International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Feature Extraction and Independent Subset Generation Using Genetic Algorithm for Improved Classification

P. Rathna Sekhar [1], Dr. B. Sujatha*[2]

Submitted: 13/11/2022 **Accepted**: 15/02/2023

Abstract: The number of traits that can be retrieved from the vast amounts of data of various forms available today is enormous. This is especially true for text data, which has benefited from the proliferation of multimedia applications. Using every available feature for each of the classification tasks can be not just time-consuming but also performance-detrimental. When a number of measurements, or features, have been acquired from a set of objects in a standard statistical pattern recognition problem, feature extraction is a frequent technique employed prior to classification. To achieve this, it is necessary to create a mapping from original representation space to a new space in which the classes may be more readily distinguished from one another. Selecting an appropriate feature set to characterize the patterns being classed is a common requirement in challenges involving knowledge discovery and pattern classification. This is because the classifier's performance and the cost of classification are both highly sensitive to the features used for the classifier's construction. To identify near-optimal solutions to such optimization issues, Genetic Algorithms (GA) present an appealing strategy. High-quality approaches to optimization and exploration issues can be quickly and easily generated using genetic algorithms, which rely on bioinspired operators including mutation, crossover, and selection. The genetic algorithm is an approach to solve optimization problems with constraints and without them, inspired by natural selection, the mechanism behind biological evolution. The genetic algorithm iteratively improves upon a pool of candidate solutions. A genetic learning and evolution model is used to pick or extract features while simultaneously designing a classifier. In this research Independent Subset Generation using Genetic Algorithm for Improved Classification (ISG-GA-IC) model is proposed for accurate selection of independent features for enhancing the classification levels. In this paper, we provide the results of our studies on the use of evolutionary algorithms for feature extraction and selection in high-dimensional data sets.

Keywords: *Feature Extraction, Feature Selection, Genetic Algorithm, Classification, High Dimensional Data, Independent Features.*

1. Introduction

One particularly effective method of machine learning is the Genetic Algorithm (GA), which draws its inspiration from Darwinism. Recent advancements in GA have covered a broad range of applications, from system identification and optimization to prognosis and diagnostics to data classification and feature selection and even image processing [1]. Although GA has been shown to be effective in carrying out heuristic optimization, its evolutionary searching nature carries the possibility of premature convergence [2]. Lack of variation in candidate selection as well as a fixed genetic operator setting are likely to blame for the prevalence of local optimums. Selection, crossover, & mutation, along with GA stopping criteria, make up the genetic operator [3]. The importance of time-consuming parameter tuning is also unknown, not only because a global optimization techniques solution is highly unlikely, but also because the number of possible combinations grows exponentially with respect towards the genetic parameters being adjusted [4].

Most statistical pattern detection algorithms need the selection of suitable features as an initial step. Classification is simplified with the aid of a competent feature selection method, which gets rid of irrelevant or distracting features that could otherwise get in the way of recognition. When there is a scarcity of training data, the inclusion of features that provide some relevant information can actually lower the accuracy of the classifier [5]. With the cost of measuring and including features and the so-called curse of dimensionality, it is clear that finding the smallest amount of characteristics that still allows a classifier to distinguish between pattern classes is valuable [6]. By defining additional features in terms of the existing feature set, this feature weighting technique extracts useful information for more precise pattern identification. Combining the classification rules with feature selection and extraction has been demonstrated to improve accuracy with the traditional classifiers [7]. This can help with the study of big datasets by isolating feature combinations that differentiate effectively between pattern classes [8].

A set of data features are employed in a decision-theoretic or statistical method to pattern recognition to classify or describe the data. Classifier design is profoundly influenced by the processes of feature selection and extraction [9], making them essential to maximizing performance. Experts in the field of application are usually needed to help define the most relevant

Research Scholar[1], Assistant Professor[2]*
Department of Computer Science and Engineering
University College of Engineering, Osmania University,
Hyderabad.

characteristics. In reality, most high-dimensional, complicated patterns contain a lot of noise and duplication [10]. For this reason, settling on a minimum or optimal set of features can be challenging for even the most knowledgeable of professionals. In statistical patterns recognition, artificial neural network, and other learning approaches, the curse of dimensionality becomes an irritating occurrence [11]. Numerous studies have shown that many learning processes do not possess the quality of scalability, meaning that they either do not work or yield unacceptable results when applied to issues of a larger size [12]. To solve this scaling issue, a method for the automatic selection and extraction of features using genetic algorithms is proposed. This method relies on a looping connection among feature assessment and classification as its foundation for functioning [13].

Genetic learning and evolution is used to do feature transformations and classifier creation simultaneously [14]. The goal of this method is to discover a subset of the original N characteristics that is sufficient for class discrimination but leaves out unnecessary class information and/or noise [15]. By reshaping the original feature space of the data, users can create a new feature space with fewer characteristics that provide better separation of the pattern classes and, in turn, improve the efficiency of the decision-making classifier [16]. Misclassification probability is typically used as a metric for determining the quality of the feature subset chosen. Exhaustive search is computationally impractical due to the vast number of possible feature subsets [17] N. Consequently, alternative methods must be investigated. It is unclear under what conditions any one heuristic should be implemented in the area of pattern recognition, as they all have their pros and cons. The general architecture of the GA model is shown in Figure 1.

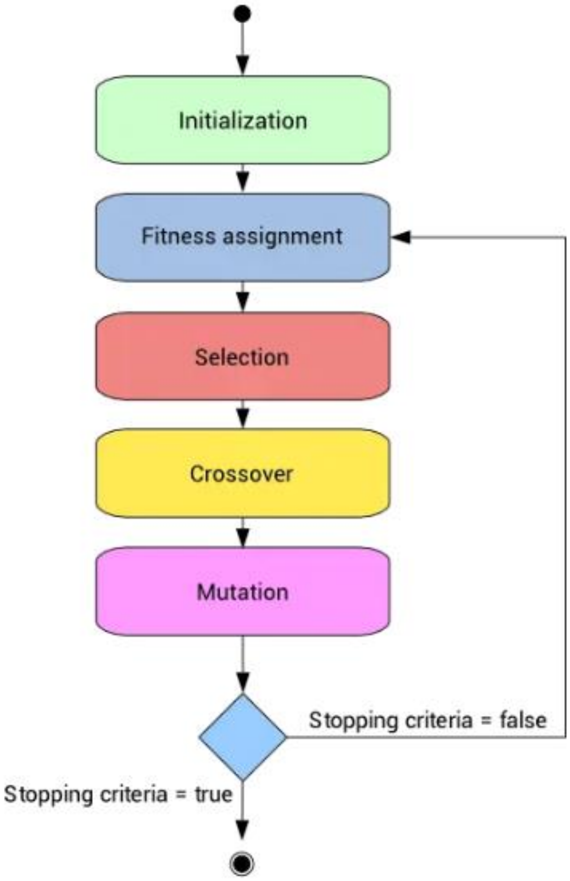

Fig 1: General Process of GA in Feature Selection

Feature subset selection is an example of a multi-criteria optimization problem that arises in the context of numerous real-world issues. Classification accuracy, cost, and risk are three of the many variables to be maximised, all of which are tied to the elements chosen to characterise the patterns. If users trying to optimise for multiple variables at once, genetic algorithms are a great option [18]. With their potential for concurrency and fault and noise tolerance, neural networks present an appealing framework for the creation of trainable patterns classifiers for real-world real-time pattern classification applications [19]. While evolutionary algorithms excel at rapidly exploring broad search areas for challenging optimization issues, neural networks provide an appealing means of honing in on possible solutions [20]. Therefore, it is tempting to investigate the use of hybrid approaches that combine global and local search strategies for solving complex design or optimization problems [21]. Given this context, it is easy to see why genetic algorithms are being explored for use in the selection of feature subsets in the development of neural network pattern classifiers.

2. Literature Survey

The use of microarray gene expression information in cancer classification and other areas of bioinformatics research has recently gained significant attention. However, the large p, small n paradigm of scarce biosamples and high-dimensional data is making gene selection, the process of choosing a small subset of genes that are strongly linked to a phenotype, a difficult and time-consuming endeavour. Most existing algorithms for selecting features rely on heuristic rules because of the nondeterministic time complexity of the problem of feature or gene selection. Here, Peng et al. [1] proposed MGRFE, a multi - layer feature elimination approach to use an encapsulated integer-coded genetic algorithm to find the smallest, most informative gene combination. MGRFE outperforms state-of-the-art algorithms for feature selection with higher cancer classifier and a smaller number of selected genes based on 19 benchmark microarray datasets, including multiclass & imbalanced datasets. For high-dimensional datasets, such as gene expression data, MGRFE may prove to be a useful technique for selecting features to focus on. Further, MGRFE-selected genes have strong biological significance to cancer phenotypes.

The fast growth of cyber physical system is due to their use in collaborative application areas of physical systems (CPS) as well as algorithms. However, due to the current state of the art, it is challenging to establish CPS with picture classification systems, as both deep learning techniques and traditional techniques for feature extraction exist as separate and distinct entities. For this reason, Wang et al. [2] proposed a quick feature fusion method to meet the need of CPS in the field of image classification. To ensure high accuracy with low training time and hardware cost, the author first used a genetic algorithm to combine features from a shallow-layer network, a big pre-trained convolutional neural network, and conventional image processing techniques. Second, the author employed a dynamically assigned weighting scheme to increase the gap between the centres, thereby enhancing class discrimination. Finally, the author proposed a partial method of selecting to further enhance classification accuracy by centralising characteristics within the same class and shorten the length of the fusion feature vectors. Finally, the experimental results demonstrate that this method can improve the efficiency and adaptability of image classification throughout cyber systems by achieving better classification accuracy with much less training time as well as hardware consumption.

Classifying texts in natural scenes is difficult due to the wide variety of content types, the presence of image degradations like noise and low contrast/resolution, and the unpredictable occurrence of center of the images and background properties. One of the most significant challenges in such endeavours is the high dimensionality of the feature space derived from the input image. The purpose of this work is to address these issues and enhance the classifier's generalisation capabilities by removing superfluous and irrelevant features. In other words, dimensionality reduction is achieved through the selection of a subjective & discriminative set of features. In this study, Ansari et al. [4] employed a genetic algorithm inspired by biology because the crossover used by such an algorithm greatly enhances the quality of a multimodal discriminative collection of features, leading to more precise classifications of images from a wide variety of natural scenes. Classification is performed using the Support Vector Machine (SVM) algorithm, with an average F-Score serving as the fitness function and target condition. At first, multimodal feature representation is used to construct the entire feature space after input images have been pre processed. As a second step, the author fused the features at the feature level. Third, a meta-heuristic optimization method employing a GA for feature selection is used to raise the classifier's average F-score. Five publicly available datasets are used to evaluate the proposed algorithm and compare it to several state-of-the-art approaches. The results revealed that the proposed methodology outperformed state-of-the-art benchmark algorithms in classifying textual & non-textual regions.

Nag et al. [5] examined genetic programming's potential to choose and retrieve linearly separable characteristics whenever the enabling model is instructed to accomplish the very same, and then proposed a unified system for doing so. Using steady-state multiple objective genetic programming to three minimisation objectives, the author break down a c-class problem in to other c binary classification and afterwards evolve c pairs of binary classifiers. Every one of the binary classifiers is a combination of a tree structure and a linear SVM. The SVM is trained using the extracted features by feature nodes and a few of the component nodes of the tree. When determining a classification, the SVM's decision is accepted as the final verdict. The SVM-weights are used to rank the relative importance of nodes during crossover and mutation. To further refine the feature selection, the author employed a fitness function focused on Golub's index. The author used unfitness functions on the feature nodes to get rid of the ones that are rarely used.

When it comes to heuristic optimization, GA is indeed a tried and true machine learning technique. The problem with this method is that it tends to converge too quickly, especially on the local optimum. Small population diversity and a constant genetic operator setting are to blame for the presence of a performance plateau. Because of the prevalence of inadequate solutions in feature selection research papers, Ooi et al. [8] presented this Self-Tune Linear Adaptive-GA (STLA-GA), an adaptive algorithm. STLA-GA uses an iterative exploration-exploitation process to fine-tune the stopping criterion, population numbers, maximum generation number, and novel convergence threshold. Based on the most recent classifier output, STLA-GA initiates an iterative cycle of exploration and exploitation. The proposed STLA-GA has many advantages over conventional feature reduction methods, such as avoiding the need for manual parameter tuning, reducing the likelihood of premature convergence, and avoiding unnecessary computational cost caused by unstable parameter estimation feedback.

Expert system (ES)-based fault diagnosis is still a topic of study in Industry 4.0 production, with higher interpretability. In the ES, the knowledge base is the central component, and the quality of the knowledge determines how well problems are diagnosed. However, it is challenging for users to make sense of the knowledge extracted from the initial dataset with the current knowledge extraction method. Because of this, it is crucial to derive clear and precise rules from of the NN architecture. To address this shortcoming, Ou et al.[9] suggested a hybrid extraction conceptual model to carry out the rule extraction. To start, LAGA, an enhanced adaptive GA that makes use of a logistic function, is proposed as a means of addressing the problem of inadequate prediction performance encountered when using conventional GAs. The experimental outcomes of optimising six chosen test operations by these GA variants demonstrate that the LAGA algorithm's convergence accuracy & speed have been significantly improved, particularly for high latitude functions, compared to the three other mainstream adaptive GAs. This manuscript discusses the LAGA-BP framework, a new feature extraction method that uses the symbol rule and NN to categorise real-valued attributes. This framework uses NN to unearth tacit knowledge and then applies rule-based reasoning to simplify the information. It is possible to divide the LAGA-BP framework's implementation into an initial phase and an iterative phase. It begins with optimising a back propagation neural network (BPNN) with the Least Absolute Gain Algorithm (LAGA) and then using that network to refine prediction classification knowledge. The second stage of this framework uses the multi-layered NN using two distinct superposed systems to reduce the attributes of the data set, and then applies the K-means clustering algorithm to obtain the if-then rule from the simplification characteristics.

Over-fitting can occur when there are too many input characteristics in an application, which can hinder this same effectiveness of the learning methodology. In most situations, the prediction target responds differently to features that have different information contents. Therefore, in this paper, Li et al. [14] proposed a feature selection method, dubbed WKNNGAFS, to compute the value of each feature. This strategy uses a GA to find the best weight vector, where the value of the ith element indicates how important that feature is to the classification as a whole. In addition, the target label is established by employing the weighted K-nearest neighbours algorithm (WKNN), which considers both the contributions of nearest neighbours and the categorization ability of each feature. Using 13 real-world datasets, including 6 high-dimensional microarray datasets, nine original feature selection techniques are compared with the proposed method to determine its efficacy. The experimental results indicate that the technique is more efficient and can boost classification accuracy.

Pattern recognition relies heavily on accurate feature extraction from range images supplied by ranging sensors. An enhanced method utilising genetic clustering VGA-clustering has been described for automatically extracting a environmental feature detected by just a 2D ranging sensor laser scanner. Instead of using a traditional clustering algorithm, Jinxia et al. [17] introduced a weighted fuzzy based algorithm that takes into account the spatial proximity of range data to accomplish feature extraction in a laser scanner. Several validation index features are used to approximate the validity of distinct clustering algorithms, and one confirmation index is chosen also as fitness value of genetic algorithm in order to automatically determine the correct clustering number based on the input data. To solve the optimum solution of the clustering algorithm, the author also proposed an enhanced genetic algorithm IVGA based on VGA. This algorithm improves upon its predecessor by increasing population diversity and refining the genetic algorithms of elitist rule in order to boost local search ability and accelerate convergence time. The efficiency of the introduced algorithm is demonstrated through comparison to other algorithms.

3. Proposed Model

Genetic algorithms, evolutionary programming, and other related methods all fall under the umbrella of evolutionary algorithms, which is a class of randomised, population-based heuristic search strategies. Methods that mimic biological evolution serve as inspiration. Fundamental to evolutionary systems of this type is the concept of a population of individuals representing potential solutions in a high-dimensional search space [22]. The users stand in for potential answers to the optimization challenge. Depending on the problem space, a wide variety of genetic representations can be employed to encode the individuals. Commonly, n-bit binary vectors are used to represent individuals in genetic algorithms [23]. It's like searching in a boolean space of n dimensions, where n is the number of possible outcomes.

Each person stands in for a feature subset in the feature subset selection issue. It is expected that a fitness function can be used to rank the potential solutions in a population by their quality. The fitness function would rank the prioritised features based on some metric in the feature subset selection problem [24]. To generate people for the next generation, evolutionary algorithms use a method of fitness-dependent probability selection from the present population. In the past, researchers have experimented with a wide range of selection methods [25]. Selection methods such as those based on physical prowess, academic achievement, or position in a league or tournament are quite popular. To maximise the likelihood of success in solving the optimization problem at hand, genetic operators are typically crafted to take advantage of specific features of the genetic representation, the search space, and the challenge at hand. The programme is able to search for optimal solutions by using genetic operators. The proposed model framework is shown in Figure 2.

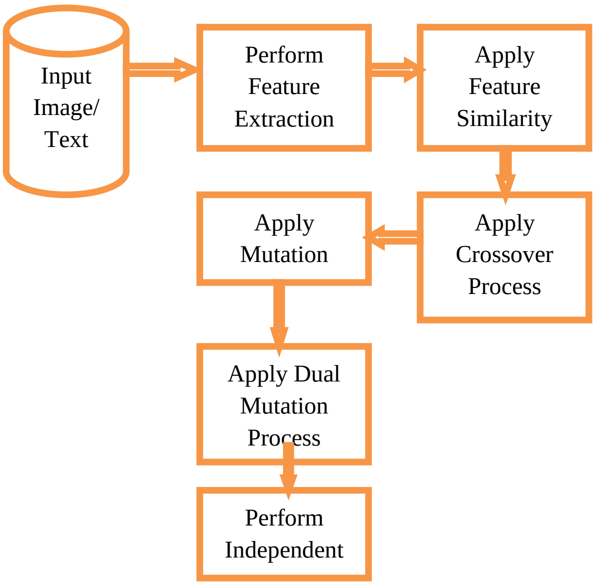

Fig 2: Proposed Model Framework

In contrast, the process of crossover uses two parent strings to generate four new strings. Two strings, 01101 and 11000, produce 01100 and 11001 when crossed at position 4. To represent other types of genetic data, special genetic operators need to be developed. Many generations of individuals are produced by applying genetic operators and engaging in fitness-dependent selection until an optimal solution is reached. Evolutionary algorithms of the type described above can be shown to effectively simulate the highly opportunistic and exploitative randomised search that searches high-dimensional search spaces. Several user-defined parameters, such as population size, possibility of application of various genetic operators, etc., affect the performance of evolutionary algorithms in practise. These include the fitness function, the information of the fitness-dependent selection method, and the genetic representation and operators chosen.

The fitness function of the model is used as the basis for the search for the best subset of features to use in the GA. As each feature is encoded with a binary value, GA provides a simple modelling to the input representation. The value 0 indicates de-selection, while the value 1 indicates selection.

3.1. The initial setup

To begin, we are going to resize all of the either positive or negative text images in the input set to 100 by 100 pixels each. Example data from the benchmark sets are used to illustrate the manual cropping process used on these samples. To proceed, we are going to grayscale the colour photos. Then, we equalise the histogram of high-resolution images to improve their contrast and better distinguish between text and non-text.

3.2. Feature Fusion and Extraction

To construct a multimodal feature representation, it is essential to extract multiple attributes from an image for use in a classification problem. To obtain sufficient supervised information for just a classification problem, the multimodal approach is usually capable of representing most image properties. Noise, non-universality, inter- and intra-class variations, as well as spoofing attacks can all affect the obtained supervised information.

3.3. Model for Combining Several Features

We have learned how to extract two distinct CNN relevant features from an image in Section A. In addition, we have many other conventionally-derived image features, such as the Haar feature for facial recognition. Here, we present the GA-based multi-feature fusion model. Different uses and requirements will place varying priorities on various features.

In this research Independent Subset Generation using Genetic Algorithm for Improved Classification (ISG-GA-IC) model is proposed for accurate selection of independent features for enhancing the classification levels. The proposed model working is clearly discussed in the algorithm.

4. Algorithm

Algorithm ISG-GA-IC

{

Input: Prediction Dataset {PDset}

Output: Independent Subset Generation {ISGset}

Step-1: The records in the dataset are loaded and the analysis is performed for extracting the features. The records in the analysis are performed to check the attributes range. The analysis is performed as

$$AttrAnal[PDset(R)] = \sum_{f=1}^{R} \frac{getattr(f)}{len(f)} + getMaxattr(f) + mean(attr(f)) - \lambda(f) + Th$$

Here λ is the model that considers the null set values that will be removed before processing. This is the threshold value,getattr model is used to read all the record values in the dataset.

Step-2: Reducing data redundancy is one of the many benefits of feature extraction. Data reduction helps to create the model with much less machine effort and accelerates the learning and generalisation phases of the learning process. The feature extraction considers the features from the dataset as

$$Fextr(AttrAnal(R)) = \sum_{f=1}^{R} \frac{\max\ (\mathrm{AttrAnal(f)})}{\min\ (\mathrm{getattr(f)})} + \sum_{f=1} getcorr(attr(f.f+1)) + \frac{maxattr(f) - minattr(f)}{len(PDset)} + M(f.f+1)$$

Here f is the current feature, getcorr is the model used for the correlation checking among the features and the maxattr models is used for consider the maximum range attributes that are similar.

Step-3: Feature Correlation is a statistical method for determining the degree of linearity between two or more independent variables. As a result of a correlation between two variables, it is possible to predict a best variable. Correlation-based feature selection is predicated on the idea that useful variables will share a high degree of similarity with the target. The feature correlation checking for similar feature detection is calculated as

$$\mathrm{FCorr(R)} = \frac{\sum_{\mathrm{f}=1}^{\mathrm{K}} \delta\left(\mathrm{Fextr(f,f+1)}\right) - \min\ \left(\mathrm{Corr}\left(\mathrm{Fextr(f,f+1)}\right)\right)}{\sum_{\mathrm{r}=1}^{\mathrm{M}} \mathrm{len}\left(\mathrm{Fextr(f)}\right) + \mathrm{maxattr}(AttrAnal(\mathrm{f}))}$$

Step-4: An individual's feature fitness level is defined by the value of their fitness function. The highest fitness value for a populace is the minimal weight factor for any feature individual within the population, and the feature subset locates the maximum of the fitness function. The fitness value is used for the best feature set selection. The process of fitness value assignment is performed by considering the feature set attributes as {F_1,F_2,…….,F_n}. The fitness attribute is assigned to each feature attribute as

μ(f)= max(Fcorr(i)) where i=1,2,3,….,N

The probability of each feature attribute is calculated as

$$Prob_f = \sum_{f=1}^{R} \frac{\mu(f)}{\sum_{i=1}^{f} f * \max\ (\mu)}$$

Step-5: The crossover operator is a type of genetic operator that results in a new chromosome by mating two existing ones from different organisms. By incorporating the best features of both parents, the resulting chromosome has the potential to outperform them both. A system's crossover acts as a filter, blocking out frequencies that aren't wanted. Mutation in GAs is used to increase variety in a population that has been sampled from. The goal of using mutation operators is to prevent the population of chromosomes from becoming excessively similar to each other and so slowing or stopping convergence to the global optimum, thereby avoiding local minima. The mutation operation and dual mutation operations are performed to get the best features. The process is of crossover and mutation and dual mutation is performed as

Input: γ feature sequences

Size 'S' of feature populations for crossover

Elitism 'E'

Rate 'M' of mutation

Iterations Count 'I' starting from 1.

Temp variable T starting from 1.

Output: Optimized Feature Set

For I in range(R) do

Create Child C=(S.E)

Calculate the best feature set 'FE' from the available S populations as

$$FE(R) = \sum_{f=1}^{R} \frac{simm(f.f+1)}{len(C)}$$

Perform crossover set as $COset(R) = \sum_{f=1}^{R} \frac{(S-C)}{2}$

For j in range(C) do

$$Fset_1 = \sum_{f=1}^{R} getRand(S) + \gamma$$

$$Fset_2 = \sum_{f=1}^{R} \gamma + getRand(S)! = Fset_1$$

Perform the crossover generation with one point crossover set to the $Fset_1$ and $Fset_2$ as

$$OPCO[(OPC_1.OPC_2)] = \sum_{f=1}^{R} Fset_f - (f-1) + E$$

Update the feature set S as max(OPCO(R))

End for

For j in range(C) do

BFset←max(S)

Mutate each feature attribute of rate M →Uset

```
If Uset is unfeasible
 Update Uset
End
Update M as M←rand(Uset,M)
S=S+Uset+max(M)
For j in range(C+M) do
BFset←max(S)
Mutate each feature attribute of rate M →min(Uset,j)
If Uset is unfeasible
 Update Uset←Uset-j
End
I=I+1
Fset=max(S)+E
end
```

Step-7: Data analysis concerns often benefit from feature selection techniques, which try to narrow down the variables describing the data to a more meaningful and concise collection. By using feature selection, results can be improved. By focusing on the most important factors and leaving out the features, it improves the algorithms' ability to make predictions. The Independent Subset Generation process considers the most important features. The Independent Subset Generation is performed as

$$\mathrm{ISubset}(Fset[R]) = \sum_{f=1}^{R} \gamma\left(\frac{\mathrm{maxattr(BFset(f))}}{\mathrm{S + Uset + max(M)}}\right) * \sum_{i=1}^{J} \frac{\sum_{i=1}^{R} Fset_i - (i-1) - minattr(Uset)}{len(BFset)}$$

4. Results

Feature selection is a crucial part of the preliminary processing of the data in the dataset. Feature selection seems to revolve around eliminating redundant or highly related traits, as well as those that provide little to no predictive value. Over the past few years, many meta-heuristic techniques have been developed to help remove as many superfluous features from high-dimensional datasets as feasible. The correlation between a subset of variables is commonly overlooked in current meta-heuristic based techniques, which is one of their key drawbacks. In this research, GA based feature selection model is proposed for accurate feature selection.

Natural selection and genetics form the basis of GAs, which are heuristic search algorithms. The goal is to develop a solution to a problem in a way that is similar to how biological mechanisms like survival of the fittest work. With GA, a population of chromosomes is used to evolve into a new population through the processes of selection, crossover, and mutation. Genes can be found on each chromosome. Crossover and mutation are analogous to the biological processes that create diversity in a population, and selection operators pick the healthiest members of the population. Crossover and mutation, on the other hand, are exploratory processes that help shape populations while selection is exploitative. In this research Independent Subset Generation using Genetic Algorithm for Improved Classification (ISG-GA-IC) model is proposed for accurate selection of independent features for enhancing the classification levels. The proposed model is compared with the traditional meta-heuristic optimization technique using genetic algorithm (MHOP-GA) [4]. The comparison results indicate that the performance of the proposed model is better than the existing models. The dataset is considered for extracting the features for training the model. The records in the dataset are analyzed for processing and attribute checking. The Dataset Records Analysis Time Levels of the proposed and existing models are shown in Figure 3.

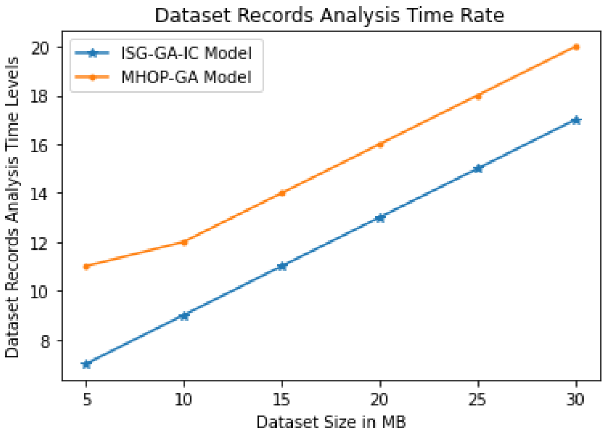

Fig 3: Dataset Records Analysis Time Levels.

The term feature extraction is used to describe the procedure of reducing unstructured data to a set of quantifiable features that may be further processed without losing any of the original data's context. When compared to using machine learning on unprocessed data, the results are far more favourable. Better performance in the categorization is only possible with the help of feature extraction and feature reduction. To improve the classifier's performance, it is necessary to extract the most relevant and concise set of features. The figure 4 represents the Feature Extraction Accuracy Levels of the proposed and traditional models.

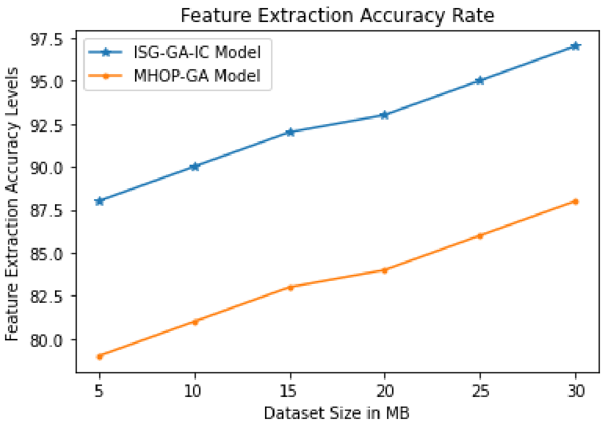

Fig 4: Feature Extraction Accuracy Levels

The feature similarity check will be done for all the extracted features to identify the most useful features that are less correlated. This feature subset helps to train the model for accurate predictions. The Feature Similarity Checking Accuracy

Levels of the proposed and existing models are shown in Figure 5.

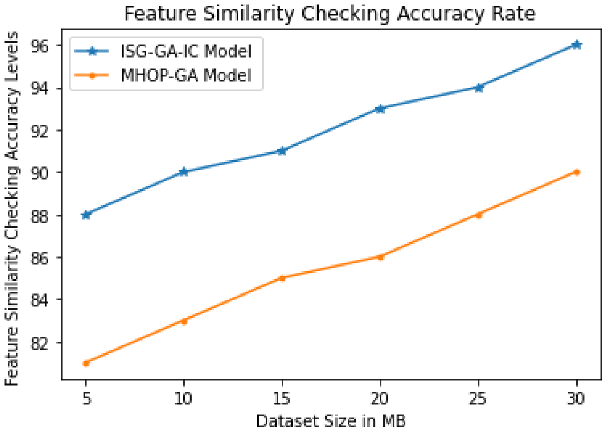

Fig 5: Feature Similarity Checking Accuracy Levels.

A fitness assignment value is a model of an objective function, and it is used to quantify, in a single score, how well one design solution compares to others in terms of how closely they come to accomplishing the specified goals. Genetic algorithms and genetic programming utilise fitness factors to optimise design simulations. The Fitness Assignment Time Levels of the existing and proposed models are shown in Figure 6.

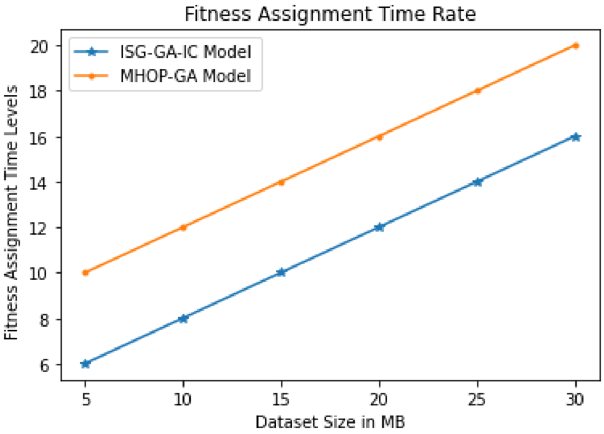

Fig 6: Fitness Assignment Time Levels

Crossover is the process by which two individuals with different designs can exchange traits to create children. With crossover, chromosomes from both parents are exchanged, increasing the likelihood of the children being superior. The process of crossover in the proposed and existing models accuracy levels are shown in Figure 7.

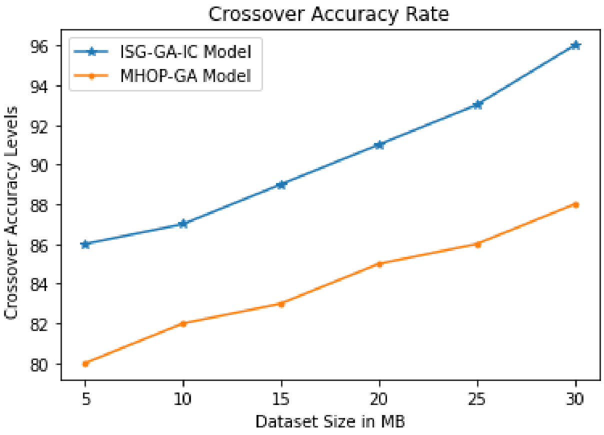

Fig 7: Crossover Accuracy Levels

In a genetic algorithm, or evolutionary algorithm (EA) more generally, mutation is a genetic operator that helps keep the chromosomes of a EA. It's like a mutation in nature. The initial step in evolution is mutation, which results in a new allele by altering the DNA sequence of a particular gene. Intragenic recombination is another way in which recombination can generate a novel DNA sequence for a given gene. The Mutation Time Levels of the proposed and traditional models are shown in Figure 8.

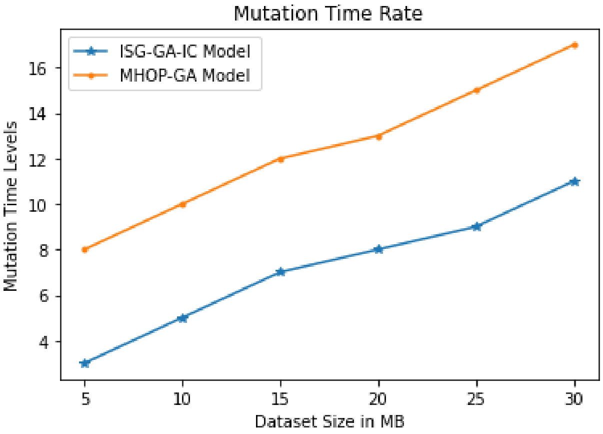

Fig 8: Mutation Time Levels

Users can swap the values of two genes on a chromosome by playing the Dual Mutation model. Dual Mutation involves randomly rearranging the numerical expression of a chosen number of genes. It's possible that the chosen genes are spread out. Intragenic recombination is another way in which recombination can generate a novel DNA sequence with the double mutation operation for a given gene. The Figure 9 shows the Dual Crossover Accuracy Levels of the proposed and existing models.

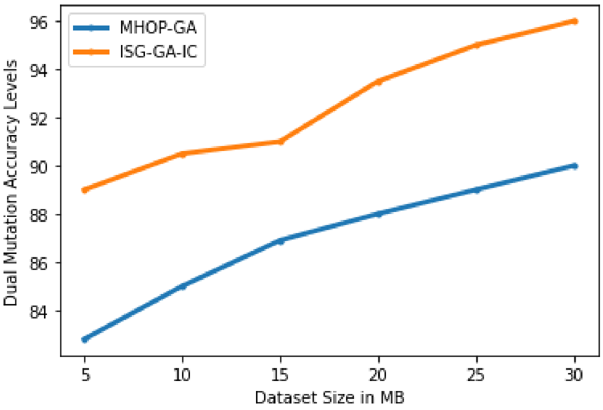

Fig 9: Dual Mutation Accuracy Levels.

No other independent set can be subset of a maximal independent feature set or maxim stable set. In other words, it is maximum with regard to the separate set property, meaning that there is not a feature outside of the independent collection that may join it. With the independent subset, the training process can be performed for accurate predictions. The Independent Subset Generation Accuracy Levels of the proposed and existing models are shown in Figure 10.

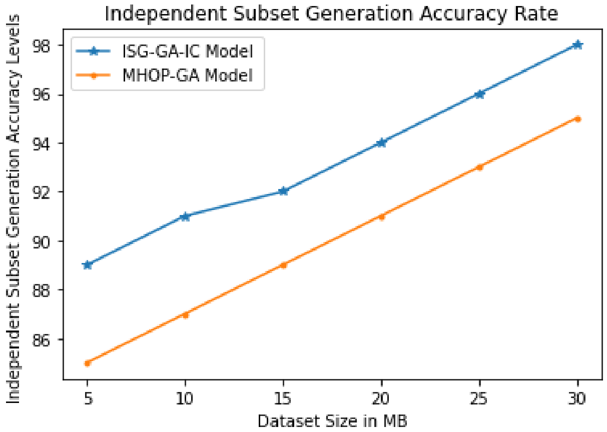

Fig 10: Independent Subset Generation Accuracy Levels.

5. Comparative Analysis

Dataset Records Analysis Time Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Time Levels	Dataset Size in MB	Time Levels
	5	11	5	4
	10	11.7	10	9
	15	13.78	15	10.25
	20	15.65	20	11.75
	25	17.85	25	13.87
	30	19.8	30	15

Table 1: Dataset Records Analysis Time Levels

Feature Similarity Checking Accuracy Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	FSC Accuracy Levels	Dataset Size in MB	FSC Accuracy Levels
	5	80	5	88
	10	82.5	10	89.5
	15	84.3	15	90.23
	20	84.5	20	92.4
	25	86.4	25	93.5
	30	88	30	95.89

Table 2: Feature Similarity Checking Accuracy Levels

Feature Extraction Accuracy Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Extraction Accuracy Levels	Dataset Size in MB	Extraction Accuracy Levels
	5	78	5	87.9
	10	80.25	10	89
	15	82.5	15	91
	20	83.5	20	92.5
	25	85	25	93.5
	30	86	30	96.3

Table 3: Feature Extraction Accuracy Levels

Fitness Assignment Time Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Fitness Assignment Time Levels	Dataset Size in MB	Fitness Assignment Time Levels
	5	10	5	6
	10	11.90	10	7.90
	15	13.5	15	9.20
	20	15.25	20	10.25
	25	17.9	25	12.30
	30	19.90	30	14.10

Table 4: Fitness Assignment Time Levels

Crossover Accuracy Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Crossover Accuracy Levels	Dataset Size in MB	Crossover accuracy Levels
	5	80	5	86
	10	82	10	86.80
	15	83.25	15	88
	20	84	20	90
	25	85	25	92
	30	86	30	96

Table 5: Crossover Accuracy Levels

Mutation Time Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Mutation Time Levels	Dataset Size in MB	Mutation Time Levels
	5	8	5	2.30
	10	9.34	10	4.20
	15	11.90	15	6.10
	20	12.25	20	7.80
	25	14.20	25	8.0
	30	16	30	8.30

Table 6: Mutation Time Levels

Dual Mutation Accuracy Levels	MHOP-GA (Existing Model)		ISG-GA-IC (Proposed Model)	
	Dataset Size in MB	Dual Crossover Accuracy Levels	Dataset Size in MB	Dual Crossover Accuracy Levels
	5	82.8	5	89
	10	85	10	90.5
	15	86.9	15	91
	20	88	20	93.5
	25	89	25	95
	30	90	30	96

Table 7: Dual Mutation Accuracy Levels

Independent Subset Generation Accuracy Levels	**MHOP-GA** (Existing Model)		**ISG-GA-IC** (Proposed Model)	
	Dataset Size in MB	**ISG Accuracy Levels**	**Dataset Size in MB**	**ISG Accuracy Levels**
	5	**84**	**5**	**89**
	10	**86.5**	**10**	**90.5**
	15	**88.20**	**15**	**91**
	20	**89**	**20**	**93**
	25	**92**	**25**	**94.50**
	30	**93**	**30**	**97**

Table 8 : Independent Subset Generation Accuracy Levels

6. Conclusion

A population of potential answers to the problem at hand is the result of running n generations of GAs. This process of recombination and mutation produces offspring from these solutions, which are subsequently passed down through generations. The objective function is used to assign a fitness value to each individual, with the fitter individuals being given a greater reproductive advantage. This conforms to the Survival of the Fittest theory put forward by Charles Darwin. Although genetic algorithms are inherently random, they outperform random local search by making use of both current and past data. This study suggests a method for choosing feature subsets for classifiers using a genetic algorithm. Inductive learning of classifiers in general, and neural network classifiers in particular, offer an intriguing method to tackling the feature subset selection problem, according to the results reported in this chapter. Traditional approaches to feature selection frequently have limited applicability to real-world classification and knowledge acquisition tasks because they make use of monotonicity assumptions. The GA-based approach to feature subset selection does not make use of these assumptions. It also provides a natural method for choosing feature subsets by accounting for the distribution of the data that is readily available.This is because feature selection is determined by estimated fitness values, which, when based on numerous partitions of the dataset into training and test data, provide an accurate indicator of how well the feature subset performed. Many greedy stepwise algorithms that choose features based on a single division of the data into training and test sets typically do not do this. As a result, it is likely that the feature subsets chosen by such methods will perform fairly badly on additional random data partitions into training and test sets. Multiple criteria can be organically incorporated into the feature selection process using the method to feature subset selection. In this research Independent Subset Generation using Genetic Algorithm for Improved Classification model is proposed for accurate selection of independent features for enhancing the classification levels. The proposed model achieves 97% accuracy in feature subset generation that considers the best features for training the models. In future multi-dimensional feature reduction techniques can be applied and optimization techniques can also be considered for enhancing the accuracy in feature reduction models.

References

[1] C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang and Y. Li,

[2] "MGRFE: Multilayer Recursive Feature Elimination Based on an Embedded Genetic Algorithm for Cancer Classification," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 2, pp. 621-632, 1 March-April 2021, doi: 10.1109/TCBB.2019.2921961.

[3] Y. Wang, B. Song, P. Zhang, N. Xin and G. Cao, "A Fast

[4] Feature Fusion Algorithm in Image Classification for Cyber Physical Systems," in IEEE Access, vol. 5, pp. 9089-9098, 2017, doi: 10.1109/ACCESS.2017.2705798.

[5] Y. Zhang, Q. Wang, D.-W.Gong and X.-F. Song,

[6] "Nonnegative Laplacian embedding guided subspace learning for unsupervised feature selection", *Pattern Recognit.*, vol. 93, pp. 337-352, Sep. 2019.

[7] G. J. Ansari, J. H. Shah, M. C. Q. Farias, M. Sharif, N.

[8] Qadeer and H. U. Khan, "An Optimized Feature Selection Technique in Diversified Natural Scene Text for Classification Using Genetic Algorithm," in IEEE Access, vol. 9, pp. 54923-54937,2021,doi: 10.1109/ACCESS.2021.3071169.

[9] K. Nag and N. R. Pal, "Feature Extraction and Selection for

[10] Parsimonious Classifiers withMultiobjective Genetic Programming," in IEEE Transactions on Evolutionary Computation, vol. 24, no. 3, pp. 454-466, June 2020, doi: 10.1109/TEVC.2019.2927526.

[11] A.Khan, A. S. Qureshi, N. Wahab, M. Hussain and M. Y.

[12] Hamza, "A recent survey on the applications of genetic programming in image processing" in arXiv preprint arXiv:1901.07387, 2019.

[13] Y. Bi, B. Xue and M. Zhang, "An automated ensemble

[14] learning framework using genetic programming for image classification", *Proc. Genet. Evol.Comput.Conf.*, pp. 365-373, 2019.

[15] C. S. Ooi, M. H. Lim and M. S. Leong, "Self-Tune Linear

[16] Adaptive-Genetic Algorithm for Feature Selection," in IEEE Access, vol. 7, pp. 138211-138232, 2019, doi: 10.1109/ACCESS.2019.2942962.

[17] Y. Ou, S. -Q. Ye, L. Ding, K. -Q. Zhou and A. M. Zain,

[18] "Hybrid Knowledge Extraction Framework Using Modified Adaptive Genetic Algorithm and BPNN," in IEEE Access, vol. 10, pp. 72037-72050, 2022, doi: 10.1109/ACCESS.2022.3188689.

[19] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin and S. L.

[20] Fernandes, "A framework for offline signature verification system: Best features selection approach", *Pattern Recognit. Lett.*, vol. 139, pp. 50-59, Nov. 2020.

[21] Y. Zhang, D.-W.Gong, X.-Z.Gao, T. Tian and X.-Y. Sun,

[22] "Binary differential evolution with self-learning for multi-objective feature selection", *Inf. Sci.*, vol. 507, pp. 67-85, Jan. 2020.

[23] Y. Bi, B. Xue and M. Zhang, "Genetic Programming With

[24] a New Representation to Automatically Learn Features and Evolve Ensembles for Image Classification," in IEEE Transactions on Cybernetics, vol. 51, no. 4, pp. 1769-1783, April 2021, doi: 10.1109/TCYB.2020.2964566.

[25] B. P. Evans, "Population-based ensemble learning with

[26] tree structures for classification", 2019.

[27] S. Li, K. Zhang, Q. Chen, S. Wang and S. Zhang, "Feature

[28] Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm," in IEEE Access, vol. 8, pp. 139512-139528, 2020, doi: 10.1109/ACCESS.2020.3012768.

[29] X. Y. Kek, C. S. Chin and Y. Li, "Multi-Timescale

[30] Wavelet Scattering With Genetic Algorithm Feature Selection for Acoustic Scene Classification," in IEEE Access, vol. 10, pp. 25987-26001, 2022, doi: 10.1109/ACCESS.2022.3156569.

[31] Y. Kataoka, T. Nakashika, R. Aihara, T. Takiguchi and Y.

[32] Ariki, "Selection of an optimum random matrix using a genetic algorithm for acoustic feature extraction," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-6, doi: 10.1109/ICIS.2016.7550890.

[33] Y. Jinxia, C. Zixing and D. Zhuohua, "Improved method

[34] for the feature extraction of laser scanner using genetic clustering," in Journal of Systems Engineering and Electronics, vol. 19, no. 2, pp. 280-285, April 2008, doi: 10.1016/S1004-4132(08)60079-1.

[35] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN:

[36] Towards real-time object detection with region proposal networks", *Proc. Adv. Neural Inf. Process.Syst. (NIPS)*, pp. 91-99, 2015.

[37] K. Nag and N. R. Pal, "A multiobjective genetic

[38] programming-based ensemble for simultaneous feature selection and classification", *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 499-510, Feb. 2016.

[39] K. Nag and N. R. Pal, "Genetic programming for

[40] classification and feature selection" in Evolutionary and Swarm Intelligence Algorithms, Cham, Switzerland:Springer, pp. 119, 2019.

[41] B. Xue, M. Zhang, W. N. Browne and X. Yao, "A survey

[42] on evolutionary computation approaches to feature selection", *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606-626, Aug. 2016.

[43] Y. Xu, Y. Sun, J. Wan, X. Liu and Z. Song, "Industrial big

[44] data for fault diagnosis: Taxonomy review and applications", *IEEE Access*, vol. 5, pp. 17368-17380, 2017.

[45] L. Kou, C. Liu, G.-W.Cai, J.-N. Zhou, Q.-D.Yuan and S.-

[46] M. Pang, "Fault diagnosis for open-circuit faults in NPC inverter based on knowledge-driven and data-driven approaches", *IET Power Electron.*, vol. 13, no. 6, pp. 1236-1245, 2020.

[47] Q. Zhou, P. Yan, H. Liu and Y. Xin, "A hybrid fault

[48] diagnosis method for mechanical components based on ontology and signal analysis", *J. Intell. Manuf.*, vol. 30, no. 4, pp. 1693-1715, Apr. 2019.

[49] R. Ramos, C. D. Acosta, P. J. R. Torres, E. I. S. Mercado,

[50] G. B. Baez, L. A. Rifón, et al., "An approach to multiple fault diagnosis using fuzzy logic", *J. Intell. Manuf.*, vol. 30, no. 1, pp. 429-439, Jan. 2019.

[51] G. Wang, F. Zhang, B. Cheng and F. Fang, "DAMER: A

[52] novel diagnosis aggregation method with evidential reasoning rule for bearing fault diagnosis", *J. Intell. Manuf.*, vol. 32, no. 1, pp. 1-20, Jan. 2021.

[53] K.-Q. Zhou, L.-P. Mo, J. Jin and A. M. Zain, "An

[54] equivalent generating algorithm to model fuzzy Petri net for knowledge-based system", *J. Intell. Manuf.*, vol. 30, no. 4, pp. 1831-1842, Apr. 2019.