Measurement 175 (2021) 109145

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

# Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning

Ching-Hsien Hsu [a,b,c,*], Xing Chen [d,e], Weiwei Lin [f], Chuntao Jiang [a], Youhong Zhang [a], Zhifeng Hao [a], Yeh-Ching Chung [g]

[a] School of Mathematics and Big Data, Foshan University, Foshan 528000, China
[b] Department of Computer Science and Information Engineering, Asia University, Taiwan
[c] Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan
[d] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
[e] Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116., China
[f] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[g] School of Science and Engineering, The Chinese University of Hong Kong, Shenzeng, China

ARTICLE INFO

*Keywords:*
Cancer
Computer-aided diagnosis
Computer modeling
Clinical trials
Machine learning

ABSTRACT

Cancer is a kind of non-communicable disease, progresses with uncontrolled cell growth in the body. The cancerous cell forms a tumor that impairs the immune system, causes other biological changes to malfunction. The most common kinds of cancer are breast, prostate, leukemia, lung, and colon cancer. The presence of the disease is identified with the proper diagnosis. Many screening procedures are suggested to find the presence of the condition under different stages. Medical practitioners further analyze these electronic health records to diagnose and treat the individual. In some cases, misdiagnosis can happen due to manual error or misinterpretation of the data. To avoid these issues, this paper presents an effective computer-aided diagnosis system supported by intelligence learning models. A machine learning-based feature modeling is proposed to improve predictive performance. From the University of California, Irvine repository, breast, cervical, and lung cancer datasets are accessed to conduct this experimental study. Supervised learning algorithms are employed to train and validate the optimal features reduced by the proposed system. Using the 10-Fold cross-validation method, the trained and performance model is evaluated with validation metrics such as accuracy, f-score, precision, and recall. The study's outcome attained 99.62%, 96.88%, and 98.21% accuracy on breast, cervical, and lung cancer datasets, respectively, which exhibits the proposed system's efficacy. Moreover, this system acts as a miscellaneous tool for capturing the pattern from many clinical trials for multiple types of cancer disease.

## 1. Introduction

Disease prediction systems are highly critical in its functionality as it involves finding the presence or absence of a medical condition in an individual. It relatively involves different factors, varying characteristics, multifaceted and real-world aspects [1,2]. In recent times, there is an increasing demand for data-driven, accurate predictive models to enhance the precise identification of future events [3]. Several medical associations and patient counseling programs provide cancer screening recommendations and guidance. Consult a doctor on the different recommendations, and together you can see what is right for you depending on your cancer risk factors. Laboratory tests, such as urine and blood tests, will help the doctor detect cancer-induced anomalies. For example, predictive models with leukemia may show the unusual number or type of white blood cell in a popular blood test called the total blood count. The doctor gathers a sample of the cells in the laboratory for examination during a biopsy. A model is obtained through any means. Dependent on the form of cancer and its location, the biopsy technique is right for you. A biopsy is the way to detect cancer certainly, in most cases.

It concerns developing systems to facilitate the end-users of the application having a more interactive and user-friendly environment. In the view of medical procedures, the physician or medical expert analyses the clinical records of the individuals to diagnose the condition

* Corresponding author.
*E-mail address:* chh@cs.ccu.edu.tw (C.-H. Hsu).

https://doi.org/10.1016/j.measurement.2021.109145
Received 1 October 2020; Received in revised form 2 January 2021; Accepted 1 February 2021
Available online 9 February 2021

with their experience, otherwise domain knowledge [4-6]. Across the globe, many healthcare providers are adopting the computer-assisted diagnosis system to facilitate medical practitioners for an accurate diagnosis [7,8]. Applications in the medical field need special attention to developing decision support systems. Clinical data contains hidden information, usually beyond human competencies and understandability [9]. Finding the pattern is difficult and raised more demand for developing new computational methodologies. In this current scenario, the data extracted from a real-time environment is highly prone to noise and erroneous information [10,11]. The existing mechanisms are not perfectly fitted to the requirement of the current challenges. Therefore, an effective solution is indeed important to address the need to make better diagnostic systems. This paper examined new techniques to fill the gap and limitations of the existing methods. In general, the outcome of a predictive model strongly depends on the input parameters [12]. Also, most of the time, the features are more chaotic than simple factors. It is not feasible to select all the features to build the model, as it might be prone to noise, incorrect inputs. The predictive model's performance solely depends on the significant features identified for effective sample categorization [13]. A small change in the parameters affects the results on different scales.

In many cases, the data is from a real-time environment, where the chance of inconsistency is high, and the quality is often not up to the mark [14-16]. Hence, this paper aims to investigate the existing models, finding a better mechanism to improve performance. The desired objective is to find the feature subsets from all the datasets incorporated in this experiment for effective disease diagnosis. Supervised machine learning algorithms were employed to test and evaluate the system's efficacy based upon its results. The healthcare industry has long been an early adopter and has greatly benefited from technological innovations. In several health fields, computer education, including innovative medical techniques, the processing of patient data and records and chronic diseases, is currently playing a key role in computer technology. Today, machine learning helps streamline administration in hospitals, map and manage infectious conditions, and customize patient care. It may affect the productivity of hospitals and health systems and decrease care costs.

This manuscript is framed with multiple sections as follows. "Background study" section discusses various algorithms and frameworks developed as a tool for disease diagnosis from previous literature. The proposed methodology is briefed in detail in multiple sub-modules that include dataset information. The proposed feature selection method's working process follows with machine learning methods with neat sketches in the "Materials and methods" section. Next, the model validation and performance evaluation process are detailed in the "Results" section. Finally, in the "conclusion" section, the findings and their significance are portrayed with proper reports and graphical analysis. In order to find the disease in various phases, a variety of screening techniques are recommended. Medics examine the electronic medical records more carefully to identify and manage the client. In certain situations, a manual mistake or misinterpretation of the data may cause an error in diagnostics. This paper provides an effective computer-aided diagnostic method with intelligence learning models to prevent these problems. In order to boost predictive efficiency a computer dependent functional simulation is proposed. This experimental research is being performed by the University of California, Irvine repository and by evidence on breast, cervical and lung cancer. Supervised learning algorithms are used for the preparation and evaluation by the proposed method of ideal features.

## 2. Background STUDY

In recent times, the predictive models have shown their importance in many fields that are not limited to healthcare, weather modeling, stock forecasting, intelligence, self-trajectory targeted missiles, etc. Many applications were constructed with the support of intelligence algorithms to perform critical operations from the past data. As the healthcare field is more sensitive over other relative fields, special attention becomes inevitable. In the absence of complex algorithms for decades back, simple models were built to handle the data with small-scale sizes. Nevertheless, these algorithms play an important role, and some of them still act as a backbone to the algorithm of recent times and provides the baseline for them.

A breast cancer detection tool is developed to perform knowledge-based statistical analysis through different algorithms. Naïve Bayes, J48, support vector machine, CART, and radial basis network were employed to make the predictive model. The algorithms' performance is validated through various performance metrics, such as accuracy, specificity, and precision [17]. In another experiment, probabilistic models are tested with a breast cancer dataset to determine the best performing method. Initially, the dataset is preprocessed to replace the empty values with a high-frequency number. Then the data is transformed into a MySQL database. Furthermore, the probability of predicting the samples based on its class label is calculated and constructed a table based on the values. New unseen samples are given as test inputs to evaluate the model's working ability [18].

Like the probabilistic evaluation, the weighted naïve Bayes model is proposed to improve the existing simple naïve Bayes algorithm [19]. Optimization algorithms gain more attention due to their precision in finding the best solution overall available candidate solutions. A breast cancer diagnosis system is developed based on the genetically optimized model on a neural network algorithm to improve the model's effectiveness. The weight and structure of the algorithm are altered for the changes in genetic operations during every cycle. Under different proportions of the dataset, the model performance is evaluated with metrics [20].

The pronostic of breast cancer is focused on past data and classificators have been established based on their trend in another work. It involves many techniques in each phase of the data mining cycle. The best model with optimal pre-processing design and feature selection method is highlighted based on the model score. The random tree and C4.5 are identified as the best over other combinations [21]. Hybrid machine learning models usually combine two or more existing models to improve their ability by sharing their properties to enhance performance. The decision tree is fused with the support vector machine algorithm to predict breast cancer in an attempt. Compared with other classifiers such as for instance-based learning, naïve Bayes, and sequential minimal optimization algorithms, the proposed model is compared. The performance of the DT-SVM model [58] is said to be effective [22].

The cells' physical behavior inspires immune algorithms in living beings to resist viruses by building a strong immunity system. Inspired by the mechanism, artificial resistant algorithms were constructed to find the optimal features for effective prediction [57] of the samples. The proposed algorithm has multiple characteristics, such as clonal selection, cell behavior, and non-linear operations, for the system's effectiveness. It acts as a semi-supervised algorithm and could handle labeled and unlabeled data at a time. This algorithm had shown significant improvement in the results with 99.51% accuracy on the Wisconsin breast cancer dataset [23].

In a related application, a multilevel lung cancer diagnostic model is proposed with fuzzy weighting based preprocessing, feature extraction with principal component analysis, and artificial immune recognition system based classification. The number of features is reduced to four principal components from 57 initial features. Once the feature vectors are identified, the data is preprocessed with a fuzzy-based weighting method. The model attained 100% accuracy on the lung cancer dataset with an artificial immune recognition algorithm and identified it as a promising model for other systems [24]. In another instance, a borderline-SMOTE fused with the AIRS model is proposed to estimate the level of brain metastasis from the data extracted from lung cancer samples. This system effectively handles the class imbalance problems

and is performing well on real-time models [25].

Expert systems are mainly aimed to find effective pipelines made to construct a better framework to diagnose multiple diseases. A machine learning-based expert system is developed for lung cancer diagnosis using general discriminant analysis and the least square support vector machine classifier. This model focuses on two phases as feature extraction and reduction, then classification. The proposed method gained 96.87% accuracy and showed better results than other experimental studies [26]. An automated diagnosis system for lung cancer is developed with a nature-inspired genetic algorithm and relative fuzzy-based extreme learning machines. This model reduces the dimension spaces with a genetic algorithm and is inputted into a fuzzy inference system, which is already trained with revolutionary learning methods. This method is more profound in the effective diagnosis of the condition and be used for other clinical systems [27].

A support vector machine-based rough set hybrid method is proposed to select the breast cancer dataset's optimal predictor features. In this method, the SVM classifier's chosen subset from the rough set model is evaluated in every cycle. Until the best subset is identified, this process generates feature sets with different combinations. This model finds the five best informative features, which are less more useful for better prediction [28].

Wrapper based feature selection methods find the best combination of features with the guidance of a learning model, which validates the performance of every subset generated. In this model, three wrapper methods such as sequential forward, backward, and optimized evolutionary-based selection are performed on multiple datasets. In addition to other models, the sequential forward selection's performance and the decision tree model proved successful [29]. The life expectancy of post-operative lung cancer affected individuals is systematically calculated through machine learning algorithms. Supervised machine learning classifiers are employed to validate the dataset's performance with a 10-fold cross-validation scheme where the multilayered perceptron model attained 82.3% accuracy, followed by J48 with 81.8% [30].

Centered on the genetically optimized model, a breast cancer detection method is built using a neural network algorithm to boost efficacy of the model. For changes in genetic operations over each stage, the weight and composition of the algorithm is affected. The model performance is evaluated with metrics in different proportions of the dataset.

## 3. Materials and methods

### 3.1. Dataset description

In order to find the disease in various phases, a variety of screening techniques are recommended. Medics examine the electronic medical records more carefully to identify and manage the client. In certain situations, a manual mistake or misinterpretation of the data may cause an error in diagnostics. This paper provides an effective computer-aided diagnostic method with intelligence learning models to prevent these problems. In order to boost predictive efficiency a computer dependent functional simulation is proposed. This experimental research is being performed by the University of California, Irvine repository and by evidence on breast, cervical and lung cancer. Supervised learning algorithms are used for the preparation and evaluation by the proposed method of ideal features

This experimental work is carried out with the dataset fetched from the UCI repository. Three datasets of different cancer types were used in this study. Those are breast [31,32], lung [33,34], and cervical cancers [35,36] with a different number of samples and features. This disease diagnosis model focuses on finding a specific subset of informative features to improve the system's predictive performance. Each dataset contains different health factors as parameters from which the condition's presence or absence is traced out. The medical experts labeled all the samples in the dataset. This proposed framework intends to learn the hidden pattern from the input features, and the trained model, the performance of the system is evaluated with unseen data. In Fig. 1, the complete workflow process of the proposed system is depicted in graphical representation. The detailed information about the datasets is given in Table 1.

### 3.2. Data preprocessing

All three datasets accessed in this experiment contains missing values and noisy information. The missing values are imputed with frequently occurred entries in all the features [37]. Cancer is a type of illness that cannot be spread, progressing through cell development in the body. The cancer cell develops tumors that affects the immune system and induces other biological changes. Breast, leukemia, lung and bowel cancer are the most prevalent forms of cancer. It is known that the disorder arises with the right diagnosis. In order to find the disease in various phases, a variety of screening techniques are recommended.

### 3.3. Feature selection

In a definitive pipeline of a machine learning framework, feature selection is an inevitable phase before training the model. This part deals with finding the most informative, useful predictor attributes to increase the predictive model's performance in training and evaluation. Many feature selection techniques are developed with different baselines in statistical and other mathematical derivatives [12,38-40]. It searches for the best subset from the entire features in

The dataset under various mechanisms. In general, three effective strategies, such as filter, embedded, and wrapper techniques [41-43], hold most of the available feature selection methods under their categories. Out of which, nature-inspired, *meta*-heuristic global optimization algorithms play an important role in identifying the best solution, i.e., optimal parameters from all the solution space features. In this paper, a genetic algorithm combined with the correlation method is proposed as a feature optimization technique for effective feature selection. The following section discusses the implemented methods in detail and exhibits the significance of the proposed technique's features based on the model performance.

### 3.4. Genetic algorithm

In the early 1962′s, the foundation of genetic algorithm (GA) on top of adaptive systems was proposed by John Holland. Later in 1975, in the book named "Adaptation in Natural and Artificial Systems," the concept of GA is published by Holland and their team [44,45]. The genetic algorithm is a nature-inspired search algorithm commonly used in computing systems to find approximate solutions to complex search and optimization problems. These algorithms are generalized as heuristics among the global search. In specific, this algorithm mimics the natural selection processes like selection, mutation(m), inheritance, crossover (c) (otherwise, recombination) from the evolutionary theory [46].

This evolution was initiated from an initial set of the population with randomly selected chromosomes in every generation. The chromosome's fitness was individually calculated and based on the fitness score; the new population is derived with strong chromosomes. This newly generated population is then used for the next iteration on finding the optimal solution. The algorithm's termination criteria are usually defined when the number of maximum generations is performed or the fitness level becomes saturated on a specific population. The main drawback is, there is no specific way to declare the best-identified population is the optimal point of convergence in a solution space. The initial phase in a genetic algorithm is defining the people with the collection of individuals as chromosomes. In this population, each chromosome represents a candidate solution. A combination of '0s and 1 s" is formed as strings, each reflecting the state of the characteristics if

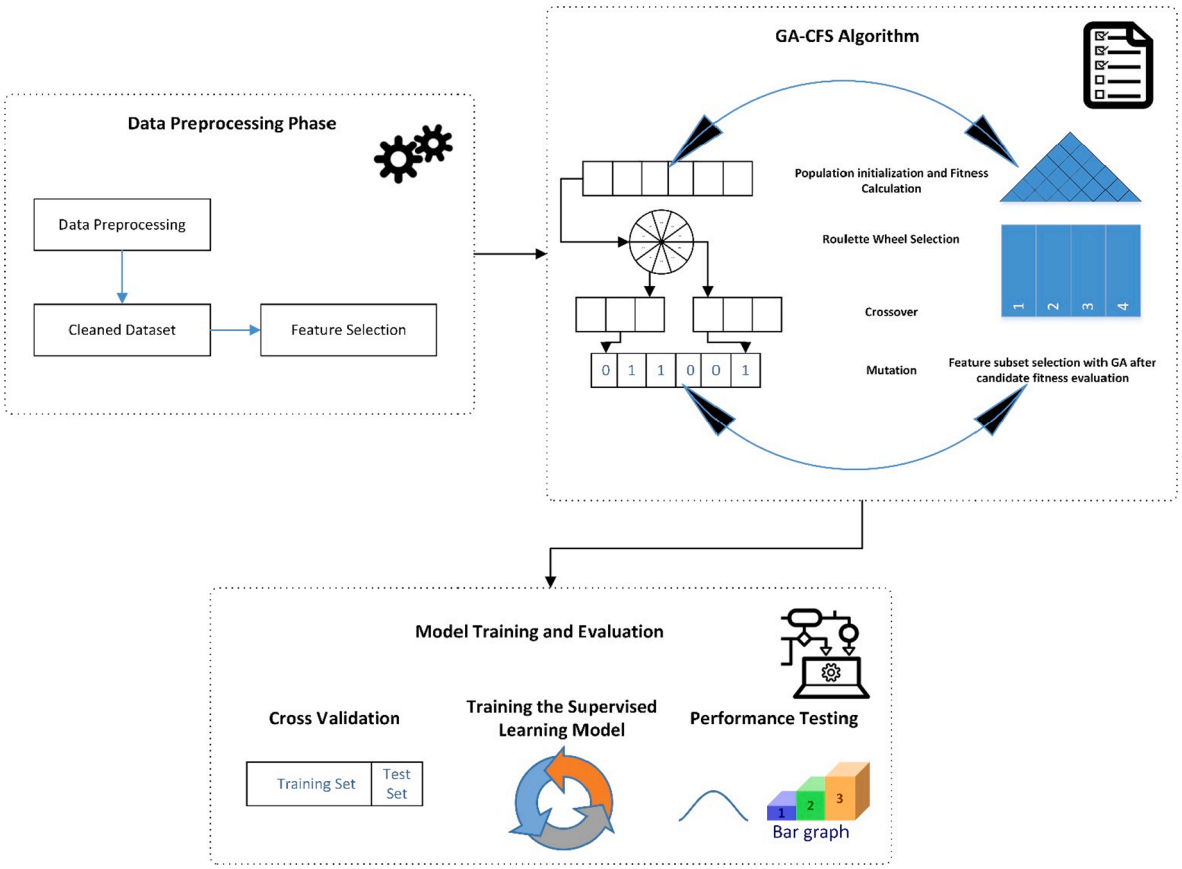


**Fig. 1.** The workflow of the Proposed GA-CFS based Cancer Diagnosis System.

**Table 1**
Dataset information.

| Description | Factors | Datasets | | |
|---|---|---|---|---|
| | F/D | Breast Cancer | Cervical Cancer | Lung Cancer |
| | Dataset Characteristics | Multivariate | Multivariate | Multivariate |
| | Attribute Characteristics | Integer | Integer, Real | Integer |
| | Attribute Count | 10 | 36 | 56 |
| | Total Instances | 699 | 858 | 32 |
| | Missing Values | Present | Present | Present |
| | Category | Classification | Classification | Classification |
| | Data Repository | University of California, Irvine | | |

a bit represents 0 in a chromosome. The fitness function(FF) for that particular chromosome will not be evaluated.

In contrast, the bit with 1 will be seen as a selected function. This bit flip in a chromosome is randomly assigned, and in the next phases, the genetic operations(GP) will be performed on the population (P) [47-49]. The working process of GA is given in Fig. 2, and the parameters are explained in Table 2.

Generic Pseudocode of GA
1: initialize GP on P
2: Check FF on Pl
3: while (*!termination_condition*) do
4: Identify(best_FF) based on GP
5: Generate (m &c)
6: Evaluate (Fitness Score)
7: Replace the individuals < fitness score over top-scored chromosomes
8: end while

The fitness function, the otherwise objective function in a genetic algorithm, calculates the individual's fitness in the population. The fitness function's definition is framed on different aspects, usually involves the maximization or minimization functions to evaluate the scores. The fitness function adopted in the genetic algorithm is given below.

$$FitnessFunctionF = Acc(x)$$

The two important genetic operators of the algorithm are crossover and mutation. A single point crossover and a bit-flip mutation method are used to perform the genetic operation. The roulette wheel selection method selects the candidates in the random population.

### 3.5. Correlation based feature selection

In statistics, correlation is an important term that implies the similarity measures to calculate the relationship between two features [50]. The features are linearly dependent if the correlation factor is one and the reverse if the value is 0. In information theory, correlation-based feature ranking is an effective method to find the feature with higher importance. The optimal subset from the given features in a dataset is identified by finding the strong correlation between the feature and target variable and at the same time weak correlation between the features, which is having an independent relationship [51]. The general formula to calculate the correlation between a pair of attributes (x, y) in a linear correlation is given equation (1).

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}} \quad (1)$$

The merit calculated for the feature subset for the k number of features in S is given in eqn. (2)

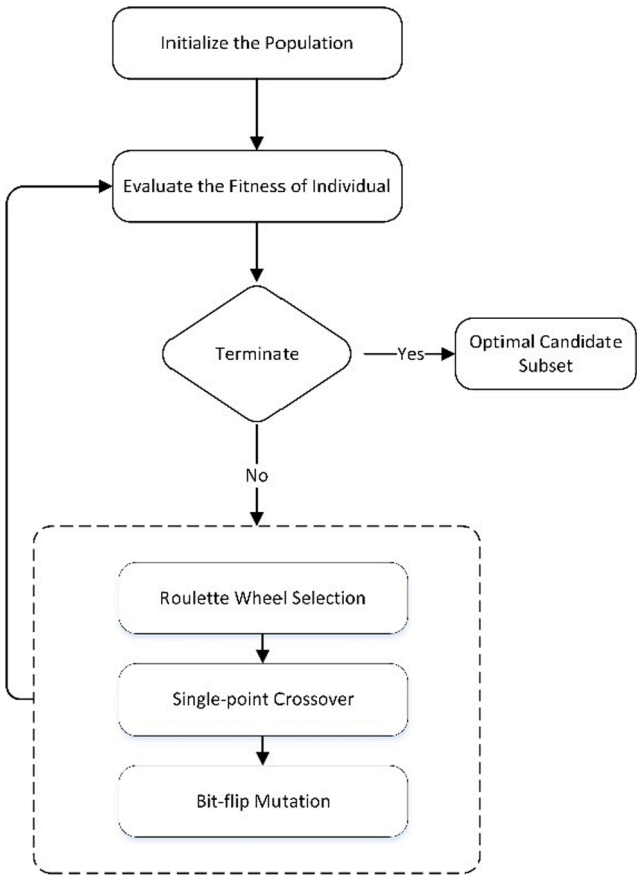


**Fig. 2.** Process of Genetic Algorithm.

**Table 2**
Parameters of genetic algorithm.

| Attributes | Inputs |
|---|---|
| Initial Population | 25 |
| Chromosome Value | Bits { 0 - Off and 1 - On} |
| Total Generations | 10 |
| Mutation Ratio | 0.05 |
| Crossover Ratio | 0.6 |
| Rate of Elitism | 3 |

$$M_{s_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2}$$

$\overline{r_{cf}}$ denotes the average of feature-target correlations, $\overline{r_{ff}}$ states the average of feature-feature mapped correlations. The features are selected based on the following equation given below. If the correlation factor is one and the opposite is 0, the features are linearly dependent. Correlation-based role ranking is an important tool for finding the function of higher significance in information theory. The optimal subset of the features given in a dataset is defined by detecting a strong correlation between the function and the target variable, while even a weak correlation between the features is found.

$$CFS = \max_{s_k}\left[\frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + \cdots + r_{fifj} + \cdots + r_{fkfk-1})}}\right] \tag{3}$$

The above equation can be represented as an optimization problem as,

$$CFS = \max_{x \in \{0,1\}^n}\left[\frac{(\sum_{i=1}^{n} a_i {x_i}^2)}{\sum_{i=1}^{n} x_i + \sum_{i \neq j}^{n} 2b_{ij}x_i x_j^{\,n}}\right] \tag{4}$$

### 3.6. Proposed GA-CFS algorithm

Let F be the complete feature sample set with attributes {$A_1$, $A_2$, $A_3$… $A_n$} where n is the individual number of attributes in a dataset. Here, the process of selecting the optimal feature is defined as finding d, the most information attribute in F >=d. The objective of the proposed algorithm is to identify the predictor attribute subset (i.e., d < F). Crossover and mutation are the two primary genetic operators of the algorithm. The genetic surgery is done using a single point crossover and bit-flip mutation process. The process of selection of roulette wheels selects the random number of applicants.

```
Algorithm: GA - CFS
initialize i = 0, pop(i) = 0
  evaluate pop(i) = 0
  while (!termination_condition()) do
  pop_p(i) = pop(i).generateParents();
  pop_c(i) = reproduce(pop_p);
  mutation(pop_c(i));
  fitness_evaluation(pop_p(i));
  calculate corr(pop_p(i_n,y)
  evaluate new_fitness(corr(pop_p(i_n,y)));
  if(finess_evaluation(pop_p(i)) < corr(pop_p(i_n,y))
  fitness_new = corr(pop_p(i_n,y))
  pop(i + 1) = generate_next(fitness_new, pop(i));
  i = i + 1
  else
  pop(i + 1) = generate_next(pop_c(i), pop(i));
  i = i + 1
  end if
  end while
```

The processed dataset is prepared to select the predictor subsets in the next phase. The proposed GA-CFS algorithm takes all the features as inputs from which the optimal solution is identified. The genetic algorithm performs as given in the above pseudo-code, and while finding the candidate's fitness, two distinct calculations are made. After calculating the fitness, the CFS technique again intends to find the optimal subset among the selected features by GA. In this way, every candidate selected by the GA is validated again with CFS to ensure the global optimum solution's attainment. In table 3, the number of features determined by the GA-CFS algorithm is given on each dataset.

### 3.7. Model training and evaluation

The selected features are evaluated using standard machine learning classifiers to find the best performing pipeline. For identifying the proposed algorithm's significance, other existing feature selection methods are employed to compare the performance variation. The training and validation sets are generated with a 10-fold cross-validation method [52]. Support vector machines, decision trees, naïve Bayes, and multilayered perceptron neural network algorithms are trained with both sets [53]. The performances of the algorithms are analyzed in detail in the next section. The suggested GA-CFS algorithm integrates all the characteristics as inputs that define the optimum solution. The genetic algorithm is as provided for in the pseudo-code above and two distinct calculations are carried out when the candidate finds fitness. The CFS technique aims, once again to find the ideal subset of GA's selected features after the fitness measurement. The effectiveness of the GA-CFS model has been shown to surpass all three comparable data sets with the neural network model over the remaining current role selection

**Table 3**
Features selected by ga-cfs in each dataset.

| Dataset | Number of features selected |
|---|---|
| Breast Cancer | 5 |
| Cervical Cancer | 7 |
| Lung Cancer | 7 |

methods. This system will be further updated in the future to support complex, high-dimensional real-time data sets. By offering an adequate diagnosis, this model acts as a testing guide for clinicians.

## 4. Results and discussion

This experimental work is carried out in Java Framework with the support of python machine learning libraries through bridges in the Windows platform. Breast, cervical, and lung cancer datasets were used to conduct the study. In every phase of the pipeline, the datasets are processed, starting with pre-processing, where the missing values are imputed. The cleaned data is then forwarded into the next phase to find the best features from the proposed GA-CFS algorithm. This method identified five breast cancer dataset features, 7 in the lung and cervical dataset [59]. Alongside the proposed algorithm, the datasets were tested against a few existing feature selection methods such as ReliefF, Recursive Feature Elimination (RFE), and Cuckoo search optimization (CSO) algorithm [54-56]. The proposed feature selection methods' performance under different classifiers is calculated and explained in Tables 4-6. In Fig. 3, the classifiers' accuracy on three datasets is plotted, and Fig. 4 shows the performance of the benchmarked feature selection techniques on the MLP-NN classifier. The formula of accuracy, precision, recall, and f-score is given as Eqn. 5,6,7,8 respectively.

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - Score = \frac{(2*(Re*Pre))}{(Re + Pre)}$$

Based on the results, it is observed that the proposed GA-CFS with MLP-NN algorithm attained better results over other combinations of feature selection methods with classification models with optimized parameter tabulated in

Moreover, the number of selected features becomes less compared with ReliefF, RFE, and CSO. It minimized the models' computational effort to find the data's discriminative patterns and revealed the most important parameters on every dataset. These parameters could act as a potential factor to effectively diagnose the condition of the individual. In all the datasets accompanied by SVM with Gaussian kernels, NB and LDA, the MLP-NN has demonstrated better performance. The effectiveness of the GA-CFS model has been shown to surpass all three comparable data sets with the neural network model over the remaining current role selection methods.

## 5. Conclusion

The computational methods have shown prominence in the medical field and can provide profound solutions for complex systems. These systems are more beneficial for medical practitioners to make a better decision based on the models' guidelines, which are represented as knowledge captured and gathered from intelligence algorithms. This study presents an effective algorithmic model for better classification of the clinical data labeled manually by the experts. The proposed algorithm finds the genetic algorithm's informative features with the support of correlation-based feature selection, where each chromosome generated by the genetic algorithm is further reduced by calculating the correlation between the features in the subset. This proposed GA-CFS algorithm method provides a promising way to find the optimal solution. Later, the identified subset is trained with various supervised classification algorithms to benchmark the models' performance. Among all, the MLP-NN has shown better results in all the datasets followed with SVM with Gaussian kernel, NB, and LDA. The GA-CFS model's efficacy is proven to outperform all three benchmarked datasets over the rest of the existing feature selection methods with the neural network model. In the future, this system is further modified to serve well on complex, high-dimensional, real-time datasets. This model

**Table 4**
Performance of the proposed ga-cfs algorithm on breast cancer dataset in (%).

| Classifier | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|
| DT | 82.06 | 84.2 | 83.71 | 87.91 |
| SVM | 84.20 | 83.1 | 84.97 | 88.29 |
| LDA | 77.15 | 78.2 | 76.51 | 80.21 |
| MLP-NN | 99.62 | 96.12 | 97.02 | 98.70 |

**Table 5**
Performance of the proposed ga-cfs algorithm on cervical cancer dataset in java framework (%).

| Classifier | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|
| DT | 86.40 | 87.22 | 88.18 | 89.71 |
| SVM | 87.03 | 85.41 | 86.81 | 86.98 |
| LDA | 81.1 | 80.29 | 81.22 | 82.51 |
| MLP-NN | 96.88 | 96.92 | 97.47 | 98.12 |

**Table 6**
Performance of the proposed ga-cfs algorithm on lung cancer dataset in (%).

| Classifier | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|
| DT | 88.05 | 89.72 | 87.14 | 89.81 |
| SVM | 86.80 | 85.61 | 84.31 | 85.69 |
| LDA | 79.61 | 80.24 | 79.81 | 80.21 |
| MLP-NN | 98.21 | 94.62 | 95.30 | 96.70 |

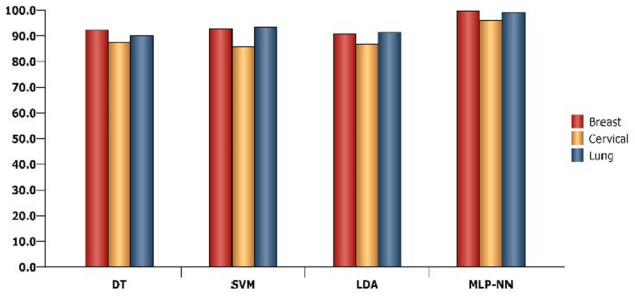


**Fig. 3.** Accuracy of the classifiers of the proposed GA-CFS Algorithm on three datasets (%).

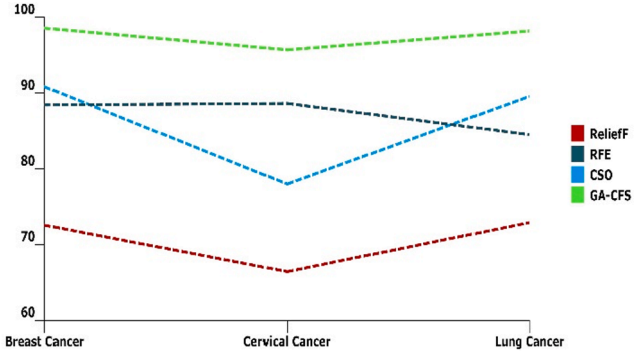


**Fig. 4.** Comparison of accuracy obtained under different feature selection methods on MLP-NN (%).

serves as a diagnostic tool for medical practitioners by assisting them with an adequate diagnosis.

## CRediT authorship contribution statement

**Ching-Hsien Hsu:** Conceptualization, Methodology, Software, Writing - original draft. **Xing Chen:** Writing - review & editing, Validation, Visualization, Investigation. **Weiwei Lin:** Investigation, Methodology, Validation, Supervision. **Chuntao Jiang:** Investigation, Validation, Supervision. **Youhong Zhang:** Investigation, Methodology, Software, Validation, Supervision. **Zhifeng Hao:** Writing - review & editing, Validation, Visualization, Investigation. **Yeh-Ching Chung:** Investigation, Methodology, Supervision.

## Declaration of Competing Interest

The authors declared that there is no conflict of interest.

## Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61872084; 61802062; 62072187) and Guangdong-Hong Kong-Macao Intelligent Micro-Nano Optoelectronic Technology Joint Laboratory (Project No. 2020B1212030010).

## References

[1] M. Berg, P.A.M. Berg, Rationalizing medical work: decision-support techniques and medical practices, MIT press, 1997.
[2] B.J. Wilson, N. Torrance, J. Mollison, M.S. Watson, A. Douglas, Z. Miedzybrodzka, A. Grant, Cluster randomized trial of a multifaceted primary care decision-support intervention for inherited breast cancer risk, Fam. Pract. 23 (5) (2006) 537–544.
[3] Thorwarth, M., & Arisha, A. (2012, December). A simulation-based decision support system to model complex demand driven healthcare facilities. In Proceedings of the 2012 Winter Simulation Conference (WSC) (pp. 1-12). IEEE.
[4] G.I. Doukidis, Decision support system concepts in expert systems: an empirical study, Decis. Support Syst. 4 (3) (1988) 345–354.
[5] R. Tsopra, J.P. Jais, A. Venot, C. Duclos, Comparison of two kinds of interface, based on guided navigation or usability principles, for improving the adoption of computerized decision support systems: application to the prescription of antibiotics, J. Am. Med. Inform. Assoc. 21 (e1) (2014) e107–e116.
[6] E. Turban, P.R. Watkins, Integrating expert systems and decision support systems, Mis Quarterly (1986) 121–136.
[7] M.A. Musen, B. Middleton, R.A. Greenes, Clinical decision-support systems, in: Biomedical informatics, Springer, London, 2014, pp. 643–674.
[8] E.H. Shortliffe, Computer programs to support clinical decision making, JAMA 258 (1) (1987) 61–66.
[9] S. Abrol, A. Kotrotsou, A. Salem, P.O. Zinn, R.R. Colen, Radiomic phenotyping in brain cancer to unravel hidden information in medical images, Top. Magn. Reson. Imaging 26 (1) (2017) 43–53.
[10] R. Goodloe, E. Farber-Eger, J. Boston, D.C. Crawford, W.S. Bush, Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record, AMIA Summits on Translational Science Proceedings 2017 (2017) 102.
[11] V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, T.I. Leung, E.P. Minty, N.H. Shah, Learning statistical models of phenotypes using noisy labeled training data, J. Am. Med. Inform. Assoc. 23 (6) (2016) 1166–1173.
[12] K. Kira, L.A. Rendell, A practical approach to feature selection, Morgan Kaufmann, 1992, pp. 249–256.
[13] N. Kwak, C.H. Choi, Input feature selection for classification problems, IEEE Trans. Neural Networks 13 (1) (2002) 143–159.
[14] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of EHR: data quality issues and informatics opportunities, Summit on Translational Bioinformatics 2010 (2010) 1.
[15] R.L. Rush, J.A. Barwick, J.A. Elsinger, D.C. Crum, M.A. Foulkes, K.H. Chantry, Maximizing detection of data inconsistency: The development of a consistency check interpreter, American Medical Informatics Association (1987), p. 848.
[16] C. Danilowicz, N.T. Nguyen, Consensus methods for solving inconsistency of replicated data in distributed systems, Distributed and Parallel Databases 14 (1) (2003) 53–69.
[17] B. Muthu, C.B. Sivaparthipan, G. Manogaran, et al., IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector, Peer-to-Peer Netw. Appl. 13 (2020) 2123–2134, https://doi.org/10.1007/s12083-019-00823-2.
[18] S. Aruna, S.P. Rajagopalan, L.V. Nandakishore, Knowledge based analysis of various statistical tools in detecting breast cancer, Computer Science & Information Technology 2 (2011) (2011) 37–45.
[19] S. Kharya, S. Agrawal, S. Soni, Naive Bayes classifiers: A probabilistic detection model for breast cancer, International Journal of Computer Applications 92 (10) (2014) 0975–8887.
[20] S. Kharya, S. Soni, Weighted naive Bayes classifier: A predictive model for breast cancer detection, International Journal of Computer Applications 133 (9) (2016) 32–37.
[21] A. Bhardwaj, A. Tiwari, Breast cancer diagnosis using genetically optimized neural network model, Expert Syst. Appl. 42 (10) (2015) 4611–4620.
[22] Jacob, S. G., & Ramani, R. G. (2012, October). Efficient classifier for classification of prognostic breast cancer data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26).
[23] K. Sivakami, N. Saraswathi, Mining big data: breast cancer prediction using DT-SVM hybrid model, International Journal of Scientific Engineering and Applied Science (IJSEAS) 1 (5) (2015) 418–429.
[24] C.B. Sivaparthipan, N. Karthikeyan, S. Karthik, Designing statistical assessment healthcare information system for diabetics analysis using big data, Multimed Tools Appl 79 (2020) 8431–8444, https://doi.org/10.1007/s11042-018-6648-3.
[25] L. Peng, W. Chen, W. Zhou, F. Li, J. Yang, J. Zhang, An immune-inspired semi-supervised algorithm for breast cancer diagnosis, Comput. Methods Programs Biomed. 134 (2016) 259–265.
[26] K. Polat, S. Güneş, Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer, Expert Syst. Appl. 34 (1) (2008) 214–221.
[27] K.J. Wang, A.M. Adrian, K.H. Chen, K.M. Wang, A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan, Comput. Methods Programs Biomed. 119 (2) (2015) 63–76.
[28] E. Avci, A new expert system for diagnosis of lung cancer: GDA—LS_SVM, J. Med. Syst. 36 (3) (2012) 2005–2009.
[29] M.R. Daliri, A hybrid automatic system for the diagnosis of lung cancer based on genetic algorithm and fuzzy extreme learning machines, J. Med. Syst. 36 (2) (2012) 1001–1005.
[30] H.L. Chen, B. Yang, J. Liu, D.Y. Liu, A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Syst. Appl. 38 (7) (2011) 9014–9022.
[31] S. Karthik, M. Sudha, Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network, Evol. Intel. (2020) 1–16.
[32] Danjuma, K. J. (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. arXiv preprint arXiv: 1504.04646.
[33] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original %29 (Accessed on 20-03-2020).
[34] Mangasarian, O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis.
[35] http://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 (Accessed on 20-03-2020).
[36] K. Fernandes, J.S. Cardoso, J. Fernandes, Transfer learning with partial observability applied to cervical cancer screening, Springer, Cham, 2017, pp. 243–250.
[37] http://archive.ics.uci.edu/ml/datasets/Lung+Cancer (Accessed on 20-03-2020).
[38] Z.Q. Hong, J.Y. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, Pattern Recogn. 24 (4) (1991) 317–324.
[39] J.W. Graham, P.E. Cumsille, A.E. Shevock, Methods for handling missing data, Second Edition, Handbook of Psychology, 2012, p. 2.
[40] Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning.
[41] Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 Science and Information Conference (pp. 372-378). IEEE.
[42] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[43] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). Ieee.
[44] K. Sekaran, M. Sudha, Predicting drug responsiveness with deep learning from the effects on gene expression of Obsessive-Compulsive Disorder affected cases, Comput. Commun. 151 (2020) 386–394.
[45] U. Stańczyk, Feature evaluation by filter, wrapper, and embedded approaches, in: Feature Selection for Data and Pattern Recognition, Springer, Berlin, Heidelberg, 2015, pp. 29–44.
[46] J.H. Holland, J.S. Reitman, Cognitive systems based on adaptive algorithms, in: Pattern-directed inference systems, Academic Press, 1978, pp. 313–329.
[47] Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning.
[48] H. Mühlenbein, July). Parallel genetic algorithms, population genetics and combinatorial optimization, in: Workshop on Parallel Processing: Logic, Organization, and Technology, Springer, Berlin, Heidelberg, 1989, pp. 398–406.
[49] W.M. Spears, V. Anand, A study of crossover operators in genetic programming, Springer, Berlin, Heidelberg, 1991, pp. 409–418.
[50] C. Kane, M. Schoenauer, Genetic operators for two-dimentional shape optimization, Springer, Berlin, Heidelberg, 1995, pp. 355–369.
[51] O. Kramer, Genetic algorithms, in: Genetic algorithm essentials, Springer, Cham, 2017, pp. 11–19.
[52] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[53] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863).
[54] T.T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recogn. 48 (9) (2015) 2839–2846.
[55] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: A review of classification techniques, Emerging artificial intelligence applications in computer engineering 160 (2007) 3–24.
[56] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine learning 53 (1–2) (2003) 23–69.
[57] A.H. Gandomi, X.S. Yang, A.H. Alavi, Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems, Engineering with Computers 29 (1) (2013) 17–35.
[58] X. Zhang, X. Lu, Q. Shi, X.Q. Xu, E.L. Hon-chiu, L.N. Harris, W.H. Wong, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data, BMC Bioinf. 7 (1) (2006) 197.
[59] https://archive.ics.uci.edu/ml/datasets.php.