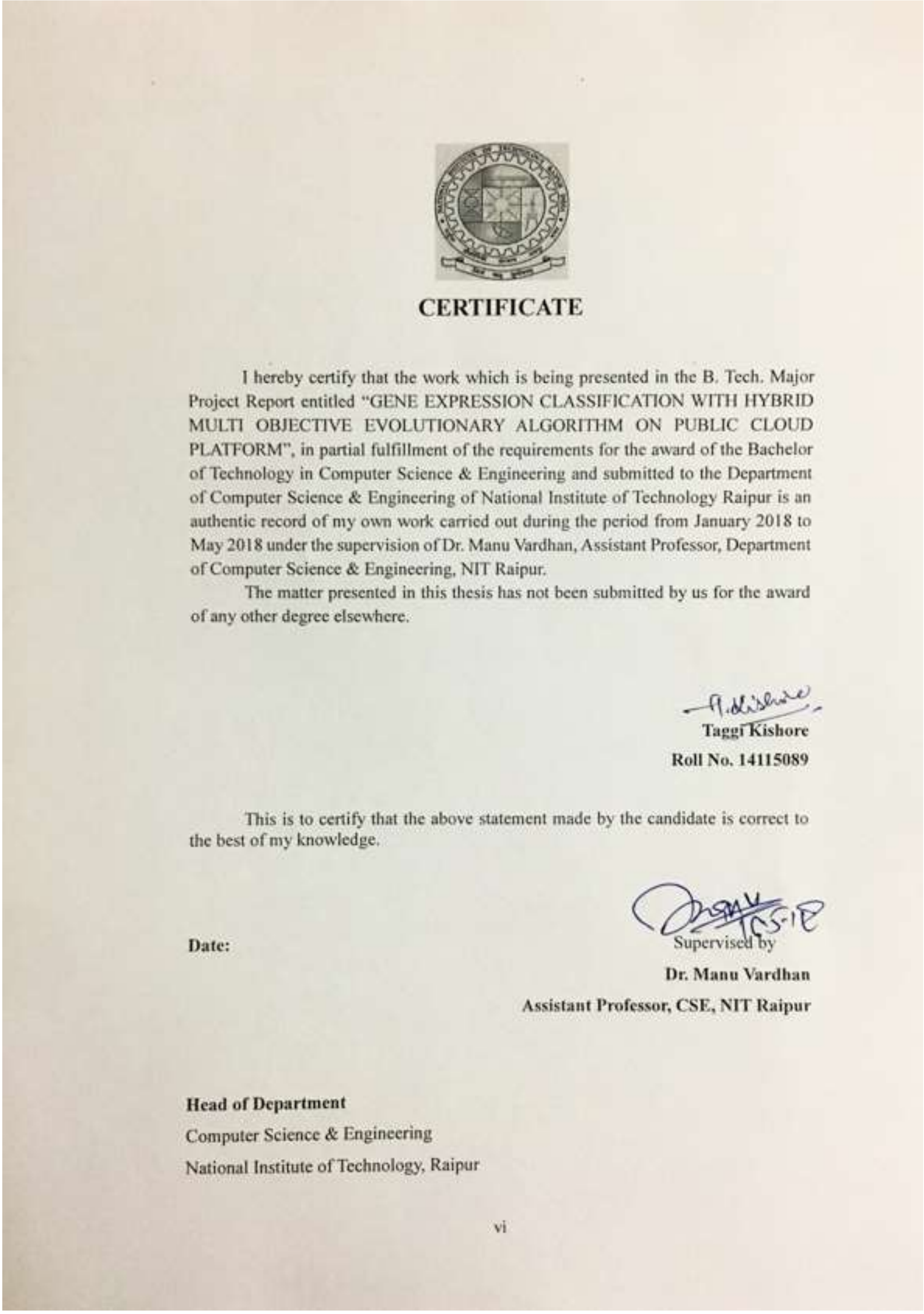
This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Supervised by
Dr. Manu Vardhan
Assistant Professor, CSE, NIT Raipur

Head of Department
Computer Science & Engineering
National Institute of Technology, Raipur

v



CERTIFICATE

I hereby certify that the work which is being presented in the B. Tech. Major Project Report entitled "GENE EXPRESSION CLASSIFICATION WITH HYBRID MULTI OBJECTIVE EVOLUTIONARY ALGORITHM ON PUBLIC CLOUD PLATFORM", in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Science & Engineering and submitted to the Department of Computer Science & Engineering of National Institute of Technology Raipur is an authentic record of my own work carried out during the period from January 2018 to May 2018 under the supervision of Dr. Manu Vardhan, Assistant Professor, Department of Computer Science & Engineering, NIT Raipur.

The matter presented in this thesis has not been submitted by us for the award of any other degree elsewhere.

Taggi Kishore
Roll No. 14115089

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Supervised by
Dr. Manu Vardhan
Assistant Professor, CSE, NIT Raipur

Head of Department
Computer Science & Engineering
National Institute of Technology, Raipur

ABSTRACT

Gene expression data helps in advancement of clinical study. The sample dimension of gene expression data is very less than gene dimension of gene expression data. This mismatch in data causes difficult to use this data for classification and prediction. Only small number of genes are strongly correlated in this data. This causes the need for selection for features. We use different multi objective optimization algorithms like NSGA-II, MOPSO and hybrid evolutionary algorithm to select features. This algorithms optimize multiple objective function and generate best solutions.

We used feature selection algorithms like F score, ReliefF etc. to select 60 genes initially and this genes are used to get better solutions. The objective functions used are number of features and AUC (Area Under Curve) score. AUC score is used because small datasets contain class imbalances which make AUC score as best measure to indicate performance of classifier used.

NSGA-II uses non-dominated sort to obtain new solutions and MOPSO uses distance calculation between present solution and global solution to generate new solutions. We use both this concepts along with crossover operators of evolutionary algorithms to implement hybrid algorithm, which generates optimal solutions in less generations. We also ran this algorithms on cloud to reduce the computing time of these algorithms.

ACKNOWLEDGEMENT

We would like to acknowledge our college National Institute of Technology, Raipur for providing a holistic environment that nurtures creativity and research based activities and we would also like to thank Dr. A. M. Rawani, Director, National Institute of Technology, Raipur for the same.

We express our sincere thanks to the head of Computer Science and Engineering branch Dr. Dilip Singh Sisodia for providing all the platform to implement the project. We also like to thank Dr. Manu Vardhan Sir, the supervisor of our project for guiding and correcting throughout the process with attention and care. He frequently suggested us creative ideas and guided through the major hurdles that occurred during the duration of the project. We also thank him for his constant feedback and expertise that paved a way forward for us.

We would also thank our institution and faculty members without whom this project would have been a distant reality. We also extend our heartfelt thanks to our family and friends who supported us.

Kappeta Srinivas Reddy (14115042)

Kaveti Anirudh Reddy (14115046)

Taggi Kishore (14115089)

LIST OF TABLES

LIST OF FIGURES

LIST OF IMAGES

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

1.1 Biological Introduction to Gene Data

Gene Data [1] is the natural biological information of life sciences data. It is collected from various scientific experiments and from the literature published, through technology having high throughput by doing experiments & computational analysis[2]. The biological data consists gene function, its structure, its localisation, its clinical effects by performing mutations, and similarities of biological structures and its sequences. Our database is a microarray database containing microarray gene expression data[3]. There are many uses of microarray database but one of the key use is its measurement data is stored, manages a search index and the data will be made available to other applications to be used for analysis[4] and interpretation. In this project, we used datasets like SRBCT, Prostate Cancer and Lung Cancer. We will look into the datasets more deeply in the section 3.1.

1.2 Introduction to Feature Selection

Feature selection[5] is a strategy to diminish dimensionality which is utilized as a part of the machine learning field, design acknowledgment, insights, and mining information, ordinarily. This approach targets to pick out a subset of important features [6] from the set of distinct features based on some condition. A few methods to select features are information gain, reliefF, chi squares, fisher score, and lasso. Selection of some features is normally utilized as a part of the areas where the datasets include a huge number of features however with generally little example measure (e.g., gene data). Gene selection is the process of feature selection applied to the gene[7] data. Selection of Gene is vital as the information generally contains numerous superfluous, irrelevant, repetitive, uproarious [8], and also for noisy articulations. It is very effective for detecting tumour earlier and discovering cancer, as it will prompt a more solid disease finding and better medical treatment [9].

The gene data could be fully labelled, unlabelled, or could be partially labelled[10]. Unlabelled information is made out of samples. Unlabelled data has features but no labels will be present which will indicate any of the explanation or info. Labelled data is a data set which will be inked with some meaningful labels or may be with classes. Supervised selection of features[11] is the procedure of choosing a feature subset in light of a few criteria for estimating significance and pertinence of the features by using labelled data.

1.2.1 Feature Selection Overview

Selection of features objective is choosing a feature subset from an original set of features in view of the feature's pertinence and repetition. Yu and Liu[10] has classified the subset of features. They are of four classes. They are (a) completely irrelevant and noisy features, (b) weakly relevant and redundant features, (c) weakly relevant and non-redundant features, and (d) strongly relevant features. Every feature in the class (c) and (d) are contained in the optimal subset[12]. Feature which are strongly relevant are absolutely vital for intensification of discriminative power and accuracy prediction.

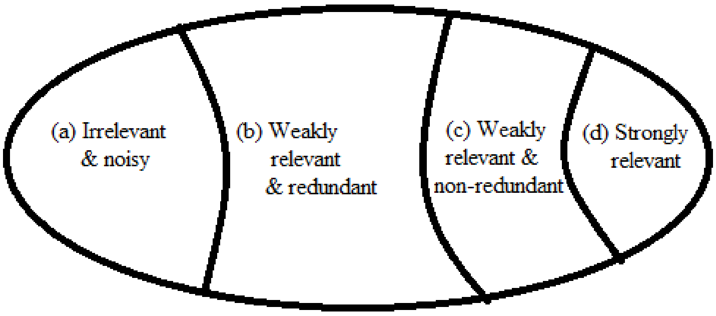

(a) Irrelevant & noisy
(b) Weakly relevant & redundant
(c) Weakly relevant & non-redundant
(d) Strongly relevant

Fig. 1. Classification of features based on relevancy and redundancy

If a feature is not a repetitive and also compatible to the assessment measures then feebly features having significance can be valuable in enhancing prediction accuracy. A feature which is irrelevant demonstrates that the feature decrease accuracy's prediction. Therefore, all irrelevant, excess or uproarious/noisy features ought to be eliminated and all the strong features which are important and some feebly features having significance ought to be chosen to fabricate a decent model prediction.

1.2.2 Process of Feature Selection

The procedure of selecting features in which choosing a subset having significant and informative features from the entire dataset of features. It is partitioned into five stages.

a) Determining Search Direction
b) Determining Search Strategy
c) Determining Evaluation Criterion

d) Defining Stopping Criteria

e) Validating the result

In determining the evaluation criteria, there are four types of evaluation methods of feature selection. They are filter, wrapper, embedded, and hybrid. In this project, we used filter and hybrid methods.

Filter[13],is the earliest method in feature selection process, which is also known as open-loop method. A procedural measure is connected to give a score to every feature by filter feature selection methods. The features are positioned with the help of score and either chose to be kept or expelled from the dataset. These all are univariate methods[7]. The features are considered independently, or they are considered with respect to the dependent variable. Feature characteristics are measured by filter algorithms based on evaluation criteria which are of four types. They are dependency, information, distance and consistency. They are known for their efficiency. Therefore, they are more easily scalable to huge databases[12]. Gaining of information can be another example.

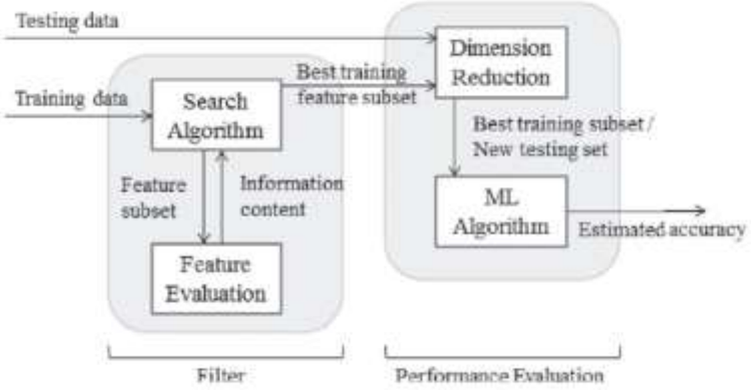


Image 1. Taxonomy of feature selection method- filter

Hybrid[14], is the latest development in the feature selection. It is formed by combining two methods, two strategies of same criteria or two selecting of features methodologies. Hybrid has the benefits of both the strategies by joining all their strengths[15].

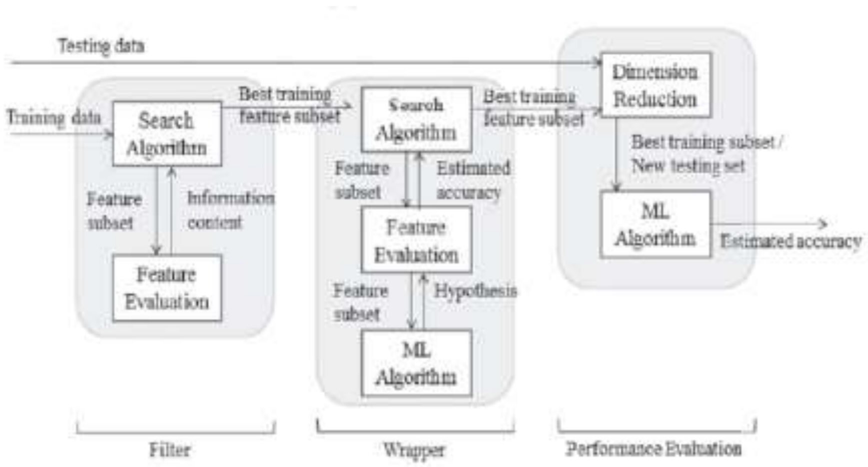


Image 2. Taxonomy of feature selection method- hybrid

In validating the result stage, different estimations of error or validating system techniques were proposed for assessing the feature sets effectiveness to be used for grouping and predicting. The common method for error estimation is Cross Validation (CV)[16]. In this project, we discuss the usage of CV method. Cross Validation is the most widely recognized and well known strategy for validation. In this strategy, the dataset is part into two sections: set for training and set for testing. Now, the classifier is trained using the set i.e., training and final evaluation is done using testing set. CV delivers a viable error estimate which is unbiased is a greater advantage using CV. This favourable position is a result of the rehashed methodology of CV for various examples taken from a populace, the average estimation of error will rough the error that is expected for the outlined classifiers over all the conceivable equivalent size examples[17]. Cross Validation has highly variable disadvantage which is its error estimation. In this project, we made use of *k*-fold CV method. The average rate of error on subsamples for test (E_i) is the formula for calculating CV error rate.

Formulae is given as $E = (1/k) \sum_{i=1}^{K} Ei$

1.2.3 Advantages of Feature Selection

The key benefits of performing feature selection on the data are:

- **Reduces Over fitting**: Having less redundant data provides less opportunity to make decisions on noise.
- **Improves Accuracy**: Having less data misleading improves the modelling accuracy.

- **Reduces Training Time**: Having less data gives the algorithms the opportunity to train faster.
- Enhances prediction performance, understand ability, versatility, and speculation capacity of the classifier.
- It likewise decreases computational complexity and capacity, gives quicker and more practical model, and assumes an essential part of the knowledge discovery.

1.3 Types of Crossover and Mutation

- **Crossover:** Crossover is process of selecting two parents and generating offspring's. Generally crossover is applied with high probability (Pc).
  - **Different crossover techniques:**
    - **One Point Crossover:** In this crossover, the parts of parents are swapped after selecting a random crossover point.

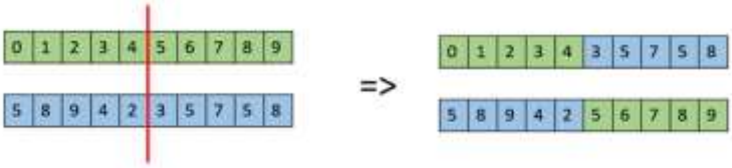


Image 3. One Point Crossover

    - **Multi Point Crossover:** In this crossover parents are divided into segments and are swapped to get the offspring's.

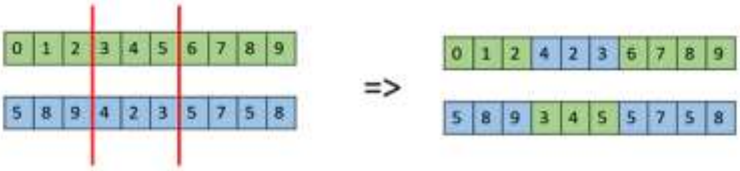


Image 4. Multi Point Crossover

    - **Uniform Crossover:** In a uniform crossover, we use the genes in the chromosome separately. We use genes taken from both parents to obtain the solution. We can also make a biased solution to get solution biased to one parent.

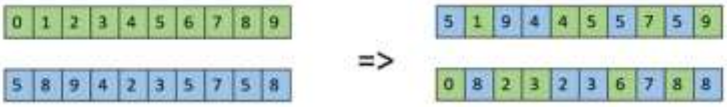


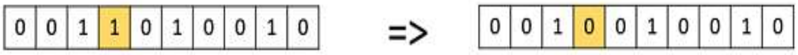
Image 5. Uniform Crossover

- **Half Uniform Crossover:** It is same as uniform crossover, but not all bits are exchanged. First number of unmatched bits between parents are calculated and only half of them are exchanged between parents to generate offsprings.

- **Mutation:** Making slight changes in the present solution to obtain a new set of solution is called Mutation. Mutation is used to obtain diverse solutions to the population with probability, p_m. It has been watched that mutation is fundamental to the merging of the GA while crossover is not.

➢ **Mutation Types:**

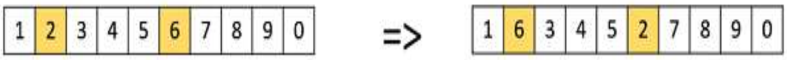
- **Bit Flip Mutation:-**This type of mutation is used in binary encoded genetic algorithm. In this method, we select bits randomly and flip them.



0 0 1 1 0 1 0 0 1 0 => 0 0 1 0 0 1 0 0 1 0

Image 6. Bit Flip Mutation

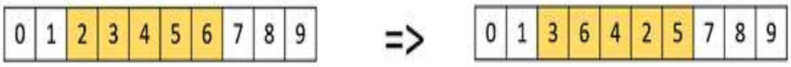
- **Random Resetting:** Augmentation of the bit flip for the representation of an integer is Random Resetting. In this, a random value in a certain range is assigned to a gene chosen randomly.
- **Swap Mutation:** In swap mutation, we select two positions on the chromosome randomly, and exchange the qualities/values.



1 2 3 4 5 6 7 8 9 0 => 1 6 3 4 5 2 7 8 9 0

Image 7. Swap Mutation

- **Scramble Mutation:** From the entire set of chromosome, values of subset of genes are scrambled randomly.



0 1 2 3 4 5 6 7 8 9 => 0 1 3 6 4 2 5 7 8 9

Image 8. Scramble Mutation

- **Inversion Mutation:** Selects a subset of genes and we replace the selected substring with reverse of subset string.

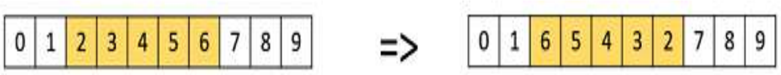


Image 9. Inversion Mutation

1.4 Introduction to NSGA-II Algorithm

Non-Dominated Sorting Genetic Algorithm (NSGA-II)[18] is an algorithm which is developed to solve the [19]problems where multiple functions are needed to be optimized. One of the most widely used algorithm for optimization. This algorithm uses a faster sorting procedure, dominant solution preserving approach.

1.4.1 General Description of NSGA-II:

A random population is taken. Population is sorted using non-dominated sort. The population is divided into parts called fronts[20]. The individuals which are in first front are non-dominant and next front contains the dominated individuals i.e., by first front. Individuals present in first front are assigned a rank as 1 & the individuals contained in the second front are assigned with rank 2 and so on.

Notwithstanding rank, another parameter which is called as crowding distance is figured for every individual. The Crowding Distance gives a measure of how close an individual is to its neighbours. It helps in producing individuals with better diversity.[21] Tournament selection is utilized to choose parents utilizing rank & also crowding distance. An individual with lesser rank is chosen (if rank is same individual with all the more swarming separation is chosen). Offsprings are generated from selected parents using crossover and mutation operator techniques. The new population is again sorted according to non-domination sort and the nPOP best solutions are selected ,where nPOP is size of the population [22].

1.4.2 Detailed Description of NSGA-II:

a) **Population Initialization:**

The population is introduced in light of the range of the problem and assuming constraints any.

b) **Non-Dominated Sort:**

The introduced population which is arranged in light of non-domination. This calculation is superior to anything the first since it use the data about the set which an individual dominate (Sp) and number of people that command the individual (np).

c) Crowding Distance:

After sorting the set using non-domination, the crowding distance[23] is allocated. As the people are selected in view of crowding distance and rank each solution in the population is allocated a value i.e., crowding-distance. Crowding distance is calculated between people in the population present in the same front and contrasting the distance between two people in various front is good for nothing.

Crowing distance's main function is finding the euclidian separation between every individual populace in a front in view of their '*m*' objectives in the '*m*'-dimensional hyperspace[24]. Crowding distance of people in the boundary are assigned infinite distance values.

d) Selection:

Once the individuals are arranged in light of non-domination and the crowding distance is used to complete the selection with the help of comparison-operator ($\prec n$).

Binary Tournament Selection is used for selection of individuals. Binary Tournament Selection (BTS)[20] is utilized, where competition is played between any two solutions and better is chosen and is used for mating. It is conveyed such that each arrangement can be made to take an interest in precisely 2 tournaments.

e) Genetic Operators:

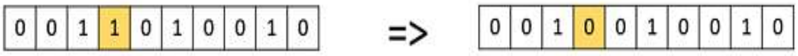
Here we use Half Uniform Crossover (HUX) for crossover operation and Bit Flip Mutation for mutation process.

i. Half Uniform Crossover (HUX):

It is same as uniform crossover, but not all bits are exchanged. First number of unmatched bits between parents are calculated and only half of them are exchanged between parents to generate offsprings.

ii. **Bit Flip Mutation:**

This type of mutation is used in binary encoded genetic algorithm. In this method, we select bits randomly and flip them.



0 0 1 1 0 1 0 0 1 0 => 0 0 1 0 0 1 0 0 1 0

Image 10. Bit Flip Mutation

f) **Recombination and Selection:**

The offspring populace is joined with the present populace and the people of the next generation is selected through selection process. Since the best offsprings are included the populace, elitism is guaranteed. Populace is currently arranged in view of non-domination. The new generation is filled by each front hence until the point when the populace estimate surpasses the present populace size. On the off chance that by including every individual in front Fj the populace surpasses N then people in front Fj are chosen in view of their crowding distance in the descending order until the point that the populace measure is N. What's more, consequently the procedure rehashes to create the subsequent generations.

1.4.3 Multi Objective Optimization Using NSGA-II:

NSGA is a prevalent non-domination based GA for MOO problems [22]. It is an extremely viable algorithm, however, has been for the most part condemned for its computational unpredictability, the absence of elitism and for picking the ideal parameter esteem for parameter sharing. A changed form, NSGA-II [18] was created, which has a superior sorting calculation, fuses elitism and sharing parameter should not be picked as a priori.

Think about the complexity of one cycle of the whole calculation. The fundamental activities and their most pessimistic scenario complexities are as per the following:

1) Non-dominated sorting is $O(M(2N)^2)$;

2) crowding-distance assignment is $O(M(2N)\log(2N))$;

3) sorting on $\prec$ is $O(2N\ log(2N))$;

where, M denotes number of objectives and N denotes population size.

The general complexity nature of the algo is $O(MN^2)$;, which is administered by the non-dominated arranging some portion of the calculation. On the off chance that performed painstakingly, the total populace of 2N size require not be arranged by non-domination. When the arranging methodology has sufficiently discovered number of fronts to have N individuals in Pt+1, there will be no motivation to proceed with the arranging strategy. The decent variety among non-dominated arrangements is presented by utilizing the crowding comparison procedure[21], which is utilized as a part of the tournament selection and amid the populace lessening stage. Since arrangements contend with their crowding distance, no additional niching parameter is required.

1.5 Introduction to MOPSO Algorithm

1.5.1 General Introduction

Various issues experienced, in actuality, can't be really defined as a single objective function[22]. Because of the complexity in such kind of issues effective heuristic strategies were required, firmly fulfilled by Swarm Intelligence (SI) systems. Particle Swarm Optimization (PSO) turned into an exceptionally develop and most well-known space in Swarm Information. MOPSO [18] has turned into a developing field for comprehending MOOs[25] with countless writing, programming, variations, codes and applications.

1.5.2 Detailed Introduction

Swarm Intelligence (SI) is fundamentally characterized as the conduct of common or manufactured self-composed frameworks. Swarms associate locally or with outside operators i.e. condition and could be as winged creature runs, ants, and honey bees and so on. Presented for enhancing constant nonlinear capacities, Particle Swarm Optimization (PSO)[26] characterized another time in Swarm Info. PSO is a populace-based technique for optimization. The number of inhabitants in the potential arrangement is called as swarm and every person in the swarm is characterized as particle. The particles fly in the swarm to look through their best arrangement in light of understanding of their own and alternate particles of a similar swarm. PSO began to hold the grasp among numerous scientists and turned into the most well-known SI technique[27] not long after in the wake of getting presented, yet because of its impediment of improvement just of single target, another idea Multi-Objective PSO (MOPSO)[28] was presented, by which enhancement can be performed for in excess of one clashing goals all the while. MOPSO was proposed to streamline in excess of one objective functions. In MOPSO rather than a solitary arrangement, an arrangement of arrangements are resolved, additionally called pareto optimal set[29]. As vector of objectives are optimized, MOO is now and again called as vector optimization. Multi-target Optimization Problem (MOP)[30]is fundamentally ordered in two ways i.e. Straight and Nonlinear MOP, Convex and Non-Convex MOP. At the point when every target capacity and limitations are straight, at that point Linear MOP is characterized, yet in the event that any of the goal or requirement work is nonlinear, at that point it is a Nonlinear MOP. Moreover if all the target capacities are arched and the attainable district is raised, at that point it is characterized as Convex MOP and for Non-Convex MOP[31] its bad habit a-versa. The applications created are in the region of condition, ventures, work shop planning, designing, science and numerous others.

1.5.3 Multi-Objective Optimization

The principle explanation behind evolutionary algorithms applications in multi-target advancement issues is to get close optimal non-dominated arrangements[32]where decision making bodies can pick a reasonable arrangement. The proficiency of multi-target improvement calculations relies upon the quality and amount of Pareto fronts delivered by them. To analyse diverse Pareto fronts[33] coming about because of various calculations, criteria are considered and connected in multi-objective issues. Every paradigm means a normal for the Pareto front. In this way, positioning methodologies are generally used to assess distinctive algorithms in light of various criteria.

MOP has various objectives and for the most part limitations[34] moreover. The requirements have been fulfilled by any feasible solution (counting the ideal arrangement). MOP[35] is defined as:

$$Max/Min\ f_n(x), n = 1,2,\dots,N;$$

$$subject\ to\ g_j(x) \geq, j = 1,2,\dots,J;$$

$$h_k(x) = 0, k = 1,2,\dots,K;$$

$$x_i^{(L)} \leq x_i \leq x_i^{(U)}, i = 1,2,\dots,m. \qquad (1)$$

An answer x is a vector of m choice factors $x = (x_1, x_2, \dots, x_m)^T$. The principal set of imperatives is inequality limitation for the minimization issue, while for maximization issue this requirement proselytes to less than equals to i.e. ≤. Next arrangement of requirements is the equality constraints took after by the last arrangement of imperatives called variable limits, confining every choice variable x_i to take an incentive inside a lower $x_i^{(L)}$and an upper $x_i^{(U)}$ bound. By and large, to solve the MOPs traditional and Artificial Intelligence (AI) procedures are utilized.

1.5.4 Particle Swarm Optimization

PSO is an AI technique for solving MOPs.

- In PSO parent data is contained inside every particle
- PSO doesn't include an express selection function from its handling
- PSO utilizes an exceptionally directional mutation operation to control individuals
- There is no system for PSO[35] to adjust its speed step size to an esteem proper to the nearby area search space

The principle objective of MOO is to locate a set of solutions which is near the ideal arrangements and sufficiently assorted to speak to the genuine spread of ideal arrangements. MOPSO calculations[36] satisfy both the past said conditions all the more specifically. The effortlessness, low calculation cost and expanding ubiquity of MOPSO upgrade its productivity to explain basic and in addition complex natured genuine issues.

1.5.5 Multi Objective Particle Swarm Optimization

Considering a search space of d-dimension and *n* particles, whose *i^th^* particle at a particular position $X_i(x_{i1}, x_{i2}, \dots, x_{id})$ is moving with a velocity $V_i(v_{i1}, v_{i2}, \dots, v_{id})$. Each particle is associated with its particular best, $P_i(p_{i1}, p_{i2}, \dots, p_{id})$ which is defined by its own best performance[37] in the swarm[38]. Additionally, a general best execution of the molecule as for the swarm characterized global best is gbest.

Each particle tries to modify its position using the following information:

- Current positions,
- Current velocities,
- Distance between the current position and pbest,
- Distance between the current position and gbest.

The movement of the particle is governed by updating its velocity and position attributes.

$$V_i^{t+1} = wV_i^t + c_1r_1\left(x_{pbest} - X_i^t\right) + c_2r_2\left(x_{gbest} - X_i^t\right) \qquad (2)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \qquad (3)$$

where w = inertia weight, c_1= cognitive acceleration coefficient, and c_2= social acceleration coefficient, r_1 and r_2 are the random values between 0 and 1, $x_{pbest}$is the personal best of the particle and x_{gbest} is the global best of the particle[39]. X_i^t is the current position of *i*th particle at iteration *t*. V_i^t is the velocity of *i*th particle at iteration *t*.

Now, the objective function in MOPSO [41]contains multi objectives as formulated in equation 1 [42].

1.5.6 Optimized MOPSO (OMOPSO)

The principle highlights of OMOPSO[32] incorporate the utilization of the crowding distance of NSGA-II[43] to sift through pioneer arrangements and the blend of two change administrators to quicken the merging of the swarm. The first

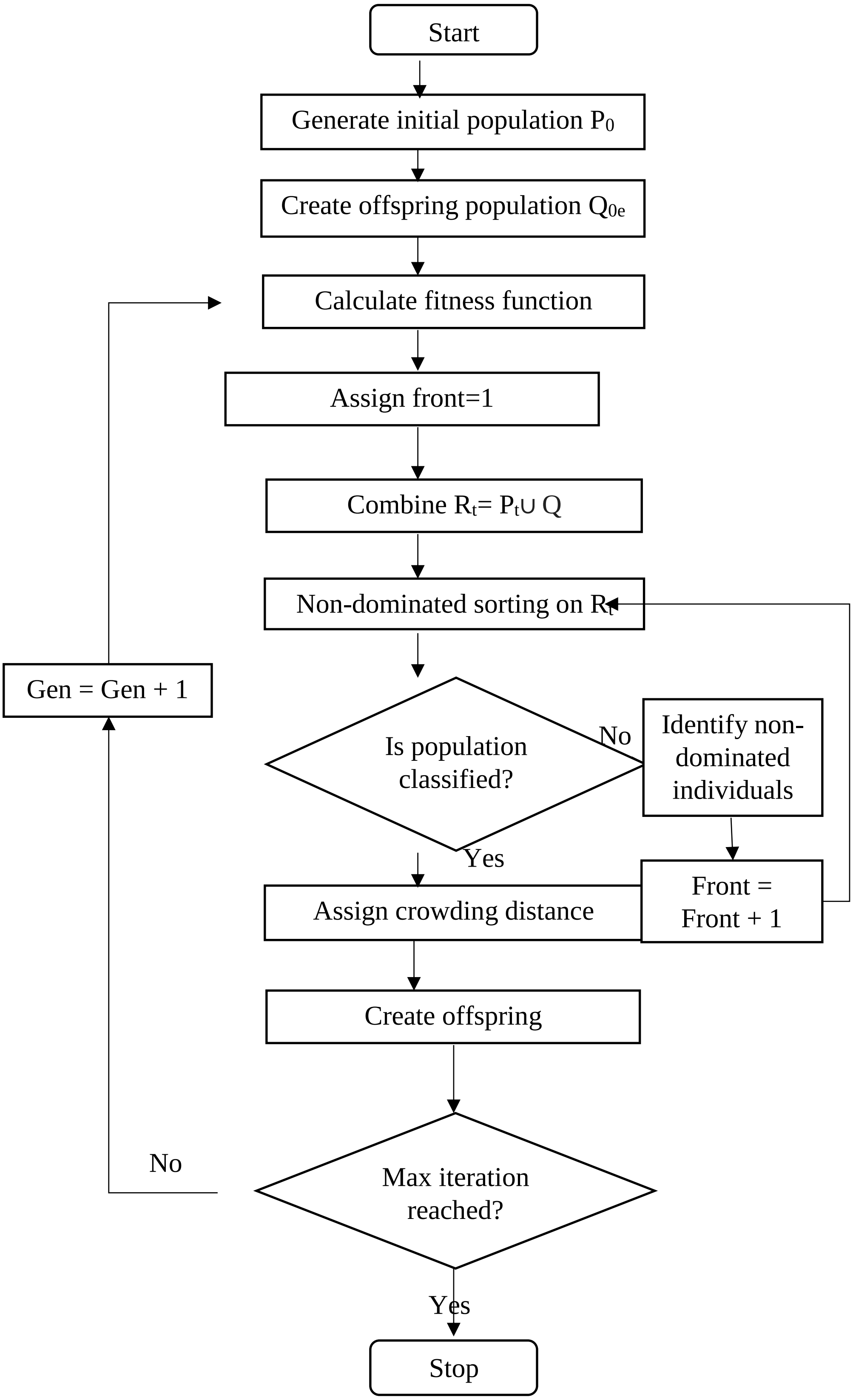
OMOPSO[44] calculation makes utilization of the idea of ϵ-dominance to restrain the arrangements delivered by this algorithm[32]. We study here a variation disposing of the utilization ϵ-dominance[40], being the pioneers chronicle the aftereffect of the execution of the procedure.

CHAPTER-2

METHODOLOGY

2.1 Flow Chart

2.1.1 Flow chart for NSGA-II algorithm implementation



Start
Generate initial population P0
Create offspring population Q0e
Calculate fitness function
Assign front=1
Combine Rt= Pt∪ Q
Non-dominated sorting on Rt
Gen = Gen + 1
Is population classified?
No
Identify non-dominated individuals
Yes
Front = Front + 1
Assign crowding distance
Create offspring
No
Max iteration reached?
Yes
Stop

Fig. 2 Working of NSGA-II

2.1.2 Flow chart for creating Offsprings in NSGA-II algorithm

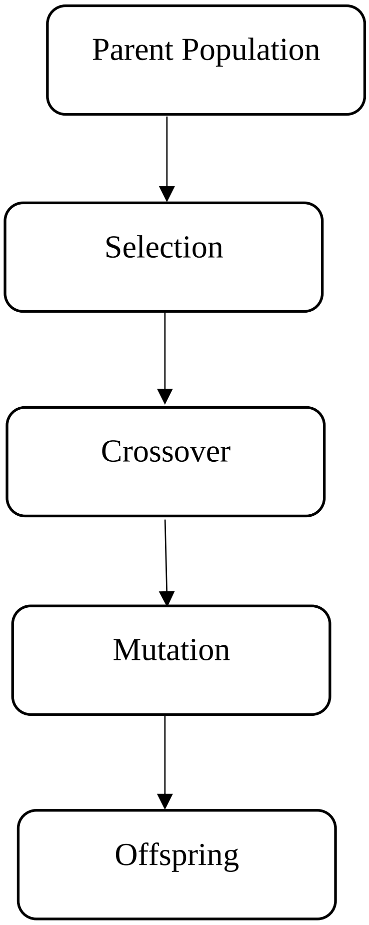


Fig. 3 Creating Offspring

2.1.3 Flow chart for MOPSO algorithm

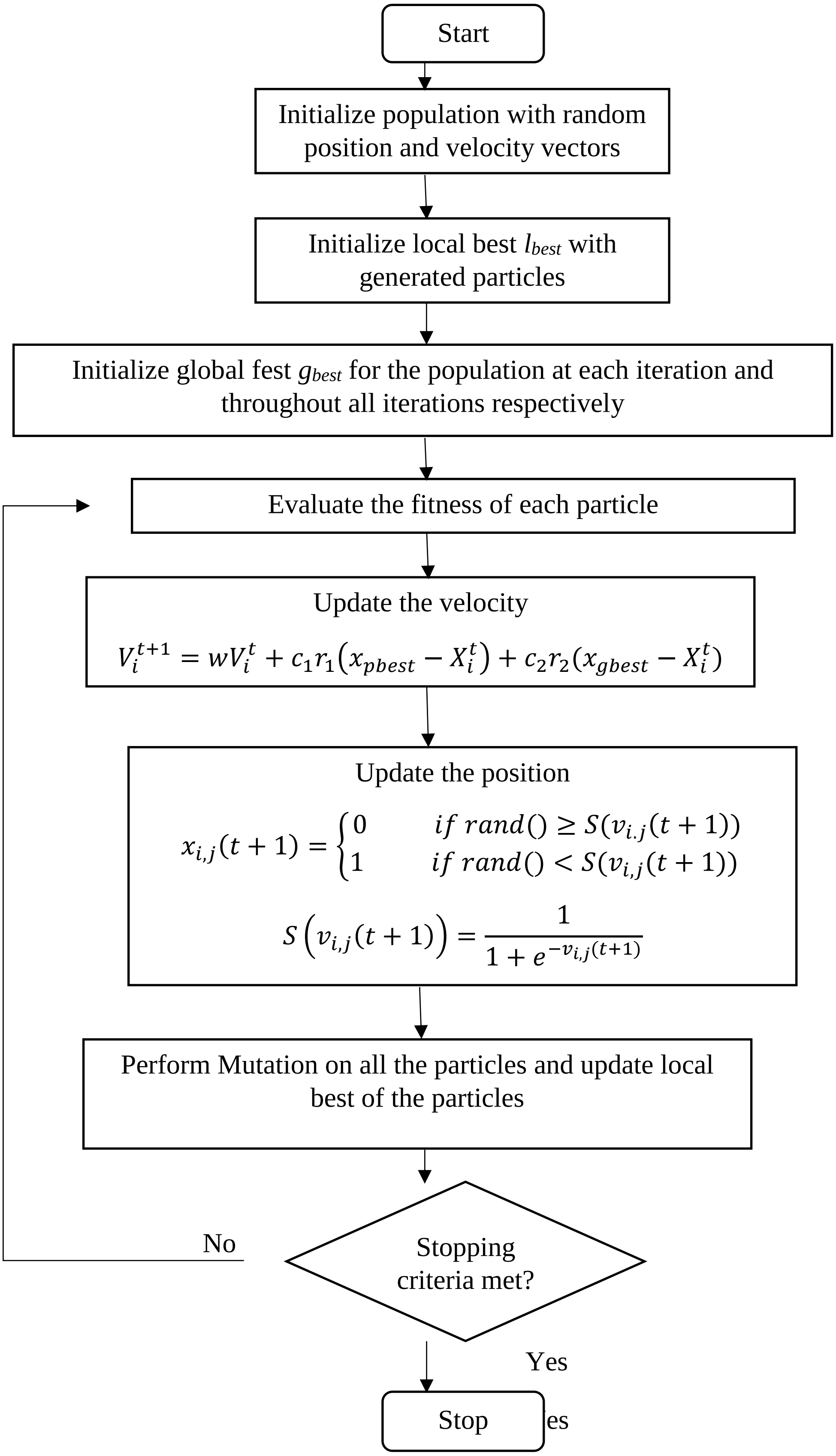

Start
Initialize population with random position and velocity vectors
Initialize local best l_{best} with generated particles
Initialize global fest g_{best} for the population at each iteration and throughout all iterations respectively
Evaluate the fitness of each particle
Update the velocity
$V_i^{t+1} = wV_i^t + c_1 r_1\left(x_{pbest} - X_i^t\right) + c_2 r_2(x_{gbest} - X_i^t)$
Update the position
$x_{i,j}(t+1) = \begin{cases} 0 & if\ rand() \geq S(v_{i.j}(t+1)) \\ 1 & if\ rand() < S(v_{i,j}(t+1)) \end{cases}$
$S\left(v_{i,j}(t+1)\right) = \frac{1}{1+e^{-v_{i,j}(t+1)}}$
Perform Mutation on all the particles and update local best of the particles
Stopping criteria met?
No
Yes
Stop
es

Fig. 4 Flowchart for Multi-objective Particle Swarm Optimization algorithm

2.1.4 Flow chart for Hybrid

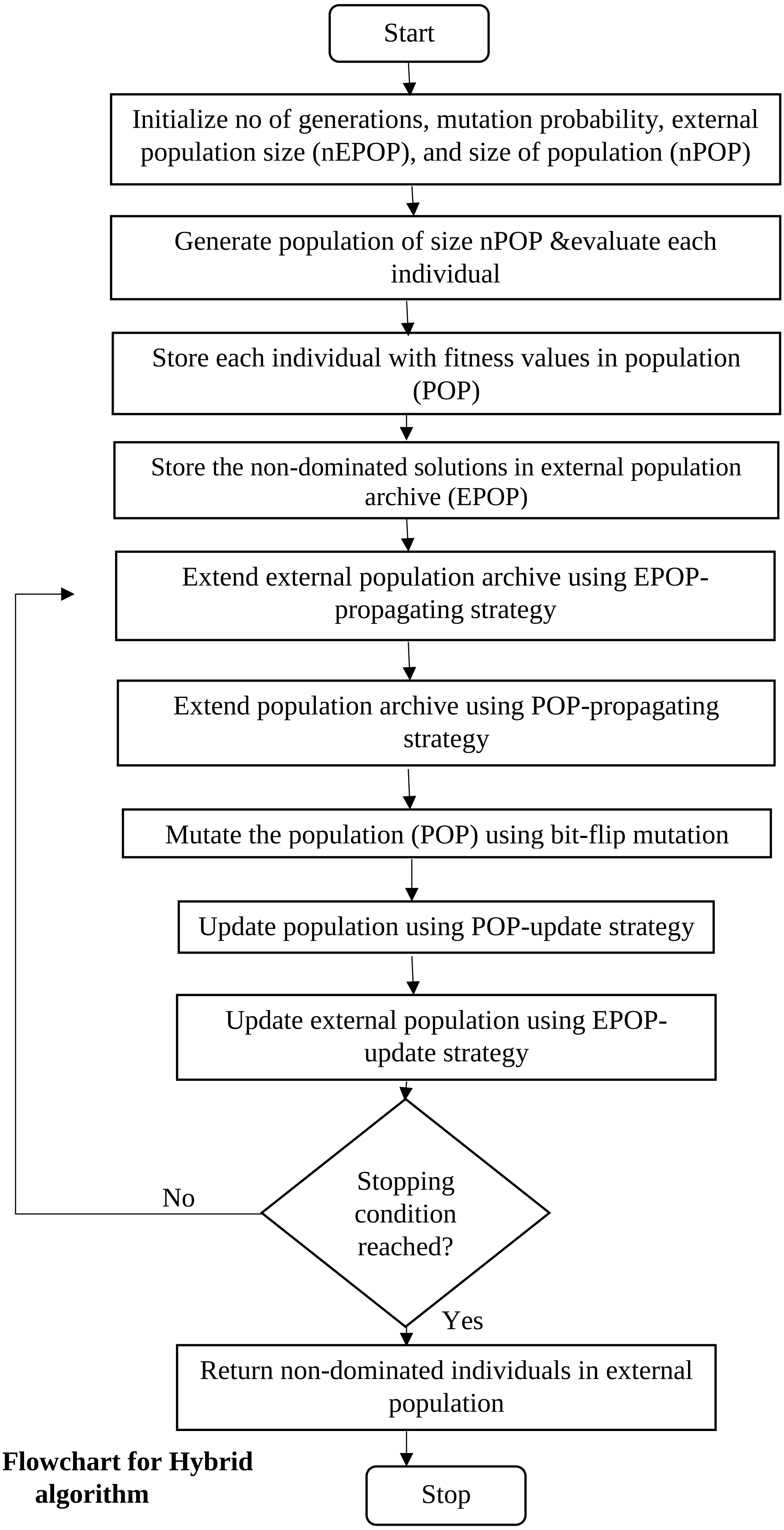

Start
Initialize no of generations, mutation probability, external population size (nEPOP), and size of population (nPOP)
Generate population of size nPOP &evaluate each individual
Store each individual with fitness values in population (POP)
Store the non-dominated solutions in external population archive (EPOP)
Extend external population archive using EPOP-propagating strategy
Extend population archive using POP-propagating strategy
Mutate the population (POP) using bit-flip mutation
Update population using POP-update strategy
Update external population using EPOP-update strategy
Stopping condition reached?
No
Yes
Return non-dominated individuals in external population
Stop

Fig. 5 Flowchart for Hybrid algorithm

2.2 Software and Libraries Used

2.2.1 Software:

- Python 2.7

2.2.2 Libraries used in Python:

i. **Scikit-learn**

This library[45] contains different algorithms like SVM, Decision Tree etc. for classification. It also contains functions to calculate metrics like accuracy score, precision score, recall etc.

ii. **Matplotlib**

Matplotlib is a plotting library in python used to draw error bars, ROC (Receiver Operating Characteristic)[46] curve, boxplots, histograms etc.

iii. **Scikit-feature**

This library is used to obtain feature selection algorithms.

iv. **Sklearn-genetic**

This library is used to obtain genetic algorithm which takes an estimator to calculate fitness values of population and to select best features based on this fitness values.

v. **Imbalanced-learn**

Imbalanced-learn is used for resampling datasets which contain imbalanced data. It contains many oversampling and undersampling techniques. We use SMOTE for oversampling the data

CHAPTER-3

IMPLEMENTATION

3.1 Datasets used in this project

3.1.1 SRBCT:

- **Description:** SRBCT is a Small Round Blue Cell Tumours. It is a four different childhood tumours named as their appearance on routine histology is similar, which making right analysis to a great degree is testing, clinically. Depending on the diagnosis, accuracy is essential in diagnosing because treatment options, reactions to treatment and forecasts shift generally.
- **Diagnostic Classes:** 4
  - Ewing’s sarcoma (EWS): 29 samples (34.9%)
  - Burkitt’s lymphoma (BL): 11 samples (13.3%)
  - Neuroblastoma (NB): 18 samples (21.7%)
  - Rhabdomyosarcoma (RMS): 25 samples (30.1%)
- **Number of genes:** 2308
- **Number of samples:** 83

3.1.2 Lung Cancer:

- **Description:** Adenocarcinoma, Squamous cell carcinoma and normal lung tissue samples from the original dataset are used to build the classification model for lung cancer[47] data set. The Adenocarcinoma class[48] consists of Adenosquamous sample.

- **Diagnostic Classes:** 5
  - Squamous cell carcinoma (Squamous): 20 samples (9.85%)
  - Adenocarcinoma (Adenocarcinoma): 17 samples (8.37%)
  - Normal lung tissue (Normal): 21 samples (10.34%)
  - Coid: 6 samples (2.95%)
  - SMCL: 139 samples (68.47%)
- **Number of genes:** 12600
- **Number of samples:** 203

3.2 Feature Selection Algorithm to select features

3.2.1 Using ReliefF:

ReliefF[49] deals with noisy, deficient and multiclass dataset. It is a based on similarity between features[50]. An instance (I) is selected from all sample instances randomly and the weight is updated by calculating the nearest hits and nearest misses[9].More weight is given to features that distinguish the instance from neighbours of different classes [51]. The weights are change based on contribution to nearest misses [15]. Prior probability of each class is used for average contribution.

3.3 Classifier Models

3.3.1 Classification using Support Vector Machine (SVM)

Support Vector Machine (SVM)[52] is a supervised ML algorithm used in finding outliers and problems where classification is needed. If data contains n features, then n-dimensional coordinate space is generated [53]. Based on training values the algorithm creates a hyper plane which isused to predict the class of new input.

Advantages of SVM's are that they are best when high number of features are present in dataset. SVM are suitable in gene data classification because of the reason that this classifier performs well on dataset with high features and less number of samples.

Disadvantage of SVM is that it causes overfitting problem on datasets like gene data.

3.4 Non-dominated Sorting Genetic Algorithm-II

Pseudocode for fast non-dominated sorting *(P)*:

```
for each $i \in Pop$
    $Set_i = \emptyset$
    $n - front_i = 0$
    for each $j \in Pop$:
        if$i \prec j$ then  :                (compare $i$ and $j$  for dominance)
```

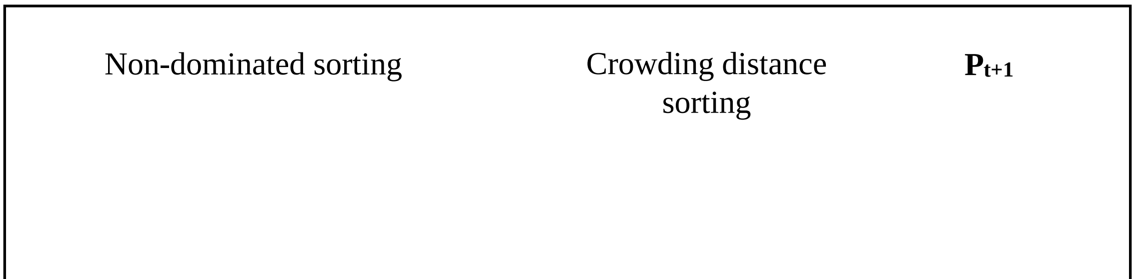
```
            $Set_i = Set_i \cup \{j\}$    ($j$ is added to the set of solutions $i$ dominates )
        else if ($j \prec i$) then :
            $n - front_i = n - front_i + 1$    (Front is updated)
    if$n - front_i = 0$ then :    ($i$ belongs to pareto front or not)
        $i_{rank} = 1$
        $Front_1 = Front_1 \cup \{i\}$
$k$= 1    (front counter is initialized)
while$F_1 \neq \emptyset$:
    $Q = \emptyset$ (stores the next front members)
    for each $i \in Front_i$
        for each $j \in Set_i$
            $n - front_j = n - front_j - 1$
            if$n - front_j = 0$ then    ($j$ is in front k)
                $q_{rank} = k + 1$
                $Q = Q \cup \{j\}$
$k$= $k$+1
$Front_k = Q$
```

Pseudocode for crowding distance assignment (pop):

```
N=size(pop)

For each p in pop:

    $dist_i = 0$

For each objective m:

    Sort(pop,m)

    $dist_1 = dist_N = \infty$

    for each p in pop(2,N-1):

        $dist_i = dist_i + (dist_{i+1} - dist_{i-1})/$(m_max - m_min)
```

NSGA-II Algorithm:

1. Randomly generate individuals to form a population.
2. Calculate the objective function values for each individual of the population
3. Initialize an empty offspring
4. Perform the below steps till the offspring size becomes the size of the population, N
    i. Select two population using Binary Tournament Selection
    ii. Apply HUX crossover and Bit Flip Mutation on the population and add these individuals to the offspring
5. Calculate the fitness values of the offspring.
6. Combine the old population into the offspring
7. Apply non-dominated sort on the offpsirng
8. Using crowding distance, we select the N individual chromosomes and rest of the offspring is rejected.
9. New population is obtained from offspring using non-dominated sort and crowding distance.
10. Go to step 2 and repeat the procedure until generations are completed.
11. The highest ranked Pareto non-dominated set from the latest population is the solution set.



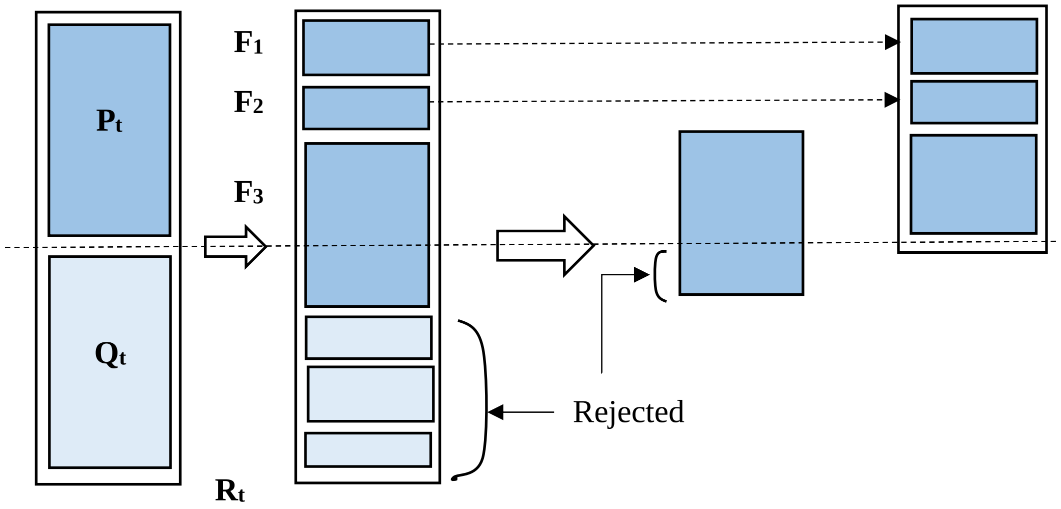


Fig. 6 NSGA-II Procedure

3.5 Multi-Objective Particle Swarm Optimization

Steps involved in MOPSO Algorithm:

i. Generate population randomly using the algorithm

ii. Initialize the local best for the generated particles.

iii. Find out the global best for the swarm i.e., leaders and determine all the leaders.

iv. Initialize all the particles of the population with velocity=0.

v. Calculate the fitness of each particle of generated population

vi. Update velocity V of the particle using the formula

$$V_i^{t+1} = wV_i^t + c_1 r_1\left(x_{pbest} - X_i^t\right) + c_2 r_2 (x_{gbest} - X_i^t)$$

vii. Using the sigmoid function update the positions of the particles using BPSO[54] concept. Formula for updating the position X is

$$x_{i,j}(t+1) = \begin{cases} 0 & if\ rand() \geq S(v_{i.j}(t+1)) \\ 1 & if\ rand() < S(v_{i,j}(t+1)) \end{cases}$$

$$S\left(v_{i,j}(t+1)\right) = \frac{1}{1 + e^{-v_{i,j}(t+1)}}$$

viii. Perform mutation on the particles and update local best for each particle by comparing it with the old values.

ix. Go to second step, until all the generations are completed.

x. The non-dominated pareto[55] set is our best population.

3.6 Hybrid Multi-Objective Feature Selection Algorithm

Procedure of Hybrid Multi-Objective Feature Selection Algorithm:

1. Initialize number of generations,mutation probability, external population size(nEPOP), size of population(nPOP).
2. Generate population of size nPOP[56] and evaluate each individual.
3. Store each individual with fitness values in population(POP)
4. Store the non-dominated solutions in external population (EPOP).
5. Repeat the following steps number of generations times:

   5.1) Extend external population archive using EPOP-propagating strategy described below.

   5.2) Extend population archive using POP-propagating strategy as described below.

   5.3) Mutate the population (POP) using bit-flip mutation.

   5.4) Update population using POP-update strategy.

   5.5) Update external population using EPOP-update strategy.
6. Return the non-dominated individuals in external population.

EPOP-propagating strategy:

1) Initialize empty list offspring

2) Repeat the following steps until size of offspring less than nEPOP:

   2.1) Select two population using binary tournament selection from EPOP.

   2.2) Generate new individuals using HUX crossover.

   2.3) Add this individuals to offspring list.

3) Evaluate the individuals in offspring.

4) Store the non-dominated solutions in EPOP.

POP-propagating strategy:

1) Initialize empty list offspring.

2) Repeat the following steps until size of offspring less than nPOP:

2.1) Select two parents randomly each from EPOP and POP.

2.2) Generate new individuals using HUX crossover.

2.3) Add this individuals to offspring list.

3) Evaluate the individuals in offspring.

4) Store the offspring individuals in POP.

POP-update strategy:

1) For each individual in POP:

1.1) If individual is dominated by any individual in POP, discard this individual.

1.2) Else remove those individuals dominated by this individual and add this to POP.

2) If POP size greater than nPOP, randomly delete individuals.

EPOP-update strategy:

1) For each individual in POP:

1.1) If individual is dominated by any individual in EPOP[19], discard this individual.

1.2) Else remove those individuals dominated by this individualadd this to EPOP.

2) If EPOP size greater than nEPOP, individuals are selected using crowding distance

3.7 Implementation in Cloud

The hybrid algorithm is now implemented on Google Cloud with high computational power.

The configurations of Google Cloud are:

- 52 Gb RAM
- 8 vCPUs

Specifications of our Personal Computer:

- 8 Gb RAM
- 2.19 GHz processor

CHAPTER-4

EXPERIMENTAL ANALYSIS

4.1 Pareto Fronts on Lung Cancer dataset

Pareto Fronts obtained using NSGAII Algorithm, MOPSO Algorithm and Hybrid Algorithm on Lung Cancer dataset and AUC score and number of features as objectives and SVM as classifier.

4.1.1 NSGA-II Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running NSGA-II algorithm for 2000 generations and size of initial population is 15.

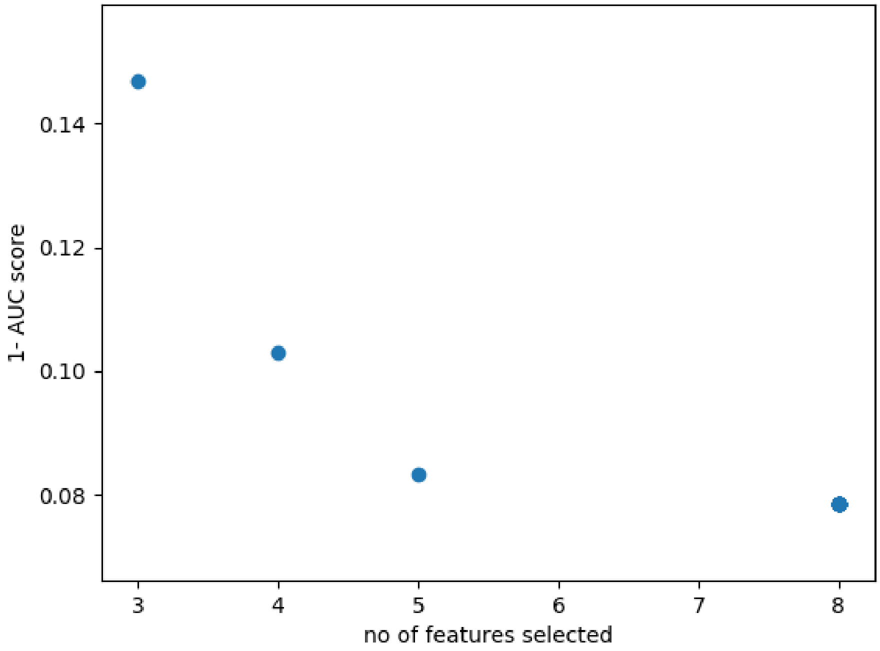


Image. 11 Pareto front obtained using NSGA-II on Lung cancer dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	8	0.921569
2	5	0.91667
3	4	0.852941
4	3	0.897059

Table. 1 AUC score & No. of Features selected using NSGA-II

4.1.2 MOPSO Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running MOPSO algorithm for 2000 generations and size of swarm is 100.

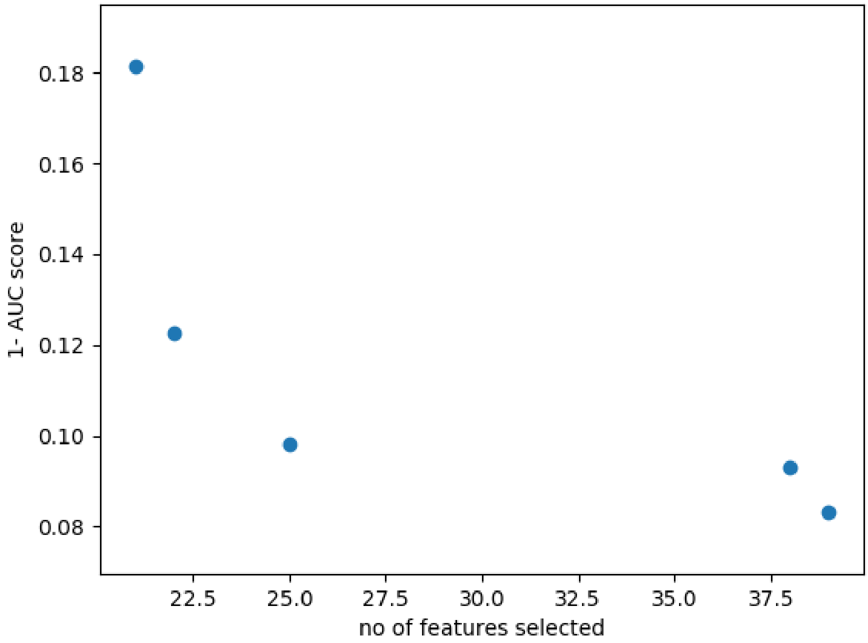


Image. 12 Pareto front obtained using MOPSO on Lung cancer dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	21	0.818627
2	22	0.877451
3	25	0.901961
4	38	0.906863
5	39	0.916667

Table. 2 AUC score & No. of Features selected using MOPSO

4.1.3 Hybrid Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running hybrid algorithm for 100 generations and size of population is 15.

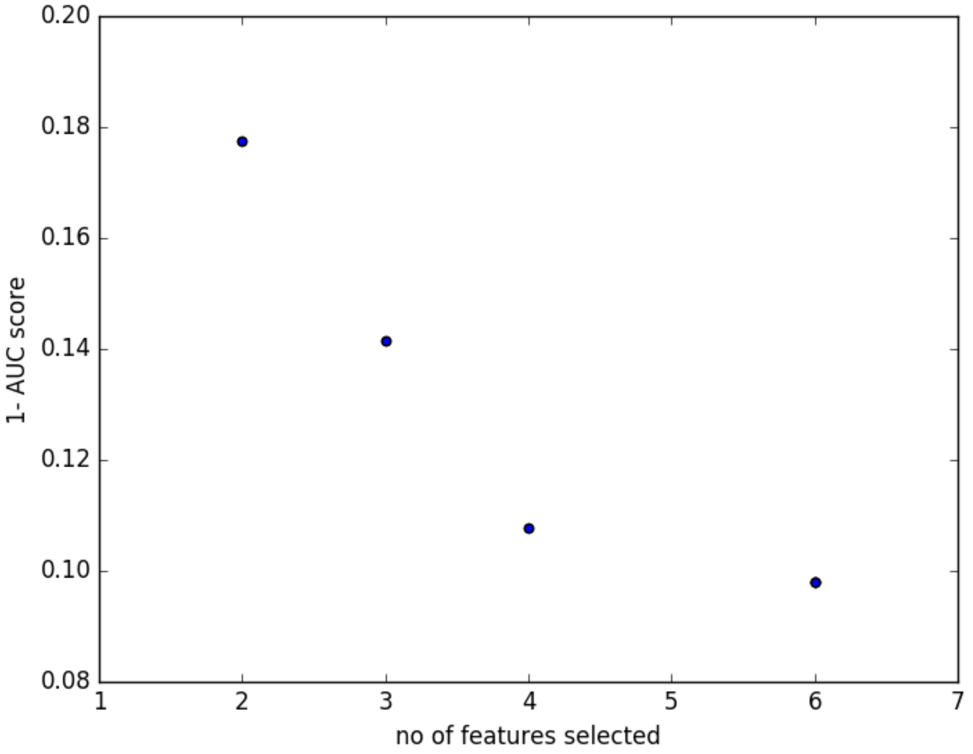


Image. 13 Pareto front obtained using Hybrid on Lung cancer dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	6	0.901961
2	4	0.892157
3	3	0.858612
4	2	0.822376

Table. 3 AUC score & No. of Features selected using Hybrid Algorithm

4.2 Comparison of computation times between PC and Cloud Platform for Lung Cancer dataset

The above algorithms are run in both cloud and personal computer. Below table shows time taken by each algorithm in personal computer and cloud.

Speedup Factor: $PC_{time} / Cloud_{time}$

Algorithm	Cloud(seconds)	PC (seconds)	Speedup Factor
NSGAII	75.49	495.053	6.55
MOPSO	120.7956	604.135	5.00
HYBRID	110.937	676.367	6.09

Table. 4 Computational times of PC and Cloud for Lung Cancer

4.3 Pareto Fronts on SRBCT dataset

Pareto Fronts obtained using NSGAII Algorithm, MOPSO Algorithm and Hybrid Algorithm on SRBCT dataset and AUC score and number of features as objectives and SVM as classifier.

4.3.1 NSGA-II Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running NSGA-II algorithm for 2000 generations and size of initial population is 15.

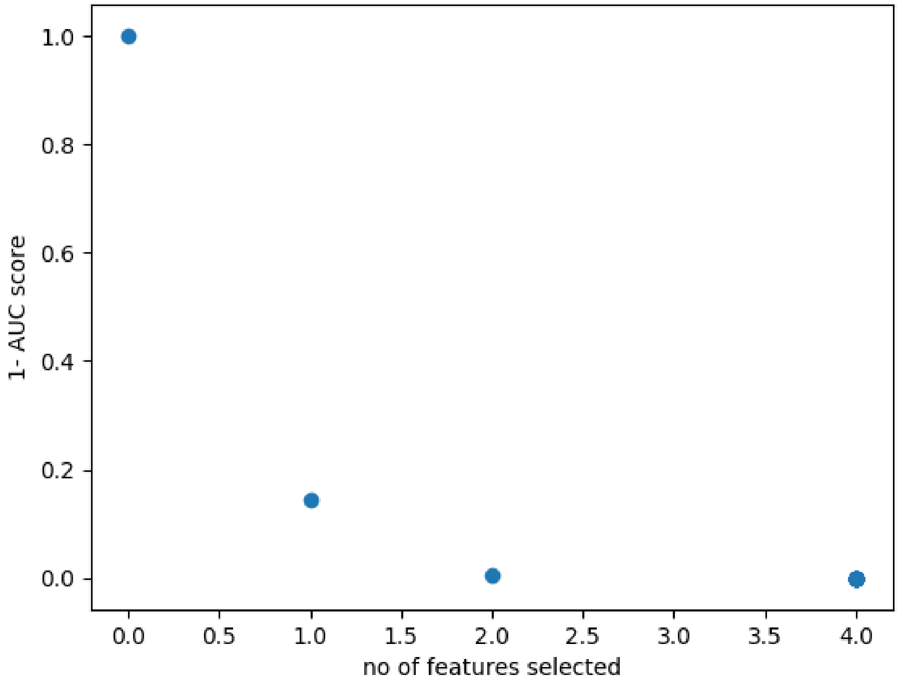


Image. 14 Pareto front obtained using NSGA-II on SRBCT dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	4	1
2	3	0.857143
3	2	0.995465
4	1	0

Table. 5 AUC score & No. of Features selected using NSGA-II for SRBCT

4.3.2 MOPSO Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running MOPSO algorithm for 2000 generations and size of swarm is 100.

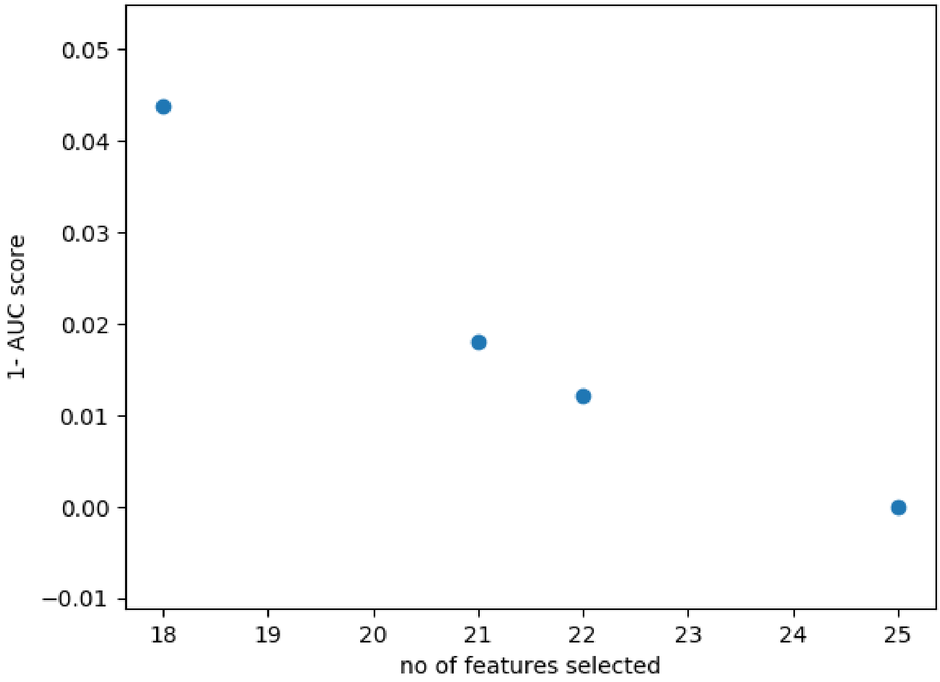


Image. 15 Pareto front obtained using MOPSO on SRBCT dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	18	0.95616
2	21	0.981859
3	22	0.987906
4	25	1

Table. 6 AUC score & No. of Features selected using MOPSO for SRBCT

4.3.3 Hybrid Algorithm:

The following figure represents graph between 1-auc score and number of features selected in individuals obtained after running hybrid algorithm for 100 generations and size of population is 15.

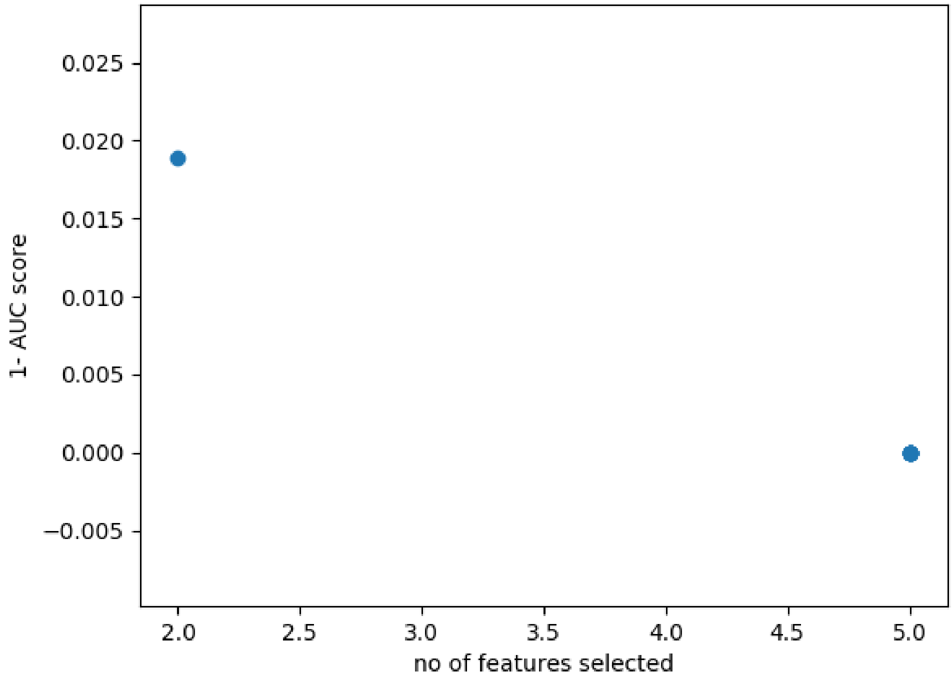


Image. 16 Pareto front obtained using Hybrid on SRBCT dataset

The following table shows above represented data in tabular form

Individual	No of Features selected	AUC score
1	5	0
2	2	0.981104

Table. 7 AUC score & No. of Features selected using Hybrid Algorithm for SRBCT

4.4 Comparison of computation times between PC and Cloud Platform for SRBCT dataset

The above algorithms are run in both cloud and personal computer. Below table shows time taken by each algorithm in personal computer and cloud.

Speedup Factor: PC_{time} / $Cloud_{time}$

Algorithm	Cloud(seconds)	PC (seconds)	Speedup Factor
NSGAII	29.31547	75.19	2.56
MOPSO	33.60758	256.286	7.6258
HYBRID	34.8744	307.634	8.8211

Table. 8 Computational times of PC and Cloud for SRBCT dataset

CHAPTER-5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The project addresses the issue of presence of redundant features or unnecessary features which cause overfitting of classifier and also errors in predicting the output.

To overcome this problem in our project we used multi objective optimization feature selection algorithms to decrease the dimensionality of the data. This also reduces the overfitting problems. In Lung cancer dataset there are 12600 features, initially we selected 60 features using relief-f feature selection and this 60 features are further reduced by multi objective optimization algorithms. AUC score has been used to study the performance of classifier.

We can observe that selecting less than 10 features using multi objective optimization algorithms, AUC score obtained is greater than 90 in most of the cases.

5.2 Future Work

In our project, we used only one type of crossover to obtain new population. In future we try to use different crossover operators and selecting them randomly using roulette wheel selection for each parent.

We also try to improve running time of hybrid algorithm used in this project.

REFERENCES

[1] V. Bolón-canedo, N. Sánchez-maroño, A. Alonso-betanzos, J. M. Benítez, and F. Herrera, “A review of microarray datasets and applied feature selection methods,” vol. 282, pp. 111–135, 2014.

[2] A. Brazma and J. Vilo, “Gene expression data analysis,” vol. 480, pp. 17–24, 2000.

[3] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, “A Review of the Stability of Feature Selection Techniques for Bioinformatics Data,” pp. 356–363.

[4] A. J. Stephenson *et al.*, “NIH Public Access,” vol. 104, no. 2, pp. 290–298, 2007.

[5] G. Agre and A. Dzhondzhorov, “A Weighted Feature Selection Method for Instance-Based Classification,” vol. 1, pp. 14–25, 2016.

[6] P. Alto, “Selection of Relevant Features and Examples in Machine Learning.”

[7] “Feature selection techniques-3.pdf.” .

[8] 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, 2009.

[9] A. J. Stephenson *et al.*, “Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy,” *Cancer*, vol. 104, no. 2, pp. 290–298, 2005.

[10] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[11] J. G. Dy and C. E. Brodley, “Feature Selection for Unsupervised Learning ,” *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, 2004.

[12] I. A. Gheyas and L. S. Smith, “Feature subset selection in large dimensionality domains,” *Pattern Recognit.*, vol. 43, no. 1, pp. 5–13, 2010.

[13] A. Manuscript, “NIH Public Access,” vol. 46, no. 2, pp. 137–151, 2012.

[14] D. Mining, “Improving accuracy of microarray classification by a simple multi-task feature selection filter Liang Lan Slobodan Vucetic *,” vol. 5, no. 2, pp. 189–208, 2011.

[15] Y. Leung and Y. Hung, “A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 7, no. 1, pp. 108–117, 2010.

[16] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, “Is cross-validation better than resubstitution for ranking genes?,” *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

[17] G. Agre and A. Dzhondzhorov, “Artificial Intelligence: Methodology, Systems, and Applications,” vol. 9883, pp. 14–25, 2016.

[18] S. Lalwani, S. Singhal, R. Kumar, and N. Gupta, “a Comprehensive Survey: Applications of Multi-Objective Particle Swarm Optimization (Mopso) Algorithm,” *Trans. Comb.*, vol. 2, no. 1, pp. 2251–8657, 2013.

[19] U. Singh and S. N. Singh, “Optimal Feature Selection via NSGA-II for Power Quality Disturbances Classification,” *IEEE Trans. Ind. Informatics*, vol. 3203, no. c, pp. 1–9, 2017.

[20] A. SESHADRI, “A Fast Elitist Multiobjective Genetic Algorithm - Research Papers - Liangyuly,” pp. 1–4, 2011.

[21] F. Fortin, “Revisiting the NSGA-II Crowding-Distance Computation,” pp. 623–630, 2013.

[22] A. E. Smith, “Multi-objective optimization using evolutionary algorithms [Book Review],” *IEEE Trans. Evol. Comput.*, vol. 6, no. 5, pp. 526–526, 2002.

[23] A. Golchha and S. G. Qureshi, “Non-Dominated Sorting Genetic Algorithm-II – A Succinct Survey,” vol. 6, no. 1, pp. 252–255, 2015.

[24] X. Li and M. Yin, “Multiobjective binary biogeography based optimization for feature selection using gene expression data,” *IEEE Trans Nanobioscience*, vol. 12, no. 4, pp. 343–353, 2013.

[25] S. S. Hosseini, S. A. Hamidi, M. Mansuri, and A. Ghoddosian, “Multi objective particle swarm optimization (MOPSO) for size and shape optimization of 2D truss structures,” *Period. Polytech. Civ. Eng.*, vol. 59, no. 1, pp. 9–14, 2015.

[26] K. Nag and N. R. Pal, “A Multiobjective Genetic Programming-Based Ensemble for Simultaneous Feature Selection and Classification,” *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 499–510, 2016.

[27] A. Khodadadi and P. Von Buelow, “Dynamic Configuration Processing and Optimization of Forms (Exploring,” no. October, 2014.

[28] Z. Beheshti, S. M. Shamsuddin, and S. Hasan, “Memetic binary particle swarm optimization for discrete optimization problems,” *Inf. Sci. (Ny).*, vol. 299, pp. 58–84, 2015.

[29] L. Tang and X. Wang, “A hybrid multiobjective evolutionary algorithm for multiobjective optimization problems,” *IEEE Trans. Evol. Comput.*, vol. 17, no. 1, pp. 20–45, 2013.

[30] H. Ali, W. Shahzad, and F. A. Khan, “Energy-efficient clustering in mobile ad-hoc networks using multi-objective particle swarm optimization,” *Appl. Soft Comput. J.*, vol. 12, no. 7, pp. 1913–1928, 2012.

[31] A. Khan and A. R. Baig, “Multi-objective feature subset selection using non-dominated sorting genetic algorithm,” *J. Appl. Res. Technol.*, vol. 13, no. 1, pp. 145–159, 2015.

[32] E. Fallah-Mehdipour, O. Bozorg Haddad, and M. a. Mariño, “MOPSO algorithm and its application in multipurpose multireservoir operations,” *J. Hydroinformatics*, vol. 13, no. 4, p. 794, 2011.

[33] C. A. Coello Coello and M. S. Lechuga, “MOPSO: A proposal for multiple objective particle swarm optimization,” *Proc. 2002 Congr. Evol. Comput. CEC 2002*, vol. 2, pp. 1051–1056, 2002.

[34] J. D. Knowles and D. W. Corne, “Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy,” *Evol. Comput.*, vol. 8, no. 2, pp. 149–172, 2000.

[35] B. Hrolenok, “Multi-objective Optimization with PSO,” pp. 1–2, 2011.

[36] K. Deb and R. B. Agrawal, “Simulated Binary Crossover for Continuous Search Space,” *Complex Syst.*, vol. 9, pp. 1–34, 1994.

[37] “Evaluating the Performance of Multi-Objective Particle Swarm Optimization Algorithms,” 2016.

[38] K. C. Tan, T. H. Lee, and E. F. Khor, “Evolutionary algorithms for multi-objective optimization: performance assessments and comparisons,” *Proc. 2001 IEEE Congr. Evol. Comput. Seoul*, vol. 2, pp. 979–986, 2001.

[39] Hui Li and Qingfu Zhang, “Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II,” *IEEE Trans. Evol. Comput.*, vol. 13, no. 2, pp. 284–302, 2009.

[40] S. Lee, S. Soak, S. Oh, W. Pedrycz, and M. Jeon, “Modified binary particle swarm optimization,” *Prog. Nat. Sci.*, vol. 18, no. 9, pp. 1161–1166, 2008.

[41] M. Hu, T. Wu, and J. D. Weir, “An Adaptive Particle Swarm Optimization With Multiple Adaptive Methods,” *IEEE Trans. Evol. Comput.*, vol. 17, no. 5, pp. 705–720, 2013.

[42] J. Durillo and J. García-Nieto, “Multi-objective particle swarm optimizers: An experimental comparison,” *5th Int. Conf. EMO 2009*, pp. 495–509, 2009.

[43] T. Sağ and M. Çunkaş, “A new ABC-based multiobjective optimization algorithm with an improvement approach (IBMO: Improved bee colony algorithm for multiobjective optimization),” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 24, no. 4, pp. 2349–2373, 2016.

[44] K. Deb, “Multi-objective optimization using evolutionary algorithms: an introduction,” *Multi-objective Evol. Optim. Prod. Des. Manuf.*, pp. 1–24, 2011.

[45] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[46] T. Fawcett and T. Fawcett, “ROC Graphs : Notes and Practical Considerations for Data Mining Researchers ROC Graphs : Notes and Practical Considerations for Data Mining Researchers,” 2003.

[47] A. De Rienzo, B. Y. Yeap, E. S. Cibas, W. G. Richards, L. Dong, and R. R. Gill, “Gene Expression Ratio Test Distinguishes Normal Lung from Lung Tumors in Solid Tissue and FNA Biopsies,” *J. Mol. Diagnostics*, vol. 16, no. 2, pp. 267–272, 2014.

[48] B. Calvo, E. Bengoetxea, and P. Larra, “Chapter 2 Machine Learning : An Indispensable Tool in Bioinformatics I naki,” pp. 25–49.

[49] S. Lecturer and A. Pradesh, “Feature Selection using ReliefF Algorithm,” vol. 3, no. 10, pp. 8215–8218, 2014.

[50] S. S. Baboo and S. Sasikala, “Multicategory Classification Using Support Vector Machine for Microarray Gene Expression Cancer Diagnosis,” vol. 10, no. 15, pp. 38–44, 2010.

[51] M. Wu and Y. Wang, “A Feature Selection Algorithm of Music Genre Classification Based on ReliefF and SFS,” 2015.

[52] D. K. Srivastava and L. Bhambhu, “Data classification using support vector machine,” *J. Theor. Appl. Inf. Technol.*, vol. 12, pp. 1–7, 2009.

[53] D. Arockia, C. Vanitha, D. Devaraj, and M. Venkatesulu, “Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection,” *Procedia - Procedia Comput. Sci.*, vol. 47, pp. 13–21, 2015.

[54] E. Zitzler, M. Laumanns, and L. Thiele, “SPEA2: Improving the Strength Pareto Evolutionary Algorithm,” *Evol. Methods Des. Optim. Control with Appl. to Ind. Probl.*, pp. 95–100, 2001.

[55] L. Ke, Q. Zhang, and R. Battiti, “MOEA/D-ACO: A Multiobjective Evolutionary Algorithm Using Decomposition and AntColony,” *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1845–1859, 2013.

[56] Q. Zhang and H. Li, “MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition,” *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, 2007.