

State of the Art of Artificial Intelligence Applications in Oncology

Idrissa Sy^{1*}, Mamadou Bousso¹, Ayoub Insa Correa¹, Madieng Dieng²

¹Faculty of Science and Technology, Iba Der THIAM University of Thies, Thies, Senegal

²Faculty of Medicine, Cheikh Anta DIOP University of Dakar, Dakar, Senegal

Email: *idrissa.sy@univ-thies.sn, mbousso@univ-thies.sn, ayoub@univ-thies.sn, madieng.dieng@ucad.edu.sn

How to cite this paper: Sy, I., Bousso, M., Correa, A.I. and Dieng, M. (2023) State of the Art of Artificial Intelligence Applications in Oncology. *Open Journal of Applied Sciences*, 13, 2245-2262.

<https://doi.org/10.4236/ojapps.2023.1312175>

Received: October 1, 2023

Accepted: December 5, 2023

Published: December 8, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Artificial intelligence (AI) operates by using algorithms and statistical models based on data, enabling computers to imitate a real form of intelligence. The structure of the data available and the aim of the research define the technique to be adopted, which will be evaluated by its degree of accuracy and its capacity for generalization. In recent years, several applications of artificial intelligence have emerged in the fight against cancer, due to its development, computing power and learning potential. This article presents the current state of AI systems, describing the techniques and innovations that have led to satisfactory results in the fight against cancer.

Keywords

Artificial Intelligence, Cancer, Prognosis

1. Introduction

Artificial Intelligence (AI) is a system of theories and techniques developed from complex programs that enable computers to imitate a real form of intelligence [1].

Machine Learning [2] is a branch of AI capable of analysing behaviour using algorithms that are themselves fed and trained by a large amount of data. Confronted with a variety of situations, the algorithm makes the appropriate decision based on its training, and creates a model. Depending on the situation, there are two main categories: supervised and unsupervised methods. Supervised Machine Learning is applied to labelled data sets. The computer is presented with examples of inputs and the desired outputs, and searches for solutions to find the best hypothesis that maps the input data to the desired output. Supervised methods are used to address two types of problem: Classification, when the

output variable is categorical (living/dead, in/out, malignant/benign, etc.) and Regression if it is continuous (survival time). The main supervised algorithms are: random forests, decision trees, k-Nearest Neighbours, linear regression, Naïve Bayes algorithm, Support vector machine, logistic regression, and neural networks.

In unsupervised Machine Learning, the data is unlabelled and there is no notion of output values. The algorithm is left to its own devices and will learn the characteristic structures and relationships buried in the data. Unsupervised approaches provide solutions to clustering and association problems. The main unsupervised machine learning algorithms are as follows: K-Means, hierarchical clustering, principal component analysis and dimensionality reduction.

Deep Learning [3] is a branch of Machine Learning. The system is based on several layers of neural networks, which combine different algorithms inspired by the human brain. Whereas a neural network comprises one to three layers of neurons and requires training, a Deep Learning model can comprise a much higher number of layers to process unstructured information such as sound, text or pictures, without prior training. The most commonly used algorithms are Convolutional Neural Networks, Recurrent Neural Networks, Radial Basis Function Networks and Long and Short-Term Memory Networks.

AI methods have been used in cancer research. But, despite the deployment of a multitude of solutions, cancer remains a major public health problem worldwide. It is one of the main causes of death and a serious obstacle to increasing life expectancy. According to estimates by the World Health Organisation, in a report on global cancer statistics [4], there will be around 19.3 million new cases of cancer and 10 million deaths worldwide in 2020. According to the same report, the global burden of cancer is expected to be 28.4 million cases in 2040, an increase of 47% compared with 2020, with a greater increase in developing countries (64% to 95%) than in developed countries (32% to 56%).

This alarming situation has prompted researchers to develop new strategies for achieving higher degrees of accuracy.

The aim of this work is to search the PubMed, Embase, Web of Science and Cochrane Library databases to present the current state of artificial intelligence techniques and their application in the field of oncology. We focused on articles published between 2017 and 2023, with a methodological innovation, tending to test several AI techniques to propose the best and/or by comparing the results obtained with those of human experts in the field.

In this manuscript, we present five categories of methods encountered: ensemble Learning models, neural networks, transfer learning, unsupervised approaches, and model mixtures.

2. Artificial Intelligence Methods Encountered

2.1. Ensemble Learning

Ensemble learning [5] is a technique that relies on the combination of multiple

machine learning algorithms to increase the performance of the model, and achieve a higher level of accuracy than would be achieved by using one of these algorithms separately. Algorithms mainly encounter two sources of error that prevent them from generalising beyond their learning set. These are Variance and Bias.

Variance is the error due to sensitivity to small fluctuations in the training set. A high variance can lead to overlearning. Bias is the error arising from incorrect assumptions in the learning algorithm. A high bias may be linked to a lack of relevant relationships between the input data and the predicted outputs (under-learning). To cover the maximum amount of information contained in the data and minimise processing errors, Ensemble learning methods are grouped into two main families: parallel methods and sequential methods.

Bagging or bootstrap aggregation (L. Kabari *et al.* [6]) is a learning method where models are trained in parallel. The aim is to reduce variance. It involves sub-sampling the data, creating a data set for each model. To determine the final result, a vote is taken for classification, or the mean for regression.

To predict mortality associated with gastric cancer, but also the occurrence of any complication following surgery, C. Neto *et al.* [7] trained with the bagging method, a decision tree algorithm, simple logistic regression and a Random Forest. The study showed that the decision tree algorithm was the best technique for predicting mortality, with an accuracy of around 74%. Furthermore, the Random Forest algorithm showed the best results for predicting the possible occurrence of complications, with an accuracy of around 83%.

For K. Kaulanjan *et al.* [8], in their study to predict biological recurrence after prostatectomy in prostate cancer patients, the Area Under the Curve (AUC) obtained were 0.65, 0.64 and 0.57 respectively for the Random Forest, Support Vector Machine (SVM) and k-Nearest Neighbour (KNN) models.

Decision Trees, Random Forest and support vectors machine algorithms were trained in parallel to identify long non-coding ribonucleic acid (RNA) as a diagnostic biomarker for gastric adenocarcinoma by Qun *et al.* [9]. The diagnostic capacity of these three models was evaluated by acquiring the area under the Receiver Operating Characteristic curve, sensitivity and specificity. The AUCs were 0.797, 0.981 and 0.983, and the specificity and sensitivity of the three models were 75.0% and 97.1%; 96.9% and 96%; 96.9% and 97.1%, respectively. The authors conclude that the machine learning models are able to diagnose adenocarcinoma of the stomach in the following decreasing order of accuracy: support vectors machine, Random Forest, Decision Trees.

The Random Forest algorithm is special in that it works on its own with the Bagging method. It is made up of several decision trees, trained independently on sub-samples of the training data-set. Each tree produces an estimate and a majority vote is used to obtain the final result.

C. Xu *et al.* [10] used a Random Forest to predict the survival status of patients with gastric cancer. External validation was performed and performance was assessed by the area under the ROC curve. The results showed an AUC of

76.6% using majority voting. The authors then decided to weight the voting process according to the accuracy tested for each tree instead of assigning the same weight as the original Random Forest (RF) model. The Random Forest algorithm with weighted voting gave an area under the ROC curve of 78.9%, an increase of 3% compared with the majority process.

Boosting is the Sequential ensemble Learning family. It helps reduce bias. The algorithms are run one after the other, and each one learns from the errors of the previous one. It improves the accuracy and predictive performance of models by converting several weak learners into a single strong learning model. Among the most common boosting techniques are Gradient boosting and Adaboosting.

Adaptive boosting (AdaBoost) [11] was one of the first boosting models developed. It adapts and attempts to self-correct with each iteration of the process. AdaBoost initially gives the same weight to each data set. It then automatically adjusts the weights of the data points after each iteration. It gives more weight to misclassified items in order to correct them for the next round. It repeats the process until the residual error, or the difference between the real and predicted values, falls below an acceptable threshold. Overall, AdaBoost is more appropriate for classification problems.

L. Fan *et al.* [12] used adaptive boosting (AdaBoost), linear discriminant analysis, and logistic regression classifiers to predict the presence of cancer cells inside a lymphatic vessel (lymph vascular invasion). Lymph vascular invasion is associated with metastases and poor survival. It is also difficult to diagnose. The data used was a combination of computed tomography images and clinical variables from patients with gastric cancer. A reference pathological examination was carried out to validate the prognosis. The AdaBoost technique obtained the best result with an AUC of 94%.

Gradient Boosting is similar to AdaBoost in that it is also a sequential technique. The difference is that Gradient Boosting does not give more weight to misclassified elements. Instead, it optimises the loss function by generating baseline learners sequentially, so that the current baseline learner is always more effective than the previous one. This method attempts to generate accurate results at the outset rather than correcting errors throughout the process. Gradient Boosting can help solve classification and regression problems.

W. Leung *et al.* [13] sequentially trained several machine learning algorithms to predict the risk of gastric cancer within 5 years of *Helicobacter pylori* eradication. The data consisted of clinical variables. Performance was measured by analysis of the area under curve (AUC) of the receiver operating characteristic. During a mean follow-up of 4.7 years, 0.21% of patients in whom *Helicobacter pylori* had been eradicated developed gastric cancer again. The gradient boosting technique performed well in terms of prediction, with an AUC of 97%. This result was superior to that of conventional logistic regression (AUC of 90%), applied independently.

The applications of ensemble learning methods in oncology, developed in this section, are recapitulated in **Table 1**.

Table 1. Application of ensemble Learning methods.

Reference	Method	Results
C. Neto <i>et al.</i> [7]	bagging -Decision trees (DT) -Logistic regression (LR) -Random Forest (RF)	RF: 83% Accuracy in predicting complications. Gastric cancer
K. Kaulanjan <i>et al.</i> [8]	bagging -RF -Support Vector Machine (SVM) -k-Nearest Neighbor (KNN)	RF: 65% Accuracy in predicting recurrence Prostate cancer
Qun <i>et al.</i> [9]	bagging -DT -RF -SVM	SVM: AUC: 98% for gastric cancer diagnosis
C. Xu <i>et al.</i> [10]	-RF	AUC: 78.9%, survival prognosis gastric cancer
L. Fan <i>et al.</i> [12]	-AdaBoost -Linear discriminant analysis -Logistic regression	AdaBoost: AUC: 94% to predict lymphovascular invasion
W. Leung <i>et al.</i> [13]	-Gradient Boosting (GB) -RL	GB: AUC: 97%, prognosis of gastric cancer occurrence

2.2. Artificial Neural Networks

Artificial neural networks (ANN) are algorithmic imitations of the functions of the human brain. Like biological neurons, they receive input information, process it through a series of operations and produce output. Once properly trained, artificial neural networks can learn from themselves and constantly update themselves to provide increasingly accurate results.

The perceptron [14] is historically the first artificial neural network model. It has two layers of neurons: an input layer and an output layer. The two layers are directly connected. This type of algorithm is useful for the linear classification of a set of information into two distinct classes (present/absent, alive/dead, benign/malignant, etc.). But given the availability of massive amounts of data, its unstructured nature and the complexity of the relationships between characteristics, neural networks are being deployed in greater depth.

Previously, image analysis has been based principally on features such as colour, brightness, shape, texture pattern and other distinguishing characteristics. However, this type of analysis is limited by image rotation, lack of brightness, adjacent or angled views of the object, or image blur [15].

Deep Learning involves introducing one or more hidden layers between the input and output layers in order to perform intermediate operations and unearth buried relationships and relevant features. Depending on the structure of the data, specific neural networks are developed.

For example, Convolutional Neural Network [16] is particularly well adapted

to image processing and is widely used in cancer research. It uses convolution operations to extract features from images by applying filters that detect characteristics in the image. Convolution layers are usually followed by pooling layers that reduce the size of images while retaining key features. With their self-learning capability, the programme works without any indication of what is required.

Neural networks have made a significant contribution to cancer diagnosis, notably for the detection of genetic types, subtypes and mutations [17], the identification of rearrangement status in prostate cancer adenocarcinoma [18], and the evaluation of pelvic echography images [19]. Some studies have reported that the performance of neural networks surpasses that of medical specialists in the fight against cancer.

Lianlian Wu *et al.* [20] have developed a system using a convolutional neural network to detect early gastric cancer. Images of different parts of the stomach were collected to train the network to monitor blind spots and automatically cover suspected cancerous areas. The results were compared with the findings of an endoscopy team. For locating cancer cells, the network achieved 90% accuracy, on a par with endoscopy experts. The model also achieved 92.5% accuracy in classifying tumours as malignant or benign, surpassing the performance of the experts.

L. Xiangchun *et al.* [21] are using a convolutional neural network to diagnose thyroid cancer. The data consisted of echography pictures of thyroid cancer patients and healthy controls. A clinical diagnosis of the training set was performed by 16 qualified radiologists and a reference pathological examination was carried out for confirmation. The model achieved almost the same sensitivity as the qualified radiologists (93.4% versus 96.9%), but much improved specificity (86.1% versus 59.4%), in identifying patients with thyroid cancer.

The authors subsequently included serological markers in addition to echographic images in a new study [22] with the same objective. The previous model was refined and now consists of two parallel branches. One branch takes the input image and outputs a vector F representing the input image. The other branch takes the abundance of serological markers as input and outputs a G vector, a learned characteristic of the serological markers. The item-by-item summation of F and G was considered the integrated multimodal feature. Labels obtained from pathology reports were used as the diagnostic reference. The performance of the new model was improved compared with that of qualified radiologists (AUC: 95% versus 89%).

Neural networks have also been applied to cancer prognosis.

Nakahira H *et al.* [23] developed a convolutional neural network using endoscopic images. The aim was to stratify the risk of gastric cancer. The model was trained using stochastic gradient descent. The system identified three distinct groups of patients: low, moderate and high risk. Endoscopic examination and serum antibody tests were performed to validate the results and the prevalence were: 2.2%, 8.8% and 16.4% respectively for each group, with significant differences ($p = 0.0017$).

L.Q. Zhou *et al.* [24] trained a convolutional neural network to predict the occurrence of clinically negative lymph node metastases in patients with primary breast cancer, based on echographic images. A reference pathological examination was performed. The model achieved a sensitivity of 82% compared with 63% for expert radiologists.

The applications of neural networks in oncology, developed in this section, are resumed in **Table 2**.

2.3. Transfer Learning

Transfer learning [25] is a machine learning technique that transfers to a model the knowledge acquired by another model previously trained for a similar task with sufficient data. The pre-trained model is called the “source” and the model receiving the transfer is the “target”. The weights and biases learned by the source are then transferred to the target. Transfer learning can be useful when the data available for the target task is limited, or to reduce the time and resources required for training. This method is increasingly used in the fight against cancer, where it can improve the accuracy of tumour classification according to type, stage or aggressiveness [26] [27], tumour detection [28], and cancer prognosis [29].

DenseNet networks [30] have an architecture called densely connected convolutional network. They are renowned for their prowess in extracting optimal features. In their study to extract relevant features for the identification of primary gastric cancer or lymphoma, B. Feng *et al.* [31] used DenseNet121 parameters on Whole slide images. The relevance of the factors selected was assessed in comparison with those selected by a team of expert clinicians. A logistic regression with the selected factors was used for validation and obtained an AUC of 96%, higher than the results of the clinical model.

A. Saber *et al.* [32] have developed a model based on transfer learning to provide effective support for the automatic diagnosis of breast cancer suspects.

Table 2. Neural network applications.

Reference	Method	Results
Lianlian Wu <i>et al.</i> [20]	Convolutional Neural Network (CNN)	Accuracy: 92.5% for classification as malignant or benign gastric tumor
L. Xiangchun <i>et al.</i> [21] [22]	CNN	[20]: <i>Sensitivity</i> 93.4% <i>specificity</i> 86.1 [21]: <i>AUC</i> : 95%. <i>For the diagnosis of thyroid cancer.</i>
Nakahira H. <i>et al.</i> [23]	CNN	Risk group: low, moderate and high Prevalences: 2.2%, 8.8% and 16.4%. $p = 0.0017$. Gastric cancer
L.Q. Zhou <i>et al.</i> [24]	CNN	82% sensitivity for predicting breast cancer metastases

Image features are extracted using a pre-trained architecture: VGG16 [33]. VGG16 is a convolutional neural network trained on a collection of more than 14 million images belonging to 22,000 categories. Experimental results show that the features selected enable breast cancer to be diagnosed with 98.96% accuracy.

ResNet-50 [34] is a convolutional neural network with 50 layers of depth. This model introduces residual connections. Unlike standard convolutional neural networks, which have a linear architecture (a stack of layers in which each output is connected only to the next layer), in a residual network the output of previous layers is connected to the output of new layers, so that they are both transmitted to the next layer. This architecture enables the creation of very deep neural networks, with greater precision, because they are able to extract more information and thus perform a more advanced analysis of pictures.

Z. Yafang *et al.* [35] used the performance of ResNet50 to extract features from echographic images to establish a prognostic model for hepatocellular carcinoma. The importance of the selected factors was compared with that of clinical parameters using a survival analysis. The results showed specificity (81.0% versus 38.1%) and accuracy (78.8% versus 51.5%) compared with the clinical model. In addition, microvascular invasion was identified as having a very poor prognosis for overall survival, with a Hazard Ratio of 6, and for relapse-free survival, with a Hazard Ratio of 3.3.

The applications of transfer learning in oncology, developed in this section, are summarized in **Table 3**.

2.4. Unsupervised Approaches

Unsupervised approaches play a major role in making accessible and explainable the information contained in masses of unlabelled data, with no notion of output values. They make it possible to learn the characteristic structures and relationships buried in the data. These algorithms are left to their own devices to discover and present the interesting structure of the data. Unsupervised techniques are used to perform clustering based on similarities or differences in the data, in order to focus the search on significantly representative sub sets. Analysis of genomic data is also made possible by reducing dimensionality.

P. Apostolou *et al.* [36] conducted an unsupervised clustering analysis based on hierarchical and k-means algorithms to discriminate between colon cancer and normal conditions as well as between different types of cancer. The data

Table 3. Transfer learning applications.

Reference	Method	Results
B. Feng <i>et al.</i> [31]	DenseNet [30]	AUC: 96% for identification of gastric cancer or lymphoma
A. Saber <i>et al.</i> [32]	VGG16 [33]	Accuracy: 98.96% for breast cancer diagnosis
Z. Yafang <i>et al.</i> [35]	ResNet-50 [34]	Accuracy: 78.8% for prognosis of hepatocellular carcinoma

consisted of blood sets from colon cancer patients, normal donors and commercial cancer cell lines representing different types of gastrointestinal cancers. Gene expression analysis was performed for more than fifty genes. The Euclidean distance metric was used to highlight dissimilarity between the data. The distribution of the data was evaluated using the Kolmogorov-Smirnov test. The study effectively separated: 1) colon cancer cell lines 2) normal sets 3) stomach, pancreatic and liver cancers.

J. Gal *et al.* [37], evaluated several unsupervised approaches to clustering according to survival in breast cancer patients, exploiting metabolomic signatures. These were Principal Component Analysis (PCA), k-means, Sparse k-means, Single-cell Interpretation via Multi-kernel Learning (SIMLR), k-sparse and Spectral clustering. The results show that three optimal groups (favourable, intermediate and unfavourable) were selected. In terms of processing times, PCA and k-means were the fastest and K-sparse was the longest. SIMLR and k-sparse methods were the most discriminants with an average silhouette value of 0.85 and 0.91, respectively. Survival analysis revealed a significant difference in predicted 5-year overall survival between the 3 groups. This approach shows the possibility of stratifying breast cancer patients using metabolomic signatures with unsupervised approaches.

Dimension reduction allows the study to focus on a finite number of significant and representative factors. For example, it is very difficult to analyse genomic data with tens of thousands of dimensions. Principal component analysis reduces dimensionality by converting a set of correlated variables into a set of principal components (*i.e.*, linearly uncorrelated variables) using an orthogonal transformation.

Gene expression data from several cancers, including gastric cancer, were exploited in the study by J. Xie *et al.* [38]. The aim was to identify optimal biomarkers for the detection of cancer cells. Principal component analysis was applied. All the characteristics are displayed in a two-dimensional space with the standard deviation defining discernability on the abscissa and the cosine similarity defining independence on the ordinate. By calculating the area bounded by the coordinate lines and the principal axes, the optimal features are located in the top right-hand corner of the space. To validate the results, SVM models are built with the detected biomarkers. The results show 100% accuracy for classification into malignant or benign tumours. As a result, the proposed technique is powerful for detecting subsets whose features are relatively independent of each other.

F. Yang *et al.* [39] studied 3385 differentially expressed genes obtained from single-cell RNA sequencing data of gastric cancer specimens. The aim was to find biomarkers with high sensitivity and specificity to accurately assess patient prognosis. Principal component analysis was applied to the dimension reduction process. Three cell subsets were identified: gastric cells, plasmacytoid dendritic cells and T-memory cells. In addition, the fatty acid binding protein (FABP1) was identified as the pivotal gene in gastric cancer progression. FABP1 was

found to be closely related to long-term survival and age at diagnosis of patients.

The applications of unsupervised approaches to oncology developed in this section are summarized in **Table 4**.

2.5. Mixtures of Models

Model mixture consists of combining different machine learning techniques, in order to exploit the advantages of each to create a more complete and more effective model. Faced with ever-growing volumes of data, the complexity of the relationships between features, and the performance requirements imposed on researchers, a single learning model may be limited in its ability to cover all the information contained in the data. The mixture strategy is like a distribution of tasks. The algorithms used each process a specific task according to its efficiency and the structure of the data. When applied to cancer data, the mixtures of model's strategy have produced very satisfactory accuracies.

O. Iizuka *et al.* [40] developed a neural network model for the classification of histopathology pictures of the stomach and colon. In the feature extraction step, they trained convolutional neural networks for each organ separately. These networks were based on the pre-trained inception-v3 architecture [41], removing the fully connected classification layer. Each image contained an arbitrary number of extracted features. Knowing that a Recurrent Neural network (RNR) [42] can take a sequence of arbitrary length and produce a single output, all the extracted features are used as input to an RNR with two Long Short-Term Memory (LSTM) layers. The final model achieved good classification performance with an AUC of 96%, compared with 86% for expert pathologists and only 41% for students.

H. Zhang *et al.* [43] combined several models to detect gastric cancer from histopathological images. First, pre-trained neural network architectures were

Table 4. Application of unsupervised approaches.

Reference	Method	Results
P. Apostolou <i>et al.</i> [36]	k-means	3 clusters identified -Colon cancer cell lines -Normal samples -Stomach, pancreas and liver cancers.
J. Gal <i>et al.</i> [37]	-Principal Component Analysis (PCA), -k-sparse -Spectral clustering	3 risk groups identified for breast cancer -Favorable -Intermediate -Unfavorable
J. Xie <i>et al.</i> [38]	PCA	Identifying optimal biomarkers for cancer cell detection
F. Yang <i>et al.</i> [39]	PCA	Three subsets are highlighted: -Gastric cells -Plasmacytoid dendritic cells -T-memory cells

used for image segmentation. Next, feature extraction is facilitated by clustering using the k-means method. Finally, different classification algorithms were trained with the selected factors. The results of this study show that using the convolutional network “U-Net” for image segmentation, feature extraction based on graphs and finally the Support Vector Machine classifier, gastric cancer can be detected with an accuracy of 94.29%.

A. Dongyao *et al.* [44] and A. Moloud *et al.* [45] have also combined convolutional neural network with Support Vector Machine, and obtained accuracies of 99.3% and 100%, respectively, to provide accurate diagnosis of cervical cancer and breast cancer, respectively.

In order to predict the occurrence of lymph node metastases, Z. Xue *et al.* [46] conducted a retrospective study using clinical and pathological data from patients with stage 1 and 2 gastric cancer who had undergone surgical treatment. Univariate logistic regression was used to detect variables significantly affecting metastasis. These variables were given as input to a Multi-Layer Perceptron (MLP) neural network to establish a prognostic model. The results were compared with the real postoperative state. Logistic regression showed that the platelet/lymphocyte ratio, systemic immune inflammation index, tumour size and clinical stage were closely related to lymph node metastasis. On the basis of these variables, the MLP model achieved an AUC of 75%, compared with 54% for post-surgical anatomopathological staging and only 30% for pre-therapeutic clinical staging. This technique makes it possible to predict metastases in the early stages of cancer.

K. Priyanka *et al.* [47] developed a hybrid model to predict the complete response of breast cancer before the start of chemotherapy. A convolutional neural network is used for image feature extraction. Mann-Whitney tests were used to evaluate the features and their relevance. Then, the selected factors were provided independently as input to different machine learning classifiers including (Random Forest, SVM, DT, KNN etc.). Classification performance was evaluated using accuracy and the area under the receiver operating characteristic curve (AUC). The results show that the combination of the CNN and the KNN classifier performed best with an accuracy of 99.8% and an AUC of 100%.

The applications of model mixing in oncology developed in this section are summarized in **Table 5**.

Recent developments in applications based on artificial intelligence offer a multitude of possibilities for exploiting data and improving levels of precision. Strategies based on ensemble learning, deep learning, transfer learning and unsupervised approaches make algorithms increasingly flexible in adapting to data and finding solutions to research questions.

Choosing the right model for a classification or regression, in the presence of labelled data, is made easier by the bagging technique, which enables several models to be trained in parallel. Boosting also means that algorithms can be run sequentially to produce a strong learning model.

Table 5. Model mixing applications.

Reference	Method	Results
O. Iizuka <i>et al.</i> [40]	-CNN -RNN -Inception-v3 [45]	Accuracy: AUC of 96%, for the classification of stomach and colon cancers
H. Zhang <i>et al.</i> [43]	-k-means -U-Net -SVM	Accuracy 94% in detecting gastric cancer
A. Dongyao <i>et al.</i> [44]	-CNN -SVM	Accuracy: 99.3% for diagnosis of cervical cancer
A. Moloud <i>et al.</i> [45]	-CNN -SVM	Accuracy: 100%, breast cancer diagnosis
Z. Xue <i>et al.</i> [46]	-Régression logistique -Multi-Layer Perceptron -(MLP)	AUC 75% for gastric cancer staging
K. Priyanka <i>et al.</i> [47]	-CNN -KNN	AUC: 100% to predict complete response in breast cancer

Deep learning enhances the self-learning capacity of neural networks. Intermediate calculation layers enable complex and unstructured data volumes to be exploited in depth. Convolutional Neural Networks are particularly well suited to image processing, and are widely used in cancer research.

Transfer Learning can be used to overcome the lack of data in some contexts. The knowledge of a pre-trained model with sufficient data can be used to solve a similar problem with few data of the same dimensions. Algorithms such as ResNet, DenseNet, VGG16, etc. are reused in new studies.

Unsupervised approaches are used to learn structures and characteristic relationships buried in the presence of unlabelled data. These algorithms are left to their own mechanisms to perform clustering based on data similarity or dimensionality reduction.

Hybrid models involve using several algorithms to solve a problem. Perfect coordination is established and each model executes a specific task according to its efficiency and the structure of the data.

All these techniques help to considerably reduce the time and resources needed to train the models, and enable high levels of accuracy to be achieved.

Despite the reported success of AI, there are still a few limits to the widespread use of artificial intelligence solutions.

Data bias is a problem faced by artificial intelligence algorithms. To illustrate, let's take the case of the Cancer Genome Atlas (TCGA), which is considered to be the largest repository of diverse cancer datasets. However, it is mainly composed of data from white individuals of European descent [48]. Consequently, the results proposed by the algorithms may be subject to errors due to selection bias [49].

Deep learning has the reputation of being a "black box". Neural networks do

not provide explanations for their results, which limits the analysis of existing phenomena [50]. Non-interpretability affects understanding and confidence in proposed solutions. Fortunately, some researchers are tackling this problem, with promising results [51] [52].

Access to sufficient high-quality training data is also a major obstacle to the adoption and widespread use of artificial intelligence solutions [53]. Although transfer learning seems to offer the beginnings of a solution to this data shortage, it is currently limited to image exploration, for which data of the same structure and dimensions are readily available. For numerical variables, on the other hand, the problem of standardization is an obstacle. In fact, there is no standard codification or unit of measurement.

Synthetic data generation in turn offers a judicious approach to this situation. It consists in generating artificial data from existing real data, based on their distributions and correlations between attributes [54].

This synthetic data can then be used for a variety of purposes, such as training machine learning models, validating data analysis methods, or preserving the confidentiality of sensitive data.

In order to allow Random Forest algorithms to continue incrementing without resorting to old data at each iteration, J. Gonzalez *et al.* [55] proposed a method consisting in the generation of synthetic data using the normal distribution.

H. Akrami *et al.* [56] have also developed a variational autoencoder for tabular datasets with categorical and continuous features that is robust to outliers in the training data.

Lei Xu *et al.* [57] have designed a CT-GAN (Conditional Tabular-Generative Adversarial Network) that relies on mode-specific normalization for continuous columns, architectural changes, and consideration of data imbalance for discrete columns.

3. Conclusion and Perspective

Artificial intelligence has been constantly evolving since its inception. Researchers are constantly learning to propose new solutions that surpass the limits encountered by existing techniques. In application to cancer data, many researchers have reported impressive AI performance, which was sometimes superior to that of human experts and standard statistical methods.

The current state of progress is presented in detail in this manuscript. All the methods encountered can be found in the model blending strategy. It offers a number of possibilities by bringing together several techniques, each dealing with a specific task according to its efficiency and the structure of the available data.

This literature review enables us to refine and select the best methodology for a study we are conducting on the prognosis of death from stomach cancer in Senegal. We have data on the clinical and pathological factors of patients with gastric cancer who have undergone treatment by surgery alone or by surgery

combined with chemotherapy. Our study sample consists of tabular data (discrete and continuous) from 262 patients with 45 explanatory variables.

First, we will use various synthetic data generation techniques to amplify our database as much as possible, while respecting the distribution of real data.

These data will then be used to train several classification algorithms using bagging, boosting and model mixing methods, in order to select the combination offering the best accuracy.

The results will be compared with those obtained with a previously developed logistic regression [58].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Dinh-Xuan, A. (2019) Intelligence artificielle, machine learning et deep learning: Nouveaux concepts et futurs acteurs clés en pneumologie. *Revue des Maladies Respiratoires*, **11**, 59-62. [https://doi.org/10.1016/S1877-1203\(19\)30031-X](https://doi.org/10.1016/S1877-1203(19)30031-X)
- [2] Azencott, C.-A. (2022) Introduction au Machine Learning. 2nd Edition, Dunod, Paris, 272 p.
- [3] Charniak, E. (2021) Introduction au Deep Learning. Dunod, Paris, 176 p.
- [4] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021) Statistiques mondiales sur le cancer 2020: Estimations GLOBOCAN de l'incidence et de la mortalité dans le monde pour 36 cancers dans 185 pays. *CA: A Cancer Journal for Clinicians*, **71**, 209-249. <https://doi.org/10.3322/caac.21660>
- [5] Dong, X., Yu, Z., Cao, W., et al. (2020) A Survey on Ensemble Learning. *Frontiers of Computer Science*, **14**, 241-258. <https://doi.org/10.1007/s11704-019-8208-z>
- [6] Kabari, L.G. and Onwuka, U.C. (2019) Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a Classifier. *International Journals of Advanced Research in Computer Science and Software Engineering*, **9**, 19-23.
- [7] Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A. and Machado, J. (2019) Application of Data Mining for the Prediction of Mortality and Occurrence of Complications for Gastric Cancer Patients. *Entropy*, **21**, Article No. 1163. <https://doi.org/10.3390/e21121163>
- [8] Kaulanjan, K., Darde, T., Auclin, E., Le Coent, Q., Blanchet, P. and Brureau, L. (2022) Approche par intelligence artificielle de la prédiction de la récurrence biologique après prostatectomie dans une population d'ascendance africaine. *Progrès en Urologie-FMC*, **32**, S55-S56. <https://doi.org/10.1016/j.fpurol.2022.07.064>
- [9] Li, Q., Liu, X., Gu, J., Zhu, J., Wei, Z. and Huang, H. (2020) Screening lncRNAs with Diagnostic and Prognostic Value for Human Stomach Adenocarcinoma Based on Machine Learning and mRNA-lncRNA Co-Expression Network Analysis. *Molecular Genetics & Genomic Medicine*, **8**, e1512. <https://doi.org/10.1002/mgg3.1512>
- [10] Xu, C., Wang, J., Zheng, T., Cao, Y. and Ye, F. (2022) Prediction of Prognosis and Survival of Patients with Gastric Cancer by a Weighted Improved Random Forest Model: An Application of Machine Learning in Medicine. *Archives of Medical Science*, **18**, 1208-1220. <https://doi.org/10.5114/aoms/135594>

- [11] Bahad, P. and Saxena, P. (2020) Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics. In: Singh Tomar, G., Chaudhari, N.S., Barbosa, J.L.V. and Aghwariya, M.K., Eds., *International Conference on Intelligent Computing and Smart Communication 2019*, Springer, Singapore, 235-244. https://doi.org/10.1007/978-981-15-0633-8_22
- [12] Fan, L., Li, J., Zhang, H., et al. (2022) Machine Learning Analysis for the Noninvasive Prediction of Lymphovascular Invasion in Gastric Cancer Using PET/CT and Enhanced CT-Based Radiomics and Clinical Variables. *Abdominal Radiology*, **47**, 1209-1222. <https://doi.org/10.1007/s00261-021-03315-1>
- [13] Leung, W.K., Cheung, K.S., Li, B., Law, S.Y. and Lui, T.K. (2021) Applications of Machine Learning Models in the Prediction of Gastric Cancer Risk in Patients after *Helicobacter pylori* Eradication. *Alimentary Pharmacology & Therapeutics*, **53**, 864-872. <https://doi.org/10.1111/apt.16272>
- [14] Saporta, G. (2018) Une brève histoire de l'intelligence artificielle.
- [15] Takiyama, H., Ozawa, T., Ishihara, S., et al. (2018) Automatic Anatomical Classification of Esophagogastroduodenoscopy Images Using Deep Convolutional Neural Networks. *Scientific Reports*, **8**, Article No. 7497. <https://doi.org/10.1038/s41598-018-25842-6>
- [16] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. <https://doi.org/10.1145/3065386>
- [17] Tang, D., Wang, L., Ling, T., Lv, Y., Ni, M., Zhan, Q. and Zou, X. (2020) Development and Validation of a Real-Time Artificial Intelligence-Assisted System for Detecting Early Gastric Cancer: A Multicentre Retrospective Diagnostic Study. *EBioMedicine*, **62**, Article ID: 103146. <https://doi.org/10.1016/j.ebiom.2020.103146>
- [18] Dadhania, V., Gonzalez, D., Yousif, M., et al. (2022) Leveraging Artificial Intelligence to Predict ERG Gene Fusion Status in Prostate Cancer. *BMC Cancer*, **22**, Article No. 494. <https://doi.org/10.1186/s12885-022-09559-4>
- [19] Gao, Y., Zeng, S., Xu, X., Li, H. and Yao, S. (2022) Deep Learning-Enabled Pelvic Ultrasound Images for Accurate Diagnosis of Ovarian Cancer in China: A Retrospective, Multicentre, Diagnostic Study. *The Lancet Digital Health*, **4**, e179-e187. [https://doi.org/10.1016/S2589-7500\(21\)00278-8](https://doi.org/10.1016/S2589-7500(21)00278-8)
- [20] Wu, L., Zhou, W., Wan, X., Zhang, J., Shen, L., Hu, S., Ding, Q., Mu, G., Yin, A., Huang, X., Liu, J., Jiang, X., Wang, Z., Deng, Y., Liu, M., Lin, R., Ling, T., Li, P., Wu, Q., Jin, P., Chen, J. and Yu, H. (2019) A Deep Neural Network Improves Endoscopic Detection of Early Gastric Cancer without Blind Spots. *Endoscopy*, **51**, 522-531. <https://doi.org/10.1055/a-0855-3532>
- [21] Li, X., Zhang, S., Zhang, Q., Wei, X. and Pan, Y. (2019) Diagnosis of Thyroid Cancer Using Deep Convolutional Neural Network Models Applied to Sonographic Images: A Retrospective, Multicohort, Diagnostic Study. *The Lancet Oncology*, **20**, 193-201. [https://doi.org/10.1016/S1470-2045\(18\)30762-9](https://doi.org/10.1016/S1470-2045(18)30762-9)
- [22] Zhang, Q., Zhang, S., Pan, Y., et al. (2022) Deep Learning to Diagnose Hashimoto's Thyroiditis from Sonographic Images. *Nature Communications*, **13**, Article No. 3759. <https://doi.org/10.1038/s41467-022-31449-3>
- [23] Nakahira, H., Ishihara, R., Aoyama, K., Kono, M., Fukuda, H., Shimamoto, Y., Nakagawa, K., Ohmori, M., Iwatsubo, T., Iwagami, H., Matsuno, K., Inoue, S., Matsuura, N., Shichijo, S., Maekawa, A., Kanesaka, T., Yamamoto, S., Takeuchi, Y., Higashino, K., Uedo, N., Matsunaga, T. and Tada, T. (2019) Stratification of Gastric Cancer Risk Using a Deep Neural Network. *JGH Open*, **4**, 466-471.

- <https://doi.org/10.1002/jgh3.12281>
- [24] Zhou, L.Q., Wu, X.L., Huang, S.Y., Wu, G.G., Ye, H.R., Wei, Q., Bao, L.Y., Deng, Y.B., Li, X.R., Cui, X.W. and Dietrich, C.F. (2020) Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. *Radiology*, **294**, 19-28. <https://doi.org/10.1148/radiol.2019190372>
- [25] Yaddaden, A., Harispe, S. and Vasquez, M. (2021) Apprentissage par transfert: Du TSP au VRP. *ROADEF 2021-22e congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision*, Mulhouse, April 2021, hal-03347689.
- [26] Hornbrook, M.C., et al. (2017) Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Digestive Diseases and Sciences*, **62**, 2719-2727. <https://doi.org/10.1007/s10620-017-4722-8>
<https://link.springer.com/article/10.1007/s10620-017-4722-8>
- [27] Zhu, Y., Wang, Q.C., Xu, M.D., Zhang, Z., Cheng, J., Zhong, Y.S., Zhang, Y.Q., Chen, W.F., Yao, L.Q., Zhou, P.H. and Li, Q.L. (2019) Application of Convolutional Neural Network in the Diagnosis of the Invasion Depth of Gastric Cancer Based on Conventional Endoscopy. *Gastrointestinal Endoscopy*, **89**, 806-815.e1. <https://doi.org/10.1016/j.gie.2018.11.011>
- [28] Giri, A., Chauhan, A. and Singh, P. (2023) An Optimized Transfer Learning and Deep Convolutional Neural Network Approach for Automated Breast Cancer Detection. *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, 25-27 August 2023, 1-6. <https://doi.org/10.1109/ASIANCON58793.2023.10270493>
- [29] Ayana, G., Dese, K. and Choe, S.-W. (2021) Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging. *Cancers*, **13**, Article No. 738. <https://doi.org/10.3390/cancers13040738>
- [30] Zhong, Z., Zheng, M., Mai, H., Zhao, J. and Liu, X. (2020) Cancer Image Classification Based on DenseNet Model. *Journal of Physics: Conference Series*, **1651**, Article ID: 012143. <https://doi.org/10.1088/1742-6596/1651/1/012143>
- [31] Feng, B., Huang, L., Liu, Y., Chen, Y., Zhou, H., Yu, T., Xue, H., Chen, Q., Zhou, T., Kuang, Q., Yang, Z., Chen, X., Chen, X., Peng, Z. and Long, W. (2022) A Transfer Learning Radiomics Nomogram for Preoperative Prediction of Borrmann Type IV Gastric Cancer from Primary Gastric Lymphoma. *Frontiers in Oncology*, **11**, Article ID: 802205. <https://doi.org/10.3389/fonc.2021.802205>
- [32] Saber, A., Sakr, M., Abo-Seida, O.M., Keshk, A. and Chen, H. (2021) A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique. *IEEE Access*, **9**, 71194-71209. <https://doi.org/10.1109/ACCESS.2021.3079204>
- [33] Yang, H., Ni, J., Gao, J., Han, Z. and Luan, T. (2021) A Novel Method for Peanut Variety Identification and Classification by Improved VGG16. *Scientific Reports*, **11**, Article No. 15756. <https://doi.org/10.1038/s41598-021-95240-y>
- [34] Behar, N. and Shrivastava, M. (2022) ResNet50-Based Effective Model for Breast Cancer Classification Using Histopathology Images. *CMES-Computer Modeling in Engineering & Sciences*, **130**, 823-839. <https://doi.org/10.32604/cmescs.2022.017030>
- [35] Zhang, Y., Wei, Q., Huang, Y., Yao, Z., Yan, C., Zou, X., Han, J., Li, Q., Mao, R., Liao, Y., Cao, L., Lin, M., Zhou, X., Tang, X., Hu, Y., Li, L., Wang, Y., Yu, J. and Zhou, J. (2022) Deep Learning of Liver Contrast-Enhanced Ultrasound to Predict Microvascular Invasion and Prognosis in Hepatocellular Carcinoma. *Frontiers in Oncology*, **12**, Article ID: 878061. <https://doi.org/10.3389/fonc.2022.878061>
- [36] Apostolou, P., Iliopoulos, A.C., Parsonidis, P. and Papatotiriou, I. (2019) Gene Ex-

- pression Profiling as a Potential Predictor between Normal and Cancer Samples in Gastrointestinal Carcinoma. *Oncotarget*, **10**, 3328-3338. <https://doi.org/10.18632/oncotarget.26913>
- [37] Gal, J., Bailleux, C., Chardin, D. and Pourcher, T. (2020) Comparison of Unsupervised Machine-Learning Methods to Identify Metabolomic Signatures in Patients with Localized Breast Cancer. *Computational and Structural Biotechnology Journal*, **18**, 1509-1524. <https://doi.org/10.1016/j.csbj.2020.05.021>
- [38] Xie, J.Y., Wang, M.Z., Xu, S.Q., Huang, Z. and Grant, P.W. (2021) The Unsupervised Feature Selection Algorithms Based on Standard Deviation and Cosine Similarity for Genomic Data Analysis. *Frontiers in Genetics*, **12**, Article ID: 684100. <https://doi.org/10.3389/fgene.2021.684100>
- [39] Yang, F., Gan, L., Pan, J., Chen, Y., Zhang, H. and Huang, L. (2022) Integrated Single-Cell RNA-Sequencing Analysis of Gastric Cancer Identifies FABP1 as a Novel Prognostic Biomarker. *Journal of Oncology*, **2022**, Article ID: 4761403. <https://doi.org/10.1155/2022/4761403>
- [40] Iizuka, O., Kanavati, F., Kato, K., et al. (2020) Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Scientific Reports*, **10**, Article No. 1504. <https://doi.org/10.1038/s41598-020-58467-9>
- [41] Wang, C., Chen, D., Hao, L., Liu, X., Zeng, Y., Chen, J. and Zhang, G. (2019) Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model. *IEEE Access*, **7**, 146533-146541. <https://doi.org/10.1109/ACCESS.2019.2946000>
- [42] De Campos, L.M.L. and Duarte, D.S. (2020) Application of Recurrent and Deep Neural Networks in Classification Tasks. *Revista Gestão & Tecnologia*, **20**, 59-79. <https://doi.org/10.20397/2177-6652/2020.v20i3.1709>
- [43] Zhang, H., Li, C., Ai, S., Chen, H., Zheng, Y., Li, Y. and Grzegorzec, M. (2022) Application of Graph Based Features in Computer Aided Diagnosis for Histopathological Image Classification of Gastric Cancer.
- [44] Jia, A.D., et al. (2020) Detection of Cervical Cancer Cells Based on Strong Feature CNN-SVM Network. *Neurocomputing*, **411**, 112-127. <https://doi.org/10.1016/j.neucom.2020.06.006>
- [45] Abdar, M. and Makarenkov, V. (2019) CWV-BANN-SVM Ensemble Learning Classifier for an Accurate Diagnosis of Breast Cancer. *Measurement*, **146**, 557-570. <https://doi.org/10.1016/j.measurement.2019.05.022>
- [46] Xue, Z., Lu, J., Lin, J., Huang, C.M., Li, P., Xie, J.W., Wang, J.B., Lin, J.X., Chen, Q.Y. and Zheng, C.H. (2022) Establishment of Artificial Neural Network Model for Predicting Lymph Node Metastasis in Patients with Stage II-III Gastric Cancer. *Chinese Journal of Gastrointestinal Surgery*, **25**, 327-335.
- [47] Khanna, P., Sahu, M., Singh, B.K. and Bhateja, V. (2023) Early Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer MRI Images Using Combined Pre-Trained Convolutional Neural Network and Machine Learning. *Measurement*, **207**, Article ID: 112269. <https://doi.org/10.1016/j.measurement.2022.112269>
- [48] Yuan, J., Hu, Z., Mahal, B.A. and Zhao, S.D. (2018) Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell*, **34**, 549-560.e9. <https://doi.org/10.1016/j.ccell.2018.08.019>
- [49] Phillips, S.P., Spithoff, S. and Simpson, A. (2022) L'intelligence artificielle et les algorithmes prédictifs en médecine: Promesses et problèmes. *Canadian Family Physician*, **68**, e230-e233. <https://doi.org/10.46747/cfp.6808e230>
- [50] Bhinder, B., Gilvary, C., Madhukar, N.S. and Elemento, O. (2021) Artificial Intelli-

- gence in Cancer Research and Precision Medicine. *Cancer Discovery*, **11**, 900-915. <https://doi.org/10.1158/2159-8290.CD-21-0090>
- [51] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision (ICCV)*, Venice, 22-29 October 2017, 618-626. <https://doi.org/10.1109/ICCV.2017.74>
- [52] Molnar, C. (2020) Interpretable Machine Learning.
- [53] Hahn, W., Schütte, K., Schultz, K., Wolkenhauer, O., Sedlmayr, M., Schuler, U., Eichler, M., Bej, S. and Wolfien, M. (2022) Contribution of Synthetic Data Generation towards an Improved Patient Stratification in Palliative Care. *Journal of Personalized Medicine*, **12**, Article No. 1278. <https://doi.org/10.3390/jpm12081278>
- [54] Van der Schaar, M. and Qian, Z. (2023) AAAI Lab for Innovative Uses of Synthetic Data.
- [55] Gonzalez, J. and Dama, F. (2021) Génération de données synthétiques à partir d'une forêt aléatoire.
- [56] Akrami, H., Aydore, S., Leahy, R.M. and Joshi, A.A. (2020) Robust Variational Autoencoder for Tabular Data with Beta Divergence.
- [57] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019) Modeling Tabular Data Using Conditional GAN. *NIPS19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 7335-7345.
- [58] Idrissa, S.Y., Bousso, M., Correa, A.I., Loum, M.A., Diop, A., Toure, K., Traore, B., Diallo, A.T. and Dieng, M. (2022) Study of Prognostic Factors in Gastric Cancer: Application of a Cox Model and Logistic Regression. *International Journal of Statistics and Applied Mathematics*, **7**, 108-113. <https://doi.org/10.22271/math.2022.v7.i5b.887>