MULTI OBJECTIVE FEATURE SELECTION WITH MISSING DATA IN CLASSIFICATION

Jagriti Sharma *ROLL NO: B190864CS*
Uttkarsh Raj *Roll No: B190955CS*
Pratyush Aggarwal *ROLL NO: B190329CS*
Himanshu Modi *ROLL NO: B190456CS*
Navnit Anand *ROLL NO: B190404CS*
Lalit Kushwah *ROLL NO: B190557CS*
Suraj Sasikumar *ROLL NO: B190382CS*
Shaik Rabnawaz *ROLL NO: B191139CS*

***Abstract*—Feature selection (FS) has high importance in machine learning research. Generally, FS is modelled to wrok as a bi-objective optimisation problem with the following objectives: 1) Accuracy in classification. 2) Number of features. Missing Data is one of the major real world issues. The databases with missing data are unreliable. Thus if we perform FS on a data set which misses data, then the FS is unreliable too. To straight up counter this issue we we propose another novel modelling of FS: We make reliability as anither objective of the problem. We make use of the Non-Dominated Sorting genetic algorithm-III. We pick six incomplete data sets from the macine learning repository of university of california Irvine (UCI). Now, to handle the issue of missing data in the experiments, we use the mean imputation method, k-nearest neighbors (K-NN) is made use of as the classifier which evaluates the subsets of features. The results of experiment reveal that using the NSGA-III along with the thee three objective approach efficiently solve the FS problem for the six data sets used.**

***Index Terms*—Feature selection, Multi-objective, Optimization, NSGA-III, Missing data.**

I. Introduction

Feature selection reduces the input size by choosing only the essential and relevant features. This reduces the computational cost of the model and hence improve the performance. Feature selection is a very important thing as the number of data sets keep gowing and increase. Very big problem in real life data sets is missing data. Problems are caused due to missing data in a feature and this results in fluctuations in the required answer. the amplitude of the effect of missing data depends upon the quantity of data that is missing. If the data which is missing is less than 1% we can ignore the effect. A slight effect is caused when the data that is missing amounts to a 1 to 5%. But this can be controlled. but now if the data that we are missing is more than 5% it seriously affect the results of the experiment. a popular method to handle missing data is the use of data imputation. by this we mean reconstructing data on the basis of the data that we have already present in the data set. in this method v interpellate the values of the missing data by taking the average of the data available. now this approach is again divided into fixed data imputation method and non distance imputation method. We know that a popular algorithm for multi objective Optimisation is NSGA-III. The main feature of this algorithm is called reference point based selection method. we can summarise its main feature as follows: this algorithm builds up a set of of reference points. then this algorithm randomly generate and initialisation with a population P made up of of N individuals. Now it generate a new population Q by using binary cross over and polynomial mutations and then it combines P and Q for fast dominating sorting. through the use of non-dominated Frank this algorithm chooses N individuals to enter the offspring population. when the algorithm cannot use the non dominated rank for selection, then in that case The selection is done through reference point mechanism.

Identify applicable funding agency here. If none, delete this.

II. Mean Imputation Method

Imputation methods refer to the methods in which the missing values in a matrix are replaced by some values according to the calculation specific to that method. The single imputation method refers to the method in which the missing values in the data matrix are replaced randomly by some number between the maximum and minimum values given in the matrix. The mean Imputation method refers to the method where the missing values in the matrix are replaced by the mean of the remaining values. The missing data in the data matrix decrease the accuracy of any model that is using that data set but these methods improve the accuracy of any algorithmic calculation that is performed on the data set by removing those missing fields by some specific value provided by the different imputation methods.

Some of the imputation methods widely used are hot-deck imputation in which missing values are replaced by a close pattern of similar variable and regression imputation in which regression of the target variable on all other variables.

To solve the problem of missing data we are using the mean imputation method here. We suppose that mean imputation method will result in the most accurate result in our case. The accuracy of different imputation methods depends heavily on the size and nature of the data set. Mean imputation method replaces the missing data in the data matrix by the mean of the other non-missing values and therefore improving the accuracy of any algorithmic model that is performed on that

data set.The formula for calculation of this mean value using mean imputation method is given below in (1)

$$Ave_j = \frac{\sum_{i=1}^{N} v_{i,j}}{N - lm_j} \tag{1}$$

The Formula used in (1) actually calculates the mean of the non-missing data which will be used to replace all the missing data in the data set. Therefore we calculate the sum of all non-missing values and divide it by the number of non-missing values to get the average of all non-missing values.This value obtained by Formula (1) is used to replace all missing values in the data set.

The advantage of using the mean imputation method is that it does not decrease the size of our data set and the whole data set remains the same. The value obtained by calculating the mean is not biased. This method is also very easy to explain and can be justified easily to anyone.

The disadvantage of using the imputation method is that we treat the missing data that we had filled by using the imputation method as real observed data which can decrease the accuracy of the data set. To increase the accuracy of the data set we use multiple imputation methods on the same data set to increase accuracy and the mean of all the different values obtained using all different imputation methods is then taken. But in the case of the mean Imputation method, we cannot be very sure about the accuracy as all the values will be replaced by a single variable which is the mean of the data set whereas in another imputation method different missing values are replaced by different variables according to the algorithm of the imputation method which can result in better accuracy.

III. THREE-OBJECTIVE FEATURE SELECTION

Multi objective optimization problems are used to solve feature selection problems. In this case three-objective feature selection is used. The candidate solution for it is represented as the vector of real numbers where values range from 0 to 1. To evaluate the candidate solution, we take a binary vector where the value 0 or 1 given is based on the threshold value. If the value of the binary vector is 1 it means that the particular feature has been selected and vice versa.

A. Three objectives are given below

1) Classification Accuracy: Classification accuracy can be determined as the total number of accurate predictions divided by the total number of predictions in that data set. Classification error rate on the other hand can be determined as 1 minus classification accuracy. We are using non dominant relationship for comparison in this case and hence we have to use the classification error rate.

2) Solution Size: This objective is calculated by adding all the feature values in the selected data sets. This objective is pretty self explanatory.

3) Missing rate: Missing rate is the rate of missing data in the selected feature set divided by the missing data in the data set.i.e it is the missing data(selected features) divided by missing data(full data set) times 100. This objective is added to contemplate for the reliability of the selected features in the data set.

IV. APPROACH AND IMPLEMENTATION

A. Algorithm 1-NSGA - III

As mentioned above, the objective of this problem is tri-fold. Along with classification accuracy and the number of features, reliability has also been added as an objective. Up to date, we got that for multi-objective optimization NSGA-III is popular a popular method for multi-objective optimization algorithms. In our research paper, we have to do three objective optimizations which we already mentioned in section III i.e. classification accuracy, solution size, and missing rate. The prerequisite of this NSGA-III algorithm is the reference directions that need to be provided to initialize this algorithm. In our implementation, we had used ”das Dennis” reference directions. These reference directions are used to assign the solutions and if reference directions are not present then solutions or the population with the smallest perpendicular distance are selected. For a better understanding, you can see Fig 2 on the right side. Another parameter of this algorithm is the parent population which is created by selecting the random values from the total individuals. And we will represent the parent population by Pt. After getting these two parameters, we can start applying this algorithm. Then we will apply binary crossover/recombination on this parent population Pt this binary crossover/recombination is in section IV.D. From this we got Rt. And then we will apply polynomial mutation to this Rt population and we got Qt population as an outcome this polynomial mutation is in section IV.E. After getting the Qt, we combine the Qt and Pt and pass this population to the non-dominated sorting algorithm this non-dominated sorting is in section IV.B. From this non-dominated sorting, we got the non-dominating rank. Up to this part, this algorithm is similar to NSGA-II. Now we are ready to select the population via the selection algorithm which is explained briefly in section IV.C. In this selection algorithm, individuals are selected with their non-dominated rank which we got from applying the non-dominating sorting algorithm, if in case the selection can not be made from the individual’s non-dominating rank, the algorithm selects via reference points. That’s why this algorithm is also known as reference point mechanism selection. For a better understanding, we had given the flow chart of this algorithm which is fig 4.

B. Alogrithm2-Fast non dominated sorting

Non-dominated sorting is used for non-dominated relationships in this algorithm. Here to compare the non-dominated relationships, we store two parameters, first, we store the two parameters to keep count of individuals which dominate p, and to store the individuals who are dominated by p.

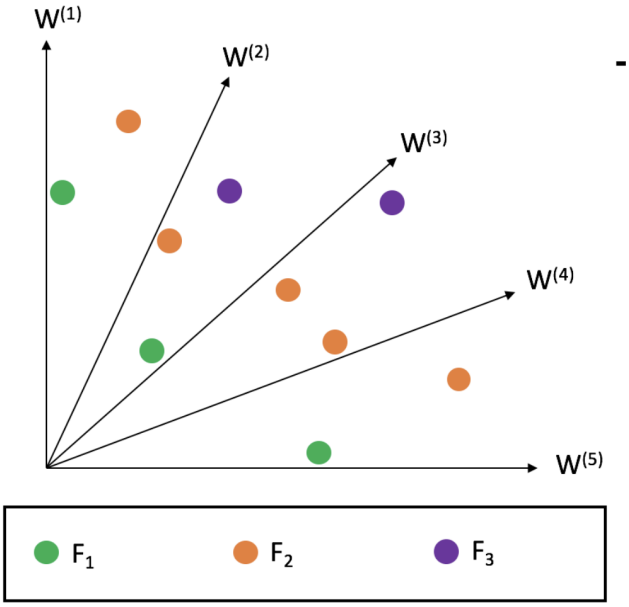

Fig. 1.

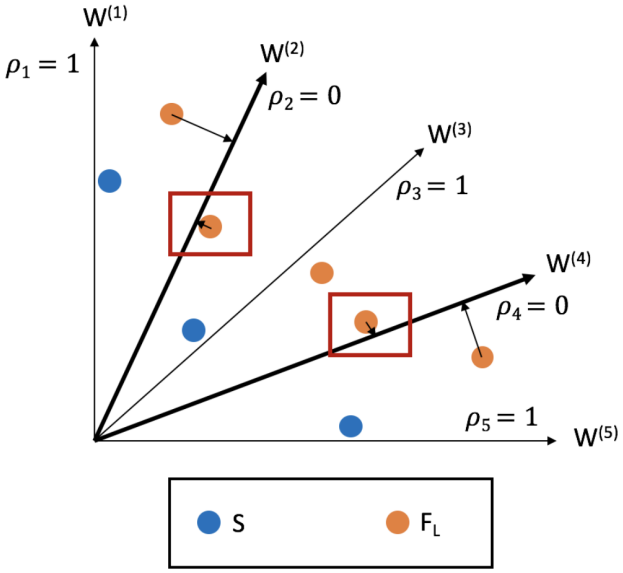

Fig. 2.

In the process, we traverse every pair of an individual in the population and change the stored parameters' values accordingly based on the relationship between each pair of individuals. Based on the parameter that store the count of the individual which dominate p equal to 0, we initialize rank and group the individuals after traversing the individuals from the stored parameter.

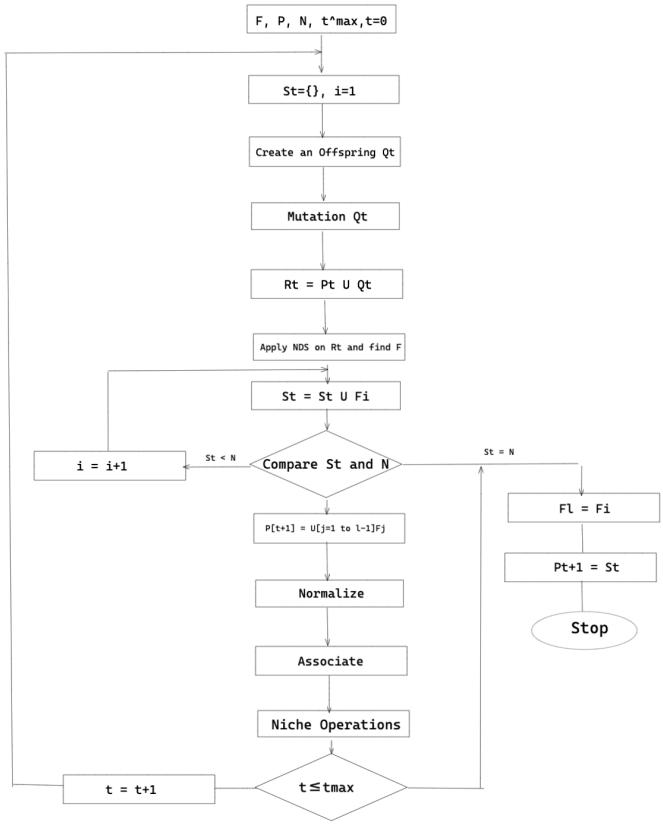

Fig. 3. flowchart of NSGA-III Algorithm

C. Algorithm3-Selection

Individuals are added to the offspring population after a quick non-dominant ranking according to the next descendant group's dominant rank. If no offspring are selected, then we use the reference point Selection method. In the reference point selection method, we create reference points according to the method developed by Das Denniss. In that we create reference points on the equilateral triangle with vertices (1,0,0), (0,1,0), (0,0,1) and the axis is based on the ideal point as considered as the origin. M is an objective dimension. Each goal is divided into P parts. The distance between each reference point is 1/P. Also coordinate of each reference point can be evaluated.

To relate each individual with a specific reference point, we begin by finding the reference line between the origin and reference point later then we calculate the vertical distance of each individual of the population from the found reference line. Then we map each individual to the reference point with the least distance from the reference line.

D. Sub-Algorithm-Recombination

Recombination is also known as binary crossover. It is a genetic algorithm based on transformative mathematics. It is an inherative operator which is used to merge the information of two parents to create a new different child/offspring which

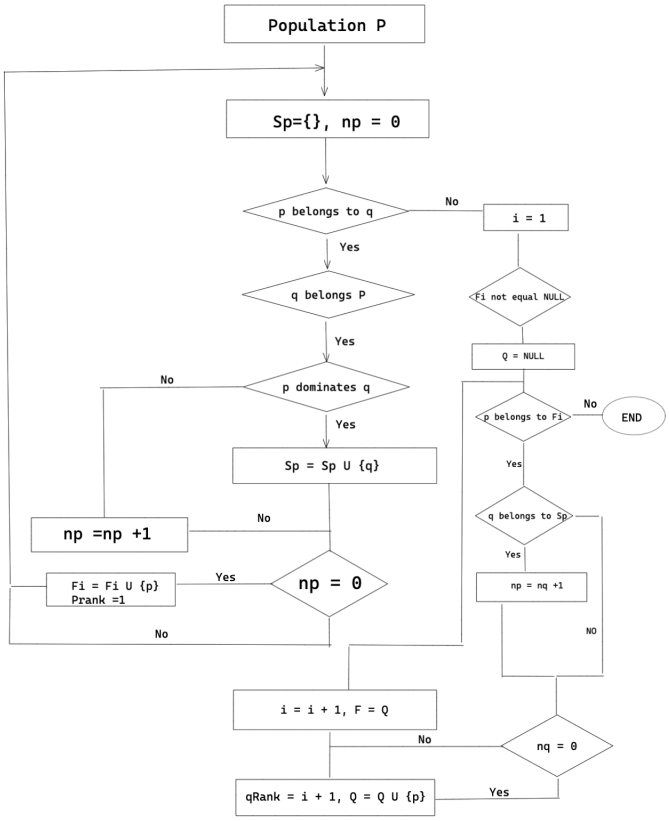

Fig. 4. Fast-non Dominating-sort

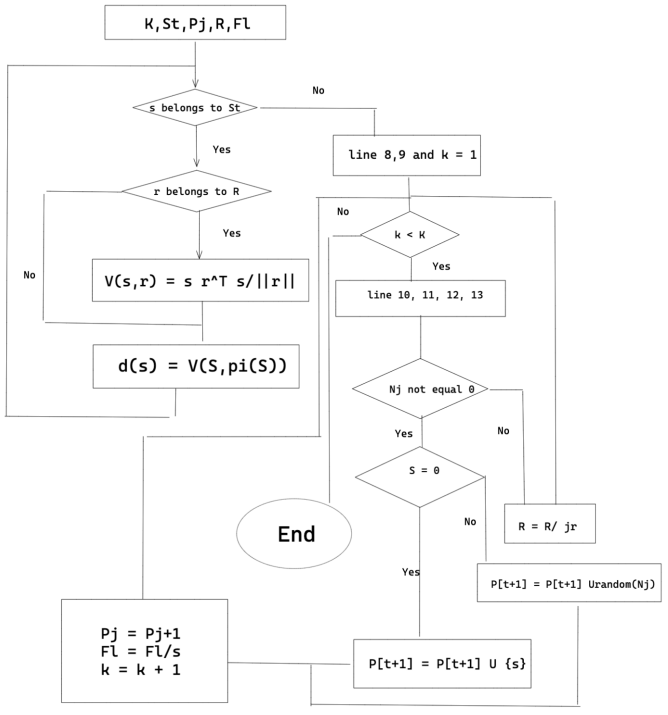

Fig. 5. Selection

means this algorithm is used to create new solutions out of existing solutions. We can also create new solutions by replicating the existing population.

There are different kinds of recombination or binary crossover algorithms and in our implementation of code, we had used a 2-point binary crossover algorithm. In this 2-point binary crossover or recombination algorithm, what we have to do is just pick arbitrarily from the parent population. We have to swap the bits between the two parents we have selected before to create a new offspring.

In fig .7 (Mutation and crossover), there are two examples to illustrate this binary crossover algorithm. In the example, one-two parents are 1.15.7.3 and 12.5.6.9 which are represented in decimals, and to create a new offspring we had swapped the values which means we had selected the first block of parent two and the second block or parent one similarly the third block of parent two and fourth block of parent one. In the end, we got a new offspring i.e. 12.15.6.3. Similarly, in the second example which is represented in a bitwise block, bits are swapped to create a new offspring. In this example, two parents are 0001.1111.0111.0011 and 1100.0101.0110.1001, and the offspring we got after applying this algorithm is 1000.1101.0110.1011. This is all we got from this algorithm.

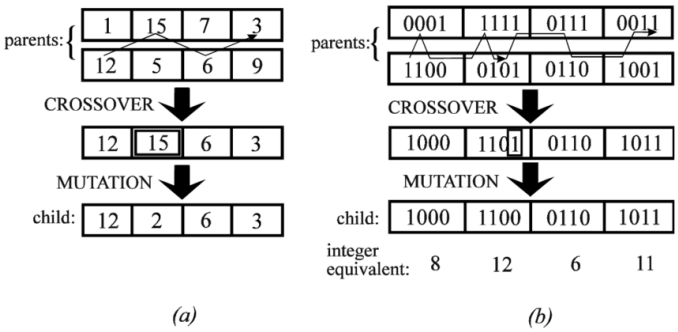

Fig. 6. Mutation and crossover

E. SubAlgorithm-Mutation

Mutation or Polynomial Mutation is carried out by flipping some digits of a string, which generates new solutions.Mutation generally changes a single bit in a bit string. This operator happens with very low probability.In the fig .5[Mutation and crossover] you can see on the right side of figure second block 1101 changes to 1100 which means 4th bit is changes from 1 to 0.

V. Analysis Of NSGA-III

We have successfully implemented the NSGA-III algorithm in this research paper. We have successfully eliminated all the downsides of the NSGA-II algorithm and produced a refined version of it. One such downside that we have been able to eliminate is the inability to maintain distinctiveness among the population members. NSGA-III solves this by using several reference points that are situated on the edges of an equilateral

triangle. The results obtained using the NSGA-III algorithm have been found to be much superior to other multi-objective algorithms that are commonly used. Inverted generational distance(IGD) and Hypervolume(HV) which are metrics used to assess the quality of multi-objective algorithms show a clear superiority for the NSGA-III algorithm. The SD and mean values also support this conclusion.

The future scope of this research paper is to improve this NSGA-III algorithm even further by including different strategies and mutation schemes.

VI. Conclusion

Form this paper we have a interpretation of the FS problem in data science and specifically for the datasets with missing data. In contrast to the classical study, which uses the accuracy and size of the solution as a quality metric, we propose the inclusion of the third metric which is the missing rate. This modeling presents three objective optimization problems addressed using the NSGA III. To demonstrate the effectiveness of the proposed approach, NSGA III was tested on six incomplete datasets and was compared with the four popular multi-objective optimization algorithms. From the obtained results it was concluded that NSGA III is the best among them in terms of IGD and HV. Hence we can say that performance of NSGA III is promising.

VII. Result

We had used das-dennis Reference direction for generating the reference points. We had used dtlz1, dtlz2 and dtlz3 problems for checking our Algorithm. DTLZ problems are made by Deb, Thiele, Laumanns and Zitzler.We are getting Fig 8, 9 and 10 as a result of the implementation of this Algorithm. We took m=3, where m represents the no of objectives as per our research paper. And we have taken population size is 400. To run this program open the terminal and reach to the folder and then run this command: python main.py and then wait for a few seconds for the program to run and then you will get a window in which the result is obtained. The obtained result is presented on the new display that pops up.

Dataset	Mean NSGA 2	Mean NSGA 3
DTLZ1	0.1670	0.1675
DTLZ2	0.4978	0.4663
DTLZ3	0.4931	0.4680

(2)

Dataset	Runtime NSGA 2 (s)	Runtime NSGA 3 (s)
DTLZ1	75.283	28.456
DTLZ2	74.513	15.441
DTLZ3	70.641	31.548

(3)

VIII. Future Scope

The NSGA-III Algorithm has become one of the most popular multi objective optimization algorithms. It uses multiple predefined reference directions to maintain diversity among its solutions. There are a lot of applications of this algorithm such as simulated binary crossover, polynomial mutation, selection, virtual mapping procedure, fitness function evaluation etc. Even though NSGA-III has so many applications, it does not perform well in large scale optimization problems. To solve this problem, we can use information feedback models to improve the ability to solve large scale optimization problems.

IX. Contribution

A. Navnit Anand implemented polynomial mutation and project report

B. Pratyush Aggarwal implemented Recombination and report

C. Himanshu Modi implemented Selection and report

D. Lalit Kushwah implemented Non dominated Sorting and report

E. Suraj Sasikumar implemented sample file and report

F. Shaik Rabnawaz, Uttakarsh Raj,Jagriti Sharma implemented the nsga3 algorithm

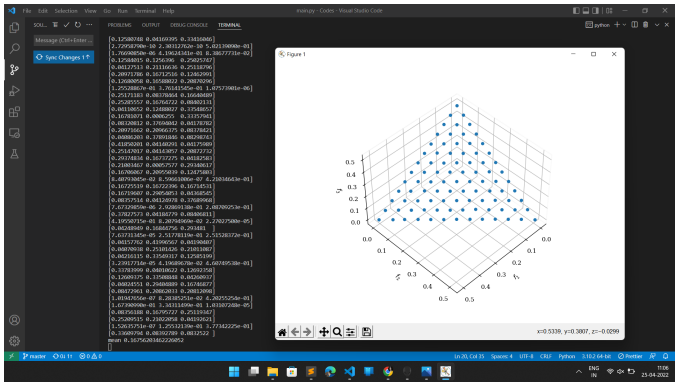

Fig. 7. Result with DTLZ1

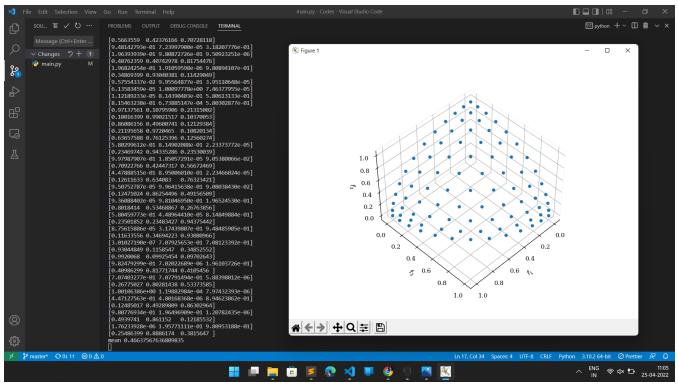

Fig. 8. Result with DTLZ2

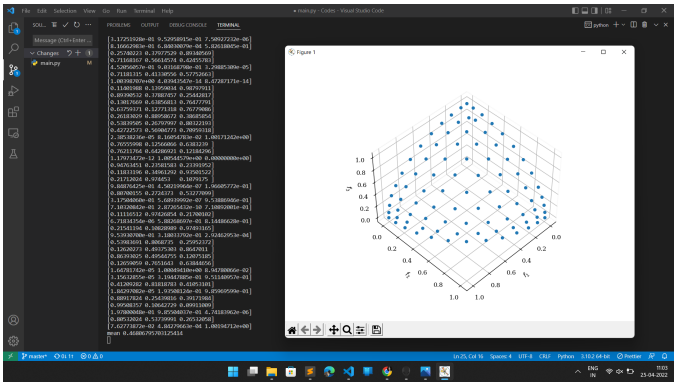

Fig. 9. Result with DTLZ3

REFERENCES

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

[2] E. Hancer, B. Xue, and M. Zhang, “Differential evolution for filter feature selection based on information theory and feature ranking,” Knowl. Based Syst., vol. 140, pp. 103–119, 2018.

[3] D. Koller and M. Sahami, “Toward optimal feature selection,” Technical Report 1996-77, 1996.

[4] Y. Zhang, D. Gong, Y. Hu, and W. Zhang, “Feature selection algorithm based on bare bones particle swarm optimization,”Neurocomputing, vol. 148, pp. 150–157, 2015.

[5] https://www.sciencedirect.com/topics/mathematics/imputation-method: :text=Mea..

[6] https://pymoo.org/algorithms/moo/nsga3.html.

[7] https://deap.readthedocs.io/en/master/examples/nsga3.html.

[8] https://en.wikipedia.org/wiki/Multi-objectiveoptimization

[9] 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.