A genetic programming-based approach for classifying pancreatic adenocarcinoma: the SICED experience

Gianni D’Angelo1 · Maria Nunzia Scoppettuolo2 · Anna Lisa Cammarota2 · Alessandra Rosati2 · Francesco Palmieri1

Accepted: 12 July 2022 / Published online: 30 July 2022 © The Author(s) 2022

Abstract
Ductal adenocarcinoma of the pancreas is a cancer with a high mortality rate. Among the main reasons for this baleful prognosis is that, in most patients, this neoplasm is diagnosed at a too advanced stage. Clinical oncology research is now particularly focused on decoding the cancer molecular onset by understanding the complex biological architecture of tumor cell proliferation. In this direction, machine learning has proved to be a valid solution in many sectors of the biomedical field, thanks to its ability to mine useful knowledge by biological and genetic data. Since the major risk factor is represented by genetic predisposition, the aim of this study is to find a mathematical model describing the complex relationship existing between genetic mutations of the involved genes and the onset of the disease. To this end, an approach based on evolutionary algorithms is proposed. In particular, genetic programming is used, which allows solving a symbolic regression problem through the use of genetic algorithms. The identification of these correlations is a typical objective of the diagnostic approach and is one of the most critical and complex activities in the presence of large amounts of data that are difficult to correlate through traditional statistical techniques. The mathematical model obtained highlights the importance of the complex relationship existing between the different gene’s mutations present in the tumor tissue of the group of patients considered.

Keywords
Ductal adenocarcinoma · Pancreas · Genetic programming · Machine learning

1 Introduction
Pancreatic cancer (PC) is the fourth most common cause of cancer death in the USA with more than 38,000 deaths in 2013 (Siegel et al. 2013) and is projected to become the second most common cause by 2030 (Rahib et al. 2014). The situation in Europe is similar in both women and men, with nearly one million deaths for year and reduced quality of life in affected people (Carrato et al. 2015). In the future, PC will overtake breast cancer as the third leading cause of cancer death (Bray et al. 2018). Pancreatic cancer is a complex malignant tumor with a poor prognosis since it does not show particular signs at an early stage, and even when they are present, they are rather vague disorders, which can be misinterpreted. For these reasons, the diagnosis is often made when the disease is already extensive.

Clearer symptoms (which vary depending on where cancer began in the pancreas) appear when cancer has started to spread to nearby organs or has blocked the bile ducts.

However, as the size of the tumor increases, the symptoms also tend to become more noticeable. They depend on many factors, including the type and location of the tumor, as well as on the impact it has on the body in general. They can result in loss of weight and appetite, jaundice, pain in the upper abdomen or back, a feeling of bloating or fullness, weakness, nausea, or vomiting.
Other fairly common symptoms include intestinal problems, diabetes, indigestion, heartburn, nausea, and vomiting, fever and chills, extreme fatigue, acute pancreatitis of unexplained origin, bulky, clear, and foul-smelling stools.

It is therefore diagnosed at an advanced stage of its natural history and is resistant to chemo- and radiotherapy (Momii et al. 2012). Surgical removal is the only potentially curative treatment, and about 80–85% of cancers are not surgically removed at the time of diagnosis. However, accurate screening of patients at high risk of developing PC could allow an early diagnosis, when surgical removal of the tumor is still possible. Therefore, identifying high-risk population is crucial, as well as understanding the risk factors for developing PC. Several factors (environmental and genetic) have been identified as being responsible for the increased risk of developing PC. Several factors (environmental and genetic) have been identified as being responsible for the increased risk of developing pancreatic cancer (D’Angelo and Palmieri 2021b), such as age, smoking (D’Angelo and Palmieri 2020; Elia et al. 2020), alimentation (D’Angelo et al. 2018), alcoholic beverages (D’Angelo and Rampone 2014), chronic pancreatitis (D’Angelo et al. 2019), diet and obesity (Duffy and Engle-Warnick 2002), diabetes (Koza 1994), genetic diseases, environmental pollution, and more.

About 70% of pancreatic cancers originate from the head of the organ and most of these develop from the ducts that carry digestive enzymes (pancreatic ductal adenocarcinoma-PDAC). Despite recent advances in surgical techniques and medical therapies, the survival on average for a PC patient is approximately 4-6 months from the diagnosis (Farthing et al. 2014), while the survival rate on average at 5-year is <8% (Lucas et al. 2016). Pancreatic mucinous cystic neoplasm (MCN) accounts for 2–5 % of all exocrine pancreas neoplasms and is characterized instead by a far better prognosis after surgical eradication. MCN is a cystic-forming, mucin-producing neoplasm and a correct clinical approach is often hampered by the difficult differential diagnosis in respect to intraductal papillary mucinous neoplasms (IPMN) or oligocystic serous adenomas before surgery (Scholten et al. 2018; Xie et al. 2021). Of note, both IPMNs and MCNs can shift toward a more aggressive form (Coriat and Pellat 2022).

In this complex picture, in pancreatic neoplasms, the most common and early events that allow cells to acquire genetic instability are the dysfunction and critical shortening of telomeres that lead to impairment of chromosomal integrity. This phenomenon is traceable both in precancerous lesions and in invasive carcinomas (van Heek et al. 2002; Gisselsson et al. 2001). Pancreatic cancer is therefore a disease caused by the accumulation of mutations. As an example, one of the first genetic analysis in 24 patients carrying pancreatic adenocarcinomas showed that the genetic basis of the neoplasm is complex and heterogeneous. In each tumor, an average of 63 genetic mutations (mainly of a single nucleotide substitution) have been identified, which are organized into twelve signal transduction pathways, whose alterations are fundamental in the oncogenetic process. Not all tumors exhibit alterations in all pathways, and key mutations in each pathway appear to differ from tumor to tumor (Gisselsson et al. 2001).

The main alterations include:

1. mutations activating the KRAS oncogene in 90% of cases;
2. mutations inactivating the p16 oncosuppressor encoded by the CDKN2A gene in almost all cases;
3. mutations inactivating the TP53 gene, encoding the oncosuppressor p53 in 60% of cases;
4. deletion of the SMAD4 DPC4 gene in 50% of cases.

Precisely because of the high genetic heterogeneity of pancreatic cancer, there is still no genetic test to accurately define the lesions. For this reason, the identification of mutations is the path that can lead to the development of reliable diagnostic tests to guarantee patients an earlier diagnosis, hopefully before the tumor manifests itself clinically. The identification of the mutations associated with the main lesions and pre-invasive forms and the description of their role in the evolution of the tumor cell can provide a fundamental contribution in the perspective of developing a rapid and accurate diagnostic system and is the first step for understanding the mechanisms of pancreatic tumorigenesis.

To better characterize and understand the genetics of pancreatic cancer, the aim of this study is to analyze the overall picture of the genetic alterations that underlie the heterogeneity of pancreatic adenocarcinoma through the use of genetic databases in order to develop innovative approaches in the early diagnosis of pancreatic adenocarcinoma.

The variety and multiplicity of mutations in the genes involved in pancreatic cancer can lead to very different phenotypic phenomena. Trying to understand how one or more mutations are linked to risk factors for pancreatic cancer is a non-trivial task.

Machine learning has proven to be a decisive approach in many sectors of the biomedical field, thanks to the ability to mine useful knowledge by directly observing biological or genetic data and processing them through numerous techniques and algorithms (D’Angelo and Palmieri 2020, 2021b; Elia et al. 2020; D’Angelo et al. 2018, 2019; D’Angelo and Rampone 2014).

Since one of the major risk factors is represented by genetic predisposition, the aim of this study is to find a mathematical model describing the complex relationship existing between genetic mutations of the involved genes and the onset of the disease.

To this end, an approach based on evolutionary algorithms (EA) is proposed. In particular, genetic programming is used, which allows solving a symbolic regression problem through the use of genetic algorithms.

The identification of these correlations is a typical objective of the diagnostic approach and is one of the most critical
and complex activities in the presence of large amounts of data that are difficult to correlate through traditional statistical techniques.

The formula obtained highlights the importance of the complex relationship existing between the different mutations present in the genes belonging to the group of patients considered.

The main contributions of this study can be synthesized as follows:

- A regression problem is transformed into a classification problem.
- A GP-based classifier is modeled by a mathematical model.
- Correlations among genetic mutations are mined and expressed through mathematical formula.
- The mechanisms of pancreatic tumorigenesis are modeled through a mathematical formula.

2 Materials and methods

2.1 Dataset

To better characterize and understand the genetics of pancreatic cancer, we have analyzed the overall picture of the genetic alterations that underlie the heterogeneity of pancreatic adenocarcinoma. To this end, we have used several clinical and biological databases in order to develop innovative approaches in the early diagnosis of pancreatic cancer.

Thanks to the study called Pan-Cancer Project which lasted about 10 years, which involved 16 working groups of the International Cancer Genome Consortium (ICGC), and the US project called The Cancer Genome Atlas (TCGA), the mapping of over 2600 genomes belonging to 38 different types of cancer has been carried out. This has made it possible to create a huge database that allows the study of multiple aspects of the development, cause, progression, and classification of cancer. The information gathered by these projects has been made public in order to accelerate the understanding of the molecular basis of cancer through the new generation of genome analysis technologies.

In our study, we have used three different databases available at the following addresses, respectively:

- https://portal.gdc.cancer.gov/
- https://cbioportal.org
- https://dcc.icgc.org

They include histological images, various information on patient history, clinical data, and all genetic information, including point mutations and copy number variants (CNV).

The three databases include 1067 pancreatic cancer patients, and they have been acquired by 5 projects, that are:

1. PAAD-US: 185 patients coming from the USA with ductal adenocarcinoma (https://dcc.icgc.org/projects/PAAD-US);
2. PACA-CA: 317 patients coming from Canada with ductal adenocarcinoma (https://dcc.icgc.org/projects/PACA-CA);
3. PACA-AU: 461 patients coming from Australia with ductal adenocarcinoma (https://dcc.icgc.org/projects/PACA-AU);
4. PAEN-AU: 67 patients coming from Australia with pancreatic neuroendocrine tumor (https://dcc.icgc.org/projects/PAEN-AU);
5. PAEN-IT: 37 patients coming from Italy with enteropancreatic endocrine tumor and rare exocrine pancreatic tumors (https://dcc.icgc.org/projects/PAEN-IT).

Due to the large number of risk factors involved in the pathogenesis of pancreatic adenocarcinoma, for each of these patients we have also extrapolated demographic data, genetic data, and risk factors. More specifically, the following factors have been considered.

Demographic:

- Age (more than 80% of adenocarcinomas affect people aged between 60 and 80 years);
- Vital state (patient dead or still alive at the time of genetic analysis);
- Age at diagnosis of pancreatic cancer.

Genetic:

- History of Acute Pancreatitis;
- Medical history Diabetes mellitus;
- Family history of cancer (pancreas or other cancer);
- Genetic mutations (point, small insertions, and small deletions);
- CNV (copy number variants).

Cancer characteristic:

- Type of pancreatic cancer:
  - Cystic, Mucinous and Serous Neoplasms;
  - Adenomas and Adenocarcinomas;
  - Ductal and Lobular Neoplasms;
  - Epithelial Neoplasms, NOS.
- Tumor stage:
  - Stage IA. The tumor is limited to the pancreas site only and in its maximum extension measures less than 2 cm (localized—resectable);
– Stage IB. The tumor is limited to the pancreas and measures more than 2 cm (localized—resectable);
– Stage IIA. The cancer extends outside the pancreas but does not involve major local arteries or lymph nodes (locally advanced—resectable or borderline resectable);
– Stage IIB. The tumor may or may not extend outside the pancreas but does not involve the major arteries. Local lymph nodes may be involved (locally advanced—resectable or borderline resectable);
– Stage III. The tumor involves important local arteries. Local lymph nodes may be involved (locally advanced—unresectable);
– Stage IV. The tumor can be of any size. The tumor has spread to other organs (metastatic—unresectable).

- Tumor grade:
  – G1: well differentiated tumor;
  – G2: moderately differentiated tumor;
  – G3: poorly differentiated tumor;
  – G4: undifferentiated tumor.

Note that the tumor grade allows us to classify the differentiation status of the cells in order to define the degree of aggressiveness of the tumor.

Other:

- Cigarette smoking (the only environmental factor definitely correlated with the risk of pancreatic adenocarcinoma);
- Alcohol consumption (data divided by frequency and intensity);
- Survival from initial diagnosis (expressed in days).

Since one of the major risk factors is represented by genetic predisposition, we preferred to focus our attention on genes mutated in high percentages and, therefore, more frequent in patients with pancreatic cancer. Furthermore, among the different somatic mutations collected in our 3 databases, we detected a high number of point mutations. Most of mutations that we found are the following: synonym, missense, nonsense, small insertions and deletions (frameshift and in frame), mutations in splicing sites or in untranslated regions (UTR).

Below are the genes used in our analysis: KRAS, Tp53, CDKN2A, TGF-1β, Smad genes, BRCA1, BRCA2, PRSS1/CFTR, APC, BAG3, STK11/LKB1, MLH1, MSH2, MSH6, PMS2, ATM, PALB2, TTN, DNM1P47, HDAC5, EGF/ERBB2, STAT3, PTEN, KLF4, PIK3CA, B RAF, ARID1A, ROBO2, KDM6A, PREX2, CDK6, GNAS, ATRX, KMT2D. A detailed description of any gene can be found in GenBank® database.¹

2.2 Features selection

Different types of mutations have been considered for each of the aforementioned genes. This led to a total number of parameters (i.e., characteristics) equal to 471.

As is known, the choice of highly informative and independent discriminating features is a crucial step to obtain an efficient pattern recognition, classification or regression algorithm (Visalakshi and Radha 2014). Besides, the number of used features should be kept low in order to obtain high performance.

Consequently, a typical preliminary step in many machine-learning applications consists in the selection of the features (Features Selection), or, more generally, in the reduction of the dimensionality of the so-called input space.

In our study, it can be seen that the available dataset includes, for each patient, a very high number of features for each gene (i.e., 471). The extraction of information from a dataset characterized by such a large number of features would require many training examples (patients). Unfortunately, as described in the previous section, the dataset includes only 1067 cases, which could negatively affect the performance of the proposed system.

For this reason, in order to evaluate the actual usefulness of all the features, the link between these features and the output was evaluated through WEKA’s attribute selection techniques. In particular, two methods were used, namely GainRatioAttributeEval and InfoGain. Note that as better described in the next subsection, our goal is to classify the cancer stage.

The results obtained for both algorithms suggested that only the features f111, f197, f279, f288, f318, f352, and f443 are relevant for the classification. Below is the correspondence between these features and their meaning:

- f111 → TP53 chr17: g.7674226A>C Replacement Missense M246R
- f197 → TTN chr2: g.178528652 del TCTGAGAG deletion Inframe Deletion E35697_S35699 del
- f279 → DNM1P47 chr15: g.101752915C>A Replacement Non Coding Transcript Exon DNM1P47
- f288 → DNM1P47 chr15: g.101754458G>C Replacement Non Coding Transcript Exon DNM1P47
- f318 → ATM chr11: g.108287646T>G Replacement Stop Gained L1347 *
- f352 → APC chr5: g.112841068A>G Replacement Missense D1825G

¹ https://www.ncbi.nlm.nih.gov/genbank/.
2.3 Genetic programming-based tumor classification

As previously mentioned, the aim of this study is to classify the type of pancreatic neoplasm. More specifically, we are interested in finding the relation existing among gene mutations and cancer stage. We remark that a classifier can be seen as a function that takes a features vector as input and assigns a class label to it. Accordingly, in our study, we considered a 471-dimensional vector as input and two classes as output, one corresponding to the presence of MCN type (Class 1) and the other corresponding to PDAC type (Class 0).

In order to build such a classifier, in this study, an evolutionary-based algorithm is proposed. More specifically, the genetic algorithm (GA) is used to perform a symbolic regression (SR) analysis.

We remark that regression is used to mine insights contained in data, and symbolic regression is a method widely known and used in the field of regression analysis. It is used to discover a mathematical formula describing an existing relation between multiple input/output pairs of variables (Duffy and Engle-Warnick 2002).

In recent decades, advances in the field of evolutionary algorithms (EAs) have allowed the spread of SRs.

The main goal of an EA is to provide solutions to many types of real-world problems by using Darwin's theory of evolution. More specifically, EA makes usage of a population of solutions, each called individual, which are combined through the operators of selection, crossover, and mutation to create a new population of solutions until the best individual (i.e., the best solution) in the population is found-out.

In order to solve a SR problem, Koza (1994) used GA by laying the foundations of the genetic programming (GP) (Affenzeller et al. 2009).

Basically, GP is an SR running by EAs. Like in AEs, in GPs there are numerous individuals which correspond to a different mathematical formula. Each formula, and then individual, is represented by a tree structure (see Fig. 1), whose leaves represent the operands of the formula, while the mathematical operations (+, −, *, /, etc.) are represented by the remaining nodes.

A fitness function is used to evaluate the goodness of a solution. Generally, it estimates the error between the actual and expected result for all the entries of the dataset provided as input.

The algorithm evolves in a similar way to the GA principles, that is through the operations of selection, crossover, and mutation. In the selection phase, the fitness function guides the choice of the best individuals corresponding to the formulas that express a result closest to the real one. These individuals are combined through the crossover operation (also called recombination), which exchanges some sub-trees (genes) of two individuals (parents). As shown in Fig. 2, the resulting new individual (offspring) includes the best genetic information of two parents, i.e., the offspring is made up of pieces of functions belonging to the best solutions.

Depending on the data structure used to store genetic information, the crossover can be performed differently (D’Angelo and Palmieri 2021a; Hasançebi and Erbatur 2000). In order to ensure genetic diversity, the mutation operation is carried out on a number of new individuals. The mutation alters one or more genes (parts of functions) based on a user-defined probability of mutation. The set of new resulting individuals represents the new population, which is used in the next cycle of evolution. The cycle is repeated until the fitness function reaches the desired value.

Figure 3 shows a sketch of the operations described. As can be seen, the system employs a supervised leaning approach, in fact both the input and the output are provided as data to the algorithm. The output is a mathematical formula that roughly represents the relationship between the inputs and the corresponding provided numerical outputs.

This leads to the need to make some assumptions in order to implement a classifier. Firstly, the class values have been changed to take on numeric values. In particular, we have assigned the numerical value 1 to the PDAC type (Class 0), while 0 to the MCN type (Class 1). In this way, GP will try to find a formula that fits all numeric inputs (vector of features) to the corresponding numeric output (1 or 0) of the data set.

As it might expect, for a given vector of input features, the outcome of the extracted formulas with an high probability will never exactly match the corresponding numeric output class, expressed as numerical values (1 or 0).

To solve this problem, the concept of threshold has been introduced. For a given formula \( f \), the threshold associated to \( f (T') \), which ranges from 0 to 1, is chosen as the point of maximum separation among the predicted numerical output of \( f \) estimated on all items of the considered dataset. More precisely, the threshold is being chosen to maximize the \( AUC \) metric (see Eq. (8) below), as shown in Eq. (1). As reported in Eq. (2), for a given vector of features \( \Theta \), if the
result of the formula \( f(\Theta) \) is greater than or equal to this threshold, then that vector belongs to Class 1; otherwise, it belongs to Class 0.

\[
T_f = \arg\max_{th \in [0,1]} AUC(f, th) \tag{1}
\]

\[
\text{Class} = \begin{cases} 
0 & f(\Theta) < T_f \\
1 & f(\Theta) \geq T_f
\end{cases} \tag{2}
\]

In order to give an idea of this approach, in Fig. 4 an example is shown. As depicted, both actual and expected results are shown for each entry of the dataset. Note that, for the sake of clarity, all results have been grouped by class. As shown, in this case, the best threshold is 0.46. As a consequence, all the outcomes above this threshold are considered to belong to Class 1, while the others to Class 0.
3 Experiments and results

3.1 Experiments flow

Figure 5 shows a flow diagram of the implementation of the experiments carried out. As can be seen, the dataset is divided into two families: the training and the testing set. The former is used to create a model, while the latter is used to validate the reliability of the extracted model. In order to avoid the phenomenon of biasing, the fivefold cross-validation was used (Witten et al. 2011), and therefore, the scheme shown in Fig. 5 was repeated 5 times, each for each folder of the cross-validation. More specifically, the overall dataset was divided into 5 disjoint subsets (named folds), each including approximately the same number of instances. In each experiment, a single fold is used as the testing set, while the aggregation of the remaining ones is used as the training set. The outcomes of each experiment are averaged to provide a single performance evaluation result.

The reliability of the outcomes is evaluated by metrics derived from the confusion matrix associated with a binary classifier, as described in the following.

3.2 Evaluation metrics

In a binary classifier, data associated with the two classes are referred to as positive and negative instances, respectively. In our case, we identified the Class 1 as positive, whereas Class 0 as negative. The instances that are correctly classified as positive and negative are called true positive (TP) and true negative (TN), respectively. While the positive instances that are incorrectly classified as negative are called false negatives (FN). Finally, negative instances that are incorrectly classified as positive are called false positives (FP). As a consequence, the overall number of positives and negatives can be represented by \( P = (TP + FN) \) and \( N = (TN + FP) \), respectively. Their representation in a table is known as confusion matrix.

Several metrics can be derived by the confusion matrix (Davis and Goadich 2006). In our study, we used the following:

- **Accuracy** correctly classified portion of data. It is given by:
  \[
  Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}
  \]

- **Sensitivity** also known as TPR (true positive rate) portion of positives that are correctly identified as such.
  \[
  TPR = \frac{TP}{TP + FN} \tag{4}
  \]

- **Specificity** also known as TNR (true negative rate) portion of negatives that are correctly identified as such.
  \[
  TNR = \frac{TN}{TN + FP} \tag{5}
  \]

- **Precision** It measures how many positives have been correctly identified against all instances classified as positive.
Fig. 5 Experiments flow diagram. It is repeated 5 times in according with the fivefold cross-validation approach

\[ \text{Prec} = \frac{TP}{TP + FP} \]  

- **F-Measure** It is the harmonic average of precision and sensitivity.

\[ F_{\text{Meas}} = \frac{2 \times TPR \times \text{Prec}}{TPR + \text{Prec}} \]  

- **AUC (area under curve)** It measures the surface subtended by the ROC curve. It is an estimate of how much greater is the probability that a classifier will be able to rank a randomly chosen positive instance over a randomly chosen negative one.

\[ \text{AUC} = \frac{TPR + TNR}{2} \]  

Notice that these metrics range from 0 to 1. A perfect classifier achieves maximum metric values, that is 1. The closer the metrics are to zero, the poorer the classifier.

### 3.3 Implementation details

The GP-based classifier was implemented in the MATLAB environment using an open-source library for symbolic data mining, namely the GPTIPS library (genetic programming toolbox for the identification of physical systems) (Searson 2014). It uses a multi-gene-based genetic programming (MGGP) as the main engine, which is a variant of the GP approach for formula discovery. In particular, in MGGP a formula is seen as a single gene of an individual, while the individual is represented by a weighted linear combination of these genes plus a bias (offset). Optimal values for weights are automatically obtained using the least-squares approach to fit the expected result to the actual output data as closely as possible. This individual representation is also known as **symbolic scaling regression**, which differs from the native SR in the presence of bias and linear combination. The linear combination of genes can capture nonlinear behavior much more effectively than pure SR.

Using MATLAB as a core platform confers to this library the advantage of having a pluggable architecture, which means that users can easily change the algorithm setting parameters and fitness function by including their own code as an add-on in the GPTIPS library. Among the large number of configuration options that can be set, there are typical ones, such as population size, maximum number of generations, number of genes included in an individual, tournament size and mutation size.

Once these setup parameters have been specified by the user, the algorithm evolves as follows. First, a population is randomly generated. Each individual can contain a random number of genes between 1 and the user-specified \( G_{\text{max}} \) value, and a tree depth level can also be specified. It is important to point out that even though a high \( G_{\text{max}} \) can capture many hidden nonlinearity behaviors in the data, it can also lead to the problem of overfitting, i.e., the output formula fits well only to the training set, but not to unknown data. Furthermore, in this phase, the algorithm tries to guarantee the diversity between the generated individuals and between the genes belonging to the same individual. This is essential to avoid duplications, which can negatively influence the evolution of the algorithm. As in the standard GP, individuals are probabilistically selected for the crossover operation at each generation. There are two types of crossover operations, namely the low and high level crossover. In the first level, the first two individuals with a higher fitness score are chosen as
parents, then a gene is chosen at random from each parent. These genes are exchanged between parents to create two children. These descendants are then used in the new population. On the other hand, the latter allows parents to swap more than one gene, i.e., they can swap entire trees, compatibly with the respective $G_{\text{max}}$ value. However, if the exchange leads to a number of genes greater than the $G_{\text{max}}$ constraint, some genes are randomly pruned. In this way, people with better fitness scores are expected to survive and evolve in the learning process.

### 3.3.1 Experimental setup

Although many experiments were performed at varying the GP parameters, only the parameters that provided best results are shown below.

- **Initialization** A population of 100 individuals was used. Each individual can include up to five levels for the depth of the tree, a $G_{\text{max}} = 5$, and can include the following arithmetic operators $F = +, -, *, /$, $\sqrt{\cdot}$, $\exp$. The bias ranges from -10 to 10.
- **Fitness function** the fitness function provides a measure of the performance of a specific individual to solve the problem and guides the evolutionary process toward the best solution. One possible metric is to measure the error between the original and expected output of a solution. Therefore, the root mean square (RMS) of the errors calculated for the entire data set is used as a fitness function.
- **Crossover operation** A low-level crossover was used. To select individuals for crossover, the tournament scheme was used. The fitness function defined above was used for parent selection. The tournament size was fixed to 3.
- **Mutation** the mutation operation can heavily influence the results, as the mutation swaps a subtree with a new one generated randomly. If the point of exchange is a terminal node, the resulting gene only mutates slightly and the mutation does not severely affect the tree. On the other hand, a point that falls on functional nodes can cause a large variation in the tree. For these reasons, a mutation probability of zero was chosen.
- **Termination** the above steps are applied recursively in order to refine the solutions during the evolution process. This iterative process is stopped when the termination condition is reached. The commonly used termination criterion is the completion of a given number of generations. In this study, the maximum number of iterations was set to 100.

### 3.4 Results

In order to evaluate the performance of the proposed approach, the whole dataset was firstly divided into two subsets, namely evaluation and testing dataset, respectively. The former, used to train, evaluate and compare the proposal with the state-of-the-art, includes data coming from PAAD-US, PACA-CA, and PAEN-IT projects containing 5 positives and 497 negatives. The latter, used to test the proposed mathematical formula, includes data coming from PACA-AU project containing 432 negatives and 6 positives.

According to the aforementioned k-fold cross-validation technique, the evaluation dataset was randomly divided into 5 disjoint subsets and the proposed classifier was trained on each training set obtained, while its effectiveness was verified on the corresponding test set.

As can be expected, since GP is based on a heuristic approach, the formulas provided in the different tests can be very different even if the algorithm is applied on the same training set. Therefore, for each test, we repeated the experiment 20 times, and the best result was only reported.

Table 1 shows the performance of the resulting best formulas. In order to demonstrate the significance of the results, the mean value, standard deviation and confidence interval (at 95% confidence level) for all metrics are also reported.

As illustrated, the proposed GP-based classifier obtained a high score for all metrics, which confirms its effectiveness in classifying the type of tumor starting from the considered genes. In particular, it can be seen that there are only a few FPs and FNs, which leads to very high scores for the metrics. Furthermore, the higher score for AUC (0.93 on average) confirms the ability of the proposed classifier to perform better than a random classifier.

### 3.5 Comparison

Below are the results obtained using some of the most common ML techniques available in the literature, such as multilayer perceptron (MLP), Naive Bayes (NB), and decision tree (J48). All reported results were obtained by using the same folds of the previous experiments. In Table 2, the metrics averaged on all folds are only reported.

As depicted, even though the accuracy achieved excellent values (0.99 on average), sensitivities were very low (0.06 on average). This means that such techniques are not able to correctly distinguish the two classes, and, further, they classify all entries as negative. Indeed, the specificity is very high (0.99 on average).

### 3.6 Modeling

Although the high scores achieved by all metrics in the previous experiments confirm the effectiveness of the proposed classifier, we remark that the potentiality offered from the proposal is to provide a mathematical model expressing the complex relationships existing between cancer typology and
Table 1  Experimental results obtained by the fivefold cross-validation technique

| Test | Confusion matrix | Acc | TPR | TNR | Prec | FMeas | AUC |
|------|------------------|-----|-----|-----|------|-------|-----|
| 1    | 1 14 0 85        | 0.86| 1.00| 0.86| 0.07 | 0.13  | 0.93|
| 2    | 1 0 0 99         | 1.00| 1.00| 1.00| 1.00 | 1.00  | 1.00|
| 3    | 1 0 1 98         | 0.99| 0.50| 1.00| 1.00 | 0.67  | 0.75|
| 4    | 1 10 0 89        | 0.90| 1.00| 0.90| 0.09 | 0.17  | 0.95|
| 5    | 1 0 0 99         | 1.00| 1.00| 1.00| 1.00 | 1.00  | 1.00|
| Mean |                 | 0.95| 0.90| 0.95| 0.63 | 0.59  | 0.93|
| StDev|                 | 0.07| 0.22| 0.07| 0.50 | 0.43  | 0.10|
| Cin@95|               | (0.87| (0.62| (0.87| (0.00| (0.06| (0.80|
|       |                  |     |     |     |     |       |     |

Table 2  Averaged results of the fivefold cross-validation obtained by state-of-the-art techniques

| Method | Confusion matrix | Acc | TPR | TNR | Prec | FMeas | AUC |
|--------|------------------|-----|-----|-----|------|-------|-----|
| MLP    | 0 3 5 494        | 0.98| 0.00| 0.99| 0.00 | –     | 0.50|
| NB     | 1 3 4 494        | 0.99| 0.20| 0.99| 0.25 | 0.22  | 0.60|
| J48    | 0 0 5 497        | 0.99| 0.00| 1.00| –    | –     | 0.50|

As depicted, three genes (Eqs. (10), (11), and (12), respectively) are included in the individual \( F \), represented by Eq. (9). Finally, the class of the cancer is obtained by comparing \( F \) with the threshold 0.01, as shown in Eq. (13).

The formula emphasizes the complex correlation existing between involved variables and between these variables and the tumor class. Namely, in order to classify the tumor, it is very important to evaluate not only the presence or absence of gene mutations but how they are correlated. Indeed, as it can be deduced by the proposed formula, many combinations of mutations of the involved genes could give the same outcomes (1 or 0).

Finally, in order to prove the accuracy of the proposed formula, we have tested it on the datasets by PACA-AU and PAEN-AU projects including 432 negatives and 6 positives. Table 3 shows the results.

Table 3  Results obtained by testing the proposed formula on data obtained by PACA-AU project

| Confusion matrix | Acc | TPR | TNR | Prec | FMeas | AUC |
|------------------|-----|-----|-----|------|-------|-----|
| TP 5 FN 497      | 1.00| 1.00| 0.99| 0.75 | 0.86  | 1.00|

As depicted, all metrics achieved the maximum score, which confirms the effectiveness of the proposal in distinguishing the typology of cancer despite the significant unbalancing between positive and negative samples.

4 Conclusions

In this paper, the task of distinguishing the typology of pancreatic adenocarcinoma has been addressed by using a genetic programming-based approach. The main aim of this work has been to find out a mathematical formula able to express the complex relationship existing between gene mutations and cancer typology. To this end, the problem has been modeled through the usage of symbolic regression and solved by a genetic algorithm. Also, the threshold concept has been provided to turn the regression problem into a classification one.

The identification of these correlations is a typical objective of the diagnostic approach, which can be used to improve the prognosis and to provide an early diagnosis.
The formula obtained highlights the complex analysis of a large amount of data, which often is not possible to achieve by traditional statistical techniques.

Although the dataset is definitely unbalanced, the proposed approach has shown to be able to generate high reliability results when evaluated on real-world cases by the cross-validation methodology. Also, comparison with the reliability results when evaluated on real-world cases by the proposed approach has shown to be able to generate high reliability results when evaluated on real-world cases by the proposed approach.

**Author Contributions** GD was involved in the conceptualization, methodology, software, validation, formal analysis and investigation. AR, MNS, and ALC contributed to the resources. GD was involved in writing—original draft preparation. GD, AR, and FP helped in writing—review and editing. GD contributed to the visualization. GD and FP were involved in the supervision. FP and AR contributed to the project administration. FP acquired the funding. All authors have read and agreed to the published version of the manuscript.

**Funding** Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement. This research was funded by the Campania Region, Italy, through the POR FESR 2014-2020 under the Project SICED (System Innovation for Cancer Early Diagnosis), CUP B53D18000150007-SURF 17061BP000000013.

**Availability of data and materials** Publicly available datasets were analyzed in this study. These data can be found here: https://portal.gdc.cancer.gov/ https://cbioportal.org https://dcc.icgc.org.

**Declarations**

**Conflict of interests** The authors declare no conflict of interest.

**Institutional review board statement** The study was conducted according to the guidelines of the Declaration of Helsinki.

**Informed consent** No explicit informed consent was necessary from all subjects involved in the study because the dataset used is publicly available.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

**References**

Affenzeller M, Winkler S, Wagner S, Beham A (2009) Genetic algorithms and genetic programming: modern concepts and practical applications, 1st edn. Chapman-Hall, Boca Raton

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6):394–424

Carrato A, Falcone A, Ducrèux M, Valle JW, Parmaby A, Djazouli K, Alnwick-Allu K, Hutchings A, Palaska C, Parthenaki I (2015) A systematic review of the burden of pancreatic cancer in Europe: real-world impact on survival, quality of life and costs. J Gastrointest Cancer 46(3):201–211

Copyright (2011). In: Witten IH, Frank E, Hall MA (eds) Data mining: practical machine learning tools and techniques (third edition). The Morgan Kaufmann series in data management systems. Morgan Kaufmann, Boston. https://doi.org/10.1016/B978-0-12-374856-0.00025-0

Coriat R, Pellat A (2022) Updates on the treatment of pancreatic diseases: focus on surgery, electrochemotherapy and rituximab. J Clin Med. https://doi.org/10.3390/jcm11010239

D’Angelo G, Palmieri F (2020) Discovering genomic patterns in SARS-CoV-2 variants. Int J Intell Syst 35(11):1680–1698. https://doi.org/10.1002/int.22268

D’Angelo G, Palmieri F (2021a) GGA: a modified genetic algorithm with gradient-based local search for solving constrained optimization problems. Inf Sci 547:136–162. https://doi.org/10.1016/j.ins.2020.08.040

D’Angelo G, Palmieri F (2021b) Enhancing COVID-19 tracking apps with human activity recognition using a deep convolutional neural network and HAR-images. Neural Comput Appl. https://doi.org/10.1007/s00521-021-05913-y

D’Angelo G, Rampone S (2014) Towards a HPC-oriented parallel implementation of a learning algorithm for bioinformatics applications. BMC Bioinform 15(5):2. https://doi.org/10.1186/1471-2105-15-S5-S2

D’Angelo G, Pilla R, Dean JB, Rampone S (2018) Toward a soft computing-based correlation between oxyrogen toxicity seizures and hyperoxic hyperpnea. Soft Comput 22(7):2421–2427. https://doi.org/10.1007/s00500-017-2512-z

D’Angelo G, Pilla R, Tascini C, Rampone S (2019) A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees. Soft Comput 23(22):11775–11791. https://doi.org/10.1007/s00500-018-03729-y

Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning. ICML ’06. Association for Computing Machinery, New York, NY, USA, pp 233–240. https://doi.org/10.1145/1143844.1143874

Duffy J, Engle-Warnick J (2002) Using symbolic regression to infer cations. BMC Bioinform 15(5):2.https://doi.org/10.1186/1471-2105-15-S5-S2

Elia S, D’Angelo G, Palmieri F, Sorge R, Massoud R, Cortese C, Har-davella G, De Stefano A (2020) A machine learning evolutionary algorithm-based formula to assess tumor markers and predict lung cancer in cytologically negative pleural effusions. Soft Comput 24(10):7281–7293. https://doi.org/10.1007/s00500-019-04344-1

Farthing M, Roberts SE, Samuel DG, Williams JG, Thorne K, Morrison-Rees S, John A, Akbari A, Williams JC (2014) Survey of digestive health across Europe: final report. Part 1: the burden of gastrointestinal diseases and the organisation and delivery of gastroenterology services across Europe. United Eur Gastroenterol J 2(6):539–543. https://doi.org/10.1177/2050640614554154

Gisselsson D, Jonson T, Petersen A, Strönbeck B, Dal Cin P, Höglund M, Mitelman F, Mertens F, Mandahl N (2001) Telomere dysfunc tion triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. Proc Natl Acad Sci USA 98(22):12683–12688
Hasançebi O, Erbatur F (2000) Evaluation of crossover techniques in genetic algorithm based optimum structural design. Comput Struct 78(1):435–448. https://doi.org/10.1016/S0045-7949(00)00089-4
Koza JR (1994) Genetic programming as a means for programming computers by natural selection. Stat Comput 4(2):87–112. https://doi.org/10.1007/BF00175355
Lucas AL, Malvezzi M, Carioli G, Negri E, La Vecchia C, Boffetta P, Bosetti C (2016) Global trends in pancreatic cancer mortality from 1980 through 2013 and predictions for 2017. Clin Gastroenterol Hepatol 14(10):1452–14624
Momi N, Kaur S, Krishn SR, Batra SK (2012) Discovering the route from inflammation to pancreatic cancer. Minerva Gastroenterol Dietol 58(4):283–297
Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM (2014) Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 74(11):2913–2921
Scholten L, van Huijgevoort NCM, van Hooft JE, Besselink MG, Del Chiaro M (2018) Pancreatic cystic neoplasms: different types, different management, new guidelines. Visc Med 34(3):173–177
Searson DP (2014) GPTIPS 2: an open-source software platform for symbolic data mining. CoRR abs/1412.4690, 1412.4690
Siegel R, Naishadham D, Jemal A (2013) Cancer statistics, 2013. CA Cancer J Clin 63(1):11–30
van Heek NT, Meeker AK, Kern SE, Yeo CJ, Lillemoe KD, Cameron JL, Offerhaus GJA, Hicks JL, Wilentz RE, Goggins MG, De Marzo AM, Hruban RH, Maitra A (2002) Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. Am J Pathol 161(5):1541–1547
Visalakshi S, Radha V (2014) A literature review of feature selection techniques and applications: review of feature selection in data mining. In: 2014 IEEE international conference on computational intelligence and computing research, pp 1–6. https://doi.org/10.1109/ICCIC.2014.7238499
Xie W, Liang H, Guo Y, Xiao S-Y (2021) Update on mucinous cystic neoplasm of the pancreas: a narrative review. J Pancreatol 4(3):115–121

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.