UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN
Ingegneria Industriale

CICLO XXXVI

Development of Artificial Intelligence based Systems for Biomedical Applications

Settore Scientifico Disciplinare
ING/IND-15

Dottorando
Dott. Magherini Roberto

Supervisore
Prof. *Governi Lapo*

Coordinatore
Prof. Ferrara Giovanni

Dedicated to ...

Abstract

The recent and more advanced applications of artificial intelligence (AI) reached a wide range of fields, transforming traditional workflows, perfecting current techniques, and introducing new paths not previously feasible. The common approach to AI-based systems relies on using an appropriately annotated database to train a model in order to identify a correlation between the input data and the desired output. With regard to the biomedical field, the literature does not provide a unified framework to follow for the creation of these tools, but rather proceeds heterogeneously. In this context, this work focuses on the study and creation of a framework intended to facilitate and enable the implementation of performing tools based on AI in the biomedical field. With the aim of providing an effective framework the production cycle of AI-based applications in the biomedical field has been studied. In particular, the framework was developed by analysing in detail all the implementation steps necessary for the development of new AI-based tools in the biomedical field and three main phases were identified: the clinical phase, the artificial intelligence engineering phase, and the application development phase.

This work was carried out in collaboration with multiple medical research centers: the joint laboratory Custom3D, which brings together the Azienda Ospedaliera Universitaria Careggi and the Department of Industrial Engineering of the University of Florence; the T3Ddy laboratory, in collaboration between the Meyer Children's Hospital and the Department of Industrial Engineering of the University of Florence; a collaboration with the Center for Biomedical Technology of the Universidad Politecnica de Madrid. Within these partnerships, four case studies were analyzed to devise, use and refine the framework. The case studies concern the following medical sectors: 1) urology - with the study of renal tumors; 2) plastic surgery - for the automation of the production process of guides used for anatomical reconstruction of the ear; 3) psychiatry - for the identification of risk factors in patients with suicidal intentions; 4) neurology - for the evaluation of a therapy for the reduction and control of brain tumors.

The implementation of the four case studies was carried out following the phases defined in the framework and using best practices for the implementation of artificial intelligence models.

With regard to the first case study a model was developed to differentiate malignant clear cell renal cell carcinoma tumors and benign oncocytoma tumors, in the event that these are very small and difficult to interpret by expert doctors with a sensitivity of 94.59%.

In the second case study, two AI-based tools were created to be used in the production process of surgical guides used by the surgeon to create anatomical replica of the patient's ear. In particular, these tools are able to generate the depth map from a simple image of the ear obtained from a normal camera, without the need to use more complex tools such as 3D acquisition scanners, with final MSE (mean square error) of $\sim$0.07 and an average SSIM (structure similarity) of $\sim$0.80, and to segment and identify the anatomical elements of interest within the depth map image of the patient's healthy ear with a 90% of accuracy considering each ear component as an independent class.

For the third case, a classification model was developed using the clinical records of psychiatric patients. This tool is able to differentiate between two types of patients, those admitted for attempted suicide and those admitted for suicidal ideation with a final accuracy of $\sim$85%. Through the creation of this tool, it was also possible to carry out a study on the major risk factors that distinguish these two types of patients.

Finally, for the last case study an application was developed that allows calculating the percentage of mouse brain volume occupied by glioblastoma multiforme tumors, reaching an average dice score of 84.48%. All for the purpose of evaluating the effects of optical hyperthermia to counteract and limit the growth and development of tumor cells through the use of nanoparticles of different materials.

Table of Contents

1. Introduction

The last decades witnessed an increasing need for the use of computer applications in various sectors. These applications aim to assist, speed up, and automate tasks and jobs, especially those that are repetitive and require a certain level of precision [1]. Among these types of applications, those based on artificial intelligence (AI) are increasingly finding their space. AI has managed to emerge in recent years thanks to the high availability and affordability of hardware resources (GPU - Graphical Processing Unit) and the increasing number of public datasets, thus becoming widely used in all sectors with different methods and purposes [2]. These include applications in the economic field for predicting possible market trends [3, 4], in the automotive sector for the development of self-driving cars [5], or for example in the industrial sector for predicting faults at the level of assembly line machinery [6, 7].

In recent years the healthcare field is rapidly progressing. In modern medicine, thanks to technological advancement, there are more and more tools available to doctors for faster diagnosis that is precise and accurate [8]. An example of this can be seen with the use of devices capable of producing diagnostic images such as computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound (US), which allow doctors to identify possible suspicious masses, such as tumors, where timely treatment, especially in cases where the mass is malignant, can make a big difference in the successful outcome of therapies adopted by doctors [9–11]. Other examples are given by all the instrumentation used for continuous patient monitoring, which continuously detect patient conditions allowing doctors to detect possible symptoms related to specific diseases and perform detailed tests to verify their actual presence [12, 13] .

In this context, thanks to methodological advancements in areas such as image processing [14], signal processing [15], and natural language processing [16], AI is able to find wide use in healthcare. Through the creation of AI-based tools it is possible to make assisted diagnoses of diseases [17], identify suspicious masses [18], segment areas of interest in the body for surgical planning purposes [19], predict patient behavior affected by disorders [20], and achieve many other things with high speed and precision thanks to continuous improvement of existing models and increased availability of public health databases [21]. The growing number of applications based on artificial intelligence highlights common and main challenges: the possibility of making these types of algorithms usable at a clinical level, overcoming ethical and regulatory issues. To this is added the strategic objective of ensuring the use of robust, reliable, and interpretable AI models for doctors, as risks and prejudices due to a model trained insufficiently or on a dataset not generalizing enough for the cases considered are to be avoided [21].

The goal of this work is to leverage the best existing artificial intelligence techniques with the aim of creating applications that can be a cornerstone for the development of new systems to be introduced into common clinical practice in the future, so they can be used directly by medical staff to improve the general level of healthcare. The idea is to provide tools to support doctors in different areas of medicine in order to simplify and improve medical decisions, reducing the overall time needed to make these choices, such as in the diagnosis of a disease, surgical planning, or evaluation of the results related to a new therapy.

In this direction, collaborations between the Department of Industrial Engineering and hospitals and research centers have allowed the development of various applications. Among these collaborations we find the Azienda Ospedaliera Universitaria di Careggi (AOUC) [22], included in the joint laboratory Custom3D, and the Meyer Children's Hospital in Florence [23], which led to the creation of the joint laboratory T3Ddy [24]. In both joint laboratories, doctors and engineers collaborate with the aim of introducing innovative and customized technzologies for patient treatment. Another important collaboration is with other European universities, as in the case of the Center for Biomedical Technology (CTB) [25] of the "Universidad Politecnica de Madrid" (UPM). Within the various collaborations, different types of case studies have been identified. As for the collaboration with AOUC, a main case study related to kidney tumors has been identified. More specifically, two main tasks have been identified: 1) being able to differentiate, in case of kidney tumors with

reduced size, between malignant and benign tumor; 2) knowing that we are dealing with a malignant type of kidney tumor, being able to diagnose its severity. In the T3Dddy laboratory, two clinical scenarios of interest were identified, the first related to plastic surgery for autologous ear reconstruction, the second concerning psychiatry, to be able to distinguish between a real attempted suicide and a fake one. Finally during collaboration with CTB, the identified case study concerns the evolution of brain tumors in rats following the use of new experimental therapies. For simplicity in dealing with the different case studies, since each one is related to a different medical area, they will from now on be identified based on this last one. Consequently there are four case studies in total: 1) "Urology" case study for collaboration with AOUC; 2) "Plastic Surgery" case study for autologous reconstruction within T3Dddy laboratory; 3) "Psychiatry" case study, also in collaboration with Meyer pediatric hospital; and finally 4) "Neurology" case study from the cooperation with CTB.

Regarding the Urology case study, the aim is to carry out an in-depth study of patients suffering from kidney cancer, using only the patients' diagnostic images, with the aim of being able to classify kidney tumors into malignant and benign. The need is to develop a tool capable of supporting the doctor during the diagnostic phase, especially in cases where the identification of a malignant tumor is very difficult. In addition, through automated grading, it is possible to obtain a real-time response, which can be of great help when choosing the type of therapy or intervention to be carried out.

In the second case study, Plastic Surgery, the aim is to automate the procedure previously created within the T3Dddy laboratory for the creation of guides, with the aim of helping the doctor in autologous ear reconstruction in children. The procedure is based on a first step of 3D scan acquisition of the healthy ear of the patient through the usage of specific 3D scanners, followed by a second step of manual elaboration of the scan where the specific ear components guides are designed and, finally, the third step in which the guides are effectively created and given to the medical staff to be used. The main challenge is to be able to obtain an artificial intelligence-based tool capable of performing some steps of the current procedure in a completely automated way, mainly intending to automate the most complex processes that require the intervention of expert personnel. All these aspects will be clarified and explained in detail in the specific case study chapter.

The Psychiatry case study mainly focuses on pediatric patients suffering from psychiatric disorders, due to which they have been led to an attempted suicide. The development of an AI-based tool aims to create an application capable, starting from patients' clinical data, of being able to distinguish between patients who have attempted suicide and those who have only staged it to determine the severity of the patient's condition.

Finally, the Neurology case study aims to be able to identify and quantify the tumor inside the rats’ brain through the sole use of diagnostic images of rats. The estimate of tumor size is used for therapeutic research purposes, as CTB is developing new therapeutic methodologies aimed at eliminating tumor cells without necessarily having to remove them surgically from patients.

The thesis is structured as follows: Chapter 2 contains all notions related to AI necessary for better understanding concepts and technologies used for developing tools shown for each case study; Chapter 3 will mainly consist of process leading to ideation and development of applications for biomedical purposes; Chapters 4, 5, 6 and 7 detail highlighted case studies; finally, conclusions are reported in Chapter 8.

2. Artificial Intelligence in medical field

Artificial intelligence is a branch of computer science that in recent years has seen its interest grow exponentially, mainly thanks to the latest technologies based on natural language processing (NLP) [16], such as ChatGPT [26], capable of automatically generating text by answering specific questions. However, AI is born and used for many purposes and is not limited exclusively to NLP. Wanting to give a definition to AI, the most recent definition of the 2018 European Commission Communication [27] can be adopted: AI refers to systems that display intelligent behavior by analyzing their environment and taking action – with some degree of autonomy – to achieve specific goals.

Considering the AI methodologies developed in recent decades, these are all based on 'data-driven' approaches. These methodologies are included within machine learning (ML) and have the peculiarity of being able to automate the learning process of algorithms, automatically learning significant relationships and patterns from the data used [28]. Even though the most common techniques used in ML use approaches that are not particularly new, the key factor that has allowed the latest advancements has been the massive increase in available data [29].

In medicine, one of the main challenges that must be considered and overcome is the very limited availability of a quantity of correctly annotated data in most of the problems that are sought to be addressed [30]. Despite this challenge and other problems that limit and hinder the development of robust tools based on AI, such as privacy issues in the case of patient data processing and ethical issues in the case of assisted diagnosis [31], many studies are conducted in all clinical areas [32]. Among these fields we find applications capable of supporting a complete management of health services and improving predictive medicine, clinical decision-making process and patient data analysis and diagnostics [33–41]. In particular, AI systems can provide healthcare operators with real-time informational updates from various sources to coordinate care and allow appropriate health risk alerts and outcome predictions [33]. Such applications allow hospitals and health services to work more efficiently, optimizing logistics, training staff and analyzing electronic medical records [34–36]. Furthermore, by identifying significant patterns in data, AI can support diagnostic and therapeutic predictions to embrace proactive disease management and personalize patient care plans [37, 38]. It also has the ability to assist in health decisions, speeding up care delivery and reducing costs [32]. As for patient data and diagnoses, AI techniques can manage vast amounts of information generated by clinical activities to discover useful insights for treatment [42]. This technology can also be used for rehabilitation therapy, robotic surgery, remote patient monitoring and protection of sensitive data [39–41].

Since this work mainly focuses on the development of new AI-based applications, rather than the creation of new AI techniques, it is good to explain all the concepts that are used within this document in such a way as to be able to understand more simply how existing methodologies have been exploited and combined for use in specific case studies. The rest of the chapter will focus on providing all the basic technical information on AI, especially on the methodology based on ML, and finally explain the more advanced concepts of ML that have been taken into consideration.

3. Basic concepts on Machine Learning

Machine learning is a subset of AI that involves the construction of computational models capable of learning and making predictions or decisions independently based on the data provided. These models continually improve their accuracy through learned data. Arthur Samuel, the first user of the term machine learning, used this phrase in 1959 to describe "the ability of computers to learn without directly programming new skills" [43].

The main types of ML are supervised, unsupervised, and reinforcement learning: i) Supervised ML assumes that the model has been trained on a dataset similar to the problem in question, consisting of input data and corresponding output data. Once the relationship between input and output is learned, the model is able to classify new unknown datasets and make predictions or decisions based on them. ii) Unsupervised machine learning differs from supervised learning in that it uses unannotated data, which has not been previously labeled by humans or algorithms. The model learns from input data without expected values, and the available dataset does not provide answers to the assigned task. Instead of labeling or predicting outputs, this algorithm focuses on grouping data based on their characteristics. The goal is to teach the machine to detect patterns and group data without a single correct answer. It mainly relies on two methods: clustering and association. Clustering involves grouping data based on their similarities and differences. Association is a method of analyzing relationships between data in a dataset. iii) Reinforcement learning assumes that the agent learns by interacting with the environment through a process of trial and error, without the need for supervised training examples. The agent selects actions and observes feedback or rewards resulting from the environment. Based on these interactions, the agent refines an action strategy or learns the value of each action to maximize a cumulative reward in the long term. In some cases, the agent can also build a model of the environment to predict future states. This type of learning aims to optimize agent behavior in complex environments through exploration and exploitation. It therefore differs from supervised learning as it does not require labeled examples, but the agent must discover the optimal strategy on its own.

In general it must be considered that all the problems can be seen as one of these three main tasks: classification, prediction, and detection. The classification task consists of all the possible cases in which the goal is to obtain a well-defined category, such in the context of diagnosis, in which the objective is to obtain, for example, a specific pathology associated with possible patient's symptoms. For the prediction the aim is to guess a certain value given different interesting features, such as predict the possible therapy to treat some pathologies or predict cancer patients' life expectancy. Finally, the detection process primarily involves identifying elements through various methods, such as segmenting. A prime example of this is the segmentation of tumors in diagnostic images, which aids in tumor identification and preoperative planning.

Added to this is the importance of the features used to train the model. These features are the independent variables used as inputs to the AI. Through various mathematical operations, linear and nonlinear, these features produce the dependent variable, i.e., the output. Given their critical role in determining the final output, there is a process that can be employed to create new features or modify existing ones to try to improve the performance of the model. This process is called feature engineering, where features are typically created using data domain knowledge to highlight patterns important to the learning algorithm.

After careful analysis, this work will mainly focus on supervised machine learning techniques for applications in the healthcare field. Although there are valid unsupervised and reinforcement approaches, current literature indicates a prevalence of supervised models in this domain [44], probably due to the necessity of annotated datasets for this specific application domain. In particular, as discussed later, supervised algorithms such as neural networks, decision trees and SVMs have been applied to multiple clinical tasks, from image diagnostics to risk prediction. Therefore it is believed that, at the current state, the supervised paradigm offers the most relevant and concrete advantages for the development of artificial intelligence systems in the medical-health field. Of course, being an active and rapidly evolving research field, new hybrid or

unconventional approaches may emerge in the future. For these reasons, this paragraph reports the most common machine learning techniques typically used with a supervised learning approach.

3.1. Basic Machine Learning techniques

Below are all the basic machine learning techniques that will be covered in this document. The aim is to provide the reader with the minimum necessary tools to understand what it is discussed in the following chapters, without having to insert too cumbersome explanations in each of them. The order in which the different techniques are reported is based on the complexity of the algorithm.

3.1.1. Linear Regression

Linear regression predicts the value of a dependent variable from an independent variable. It generates a simple, interpretable formula for predictions and is widely used across a range of fields like science, biology, business, and behavioural science [45], an example is depicted in Figure 1.

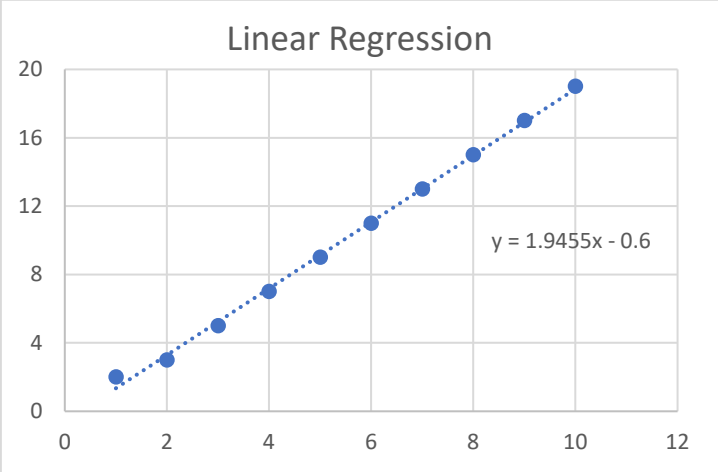


Figure 1 - Simple linear regression example: the dashed line is the linear regression line, and the equation is its formula.

An example of the use of these linear regression algorithms is combining multiple linear regression to estimate the carotid-to-femoral pulse wave velocity [46].

3.1.2. Logistic Regression

Logistic regression [47] is a supervised machine learning method used to predict the probability that an observation belongs to a particular class, typically binary (0/1, true/false, positive/negative). Unlike linear regression, logistic regression is suitable for categorical dependent variables and does not require a linear relationship between independent and dependent variables. It works by modeling the probability of belonging to a class through the logistic function, which returns values between 0 and 1, in Figure 2 an example.

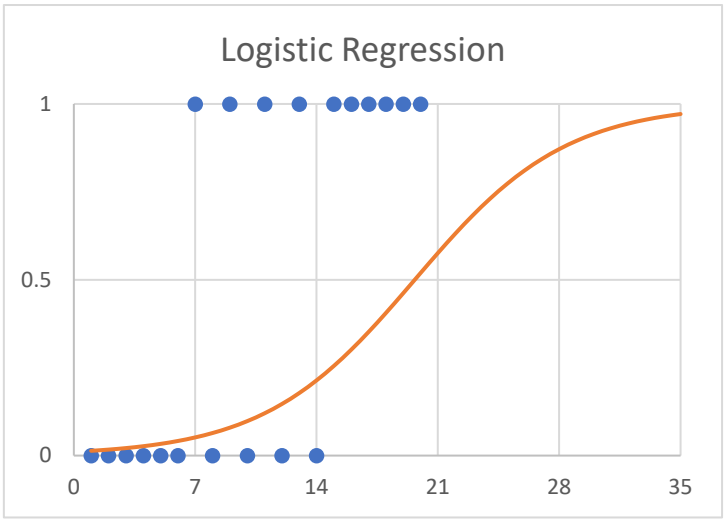


Figure 2 - Simple logistic regression example: the blu points are the samples to which are associated the corrispective binary value, in orange the logit function.

The parameters of the model are estimated through the maximum likelihood method. Logistic regression is used in much scientific research concerning medicine to classify observations into groups and make predictions [48]. For example, in the medical field it can predict the presence of a disease based on symptoms and risk factors and differentiate its type [49] or estimate the probability of survival given the patients' characteristics [50].

3.1.3. Decision Trees

Decision trees [51] are a non-parametric supervised learning approach used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees create a flowchart-like structure where each internal node represents a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent class labels or regression values. Decision trees are easy to interpret and visualize, require little data preparation, and can capture non-linear relationships.

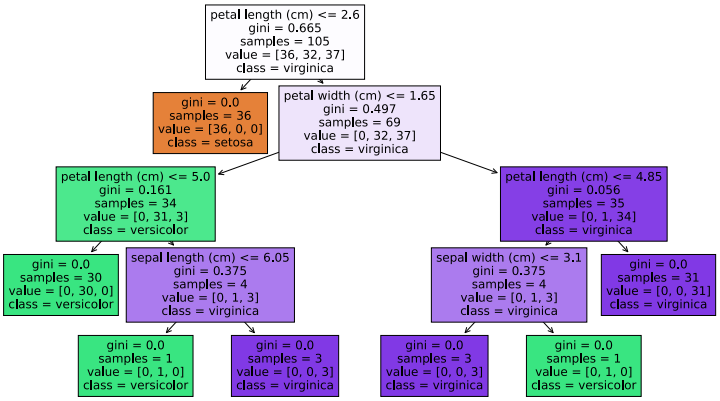


Figure 3 - Example of a decision tree trained using the iris dataset [52]

They have been used in healthcare for diagnosis, survival analysis, risk prediction, and treatment selection. As in the case of identifying homogeneous subgroups defined by combinations of individual characteristics [53] or to detect and predict urine infections [54].

3.1.4. Random Forest

Random forests [55] are an ensemble machine learning technique that operate by constructing a multitude of decision trees at training time. For each tree in the forest, a random sample of the data points and features is used. The predictions from all the individual decision trees are then averaged to produce the final random

forest prediction. By combining many decision trees, random forests overcome the tendency of individual trees to overfit their training set. The randomness introduced also decorrelates the trees so that the final model has reduced variance over a single estimator.

Random forests are robust to noise, can model complex nonlinear relationships, and are commonly used in medical applications like disease diagnosis [56], like diabetes diagnosis [57], clinical risk assessment [58], image analysis, and biomarker discovery [59], and other cases like incident stroke prediciton [60].

3.1.5. K-Nearest Neighbour

The k-nearest neighbour (kNN) [61] method is a popular method used in data mining and statistics. The kNN method is a type of algorithm that predicts the correct class of the test data by calculating the distance between the test data and all the training points. It then returns the number of k (training) points that are close to the test data. In Figure 4 is represented an example of k-NN usage.

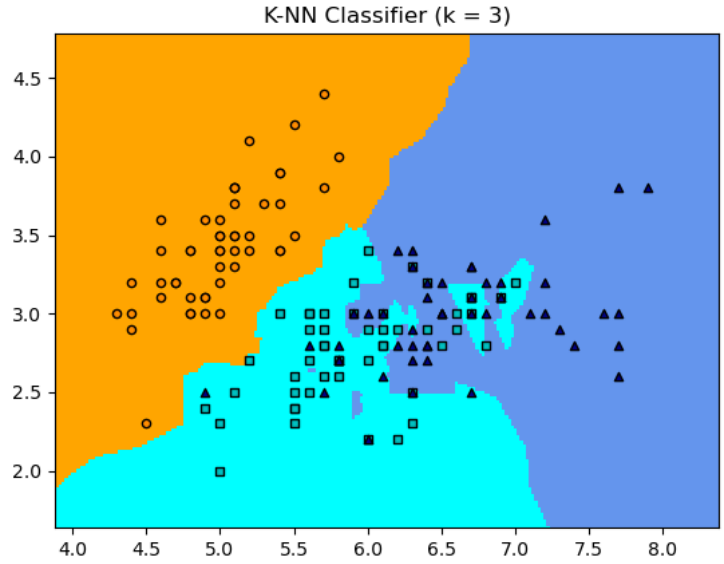


Figure 4 - K-NN classifier example with K equal to 3

kNN can be used for several medical application, like in the classification of COVID-19 cases [62] using a variant algorithm based on it, and in the classification of three medical UCI datasets [63], containing numerical and statistical data [64].

3.1.6. Naïve Bayes

Naïve Bayes classifiers [65] work by calculating the probability of a data point belonging to each class, given the values of the data point's features. This is done using Bayes' theorem which states that:

$$P(A \mid B) = P(B \mid A) * \frac{P(A)}{P(B)}$$

Where P(I | J) is the probability that event I happens, knowing that event J has already happened, and P(I) the probability that event I happens. In the context of naïve Bayes classifiers, event A is the data point belonging to a particular class, and event B is the data point having certain values for its features.

It can be used for text analysis determining whether the text is positive, negative, or neutral. The Naive Bayes classifier assumes feature independence and quickly categorises information based on this assumption. It is widely used in spam filtering, text classification, and sentiment analysis [66], and it can be also used in diagnosis of clinical disease [56].

3.1.7. Support Vector Machines

Support Vector Machines [67] (SVMs) are a type of supervised learning algorithm that can be used for both classification and regression tasks. SVMs work by finding a hyperplane in the data that best separates the data points into two classes. A three class example is shown in Figure 5.

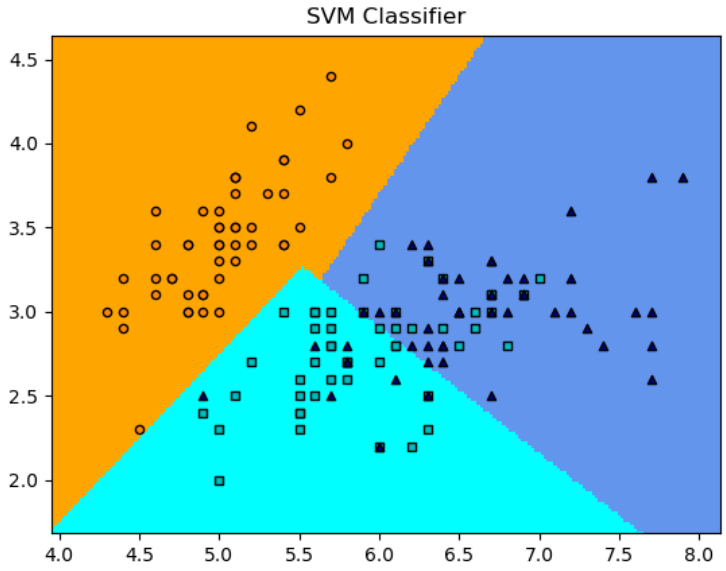


Figure 5 - A simple example of SVM classifier: in this simplified 2D visualization, it is possible to see how the data is divided by linear elements.

Their ability to handle non-linear data and high-dimensional classification tasks, makes SVMs suitable for processing large datasets [68]. SVMs can be used also for diagnostic application, like to diagnose diabetes using tongue photographs [69], diagnose heart disease and cancer [70].

3.1.8. AdaBoost

AdaBoost [71], short for adaptive boosting, is a machine learning algorithm used for binary classification and regression tasks. It is a boosting technique that iteratively trains multiple weak learners on different subsets of the training data and assigns higher weights to the misclassified instances in each iteration. In subsequent iterations, the algorithm focuses more on the misclassified samples, allowing the weak learners to learn from their mistakes and improve on their performance. The weak learners are then combined to form a single strong classifier. An example is shown in Figure 6.

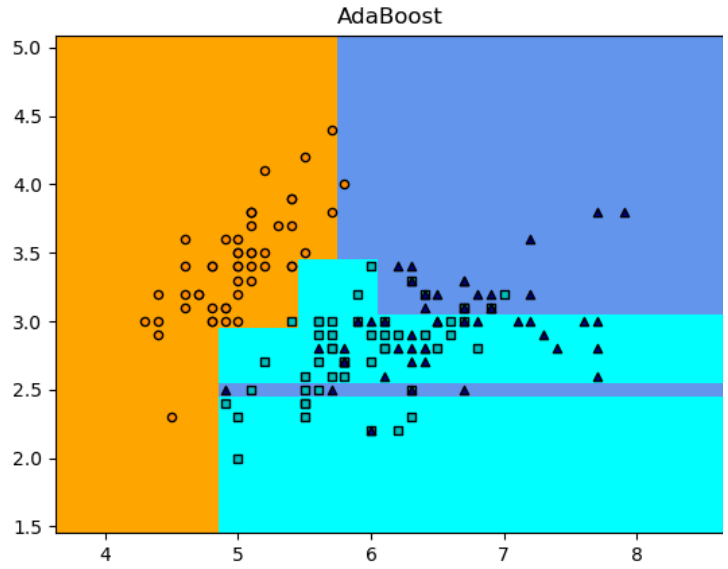


Figure 6 - Example of an AdaBoost classifier.

AdaBoost is widely used in computer-aided diagnosis (CAD) and can support medical practitioners to make critical decisions regarding their patients' disease conditions, such as diabetes, Alzheimer's disease, cancers, and hypertension[72, 73].

3.1.9. XGBoost

Extreme gradient boosting, also known as XGBoost, is a powerful machine learning algorithm that has gained significant popularity across various domains, including medicine. It is an ensemble learning method that aggregates the predictions of many individually trained weak decision trees to create a more accurate and more powerful model [74] and it is particularly well-suited for tasks with large and complex datasets [75].

In medicine, XGBoost has been applied to a wide range of tasks, like disease diagnosis, prognosis, treatment selection, and patient outcome prediction, such as the case of predicting myocardial infarction [76], or the case of treatment typology prediction [77].

3.1.10. Neural Networks

Neural networks (NN) are widely used in any kind of task nowadays, regarding classification, prediction, detection, natural language processing, speech recognition, and many other fields [78]. They are inspired by the human brain and consist of interconnected neurons that generates a response (called output) to some kind of stimulus (known as input) [79]. Nowadays there exists many types of neural networks, some of the more common are the following:

Perceptron networks: the simplest type of NN, with an input and output layer composed of perceptrons [80]. Perceptrons assign a value of one or zero based on the activation threshold, dividing the set into two.

Layered networks (feed forward): multiple layers of interconnected neurons where the outputs of the previous layer neurons serve as the inputs for the next layer [81]. The neurons of each successive layer always have an input of one element more from the previous layer. Enables the classification of non-binary sets, and are used in image, text, and speech recognition.

Recurrent networks [82]: neural networks with feedback loops where the output signals feed back into the input neurons. Can generate sequences of phenomena and signals until the output stabilizes. Used for sentiment analysis and text generation.

Convolutional neural networks (CNNs) [83]: a type of neural network that is well-suited for processing data that has a grid-like topology, such as images. CNNs use a series of convolution layers to extract features from the input data, which are then fed into fully connected layers for classification or regression. CNNs are widely used in computer vision tasks such as image classification, object detection, and segmentation.

Transformer networks [84]: a type of neural network that has become popular in recent years for natural language processing tasks such as machine translation and text summarization. Transformer networks use a self-attention mechanism to learn long-range dependencies in the input data.

It is important to understand that in order to learn NNs need, aside from the data, a training algorithm. From the possible training algorithm, the most common used are training algorithms that are composed by two fundamental elements: the loss function and the optimizer. The loss function is used as a measure of how well the neural network's predictions match the actual data, and it is used to compute the error of the model, that is finally used by the optimizer to update the model's weights. The optimizer is an algorithm that updates the neural network's weights in order to minimize the loss function. So far many different optimizers have been realized, each with its own strengths and weaknesses. The most common are based on a gradient descent mechanism, like SGD [85] and Adam [86].

3.2. Deep Learning

Deep learning is a subset of machine learning based on artificial neural networks with multiple layers that enable progressive extraction of higher-level features (also known as “deep features”) from raw input data. This allows the network to understand and mimic more complex and abstract behaviours [81]. Given the greater architectural complexity afforded by deep learning models and their ability to tackle more challenging tasks, numerous novel deep learning-based methods have been developed in order to effectively address a wide array of complex real-world problems.

3.2.1. Deep Neural Network

Deep neural networks (DNNs) are a type of machine learning model comprised of multiple hidden layers. They process information from input to output, utilizing weights and backpropagation to minimize errors. Increasing the number of hidden layers enables DNNs to achieve better results, but also increases their computational and memory requirements [81]. In Figure 7 is illustrated a generic structure of a DNN.

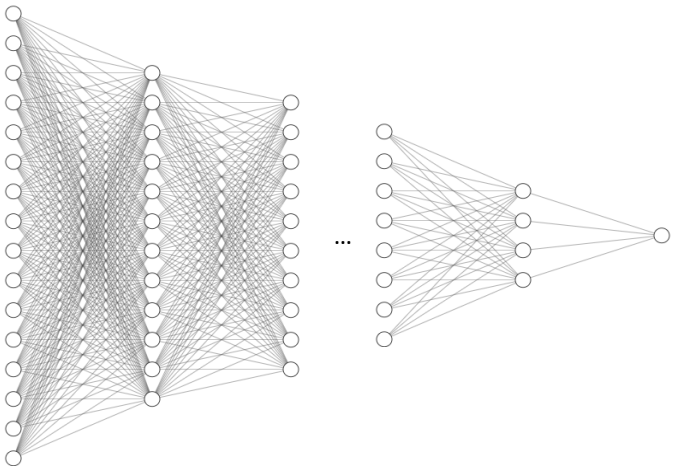


Figure 7 - Generic example of a deep neural network.

Deep Neural Networks (DNNs) have shown promising results in predicting drug dissolution times, outperforming Artificial Neural Networks (ANNs) in generalizing to new data. Their applications in healthcare are extensive, including medical imaging, diagnosis, drug development, prognosis, risk assessment, remote monitoring, and sports medicine [87, 88]. They have been used to analyze radiological images for detecting various conditions [89–91], evaluate endomyocardial biopsy data, and identify hypertension from ballistocardiogram signals [92, 93]. DNNs are also used in diagnostic models for diseases like cancer, diabetic retinopathy, and cardiovascular diseases [94–96].

3.2.2. Convolutional Neural Networks

Convolutional neural networks (CNNs) are comprised of specific block of layers. These blocks are composed of three main elements convolutional layers, the nonlinear activation function (e.g. rectified linear unit, ReLU), and the pooling layer. They detect visual patterns from raw image pixels using hidden layers. The convolutional layers analyze receptive fields with filters [97]. Nonlinear functions extract meaningful features about image features. Pooling reduces data, selecting features, and speeds up computation, enabling CNNs to identify similar features across an image for pattern analysis [98]. A simple CNN is shown in Figure 8.

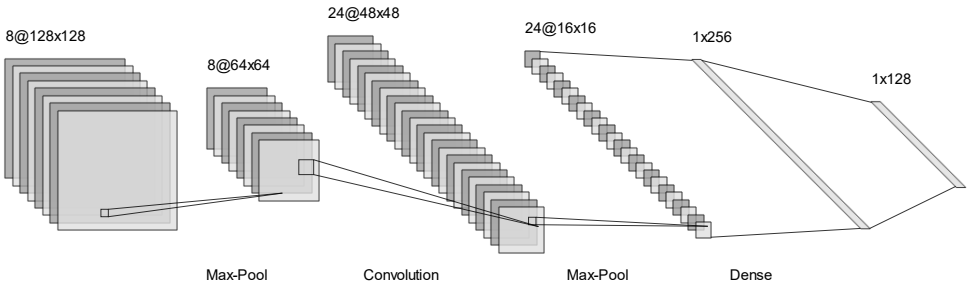


Figure 8 - Simple CNN example with a single convolutional layer.

Convolutional neural networks (CNNs) have demonstrated a remarkable capacity to interpret and analyze medical images across a variety of tasks. For example, CNNs can match or even exceed radiologist performance in identifying pulmonary nodules or measuring coronary artery calcium on CT scans [99]. They are also able to classify chest x-rays and detect various lesions with a high degree of accuracy [100, 101]. More broadly, CNNs have achieved excellent results in medical image classification, segmentation, reconstruction, and other areas.

3.2.3. Segmenting Neural Networks

Segmenting Neural Networks can be recognized by their characteristic U shape, like the U-Net [102] shown in Figure 9. The U-Net model is characterized by its symmetric encoder-decoder design and minimal connections. The encoder part of the model extracts deep features with large receptive fields through convolutional and downsampling layers. These features are then upscaled by the decoder to match the input resolution, allowing for pixel-level semantic prediction. The minimal connections primarily combine high-resolution features and different scales at the end, thereby reducing data loss from downsampling.

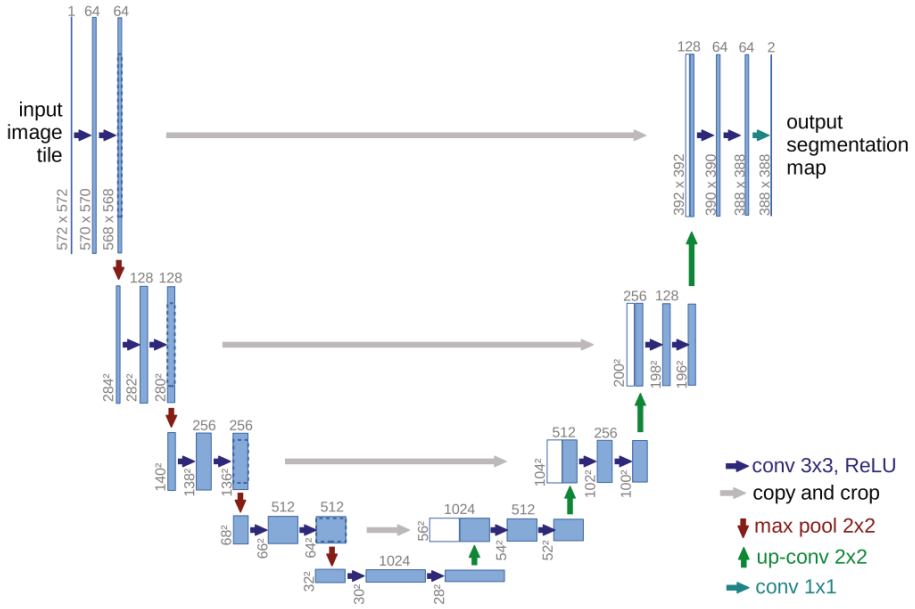


Figure 9 - Original U-Net model [102].

U-Net has been applied in a variety of contexts and in particular it offers an efficient backbone for segmenting medical images and generating semantic maps to aid diagnosis. Starting from these reasons the 3D U-Net version [103] has been implemented and it is currently used in all the main segmentation tasks that work directly with diagnostic images, like CT, MRI, US, histopathological images [104, 105].

3.2.4. Generative Adversarial Networks

Generative adversarial networks (GANs) [106] are structured in a very particular way, they are based on two competing networks: a generator and a discriminator. The first one learns to generate data that could be

considered realistic, while the second one tries to distinguish between the data real and the generate one, also known as the data fake. The trained mechanism is also particular, and it is done by having a competition between the generator and the discriminator. The generator tries to generate data that is realistic enough to fool the discriminator, while the discriminator tries to correctly identify real and fake data.

GANs have been successfully applied in medicine to generate various types of medical images, such as mammograms, CT scans, and MRIs. This allows for the training of models that require diverse image data. For example, GANs can be used to generate synthetic medical images that can be used to train models for disease detection and diagnosis [107].

3.2.5. Transfer Learning

Transfer Learning [108] is one of the most used machine learning techniques in deep learning. It consists of developing a model for one task and then reuse it as the starting point for a second task. This is useful because it can help save time and computational resources when training a model on a new task, especially if the new task is similar to the original task. Transfer learning is particularly important in deep learning because deep learning models can be very computationally expensive to train. By using transfer learning, we can leverage the knowledge that a deep learning model has already learned from one task to solve a new task more quickly and efficiently.

One of the most common ways to perform it, is to use a pre-trained model as a feature extractor. These features can then be used to train a new model for a different task. For example, we could use a pre-trained convolutional neural network (CNN) as a feature extractor to train a new model for image classification. The CNN would have already learned to extract features from images, such as edges and shapes. The new model could then use these features to learn to classify images into different categories.

Transfer learning has been used to achieve state-of-the-art results on a wide range of deep learning tasks, including image classification [109], object detection [110], natural language processing [111], and machine translation [112].

4. Artificial Intelligence-based tools for biomedical applications

At the moment of realization of an application based on artificial intelligence it is crucial to considerate the overall process. This means that all the steps that lead to the generation of the final result can be considered crucial, starting with the problem definition till the final result. This is strategic in order to avoid the arising of unexpected problems or misunderstandings of the real capabilities of the final tools.

In this work, it has been proposed, after a deep analysis and through a refining process carried out during the last three years, a framework based on three main phases for AI-based tools development, it is depicted in Figure 10. The phases are: the clinical phase, the artificial intelligence engineering phase, and the final application development. A complete definition of these three phases is generously explained below in a general meaning, with the aim of giving all the possible steps that could be performed to reach the final goal.

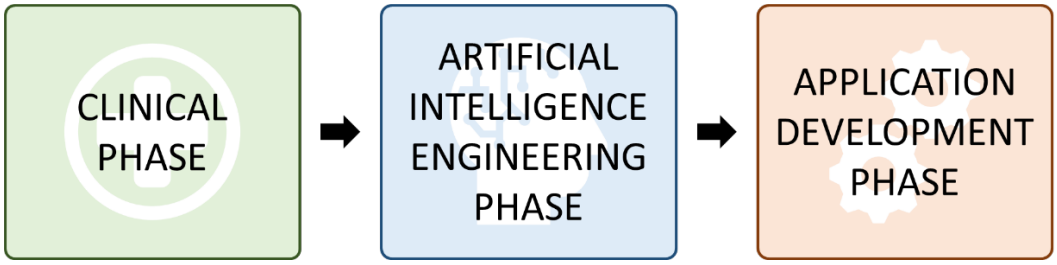


Figure 10 - General framework representation: the three main phases 1) clinical phase, in which is majorly involved in the clinical staff; 2) artificial intelligence engineering phase, where the AI model is selected through deep analysis of the literature and after several testing steps; 3) application development phase, that consists of developing the final application integrating the AI model that will be used by the physicians.

It must be noted that this work has not the aim to realize all the steps necessary to satisfy the requirements of the legislative regulations and the ethical ones given the evolving situation present these years for this specific sector.

4.1. Clinical phase

The clinical phase consists of all the steps that need to be performed during the cooperation of an interdisciplinary team comprising physicians, engineers and other possible expert correlated figures. These steps are the fundamental aspect to obtain a final result that could be considered valid.

The initial step in this phase involves identifying a clinical task that could benefit from the implementation of an AI-based tool. This need typically arises from an unresolved issue or a repetitive task that physicians have identified for enhancement. In general, medical professionals can pinpoint crucial aspects of their clinical work that could be improved with AI, aiming for better performance across all tasks.

Upon identifying the potential task, several important meetings are held to deeply understand the requirements and possibilities. This phase involves clarifying several elements: the specific task that AI could potentially handle; the pathology involved; the individuals affected by this tool's introduction; and the data available for training the AI.

It is very important to correctly identify all these elements and are critical to the next step: the database creation process. Before describing this step, it is important to understand that, when dealing with any kind of medical data that could potentially reveal some type of information of some patients, it is mandatory to get the patient consent. Furthermore, when doing a campaign of medical data harvesting, diagnostical or clinical information of the patient, it is necessary to follow all the regulations for handling any kind of confidential information of the country where the data is acquired. Therefore, gathering new medical data is a very complicated process, not only to obtain many cases, but principally doubt to all the bureaucracy necessary.

There could be many ways to perform the creation of the database and they could be all performed in parallel considering the many difficulties that usually occur for the medical tasks. All these possibilities could be grouped in three main categories: i) gathering new data, performing a campaign of data gathering, in which the hospital involved, therefore the physicians of that hospital, do all the possible steps necessary to be able to gather the information of the new patients from the moment the campaign start; ii) gathering data from publicly available datasets, there exists many publicly available datasets of diagnostic data that contain many kind of data; iii) request data from private study centers, try to gather data directly from other study or hospital centers. Considering the main goal of acquiring as much data as possible, all these ways have some positive and negative aspects, but there is not one that can be considered as the best one, notwithstanding further details will be discussed. The idea of creating a new data gathering campaign that involves the medical staff's hospital center could be considered as the best one, mainly because all the data gathered would be exactly as the team have decided, but this is not simple because, as said before, the time necessary to obtain the concessions is undefined and could last more than the project time itself and obviously at this it is necessary to add the time necessary to effectively perform the new data acquisition process. Considering this aspects, the other two options seem more easier to be done, but they are also very optimistic like in the request of private data in the majority of the cases drives to nothing because the centers ultimately or does not answer or cannot share the confidential data. Instead for what concern the research of publicly available dataset, it does not exist the certainty to find any dataset for the specific task planned, and if it does exist, it is not guaranteed that it contains the data previously decided as the data to be used.

Given all these factors, the database creation process for developing AI-based solutions for clinical applications is the most important and difficult step, especially being time-consuming without the possibility of avoiding this step considering the importance of the typology of the final data that would be effectively gathered.

4.2. Artificial intelligence engineering phase

Considering completed all the above steps, this phase is carried on mainly by the engineering team with the goal of developing the artificial intelligence tool that could satisfy as much as possible the requirements decided by the multidisciplinary team during the clinical phase. Given this, the steps during this phase will have a technical character.

The first step is the database cleaning. During this step the database gathered will be cleaned and prepared to be effectively useful at practical level. It is fundamental to perform this phase with a painstaking precision because any error will be reflected to the final results leading to possible false conclusions and requiring the restarting of the overall cycle of development. In this phase the following steps are performed: *correction*, the dataset created is checked in order to identify potential anomalies or unacceptable values, *completion*, then it is controlled to find possible values that does not exist and to appropriately handle all these situations deciding whether to remove the cases found or to change them in order to fit correctly the dataset structure. Following these two steps, the creation phase begins. This involves the use of feature engineering to develop new features. Lastly the *conversion* process is performed in order to obtain a domain of features that is all coherent, limited and well structured, e.g., all the non-numerical data is converted to numerical values, the categorical data is converted to a list of new binary features, the domain of each feature is scaled to be inside the interval $[0,1]$ or $[-1,1]$.

After the database cleaning the second step is typically based on the exploratory analysis of all the features using statistical methodologies. This is done in order to get a deeper grade of understanding of the information available and to decide which approach could be used in the next step of this phase. Usually during this step, it should be possible to identify possible correlations between the available features in order to obtain the desired outcomes, and in this way discard useless data. This is mainly useful for the cases in

which there a lot of features available and there is uncertainty on which could be the most suitable to the problem addressed.

Next for the third step there is the one that involves the creation of the model data or models data that will be used finally. Based on the final typology of task identified that must be done, classification, prediction, or regression, several different machine learning models, described in the previous chapter, could be chosen. For this reason, the first process of this step consists of reducing the number of possible models to use to only the ones that are more promising, based on the main usage, the type of data available, and the state-of-the-art on that specific task. Once the set of possible models is created, the next step is the training of the model followed by its validation. In order to perform these steps in the best possible way, there exist some well-known practices that should be followed. These practices consist of splitting the available data in two main different subsets, the set for the training process and the set for the testing one, named correspondingly training set and testing set. The training set is then divided into two subset the effectively used for the model training and the one used to validate it during the training, after each epoch is done. The so-called validation set is typically created using some specific split algorithm, like k-fold cross-validation and leave-one-out, done to obtain several different splits of training set and validation set. All these splits are used to train different instances of the same and to validate all of them in different sets of the same data, in order to obtain statistically more accurate information about the effectiveness of the data and of the model selected. Once all the possible subsets are created, before the training of the model, it is not only mandatory but also strategic to decide the final metrics that will be used to evaluate the model. This activity is fundamental in obtaining the right result when training the model, because it will be the goal of the training process to minimize or maximize the metrics selected. The metric that must be used depend on the task that must be performed, for example in the classification task some metrics could be the accuracy, precision, recall, f-score, etc., but in regression problems other metrics are typically used like mean square error, mean absolute error, r-square, etc., and also for the detection ones some other like mean average precision, intersection over union, false positive rate, etc. Thereby once the data and the metrics are ready the training is done with in parallel the validation. To obtain the best result the hyperparameters of the optimization and loss functions must be tuned. Once the validation metrics are obtained the next step can start. The best performing model based on the validation results is selected, one per each model typology, and tested on the testing set to identify the testing metrics. The real model capability is measured based on the performance resulting over the testing data. This step is crucial because in case the final values are not sufficiently to satisfy the expected ability of the model, it is necessary to go back to the clinical phase to decide whether to change some aspects of the current task, maybe simplifying it, or to perform a new data acquisition step in order to have a greater dataset that could be more representative for the given problem.

4.3. Application development phase

This last phase consists of developing the application that will be actually used for the task required. Reaching this last phase is not so straightforward, and it will not be seen in all the cases that will be presented in the next chapters. This is due to different factors, from the metrics not being sufficiently effective in the medical practice, or because there is not enough data to get a statistically stronger result, being in this way some preliminary study that will require more efforts for the data acquisition part.

The first step of this phase consists mainly of integrating the AI model that has been elected as the best one with the code of an application. Secondly the user interface is designed to facilitate the usage of the application for the users. Usually this step have a prior step in which there are some meetings in order to choose the best user interface possible based on the user needs and the feasibility of the requests. Subsequently the test and debugging step is performed to be sure that the final product will not fail in future scenarios. Finally the deployment step is done. This last step consists in releasing the final application for a specific platform, that could be a web server, the app store for mobile applications or the desktop environment. Also this step is chosen based on the customer needs and the application requirements.

4.4. Artificial intelligence pipeline definition

With the idea of summarizing and facilitate the usage of all the explained phases and steps, all the above can be reduced inside a framework. In this way each time an application based on artificial intelligence must be developed for a biomedical application, it is possible to use the proposed framework with the goals of facilitating the creation of the application and reaching possible state-of-the-art performance with the AI model following the best practices existing in this field.

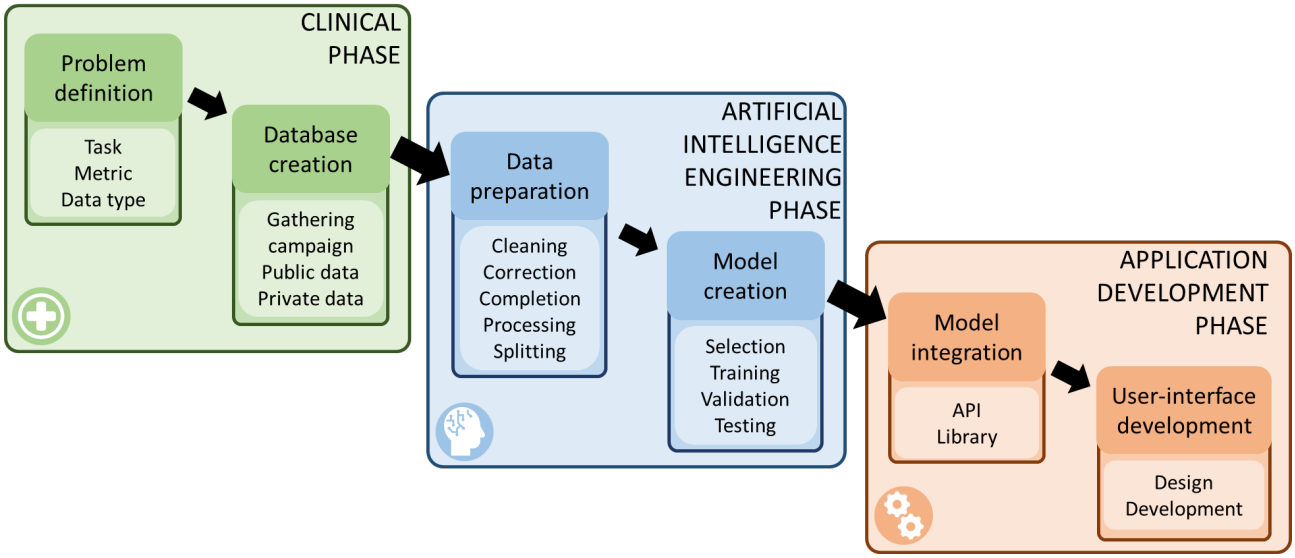


Figure 11 – Detailed view of the proposed framework, all the phases are divided into two main sub-phases with their own set of specific steps that must be done in order to go to the next phase.

The complete workflow, depicted in Figure 11, is composed of these steps: the *problem definition* is the phase in which the clinical problem is defined by the cooperation between the engineering team and the medical staff, identifying the main goal of the application that would be created. Followed by these three main aspects: *task selection*, with this step the effectively task type is defined and so the main objective of the final AI model that would be selected. *Metrics selection* is the next step and it is crucial to understand the actual performances of the models used on the task selected. Finally *data selection*, in this step the typology of the data that will be used as the input of the model is chosen based on the expertise of the physicians on the topic addressed.

For the *database creation* phase the principal aspect consist in the effective creation of the database that will be used to train, validate, and test of the machine learning model. This step can be carried out by the means of three different procedures: the medical staff after getting all the necessary permits start a *data gathering campaign*, in which the data of new patients are acquired. Other than this, the other two ways are the *public data gathering* and the *private data gathering*. The final result of both these two possibilities is the same, in which an already existing dataset is obtained. The real difference between the two is that the private data gathering usually is not easily doable, and almost always ends in nothing.

After creating the database, the next important step is the *data cleaning* phase. During this phase all the data is prepared in order to be sure that it is really usable and does not present any kind of problem in the next phases. Several steps are done for this, starting with the *data correction and completion*. The data is checked to avoid possible inconsistencies in the values contained and if there are any missing value, in both cases a specific strategy is performed to replace the value or remove the sample or remove the feature. After this the *data preprocessing* is realized to move all the features to the same limited domain of values. Finally, the *data splitting* is done, in such a way that three different subsets are created, training, validation and testing.

Once the dataset is created and all the necessary are finished the following phase is the one that consists of the *AI model creation*. In this phase after an initial *model selection* step, in which a list of possible models is created to be sure to identify the best model to use for the selected task. All the selected models are the used in the *model training and validation* step. As the name of this step suggests, all the models are training on the training set and validated over the validation set. The best model elected after the results on the validation set is used during the *model testing*. This last step is crucial to identify the final model performance on unseen data, from this the life of the model is decided based on the possible criteria imposed by the interdisciplinary equipe. If the final results are not sufficiently satisfactory, usually the cycle of the life of this application goes back to the database creation process, and if it is too difficult to be done, a change in the final task could be performed by the interdisciplinary team.

Eventually, after all these steps are completed in a satisfactory way the final phase consists in the *application development*. During this last phase the model trained is integrated – *model integration* - into the final application, with the goal of being really usable in practice. This is obviously followed by the *user-interface development*, necessary in other to create a tool usable by everybody, even the ones without any kind of expertise in informatics. Finally after all this is phases and steps the *application deployment* must be realize. In this latest step the application is released for the specific final platform in which the code should be used.

Therefore is to be expected that this framework can be applied generally, with almost all these steps given the sufficient amount of time, to all the problems in which an artificial intelligence tool will be developed as the solution. In the following chapters, the framework will be applied separately to all the cases considered. In particular the following chapters will be about the following case studies: urology, with a focus on kidney tumors classification; plastic surgery, with the goal of automating the procedure of creating surgical guides for the surgery; psychiatry, in which the aim is to distinguish between fake and real suicide attempt; neurology, for which the objective is to obtain a tool to identify brain tumor cells to compute the volume of the tumor with to the volume of the brain.

5. Artificial intelligence for urology – case study kidney tumor

The case study on kidney tumors was conducted inside the custom3D conjunct laboratory at the Department of Industrial Engineering of Florence (DIEF). The first case study during the overall time of the PhD involved a deep study of the literature on kidney AI application, which has been performed and reported here to understand what the state-of-the-art was and if the thought framework was good enough or it needed some enhancements. The task analyzed concerning the classification of kidney cancer is reported. For this, the framework proposed and explained in the previous section will be used, and therefore the structure of these chapters will be: explanation of the clinical scenario, followed by the definition of the task to be performed and the data type to be used, with the description of the data gathered and available; after this steps, the proposed model is illustrated with the usage strategy, training methodology, and obtained results. Finally, some discussion is made for both and for the overall clinical case study.

5.1. State-of-the-art kidney AI-based applications

Kidney diseases, such as renal tumors, acute kidney injury (AKI), and chronic kidney disease (CKD), are important issues for nephrology and public health worldwide, as they are associated with high mortality and morbidity rates [113, 114]. These diseases, if not identified and treated preventively, can degenerate and lead to severe renal dysfunction, comorbidities, and, in the worst case, death [115–117]. Currently, in order to detect and prevent the degeneration of kidney disease, continuous monitoring of specific parameters obtained through diagnostic tests is performed [118]. Given that statistical models are used to deter-mine the actual presence or absence of disease [119], its severity [120], or its degeneration [121], it is natural to think that models based on artificial intelligence (AI) and machine learning (ML) [122] could also be used to achieve this same goal, to obtain statistically better results or more high-performing solutions.

As said in chapter two, in the last decade ML techniques have been increasingly employed in a variety of research areas. In this chapter, there will be examined with deeper detail the usage of ML in urology, in particular there will be examined the applications that involve the kidney.

In nephrology, ML techniques are used for several purposes:

- segmentation and identification of the anatomy of interest within the diagnostic images (e.g., kidney masses such as tumors, cysts, etc.);
- classification of a kidney mass type, or of the stage in which a specific tumor is found;
- prediction of the evolution of kidney functionality, which can highlight the presence of pathologies.

Among others, ML techniques can be used in the analysis of suspicious renal masses. In such cases, it is nowadays necessary to surgically remove the tumor to identify if it is of a malignant or benign nature, but, due to its position, surgical removal is impossible without risking permanently compromising the patient's urological function. For this rea-son, by working directly with diagnostic data and images, machine learning techniques can be crucial alternative solutions for segmenting and identifying masses.

Furthermore, some techniques can be used to help physicians to distinguish between particular cases of some pathologies that are very difficult to distinguish. In these cases, features obtained from diagnostic exams are used to classify the single cases; in this way, the physicians can reach a more precise diagnosis.

In addition to these applications, there are also techniques realized to prescribe specific therapies, or to detect a pathology in advance, in order to prevent it or any of the possible degenerative side effects (e.g., chronic kidney disease, acute kidney injury). In these applications are included also tasks with the aim to predict the compatibility and the outcome of a surgical operation, such as a kidney transplant.

Recently, the number of works related to this area has dramatically increased, rising from a few dozen papers before 2018, to a few hundred presently (based on papers indexed on the Scopus® database from Elsevier). For this reason, it is crucial to carry out an updated survey summarizing the most promising opportunities

offered by ML in this area. Accordingly, the present work aims to propose an updated and schematic survey of the most effective existing techniques and to draft possible future research lines based on ML.

First, the most promising articles are selected from the overall literature and classified based on their different applications. Then there is a description and a comparison of all the used datasets relative to the works selected. After that the implemented methods and the possible future developments are analyzed. Finally, conclusions are drafted based on the literature.

The contribution that it is intended to make with this work is to give a macroscopic view of the existing works concerning nephrology. In particular, the aim is to understand the state of the art of the methods that employ ML techniques to deal with some of the most common kidney diseases, reporting the various resulting metrics for each method. In addition, dimensional analysis of the various types of existing datasets that have been used so far is carried out and a generic comparison is made from the point of view of the type of data.

5.1.1. Article selection

A study of the literature related to publications spanning from 1992 to February 2022 was carried out using Elsevier's abstract and citation database, Scopus®, by entering keywords, "Artificial Intelligence", "Machine Learning", "Kidney", to identify the most common and effective artificial intelligence (AI) and ML techniques that directly involve the kidney. In particular, the entered query was as follows:

$$\begin{aligned} &\text{TITLE-ABS(artificial OR intelligence OR machine OR learning OR kidney)} \\ &\text{AND} \\ &\text{KEY(artificial AND intelligence AND machine AND learning AND kidney)} \end{aligned} \tag{1}$$

The research thus performed allowed the identification of papers that use AI and ML in kidney analysis contexts. Figure 12 shows a significant increase in recent years (after 2017) in the interest and production of papers by the scientific community—in general, there was an overall number of 224 papers dealing with the selected topic.

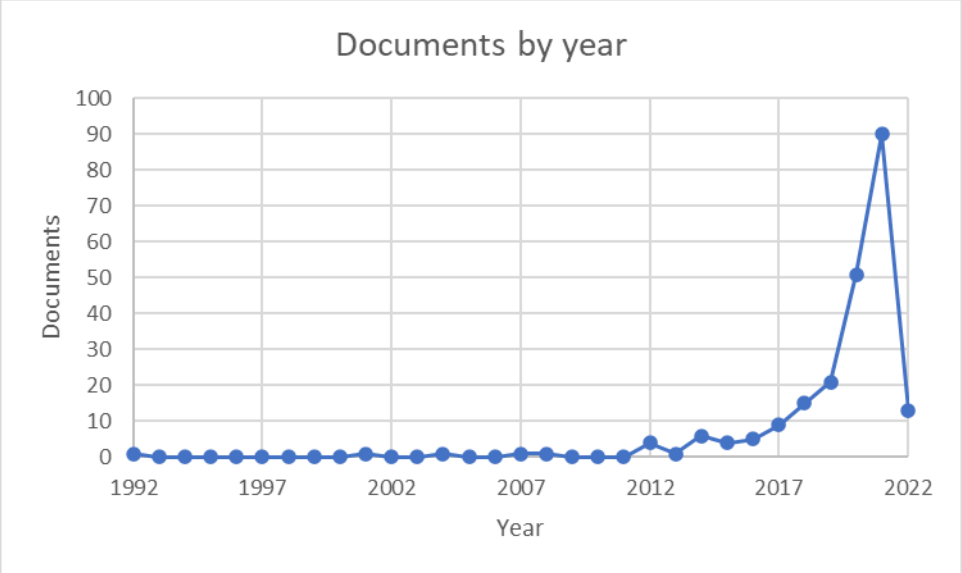


Figure 12 – Trend of documents per year.

To focus on the most relevant works, the literature analysis was carried out according to the following inclusion and exclusion criteria.

Inclusion criteria: (1) articles dealing with ML and AI techniques applied to the kidney were considered; (2) original articles concerning one or more of the following aspects were taken into consideration - segmentation, classification, and prediction of diseases directly related to the kidney; (3) reviews related to these topics were studied to perform a final check of the selected articles.

Exclusion criteria: (1) editorials, commentaries, and abstracts were not included in this study; (2) studies related to animals or carried out only at a laboratory level were excluded; (3) research studies that were not applied in clinical practice were not considered.

According to the aforementioned procedure, fifty-nine studies were found to be eligible to be part of this survey.

5.1.2. Machine Learning approaches for nephrology

In the following, the studies are grouped based on the nature of the kidney disease. In detail, the analyzed pathologies are “kidney masses”, “acute kidney injury”, “chronic kidney disease”, “kidney stone”, “glomerular disease”, “kidney transplant”, and “other kidney pathologies”. From the analysis of the selected articles, three main research tasks are identified across the application areas:

1. segmentation and identification, which intends to analyze diagnostic images with the purpose of highlighting or detecting one or more specific elements;
2. classification, which aims to perform a diagnosis or to determine the degree of severity of disease;
3. prediction, which aims to prevent or forecast some future event, e.g., predict either the degeneration of a disease or the outcome of a specific therapy.

In the next subsections are reported, for each disease, a brief description of the symptoms to provide the reader with a simple explanation of the clinical scenario, and the various ML techniques used in the state of the art, grouped according to the research tasks described above, highlighting the type of database used. Figure 13 shows a graph schematically outlining the several analyzed pathologies (red color). From each pathology, one or two branches may be amplified according to the type of data available in the available studies (green color), and finally from these as many branches as the ML methods used on that type of data for that specific renal pathology (blue color). The following sections are based on the schematization depicted in the graph.

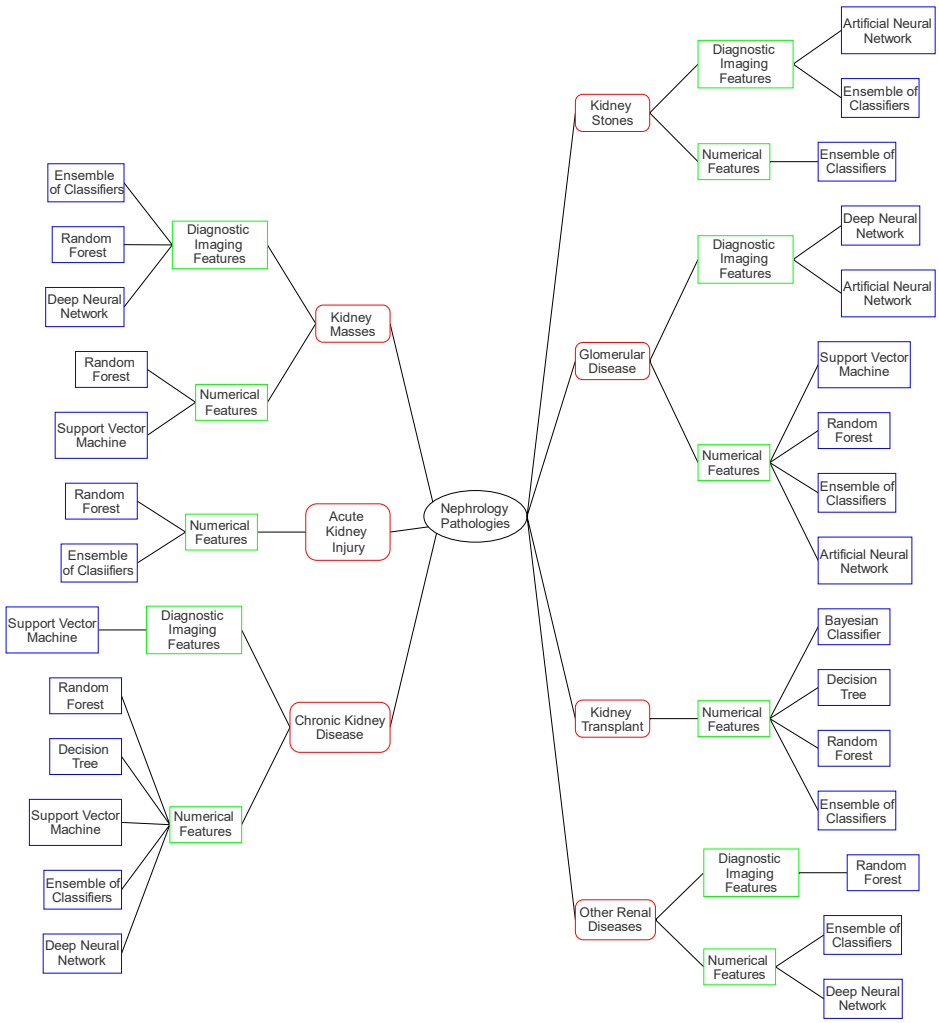


Figure 13 - Scheme of pathologies with ML techniques applied according to the type of dataset available. Kidney disease addressed in red; type of available data in green; ML technique used in blue.

5.1.2.1. Kidney Masses

Kidney masses are abnormal growths within the kidney. They are mainly subdivided into two main categories: solid and cystic. Generally, the presence of a kidney mass, in vivo, is determined by relying on imaging techniques such as CT, MRI, or US.

In general, cystic kidney masses are, in most cases, benign [123], while solid kidney masses are generally malignant; therefore, the kidney is generally partially or totally removed to perform the histological exam. However, approximately 16% of surgically removed solid kidney masses are benign [124] and surgical removal would not have been necessary. Unfortunately, the distinction of the nature of the solid renal mass, using diagnostic imaging, is very complex, even for specialized physicians, given the significant similarities in the appearance of some types of malignant and benign renal masses, in terms of texture, size, volume, and position. To face this challenge, modern ML techniques have been employed to process image data, proving to help physicians in making a more precise and accurate diagnosis. To classify and distinguish between malignant and benign masses, [125] some use a Bayesian classifier [126], a learning algorithm based on the statistical relationship between radiomics features (relational functional gradient boosting), and [127] an algorithm based on CT texture analysis. Many works focus on the analysis of renal cell carcinoma (RCC), which is the cause of 80% of kidney cancer deaths [123], either to distinguish different types of RCCs or to differentiate them from benign tumors. In [128–132], the main goal is to diagnose the most common malignant tumor, the clear cell RCC, using radiomic features and ML-based classifiers (e.g., random forest, CatBoost). Using radiomic features extracted from multiphoton microscopy images of kidney tissue sections,

[133] try to distinguish RCC chromophobes and oncocytomas, while [134] try to classify the stage of a particular type of malignant tumor, the papillary RCC, using microarray datasets [135] and clinical information of the patients. Some more recent research, such as that of [136–138], focuses not only on tumor classification, but also on automatic tumor identification through diagnostic images, by using three-dimensional image processing with ML techniques such as 3D U-Net, and 3D V-Net; with these solutions, they are able to automatically segment the tumor inside the CT. Table 1 shows these works, explaining the main objective of each one, the adopted ML techniques, the database exploited, the best result achieved, and finally the year of publication; the reported metrics should be read from the perspective that the higher the reported value, the better the obtained performance.

Table 1 – Renal mass research.

Paper	Objective	Method	Database	Results	Year
[125]	Malignant renal cyst prediction	Bayesian classifier	[125]	AUC 0.96	2009
[126]	Identify malignant renal masses	Statistical relational learning—RFGB: relational functional gradient boosting	[126]	Accuracy 82%	2018
[127]	Differentiate between malignant and benign masses	CT texture analysis with random forest	[127]	Accuracy 90.5% AUC 0.915	2020
[128]	Diagnose ccRCC	WEKA with and without SMOTE	[139]	AUC contour-focused 0.865–0.984 AUC margin shrinkage 0.745–0.887	2019
[129]	Diagnose ccRCC	Pyradiomics and random forest	[139]	Accuracy 84.6% Sensitivity 90.4% Specificity 78.8% Precision 81%	2020
[130]	Diagnose ccRCC	Radiomics and CatBoost	[130, 139]	MR accuracy 73% internal 74% external CT accuracy 79% internal 69% external	2020
[131]	Diagnose ccRCC	MaZda and WEKA toolkit	[131]	Accuracy 85.1%	2018
[132]	Diagnose ccRCC	Proteomics-based random forest and imaging-based VGG16	[139]	Proteomics accuracy 98% image accuracy 83% validation, 95% testing set	2019
[133]	Differentiate between kidney chromophobe renal cell carcinoma and oncocytoma	Linear SVM	[133]	Accuracy 80%	2016
[134]	Classify papillary renal cell carcinoma stages	Feature extraction and random forest	[140, 141]	Accuracy 88.5%	2018
[136]	Kidney and tumor segmentation	3D U-Net	[142]	Mean Kidney Tumor Dice 0.9168	2019
[137]	Kidney and tumor segmentation	Cascade 3D U-Net	[142]	Mean Kidney Tumor Dice 0.9064	2019
[138]	Kidney and tumor segmentation	Multi-resolution 3D V-Net	[142]	Mean Kidney Tumor Dice 0.8815	2019

5.1.2.2. Acute Kidney Injury

During an episode of acute kidney injury (AKI), the kidneys show difficulty in maintaining the proper fluid balance in the body, due to an accumulation of waste products. Given the speed with which it strikes and the damage that it causes, being able to detect it early can be of great significance. In this type of critical situation, AI is demonstrated to be one of the best solutions to correctly identify a patient with AKI. In studies by [143–145], the goal is to predict AKI based on early symptoms to prevent a possible degeneration of the disease, analyzing electronic health records (HER) and other clinical data, such as laboratory tests, vital signs, and patient demographics. AI techniques, thanks also to the speed of response, can be decisive, as in the case of [146], in which the authors try to detect AKI in burn patients using a k-nearest neighbor classifier on numerical features obtained from plasma creatinine testing [147].

Some research, such as [148, 149], focuses on predicting an episode of AKI in patients undergoing examinations that require contrast agents, specifically coronary angiography. It has been observed that the use of such agents can lead to AKI episodes; in these studies, the authors aim to predict the AKI episode with AI approaches by using clinical variables collected before the examination and by the results of the coronary angiography that they undergo [150].

Recent studies focus on predicting AKI episodes' insurgence within different periods from its manifestation. The most common prediction time intervals vary from 48 h to a maximum of 90 days, as in [151]; in this work, the authors evaluate their solution based on the analysis of time-series data over these time intervals. It is possible to find another ex-ample in [152], in which the authors, through numerical features extracted from multiple blood tests per single patient, attempt to predict AKI within 30 days from its manifestation. Finally, in [153], the authors, using daily collected patients' clinical data, propose a particular type of deep learning algorithm, based on time series, which is able to predict AKI within 48 h from its occurrence, as well as classify the stage of the AKI disease if it is already present. In Table 2, analogously to Table 1, are reported all the related objectives, methods, used databases, and results.

Table 2 - AKI research.

Paper	Objective	Method	Database	Results	Year
[143]	Predict AKI in adult and children	Boruta [56] (selection algorithm) + random forest	[154, 155]	AUC 0.796	2018
[144]	Predict AKI in adult and children	Gradient boosted machine	[156]	AUC 0.85	2021
[145]	Predict AKI	Gradient boosted machine	[145]	AUC 0.76	2021
[146]	Predict AKI in burn patients	K-NN	[146]	Accuracy 97%	2019
[148]	Predict AKI	Lasso + logistic regression	[157]	AUC 0.79 AUC 0.82 [$p < 0.001$]	2019
[149]	Predict AKI	RF + XGboost	[149]	AUC 0.843	2020
[151]	Predict AKI	Streams	[151]	Accuracy 56% in 48 h Accuracy 84% in 30 d Accuracy 90% in 90 d	2019
[152]	Prediction of AKI from blood test	Feature selection + random forest	[152]	AUC 0.881 in 30 d	2021
[153]	Predict AKI	Gradient boosting tree-based machines	[158]	AUC 76% in 48 h AUC 81% stage 2 AUC 87% stage 3	2020

5.1.2.3. Chronic Kidney Disease

Chronic kidney disease (CKD) is a condition characterized by the gradual loss of kidney function over time. CDK damages the kidneys by decreasing their ability to filter waste from the blood. In severe conditions, waste can reach high levels and lead to the development of other complications, which, in the most extreme cases, will require periodic medical treatment, such as dialysis, or even a kidney transplant [159]. CDK is a disease that can be diagnosed by physicians through the study and analysis of a variety of indices (e.g., eGFR [160]); thus, it is suitable for the application of ML methods. An example of using AI for this purpose can be seen in the study by [124], where the stage of pathology is classified using radiomic features obtained from ultrasound images of the kidney.

The general interest and applications to diagnose CKD underwent an abrupt increase with the creation and public release in 2015 of a database containing characteristic features (i.e., age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia) related to 400 patients during the early symptoms of the disease [63]. Different methods based on the analysis and classification of patient features are adopted by [124, 161–169].

In addition to the diagnosis of CKD, there are some related studies in the literature, such as [170], in which the authors try to predict a possible plan for the patients' diet, given the fact that following a proper and suitable diet plan can help to slow down the progress of CKD [171]. In [172], since maintaining appropriate hemoglobin levels during treatment for CKD is critical, the authors try to predict the hemoglobin level in the blood during anemia treatment in predialysis CKD patients, to intervene more quickly.

This information, the used databases, and the obtained accuracy results are shown in Table 3, analogously to the others.

Table 3 – CKD research.

Paper	Objective	Method	Database	Results	Year
[124]	Diagnose CKD based on patient stage	Support vector machine—SVM	[124]	Accuracy 82% on 2 stages Accuracy 67.21% on 3 stages Accuracy 51% on 5 stages	2014
[161]	CKD diagnosis	Random forest	[63]	Accuracy 99.3%	2016
[162]	CKD diagnosis	Decision tree C4.5	[63]	Accuracy 63%	2016
[163]	CKD diagnosis	SVM	[63]	Accuracy 98.3%	2016
[164]	CKD diagnosis	k-NN with CFS and AdaBoost	[63]	Accuracy 98.1%	2017
[165]	CKD diagnosis	Random forest	[63]	Accuracy 100%	2017
[166]	CKD diagnosis	RPART	[63]	AUC 0.995 Sensitivity 0.9897 Specificity 1	2018
[167]	CKD diagnosis	PSODP + DL-RNN	[63]	Accuracy 99.5%	2018
[168]	CKD diagnosis	PNN [77]	[63]	Accuracy 96.7%	2019
[169]	CKD diagnosis	RFE and Random Forest	[63]	F1 score 100%	2021
[170]	Predict diet plan for CKD patients	Multiclass Decision forest	[63]	Accuracy 99.17%	2017
[172]	Predict hemoglobin levels in CKD patients	Extraction rule—Re-RX + J48graft	[63]	Accuracy 95.18%	2019

5.1.2.4. Kidney Stone

Nephrolithiasis, or kidney stones, is a condition characterized by the presence of deposits in the kidney, caused by an alteration in the balance between the solubility and precipitation of salts in the urinary tract and kidneys [173]. One crucial point is given by the fact that surgery is required in 20% of patients with this condition [174]. In this context, AI is applied to identify the correct type of treatment to be followed based on parameters such as sediment composition, location, and size [175]. Some research focuses on the detection of kidney stones, such as [176, 177], which use radiomic features extracted from manually segmented CT, with the goal of the early detection of stone deposits before they reach a size greater than 2 cm, allowing the use of non-invasive treatments. Other research, such as [178–180], focuses on predicting the outcome of shock wave treatment without the use of diagnostic imaging techniques, by analyzing the preoperative parameters of patients (such as age, sex, presence of related diseases, and stone characteristics including stone laterality, location, and maximum length). Similar to the other tables, Table 4 reports this information, the databases used, and the accuracy of the obtained results.

Table 4 - Kidney stone research.

Paper	Objective	Method	Database	Results	Year
[176]	Renal stone detection	Segmentation + ANN	[176]	Accuracy 86%	2019
[177]	Renal stones vs. phleboliths	Radiomics + AdaBoost classifier	[177]	Accuracy 85.1%	2019
[178]	Kidney stone removal, prediction of postoperative variables	ANN	[178]	Accuracy 81–98.2%	2017
[179]	Predict stone-free status after the first treatment	Feature extraction + sequential forward selection + multiple classifier scheme	[179]	Accuracy 60%	2019
[180]	Stone-free prediction	Light gradient boosting method	[180]	Accuracy 87.9%	2020

5.1.2.5. Glomerular Diseases

Glomerular diseases are diseases that affect the glomeruli, whose function is to filter blood and, at the same time, to retain proteins and blood that the body needs. Many diseases, such as diabetes, affect kidney

function by attacking the glomeruli [181]. In this regard [182–184], use methods based on the analysis of patients' clinical data to predict type II diabetes. Some studies focus on specific conditions and causes of glomerular diseases, such as Immunoglobulin A Nephropathy (IgAN), which is the most common biopsy-proven primary glomerulonephritis in the world [185]; it damages not only the kidneys, but also the immune system response [186]. In [187–189], the authors implement applications able to predict IgAN using a renal immunofluorescent image obtained by fluorescence microscopes relative to a renal biopsy. Other works, such as [190–192], focus on detecting type II diabetes directly from diagnostic images, using radiomic features. Finally, [193] try to predict the weight of children with glomerular disease to avoid possibly dangerous weight loss, using diagnostic numerical features obtained from blood monitoring and analysis.

All the useful information is reported in Table 5, analogously to the previous tables.

Table 5 – Glomerular disease research

Paper	Objective	Method	Database	Results	Year
[182]	Predict diabetic kidney disease	SVM radial	[182]	Accuracy 94%	2013
[183]	Predict diabetic kidney disease	Unbalanced random forest	[183]	Accuracy 83.8%	2018
[184]	Predict diabetic kidney disease	Knime + WEKA	[184]	Accuracy 83.5%	2019
[187]	Resolution image-based renal pathology	Convolutional neural network	[187]	Accuracy > 80%	2021
[188]	Predict ESKD in patients with IgAN	ANN	[188]	AUC 0.82 with 5-year follow-up AUC 0.89 with 10-year follow-up	2021
[189]	Predict deterioration of kidney function in IgAN patients	SVM	[189]	Accuracy 79.8%	2021
[190]	Diagnose glomerular disease	Disjunctive least generalization—DLG algorithm	[190]	Accuracy 81.26–96.5%	1992
[191]	Detect pathogenic and non-pathogenic glomerulus and tubulus	RatSnake—ML automatic segmentation	[191]	Accuracy 94.7%	2014
[192]	Diagnose glomerular disease	Decision tree with J48 algorithm	[192]	Accuracy 89.47%	2021
[193]	Predict weight of children in renal dialysis	ANN	[193]	Mean difference 0.497	2018

5.1.2.6. Kidney Transplant

Even if kidney transplantation is not a pathology but rather a specific surgical treatment, some authors considered creating a dedicated section since there are several studies regarding this topic, and it is one of the most common treatments for patients with severe kidney pathologies.

In detail, kidney transplantation is a surgical procedure that involves taking a healthy kidney from a living or cadaveric donor and implanting it into the recipient patient. For the transplant to be successful, many factors must be considered, including the compatibility of the donor with the human leukocyte antigen (HLA) proteins of the recipient. Although, nowadays, there is a method that reduces the risk of rejection, in the case of mismatched HLA [194, 195], approximately 40% of donated kidneys are rejected [196]. The ML techniques ap-plied by [197–201] focus on predicting the probability of success and survival in these types of interventions using numerical features (e.g., age, sex, time in dialysis, donor type, donor age, HLA mismatches, delayed graft function, acute rejection episode, and chronic allograft nephropathy). Table 6 reports all the necessary information, analogously to the others.

Table 6 – Kidney transplant research

Paper	Objective	Method	Database	Results	Year
[197]	Predict transplant failure probability	Decision tree	[197]	Specificity 73.8% Sensitivity 88.2%	2010
[198]	Predict post-transplant survivability	Bayesian belief network	[198]	Accuracy 52% after 1 year Accuracy 56% after 3 years	2012
[199]	Classify risk levels for kidney graft survival after transplant	ElasticNet + Bayesian belief network	[202]	Accuracy 68.4%	2018
[200]	Predict early transplant rejection	Decision tree and random forest	[200]	Accuracy 85%	2019
[201]	Predict kidney transplantation compatibility Predict renal function worsening 1 year after transplant	Elderly KTbot	[201]	Precision 90% Sensitivity 71% F1 score 0.79	2020

5.1.2.7. Other Renal Diseases

In this group are reported other renal diseases that do not fit within the classification provided so far. These studies focus on uncommon objectives, such as [203], which aims to predict the level of hemoglobin in patients with renal dysfunction, using numerical characteristics obtained from clinical data related to dialysis [204]; in [205], an application is developed that intends to define the need to perform or not a renal biopsy by analyzing physicians' annotations through a natural language processing ML algorithm; [206] try to predict the survival of hemodialysis patients using numerical characteristics (age, sex, diabetes mellitus, chronic glomerulonephritis or nephrosclerosis, body mass index, albumin, sodium, potassium, calcium, phosphorus, creatinine, total cholesterol, etc.). In [207], the authors extract radiomics features from three-dimensional ultrasound images to identify renal and liver tissue in patients with hydronephrosis. Finally, [208] use numerical features extracted from patients' EHRs with the corresponding acquisition time, to predict the risk of stratification of renal function deterioration.

Table 7 is presented analogously to the previous ones.

Table 7 – Other renal diseases research

Paper	Objective	Method	Database	Results	Year
[203]	Predict hemoglobin in patients with kidney disfunction	Data merging + clustering + ensemble of classifiers	[203]	Mean absolute error 0.662—Italy, mean absolute error 0.673—Spain	2014
[205]	Recommend renal biopsy	Tokenization + NLP machine learning classifier	[205]	Accuracy 83.5% Precision 80.6%	2019
[206]	Prediction of 1-year survival in hemodialysis patients	Ensemble artificial intelligence model	[206]	Accuracy 94.8%	2020
[207]	Detect kidney and liver tissue for hydronephrosis patient	Homodyne-K feature extraction + random forest	[207]	Accuracy 94%	2015
[208]	Predict risk stratification of renal function deterioration based on eGFR threshold	Multitask temporal-based classifier	[208]	Specificity 0.828 with 10% threshold Specificity 0.786 with 20% threshold	2015

5.1.3. Databases used in reviewed research

In this section, two tables contain information about the databases used in the research considered. This information includes the name of the database, when available, or otherwise a distinctive name related to the type of data and the organization in which they were collected; the number of elements that make up

the dataset; a brief description of the type of data present; the year in which the database was made public, when available, otherwise the year in which it was used for the first time in a paper; and, finally, whether the database is open access.

Specifically, in Table 8 are reported all databases that have as the data type diagnostic images; this can be CT, MRI, US, or images obtained through analysis in the laboratory with instruments such as a digital microscope. This second type of technique is mainly used for the detection of masses or malformations within the kidneys. It is possible to note that these types of databases have very different volumes; in the case of 3D US images, there are, for example, databases of nine patients; for CT and MRI, there are databases with a minimum of 50 cases up to a few hundred, and finally, with regard to other imaging techniques, there are databases from a minimum of 24 up to a maximum of 1321 cases. This discordance at the numerical level is given mainly by the effectiveness and invasiveness of the different examinations and therefore by the frequency of their use in clinical practice. US is a less effective imaging technique in this field, compared to CT and MRI, and, therefore, the studies concerning the application of this technique are very small and dated. As for the examinations performed on biopsies, the number of samples is much larger because it is an examination that is compulsorily performed in every case to define with absolute certainty the type of mass removed. Among the reported databases, only two are publicly accessible: the CPTAC Clear Cell Renal Cell Carcinoma Discovery Study [139, 209], released in 2018 by the U.S. National Cancer Institute, and kits2019 [142], released in 2019 by grand-challenge.org, hosted by MICCAI.

Table 8 – Diagnostic image databases.

Database	Number of Patients	Description	Year (First Use/Published)	Open Access
[125]	93	Patients' MDCT. Patients with complicated cysts: cyst with at least one focus of septa, a solid nodule, and any calcification or wall thickening on MDCT	2009	No
[126]	150	Patients' CT. 100 malignant tumors: 70 clear cell renal cell carcinoma (ccRCC), 20 papillary renal cell carcinoma (pRCC), and 10 chromophobe renal cell carcinoma (chRCC); 50 benign tumors: 20 lipid-poor angiomyolipoma (lpAML), 30 renal oncocytoma	2018	No
[127]	79	84 renal masses: 63 malignant (25 clear cell RCC, 23 papillary cell RCC, 15 chromophobe RCC), 21 benign (10 oncocytomas, 11 fat-poor angiomyolipomas)	2020	No
[130]	440	440 MRI and CT of patients with ccRCC	2020	No
[131]	54	Patients' CT. All patients have ccRCC.	2019	No
[139]	216	216 proteomics data and 783 slide images (524 tumoral)	2018	Yes
[142]	300	CT of patients with one or more kidney tumors. Segmentation of kidneys and tumors.	2019	Yes
[124]	188	The database is composed of 40, 16, 38, 60, 28, and 6 entries for healthy, stage 1, 2, 3, 4, 5, respectively. These images are obtained from 35 observers taken at different times. The kidney ultrasonic images are segmented and annotated into three regions of interest (ROIs)	2014	No
[176]	200	200 kidney stones harvested from nondestructive stone extraction at three different sites. Stone size was measured using a digital caliper	2020	No
[177]	412	LDCT of 235 kidney stones and 224 phleboliths	2019	No
[178]	254	Preoperative abdominopelvic ultrasound and intravenous urography or CT scan of PCNL patients.	2017	No
[207]	9	This dataset contains the 3D US abdominal images from 9 pediatric patients with hydronephrosis	2015	No
[191]	1321	Biopsy images of pathogenic (338) and nonpathogenic (396) glomerulus and some of pathogenic (338) and nonpathogenic (248) tubulus	2014	No

[192]	584	Renal biopsy reports, each of 4 or 5 slides with different stains, for each case: clinical and laboratory data, diagnostic hypothesis, histological biopsy study, histological report of glomerular disease	2021	No
[187]	422	Renal immunofluorescent images obtained by fluorescence microscopes relative to a renal biopsy of 162 patients with IgAN and 260 without	2021	No
[133]	24	24 unstained deparaffinized formalin-fixed kidney tissue sections of chRCC and oncocytoma, 12 of each type	2016	No

In Table 9 are reported all the databases exclusive of numerical type, relating to in-formation obtained from diagnostic tests, such as blood tests, genetic tests of kidney tissues, or data from patient history. For these databases, the volume varies; for more complex tests, such as genetic tests, there is a variation ranging from a few tens up to a few hundred cases; for medical histories, this ranges from a few hundred up to 269,999 cases; for simpler diagnostic tests, from a few tens up to several thousand cases. Of these databases, only three are publicly available; for some, access is limited to a specific country (in the table, these are reported as "only in the USA"). Among the public databases, two contain RNA sequences of renal tumors, which are used to identify the pathological stage of the tumor. Finally, the third public database contains data on blood tests, patient history, and information about CDK-related diseases.

Table 9 – Numerical databases.

Database	Number of Patients	Description	Year (First Use/Published)	Open Access
[140]	260	Tumor RNASeq and pathological stage (I, II, III, and IV): Stage I—172, Stage II—22, Stage III—51, and Stage IV—15.	2010	Yes
[141]	34	This dataset was obtained using Affymetrix HGU133 Plus 2.0 array platform and includes 19 and 15 samples in early (excellent survival) and late (poor survival) stages of PRCC.	2005	Yes
[154]	269,999	6.1% of patients in the dataset had a clinical deterioration event: 424 cardiac arrests, 13,188 intensive care unit (ICU) transfers, and 2840 deaths on the wards. For each patient, there are a total of 29 features.	2014	No
[156]	108,441	Australian and New Zealand Society of Cardiac and Thoracic Surgeons Database registry recorded 110,342 cardiac surgery events in 108,441 unique patients.	2018	No
[145]	780	Medical data collected by natural language process module from EMRs including demographic data, daily documentation, laboratory and imaging results, anesthesia records, medications, interventions, and diagnosis. TRIPOD guidelines were followed.	2021	No
[146]	50	Serial creatinine testing of patients with ≥20% total body surface area (TBSA) burns at risk for AKI. AKI was defined using the Kidney Disease: Improving Global Outcomes (KDIGO) criteria.	2019	No
[158]	153,821	153,821 patients from 6 different sites. Each patient had a mean of 67 (SD = 46) clinical facts per day.	2020	No (only in USA)
[149]	671	Information related to demographic characteristics, clinical condition, preoperative biochemistry data, preoperative medication, and intraoperative time-series hemodynamic features (systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial blood pressure (MAP), and heart rate (HR)) from electronic medical records and records on intraoperative variables.	2020	No
[152]	51,869	618,719 blood test occurrences for 51,869 distinct patients.	2021	No
[63]	400	The CKD dataset was collected from 400 patients from the University of California, Irvine Machine Learning Repository.	2015	Yes
[179]	254	This dataset includes information on preoperative, intraoperative, and postoperative parameters from 254 patients who underwent kidney surgery.	2019	No

[188]	1015	The variables contained per IgAN patient are age, sex, hypertension, serum creatinine, daily proteinuria, kidney biopsy, therapy—RASBs or corticosteroids. The primary outcome is ESRD, dialysis, or transplantation.	2020	No
[189]	80	Features of 80 IgAN patients: secondary IgA deposition, eGFR, MEST-C scores.	2021	No
[190]	284	38 features for each patient and biopsy diagnosis.	1992	No
[197]	194	Features for each patient: age, sex, time in dialysis, donor type, donor age, HLA mismatches, delayed graft function, acute rejection episode, and chronic allograft nephropathy.	2010	No
[198]	7348	A total of 793 pre- and post-transplant variables per patient.	2004	No (only in USA)
[155]	6564	First 12 h of 6564 HER from critically ill children admitted to a pediatric ICU without evidence of AKI; 4% of the patients developed AKI by 72 h.	2016	No
[151]	2642	The dataset contains the data relative to 1781 patients pre-implementation and 861 patients post-implementation of a digital intervention system, with the relative alert severity.	2019	No
[180]	358	This dataset includes 42 features including the two target variables, stone-free and one-session success, for all 358 cases. The number of cases with stone-free and one-session success was 253 (70.7%) and 154 (43.0%).	2020	No
[157]	1250	Several serum markers per patient undergoing angiography as clinical standard care.	2015	No (only in USA)
[200]	80	80 patients who received HLA-incompatible renal allografts; 14 features measured before transplantation.	2019	No
[201]	118	Medical records of 18 elderly and 100 younger patients.	2020	No
[182]	1386	Anthropometric measurements and blood pressure (BP), drug use and past medical history, physical assessment for retinopathy, sensory neuropathy, and peripheral arterial disease. eGFR calculated using the Chinese-modified Modification of Diet in Renal Disease equation.	2013	No
[183]	1000	1000 T2DM patients' data collected by the IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) of the Hospital of Pavia.	2018	No
[184]	~32,000	Diabetes of type 2 patients with a 24-month analysis window.	2019	No
[205]	3149	This dataset contains a total of 3149 admission notes from the nephrology department. For the ground truth, there are recommendations given by physicians in first-day progress notes.	2019	No
[208]	6435	Electronic health records of patients with hypertension, diabetes, or both.	2015	No
[202]	~31,000	United Network for Organ Sharing, a private, non-profit (UNOS) dataset including information on all kidney waiting-list registrations and transplants that had been recorded in the U.S.	2014	No (only in USA)
[203]	13,011	125 features from dialysis clinical practice of 13,011 patients.	2014	No
[193]	14	ESRD patients on chronic hemodialysis or hemodiafiltration weighing 20 kg or more.	2018	No
[206]	79,860	Various features for each patient are presented with the relative risk score based on mass, serum albumin level, cholesterol level, and creatinine.	2020	No

5.1.4. Discussion

After having reported in the previous sections the existing methods in the literature to address renal pathologies with machine learning methods and analyzed the available databases, we summarize in this section what has been found for each pathology; in particular, the limitations of the studies carried out so far and possible future developments will be indicated.

Regarding renal masses, the goal of the analyzed works is to find a method to non-invasively discriminate benign and malignant masses [127], and artificial intelligence has the potential to become a very important

tool for assisted diagnosis. This is motivated by the results of identified research, in which are obtained accuracies ranging from 79% [130] to a peak of approximately 90% [127] (these results are from private single-center databases). Currently, the gold standard for the detection of a renal mass is based on the analysis, by an experienced physician, of CT images before and after dosing with a contrast medium [210]. AI can perform the discrimination function because it can analyze diagnostic images, such as CT, at a very high or equal level of detail as an expert [211]. This is because it can also take into account multidimensional characteristic features, such as texture. However, using CT, the various parameters used for the acquisition and the timing with which it is done assume an important role [127]. In fact, from the articles analyzed, it emerges that, according to the CT acquisition phase taken into consideration, the results obtained change; specifically, the most used phase is the corticomedullary phase [128]. Furthermore, as regards the use of CT for the extraction of characteristic features, the literature considers the three-dimensional use of CT to be better and more representative [212], but in the research identified [125–134], to reduce the workload of manual segmentation and facilitate the repeatability of this operation, a limited number of slices or only the two-dimensional slice containing the largest portion of the mass considered is used. In addition to how CT is used, it is also important to control the method by which features are extracted; in some research [126, 127, 129–134], radiomic features are used, after manual segmentation by at least one experienced physician, to classify tumors. One of the major limitations introduced, in doing so, is the bias of the operator who performs the segmentation [213]. For this reason, more recent studies [136–138] have focused on overcoming manual segmentation by creating deep learning algorithms capable of automatically segmenting kidneys and tumors present in CT; the results obtained from these studies are positive, as they achieve a mean kidney tumor size–mean per CT of the testing set of (Kidney Sørensen-Dice + Tumor Sørensen-Dice)/2 [142], with a maximum of 0.9168. In particular, one solution proposed in the literature to deal with operator-introduced bias is for a team of clinicians to collaborate on the kits2019 database in a way that reduces the risk of bias as much as possible.

Regarding AKI, this pathology is very widespread, with consequences that, if not treated in time, can even lead to death. Currently, there is no specific intervention that can prevent AKI; there are only general measures that can be taken to delay more critical procedures such as surgery [153]. For this reason, most of the recently developed research focuses on predicting the prognosis of this disease [143–146, 148, 149, 151–153], being able to predict AKI with good accuracy even 30 days in advance [152]. The solutions implemented depend not only on the task but also on the actual number of data available for each patient [214]. Maintaining a large amount of data for each patient has an economic cost and features used in one center may not be available in other centers [143]. ML techniques can outperform clinical tools used to estimate AKI risk, as we see in [144], with an AUC of 0.85. The performance of solutions exploiting ML for the prediction of AKI is positive: AUC 0.76 [145], in liver transplant patients; 97% accuracy [146] and AUC 0.76 [153], for burn patients; AUC [0.79–0.843] [148, 149], for patients undergoing coronary angiography. However, de-spite the various existing applications, there is a lack of a ML-based prediction systems that can be recognized as state of the art for AKI prediction [145].

Regarding CKD, this is a very common type of disease, which, if detected in time, can be managed through periodic therapies. Thanks to the University of California, Irvine (UCI), which made public the database known as UCI CKD [63] (containing 24 characteristics, derived from patient history and diagnostic tests, plus information regarding the presence or absence of CKD), many studies have been developed to diagnose CKD. Since this database was made public, various studies have used it to test multiple different types of solutions, obtaining increasingly impressive results for accuracy (63–100%) [161–165, 167, 168, 170, 172], AUC (0.995) [166], and F1 score (100%) [169]. Being the only public database available for this pathology, the research has been mainly focused on the analysis of numerical features; this is also due to the fact that patients suffering from CKD, or otherwise at risk, cannot undergo all the existing diagnostic imaging techniques. In this case, techniques that require the use of radiation, such as CT, are strongly discouraged, because they can easily worsen the patients' condition. Therefore, imaging techniques such as US, used in [124] with 82% accuracy

in predicting the stage of CKD, and MRI are preferred. The latter has been shown to have the ability to allow assessment of both renal function and structure [215]. Major future developments may shift in this direction and focus on the development of methods that take advantage of MRI to be able to determine CKD.

If radiation imaging techniques cannot be used to determine CKD, the same is not true for detecting and analyzing kidney stones. In particular, for kidney stones, it is possible to use not only CT but also low-dose CT (LDCT), which exposes the patient to approximately five times less radiation than regular CT [216]. The independence of the dosage used to acquire CT is demonstrated in several studies: in [176], ML techniques are applied to process LDCT and CT and identify the composition of a kidney stone, achieving 86% accuracy for both assays used; in [177], LDCT is analyzed to differentiate between kidney stones and phleboliths in patients with acute flank pain, with 85.1% accuracy. The applicability of these methods ensures that low-dose radiation CT acquisitions can be used for the detection of a kidney stone, reducing any risks associated with the radiation expo-sure of normal CT. In addition to the detection and analysis of kidney stones, researchers are also studying the prediction of success in removing a kidney stone. Successful selection of the most appropriate method can lead to a higher rate of kidney stone clearance, lower risk of associated morbidities, higher probability of survival, faster recovery, and lower overall cost of care [217]. Depending on the procedure chosen [179, 180], and for the prediction of stone removal, there is 60% accuracy [179] for predicting success after the first treatment, and 87.9% for predicting success when a shock wave is used for kidney stone clearance [180]. Accuracies ranging from 81% to 98.2% have been obtained for predicting a patient's condition and possible complications following renal stone removal [178].

Since glomerular disease is a condition that worsens over time, the machine learning techniques implemented are primarily focused on predicting the prognosis of the condition and identifying the consequences caused by the presence of the disease [87–89,92–98]. The most common glomerular disease prevalent in the world is Immunoglobulin A Nephropathy (IgAN) [218]. IgAN is caused by renal dysfunction and can be diagnosed by diagnostic imaging of the kidney, particularly immunofluorescence imaging. Some re-searchers have focused on diagnosing IgAN from diagnostic images with different resolutions, with an accuracy of at least 80% [187] and an accuracy of 80.27% [190], using only clinical and laboratory analysis data. Around 30–40% of IgAN patients carry the risk of the disease degenerating into ESRD (end-stage renal disease) [188]; for this reason, some re-search tries to predict this degeneration to allow the efforts of physicians to focus mainly on patients who are more at risk, as, for example, in [188], where it predicts the degeneration of the disease in the next 5 years, with AUC of 0.82, and after 10 years with AUC of 0.89, and as in [189], with 79.8% accuracy. Another particular type of glomerular disease is caused by diabetes. Since diabetes is very common, it is very important to prevent its de-generation into diabetes kidney disease, and in [87–89], the authors focus precisely on this aspect by creating algorithms that can predict the prognosis, with an accuracy of 83.5–94%.

Regarding the literature inherent to renal transplantation, it is possible to identify three possible applications of AI [218]: (i) diagnosis, using AI to diagnose the level of transplant risk by detecting parameters associated with renal transplant rejections, and identifying abnormal patterns within them, as in [199], with 68.4% accuracy, and in [201]; (ii) prescription, using AI to prescribe postoperative therapies [219] to prevent complications or rejection, or to prescribe diets that may improve quality of life after renal transplantation [220]; (iii) prediction, using AI to predict mortality, and possible rejection, as in [197], with 73.8% specificity and 88.2% sensitivity; in [198], with 56% accuracy over a 3-year timeframe from possible rejection, and in [200], with 85% accuracy. It is important to note that for this specific task, the main limitation for the application of AI is given by the fact that the type of database is very patient-specific [103–106], as the values are highly dependent on both the recipient and the donor(s) available, resulting in a limitation that makes it difficult to generalize the solutions devised [221].

Before concluding, we believe that it is also necessary to analyze the ML algorithms used in nephrology, to address a possible reader interested in a specific type of algorithm rather than another, depending on the

type of application that they would like to achieve. First of all, it is possible to notice that all the ML algorithms used are based on the use of supervised learning techniques. This is mainly due to the fact that the realized tasks are formulated and viewed in the form of classification problems. In particular, with regard to the research identified in this work, in Table 10, all the methods used have been grouped by algorithm type.

Table 10 – Searches grouped by type of ML algorithm applied.

Method—ML Algorithm (Based)	Authors	Year
Bayesian classifier	[125]	2009
	[198]	2012
	[199]	2018
Logistic regression	[148]	2019
	[201]	2020
Decision tree	[162]	2016
	[192]	2021
	[197]	2010
Random forest	[207]	2015
	[161]	2016
	[165]	2017
	[143]	2018
	[134]	2018
	[132]	2019
	[200]	2019
	[127]	2020
	[129]	2020
	[169]	2021
	[152]	2021
	[183]	2018
SVM	[182]	2013
	[124]	2014
	[133]	2016
	[163]	2016
	[189]	2021
ANN	[176]	2019
	[178]	2017
	[193]	2018
	[188]	2021
Ensemble of classifiers	[179]	2019
	[206]	2020
	[203]	2014
	[164]	2017
	[146]	2019
	[177]	2019
	[172]	2019
	[131]	2018
	[184]	2019
	[128]	2019
	[126]	2018
	[144]	2021
	[145]	2021
	[153]	2020
	[180]	2020
	[149]	2020
	[130]	2020
DNN	[191]	2014
	[208]	2015
	[205]	2019
	[151]	2020
	[190]	1992
	[166]	2018
	[167]	2018
	[168]	2019
	[132]	2019
	[187]	2021

	[136]	2019
	[137]	2019
	[138]	2019

From the table, it can be observed that the simplest and most common classification algorithms, such as random forest and support vector machine, and ensemble algorithms, such as gradient boosting machine, are the most used in these types of studies. However, more complex ML algorithms, such as artificial neural network, and deep neural networks, such as convolutional neural network, autoencoder, and more sophisticated approaches based not only on feature or image analysis, but also on natural language processing and the temporal evolution of features (temporal-based approaches, e.g., recursive neural network) are not missing. This could be due to the lack of very large public databases that would allow better use of the more complex ML techniques [222].

It is also possible to note that the methods applied by the authors differ mainly with respect to the type of the used data and the techniques of analysis and data processing. In particular, in cases where the database is composed exclusively of numerical features, derived from patients' medical records, classifiers such as support vector machine, random forest, and artificial neural network are the most frequently applied. Whenever diagnostic images are present, instead, the type of ML technique varies according to the preprocessing applied to the data. In the case of minimal or null preprocessing, techniques such as convolutional neural network are used, in which the model directly analyzes the image and finds the most relevant features in order to classify it. Instead, when algorithms are used for the extraction of radiomic features from specific anatomical regions, algorithms generally applied to numerical features are used; in particular, ensemble algorithms are exploited, which typically, in these cases, guarantee a better result in terms of metrics.

Finally, for the evaluation of algorithms' performance, the authors feel that it could be misleading to compare methods applied to the same objective based on the values obtained from the evaluated metrics computed with different data. However, it is possible offer some considerations about the various metrics used, in order to understand in which cases some metrics are used instead of others. Since the most commonly used metrics are accuracy and AUC, we consider it appropriate to briefly discuss what the differences are: accuracy is a metric that represents the ratio of the number of correctly predicted samples to the total number of samples present; AUC, on the other hand, represents the area under the receiver operating characteristic (ROC) curve that shows, for different probability thresholds, the relationship between the false positive rate (ratio of the number of false positives to the total number of negative cases) and the true positive rate (ratio of the number of true positives to the total number of positive cases). Looking at the two definitions, it may be deduced that the accuracy is a more intuitive metric and therefore more frequently used, but its simplicity has drawbacks, since it cannot be used in all cases—for example, in the case of unbalanced datasets, where it is preferable to use metrics such as the F1 score or AUC, or in case it is desired to take into account the probability associated with the various classes predicted, in which case only AUC takes this aspect into account. With the above in mind, the use of AUC is strongly recommended as it encapsulates increasingly confident information than accuracy alone.

Despite the limits of this work, given the continuous evolution of research in this area, based on what has been analyzed so far, it is possible to conclude that, given the many existing applications of ML in nephrology, AI has great potential and versatility in this field. An example of a possible application for kidney image analysis can be based on the combination of the multiple methodologies that currently exist, such as the use of deep learning to detect kidneys and tumors [38–40], followed by the use of other machine learning techniques to classify the nature and/or severity of tumors, or the presence of any kidney disease and/or other possible masses. However, this does not mean that limitations are not still present. Most of the studies identified end before moving to a clinical trial, remaining only single-center retrospective studies, reducing

their external validity [223, 224]. Consequently, the main and most urgent gap that should be addressed as soon as possible is that of the public availability of data; this will not only allow studies to be compared with each other but will ensure that there are improvements in nephrology itself [114]. To this end, the guidelines for conducting clinical trials in nephrology, reported at the Kidney Disease-Improving Global Outcomes (KDIGO) conference, could be followed [221].

5.1.5. Final remarks

Fifty-nine, from a total of 224, studies concerning the application of ML techniques for the segmentation, prediction, and classification of renal diseases were analyzed. First, the studies were divided, analyzed, and presented based on the addressed pathology and the main goal of the research. Then, the existent datasets were analyzed in terms of data typology, size, and public availability; the main concept derived from this analysis is the importance of a large dataset and public availability to allow research to go as far as possible for a specific objective. Finally, the various pathologies were discussed in terms of what does not exist and what can be done to achieve further developments in this specific sector. In conclusion, from the analysis of the literature, it can also be noted how the introduction of modern ML techniques in the nephrological field allows the achievement goals not obtainable with traditional techniques, such as speeding up and automating CT segmentation processes, the possibility to perform non-invasive and reliable diagnosis, and to create predictive models - for example, to evaluate surgical or transplant outcomes and create predictive models to monitor patient's parameters in order to act promptly.

Among all the works analyzed, it can be seen that the practical purposes of the use of AI in urology range from the diagnosis of a disease, to the analysis of diagnostic images, to the prediction of prognosis, etc., and generally aim to aid doctors in making more accurate decisions, without attempting, in any way, to replace them [225–228]. The physician's attendance remains essential both from a human point of view, in establishing a deep doctor-patient bond of trust that can improve the success of any therapies and treatments [229], and from an ethical and accountable point of view for diagnoses [230].

5.2. Kidney tumor classification

According to GLOBOCAN 2020 estimations, renal tumors are among the most prevalent cancers in the world and reach a mortality rate of 42% [231]. Renal tumor is a particular type of kidney mass that can be either malignant or benign. The majority of malignant renal tumors are RCC, a type of malignant tumor (carcinoma) that originates from the glandular epithelial cells of the kidney and has the potential to invade the surrounding tissues tending to metastasize to other anatomical sites [232], being the cause of 80% of renal cancer deaths [123]. As for benign renal tumors, one of the most common is oncocytoma, which is characterized by the presence of large cells with abundant eosinophilic granular cytoplasm [233]. Although oncocytoma does not entail long-term risks [234], it accounts for approximately 16% of surgically removed renal masses [235] because of its high similarity to clear cell RCC [236]. In fact, currently the standard procedure for the treatment of renal tumor involves the use of diagnostic imaging techniques to determine the presence of suspicious renal masses and their characterization by analyzing the tissues with visual inspection, comparing the mass-related parameters obtainable from the diagnostic image (e.g., size, shape of the mass, etc.) [237]. Unfortunately, distinguishing the nature of a renal mass is a very complex task even for an experienced physician due to the similarity between oncocytomas and RCCs at the radiological imaging level [236]. Therefore, the best current method to diagnose the nature of the tumor is based on histological analysis of the tissue of the mass [238] collected by biopsy after a radical nephrectomy of the kidney containing the mass or a partial nephrectomy of the tumor [239].

Being able to detect a renal tumor in advance and being able to classify it correctly would be a crucial step, as it would allow the introduction and use of medical procedures that can safeguard the patient's life and renal function [240]. In fact, when a mass is identified with certainty as benign, the surgical solution can be avoided and treatment involves continuous monitoring and control. Therefore, it becomes strategic to be able to implement an analysis tool, which can overcome these limitations. Given the wide use of ML techniques in image analysis and their evolution in the analysis of diagnostic images [211, 241–243] the idea behind this work is to rely on these techniques for the realization of this tool. ML procedures generally change according to the type of available data. In particular, for medical ML applications usually the available data can be of two types: (i) clinical information, derived from routine clinical examinations, e.g. blood tests and medical history, (ii) information derived from diagnostic images, from specific clinical examinations, such as CT and MRI. With regard to diagnostic images, there are various approaches and the most common is based on identifying, by manual segmentation, the areas of interest within the CT and extracting from it specific features (e.g. texture, size, volume, etc.), named radiomic features. In addition, there is also the possibility of directly using diagnostic images with deep learning algorithms to extract characteristic features, called deep features. While the radiomic features are more interpretable, the effort required in terms of segmentation is considerable and the results can vary depending on how this is performed, as proved in [128].

In this work the available data is a set of CTs from renal cancer patients. Using these data, this paper proposes two approaches for the classification of renal tumors, with a special focus on the distinction between oncocytomas and clear cell RCCs, to provide an assisted diagnosis tool. The two approaches were developed and compared, seeking to create a tool that could be generalizable, updateable to new data, and fully automatic (without the need to manually segment regions of interest). It has been paid close attention to the results analysis and comparison between the two approaches, trying to eliminate bias caused by unbalanced data and incorrect readings of output values. Moreover the focus was placed on building an algorithm capable of extracting in a fully automated framework both radiomic and deep features, exploiting automatic CT segmentation using a Convolutional Neural Network (CNN). The obtained results show that automatic segmentation can be used as a starting point for the extraction of radiomic features for this type of task, and also that the obtained deep features are sufficiently representative. In particular, comparable results were obtained with the two types of features, achieving an accuracy above 85% for both methods.

5.2.1. Background existing methods

Various techniques have been developed to be able to distinguish malignant and benign renal tumors through the use of ML-based techniques that have evolved over time along with ML algorithms and their image processing capabilities [244–246].

Early works in the literature with significantly relevant results use radiomic features extracted directly from manually segmented areas within diagnostic images such as CT and MRI. Relying on the use of radiomic features, in [247] starting from the CT slice on which the size of the mass is largest, 5 to 10 slices were selected to extract the features to be used with a random forest classifier with the purpose of classifying various solid renal masses (oncocytomas, ccRCC, cysts, and papillary RCC). In order to identify the actual value of texture analysis, in [248] are extracted radiomic features from 10 consecutive CT slices to be used with a support vector machine classifier to analyze the radiomic features related to each of the 10 slices with a majority voting algorithm to decide the actual tumor class. [126] aims to investigate the usefulness of relational statistical machine learning algorithms to differentiate benign and malignant tumors, using radiomic features extracted from CT, analyzing the slice in which the tumor diameter is largest to obtain the two-dimensional features and the entire volume to derive the three-dimensional ones. [249] is among the first research for this specific task to compare the ability in differentiating benign and malignant solid renal masses of experienced radiologists and machine learning algorithms trained on radiomic features extracted from tumor CT by MCNemar test [250]. Unlike other works, [251], extracts radiomic features from contrast-enhanced MRI images related to 3 consecutive slices to differentiate various types of renal tumors, selected from the section where the tumor area is largest, using them with a random forest classifier. [126, 247–249, 251] follow a similar pipeline: manual segmentation of the tumor within the diagnostic images by one or more experts with an internal validation by another expert; feature extraction; selection of the most relevant features using specific algorithms (e.g., recursive feature elimination); feature processing, with the aim of normalizing the values (e.g., z-score normalization); classifier training; and finally, validation with algorithms specific to the identified method (k-fold cross validation).

In recent years, thanks to the increase in computing power due to the large-scale deployment and affordability of GPUs, studies began to introduce the use of deep learning algorithms using CNN capable of performing direct image processing. [252] studies the diagnostic value and feasibility of a renal tumor classifier based on deep learning, taking advantage of a CNN (Inception v3 [253]), and the four phases of tumor CT. [254] introduces a mixed classifier that is divided into three main branches: one using a logistic regression classifier with radiomic features, and the other two using a CNN (ResNet50 [255]) by taking as input an image related to a slice of the tumor in which the three channels R, G, and B are matched to the axial, sagittal, and coronal gray-scale slice in which the tumor had the largest diameter, following the 2.5D model [256].

A compact view of the performance and methods found in the state of the art can be seen in Table 11; the performances are given only as a reference since the results are not directly comparable as the used databases are different. For each paper is indicated the used model or classifier, the addressed task, the type of input, the amount of data, and a brief summary of how the data were used and how the results were obtained (in cases where a standard procedure was not followed). In addition, the metrics obtained for each classifier are provided. For consistency, only the results related to classifications that involved oncocytomas and ccRCC are reported.

Table 11 Comparison of state-of-the-art algorithms to distinguish malignant and benign renal tumors through the use of ML-based techniques.

Paper	Model	Data Type	DB	Classification Task	Metrics						
					Sensitivity	*Specificity*	*Accuracy*	*F-Score*	*AUC*	*PPV*	*NPV*

[247]	RF	CT Arterial Phase	24 Onco 25 ccRCC	Onco	81%	97,80%	-	-	-	-	-
				ccRCC	93,50%	95,10%	-	-	-	-	-
	Summary	The authors consider individual slices containing the element of interest as a single input. The percentages in the results are relative to the number of total slices and not on the actual number of masses studied. Results refers to the slices of the testing set composed by 4 oncocytomas and 5 ccRCC.									
[248]	SVM (linear)	CT Portal venous phase	46 ccRCC 63 other RCC 10 Onco	ccRCC vs all	-	-	-	-	91%	-	-
				Onco vs all	-	-	-	-	86%	-	-
	Summary	The authors considered 10 slices containing tumor per CT. From these they extracted 43 texture features to be used with a classifier. The results are reported on all cases despite using 5-fold-cross validation.									
[126]	RFGB	CT - 4 phases	70 ccRCC 30 other RCC 20 angio 30 Onco	Benign vs Malignant	-	-	82%	87%	83%	-	-
	Summary	The authors use the slice where the tumor diameter is largest to obtain 2D features and use the entire volume for 3D ones. This process is repeated for each CT stage, and the obtained features are concatenated to form a single descriptor.									
[249]	SVM	CT Arterial phase	190 ccRCC 64 other RCC 26 angio 10 onco	ccRCC vs (angio and onco)	86.3%	83.3%	85.8%	91.1%	-	96.5%	53,6%
	Summary	The authors analyze radiomic features extracted from tumor CT by MCNemar test. Metrics are considered on the full dataset.									
[251]	RF	MRI	90 ccRCC 22 other RCC 30 onco	onco vs ccRCC	67,3%	88,9%	79,3%	-	-	-	-
	Summary	The authors analyze features extracted from 3 consecutive slices of an MRI starting with the slice where the tumor has the largest diameter									
[252]	Inception V3	CT multiphases	128 ccRCC 51 onco	ccRCC vs Onco	88,3%	52,9%	75,4%	-	-	80,3%	-
	Summary	The authors create RGB images to be used as network inputs by inserting one slice per phase into the R-G-B channels. Different combinations have been tested varying the phases and the type of slice, but the results refer only to the best ones.									
[254]	LR + 2x ResNet50	MRI	425 ccRCC 230 others RCC 92 Onco 406 Angio	Benign vs Malignant	40,81%	92,06%	69,64%	54,05%	-	80%	66,67%
	Summary	The authors use the 3 slices in which the tumor is largest on the axial, coronal, and sagittal axis, fitting them into the three R-G-B channels to create the images to be used as input. An ensemble model is built using 2 Resnet50, one trained with MRI T1C									

		and the other with MRI T2WI, and a logistic regression classifier trained with radiomic features extracted from the lesion present in the MRI.

5.2.2. Materials

In this work, a new framework for the study and analysis of diagnostic images for the classification of renal tumors through the use of ML techniques was developed and validated. Specifically, CTs of renal tumor patients with proven histological findings were used, no other data types are included in the study, and all data was provided or retrieved anonymized. It is important to point out that CTs performed with contrast agents are multiphase, i.e., multiple acquisitions are performed, up to a maximum of 4 phases for the kidney, depending on the time elapsed since contrast agent injection: i) before the use of contrast agent unenhanced phase, ii) after 30 seconds corticomedullary (or arterial) phase, iii) after 90 seconds nephrographic phase, iv) excretory phase. Given the possibility of having multiple phases, it is necessary to keep in mind that it is not always possible to have them all available, as they are acquired according to the clinics' individual protocols.

The information was retrieved from two different sources: (1) “Careggi” Florence Hospital (Azienda Ospedaliera Universitaria Careggi - AOUC) [22], (2) the 2019 Kidney Tumor Segmentation Challenge (kits2019) [257]. The data from AOUC are related to 61 patients of which 29 with ccRCC and 34 having an oncocytoma, with one or more acquisition phases and histological result. On the other hand, as for the kits2019 data, these are publicly accessible [258], and are related to 300 patients with at least one renal tumor, in particular there is the CT of the arterial phase for each patient, the histological result and some information about the patients, e.g. if they are smokers, if they have other diseases, if they have already had surgery, gender, more information can be found in [259]. Details of the data available for this study can be seen in Table 12.

Table 12 - Dataset included in the study.

Data source	**Number of Patients**	**Data**	**Description**
AOUC	61	CT multiphase	The data are referred to patients with only one kidney tumor that can be an oncocytoma or a ccRCC. In both cases the size is limited to a few millimeters. The data have been acquired using different scanners and different protocols, for this reason for each patient there could be from 1 to 4 phases.
KITS2019	300	CT corticomedullary	The data are referred to patients with at least one kidney tumor of any kind. The size is heterogeneous, but all the CT are referred to the same phase. The data in this dataset have been acquired from two different clinics and with different scanners.

To standardize the type of available data, the following choices were made: (1) use only diagnostic images, discarding any registry data, and the corresponding histologic result as the ground truth. In particular for data from AOUC, for patients in whom more than one acquisition phase was present, the corticomedullary phase was chosen, being the one known to be most used for diagnostic imaging in the case of CT contrast enhanced; (2) only patients with a single renal tumor of ccRCC or oncocytoma type were considered, discarding any patients with multiple renal tumors or renal tumors of different nature at the same time. As a result, the final database comprises 271 patients. Figure 14 shows the database information used in this study after appropriate exclusions.

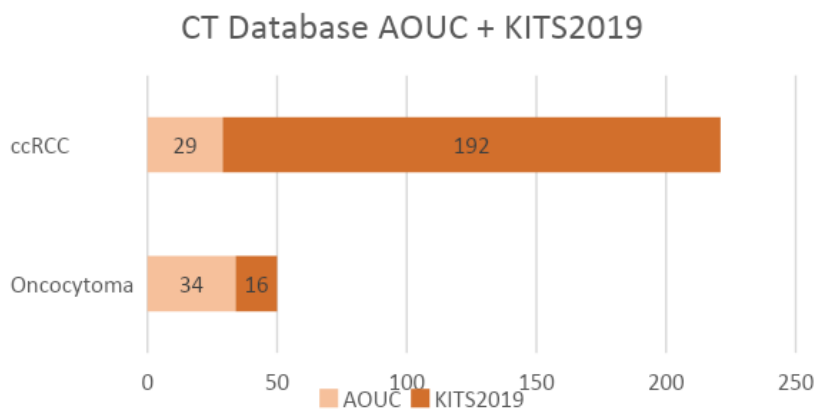


Figure 14 - CT dataset composition derived from the union of AOUC and KITS2019 data.

5.2.3. Methods

At a general level, the structure of the implemented application can be divided into two modules (Figure 15): (i) the feature generator that is in charge of creating, from the patients' CTs, a database of features extracted with specific algorithms; (ii) the classifier in charge of distinguishing the type of the tumor by the analysis of the extracted features.

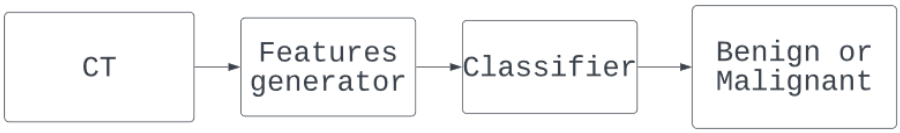


Figure 15 - General structure of the application.

In particular, in the feature generation process the extraction of two different types of features was implemented, one based on statistical methods (radiomic features) and the other based on deep learning (deep features). Finally, classification algorithms of different nature were examined with the aim of identifying the best performance for feature type.

In the following sections, the radiomic and deep feature extraction methods are explained in detail, specifying with which algorithm and techniques they were obtained and how they were processed before being included in the final database.

5.2.3.1. Radiomic Features Extraction – Type 1 Features

The main goal of radiomics is to extract quantitative and ideally reproducible features from diagnostic images, thereby including complex patterns that are difficult to recognize or quantify by the human eye [260]. Radiomic features represent properties related to the characteristics of tissues and lesions present in diagnostic images e.g., heterogeneity, shape, etc., and within a sufficiently large dataset these data turn out to be minable, i.e., usable to uncover hidden patterns to discover diseases or implement specific therapies.

Broadly speaking, radiomic features can be divided into statistical features, e.g., histograms, texture analysis, analysis by parameterized models, analysis by transformations (Fourier, Gabor, Haar, etc.), and shape-derived features, which can be extracted either two- or three-dimensionally.

The commonly used procedure for radiomic feature extraction consists of the following steps: (1) deciding which is the region or volume of interest (ROI) (2) segmentation – manual or automated – of the area of interest within each slice under consideration; (3) feature extraction using specific algorithms; and (4) feature post-processing comprising one or more of the existing feature processing techniques, e.g., feature

harmonization, selection, and reduction [261]. After these steps the features are ready to be used with a statistical model to accomplish the required task.

Several strategies have been proposed to define the portion of anatomy from which to extract radiomic features. In this work the standard procedure that takes into account the whole region is considered along with the strategy that uses only the slice with maximum tumor size on the three CT planes.

In order to automatically identify and segment ROIs, avoiding time-consuming and non-reproducible processes, it was decided to use the nnU-Net [136] trained to segment kidney and kidney tumors.

Regardless of the procedure used to obtain radiomic features, they must go through a pre-processing step before they can be used. This consists of a normalization process according to the following equation $X_{new} = \frac{X-\mu}{\sigma}$, in which the normalized value is obtained by dividing the subtraction between the original value (X) and the mean value of the data (μ) by the value of the standard deviation of the data (σ). This is equivalent to performing a z-score normalization.

In addition, given the high number of extractable features, a feature selection algorithm, namely recursive feature elimination (RFE), was applied to reduce the number of features to be analyzed and consider only the most significant ones. Specifically a variant of RFE was used which exploits an SVM and has been shown to be more efficient when analyzing biomedical data [262]. The number of selected features ranged from a minimum of 5 to a maximum of 45.

5.2.3.2. Deep Features Extraction - Type 2 Features

Deep features correspond to the output of a layer belonging to a deep neural network. The "depth" of these features will depend solely on the depth level of the layer from which they are extracted. "Depth" is directly proportional to the level of detail and inversely proportional with generalizability; therefore, the deeper the extracted features the greater the level of detail of these features and the lower the level of generalizability.

Taking into consideration the definition of deep features and how they are obtained, it was considered as one of the main strategic elements the choice of the deep neural network to be used for feature creation. A network trained in a task similar or at least related to the one addressed was selected for this purpose. Specifically, the nnU-Net [136] was used in its variant trained with the goal of segmenting from a CT the kidneys and kidney tumors.

After some careful evaluations, it was decided to use this network, which, by identifying the location of kidney and kidney tumors, allows deep features to be extracted directly from these regions, providing characteristic and specific details useful in classifying the type of tumor.

In order for the model to analyze a CT with any number of slices, a patches approach was exploited, by which the CT is divided into sub-portions called patches. Each patch generated goes through various convolutional blocks until it reaches a minimum size of 320x4x4x4, which corresponds to the lowest point of the nnU-Net. With the aim of identifying the most specific features, the level chosen for feature extraction is the one in which the achieved dimension is the smallest possible and consequently the number of features is also the smallest possible. By doing so, it is possible to obtain features for each of the patches and consequently for the entire CT. Finally, by reverse mapping the patches using the final output of the network (the segmentation of the tumor and kidney) all features related to patches in which no parts of the tumor are present are discarded. The deep feature extraction procedure just described can be seen in Figure 16.

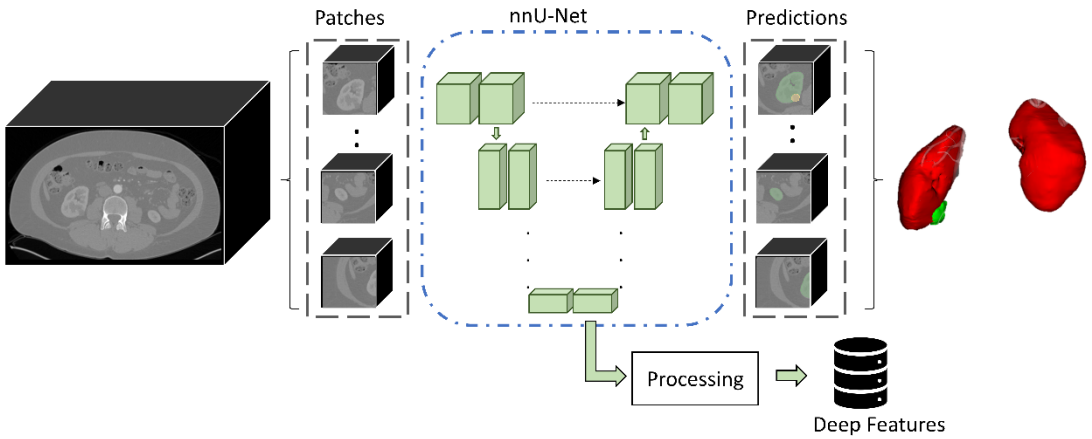


Figure 16 - Deep feature extraction scheme.

Once the deep features matrices are obtained, a processing of the matrices is carried out, with the aim of obtaining a single vector of unique size independent of the size of the CT input to the network. This processing consists of the following steps for each CT: for each matrix of deep features obtained from the single patch, of size 320×4×4×4, these are transformed into vectors of size 20480×1 and join together to create a matrix of size N×20480×1, with N equal to the number of patches considered valid; once this matrix is made to obtain a single vector, a statistical operation is used to reduce the information into a single vector, of size 20480×1 to be used as input for a final classifier. Figure 17 illustrates this feature reduction procedure.

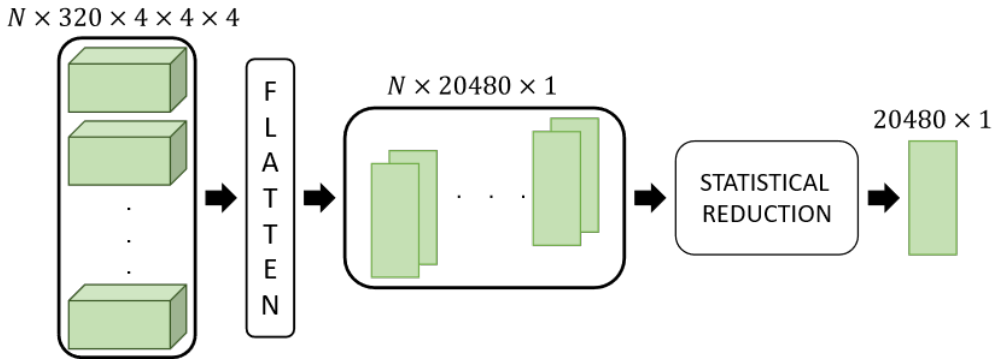


Figure 17 - Feature reduction procedure.

Figure 18 shows the framework for extracting the two types of features and thus creating the dataset exploiting in both cases the nnU-Net for automatic CT segmentation.

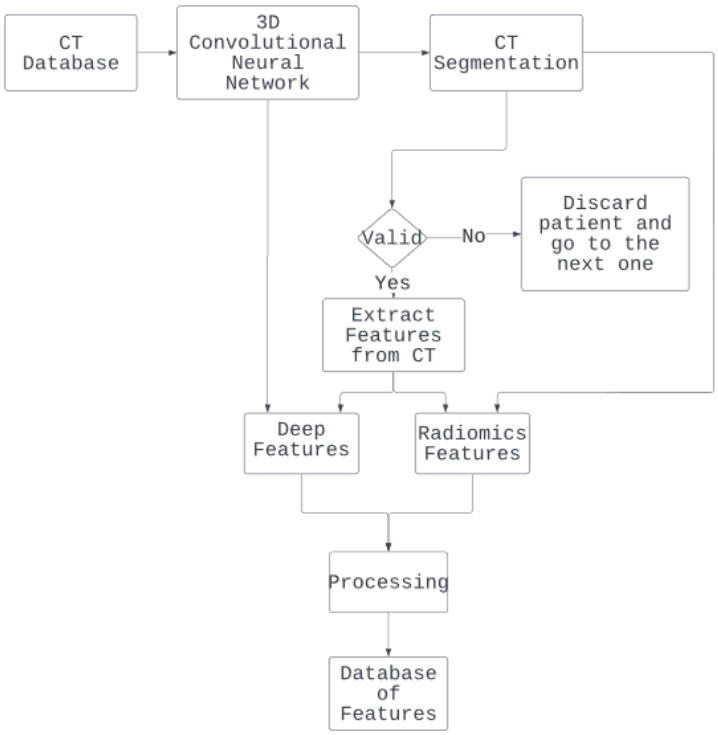


Figure 18 - Feature extraction process.

5.2.3.3. Classification

At this stage, the previously generated data is divided to form two sets, the training set (66.79%), and the testing set (33.21%). The training set was further divided into two sets by k-fold cross-validation (k=5). The most common types of classifiers (random forest (RF), k nearest neighbor (K-NN), support vector machine (SVM), artificial neural network (ANN)) were tested.

Concerning the parameters and variances of the classifiers, for RF it was set an internal decision tree number of 1000, with a minimum number of splits of 2 and Gini impurity as the criterion for measuring the quality of splits. Considering K-NN the number of k neighbors varied between 2 and 8. For SVM it was decided to test the possible variants of the kernels, namely linear, polynomial (by varying the degree from 2 to 5), radial basis function, and sigmoid. Finally, with regard to the artificial neural network an approach based on the following thumb rule on the number of hidden nodes to be used for this type of classifier was used:

$$Hidden_{nodes} \leq \sqrt{Input_{nodes} \times Output_{nodes}}$$

With $Input_{nodes}$ equal to the number of input features and $Output_{nodes}$equal to 1. The number of internal nodes in the network is limited to reduce the risk of overfitting the model on the training set. Regarding the number of layers and the distribution of these nodes, a trial-and-error strategy was implemented, thanks to which it was possible to identify the best configuration taking into account the tradeoff between sensitivity and specificity while looking at the balanced_accuracy:

$$Sensitivy = \frac{TP}{TP + FN}, \qquad Specificity = \frac{TN}{TN + FP}$$

$$BalncedAccuracy = \frac{Sensitivity + Specificity}{2}$$

With TP = malignant tumors correctly classified, TN = benign tumors correctly classified, FP = benign tumors incorrectly classified, and FN = malignant tumors incorrectly classified.

Table 13 provides a schematic of the above to summarize the used strategies for each classifier and input data type.

Table 13 - Classifier parameters for the training phase

Classifier	Description
RF	*Up to 1000 predictors, with a minimum number of splits of 2, based on Gini impurity criterion for split.*
KNN	*K* ∈ [2,8]*, Euclidean distance used as weight for the query result.*
SVM	*Linear, polynomial, radial basis function, and sigmoid kernels used, for the polynomial the grade varied from 2 up to 5.*
ANN	*Number of nodes decided using the Formula above, the input nodes depend on the number of input features and the number of output nodes is 1, because it is a binary classification task.*

In Figure 19 it is shown in detail the workflow of the implemented application.

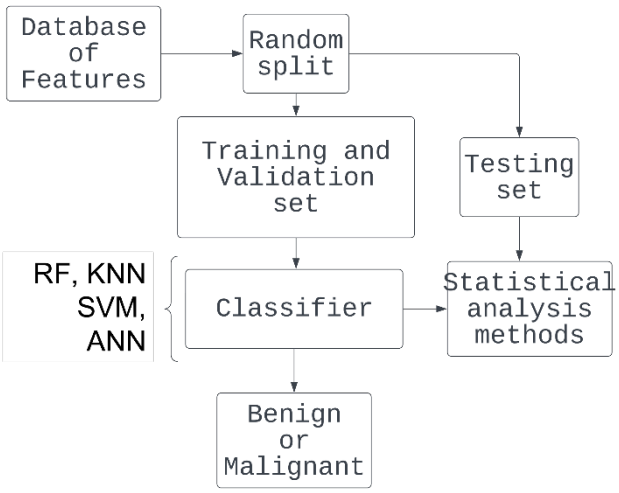


Figure 19 - Classification workflow

Once the various predictors were trained, a statistical analysis of the metrics of interest was performed. Specifically, the most common metrics, i.e. accuracy and precision, were studied, and in addition, sensitivity and specificity were analyzed, as the former allows to get a sense of the predictor's ability to be able to correctly identify malignant tumor [263], the latter refers to the classifier's ability to correctly discard benign tumors. A sensitivity of 100 percent ensures certainty in the case of negativity that a patient is "healthy", conversely a specificity of 100 percent in the case of positivity ensures certainty that the patient is "sick". The goal to be pursued in this study is to seek the best possible trade-off between sensitivity and specificity while

trying to obtain a sensitivity that is as close to 100% as possible. This means that the solution sought will have the fewest number of malignant cases misclassified as benign at the expense of benign cases identified as malignant. In other words the goal is not to misclassify a malignant case, since on a practical level it would bring the greatest disadvantage to the patient.

5.2.4. Results

Below are the results obtained through the use of the previously described approaches. All reported numerical results refer to the testing set.

Regarding the use of radiomic features, in particular the case where total tumor segmentation is used, Table 14 shows the best results obtained by the individual classifiers, specifying for each the number of features selected through the use of RFE. In particular, it is possible to see that in terms of sensitivity classifiers such as K-NN, SVM, and RF achieve better results than ANN, reaching about 96% at the expense, however, of specificity, in which only the ANN classifier is able to reach 58.82% with a sensitivity of 87.84%. In terms of accuracy the various classifiers all stand more or less in the same range between 81% and 85%, and in terms of precision only the ANN manages to reach 90%. Finally, considering the metric related to balanced accuracy we see that the best still turns out to be ANN reaching 73.31%.

Table 14 - Performance of the classifiers used with the radiomic features of the testing set, considering the approach with complete tumor segmentation

Classifier	***RFE***	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Balanced Accuracy**
K-NN (K=4)	20	94.59%	41.18%	**84.62%**	87.50%	67.89%
SVM (Linear)	43	94.52%	29.41%	82.22%	85.19%	61.97%
RF	7	**95.95%**	17.65%	81.32%	83.53%	56.80%
ANN	40	87.84%	**58.82%**	82.42%	**90.28%**	**73.31%**

Whereas, considering using only the segmentation relative to the slice where the area is largest in the three dimensions, the best obtained results are shown in Table 15. As in the previous case K-NN, SVM and RF achieve better results in terms of sensitivity than ANN, with results ranging from 87% up to about 95 percent. For specificity there is a maximum of 56.25% achieved by ANN. On the other hand, with regard to the other metrics these do not turn out to be so distant from each other, in particular it can be seen that compared to the previous approach there is a general decline in performance, except for RF which achieves better performance than it did before, while still remaining at a lower level than the best model in the previous case.

Table 15 - Performance of the classifiers used with the radiomic features of the testing set related to the approach with single slices of maximum area in the three axes

Classifier	***RFE***	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Balanced Accuracy**
K-NN (K=2)	9	87.01%	43.75%	79.57%	88.16%	65.38%
SVM (Linear)	39	93.51%	12.50%	79.57%	83.72%	53.00%
RF	31	**94.81%**	31.25%	**83.87%**	86.91%	63.03%

ANN	32	80.26%	**56.25%**	76.09%	**89.71%**	**68.26%**

Finally, the results obtained from the approach using deep features are discussed. It is possible to see by analyzing Table 16 how in this case sensitivity results exceed 94% for both KNN, RF and ANN, specificity peaking with ANN at 52.94%, and accuracy and precision between 78% and 87% the former and between 85% and 90% the latter. Finally, balanced accuracy peaks with ANN at 73.77%.

Table 16 - Performance of classifiers using deep features obtained from the proposed approach

Classifier	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Balanced Accuracy**
K-NN (K=6)	95.95%	31.25%	84.44%	86.59%	63.60%
SVM (Linear)	86.49%	43.75%	78.89%	87.67%	65.12%
RF	**97.26%**	29.41%	84.44%	85.54%	63.34%
ANN	94.59%	**52.94%**	**86.84%**	**89.74%**	**73.77%**

5.2.5. Discussion

To further discuss the proposed solutions Table 17 shows all the metrics regarding the performance of the best solutions obtained, taking into consideration the trade-off between sensitivity and specificity, thus selecting the best classifiers on the basis of the maximum value of balanced accuracy.

Table 17 - Best performance for each of the tested methods, selecting the classifiers with the best balanced accuracy for the three types of features considered: for all three cases this was found to be the ANN classifier

Method	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Balanced Accuracy**
Radiomic feature - standard procedure	87.84%	**58.82%**	82.42%	**90.28%**	73.31%
Radiomic feature - maximum tumor size	80.26%	56.25%	76.09%	89.71%	68.26%
Deep feature	**94.59%**	52.94%	**86.84%**	89.74%	**73.77%**

From the table it is possible to compare the two solutions that exploit radiomics: it can be seen that the use of all spatial information can result in increased performance for all metrics, especially if we consider

sensitivity and accuracy with more than 6 percentage points for both metrics. This result emphasizes the importance of features derived from the entirety of the analyzed tumor mass, which is lost in the second radiomic approach where only the planar information is considered. Comparing the results obtained by radiomic features and deep features it is possible to note some important aspects: i) taking into account the trade-off between sensitivity and specificity in the case of deep features the best balanced accuracy, of 73.77%, is obtained, taking more into account sensitivity, than specificity, reaching 94.59% and 52.94%, respectively. Which, on the other hand, is not the case in the two previous cases as in the best of the two a maximum of 87.84% sensitivity is achieved and specificity peaks at 58.82%. So, if we take into consideration the use of balanced accuracy as a tool to realize the trade off between sensitivity and specificity we have that the best approaches, with directly comparable results, are the use of deep features and the use of radiomic features extracted with the first approach described.

Objective comparisons with the other state-of-the-art approaches are difficult to make due to the diversity of the datasets used, both in terms of numerosity and from the point of view of the representativeness of the tumor kinds. Also to be considered is the fact that the numerical results in the state of the art, Table 11, can also be calculated by making various considerations depending on how the data were used. Notwithstanding this, the results can be compared by checking if at least the final metrics follow a similar behaviour, for example with the solutions using a workflow similar to the one proposed. In Table 18, the metrics for [252] and our solution are shown. The table shows all the metrics available, and it is possible to see how the behaviour is similar between the two methods, with sensitivity being the metric with the highest value and specificity the worst. It is also possible to see that our approach obtains better values in most of the metrics, with the exception of specificity is comparable. Considering the numerosity of the two datasets used, it is possible to see that in both cases the number of oncocytomas is significantly lower than in ccRCC, thus most likely being the cause of the low specificity value in both cases.

Table 18 - Metrics of one of the best approach in literature and of the best approach realized in this work

Method	**Metrics**						
	Sensitivity	***Specificity***	***Accuracy***	***F-Score***	***AUC***	***PPV***	***NPV***
[252]	88.30%	52.90%	75.40%	-	-	80.30%	-
Our	94.59%	52.94%	86.84%	93.58%	71.34%	89.74%	69.23%

A major problem that was addressed during the training of the classifiers was the imbalance of the available dataset, with the number of ccRCC cases being about 4:1 in proportion to oncocytomas, which was solved through the use of an appropriate weight for the elements belonging to the various classes. Indeed, one of the main future developments will be to try to reduce this imbalance and increase the number of cases available as much as possible so that it will be more statistically significant. In addition, with the introduction of new cases, it is expected that the deep features-based solution, as it is based on feature extraction exclusively based on deep learning, will be able to generalize more across the two tumor types and thus be able to achieve higher results than currently achieved. Other future developments will focus on using this type of approach for a larger number of renal tumor types, or considering using other deep neural network variants that can identify other tumors. In addition, one can also consider mixing radiomics with deep learning approaches, or approaches in which one previously demarcates the area of interest to use convolutional neural networks only on the marked input with the aim of directly classifying the nature of the tumor.

5.2.6. Final Remarks

In this work, a comparison was made on the use of two distinct types of features, radiomic and deep, obtained by automated procedures, with the aim of being able to build a tool that can provide support at the medical diagnostic level. Specifically with regard to radiomic features, two methods were tested that take inspiration from the state of the art. Instead, for deep features, a new method of extraction from patient CTs was proposed with an approach based on the exclusive use of a deep neural network (nnU-net) capable of directly segmenting diagnostic images, minimizing the human intervention required to identify areas of interest within the diagnostic type images. Comparing the various features is not only useful in understanding the feasibility of using tools based on them, but also in testing the actual validity of the proposed new method. To do this, once the features for the available CTs were obtained, performance was compared through the use of the most common classification methods (RF, KNN, SVM, ANN). In particular, it was possible to note that the use of radiomic features with classifiers such as RF, KNN, and SVM allows us to maintain a sensitivity in most cases better than ANN, at the expense, however, of specificity, which is why considering balanced accuracy, ANN turns out to be the best for both approaches. On the other hand, if we consider all the proposed features, the best results are obtained by ANN with 73.77% balanced accuracy, 94.59% sensitivity, 52.94% specificity and 86.84% accuracy.

Taking into consideration all the results obtained and especially taking into account that the best metrics obtained for deep features are similar to or better than those obtained by exploiting the radiomic features approaches, it is possible to conclude that the deep features obtained according to the method presented in this article are valid for differentiating renal tumors, especially in cases where only the CT of a patient with renal cancer is available.

6. Artificial Intelligence for plastic surgery – case study autologous ear reconstruction

This case study concentrates mainly on the realization of AI-based tools in order to realize a better workflow starting from another already existing and created inside the T3Ddy lab [24]. For this reason, it is focused mainly on the automatization of two main task that usually requires external interaction performed by an expert engineer. First an introduction to the specific case study analyzed is reported, then the two tools realized are presented, which are one for the generation of the ear depth map starting from images and the second one is for the automatic segmentation of the ear components. More in particular both the tools will be analyzed following the proposed framework, so the specific task is identified, with its corresponding metric and its specific data type. Then for each one the data gathered will be presented and where it is significant the corresponding processing will be described. Finally for both the model proposed and the relative results will be displayed and discussed.

6.1. Autologous Ear Reconstruction

In recent years, personalized medicine has rapidly transformed the healthcare sector by overcoming the paradigm of standardized medicine. This approach offers treatments built on the specific characteristics of each patient, which has proven to make treatment more effective and outcome more predictable in many medical fields [264]. In particular, the surgical field has significantly benefited from the creation of medical devices tailored to each patient using modern 3D modeling and manufacturing techniques [265]. In particular, autologous ear reconstruction (AER) surgical procedure is also taking advantage of these technologies in order to personalize and improve the treatment. The AER procedure is performed in case of total or partial absence of the external ear and involves: 1) the harvest of cartilaginous material from the costal region of the patient 2) the manual modeling of the tissue in order to create a framework that reproduces the features of the ear and 3) the insertion of this framework in correspondence of the pathological auricular region in a subcutaneous pocket [266]. To ensure the harmony of the patient's face after surgery, the framework is modelled using the contralateral mirrored ear as guide. The procedure is very complex for the surgeon who has to manually model the cartilages to reproduce the auricular anatomy and the final outcome is therefore strongly related to the "artistic skills" of the physician performing the surgery. For this reason, in the last years several techniques are being investigated to realize surgical aids able to support and help the physician in this procedure [267]. Recently 3D acquisition and CAD-based modeling approaches have allowed great progress in the realization of guides for AER outperforming the results obtained using two-dimensional aids [268, 269].

The AER surgical procedure involves the realization of a 3D structure of the ear, obtained by carving and sculpting the patient's costal cartilage, and its implant in a subcutaneous pocket located in the auricular pathological region. According to the technique proposed in [266], the auricular elements to be reconstructed are helix, antihelix, tragus-antitragus (colored elements in Figure 20), plus a support base.

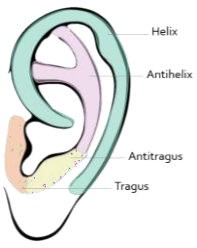


Figure 20 - Anatomical elements of the ear.

The result of the surgery, however, strongly depends on the surgeon's manual skills in modeling auricular 3D geometries. The surgery aims at achieving a result that ensures symmetry of the face with the contralateral ear, but the operation is a real challenge for the surgeon since the geometry of the ear is actually very difficult to reproduce. In order to help the surgeon in this procedure it had been studied and realized 3D printed

surgical guides that provide a simplified representation of the patient's ear anatomy and guide the surgeon in cutting the different anatomical elements. An example of the cutting guides is shown in Figure 21. The realization of the patient-specific medical devices involves the acquisition of the 3D anatomy of the healthy ear (with optical scanning techniques or from CT scan), a mirroring operation to obtain the target anatomy to be reconstructed, and the CAD modeling with appropriate modeling tools.

Figure 21 - Example of CAD models of the surgical guides created by simplifying the original anatomy.

In detail, the CAD procedure is performed on the correctly oriented 3D model, i.e. the ear must be oriented in such a way that all the elements involved in the reconstruction are visible to the user view point (coincident with one of the global reference system planes e.g. the XY-plane), then the procedure takes place on the 2D sketch created on such plane (example in Figure 22). On the so defined sketch, through well-known CAD operations, the printable models of the surgical guides are created.

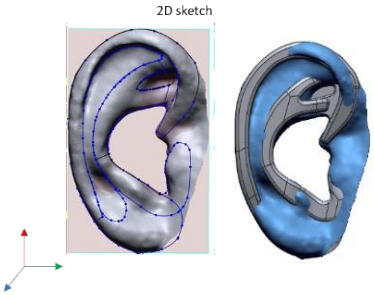


Figure 22 - Main phases of the manual modelling procedure of the surgical guides. a) initial anatomy orientation; b) result of the cad modelling procedure.

With the aim of making the physician autonomous in the realization of patient-specific cutting guides for AER surgery, the research vision is to develop software tools for designing the semi-automatic design of the medical devices' CAD models. The development of tools easily manageable at hospital level would allow to realize a streamlined and fast production process, to be used within the common clinical practice. To this end, through a software routine, the CAD modeling of the surgical guides was automated [270]. In particular, two new automatizations have been proposed, the first one based on the necessity of a 3D model to obtain depth information, and the second one based on the fact that the algorithm requires the contours of each anatomical element as input.

6.2. Ear depth map generation

With the idea of making the process of acquisition and realization of the depth map more accessible to non-expert users such as hospital staff, this work focuses on the development of a system based on deep learning techniques that, through a single RGB image of the ear, can create the depth image which by definition is an image that contains information about the distance of the surfaces of objects in the scene from a point of view.

Searching for geometric relationships between objects in the scene by using neural network-based approaches is a topic extensively investigated in the literature [271, 272]. This is because depth perception is a key component for example for autonomous systems that interact in the real world, such as warehouse robots, self-driving cars, etc. [272]. Many systems study the arrangement of objects in the scene, but they are usually applied to external scenario contexts that require a much lower resolution compared to the one required for the application under consideration. For this reason, the systems available in the literature are not directly applicable for the realization of depth maps of the auricular anatomy. Accordingly, the present work focuses on the study of alternative approaches suitable to create depth maps with higher resolution. In addition, to make the realization of the depth map as simple as possible this work focuses on approaches based on monocular depth estimation [273, 274], in which the depth image is created from a single photo. This aspect is of considerable relevance if we contextualize the acquisition of the ear on pediatric subjects (generally uncooperative) and in a hospital environment (not always equipped with modern acquisition technologies).

The devised approach, detailed in the following chapters, takes as input the side picture of the face of subject and returns the depth map of the ear using three different architectures: 1) a Faster Region-based Convolutional Neural Network (R-CNN) [272] that has the task of isolating the ear region from the face image; 2) a cycle Generative Adversarial Network (GAN)[275] that takes as input the cropped anatomical region and creates the depth map; 3) a semantic-aware approach [276] to refine the depth map created in the previous step. If on the one hand the use of Faster R-CNN is well established for this type of task (object recognition), to determine the architecture for the creation of the depth map it was necessary to study the recent literature in order to identify the best approach. The cycle GAN architecture was identified as the best choice as it supports relatively high resolutions compared to other approaches [274], not negligible aspect for the success of the workflow.

6.2.1. Data description

The present work is intended to use machine-learning systems for the creation of a depth map of a human ear from a single RGB image with a level of accuracy acceptable for the creation of custom surgical guides for auricular reconstruction.

In order to train and test the system, a subset of the ND-Collection F [277] and ND-Collection G [278] made available by the University of Notre Dame was employed, containing depth maps and corresponding 2D images of 302 and 235 human subject profiles, respectively.

6.2.2. Model description

Considering the difficulty of the task it has been decided to create a customized workflow to reach the final goal and Figure 23 shows the general structure defined to create the depth map of the ear region from a profile photo.

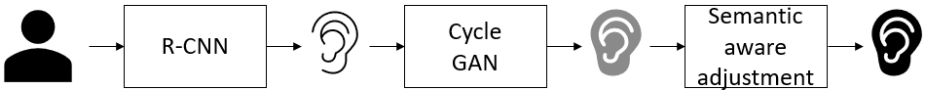


Figure 23 - General structure diagram: profile image of a person given as input to the R-CNN that extracts the ear, that it is used by the Cycle GAN to make the depth map, which is finally adjusted using a semantic-aware module.

All the architectures used are explained in more detail following the order presented in the workflow.

6.2.2.1. Faster R-CNN

The first step of the workflow involves the extraction of the ear region from the RGB images and from the depth map images in order to use all the resources on the creation of the ear depth map without considering the remaining portion of the scene. To do this, the algorithm is based on the use of a R-CNN, which was trained specifically for the detection of ears, implementing procedures known in the literature to train this type of network to specific tasks [279].

R-CNNs are a class of techniques that share the common trait of using deep models to pursue object detection. An R-CNN model first selects several proposed regions from an image and then labels their categories and bounding boxes. Then, a CNN is used to perform features extraction for each region and use them to predict the regions' categories. In our case, a Faster R-CNN Inception v2 [253], with the architecture illustrated in Figure 24, was used; the specific CNN architecture used is the Inception v2. This paper does not report implementation details of these architectures as they are well known in the literature.

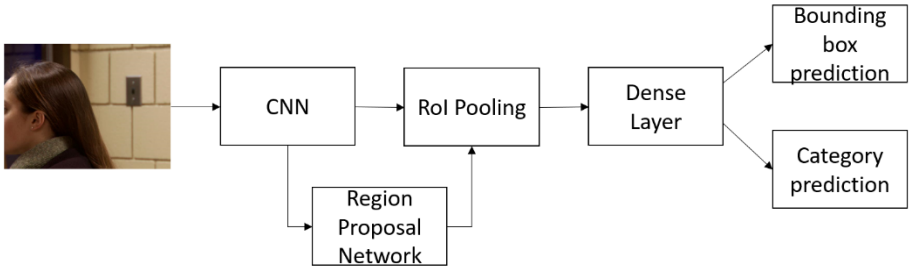


Figure 24 - Faster R-CNN for ear detection

6.2.2.2. Cycle GAN

The region of the ear cropped from the original image is used to train a cycle GAN that has the task of creating the depth image from the single photo. The GAN is a class of machine learning frameworks, in which two neural networks contest with each other in a game (zero-sum game) [280]. The cycle GAN, depicted in Figure 25, is a variation that combines the properties of condition constraints and cycle consistencies to use the GAN architecture in Image-to-Image translation task effectively.

Cycle GANs can be used to alter a given input image to a distribution of the target domain. Specifically, an image is taken from input domain X and transformed into an image of the target domain Y without a one-to-one mapping between the images. A generator is used to map the image to the target domain and the quality of the result image is checked using a discriminator which pushes the generator to perform better.

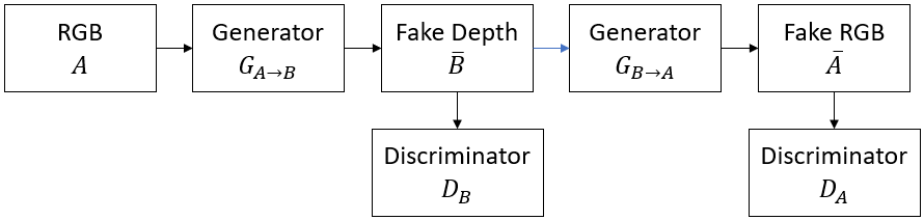


Figure 25 - Simplified scheme of a cycle GAN architecture.

6.2.2.3. Depth Map Refinement

Finally, to increase the definition and accuracy of the generated depth maps, a semantic-aware module is used, which receives as input both the RGB image and the estimated depth map. The module uses the RGB image to get the semantic context, that is used to create several masks which are used to perform a max pooling operation over the pixels inside the same mask within the estimated depth map. This approach has already been used in the literature to correct errors generated in depth images [276].

Figure 26 shows an example of the entire process applied to a profile photo, showing the results for each step.

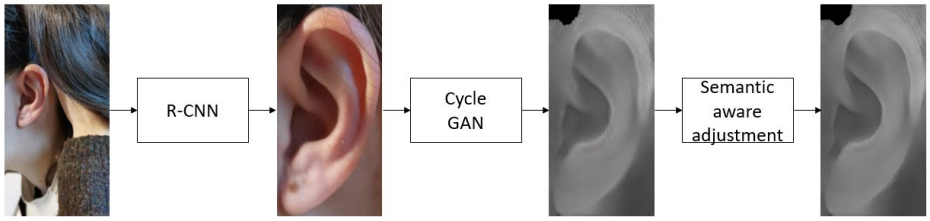


Figure 26 - An example of the output of each module.

6.2.3. Results

In this section the results obtained through the proposed architecture are shown, divided in two sections: the first one related to the ear detection task, and the second one related to the depth map generation from RGB images. For both tasks the dataset was randomly split into two parts, 70% for training, and 30% for testing.

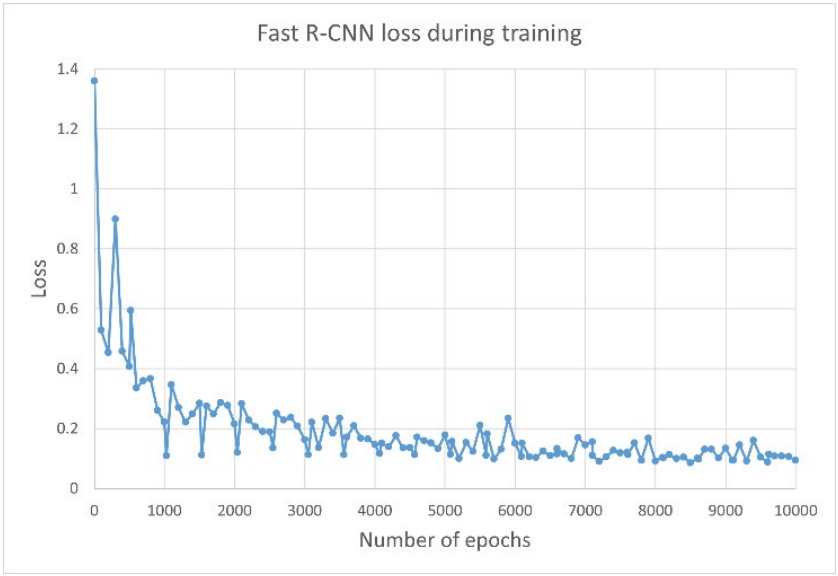
6.2.3.1. Faster R-CNN

The Faster R-CNN architecture was implemented exploiting the Tensorflow APIs [281] and the model was fine-tuned over the ear dataset. To generate the labels for the training, it was used the free software LabelImg [282].

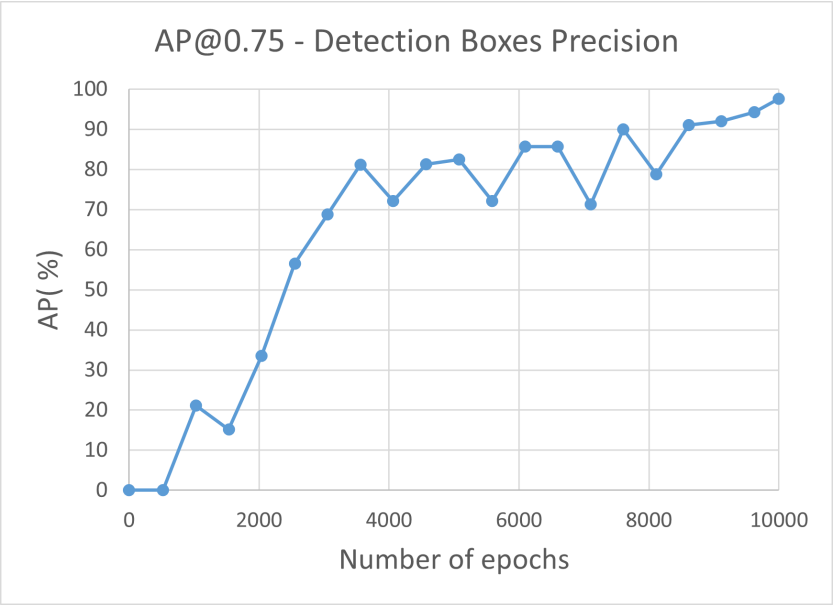
The obtained results are shown in Figure 27. The first graph shows how the loss of the network decreases with the increasing number of training epochs. In the second graph is represented the function AP@0.75IOU [283]. The AP represents the Average Precision (AP) of the network, i.e. the average of the precision scores. The AP@0.75IOU corresponds to the AP considering only the bounding boxes for which the Intersection Over Union (IOU) is greater than 75%, where IOU is computed as follow:

$$IOU = \frac{Area(B_{predicted} \cap B_{ground\_truth})}{Area(B_{predicted} \cup B_{ground\_truth})}$$

At the end of the overall training process an AP of 97.63% is reached over the testing set.



(a)



(b)

Figure 27 - Graph of the loss (a) and of the average precision (b) achieved by the Faster R-CNN.

Figure 28 shows an example of the output at this stage where, from the profile, the ear region is detected.

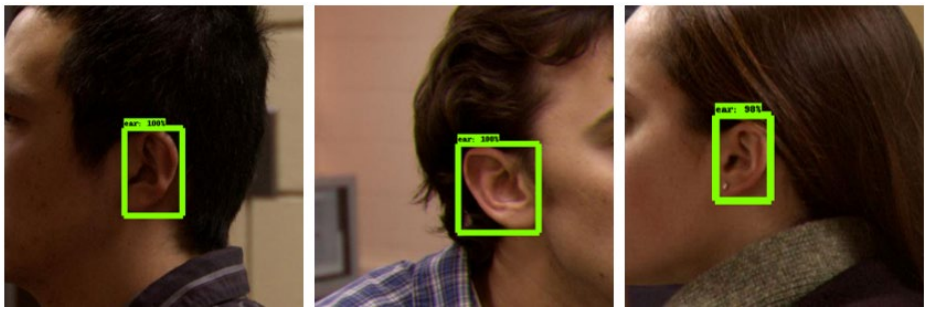


Figure 28 - Output of the Faster R-CNN module.

6.2.3.2. Cycle GAN and Depth Map Refinement

The architecture of the cycle GAN was realized by exploiting a standard PyTorch implementation of the 4 necessary components, the two Generators and the two Discriminants. The Adam optimizer was exploited as suggested by Goodfellow et al. [280]. Regarding GAN results, it is quite complex to evaluate the accuracy of the generated images since there is no standard measure described in literature [284].

Considering the difficulty in evaluating the results of GAN architectures, in this work it was decided to use two standard metrics, the mean square error (MSE) and the structural similarity (SSIM), often used to assess the quality of images after they have undergone changes. The MSE indicates the cumulative error between the generated image and the original image therefore lower values indicate more similar images [285]. SSIM in recent years has become an accepted standard among image quality metrics. This technique evaluates the visual impact of changes in image luminance, contrast and structure [286], and can assume values in the range [0,1]. MSE and SSIM values were evaluated comparing the original image and the output image at the final step, i.e. after the depth map refinement with the semantic-aware adjustment module implemented using the code available in [276].

Figure 29 shows a qualitative comparison of the output and original depth maps. The MSE and SSIM values were averaged over the test data resulting in an average MSE of ~0.07 and an average SSIM of ~0.80.

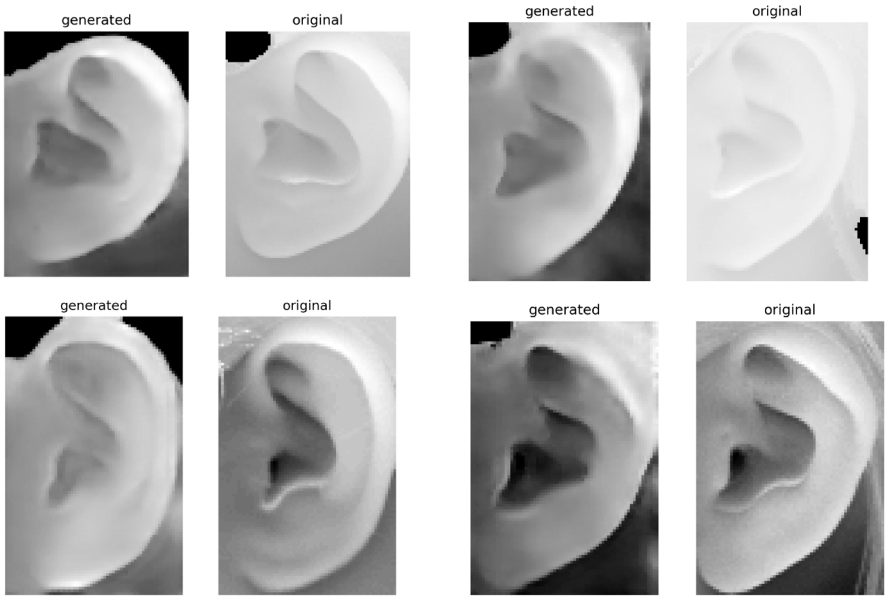


Figure 29 - A subset of the depth map generated on the test set.

6.2.4. Discussions

Autologous ear reconstruction has been the subject of several studies in the literature involving medical and engineering fields. The great interest is due on one hand to the importance from the psychological point of view of a satisfactory outcome for the patient, on the other hand to the intrinsic difficulty in the manual construction of the ear reported by surgeons. Work is being carried out on the development of a user friendly and fast system for the fabrication of patient-specific devices that can assist the physician during surgery. In this context, this work aimed to define a system to create depth map images of the patient's ear from a single profile photo. The analysis of the literature revealed the lack of approaches that can provide a sufficiently detailed depth map. In this work the most suitable architectures for the purpose were first identified, they were then implemented and trained on a specific dataset and the interfaces between the different architectures were defined in order to obtain an efficient workflow. The results obtained for the single steps were shown and the quality of the result was quantitatively evaluated in terms of MSE and SSIM, obtaining satisfying results considered suitable for the final application. The implemented system constitutes a disruptive innovation that considerably streamlines the workflow currently foreseen for the realization of surgical guides. Future developments envisage the evaluation of the depth maps produced by the network by performing an acquisition campaign on an adequate number of subjects, reconstructing the corresponding point clouds from the depth images by controlling the intrinsic acquisition parameters.

6.3. Ear components identification

The first AI-based automation is a 2D segmentation algorithm starting from the depth map of the ear, obtained from the correctly oriented 3D model. As well-known the depth map is an image that contains information relating to the distance of the surfaces of scene objects from a viewpoint. Therefore, the proposed algorithm exploits the 2D characteristics of the depth map, which contains, in its definition, depth information defined by the 3D model.

In this perspective, a first segmentation software was proposed by [287], where state-of-the-art segmentation algorithms based on image processing techniques were analyzed, without being able to find a suitable algorithm to perform segmentation of ear elements. As a result, a very accurate ad-hoc algorithm based on image processing techniques was developed, which however requires setting some initial parameters. To overcome this shortfall, the feasibility of using deep learning techniques, specifically the U-Net architecture, for the ear segmentation is evaluated.

6.3.1. Data description

The dataset for training and testing the network consists of 131 ear depth maps. To create the dataset were used 62 computerized tomographies of the head (CTs) and 18 ear scans (the number of retrieved scans exceeds 131 since not all CT scans allow both left and right ear anatomies to be used due to the patient's position during the scan or to congenital malformations of ear). In detail, the anonymized CT scans were retrieved from the CQ500 dataset [288] and further processed with the Mimics Materialise software package [289] to obtain the 3D model of the ears. The remaining 18 ear models were already in 3D form as they were obtained with a professional 3D scanner. The dataset can be considered heterogeneous since the ear constitutes a biometric element whose shape and size are independent of age, gender, and ethnic group [287]. Starting from such three-dimensional models of the ear, to create the depth maps of the dataset was used the algorithm implemented in [287] able to properly orient the ear and create the depth map. The depth maps were then manually annotated in collaboration with a physician. For this feasibility study, it was decided to combine in a single element i) the tragus with the antitragus elements and ii) the antihelix with the triangular fossa and the root of the helix. Moreover, to implement a more comprehensive algorithm to be used in a variety of applications, the concha element (blue in Figure 30) is also considered. Figure 30 shows the target segmentation of the anatomical elements and an example of manual annotation.

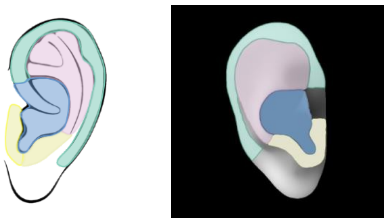


Figure 30 – In the left picture the ear elements definition and in the right an example ear manual segmentation.

6.3.2. Model architecture

The neural network model chosen for the ear segmentation is based on the U-Net architecture [102].

A U-Net consists of an encoder (downsampler) and a decoder (upsampler); in fact the original architecture of the network consists of a contraction path and an expansion path. As for the contraction it follows the typical architecture of convolutional neural networks, i.e. repeated application of two 3x3 convolutions followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At this stage, each downsampling step doubles the number of channels. The expansion path consists of an upsampling of feature maps followed by a 2x2 convolution which halves the number of channels, a concatenation with the corresponding feature map of the contracting path, and two 3x3 convolutions followed by a ReLU. As a final layer, a 1x1 convolution is used to map each component feature vector to the desired number of classes. The

U-Net also provides skip connections in the encoder decoder architecture, this way fine-grained details can be retrieved in the prediction.

In this work, a modified version of the standard U-Net was used. As said, in the first half of the network the characteristics of the input images are extracted using the encoder. Since the task of the encoder is to extract generic characteristics, the initial learning phase based on random input parameters can be replaced with a pre-trained model in order to learn robust features and reduce the number of trainable parameters. More precisely, the intermediate layers outputs of a pre-trained MobileNetV2 model [290] are used as the encoder. As for the decoder, it follows the general structure of the original U-Net. The final transposed convolution has six output channels, since there are six possible labels for each pixel, corresponding to the four anatomical elements (see Figure 20), the rest of the ear anatomy and the background. The network architecture is shown in Figure 31. As far as training is concerned, taking advantage of the use of already pre-trained levels, it is necessary to train only the decoder and the final classifier levels. Moreover, a data augmentation process was applied on the dataset by mirroring all training images and changing the image brightness. The orientation variations were not tested since the segmentation task is embedded within a workflow that provides for robust automatic orientation of the 3D model prior to creating the depth map, thus resulting in the standardization of the orientation of the images to be segmented. Taking advantage of transfer learning and data augmentation, it is possible to obtain good results by using a reduced number of input data.

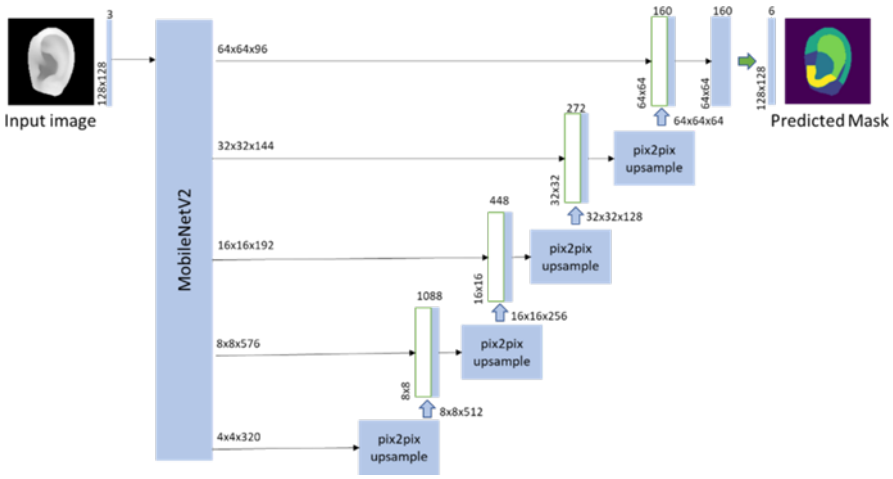
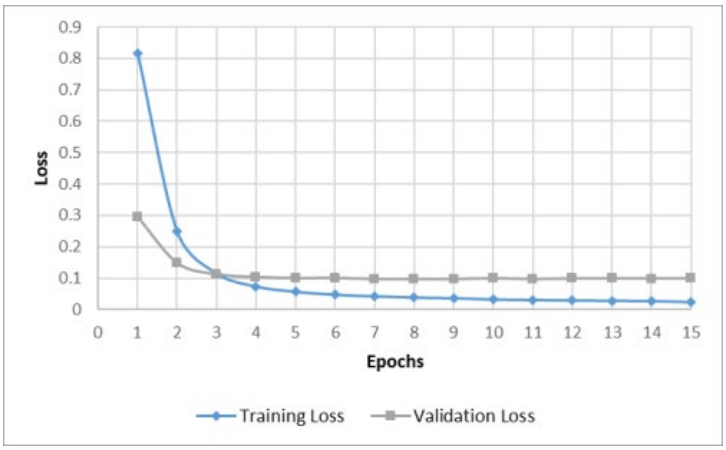


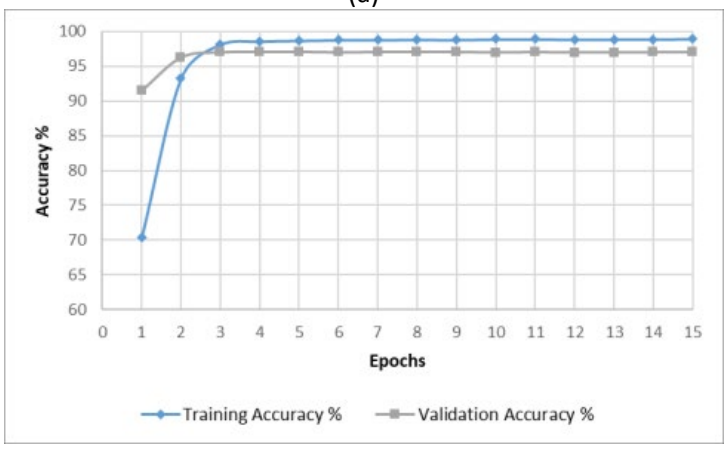
Figure 31 - Architecture of the implemented network: U-Net model with MobileNetV2 backbone in order to obtain better results with a lower number of training samples.

6.3.3. Results and discussion

The neural network was implemented with Tensorflow [291] using the high-level API Keras. The Adam optimizer was used and as loss the sparse categorical crossentropy was evaluated, since there are more than two classes as network output. To train the network the dataset was divided as follows: 70% of the data are used as the training set, 15% composes the validation set, and the remaining 15% is used as testing set. The number of epochs was set to 15, considering that subsequently the accuracy of the network remains stable and the loss increases as shown in Figure 32. In particular, in Figure 32 is possible to observe the loss trend (a) and the accuracy trend (b) obtained during the 15 epochs of the training phase, both for the training set and the validation set. As can be seen in the graphs the network reaches an accuracy over 95% after few epochs on both sets, reaching finally an accuracy of about 99% on the training set and about 97% on the validation set. In Figure 32a it can be seen how the loss on the training set has a decreasing trend as the training epochs increase and how this does not happen in the validation set: for this reason, and to avoid overfitting, it is not necessary to carry out a higher number of training epochs.



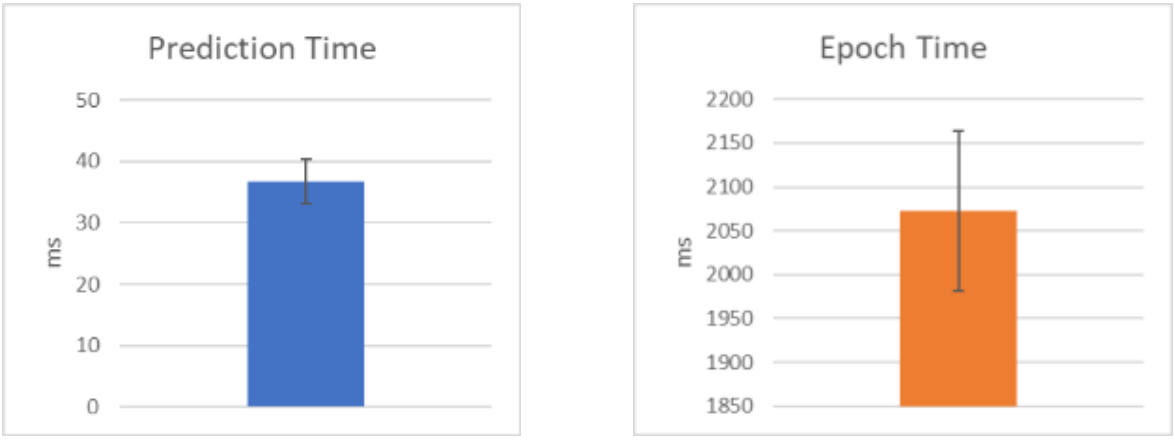
(a)



(b)

Figure 32 - a) loss and b) accuracy graphs for the training epochs.

The network was developed using Colab (a service provided by Google) as it offers many advantages including free access to GPUs so as not to overload personal machines, but with all the disadvantage of not being able to choose and therefore not having a fixed configuration of processing machine. For this reason with the aim of providing reference times, both in terms of training time and prediction time, a fixed configuration was also used. In particular, the machine has a Nvidia GeForce MX150 GPU. Figure 33 shows the average prediction time, calculated on all the dataset images,and the average time per epoch, calculated on 1000 epochs.



(a) (b)

Figure 33 - Average prediction time a) and average time for epoch b) obtain with Nvidia GeForce MX150 GPU.

Figure 34 shows some of the segmentations predicted by the network (predicted mask) compared with the corresponding ground truth (true mask).

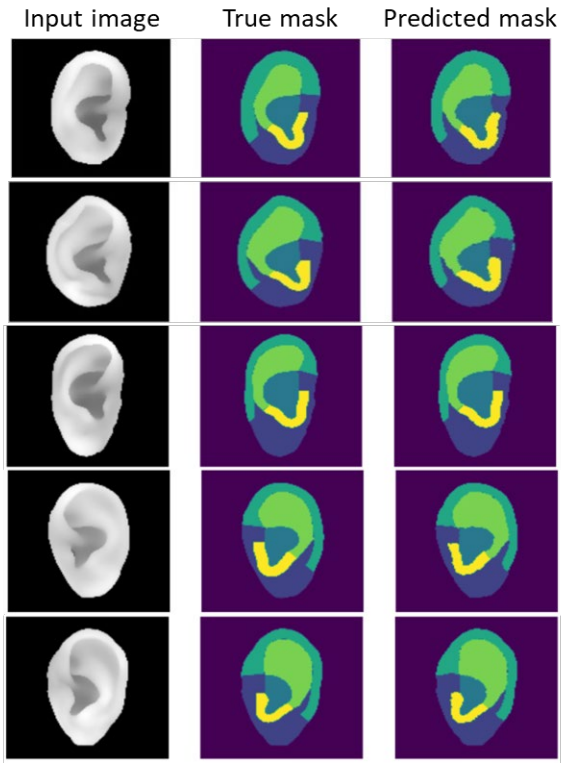


Figure 34 - Subset of network experimental results

The network is able to achieve excellent performance, as can be seen both from the accuracy values in Figure 32 and qualitatively in Figure 34. These results are in any case strongly related to the ear's position within the image: the correct segmentation is in fact strongly dependent on the correct orientation of the ear and, at the moment, the network is not able to produce correct segmentations with input scans not correctly oriented, both on the frontal plane and in space before creating the depth map.

6.3.4. Final remarks

The increasing use of artificial intelligence techniques in the medical field has allowed the development of fast, accurate, and reliable systems to support clinical practice, including the use of convolutional neural networks for automated medical image segmentation. This work demonstrates the potential of using a U-Net architecture for ear segmentation, which can recognize anatomical elements to assist with autologous ear reconstruction surgery. The network was trained on depth map images and achieved 97% accuracy in segmenting the ear into main components. In parallel, we developed an innovative system to create patient-specific ear depth maps from profile photos in order to streamline the surgical planning process. Quantitative evaluation showed the depth maps provide sufficient detail for the application. Future work will focus on extending the datasets to improve robustness, evaluating network performance on non-oriented ears, reconstructing 3D point clouds from the depth maps, and translating the systems to clinical use. The ultimate goal is to provide an integrated, fast, and user-friendly platform to design custom surgical guides that can assist physicians during autologous ear reconstruction.

7. Artificial intelligence for psychiatry – case study suicidal patients

Also this case study is developed inside the T3Dddy laboratory [24] and in this case the psychiatry department of the Meyer Hospital [23] is involved. The tool has been developed with the main objective of been an extension to the information obtainable from the standard psychology procedures. This means that the goal is to analyze the clinical features of the patients to increase the knowledge of the current patient state. The specific study is based on the identification of suicidal patients that have effectively tried to suicide and distinguish the ones that only faked to do it. This chapter similarly to the previously ones is structured based on the proposed framework, with the first part relative to the clinical phase and the latter one to the artificial intelligence engineering phase. So firstly the clinical scenario of this specific case study will be introduced to give the reader a little deeper understanding on what the problem is, then the specific task and data used are defined and finally all the artificial engineering phase is illustrated with a statistical analysis on the data available and all the part relative to model selection, training, evaluation and final testing is reported. Finally some remarks are made in the results and discussion subchapters.

7.1. Suicidal patients

The World Health Organization recognizes suicide as a critical public health issue [292]. Suicide is one of the leading causes of death across all age groups, but it is particularly important during developmental stages because the sharpest increase in suicide-related deaths happens during the transition from adolescence to early adulthood, and even in most subjects who attempt suicide later in life, suicidal thoughts and behaviors often begin before the age of 25 [293]. The prevalence of suicidal ideation and attempted suicide increases dramatically during adolescence [294] and it represents the second leading cause of death among youth of 10–19 years old [295].

Regarding the terminology used in the suicide literature, there is still no agreement. The most prevalent tendency is to label a variety of occurrences with a spectrum of symptom severity and sharing the same proximal and distal risk variables as suicidal behaviors and thoughts (SBTs) [294–296]. Although SBTs are closely related to one another, they differ in terms of lethality, strength of intention, and prevalence. Suicidal ideation is a risk factor for suicidal behavior later on: those who have previously entertained suicidal thoughts have a 12-fold increased chance of trying suicide by the time they are 30 [297]. Suicides can be prevented to some extent in any age group. Prevention techniques necessitate coordinated and integrated measures combining healthcare, education, politics, and the media due to their multifaceted etiopathogenesis [292].

WHO Member States have pledged to strive toward the global goal of lowering the suicide rate in nations by one-third by 2030 under the WHO Mental Health Action Plan 2013–2030 [298]. The WHO also emphasized the importance of identifying persons who are at risk, developing preventative plans, and putting certain therapeutic approaches into practice by identifying early indicators of suicide tendencies [292, 298].

Artificial intelligence has shown itself to be a successful method in recent years for automating the analysis of medical data, identifying suicidal behaviors among adolescents in mental hospitals [299] and extracting novel combinations of biomarkers helpful for early diagnosis [300–303].

One of the primary issues in suicidology in this scenario is identifying risk variables for the passage from suicidal thinking to attempted/failed suicide. Suicide and self-harming behaviors that are not meant to be fatal must be recognized from one another. These behaviors include, for instance, bravery exercises and, among minors, the generally prevalent practice of refraining from suicidal self-harm, which frequently functions to control negative emotions [304]. Moreover, there is a common propensity in clinical practice to differentiate between failed and attempted suicides. Attempted suicides are defined as those occurrences that, even in the presence of a proven suicidal purpose, do not result in death because low-cost method lethality is used unintentionally or because the subject ends the event before it is deadly owing to their own volition. However, despite confirmed suicidal intentionality, high lethality, method choice, and other factors,

failed suicides are defined as incidents that do not result in death due to chance circumstances, irrespective of the subject's decision [305] (e.g., third parties intervening).

In order to examine risk factors for SBTs and assess the distinctions between suicidal ideation and attempted/failed suicide in children and adolescents admitted to the Child and Adolescent Psychiatry Emergency Unit (CAPEU) of the Meyer Children's Hospital in Florence, this case study combined clinical evaluation, statistical analysis, and a neural network approach. A neural network technique can be used to predict or categorize a certain behavior, and the variability found in SBT patients may be utilized to train a multivariate model. One of the primary goals of this work is to develop an instrument that can detect individuals who are at-risk and provide a probabilistic indicator of the characteristics of a suicide episode.

7.2. Data gathering

For this specific case study, the data gathering procedure was approved by the Pediatric Ethics Committee of the Tuscany Region (number 112/2022). In this single-center retrospective observational cohort study, 237 patients who were hospitalized to the Meyer Children's Hospital's CAPEU for SBTs between January 1, 2016, and June 30, 2020, were included. The sequential patient enrollment reduced the selection bias [306]. The following two requirements had to be met for inclusion: (1) a mental diagnosis in line with the Diagnostic and Statistical Manual of Mental Disorders DSM-5 criteria [307]; and (2) the need for inpatient treatment for the participants. The exclusion criteria were intellectual impairment and neurological illnesses ranging from moderate to severe. As part of the study methodology, every participant completed a thorough diagnostic evaluation that collected psychopathological and sociodemographic data. From each patient's medical records (clinical software C7), we retrospectively gathered clinical data, including sex, date of birth, age at first hospital admission, ethnicity, family characteristics, type of school, academic performance, method, previous treatments, hospitalization outcome, and past stressful or traumatic events (STEs). A traumatic or stressful event that happens during childhood or adolescence and may have a detrimental effect on a person's neurobiological or psychosocial development is known as a STE [307, 308].

Furthermore, we split the sample into two groups according to the admission referral's justification. The first group, referred to as "suicidal volition patients," contained people with unstructured suicidal thoughts and had low-damaging SBTs. Those with suicidal intent and a high risk of injury were included in the second group, who were hospitalized for attempted or unsuccessful suicide (henceforth referred to as "suicidal motivation patients"). All investigations, including statistical and neural network ones, focused on this classification. The Kiddie Schedule for Affective Disorders and Schizophrenia-Present and Lifetime Interview [309] and the DSM-5 [310] served as the foundation for psychiatric diagnosis. SBT phenomena were assessed using the Italian version of the Columbia-Suicide Severity Rating Scale (C-SSRS), which is administered to psychiatric residents during the first two days of their admission to the mental hospital. We distinguished between the existence of suicidal thoughts, the degree of ideation, self-harming acts, and suicide attempts using the C-SSRS scores [311]. Based on actual fatality or medical harm, the lethality of suicide attempts was categorized as shown in Table 19.

Code	Lethality of suicide attempts
0	No suicidal ideation or suicidal behavior with no damage
1	Thoughts of death but not suicidal ideation and not suicidal behavior
2	Sporadic unstructured suicidal ideation or minor suicidal behavior, such as superficial self-cutting with minor physical damage (slight bleeding, scratching, bruising)
3	Unplanned suicidal ideation or persistent thoughts of death or suicidal behavior with moderate physical damage, need for medical attention (e.g. second-degree burns, major vessel bleeding)
4	Active suicidal ideation with some intent to act, without specific plan or preparatory acts or behavior (anything beyond verbalization or thought, like assembling the specific method (e.g. buying pills or a gun) or preparing for death by suicide (e.g. giving things away, writing suicide notes)

5	Active suicidal ideation with a specific plan and intent or suicide attempt with minor physical damage and medical hospitalization required
6	Repeated major self-injurious behaviors, suicide attempts with severe physical harm and repeated suicide attempts

Table 19 – Codification of lethality of the suicide attempts from the less severe (0), to the most lethal (6).

7.3. Statistical Analysis

Statistical analysis envisaged the extraction of absolute and relative frequencies for the categorical variables and the evaluation of mean and standard deviation for the numerical variables. Pearson's chi-squared test (χ^2) [312] was performed to assess the statistical significance of the observations. In particular, the independence or link between the available categorical data and the variable referral reason for admission was assessed. In order to pass the test, two-entry tables known as contingency tables between two variables must be created. In these tables, the number of cases with positive values of the two variables (joint frequencies) is recorded in the cells defined by the intersection of rows and columns.

The χ^2 test assumes as a null hypothesis that the two variables are statistically independent, whilst the alternative hypothesis is that there is a relationship between the two variables. The χ^2 test uses the χ^2 distribution to decide whether to reject the null hypothesis and is widely used to verify if the frequencies of the observed values fit the theoretical frequencies of a fixed probability distribution. The χ^2 test makes it possible to establish, after setting the maximum permissible error, whether discrepancies between observed and theoretical frequencies are due entirely to chance or are justified by statistical dependence. The tolerated error was set at 5% [313]. To carry out this test, each variable of interest was transformed into a dichotomous variable (e.g. the variable "Presence/absence of previous specialist care" was analyzed as "Caretaking YES/ Caretaking NO") and its dependence on the variable referral reason for admission was evaluated, which was also transformed into a dichotomous mode ("suicidal motivation patients" versus "suicidal volition patients").

7.4. Neural network approach

Statistical analysis evaluated the single variables independently and did not allow the creation of a predictive model that considers different factors at the same time and therefore predicts a possible suicidal event based on the available information. This failure could be overcome with the use of artificial neural networks, which had the advantage of not requiring the assumption of a linear relationship between values and could be used for modeling prediction problems that had several characteristics following undefined functions [314, 315].

A neural network is an interconnected system of perceptron. The perceptron is a type of binary classifier that maps its inputs into an output value calculated with an activation function (χ) that evaluates the scalar product between the input vector (x) and a vector of weights (w) added to a constant bias value (b). The activation formula is therefore:

$$f(x) = \chi\,(\ \langle \mathrm{w}, \mathrm{x} \rangle\ + \mathrm{b})$$

By modifying the vector of the weights w through a specific learning algorithm, it is possible to modulate the output of a perceptron, with the aim of obtaining learning properties. Several layers of perceptrons form a neural network.

Preparation of data. The preparation of data is a crucial step for the success of the system as it can strongly influence the results of the analysis and the simplicity of the data management. The dataset based on the medical records at the CAPEU of the Meyer Children's Hospital contains 53 variables, including numerical data (e.g. age) and categorical features (e.g. nationality), that is variables represented by labels with predefined values. In this work the pre-processing of data involved: (1) normalization of numerical data in the interval $[-1, 1]$, with the aim of avoiding features with wider ranges weighing more on the output; (2) one-hot encoding of categorical features (for each example all bits are set to 0 except one, which indicates the category to which it belongs) was applied to avoid the forced creation of ordinality relations where these do

not exist; (3) dichotomous variables that assumed value $\{0,1\}$ were remapped to $[-0.5, 0.5]$ to avoid the risk of setting some weights to zero.

Feature selection. The size of the dataset after the above-mentioned preprocessing required a feature selection phase to reduce the network inputs. In general, feature reduction processes are applied to avoid resource consumption by weak features, optimize model performance by avoiding the noise generated by unnecessary fields, and generally identify the strongest predictors. A widely used technique for this task is recursive feature elimination (RFE) [316]. Given an external estimator that assigns weights to features, and a desired number of features to select, the goal of feature selection is to select features by recursively considering smaller and smaller sets of features. The estimator chosen in this work is Random Forest Classifier [317] a machine learning method that generally works well with high-density problems and allows non-linear relationships between predictors. In our work, the desired number of features was fixed at 30. Minimizing the number of features is a practice generally executed to facilitate the learning process, but at the same time, it is necessary to avoid reducing the amount of input information too much. In the present study, the value of 30 features was empirically found, as the minimum number of features required by the network to maximize accuracy.

Model architecture. As mentioned above, a neural network is a multilayer system of perceptrons, the choice of the number of neurons in the hidden layers is a very important part of the decision on the overall architecture of the neural network. Using too few neurons in the hidden layers causes so-called underfitting, i.e. when there are too few neurons in the hidden layers to adequately detect signals in a complicated data set. Using too many neurons in the hidden layers can cause several problems. First, it can cause overfitting, i.e. when the neural network has such a high information processing capacity that the limited amount of information in the training set is not enough to train all the neurons in the hidden layers. An excessive number of neurons in the hidden layers may also increase the time required to train the network. A trade-off must be reached between too many and too few neurons in the hidden layers. There are many rules of thumb for determining the correct number of neurons to use in the hidden layers [314]. However, choosing an architecture for the neural network usually comes down to trial and error. The choice of the number of neurons during the neural network's trial and error process can be guided by some considerations regarding the network's ability to learn the complexity of the problem. Also, for the learning algorithm that defines the data model by changing the weights, an empirical procedure is generally used [314].

Iterative empirical tests were performed to identify the final and optimized configuration of the neural network by varying both the number of neurons per layer and the number of hidden layers. The best configuration consists of a single layer with 15 neurons and uses the Adam Optimizer [86]. In this research, neural networks were used as a tool to predict the type of suicidal behavior ("suicidal motivation patients" versus "suicidal volition patients"), and therefore the variable referral reason to admission was considered as the single variable in the output layer. Model evaluation. The loss graph and the accuracy parameter were used to assess the learning ability and accuracy of the network. The loss value indicates the optimization error, so a gradual decrease in this function is to be expected. The accuracy of a model is usually determined after the model parameters are learned and fixed and no learning is taking place and indicates the percentage of the number of correct predictions. Often to evaluate the performance of a model and its generalization ability an approach involving dividing the data into three parts, Train, Validation and Test, is followed. However, this technique generally does not work well in cases where large datasets are not avail-able. Thus, when limited datasets are available, as is the present case, splitting the dataset may result in some useful information being excluded from the training procedure and the model failing to learn the data distribution correctly. Therefore, after this first stage of model evaluation, the K-fold Cross-validation (CV) method was used to validate the network's generalization ability. In K-Fold CV, the "K" parameter decides the number of folds into which the dataset will be divided. Each fold has a chance to appear in the training set (K-1) times, which ensures that every observation appears in the dataset, thus allowing the model to better learn the

distribution of the underlying data. The results of the K-fold test were also compared with the estimation obtained by leave-one-out cross-validation (LOOCV), a computationally expensive version of cross-validation in which K = N and N is the total number of examples in the training dataset. In other words, each sample in the training set is assigned an example to be used alone as the test evaluation dataset. This procedure is rarely used for large datasets because it is computationally expensive, but it can be used in the case presented in this article. Thus, a comparison could be made between the average classification accuracy for different values of K and the average classification accuracy of LOOCV on the same dataset.

7.5. Results

Statistical analysis

The main description of the population considered in this clinical case is summarized in Table 20.

	Total patients N = 237	**Patients N = 111**	**Suicidal volition patients N = 126**	
Features	***N (%)***	***(%)***	***(%)***	***p-value***
Patients 2016-2020	42.62	21.09	21.51	0.478
Female	74.69	33.76	40.93	0.386
Age 125	6.30	4.64	1.68	0.034
Italian	57.38	26.16	31.22	0.655
Resident in Florence	31.64	12.23	19.41	0.086
Parental divorce/death	42.60	21.09	21.51	0.478
Only child	21.52	9.28	12.23	0.550
Maternal PD	33.75	15.19	18.56	0.686
Paternal PD	35.85	17.29	18.56	0.747
Family history of SBTS	8.00	3.79	4.21	0.961
Poor school performance	20.68	11.39	9.28	0.193
School failure	17.30	8.86	8.43	0.536
Bullying	38.40	17.29	2.09	0.665
RCCI	17.30	6.75	10.54	0.270
Winter/spring ADM	58.65	28.27	30.37	0.616
School day ADM	69.20	30.80	38.39	0.283
Self-cutting	41.34	4.21	37.13	0.000
Intoxication	27.43	24.47	2.95	0.000
Previous IS	83.54	39.24	44.30	0.926
Previous AS	23.62	14.34	9.28	0.017
Previous specialist care	83.54	35.86	47.67	0.007
Drug	61.60	27.00	34.60	0.241
MD	55.27	27.00	28.27	0.489
DICCD	8.01	6.33	1.69	0.003
ID	2.90	0.42	2.54	0.080
Anxiety	24.05	11.39	12.65	0.926
ND	12.23	7.17	5.06	0.175
Substance abuse	5.90	2.95	2.95	0.807
RCCI after discharge	24.05	13.08	10.97	0.190

Table 20 - Summary of the clinical data of 237 patients involved in this study

This study included 237 hospitalizations per SBT and represented 33.86% of the total hospitalizations of the CAPEU. In addition, 14.77% of patients exhibited SBTs twice or three times during the study period. Regarding

temporal trends, the hospitalization per-centage for SBTs switched from 26.92% in 2016 to 52.83% in the first half of 2020. Specifically, hospitalizations are shown in Table 20.

The female/male ratio among our population was 2.9:1 (177 females and 60 males), whilst the age ranged between 7.4 and 17.9 years (mean 15.44 ± 1.67). Age-related strati-fication showed that 43.04% (102 patients) of SBTs occurred in 16 to 18-year-old patients, followed by 14 to 16-year-olds (37.13%; 88 patients), 12 to 14-year-olds (16.88%; 40 patients) and those less than 12 years old (2.95%; 7 patients). We detected a female prevalence in patients older than 12 years old (75.65%; 173 females) and a male prevalence (57.14%; 4 males) in subjects younger than 12 years old.

The most common referral reasons were, in order of frequency, suicidal ideation (53.16%; 126 patients), followed by attempted suicide (32.07%; 76 patients) and failed suicide (14.77%; 35 patients). Forty-eight-point-five percent of the entire population (105/237) showed a family history of neuropsychiatric disorders, 5.5% of them (13 patients) having a family history of suicidal behaviors. In our study, 81.01% (191 patients) of the population reported one or more STEs during their life. Previous suicidal behaviors were a common finding and 64.98% (154 patients) revealed a lifelong history of non-suicidal self-injury.

The most frequent diagnosis at discharge was mood disorder (55.27%; 131 of patients), with a significant prevalence in females (70.99%), followed by feeding and eating disorders (11.81%; 28 of patients) and trauma and stress-related disorders (9.70%; 23 of patients). Considering all patients, 162 (68.35%) showed comorbidities: anxiety disorder (16.88%; 40 patients), multimorbidity (14.77%; 35 patients) and neurodevelopmental disorder (6.33%; 15 patients). Sixty-one-point six percent (146 patients) already had a drug prescription in mono and polytherapy including atypical antipsychotics (39.24%), mood stabilizers (33.76%), antidepressants (29.96%), and anxiolytics (21.52%). Most patients (83.54%; 197 patients), when admitted, were already on psychiatric or psychological care. “Suicidal volition patients” versus “suicidal motivation patients”. Comparing patients admit-ted for “suicidal volition” (53.16%; 126/237) and subjects hospitalized for “suicidal motivation” (46.84%; 111/237), we found statistically significant differences in the method of suicide and previous specialist care (see Table 20). Furthermore, a correlation between people admitted for “suicidal motivation” and aged younger than 12.5 years old, $\chi^2(1, \mathrm{N} = 237) = 4.516, \mathrm{p} = .034$, previously attempted suicide, $\chi^2(1, \mathrm{N} = 237) = 5.672, \mathrm{p} = .017$, intoxication, $\chi^2(1, \mathrm{N} = 237) = 64.651, \mathrm{p} < .01$, and disruptive, impulse-control or conduct disorder diagnosis, $\chi^2(1, \mathrm{N} = 237) = 8.554, \mathrm{p} = .003$, was demonstrated. A higher risk in the group of “suicidal volition,” instead, was demonstrated for self-cutting behavior, $\chi^2(1, \mathrm{N} = 237) = 90.047, \mathrm{p} < .01$, and previous specialist care, $\chi^2 = (1, \mathrm{N} = 237) = 7.373, \mathrm{p} = .007$ (see Table 20).

Artificial intelligence: Neural network approach. The dataset, consisting of 237 subjects, was divided into a train set (166 cases, equivalent to 70% of the total) and a test set (71 cases, equivalent to 30% of the total) to train and test the implemented network. The number of epochs fixed at 100 allows us to avoid the phenomenon of data overfitting. This phenomenon occurs when the weight model follows the test set too specifically and cannot adapt to other examples, which leads to a loss of accuracy. The network described achieved a final accuracy of 86.7% (Figure 1).

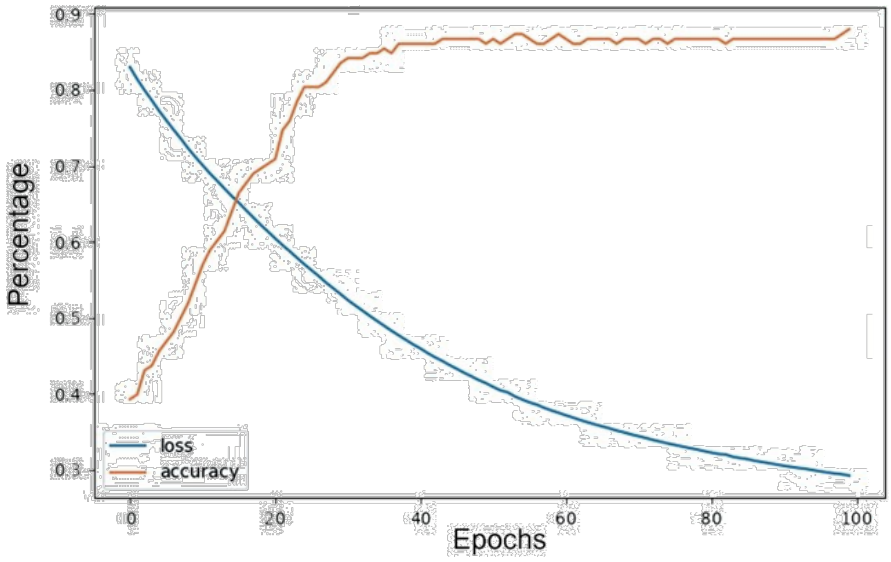


Figure 35 - Neural network approach: Graph showing loss (mean squared error) and accuracy versus number of training epochs. The highest accuracy peek achieved is 89%

K-fold CV and LOOCV generally produce a less biased model than other methods as they ensure that every observation in the original dataset has a chance to appear in the training set and the test set. Sometimes the value obtained with LOOCV is considered to be the "ideal" value achievable by the model despite the fact that in general LOOCV can be subject to high variance (so that very different estimates would be obtained if the estimation were repeated with different initial samples of data from the same distribution) or overfitting since the model is being provided with almost all of the training data to learn and only a single observation to evaluate. In the present case, for K = 5 the model obtains an average accuracy of 84%, which is consistent with the ideal line obtained with LOOCV, i.e. indicating the model's stability in predicting with this accuracy value, which is slightly lower than that obtained in the previous test.

7.6. Discussion

Overall, out of 700 inpatients admitted to CAPEU between January 2016 and June 2020, 33.86% were for STBs with a significant increase over the years and hospitalization for SBTs switched from 26.92% in 2016 to 52.83% in the first half of 2020. More than 75% of patients were female, except for subjects younger than 12 years old, who showed a male/female ratio of 1.3:1. Male prevalence in younger patients seems to be related to the common finding of Disruptive, Impulse-control and Conduct-Disorders, which are known to have a male-prevalence [318].

The greatest incidence of SBTs is observed in spring and winter (56.96% of cases), probably due to the potentially stressful role of school. This data, already highlighted by the international literature [319], is confirmed by the higher incidence of SBTs during weekdays (69.20%) compared to weekends and school holidays (30.8%). In addition, school difficulties are the second leading cause of triggering SBTs (8.44). More than 14% of our patients had a second or third re-hospitalization for SBTs. Determining the rate of rehospitalization with a diagnosis of suicidal ideation or attempted suicide within a year could be useful for the implementation of preventive measures. The presence of psychiatric disorders in 99.58% of patients could be related to their enrollment in an emergency department of a mental health third-level center. Additionally, mood dis-orders are the most frequent diagnosis at discharge, often associated with anxiety, behavior and eating disorders. In our sample, 68.35% of patients had one or more psychiatric comorbidities confirming that suicidal risk is strictly connected with the number of psychiatric diagnoses [320].

Suicidal ideation was the main reason for hospitalization in our series (45.15%), followed by suicide attempts (32.07%). Patients with suicide attempts/failed suicide did not show an increased presence of suicidal behaviors such as previous ideation or self-cutting, unlike patients with suicidal ideation. Furthermore, suicide

attempts in the group of patients hospitalized for attempted suicide occurred mostly through intoxication (voluntary drug ingestion), and this figure is also consistent with that of other studies [321].

However, it appears that more patients with “suicidal volition” than those with “suicidal motivation” have greater access to previous specialist treatments. This datum is coherent and explicable with the fact that in many psychiatric illnesses with developmental onset and in particular, depressive disorder and bipolar disorder, acute psychotic dis-order, posttraumatic stress disorder, eating disorders, anxiety, personality disorders with high impulsivity, and chronic or repetitive suicide may be a component of the syndrome [322].

Our observation of a strong correlation between age <12.5 years and a higher risk of “suicidal motivation” is scarcely observed in the literature. Consistent with this finding, a few other studies have identified a similar prevalence of suicidal ideation in boys and girls through age 12 and a higher prevalence of suicide attempts in boys than in girls in this age group [323]. It may be related to a higher rate of diagnoses of Disruptive, Impulse Control and Conduct Disorders and underreporting of thoughts of death and suicidal ideation in this age group.

Children with these conditions may have a diminished view of their emotions and an understanding of their own frustration, which can lead them to internalize their difficulties and experience self-harming thoughts and impulsive behaviors. This may be related to their difficulty recognizing or communicating their distressing thoughts and sometimes referral adults and clinicians may have difficulty assessing the nature and intent of their behavior. Accurate estimation of suicidal risk remains one of the most difficult and most important tasks that clinicians face. Estimation of suicidal risk also requires taking into consideration specific factors associated with the progression from suicidal ideation to attempted suicide.

Finally, the information collected in the database for each patient has proved to be relevant for the creation of a predictive model using machine learning tools. Although it is not possible to trace the characteristics of greatest impact clearly and directly after RFE, we can state that the information used to train the model is significant, as it yields an accuracy ranging between about 84 and 86%, on a relatively small input case study, proving to be a reliable monitoring and prediction tool. This work has, therefore, allowed the development of a key instrument able to predict, with good reliability, a suicidal event, giving the possibility to define a new intervention strategy, thus preventing and reducing the risk of suicide.

7.7. Final remarks

This study reveals a significant increase in the hospitalization rate for SBTs among females aged 16–18 years old at the CAPEU of the Meyer Children’s Hospital. Risk factors include males under 12 years with disruptive, impulse control and conduct dis-order, individuals using intoxication as a suicide method, those with previous suicide attempts, and those with prior specialist care. Early identification of suicide risk factors during childhood and adolescence is crucial. This study contributes to knowledge and prevention efforts, utilizing machine learning techniques.

The limitation of the study is the retrospective nature of the study; however, it is a single-center recruitment, and all patients have been investigated with a standardized protocol.

Future work to improve the study is needed for further exploration, and it will have to consider a greater number of subjects in order to improve the performance of artificial intelligence.

8. Artificial intelligence for neurology – case study glioblastoma multiforme in rats

This last case study is inserted in the cooperation born inside the visiting period of my PhD research to the CTB of the UPM in Madrid. The need of this group was to automatize a process performed to evaluate the effectiveness of the new technology implemented by this center used to treat the brain tumor, glioblastoma multiforme, without the need for a surgery. After a further explanation of the clinical problem, the AI tool developed will be presented as follows, remembering the AI framework: the clinical scenario to which the case study belongs is presented, the task to be addressed, with the relative metric and the utilized data type are described; the data used are illustrated and presented with the necessary processing; the model used is presented and the final application developed is shown and described.

8.1. Glioblastoma multiforme

Glioblastoma multiforme (GBM) is one of the most aggressive types and the most common primary brain tumors [324]. It is also the deadliest form of glioma, accounting for approximately 45% of all brain tumors, with a median overall survival time ranging from 14.6 to 20.5 months. This is given by several characteristics of GBM, among which are its rapid growth, invasiveness and resistance to treatment [324]. For these reasons the treatment is complex and challenging, and so several standard treatments exist, including surgery, radiation therapy and chemotherapy [325]. However these treatments are often ineffective, and the recurrence rate is high given the GBM's ability of invading surrounding brain tissue.

Given the limited success of current methods, research is progressing in other directions with the aim of finding an effective therapy, such as tomotherapy, hyperthermia, and oncolytic virotherapy. Among the possible new therapies the one considered in this work is optical hyperthermia. This technique is based on heating the tumor tissue to temperatures ranging from 40 to 45°C, which can induce the apoptosis process and slowdown intracranial glioma progression [326]. More specifically, this therapy has been carried out by the CTB on rats to which the GBM was previously stereotaxically inoculated in the previous two weeks and finally using gold nanorods the hyperthermia was performed. Given the innovative aspects of this treatment that does not need to perform a surgical procedure or a radiotherapy it is important to understand the effectiveness of the overall procedure using some specific metrics, such as volume of the GBM and percentage of brain volume occupied.

8.2. Automatic glioblastoma multiforme volume computation in rats' brain

In order to compute the volume of 3D structures obtained by means of MRI images, it is necessary to have the 3D delineation, also called segmentation. In literature, there are many cases of automatic segmentation of tumors [327–329], but for the majority of them, the studies are based on human databases. This is one of the main reasons why, based on deep research of the literature, it does not exist a public available tool to automatically segment and compute the volume of GBM tumors in rat's brain. Even if there are not research that deal directly with this problem, in the literature there exist some studies that try to segment the rat's brain from the MRI, such as in the case of [330], in which the authors proposed a framework that used a multi-atlas similarity-based and multi-label fusion algorithm for the propagation of the segmentation to perform the segmentation of multiple brain areas. Also, in most recent years the proposed solutions are based on machine learning algorithms, more specifically deep learning algorithms that uses as baseline the 3D U-Net model [331–334], other solutions exist that uses generative adversarial networks such as [335]. Furthermore, public dataset contains MRI data of rats, but there are three main problems, firstly they mainly consist of rats that do not have a GBM developed in the brain, secondly there are no segmentation provided, finally they are not coherent and consistent with the type of data produced by our laboratory.

The aim of this work is to develop a completely automated tool through the use of deep learning to compute the volume of a GBM brain tumor, and the volume of the brain itself, to obtain the percentage of the brain's

volume occupied by the GBM. The 3D U-Net model has been selected and trained over two different strategies, based on the type of used input to automatically segment both brain and tumor. The best results were obtained using the nnU-Net framework with an ensemble of the trained models with a testing dice score of 99.07% for the brain segmentation and of 97.47% for the tumor segmentation. A 3D Slicer's extension has been implemented to create an easy to use and fast tool.

8.3. Materials

All the procedure of data acquisition, starting from the rats' management, the therapeutic approach, to the final MRI acquisition of the rats. Here there will be discussed only the steps that follow the data acquisition, in order to compute the volume and the percentage of brain volume occupied.

A total of 26 MRI of rats with a GBM tumor were acquired, obtained from 19 different rats. For each MRI both the TC1 and TC2 phases were available (an example is depicted in Figure 36) the first one is useful to examine the normal anatomy of the brain, and the latter is used to detect the pathological changes in the neural tissue [336]. The dimensions of the diagnostic images are of 170x170 with a total number of 15 slices and a spacing of 0.1 mm x 0.1 mm x 1.0 mm.

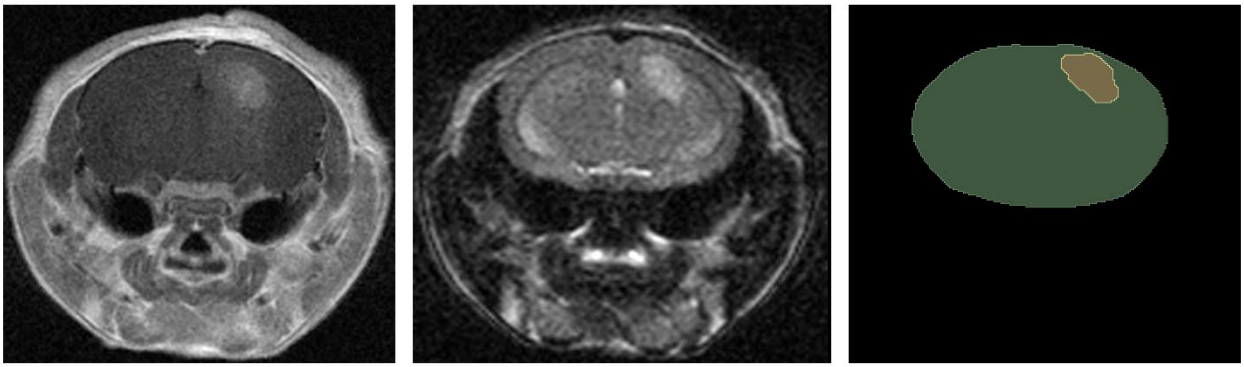


Figure 36 Representation of a single slice corresponding to the MRI of one rat with GBM: (a) TC1 phase in which it can be noted like the GBM has a brighter color than the brain; (b) TC2 phase where it can be observed how the brain contours are crisper; (c) the ground truth manual segmentation of the brain and of the GBM.

Through manual segmentation a ground truth of the brain and the GBM of each MRI was realized by a team of experts to reduce potential biases. In particular it was done using the "3D Slicer" software [337], combining manual segmentation with the tools provided by the program (e.g., "grow from seed", "smoothing", and "fill holes").

8.4. Initial Analysis

A first analysis phase on the available data was done to understand the range of possible approaches that could be used for this specific task. In particular, it was used the ground truth segmentation of the tumors and the brains in order to analyze the overall pixels' distribution. In Figure 37 are shown, for a subset of the available data and only for the TC1 phase, the pixels' intensity distribution, normalized through min-max normalization, relative to the ground truth of the tumor segmentation (a) and of the brain segmentation without the tumor's pixels (b). From the two graphs is possible to see that the distribution of the brain pixels is similar for all the cases having the majority of the values below 0.6, but if we check the tumor's pixel distribution there is just one case that could be easily divided from the tumor, instead for the majority of the tumors they would have a pixels' distribution indistinguishable from the brain's one. From this it is possible to understand that this problem is not straightforward, not solvable with a threshold-based approach, given the variability on the tumors' pixel distributions.

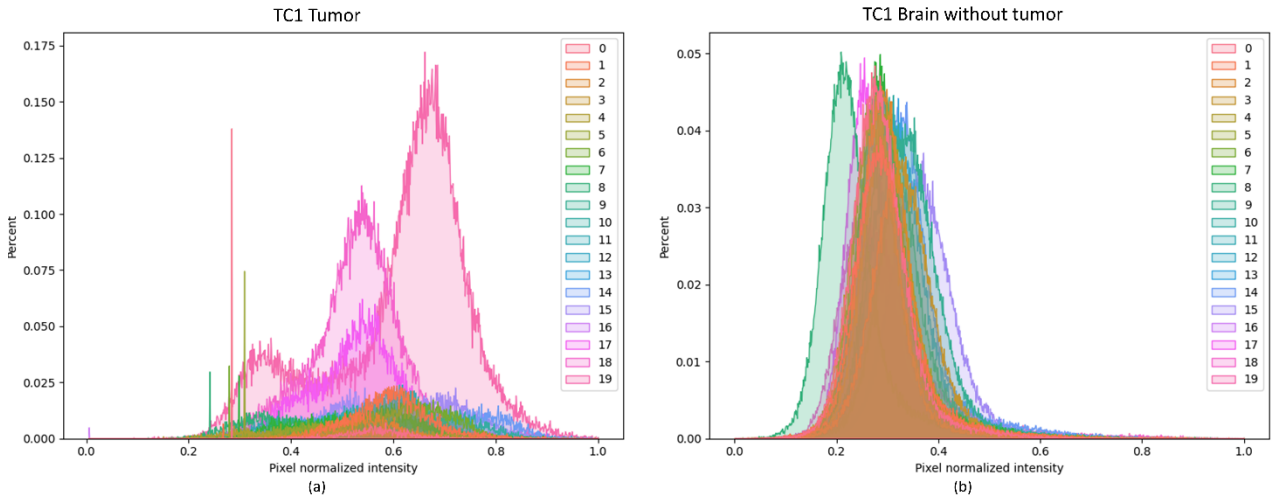


Figure 37 The TC1 phase distribution of the pixels' intensity inside the ground truth of the tumor (a) and of the brain (b) removing the tumor ground truth values. To increase the readability of the graphics it has been decided to plot only a small randomly selected subset of the dataset. The tumors intensities values belong to the interval of values greater than 0.2, and the brain intensities are mainly below 0.6. Even if some cases could seem to be easily split with a threshold, the majority of the tumors pixels overlaps with the relative brain pixels.

Looking at the TC1 phase pixels' intensity distribution it seems difficult to apply a statistical-based approach, and checking the one of the TC2 phase, depicted in Figure 38, it seems impossible to distinguish the tumor from the brain just using the pixel values. This is a very important aspect that must be taken into consideration at the moment of strategy development.

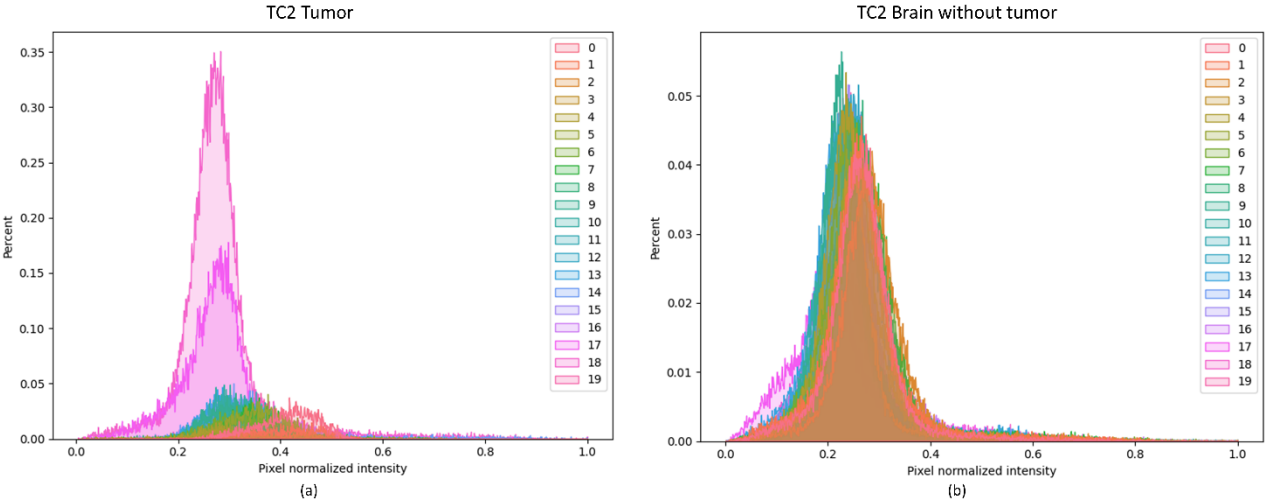


Figure 38 The TC2 phase distribution of the pixels' intensity inside the ground truth of the tumor (a) and of the brain (b) removing the tumor ground truth values. To increase the readability of the graphics it has been decided to plot only a small randomly selected subset of the dataset. The tumors intensities values and the brain ones belong practically to the same identical interval without showing differences in the behavior for all the cases.

Other than this, another challenge is given by the fact that the dataset has very few cases, this problem must be taken into account particularly to avoid creating solutions that will not be able to generalize well the GBM characteristics and therefore be useless.

8.5. Methods

As stated in the introductory section, the aim of this work is to create an automatic system that is able to compute the volume of GBM in rat MRI. Also the problem of computing the volume of the brain occupied by the GBM will be addressed and therefore taken into account in algorithm selection phase. The main difficulty for this task consists in the segmentation of the interesting areas (GBM and brain), once this is done the

computation of the volume is straightforward knowing the spacing of the diagnostic images. In particular it will be simply computed using the following formula:

$$Volume = N_{Seg} \times Voxel_{size}$$

where N_{seg} is the number of all segmented voxels, and $Voxel_{size}$ is the size of the single voxel simply obtaining by multiplying the pixel size in both dimensions with the slice thickness.

Being this an automatic segmentation task and knowing the particularity and the challenges of this type of dataset, considering the literature on segmenting tumors and brains, an AI-based solution seems to be the most appropriate option. Considering the limited cases available the U-Net model is the best choice, being able to achieve good performances even in these situations [102]. More in detail the 3D U-Net [338] was chosen, considering the type of input and with the aim of elaborate the full images without dividing it into single slices, trying to achieve better results. The goal is to select and implement the best version of the 3D U-Net for this specific problem and since several implementations of the 3D U-Net exists, it was chosen to compare the results obtained using the 3D U-Net with variational autoencoders (VAE) [339] provided by the MONAI API [340] and the 3D U-Net implementation of the no-new-Net (nnU-Net) framework [341]. The 3D U-Net with VAE was selected because the combination between U-Net and VAE has the potential to improve the final segmentation performance in small dataset scenarios by leveraging the generative ability of the VAE to augment the data and reduce overfitting [342]. The nnU-Net was picked because it is able to obtain state-of-the-art performance in the vast majority of biomedical imaging segmentation tasks with its particular type of training that is able to handle any diagnostic image shape type as input, its data augmentation policy, the preprocessing and the final postprocessing.

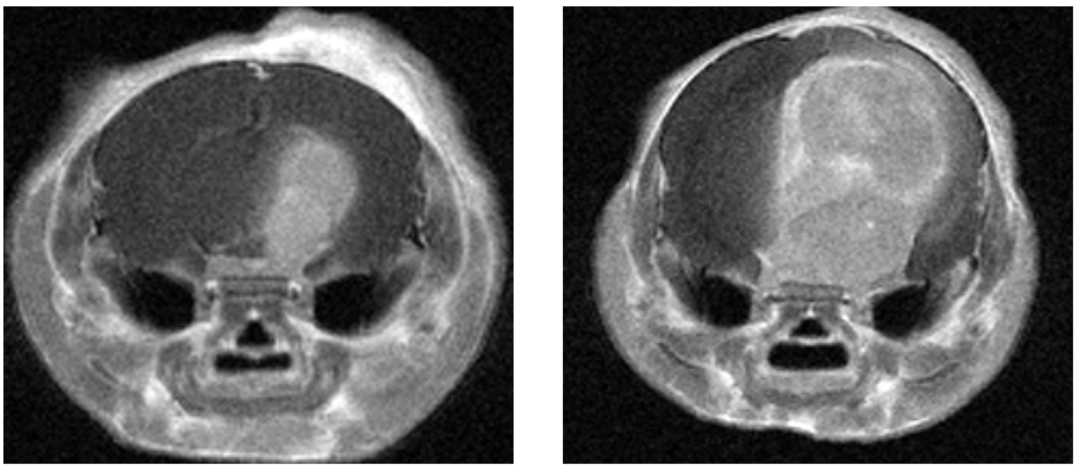
After choosing the model, there was a main aspect to take into account which is the type of chosen input: considering the above, the TC1 phase of the MRI is a necessary input to give relevant information to the model in order to identify the tumor; instead for what concern the TC2 it is useful only to have a better delimitation of the brain and it could discarded.

Given the type of available data and chosen architecture, two different strategies were selected in order to determine the best input that should be used. The two strategies vary by the type of the used input as shown in Table 21. The main difference between them is given by the fact that the first one uses as input only the TC1 phase of the MRI, the second uses all the available information.

Strategy	*Input*	*Task*
Tumor and Brain – TC1	MRI TC1	Segmentation of tumor and brain
Tumor and Brain – TC1 TC2	MRI TC1 & TC2	Segmentation of tumor and brain

Table 21 The two strategies developed in this study: the first one is the only that has the aim of segment purely the tumor not considering the brain, and as the second one it uses as input just the TC1 phase of the MRI; the second and the third one have the same task, but the last strategies uses as input both the TC1 and TC2

Another important aspect is the fact that some MRI in the dataset corresponds to the same rat (taken in different days). They were not removed considering that there is a clear difference between the two (an example is visible in Figure 39) but they were handled with a specific criterion detailed in the following. The dataset to train and test the model was split into two main sets, the training set, that will contain 24 cases, and the testing case that will contain 2 cases. In addition to this, the training set was used with a k-fold cross-validation (k=4, so 18 cases for training and 6 to validate) method to obtain statistically accurate results. All the splits were not completely random but made in such a way that all the MRI data relative to one specific rat are always in the same set and the 2 cases in the testing set are relative to two different rats.



(a) (b)

Figure 39 Example of GBM evolution over one week of time for one rat: (a) is the first MRI made to the rat, (b) is the second MRI made to the rat after one week from the other one.

To compare the two different implementations of 3D U-Net the average dice score was used obtained using the k-fold cross-validation. To evaluate the capability of the best proposed solutions on the testing set in order to pick the best solution not only the dice score, but also the jaccard score, precision, recall, and false positive rate (FPR) was computed.

8.5. Results

The results obtained during the training phase of all the strategies are depicted in Table 22. For each one of the strategies is reported for the specific framework, the considered output and the dice score obtained for each one of the k-folds and the average of the folders dice scores.

Strategy	*Framework*	*Output*	*Metrics Validation*				
			K-Fold	0	1	2	3
Tumor and Brain TC1	MONAI	Brain	Dice Score	82.20%	81.91%	88.06%	90.07%
			Average Dice Score	85.56%			
		Tumor	Dice Score	40.4%	4561%	47.06%	48.40%
			Average Dice Score	45.37%			
	nnU-Net	Brain	Dice Score	95.62%	94.09%	95.00%	94.72%
			Average Dice Score	94.86%			
		Tumor	Dice Score	83.40%	88.60%	83.31%	82.62%
			Average Dice Score	84.48%			
Tumor and Brain TC1 TC2	MONAI	Brain	Dice Score	94.85%	90.12%	91.85%	90.85%
			Average Dice Score	91.92%			
		Tumor	Dice Score	66.83%	64.37%	64.05%	43.33%
			Average Dice Score	59.65%			

nnU-Net	Brain	Dice Score	97.00%	96.45%	95.66%	95.09%
		Average Dice Score	96.05%			
	Tumor	Dice Score	84.91%	87.98%	84.66%	81.80%
		Average Dice Score	84.84%			

Table 22 Results of the training phase: per each strategy the dice score evaluated on the validation set of each folder is reported in such a way that it is divided by the framework used and the considered output.

The results obtained using MONAI are for the mean dice score of brain segmentation 85.56% (C.I. $\pm$3.51, α = 95%) using only TC1 phase and 91.92% (C.I. $\pm$1,18, α = 95%) using both phases and for GBM segmentation an average of 45.37% (C.I. $\pm$2.97, α = 95%) with TC1 59.65% (C.I. $\pm$9.29, α = 95%) with TC1 and TC2. In contrast, using nnU-Net the results, obtained with the strategy "Tumor and brain - TC1," are 94.86% (C.I. $\pm$5.40, α = 95%) for brain segmentation and 84.48% (C.I. $\pm$2.35, α = 95%) for tumor segmentation. Finally, for the last strategy with CT2, the average dice score of nnU-Net for brain and tumor segmentation is 96.05% (C.I. $\pm$0.72, α = 95%) and 84.84% (C.I. $\pm$2.14, α = 95%), respectively.

Reaching the nnU-Net model the best performances for any type of strategy it has been decided to pick it as the best model, more in detail the ensemble of the trained models has been evaluated on the testing set. The testing cases given the limited size of the available data is composed just by two cases shown in

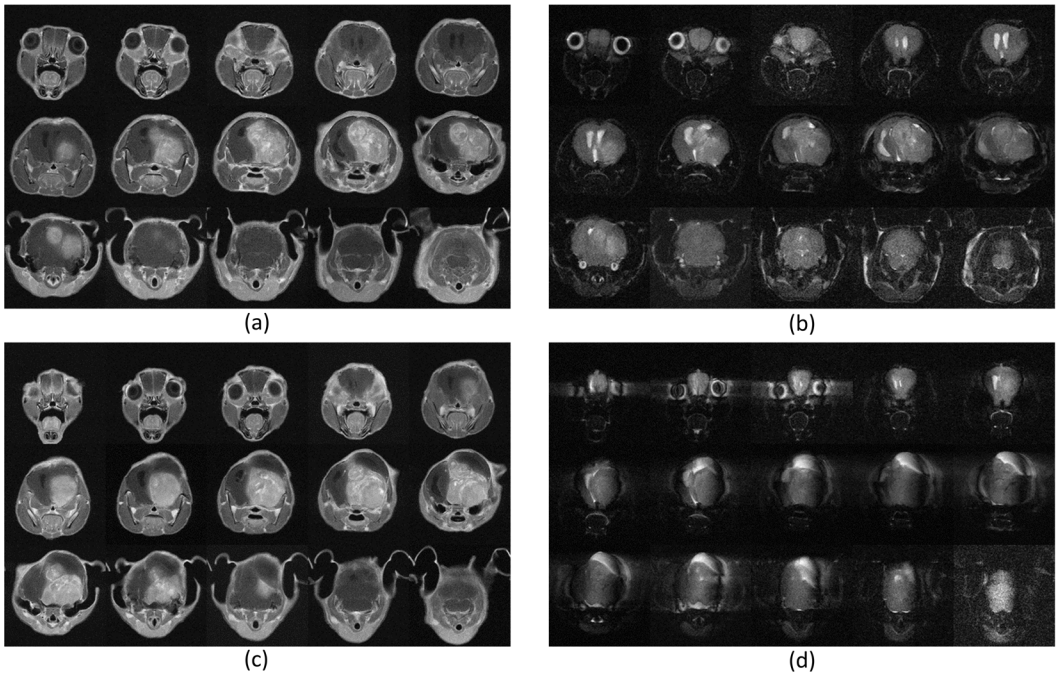


Figure 40, in which the first two images (a) and (b) represent the TC1 and the TC2 of the first rat, RatTest1, that is a rat with a GBM developed and a good MRI acquisition for both phases; the (c) and (d) images in the last row, of RatTest2, show and example of rat with GBM, but with an error in the acquisition of the TC2 phase.

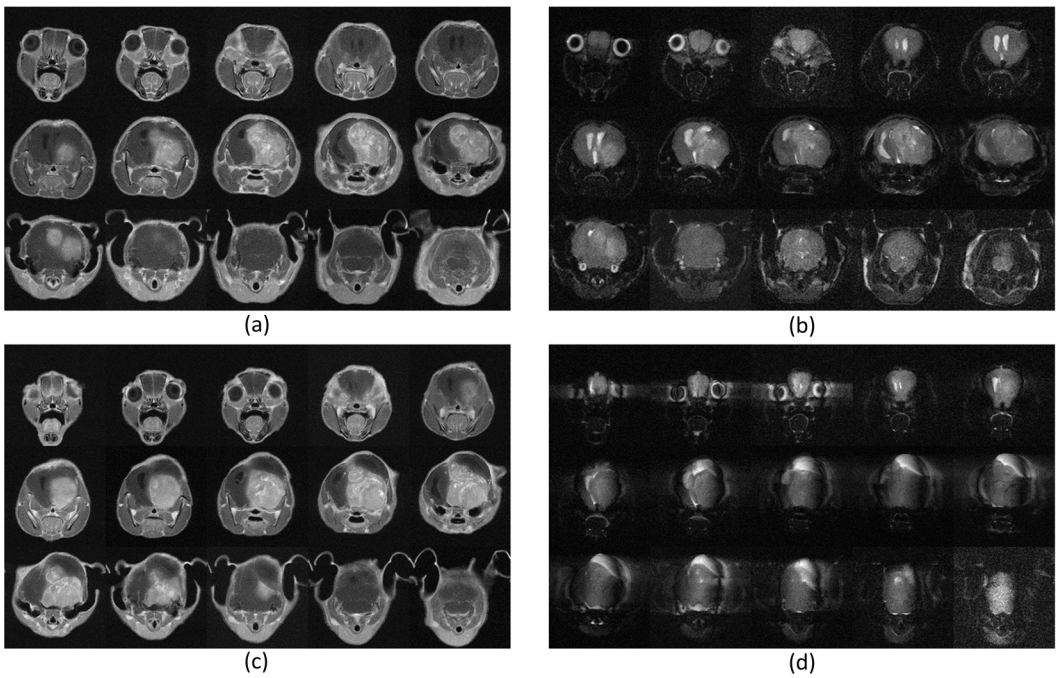


Figure 40 The slices of the two rats' MRIs of the testing set: (a) and (b) are respectively the TC1 and TC2 phases of RatTest1;(c) and (d) the TC1 and TC2 of RatTest2. From (d) it can be seen that the image of the TC2 phase is a little bit noisy, this could influence negatively the results when it is used.

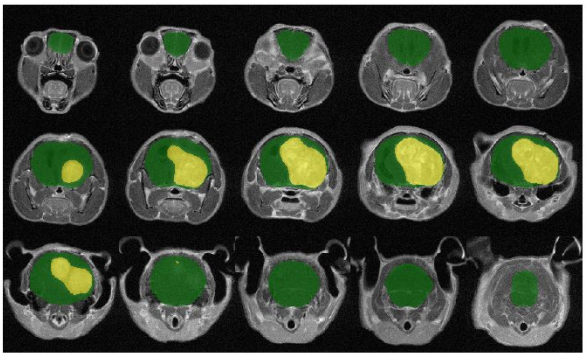
The metrics are reported in Table 23 for both cases, also the results for the RatTest2 of the last strategy are not considered valid because the TC2 image contains noise due to the acquisition, and this would negatively affect the comparison.

		Metrics Testing									
Strategy	*Output*	Dice Score		Jaccard Score		Precision		Recall		FPR	
		Rat Test1	Rat Test2	Rat Test1	Rat Test2	Rat Test1	Rat Test2	Rat Test1	Rat Test2	Rat Test1	Rat Test2
Brain and Tumor TC1	Brain	99.29%	98.84%	98.59%	97.71%	99.55%	98.26%	99.03%	99,43%	0,07%	0,30%
	Tumor	98.27%	96.66%	96.60%	93.54%	97.43%	93.82%	99.12%	99,68%	0,10%	0,39%
Brain and Tumor TC1 TC2	Brain	96.55%	-	93.33%	-	96.45%	-	96.65%	-	0.57%	-
	Tumor	97%	-	94.18%	-	97.48%	-	96.53%	-	0.09%	-

Table 23 Metrics of the ensemble nnU-Net model evaluated over the testing set for each strategy considered and divided into tasks, '-' no meaningful data.

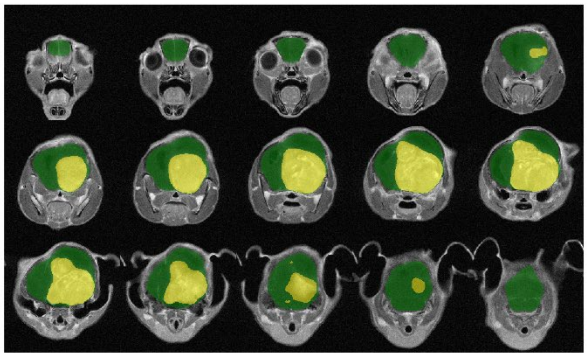
Finally, considering that the first strategy yields the best results, in Figure 41 displays the segmentation outcome obtained by applying the nnU-Net model to the two testing cases.

RatTest1



Brain Volume: 600.40 mm^3 – Tumor Volume: 156.85 mm^3

RatTest2



Brain Volume: 744.00 mm^3 – Tumor Volume: 289.80 mm^3

Figure 41 The results relative to the usage of the nnU-Net with the "Brain and Tumor TC1" strategy for both the testing cases: under the images there are the volumes computed.

8.6. Discussions

From the results is possible to see how using deep learning, in particular using the 3D U-Net proposed by nnU-Net, it is possible to segment diagnostic images of rats' brains with a high level of accuracy and detail. Even if this study is limited by the number of used cases, through the use of k-fold cross-validation it has been demonstrated that the performances are good and the results obtained are considerable as a valid starting point to correctly compute the volume. Considering the metrics measured it is possible to say that the nnU-Net framework obtains easily better results than the MONAI implementation. Furthermore, comparing the results obtained with the nnU-Net between the two strategies it seems like the usage of the TC2 does not provide enough information to be deeply influent in the segmentation task, or at least in this limited dataset it does not show interesting improvements when it was used. For this reason, the "Brain and Tumor TC1" is the best strategy so far, being more consistent considering all the possible situations.

Additionally, further steps were performed in order to made a usable tool, i.e. a 3D Slicer extension has been implemented. The extension is able to use any of the proposed strategies with the ensemble of the corresponding models and can work with GPU or directly on the CPU, with an approximate time of one minute on a small budget GPU, such as the NVIDIA MX150, and about three minutes on a CPU Intel Core i7-8550U. Once the segmentation is done a table is created containing the volume expressed in mm^3 of the GBM, of the brain and finally the percentage of brain occupied by the GBM. In Figure 42 is shown an example of usage of the extension having as input the RatTest1 and using as strategy "Brain and Tumor TC1".

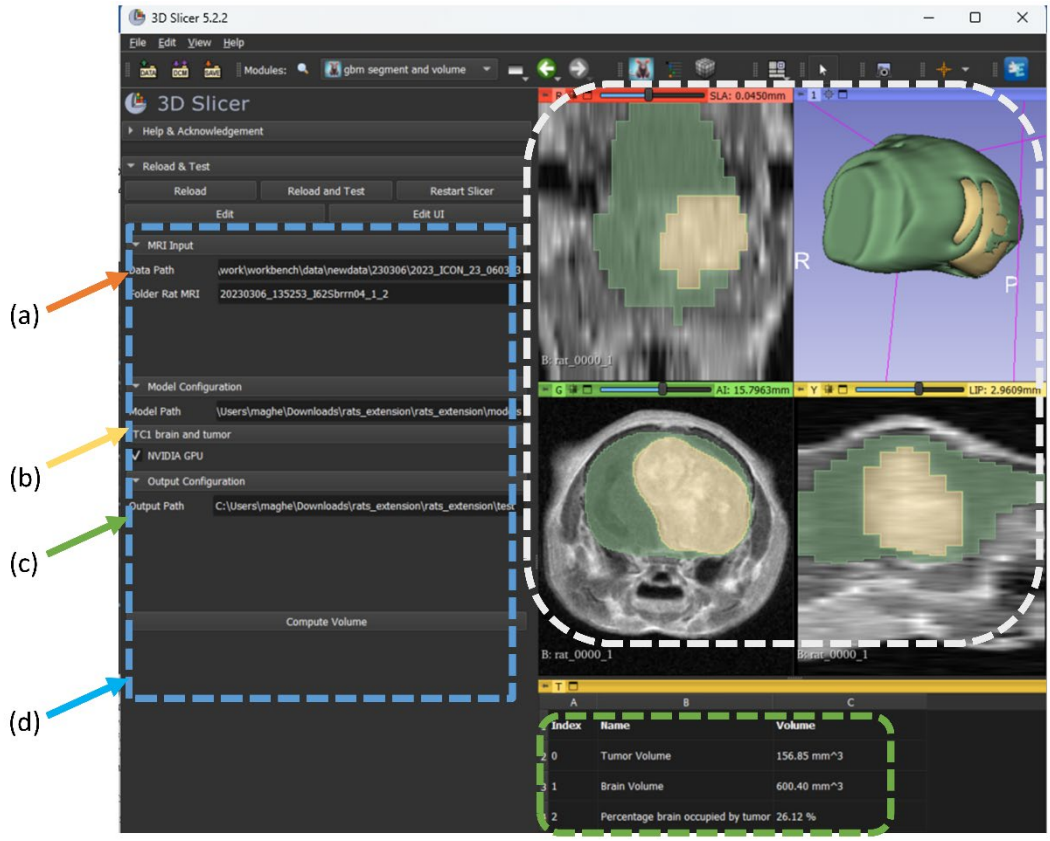


Figure 42 An example of the extension's functioning. In the blue box are reported the field of the extension that must be filled in order to specify (a) the data path containing the rats' MRI folders and the folder name relative to the MRI saved with the standard bruker format; (b) the folder containing the checkpoints of the usable models, the strategy to be used (by default the "brain and tumor TC1") and if it is present and usable an NVIDIA GPU; (c) the output path that will contain the starting MRI and the final segmentation both as nifti files, and the final computed values of the volumes of the brain and the GBM, and the percentage of the brain's volume occupied by the GBM. In the white box the visualization of the MRI used as input, with the automatically created segmentation superimposed to check if it is consistent with the MRI. Finally, in the green box a table is generated containing the computed tumor and brain volumes and the percentage of the brain occupied by the tumor.

Using this extension, it is possible not only to automatize a very time consuming and repetitive task, but also enable the usability of an objective system that can be used even by non-expert operators to obtain in a few minutes the desired result, also using low-budget devices. Furthermore, having the possibility of using this extension within 3D Slicer enable the use of the other tools of the same, and in the cases in which the final results is no good enough it can be manually refined to obtain a better final result. Considering that this extension could be beneficial for others study, it will be freely available on request. Furthermore, being publicly available it is really simple to change the model used to generate the segmentation if better proposal might arise. Future steps will aim to the acquisition of new data to increase the number of case studies and to improve the performance of the model proposed. Finally, given the efforts made to collect and to label the data used in this study, and also given the value of the data availability it has been decided to publicly release the dataset used with its ground truth.

8.7. Final remarks

This study demonstrates the potential of deep learning, specifically the 3D U-Net architecture, for accurate automated brain tumor segmentation in rat MRI data. Although limited by sample size, the nnU-Net framework achieved high Dice similarity scores, outperforming the MONAI implementation. Multi-contrast MRI incorporating T2 and TC2 did not substantially boost segmentation performance versus T1 alone in this small dataset; the T1-only strategy provided the most consistent results. To enable easy quantification of tumor volumes, an automated segmentation pipeline was developed as a 3D Slicer extension, allowing quick analysis even on low-power devices. While additional data is needed to optimize and thoroughly evaluate the

deep learning models, this pilot study shows the feasibility of leveraging artificial intelligence for glioblastoma quantification in rodent models. The annotated MRI dataset and ground truth labels have been publicly released to facilitate future research. In conclusion, deep learning-based segmentation approaches show strong potential for efficient quantitative analysis of preclinical brain tumor studies, though larger datasets and continued algorithm refinement is warranted to improve generalizability.

9. Conclusions

Considering the growth of AI research and its ubiquitous usage in all the fields existing with its great capabilities of adapt to very complex problems, AI is becoming increasingly used in the research field for medical related tasks. This work analyzes the implementation of AI-based tools for medical scenarios with the aim of creating a general framework useful for the realization of every application that use AI with a clinical field. The common approach to realize these tools involves the creation of a specific task-oriented dataset, followed by the selection and the training of an AI model.

Aiming to the goal of implementing AI-based tools that could be usable in clinical practice and used directly by the clinical equipe, a framework of the overall development of the application has been designed to define all the phases that should be followed in order to obtain a valid final result. The following are all the phases taken into account: the definition of the clinical problem challenged, with the analysis of the type of task, metrics, and data necessary; the creation of the database, considering all the data gathering approaches; the cleaning of the data, removing not-valid values, replacing missing one and preprocessing the data to be usable; the creation, training, and testing of the AI model; finally, the implementation of the final application with the user-interface and the AI model ready for the usage of the clinical staff.

The proposed framework has been tested, given all the expected and unexpected limitations, in four case studies identified in collaboration with different groups that vary depending on the case study, like the hospital Careggi in the Custom3D joint laboratory, the hospital Meyer in the T3Ddy joint laboratory, and the Biomedical Technology Center at the Universidad Politecnica de Madrid. These case study are: 1) the analysis of kidney tumor with the cooperation of the urology department of AOUC; 2) the automatization of the design of surgical cut guides with the cooperation of the plastic surgery department of Meyer; 3) distinguishing between possible suicidal patients to understand their mind health in collaboration with the psychiatry group of Meyer; 4) the automatic computation of the brain tumor in rats through the usage of automatic segmentation with the collaboration of the CTB.

Regarding the urology case, after a very specific overview and an in-depth study of the state of the art regarding artificial intelligence applications, an AI-based tool was created following the described framework. In particular, regarding the classification between benign and malignant renal tumor. More specifically, for the first case, more types of metrics were taken into consideration for the final evaluation of the results, but the model's sensitivity was considered a priority as it is able to give a very precise measure of how many cases of malignant tumors are correctly classified. To perform this task, it was decided to use only the diagnostic images (CT) of the patients, in particular a total of 271 CT scans of different patients were collected, of which 221 ccRCC and 50 oncocytomas. Starting from these data, deep features and radiomic features were obtained using a 3D U-Net. Using these features, two appropriate classifiers were trained, and the best one, an ANN based on the use of deep features, obtained the following final performance 73.77% balanced accuracy, 94.59% sensitivity, 52.94% specificity and 86.84% accuracy.

In the case of plastic surgery, the aim was to create tools capable of automating an already existing procedure, but fundamentally dependent on the presence of highly expert and specialized personnel. The identified tasks were two: the generation of depth map type images and the segmentation of the anatomical elements of the ear. For the generative task, MSE and SSIM were used to evaluate the result, having to deal only with RGB images made using a simple camera. In particular, 302 correlated RGB images of specific depth maps were collected; these were specifically labeled to be able to train a first model, a Faster R-CNN, to identify the ear within the images and thus be able to actually use it to train a Cycle GAN model to map only the RGB ear to its equivalent depth map. A final AP of 97.63% was obtained for the Faster R-CNN and for the final depth map generated through Cycle GAN an MSE of ~0.07 and an average SSIM of ~0.80. As for the segmentation of the ear anatomy, the results were evaluated through the model's accuracy in classifying the individual pixels corresponding to the various ear components. In particular, for this specific case, depth maps were used

directly, for a total of 131 depth maps. The model used is a modified version of the U-Net which was able to achieve a final accuracy of 97%.

Moving on to the psychiatry case, the task of classifying between patients with real suicidal intentions and not was identified. Accuracy was used as a measure of the results obtained and only clinical data relating to patients were used. A total of 237 patients were examined, and an ANN was used for the final classification which achieved an overall accuracy of about 86%.

Finally, regarding the case study related to neurology, starting from the objective of calculating the percentage volume of the mouse brain occupied by glioblastoma multiforme tumors, the closely related task of segmenting the brain and tumor of rats having to deal exclusively with MRI was identified. The average dice score was used to evaluate the final segmentations obtained. In particular, a total of 26 MRI scans were obtained from 19 mice. A 3D U-Net was trained for this task obtaining an average dice score of 96.05% for brain segmentation, and an average dice score of 84.84%. Finally, an application was created for the use of this tool, to speed up the creation of the graphical interface and the visualization of the results, it was developed as an extension for 3D Slicer, and it is able through simple steps to generate the required segmentations, show them to the user and calculate the required volume.

9.1 Limitations and future works

Despite the efforts made to ensure the accuracy and relevance of the results, this work has certain limitations that must be acknowledged. It must be considered that the application of AI techniques in clinical practice is still an emerging field, and as such, the models and algorithms used may require further validation and optimization. Except for the plastic surgery task, the others have been explored with a limited dataset, which may affect the generalizability of the results.

Future works in the future will concentrate on overcoming the above limitations by adding an external validation step that verifies the reliability and efficacy of the proposed tools using larger datasets. Another fundamental aspect that must be addressed is the ethical one. As said in previous chapters the ethical aspect of the use of the tools implemented in the clinical environment has not been considered, as none of them have actually been used in the clinical environment and mainly because the development of AI-based technologies has evolved at a much faster pace than legislation and consequently, at the time the overall work was carried out, there was no well-defined regulation for this technology. Therefore, it is necessary to concentrate in great detail in order to make the models more explicable in order to facilitate their understanding and, consequently, to create tools that meet all possible requirements imposed from an ethical point of view. Complementary to these wider advances, new developments in the sectors under consideration will be made. In the case of kidney cancer, for example, the intention is to create a tool that can predict the grade according to the WHO/ISUP criteria for malignant kidney tumors.

Bibliography

1. Ravipati T, Andrew NE, Srikanth V, Beare R (2022) Challenges in public healthcare research data warehouse integration and operationalisation. Int J Popul Data Sci 7:. https://doi.org/10.23889/IJPDS.V7I3.1859
2. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data 2021 8:1 8:1–74. https://doi.org/10.1186/S40537-021-00444-8
3. Enholm IM, Papagiannidis E, Mikalef P, Krogstie J (2022) Artificial Intelligence and Business Value: a Literature Review. Information Systems Frontiers 24:1709–1734. https://doi.org/10.1007/S10796-021-10186-W/TABLES/8
4. Furman J, Seamans R (2019) AI and the economy. Innovation Policy and the Economy 19:161–191. https://doi.org/10.1086/699936/ASSET/IMAGES/LARGE/FG9.JPEG
5. Othman K (2021) Public acceptance and perception of autonomous vehicles: a comprehensive review. AI and Ethics 2021 1:3 1:355–387. https://doi.org/10.1007/S43681-021-00041-8
6. Papageorgiou K, Theodosiou T, Rapti A, Papageorgiou EI, Dimitriou N, Tzovaras D, Margetis G (2022) A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing. Frontiers in Manufacturing Technology 2:972712. https://doi.org/10.3389/FMTEC.2022.972712
7. Samatas GG, Moumgiakmas SS, Papakostas GA Predictive Maintenance-Bridging Artificial Intelligence and IoT
8. Seyhan AA, Carini C (2019) Are innovation and new technologies in precision medicine paving a new era in patients centric care? J Transl Med 17:1–28. https://doi.org/10.1186/S12967-019-1864-9/FIGURES/4
9. Fowler KJ, Cunha GM, Kim TK (2020) Diagnosing non-hepatocellular carcinoma malignancies on CT/MRI and contrast enhanced ultrasound: the Liver Imaging Reporting and Data System approach. Hepatoma Res 6:null-null. https://doi.org/10.20517/2394-5079.2020.21
10. Tsikitas LA, Hopstone MD, Raman A, Duddalwar V (2023) Imaging in Upper Tract Urothelial Carcinoma: A Review. Cancers 2023, Vol 15, Page 5040 15:5040. https://doi.org/10.3390/CANCERS15205040
11. Jiang H, Qin Y, Wei H, Zheng T, Yang T, Wu Y, Ding C, Chernyak V, Ronot M, Fowler KJ, Chen W, Bashir MR, Song B (2023) Prognostic MRI features to predict postresection survivals for very early to intermediate stage hepatocellular carcinoma. European Radiology 2023 1–20. https://doi.org/10.1007/S00330-023-10279-X
12. Malasinghe LP, Ramzan N, Dahal · Keshav (2019) Remote patient monitoring: a comprehensive study. 10:57–76. https://doi.org/10.1007/s12652-017-0598-x
13. Lee SM, Lee DH (2020) Healthcare wearable devices: an analysis of key factors for continuous use intention. Service Business 14:503–531. https://doi.org/10.1007/S11628-020-00428-3/TABLES/8
14. Dufaux F (2021) Grand Challenges in Image Processing. Frontiers in Signal Processing 1:675547. https://doi.org/10.3389/FRSIP.2021.675547
15. Machine Learning Algorithms for Signal and Image Processing | IEEE eBooks | IEEE Xplore. https://ieeexplore.ieee.org/book/9960833. Accessed 24 Oct 2023
16. Khurana D, Koli A, Khatter K, Singh S (2023) Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82:3713–3744. https://doi.org/10.1007/S11042-022-13428-4/FIGURES/3
17. Kumar Y, Koul A, Singla R, Ijaz MF (2022) Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of Ambient Intelligence and Humanized Computing 2021 14:7 14:8459–8486. https://doi.org/10.1007/S12652-021-03612-Z
18. Egert M, Steward JE, Sundaram CP (2020) Machine Learning and Artificial Intelligence in Surgical Fields. Indian J Surg Oncol 11:573–577. https://doi.org/10.1007/S13193-020-01166-8/METRICS
19. Dundar TT, Yurtsever I, Pehlivanoglu MK, Yildiz U, Eker A, Demir MA, Mutluer AS, Tektaş R, Kazan MS, Kitis S, Gokoglu A, Dogan I, Duru N (2022) Machine Learning-Based Surgical Planning for Neurosurgery: Artificial Intelligent Approaches to the Cranium. Front Surg 9:863633. https://doi.org/10.3389/FSURG.2022.863633/BIBTEX
20. Chung J, Teo J (2022) Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. Applied Computational Intelligence and Soft Computing 2022:. https://doi.org/10.1155/2022/9970363
21. Fisher S, Rosella LC (2022) Priorities for successful use of artificial intelligence by public health organizations: a literature review. BMC Public Health 22:1–14. https://doi.org/10.1186/S12889-022-14422-Z/FIGURES/1
22. Careggi. https://www.aou-careggi.toscana.it/internet/index.php?lang=it. Accessed 25 Oct 2023
23. Meyer - Azienda Ospedaliera Universitaria. https://www.meyer.it/index.php/en/. Accessed 26 Oct 2023
24. T3Ddy. https://www.t3ddy.org/. Accessed 25 Oct 2023
25. CTB | Center for Biomedical Technology. http://www.ctb.upm.es/. Accessed 27 Oct 2023
26. OpenAI (2023) GPT-4 Technical Report
27. Communication Artificial Intelligence for Europe | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe. Accessed 24 Oct 2023
28. Janiesch C, Zschech P, Heinrich K Machine learning and deep learning. https://doi.org/10.1007/s12525-021-00475-2
29. Deep Learning. https://www.deeplearningbook.org/. Accessed 24 Oct 2023
30. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, Czogalik Ł, Dudek P, Magiera M, Lis A, Paszkiewicz I, Nawrat Z, Cebula M, Gruszczyńska K (2023) What Is Machine Learning, Artificial Neural Networks and Deep Learning?—

Examples of Practical Applications in Medicine. Diagnostics 2023, Vol 13, Page 2582 13:2582. https://doi.org/10.3390/DIAGNOSTICS13152582
31. Murdoch B (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 22:1–5. https://doi.org/10.1186/S12910-021-00687-3/PEER-REVIEW
32. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P (2021) The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 21:1–23. https://doi.org/10.1186/S12911-021-01488-9/FIGURES/12
33. Agrawal A, Gans JS, Goldfarb A (2019) Exploring the impact of artificial Intelligence: Prediction versus judgment. Information Economics and Policy 47:1–6. https://doi.org/10.1016/J.INFOECOPOL.2019.05.001
34. Chakradhar S (2017) Predictable response: Finding optimal drugs and doses using artificial intelligence. Nat Med 23:1244–1247. https://doi.org/10.1038/NM1117-1244
35. Fleming N (2018) How artificial intelligence is changing drug discovery spotlight /631/45 /639/705/117 /631/154 /706/703/559 n/a. Nature 557:S55–S57. https://doi.org/10.1038/D41586-018-05267-X
36. Guo J, Li B (2018) The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries. Health Equity 2:174–181. https://doi.org/10.1089/HEQ.2018.0037
37. Mehta N, Pandit A, Shukla S (2019) Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. J Biomed Inform 100:103311. https://doi.org/10.1016/J.JBI.2019.103311
38. Collins GS, Moons KGM (2019) Reporting of artificial intelligence prediction models. Lancet 393:1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6
39. Winter JS, Davidson E (2018) Big data governance of personal health information and challenges to contextual integrity. https://doi.org/101080/0197224320181542648 35:36–51. https://doi.org/10.1080/01972243.2018.1542648
40. Novak D, Riener R (2015) Control Strategies and Artificial Intelligence in Rehabilitation Robotics. AI Mag 36:23–33. https://doi.org/10.1609/AIMAG.V36I4.2614
41. Tarassoli SP (2019) Artificial intelligence, regenerative surgery, robotics? What is realistic for the future of surgery? Ann Med Surg (Lond) 41:53–55. https://doi.org/10.1016/J.AMSU.2019.04.001
42. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2:230–243. https://doi.org/10.1136/SVN-2017-000101
43. Samuel AL (2010) Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress. IBM J Res Dev 11:601–617. https://doi.org/10.1147/RD.116.0601
44. Mak KK, Lee K, Park C (2019) Applications of machine learning in addiction studies: A systematic review. Psychiatry Res 275:53–60. https://doi.org/10.1016/J.PSYCHRES.2019.03.001
45. About Linear Regression | IBM. https://www.ibm.com/topics/linear-regression. Accessed 24 Oct 2023
46. Garcia JMV, Bahloul MA, Laleg-Kirati TM (2022) A Multiple Linear Regression Model for Carotid-to-Femoral Pulse Wave Velocity Estimation Based on Schrodinger Spectrum Characterization. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2022-July:143–147. https://doi.org/10.1109/EMBC48229.2022.9871031
47. Cox DR (1958) The Regression Analysis of Binary Sequences. J R Stat Soc Series B Stat Methodol 20:215–232. https://doi.org/10.1111/J.2517-6161.1958.TB00292.X
48. Wallisch C, Bach P, Hafermann L, Klein N, Sauerbrei W, Steyerberg EW, Heinze G, Rauch G (2022) Review of guidance papers on regression modeling in statistical series of medical journals. PLoS One 17:e0262918. https://doi.org/10.1371/JOURNAL.PONE.0262918
49. Lynam AL, Dennis JM, Owen KR, Oram RA, Jones AG, Shields BM, Ferrat LA (2020) Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagnostic and Prognostic Research 2020 4:1 4:1–10. https://doi.org/10.1186/S41512-020-00075-2
50. Suresh K, Severn C, Ghosh D (2022) Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol 22:1–18. https://doi.org/10.1186/S12874-022-01679-6/FIGURES/5
51. Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37. https://doi.org/10.1007/S10115-007-0114-2/METRICS
52. Iris - UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/53/iris. Accessed 27 Oct 2023
53. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, Jaka M, French S (2017) Decision trees in epidemiological research. Emerg Themes Epidemiol 14:1–12. https://doi.org/10.1186/S12982-017-0064-4/FIGURES/6
54. Anisha, Sabharwal M, Tripathi R (2023) A Novel IoT-based Framework for Urine Infection Detection and Prediction using Ensemble Bagging Decision Tree Classifier. International Journal on Recent and Innovation Trends in Computing and Communication 11:416–423. https://doi.org/10.17762/IJRITCC.V11I5S.7081
55. Ho TK (1995) Random decision forests. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR 1:278–282. https://doi.org/10.1109/ICDAR.1995.598994
56. Jackins V, Vimal S, Kaliappan M, Lee MY (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. Journal of Supercomputing 77:5198–5219. https://doi.org/10.1007/S11227-020-03481-X/FIGURES/10

57. Benbelkacem S, Atmani B (2019) Random forests for diabetes diagnosis. 2019 International Conference on Computer and Information Sciences, ICCIS 2019. https://doi.org/10.1109/ICCISCI.2019.8716405
58. Wongvibulsin S, Wu KC, Zeger SL (2019) Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. BMC Med Res Methodol 20:1–14. https://doi.org/10.1186/S12874-019-0863-0/FIGURES/3
59. Acharjee A, Larkman J, Xu Y, Cardoso VR, Gkoutos G V. (2020) A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genomics 13:1–14. https://doi.org/10.1186/S12920-020-00826-6/FIGURES/5
60. Wallace ML, Mentch L, Wheeler BJ, Tapia AL, Richards M, Zhou S, Yi L, Redline S, Buysse DJ (2023) Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med Res Methodol 23:1–12. https://doi.org/10.1186/S12874-023-01965-X/FIGURES/3
61. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. American Statistician 46:175–185. https://doi.org/10.1080/00031305.1992.10475879
62. Hamed A, Sobhy A, Nassar H (2021) Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm. Arab J Sci Eng 46:8261–8272. https://doi.org/10.1007/S13369-020-05212-Z/TABLES/6
63. Dua D, Graff C (2017) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
64. Hu LY, Huang MW, Ke SW, Tsai CF (2016) The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus 5:1–9. https://doi.org/10.1186/S40064-016-2941-7/FIGURES/8
65. Vikramkumar, B V, Trilochan (2014) Bayes and Naive Bayes Classifier
66. Možina M, Demšar J, Kattan M, Zupan B (2004) Nomograms for visualization of naive Bayesian classifier. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3202:337–348. https://doi.org/10.1007/978-3-540-30116-5_32/COVER
67. Cortes C, Vapnik V, Saitta L (1995) Support-vector networks. Machine Learning 1995 20:3 20:273–297. https://doi.org/10.1007/BF00994018
68. Winters-Hilt S, Merat S (2007) SVM clustering. BMC Bioinformatics 8:1–12. https://doi.org/10.1186/1471-2105-8-S7-S18/FIGURES/8
69. Zhang N, Jiang Z, Li JX, Zhang D (2023) Multiple color representation and fusion for diabetes mellitus diagnosis based on back tongue images. Comput Biol Med 155:. https://doi.org/10.1016/J.COMPBIOMED.2023.106652
70. Jackins V, Vimal S, Kaliappan M, Lee MY (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. Journal of Supercomputing 77:5198–5219. https://doi.org/10.1007/S11227-020-03481-X/FIGURES/10
71. Schapire RE (2013) Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N Vapnik 37–52. https://doi.org/10.1007/978-3-642-41136-6_5/COVER
72. Hatwell J, Gaber MM, Atif Azad RM (2020) Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences. BMC Med Inform Decis Mak 20:1–25. https://doi.org/10.1186/S12911-020-01201-2/TABLES/24
73. Takemura A, Shimizu A, Hamamoto K (2010) Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the adaboost algorithm with feature selection. IEEE Trans Med Imaging 29:598–609. https://doi.org/10.1109/TMI.2009.2022630
74. Chen T, Guestrin C XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672
75. Shwartz-Ziv R, Armon A (2021) TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED
76. Moore A, Bell M (2022) XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction, a UK Biobank cohort study. medRxiv 2022.04.08.22273600. https://doi.org/10.1101/2022.04.08.22273600
77. Price J, Yamazaki T, Fujihara K, Sone H (2022) XGBoost: Interpretable Machine Learning Approach in Medicine. WSCE 2022 - 2022 5th World Symposium on Communication Engineering 109–113. https://doi.org/10.1109/WSCE56210.2022.9916029
78. An Q, Rahman S, Zhou J, Kang JJ (2023) A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. Sensors 23:. https://doi.org/10.3390/S23094178
79. Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: A tutorial. Computer (Long Beach Calif) 29:31–44. https://doi.org/10.1109/2.485891
80. Haykin S (1999) Neural networks: a comprehensive foundation by Simon Haykin. Knowl Eng Rev 13:409–412
81. Sarker IH (2021) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci 2:1–20. https://doi.org/10.1007/S42979-021-00815-1/FIGURES/6
82. Mandic D, Chambers J (2001) Recurrent neural networks for prediction: learning algorithms, architectures and stability
83. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278–2323. https://doi.org/10.1109/5.726791
84. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention Is All You Need. Adv Neural Inf Process Syst 2017-December:5999–6009
85. Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. 1139–1147
86. Kingma DP, Ba JL (2014) Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings

87. Egger J, Gsaxner C, Pepe A, Pomykala KL, Jonske F, Kurz M, Li J, Kleesiek J (2020) Medical Deep Learning -- A systematic Meta-Review. Comput Methods Programs Biomed 221:106874. https://doi.org/10.1016/j.cmpb.2022.106874
88. Hou JJ, Tian HL, Lu B (2022) A Deep Neural Network-Based Model for Quantitative Evaluation of the Effects of Swimming Training. Comput Intell Neurosci 2022:. https://doi.org/10.1155/2022/5508365
89. Singh A, Ardakani AA, Loh HW, Anamika P V., Acharya UR, Kamath S, Bhat AK (2023) Automated detection of scaphoid fractures using deep neural networks in radiographs. Eng Appl Artif Intell 122:106165. https://doi.org/10.1016/J.ENGAPPAI.2023.106165
90. Gülmez B (2022) A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images. Ann Oper Res 328:617–641. https://doi.org/10.1007/S10479-022-05151-Y/TABLES/15
91. Tsai K-J;, Chou M-C;, Li H-M;, Liu S-T;, Hsu J-H;, Yeh W-C;, Hung C-M;, Yeh C-Y;, Hwang S-H, Cao J, Bhatt C, Bhuyan MH, Ghoraani B, Prasad M, Tsai K-J, Chou M-C, Li H-M, Liu S-T, Hsu J-H, Yeh W-C, Hung C-M, Yeh C-Y, Hwang S-H (2022) A High-Performance Deep Neural Network Model for BI-RADS Classification of Screening Mammography. Sensors 2022, Vol 22, Page 1160 22:1160. https://doi.org/10.3390/S22031160
92. Rajput JS, Sharma M, Kumar TS, Acharya UR (2022) Automated Detection of Hypertension Using Continuous Wavelet Transform and a Deep Neural Network with Ballistocardiography Signals. International Journal of Environmental Research and Public Health 2022, Vol 19, Page 4014 19:4014. https://doi.org/10.3390/IJERPH19074014
93. Voigt I, Boeckmann M, Bruder O, Wolf A, Schmitz T, Wieneke H (2022) A deep neural network using audio files for detection of aortic stenosis. Clin Cardiol 45:657–663. https://doi.org/10.1002/CLC.23826
94. Ma L, Yang T (2021) Construction and Evaluation of Intelligent Medical Diagnosis Model Based on Integrated Deep Neural Network. Comput Intell Neurosci 2021:. https://doi.org/10.1155/2021/7171816
95. Ragab M, Al-Ghamdi ASAM, Fakieh B, Choudhry H, Mansour RF, Koundal D (2022) Prediction of Diabetes through Retinal Images Using Deep Neural Network. Comput Intell Neurosci 2022:. https://doi.org/10.1155/2022/7887908
96. Min JK, Yang HJ, Kwak MS, Cho CW, Kim S, Ahn KS, Park SK, Cha JM, Park D Il (2021) Deep Neural Network-Based Prediction of the Risk of Advanced Colorectal Neoplasia. Gut Liver 15:85. https://doi.org/10.5009/GNL19334
97. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK (2018) Medical Image Analysis using Convolutional Neural Networks: A Review. J Med Syst 42:1–13. https://doi.org/10.1007/S10916-018-1088-1/TABLES/4
98. Mohamed EA, Gaber T, Karam O, Rashed EA (2022) A Novel CNN pooling layer for breast cancer segmentation and classification from thermograms. PLoS One 17:e0276523. https://doi.org/10.1371/JOURNAL.PONE.0276523
99. Chamberlin J, Kocher MR, Waltz J, Snoddy M, Stringer NFC, Stephenson J, Sahbaee P, Sharma P, Rapaka S, Schoepf UJ, Abadia AF, Sperl J, Hoelzer P, Mercer M, Somayaji N, Aquino G, Burt JR (2021) Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value. BMC Med 19:1–14. https://doi.org/10.1186/S12916-021-01928-3/FIGURES/6
100. Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data 6:1–18. https://doi.org/10.1186/S40537-019-0276-2/TABLES/16
101. Sarvamangala DR, Kulkarni R V. (2022) Convolutional neural networks in medical image understanding: a survey. Evol Intell 15:1–22. https://doi.org/10.1007/S12065-020-00540-3/FIGURES/2
102. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351:234–241. https://doi.org/10.1007/978-3-319-24574-4_28/COVER
103. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS:424–432. https://doi.org/10.1007/978-3-319-46723-8_49
104. Feng X, Tustison NJ, Patel SH, Meyer CH (2020) Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. Front Comput Neurosci 14:488825. https://doi.org/10.3389/FNCOM.2020.00025/BIBTEX
105. Cui Y, Arimura H, Yoshitake T, Shioyama Y, Yabuuchi H (2023) Deep learning model fusion improves lung tumor segmentation accuracy across variable training-to-test dataset ratios. Phys Eng Sci Med 46:1271–1285. https://doi.org/10.1007/S13246-023-01295-8
106. Openai IG (2016) NIPS 2016 Tutorial: Generative Adversarial Networks
107. Skandarani Y, Jodoin PM, Lalande A (2021) GANs for Medical Image Synthesis: An Empirical Study. J Imaging 9:. https://doi.org/10.3390/jimaging9030069
108. Transfer Learning Vs. Designing CNN cons and pros. https://www.linkedin.com/pulse/transfer-learning-vs-designing-cnn-cons-pros-dr-wafaa-shalash. Accessed 24 Oct 2023
109. Plested J, Gedeon T (2022) Deep transfer learning for image classification: a survey
110. Athanasiadis I, Mousouliotis P, Petrou L (2018) A Framework of Transfer Learning in Object Detection for Embedded Systems
111. Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer Learning in Natural Language Processing. Proceedings of the 2019 Conference of the North 15–18. https://doi.org/10.18653/V1/N19-5004

112. Zoph B, Yuret D, May J, Knight K (2016) Transfer Learning for Low-Resource Neural Machine Translation. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings 1568–1575. https://doi.org/10.18653/v1/d16-1163
113. Sutherland SM, Goldstein SL, Bagshaw SM (2017) Leveraging Big Data and Electronic Health Records to Enhance Novel Approaches to Acute Kidney Injury Research and Care. Blood Purif 44:68–76. https://doi.org/10.1159/000458751
114. Thongprayoon C, Kaewput W, Kovvuru K, Hansrivijit P, Kanduri SR, Bathini T, Chewcharat A, Leeaphorn N, Gonzalez-Suarez ML, Cheungpasitporn W (2020) Promises of big data and artificial intelligence in nephrology and transplantation. J Clin Med 9
115. Zitt E, Pscheidt C, Concin H, Kramar R, Peter RS, Beyersmann J, Lhotta K, Nagel G (2018) Long-term risk for end-stage kidney disease and death in a large population-based cohort. Sci Rep 8:1–8. https://doi.org/10.1038/s41598-018-26087-z
116. Thomas R, Kanso A, Sedor JR (2008) Chronic Kidney Disease and Its Complications. Primary Care - Clinics in Office Practice 35:329–344. https://doi.org/10.1016/j.pop.2008.01.008
117. Thompson S, James M, Wiebe N, Hemmelgarn B, Manns B, Klarenbach S, Tonelli M (2015) Cause of death in patients with reduced kidney function. Journal of the American Society of Nephrology 26:2504–2511. https://doi.org/10.1681/ASN.2014070714
118. Chen TK, Knicely DH, Grams ME (2019) Chronic Kidney Disease Diagnosis and Management: A Review. JAMA - Journal of the American Medical Association 322:1294–1304. https://doi.org/10.1001/jama.2019.14745
119. Yang HC, Zuo Y, Fogo AB (2010) Models of chronic kidney disease. Drug Discov Today Dis Models 7:13–19. https://doi.org/10.1016/j.ddmod.2010.08.002
120. Thompson RH, Kurta JM, Kaag M, Tickoo SK, Kundu S, Katz D, Nogueira L, Reuter VE, Russo P (2009) Tumor Size is Associated With Malignant Potential in Renal Cell Carcinoma Cases. Journal of Urology 181:2033–2036. https://doi.org/10.1016/j.juro.2009.01.027
121. Chen J, Remulla D, Nguyen JH, Aastha D, Liu Y, Dasgupta P, Hung AJ (2019) Current status of artificial intelligence applications in urology and their potential to influence clinical practice. BJU Int 124:567–577
122. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R (2019) Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. Acad Pathol 6:. https://doi.org/10.1177/2374289519873088
123. Ballard BD, Guzman N (2022) Renal Mass. In: StatPearls [Internet]. StatPearls Publishing
124. Chen CJ, Pai TW, Fujita H, Lee CH, Chen YT, Chen KS, Chen YC (2014) Stage diagnosis for Chronic Kidney Disease based on ultrasonography. 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2014 525–530. https://doi.org/10.1109/FSKD.2014.6980889
125. Lee Y, Kim N, Cho KS, Kang SH, Dae YK, Yoon YJ, Kim JK (2009) Bayesian classifier for predicting malignant renal cysts on MDCT: Early clinical experience. American Journal of Roentgenology 193:106–111. https://doi.org/10.2214/AJR.08.1858
126. Kunapuli G, Varghese BA, Ganapathy P, Desai B, Cen S, Aron M, Gill I, Duddalwar V (2018) A Decision-Support Tool for Renal Mass Classification. J Digit Imaging 31:929–939. https://doi.org/10.1007/s10278-018-0100-0
127. Erdim C, Yardimci AH, Bektas CT, Kocak B, Koca SB, Demir H, Kilickesmez O (2020) Prediction of Benign and Malignant Solid Renal Masses: Machine Learning-Based CT Texture Analysis. Acad Radiol 27:1422–1429. https://doi.org/10.1016/j.acra.2019.12.015
128. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O (2019) Influence of segmentation margin on machine learning–based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. Eur Radiol 29:4765–4775. https://doi.org/10.1007/s00330-019-6003-8
129. Kocak B, Durmaz ES, Kaya OK, Kilickesmez O (2019) Machine learning-based unenhanced CT texture analysis for predicting BAP1 mutation status of clear cell renal cell carcinomas. Acta radiol 61:856–864
130. Cui E, Li Z, Ma C, Li Q, Lei Y, Lan Y, Yu J, Zhou Z, Li R, Long W, Lin F (2020) Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. Eur Radiol 30:2912–2921. https://doi.org/10.1007/s00330-019-06601-1
131. Bektas CT, Kocak B, Yardimci AH, Turkcanoglu MH, Yucetas U, Koca SB, Erdim C, Kilickesmez O (2018) Clear Cell Renal Cell Carcinoma: Machine Learning-Based Quantitative Computed Tomography Texture Analysis for Prediction of Fuhrman Nuclear Grade. Eur Radiol 29:1153–1163. https://doi.org/10.1007/s00330-018-5698-2
132. Azuaje F, Kim S-Y, Perez Hernandez D, Dittmar G (2019) Connecting Histopathology Imaging and Proteomics in Kidney Cancer through Machine Learning. J Clin Med 8:1535. https://doi.org/10.3390/jcm8101535
133. Wu B, Mukherjee S, Jain M (2016) A new method using multiphoton imaging and morphometric analysis for differentiating chromophobe renal cell carcinoma and oncocytoma kidney tumors. Multiphoton Microscopy in the Biomedical Sciences XVI 9712:97121O. https://doi.org/10.1117/12.2213681
134. Singh NP, Bapi RS, Vinod PK (2018) Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med 100:92–99. https://doi.org/10.1016/j.compbiomed.2018.06.030
135. Alonso-Betanzos A, Bolón-Canedo V, Morán-Fernández L, Sánchez-Maroño N (2019) A Review of Microarray Datasets: Where to Find Them and Specific Characteristics. pp 65–85
136. Isensee F, Maier-Hein KH (2019) An attempt at beating the 3D U-Net. ArXiv 1–8. https://doi.org/10.24926/548719.001
137. Hou X, Xie C, Li F, Nan Y (2019) Cascaded Semantic Segmentation for Kidney and Tumor. 2–6. https://doi.org/10.24926/548719.002

138. Mu G, Lin Z, Han M, Yao G, Gao Y (2019) Segmentation of kidney tumor by multi-resolution VB-nets. 1–5. https://doi.org/10.24926/548719.003
139. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2018) Radiology data from the Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma CPTAC-CCRCC collection
140. Genes. https://portal.gdc.cancer.gov/genes/ENSG00000143294. Accessed 25 Oct 2023
141. Yang XJ, Tan M-H, Kim HL, Ditlev JA, Betten MW, Png CE, Kort EJ, Futami K, Furge KA, Takahashi M, Kanayama H-O, Tan PH, Teh BS, Luan C, Wang K, Pins M, Tretiakova M, Anema J, Kahnoski R, Nicol T, Stadler W, Vogelzang NG, Amato R, Seligson D, Figlin R, Belldegrun A, Rogers CG, Teh BT (2005) A molecular classification of papillary renal cell carcinoma. Cancer Res 65:5628–37. https://doi.org/10.1158/0008-5472.CAN-05-0533
142. (2022) KiTS19 - Grand Challenge. In: grand-challenge.org. https://kits19.grand-challenge.org/home
143. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM (2018) Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 116:10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006
144. Penny-Dimri JC, Bergmeir C, Reid CM, Williams-Spence J, Cochrane AD, Smith JA (2021) Machine Learning Algorithms for Predicting and Risk Profiling of Cardiac Surgery-Associated Acute Kidney Injury. Semin Thorac Cardiovasc Surg 33:735–745. https://doi.org/10.1053/j.semtcvs.2020.09.028
145. Zhang Y, Yang D, Liu Z, Chen C, Ge M, Li X, Luo T, Wu Z, Shi C, Wang B, Huang X, Zhang X, Zhou S, Hei Z (2021) An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation. J Transl Med 19:1–15. https://doi.org/10.1186/s12967-021-02990-4
146. Tran NK, Sen S, Palmieri TL, Lima K, Falwell S, Wajda J, Rashidi HH (2019) Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. Burns 45:1350–1358. https://doi.org/10.1016/j.burns.2019.03.021
147. Fix E, Hodges JL (1989) Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. Int Stat Rev 57:238. https://doi.org/10.2307/1403797
148. Ibrahim NE, McCarthy CP, Shrestha S, Gaggin HK, Mukai R, Magaret CA, Rhyne RF, Januzzi JL (2019) A clinical, proteomics, and artificial intelligence-driven model to predict acute kidney injury in patients undergoing coronary angiography. Clin Cardiol 42:292–298. https://doi.org/10.1002/clc.23143
149. Tseng PY, Chen YT, Wang CH, Chiu KM, Peng Y Sen, Hsu SP, Chen KL, Yang CY, Lee OKS (2020) Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 24:1–13. https://doi.org/10.1186/s13054-020-03179-9
150. Azzalini L, Candilio L, McCullough PA, Colombo A (2017) Current Risk of Contrast-Induced Acute Kidney Injury After Coronary Angiography and Intervention: A Reappraisal of the Literature. Canadian Journal of Cardiology 33:1225–1228. https://doi.org/10.1016/j.cjca.2017.07.482
151. Connell A, Montgomery H, Martin P, Nightingale C, Sadeghi-Alavijeh O, King D, Karthikesalingam A, Hughes C, Back T, Ayoub K, Suleyman M, Jones G, Cross J, Stanley S, Emerson M, Merrick C, Rees G, Laing C, Raine R (2019) Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions. NPJ Digit Med 2:1–9. https://doi.org/10.1038/s41746-019-0100-6
152. Scanlon LA, O'hara C, Garbett A, Barker-Hewitt M, Barriuso J (2021) Developing an agnostic risk prediction model for early aki detection in cancer patients. Cancers (Basel) 13:1–12. https://doi.org/10.3390/cancers13164182
153. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, Hu Y, Liu M (2020) Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 11:1–12. https://doi.org/10.1038/s41467-020-19551-w
154. Churpek MM, Yuen TC, Winslow C, Robicsek AA, Meltzer DO, Gibbons RD, Edelson DP (2014) Multicenter Development and Validation of a Risk Stratification Tool for Ward Patients. Am J Respir Crit Care Med 190:649–655. https://doi.org/10.1164/rccm.201406-1022OC
155. Sanchez-Pinto LN, Khemani RG (2016) Development of a Prediction Model of Early Acute Kidney Injury in Critically Ill Children Using Electronic Health Record Data. Pediatric Critical Care Medicine 17:508–515. https://doi.org/10.1097/PCC.0000000000000750
156. (2022) National Database | ANZSCTS. https://anzscts.org/database
157. (2015) The CASABLANCA Study: Catheter Sampled Blood Archive in Cardiovascular Diseases - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/study/NCT00842868
158. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR The Greater Plains Collaborative: a PCORnet Clinical Research Data Network. J Am Med Inform Assoc 21:637–41. https://doi.org/10.1136/amiajnl-2014-002756
159. Facts About Chronic Kidney Disease. Accessed 7 Jun 2022
160. Botev R, Mallié J-P (2008) Reporting the eGFR and Its Implication for CKD Diagnosis. Clinical Journal of the American Society of Nephrology 3:1606–1607. https://doi.org/10.2215/CJN.04560908
161. Salekin A, Stankovic J (2016) Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016 262–270. https://doi.org/10.1109/ICHI.2016.36
162. Boukenze B, Haqiq A, Mousannif H (2017) Predicting chronic kidney failure disease using data mining techniques. Lecture Notes in Electrical Engineering 397:701–712. https://doi.org/10.1007/978-981-10-1627-1_55

163. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N (2017) Predictive analytics for chronic kidney disease using machine learning techniques. 2016 Management and Innovation Technology International Conference, MITiCON 2016 MIT80–MIT83. https://doi.org/10.1109/MITICON.2016.8025242
164. Wibawa MS, Maysanjaya IMD, Putra IMAW (2017) Boosted classifier and features selection for enhancing chronic kidney disease diagnose. 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017. https://doi.org/10.1109/CITSM.2017.8089245
165. Subasi A, Alickovic E, Kevric J (2017) Diagnosis of Chronic Kidney Disease by Using Random Forest. pp 589–594
166. Aljaaf AJ, Al-Jumeily D, Haglan HM, Alloghani M, Baker T, Hussain AJ, Mustafina J (2018) Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics. 2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings. https://doi.org/10.1109/CEC.2018.8477876
167. Vanaja R, Mukherjee S (2018) Novel wrapper-based feature selection for efficient clinical decision support system. Communications in Computer and Information Science 941:113–129. https://doi.org/10.1007/978-981-13-3582-2_9
168. Rady EHA, Anwar AS (2019) Prediction of kidney disease stages using data mining algorithms. Inform Med Unlocked 15:100178. https://doi.org/10.1016/j.imu.2019.100178
169. Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N, Uddin MI, Alahmadi AH, Jadhav ME, Alzahrani MY (2021) Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. J Healthc Eng 2021:. https://doi.org/10.1155/2021/1004767
170. Wickramasinghe MPNM, Perera DM, Kahandawaarachchi KADCP (2018) Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms. 2017 IEEE Life Sciences Conference, LSC 2017 2018-Janua:300–303. https://doi.org/10.1109/LSC.2017.8268202
171. Mitch WE, Remuzzi G (2016) Diets for patients with chronic kidney disease, should we reconsider? BMC Nephrol 17:80. https://doi.org/10.1186/s12882-016-0283-x
172. Hayashi Y, Nakajima K, Nakajima K (2019) A rule extraction approach to explore the upper limit of hemoglobin during anemia treatment in patients with predialysis chronic kidney disease. Inform Med Unlocked 17:100262. https://doi.org/10.1016/j.imu.2019.100262
173. Han H, Segal AM, Seifter JL, Dwyer JT (2015) Nutritional Management of Kidney Stones (Nephrolithiasis). Clin Nutr Res 4:137. https://doi.org/10.7762/cnr.2015.4.3.137
174. Yarnell J, O'Reilly D (2013) Epidemiology and disease prevention, 2nd ed. Oxford University Press, London, England
175. Miernik A, Hein S, Wilhelm K, Schoenthaler M (2017) Harnsteindiagnostik – Was bringt uns die Zukunft? Aktuelle Urol 48:127–131. https://doi.org/10.1055/s-0042-120468
176. Große Hokamp N, Lennartz S, Salem J, Pinto dos Santos D, Heidenreich A, Maintz D, Haneder S (2019) Dose independent characterization of renal stones by means of dual energy computed tomography and machine learning: an ex-vivo study. Eur Radiol 30:1397–1404. https://doi.org/10.1007/s00330-019-06455-7
177. De Perrot T, Hofmeister J, Burgermeister S, Martin SP, Feutry G, Klein J, Montet X (2019) Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning. Eur Radiol 29:4776–4782. https://doi.org/10.1007/s00330-019-6004-7
178. Aminsharifi A, Irani D, Pooyesh S, Parvin H, Dehghani S, Yousofi K, Fazel E, Zibaie F (2017) Artificial Neural Network System to Predict the Postoperative Outcome of Percutaneous Nephrolithotomy. J Endourol 31:461–467. https://doi.org/10.1089/end.2016.0791
179. Shabaniyan T, Parsaei H, Aminsharifi A, Movahedi MM, Jahromi AT, Pouyesh S, Parvin H (2019) An artificial intelligence-based clinical decision support system for large kidney stone treatment. Australas Phys Eng Sci Med 42:771–779. https://doi.org/10.1007/s13246-019-00780-3
180. Yang SW, Hyon YK, Na HS, Jin L, Lee JG, Park JM, Lee JY, Shin JH, Lim JS, Na YG, Jeon K, Ha T, Kim J, Song KH (2020) Machine learning prediction of stone-free success in patients with urinary stone after treatment of shock wave lithotripsy. BMC Urol 20:1–8. https://doi.org/10.1186/s12894-020-00662-x
181. (2017) Understanding Glomerular Diseases. In: National Kidney Foundation. https://www.kidney.org/atoz/content/understanding-glomerular-diseases
182. Leung RKK, Wang Y, Ma RCW, Luk AOY, Lam V, Ng M, So WY, Tsui SKW, Chan JCN (2013) Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: A prospective case-control cohort analysis. BMC Nephrol 14:. https://doi.org/10.1186/1471-2369-14-162
183. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R (2018) Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol 12:295–302. https://doi.org/10.1177/1932296817706375
184. Bennett CC (2019) Artificial intelligence for diabetes case management: The intersection of physical and mental health. Inform Med Unlocked 16:. https://doi.org/10.1016/j.imu.2019.100191
185. Schena FP, Nistor I (2018) Epidemiology of IgA Nephropathy: A Global Perspective. Semin Nephrol 38:435–442. https://doi.org/10.1016/j.semnephrol.2018.05.013
186. (2022) IgA Nephropathy. In: National Kidney Foundation. https://www.kidney.org/atoz/content/iganeph
187. Takahashi K, Kitamura S, Fukushima K, Sang Y, Tsuji K, Wada J (2021) The resolution of immunofluorescent pathological images affects diagnosis for not only artificial intelligence but also human. J Nephropathol 10:e26–e26. https://doi.org/10.34172/jnp.2021.26

188. Schena FP, Anelli VW, Trotta J, Di Noia T, Manno C, Tripepi G, D'Arrigo G, Chesnaye NC, Russo ML, Stangou M, Papagianni A, Zoccali C, Tesar V, Coppo R, Tesar V, Maixnerova D, Lundberg S, Gesualdo L, Emma F, Fuiano L, Beltrame G, Rollino C, Coppo R, Amore A, Camilla R, Peruzzi L, Praga M, Feriozzi S, Polci R, Segoloni G, Colla L, Pani A, Angioi A, Piras L, Feehally J, Cancarini G, Ravera S, Durlik M, Moggia E, Ballarin J, Di Giulio S, Pugliese F, Serriello I, Caliskan Y, Sever M, Kilicaslan I, Locatelli F, Del Vecchio L, Wetzels JFM, Peters H, Berg U, Carvalho F, da Costa Ferreira AC, Maggio M, Wiecek A, Ots-Rosenberg M, Magistroni R, Topaloglu R, Bilginer Y, D'Amico M, Stangou M, Giacchino F, Goumenos D, Papasotiriou M, Galesic K, Toric L, Geddes C, Siamopoulos K, Balafa O, Galliani M, Stratta P, Quaglia M, Bergia R, Cravero R, Salvadori M, Cirami L, Fellstrom B, Smerud HK, Ferrario F, Stellato T, Egido J, Martin C, Floege J, Eitner F, Rauen T, Lupo A, Bernich P, Menè P, Morosetti M, van Kooten C, Rabelink T, Reinders MEJ, Grinyo JMB, Cusinato S, Benozzi L, Savoldi S, Licata C, Mizerska-Wasiak M, Roszkowska-Blaim M, Martina G, Messuerotti A, Canton AD, Esposito C, Migotto C, Triolo G, Mariano F, Pozzi C, Boero R, Mazzucco, Giannakakis C, Honsova E, Sundelin B, Di Palma AM, Gutiérrez E, Asunis AM, Barratt J, Tardanico R, Perkowska-Ptasinska A, Terroba JA, Fortunato M, Pantzaki A, Ozluk Y, Steenbergen E, Soderberg M, Riispere Z, Furci L, Orhan D, Kipgen D, Casartelli D, GalesicLjubanovic D, Gakiopoulou H, Bertoni E, Ortiz PC, Karkoszka H, Groene HJ, Stoppacciaro A, Bajema I, Bruijn J, FulladosaOliveras X, Maldyk J, Ioachim E, Abbrescia D, Kouri N, Scolari F, Delbarba E, Bonomini M, Piscitani L, Stallone G, Infante B, Godeas G, Madio D, Biancone L, Campagna M, Zaza G, Squarzoni I, Cangemi C (2021) Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with immunoglobulin A nephropathy. Kidney Int 99:1179–1188. https://doi.org/10.1016/j.kint.2020.07.046
189. Konieczny A, Stojanowski J, Krajewska M, Kusztal M (2021) Machine learning in prediction of iga nephropathy outcome: A comparative approach. J Pers Med 11:. https://doi.org/10.3390/jpm11040312
190. Agar JWM, Webb GI (1992) Application of machine learning to a renal biopsy database. Nephrology Dialysis Transplantation 7:472–478. https://doi.org/10.1093/oxfordjournals.ndt.a092173
191. Iakovidis DK, Goudas T, Smailis C, Maglogiannis I (2014) Ratsnake: A versatile image annotation tool with application to computer-aided diagnosis. The Scientific World Journal 2014:. https://doi.org/10.1155/2014/286856
192. Aldeman NLS, de Sá Urtiga Aita KM, Machado VP, da Mata Sousa LCD, Coelho AGB, da Silva AS, da Silva Mendes AP, de Oliveira Neres FJ, do Monte SJH (2021) Smartpathk: a platform for teaching glomerulopathies using machine learning. BMC Med Educ 21:1–8. https://doi.org/10.1186/s12909-021-02680-1
193. Niel O, Bastard P, Boussard C, Hogan J, Kwon T, Deschênes G (2018) Artificial intelligence outperforms experienced nephrologists to assess dry weight in pediatric patients on chronic hemodialysis. Pediatric Nephrology 33:1799–1803. https://doi.org/10.1007/s00467-018-4015-2
194. Higgins R, Hathaway M, Lowe D, Lam F, Kashi H, Tan LC, Imray C, Fletcher S, Zehnder D, Chen K, Krishnan N, Hamer R, Briggs D (2007) Blood Levels of Donor-Specific Human Leukocyte Antigen Antibodies After Renal Transplantation: Resolution of Rejection in the Presence of Circulating Donor-Specific Antibody. Transplantation 84:876–884. https://doi.org/10.1097/01.tp.0000284729.39137.6e
195. Gloor JM, Stegall MD (2007) ABO incompatible kidney transplantation. Curr Opin Nephrol Hypertens 16:529–534. https://doi.org/10.1097/MNH.0b013e3282f02218
196. Khovanova N, Daga S, Shaikhina T, Krishnan N, Jones J, Zehnder D, Mitchell D, Higgins R, Briggs D, Lowe D (2015) Subclass analysis of donor HLA-specific IgG in antibody-incompatible renal transplantation reveals a significant association of IgG4 with rejection and graft failure. Transplant International 28:1405–1415. https://doi.org/10.1111/tri.12648
197. Greco R, Papalia T, Lofaro D, Maestripieri S, Mancuso D, Bonofiglio R (2010) Decisional Trees in Renal Transplant Follow-up. Transplant Proc 42:1134–1136. https://doi.org/10.1016/j.transproceed.2010.03.061
198. Brown TS, Elster EA, Stevens K, Graybill JC, Gillern S, Phinney S, Salifu MO, Jindal RM (2012) Bayesian modeling of pretransplant variables accurately predicts kidney graft survival. Am J Nephrol 36:561–569. https://doi.org/10.1159/000345552
199. Topuz K, Zengul FD, Dag A, Almehmi A, Yildirim MB (2017) Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. Decis Support Syst 106:97–109. https://doi.org/10.1016/j.dss.2017.12.004
200. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N (2017) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 52:456–462. https://doi.org/10.1016/j.bspc.2017.01.012
201. Elihimas Júnior UF, Couto JP, Pereira W, Barros De Oliveira Sá MP, Tenório De França EE, Aguiar FC, Cabral DBC, Alencar SBV, Feitosa SJDC, Claizoni Dos Santos TO, Santos Elihimas HC Dos, Alves EP, José De Carvalho Lima M, Branco Cavalcanti FC, Schwingel PA (2020) Logistic Regression Model in a Machine Learning Application to Predict Elderly Kidney Transplant Recipients with Worse Renal Function One Year after Kidney Transplant: Elderly KTbot. J Aging Res 2020:. https://doi.org/10.1155/2020/7413616
202. Shih DT, Kim SB, Chen VCP, Rosenberger JM, Pilla VL (2014) Efficient computer experiment-based optimization through variable selection. Ann Oper Res 216:287–305. https://doi.org/10.1007/s10479-012-1129-y
203. Martínez-Martínez JM, Escandell-Montero P, Barbieri C, Soria-Olivas E, Mari F, Martínez-Sober M, Amato C, Serrano López AJ, Bassi M, Magdalena-Benedito R, Stopper A, Martín-Guerrero JD, Gatti E (2014) Prediction of the hemoglobin level in hemodialysis patients using machine learning techniques. Comput Methods Programs Biomed 117:208–217. https://doi.org/10.1016/j.cmpb.2014.07.001

204. Stopper A, Amato C, Gioberge S, Giordana G, Marcelli D, Gatti E (2007) Managing Complexity at Dialysis Service Centers across Europe. Blood Purif 25:77–89. https://doi.org/10.1159/000096402
205. Lu Y, Jia Z, Zeng X, Feng C, Lu X, Duan H, Li H (2019) Renal biopsy recommendation based on text understanding. Stud Health Technol Inform 264:689–693. https://doi.org/10.3233/SHTI190311
206. Kanda E, Epureanu BI, Adachi T, Tsuruta Y, Kikuchi K, Kashihara N, Abe M, Masakane I, Nitta K (2020) Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan. PLoS One 15:1–23. https://doi.org/10.1371/journal.pone.0233491
207. Aalamifar F, Rivaz H, Cerrolaza JJ, Jago J, Safdar N, Boctor EM, Linguraru MG (2015) Classification of kidney and liver tissue using ultrasound backscatter data. Medical Imaging 2015: Ultrasonic Imaging and Tomography 9419:94190X. https://doi.org/10.1117/12.2082300
208. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag J V. (2015) Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform 53:220–228. https://doi.org/10.1016/j.jbi.2014.11.005
209. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F (2013) The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 26:1045–1057. https://doi.org/10.1007/s10278-013-9622-7
210. Higgins JC, Fitzgerald JM (2001) Evaluation of incidental renal and adrenal masses. Am Fam Physician 63:288–94, 299
211. Martin-Isla C, Campello VM, Izquierdo C, Raisi-Estabragh Z, Baeßler B, Petersen SE, Lekadir K (2020) Image-Based Cardiac Diagnosis With Machine Learning: A Review. Front Cardiovasc Med 7:. https://doi.org/10.3389/fcvm.2020.00001
212. Ng F, Kozarski R, Ganeshan B, Goh V (2013) Assessment of tumor heterogeneity by CT texture analysis: Can the largest cross-sectional area be used as an alternative to whole tumor analysis? Eur J Radiol 82:342–348. https://doi.org/10.1016/j.ejrad.2012.10.023
213. Starmans MPA, van der Voort SR, Castillo Tovar JM, Veenland JF, Klein S, Niessen WJ (2020) Radiomics. In: Zhou SK, Rueckert D, Fichtinger G (eds) Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier, pp 429–456
214. Raschka S (2018) Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. https://doi.org/10.48550/arXiv.1811.12808
215. Alnazer I, Bourdon P, Urruty T, Falou O, Khalil M, Shahin A, Fernandez-Maloigne C (2021) Recent advances in medical image processing for the evaluation of chronic kidney disease. Med Image Anal 69:101960. https://doi.org/10.1016/j.media.2021.101960
216. (2022) Computed Tomography (CT) Scans and Cancer Fact Sheet. In: National Cancer Institute. https://www.cancer.gov/about-cancer/diagnosis-staging/ct-scans-fact-sheet
217. Khan SR, Pearle MS, Robertson WG, Gambaro G, Canales BK, Doizi S, Traxer O, Tiselius H-G (2016) Kidney stones. Nat Rev Dis Primers 2:16008. https://doi.org/10.1038/nrdp.2016.8
218. Burlacu A, Iftene A, Jugrin D, Popa IV, Lupu PM, Vlad C, Covic A (2020) Using Artificial Intelligence Resources in Dialysis and Kidney Transplant Patients: A Literature Review. Biomed Res Int 2020:1–14. https://doi.org/10.1155/2020/9867872
219. Thishya K, Vattam KK, Naushad SM, Raju SB, Kutala VK (2018) Artificial neural network model for predicting the bioavailability of tacrolimus in patients with renal transplantation. PLoS One 13:e0191921. https://doi.org/10.1371/journal.pone.0191921
220. Stachowska E, Gutowska I, Strzelczak A, Wesołowska T, Safranow K, Chlubek D (2006) The Use of Neural Networks in Evaluation of the Direction and Dynamics of Changes in Lipid Parameters in Kidney Transplant Patients on the Mediterranean Diet. Journal of Renal Nutrition 16:150–159. https://doi.org/10.1053/j.jrn.2006.01.003
221. Baigent C, Herrington WG, Coresh J, Landray MJ, Levin A, Perkovic V, Pfeffer MA, Rossing P, Walsh M, Wanner C, Wheeler DC, Winkelmayer WC, McMurray JJV, Abu-Alfa A, Archdeacon P, Block GA, Caskey FJ, Cheung AK, Cooper B, Craig JC, Dember LM, Eknoyan G, Gansevoort RT, Gill JS, Gillespie B, Greene T, Harris DC, Haynes R, Hemmelgarn BR, Herzog CA, Hiemstra TF, Inker LA, Jardine MJ, Jha V, Jiang L, Johansen KL, Kewalramani R, Lambers Heerspink HJ, Lefkowitz M, Lok CE, Loud F, Mačiulaitis R, Maddux DW, Maddux FW, Madero M, Mariz S, Mauer M, Nally J V., Nangaku M, Okpechi IG, Parfrey PS, Pecoits-Filho R, Pereira BJG, Rocco M V., Rossignol P, Schaefer F, Tentori F, Thompson A, Tonelli M, Tong A, Toto RD, Tuttle KR, Vetter T, Moon Wang AY, Zannad F (2017) Challenges in conducting clinical trials in nephrology: conclusions from a Kidney Disease—Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int 92:297–305. https://doi.org/10.1016/j.kint.2017.04.019
222. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R (2020) A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc 2020:191–200
223. Yuan Q, Zhang H, Deng T, Tang S, Yuan X, Tang W, Xie Y, Ge H, Wang X, Zhou Q, Xiao X (2020) Role of Artificial Intelligence in Kidney Disease. Int J Med Sci 17:970–984. https://doi.org/10.7150/ijms.42078
224. Gameiro J, Branco T, Lopes JA (2020) Artificial Intelligence in Acute Kidney Injury Risk Prediction. J Clin Med 9:678. https://doi.org/10.3390/jcm9030678
225. Shortliffe EH, Sepúlveda MJ (2018) Clinical Decision Support in the Era of Artificial Intelligence. JAMA 320:2199–2200. https://doi.org/10.1001/JAMA.2018.17163

226. Obermeyer Z, Lee TH (2017) Lost in Thought — The Limits of the Human Mind and the Future of Medicine. New England Journal of Medicine 377:1209–1211. https://doi.org/10.1056/NEJMp1705348
227. Gampala S, Vankeshwaram V, Gadula SSP (2020) Is Artificial Intelligence the New Friend for Radiologists? A Review Article. Cureus. https://doi.org/10.7759/cureus.11137
228. Mistry NS, Koyner JL (2021) Artificial Intelligence in Acute Kidney Injury: From Static to Dynamic Models. Adv Chronic Kidney Dis 28:74–82. https://doi.org/10.1053/j.ackd.2021.03.002
229. Chipidza FE, Wallwork RS, Stern TA (2015) Impact of the Doctor-Patient Relationship. Prim Care Companion CNS Disord. https://doi.org/10.4088/PCC.15f01840
230. Neri E, Coppola F, Miele V, Bibbolino C, Grassi R (2020) Artificial intelligence: Who is responsible for the diagnosis? Radiol Med 125:517–521. https://doi.org/10.1007/s11547-020-01135-9
231. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249. https://doi.org/10.3322/CAAC.21660
232. renal cell carcinoma | Monarch Initiative. https://monarchinitiative.org/MONDO:0005086. Accessed 25 Oct 2023
233. kidney oncocytoma | Monarch Initiative. https://monarchinitiative.org/MONDO:0003825. Accessed 25 Oct 2023
234. Oncocytoma | Risk Factors, Signs, Symptoms, Diagnosis and Staging. https://www.knowcancer.com/oncology/oncocytoma/. Accessed 25 Oct 2023
235. Kay FU, Pedrosa I (2018) Imaging of Solid Renal Masses. Urologic Clinics of North America 45:311–330. https://doi.org/10.1016/j.ucl.2018.03.013
236. Jones J, D'Souza D (2008) Renal oncocytoma. Radiopaedia.org. https://doi.org/10.53347/RID-1969
237. Garg S, Bhagyashree SR (2020) Detection and Classification of Tumors Using Medical Imaging Techniques: A Survey. Lecture Notes on Data Engineering and Communications Technologies 33:363–372. https://doi.org/10.1007/978-3-030-28364-3_35/FIGURES/4
238. Shen SS, Ro JY (2019) Histologic Diagnosis of Renal Mass Biopsy. Arch Pathol Lab Med 143:705–710. https://doi.org/10.5858/ARPA.2018-0272-RA
239. Kidney Cancer Surgery | American Cancer Society. https://www.cancer.org/cancer/types/kidney-cancer/treating/surgery.html. Accessed 25 Oct 2023
240. Lopes Vendrami C, Parada Villavicencio C, Dejulio TJ, Chatterjee A, Casalino DD, Horowitz JM, Oberlin DT, Yang GY, Nikolaidis P, Miller FH (2017) Differentiation of Solid Renal Tumors with Multiparametric MR Imaging. https://doi.org/101148/rg2017170039 37:2026–2042. https://doi.org/10.1148/RG.2017170039
241. Latif J, Xiao C, Imran A, Tu S (2019) Medical imaging using machine learning and deep learning algorithms: A review. 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, iCoMET 2019. https://doi.org/10.1109/ICOMET.2019.8673502
242. Saha A, Tso S, Rabski J, Sadeghian A, Cusimano MD (2020) Machine learning applications in imaging analysis for patients with pituitary tumors: a review of the current literature and future directions. Pituitary 23:273–293. https://doi.org/10.1007/S11102-019-01026-X/TABLES/2
243. Windisch P, Koechli C, Rogers S, Schröder C, Förster R, Zwahlen DR, Bodis S (2022) Machine Learning for the Detection and Segmentation of Benign Tumors of the Central Nervous System: A Systematic Review. Cancers (Basel) 14:2676. https://doi.org/10.3390/CANCERS14112676/S1
244. Magherini R, Mussi E, Volpe Y, Furferi R, Buonamici F, Servi M (2022) Machine Learning for Renal Pathologies: An Updated Survey. Sensors 2022, Vol 22, Page 4989 22:4989. https://doi.org/10.3390/S22134989
245. Kocak B, Kus EA, Yardimci AH, Bektas CT, Kilickesmez O (2020) Machine Learning in Radiomic Renal Mass Characterization: Fundamentals, Applications, Challenges, and Future Directions. American Journal of Roentgenology 215:920–928. https://doi.org/10.2214/AJR.19.22608
246. Bhandari A, Ibrahim M, Sharma C, Liong R, Gustafson S, Prior M (2021) CT-based radiomics for differentiating renal tumours: a systematic review. Abdominal Radiology 46:2052–2063. https://doi.org/10.1007/s00261-020-02832-9
247. Raman SP, Chen Y, Schroeder JL, Huang P, Fishman EK (2014) CT Texture Analysis of Renal Masses: Pilot Study Using Random Forest Classification for Prediction of Pathology. Acad Radiol 21:1587–1596. https://doi.org/10.1016/J.ACRA.2014.07.023
248. Yu HS, Scalera J, Khalid M, Touret AS, Bloch N, Li B, Qureshi MM, Soto JA, Anderson SW (2017) Texture analysis as a radiomic marker for differentiating renal tumors. Abdominal Radiology 42:2470–2478. https://doi.org/10.1007/S00261-017-1144-1/TABLES/4
249. Sun XY, Feng QX, Xu X, Zhang J, Zhu FP, Yang YH, Zhang YD (2020) Radiologic-radiomic machine learning models for differentiation of benign and malignant solid renal masses: Comparison with expert-level radiologists. American Journal of Roentgenology 214:W44–W54. https://doi.org/10.2214/AJR.19.21617
250. Brunner HI, Giannini EH (2011) Chapter 7 – TRIAL DESIGN, MEASUREMENT, AND ANALYSIS OF CLINICAL INVESTIGATIONS. Textbook of Pediatric Rheumatology 127–156. https://doi.org/10.1016/B978-1-4160-6581-4.10007-X
251. Hoang UN, Mirmomen SM, Meirelles O, Yao J, Merino M, Metwalli A, Linehan WM, Malayeri AA (2018) Assessment of multiphasic contrast-enhanced mr textures in differentiating small renal mass subtypes. Abdominal Radiology 43:3400–3409. https://doi.org/10.1007/S00261-018-1625-X/FIGURES/6

252. Coy H, Hsieh K, Wu W, Nagarajan MB, Young JR, Douek ML, Brown MS, Scalzo F, Raman SS (2019) Deep learning and radiomics: the utility of Google TensorFlow™ Inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic CT. Abdominal Radiology 44:2009–2020. https://doi.org/10.1007/S00261-019-01929-0/TABLES/3
253. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December:2818–2826. https://doi.org/10.1109/CVPR.2016.308
254. Xi IL, Zhao Y, Wang R, Chang M, Purkayastha S, Chang K, Huang RY, Silva AC, Valliéres M, Habibollahi P, Fan Y, Zou B, Gade TP, Zhang PJ, Soulen MC, Zhang Z, Bai HX, Stavropoulos SW (2020) Deep learning to distinguish benign from malignant renal lesions based on routine MR imaging. Clinical Cancer Research 26:1944–1952. https://doi.org/10.1158/1078-0432.CCR-19-0374/75136/AM/DEEP-LEARNING-TO-DISTINGUISH-BENIGN-FROM-MALIGNANT
255. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December:770–778. https://doi.org/10.1109/CVPR.2016.90
256. Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, Kavouridis VK, Senders JT, Boaro A, Beers A, Zhang B, Capellini A, Liao W, Shen Q, Li X, Xiao B, Cryan J, Ramkissoon S, Ramkissoon L, Ligon K, Wen PY, Bindra RS, Woo J, Arnaout O, Gerstner ER, Zhang PJ, Rosen BR, Yang L, Huang RY, Kalpathy-Cramer J (2018) Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from mr imaging. Clinical Cancer Research 24:1073–1081. https://doi.org/10.1158/1078-0432.CCR-17-2236/87333/AM/RESIDUAL-CONVOLUTIONAL-NEURAL-NETWORK-FOR
257. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, Nan Y, Mu G, Lin Z, Han M, Yao G, Gao Y, Zhang Y, Wang Y, Hou F, Yang J, Xiong G, Tian J, Zhong C, Ma J, Rickman J, Dean J, Stai B, Tejpaul R, Oestreich M, Blake P, Kaluzniak H, Raza S, Rosenberg J, Moore K, Walczak E, Rengel Z, Edgerton Z, Vasdev R, Peterson M, McSweeney S, Peterson S, Kalapara A, Sathianathen N, Weight C, Papanikolopoulos N (2019) The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge. ArXiv
258. neheller/kits21: The official repository of the 2021 Kidney and Kidney Tumor Segmentation Challenge. https://github.com/neheller/kits21. Accessed 25 Oct 2023
259. Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M, Dean J, Tradewell M, Shah A, Tejpaul R, Edgerton Z, Peterson M, Raza S, Regmi S, Papanikolopoulos N, Weight C (2019) The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes
260. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, Cook G (2020) Introduction to radiomics. Journal of Nuclear Medicine 61:488–495. https://doi.org/10.2967/JNUMED.118.222893
261. Jia W, Sun M, Lian J, Hou S (2022) Feature dimensionality reduction: a review. Complex & Intelligent Systems 2022 8:3 8:2663–2693. https://doi.org/10.1007/S40747-021-00637-X
262. Sanz H, Valim C, Vegas E, Oller JM, Reverter F (2018) SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics 19:1–18. https://doi.org/10.1186/S12859-018-2451-4/FIGURES/16
263. Altman DG, Bland JM (1994) Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ 308:1552. https://doi.org/10.1136/BMJ.308.6943.1552
264. Dodziuk H (2016) Applications of 3D printing in healthcare. Kardiochirurgia i Torakochirurgia Polska/Polish Journal of Thoracic and Cardiovascular Surgery 13:283–293. https://doi.org/10.5114/KITP.2016.62625
265. AlAli AB, Griffin MF, Butler PE (2015) Three-Dimensional Printing Surgical Applications. Eplasty 15:e37
266. Nagata S (1993) A new method of total reconstruction of the auricle for microtia. Plast Reconstr Surg 92:187–201. https://doi.org/10.1097/00006534-199308000-00001
267. Storck K, Staudenmaier R, Buchberger M, Strenger T, Kreutzer K, Von Bomhard A, Stark T (2014) Total reconstruction of the auricle: our experiences on indications and recent techniques. Biomed Res Int 2014:. https://doi.org/10.1155/2014/373286
268. Facchini F, Morabito A, Buonamici F, Mussi E, Servi M, Volpe Y (2021) Autologous Ear Reconstruction: Towards a Semiautomatic CAD-based Procedure for 3D Printable Surgical Guides. Comput Aided Des Appl 18:357–367. https://doi.org/10.14733/cadaps.2021.357-367
269. Zhu P, Chen S (2016) Clinical outcomes following ear reconstruction with adjuvant 3D template model. Acta Otolaryngol 136:1236–1241. https://doi.org/10.1080/00016489.2016.1206967
270. Mussi E, Servi M, Facchini F, Carfagni M, Volpe Y (2021) A computer-aided strategy for preoperative simulation of autologous ear reconstruction procedure. International Journal on Interactive Design and Manufacturing 15:77–80. https://doi.org/10.1007/S12008-020-00723-3/FIGURES/4
271. Alonso Jr M (2019) Y-GAN: A Generative Adversarial Network for Depthmap Estimation from Multi-camera Stereo Images
272. Gwn K, Reddy K, Giering M, Bernal EA (2018) Generative adversarial networks for depth map estimation from RGB video. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2018-June:1258–1266. https://doi.org/10.1109/CVPRW.2018.00163
273. Eigen D, Puhrsch C, Fergus R (2014) Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. Adv Neural Inf Process Syst 3:2366–2374
274. Kwak DH, Lee SH (2020) A Novel Method for Estimating Monocular Depth Using Cycle GAN and Segmentation. Sensors 2020, Vol 20, Page 2567 20:2567. https://doi.org/10.3390/S20092567
275. Kwon Y-H, Park M-G (2019) Predicting Future Frames Using Retrospective Cycle GAN. 1811–1820

276. Niklaus S, Mai L, Yang J, Liu F (2019) 3D Ken Burns effect from a single image. ACM Transactions on Graphics (TOG) 38:. https://doi.org/10.1145/3355089.3356528
277. Yan P, Bowyer KW (2005) Empirical evaluation of advanced ear biometrics. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2005-September: https://doi.org/10.1109/CVPR.2005.450
278. Yan P, Bowyer KW (2006) An automatic 3D ear recognition system. Proceedings - Third International Symposium on 3D Data Processing, Visualization, and Transmission, 3DPVT 2006 326–333. https://doi.org/10.1109/3DPVT.2006.25
279. Bizjak M, Peer P, Emeršic Ž (2019) Mask R-CNN for ear detection. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings 1624–1628. https://doi.org/10.23919/MIPRO.2019.8756760
280. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y Generative Adversarial Nets
281. Chen X, Gupta A (2017) An Implementation of Faster RCNN with Study for Region Sampling
282. HumanSignal/labelImg: LabelImg is now part of the Label Studio community. The popular image annotation tool created by Tzutalin is no longer actively being developed, but you can check out Label Studio, the open source data labeling tool for images, text, hypertext, audio, video and time-series data. https://github.com/HumanSignal/labelImg. Accessed 25 Oct 2023
283. Zhang E, Zhang Y (2009) Average Precision. Encyclopedia of Database Systems 192–193. https://doi.org/10.1007/978-0-387-39940-9_482
284. Shmelkov K, Schmid C, Alahari K (2018) How Good Is My GAN? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11206 LNCS:218–234. https://doi.org/10.1007/978-3-030-01216-8_14/TABLES/4
285. Zhang L, Wu X (2005) Color demosaicking via directional linear minimum mean square-error estimation. IEEE Transactions on Image Processing 14:2167–2178. https://doi.org/10.1109/TIP.2005.857260
286. Dosselmann R, Yang XD (2011) A comprehensive assessment of the structural similarity index. Signal Image Video Process 5:81–91. https://doi.org/10.1007/S11760-009-0144-1/METRICS
287. Mussi E, Servi M, Facchini F, Furferi R, Governi L, Volpe Y (2021) A novel ear elements segmentation algorithm on depth map images. Comput Biol Med 129:104157. https://doi.org/10.1016/J.COMPBIOMED.2020.104157
288. Download Head CT CQ500 dataset. http://headctstudy.qure.ai/dataset. Accessed 25 Oct 2023
289. Orthopaedic and CMF Solutions for Medical Device Companies. https://www.materialise.com/en/healthcare/medical-device-companies/ortho-cmf. Accessed 25 Oct 2023
290. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
291. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016 265–283
292. World Health Organization Preventing suicide: A global imperative. https://www.who.int/publications/i/item/9789241564779. Accessed 26 Oct 2023
293. Cha CB, Franz PJ, M. Guzmán E, Glenn CR, Kleiman EM, Nock MK (2018) Annual Research Review: Suicide among youth – epidemiology, (potential) etiology, and treatment. Journal of Child Psychology and Psychiatry 59:460–482. https://doi.org/10.1111/JCPP.12831
294. Andover MS, Morris BW, Wren A, Bruzzese ME (2012) The co-occurrence of non-suicidal self-injury and attempted suicide among adolescents: distinguishing risk factors and psychosocial correlates. Child Adolesc Psychiatry Ment Health 6:. https://doi.org/10.1186/1753-2000-6-11
295. Centers for Disease Control and Prevention. https://www.cdc.gov/index.htm. Accessed 26 Oct 2023
296. Naghavi M (2019) Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the Global Burden of Disease Study 2016. BMJ 364:l94. https://doi.org/10.1136/BMJ.L94
297. Harmer B, Lee S, Duong T vi H, Saadabadi A (2023) Suicidal Ideation. Acute Medicine: A Symptom-Based Approach 415–420. https://doi.org/10.1007/9781139600354.061
298. World Health Organization (2021) Comprehensive Mental Health Action Plan 2013-2030. https://www.who.int/publications/i/item/9789240031029. Accessed 26 Oct 2023
299. Carson Id NJ, Mullin B, Jose Sanchez Id M, Id FL, Yang K, Menezes M, Cook BL (2019) Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. https://doi.org/10.1371/journal.pone.0211116
300. Navarro MC, Ouellet-Morin I, Geoffroy MC, Boivin M, Tremblay RE, Côté SM, Orri M (2021) Machine Learning Assessment of Early Life Factors Predicting Suicide Attempt in Adolescence or Young Adulthood. JAMA Netw Open 4:. https://doi.org/10.1001/JAMANETWORKOPEN.2021.1450
301. George A, Johnson D, Carenini G, Eslami A, Ng R, Portales-Casamar E (2021) Applications of Aspect-based Sentiment Analysis on Psychiatric Clinical Notes to Study Suicide in Youth. AMIA Summits on Translational Science Proceedings 2021:229

302. Fortuna LR (2021) Editorial: Disrupting Pathways to Self-Harm in Adolescence: Machine Learning as an Opportunity. J Am Acad Child Adolesc Psychiatry 60:1459–1460. https://doi.org/10.1016/J.JAAC.2021.05.004
303. Nordin N, Zainol Z, Mohd Noor MH, Chan LF (2022) Suicidal behaviour prediction models using machine learning techniques: A systematic review. Artif Intell Med 132:102395. https://doi.org/10.1016/J.ARTMED.2022.102395
304. Plener PL, Kaess M, Schmahl C, Pollak S, Fegert JM, Brown RC (2018) Non-suicidal self-injury in adolescents. Dtsch Arztebl Int 115:23–30. https://doi.org/10.3238/ARZTEBL.2018.0023
305. P S, FA S, N G, DR B (2006) Failed suicide and deliberate self-harm: A need for specific nomenclature. Indian J Psychiatry 48:78. https://doi.org/10.4103/0019-5545.31594
306. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP (2016) ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355:. https://doi.org/10.1136/BMJ.I4919
307. Herzog JI, Schmahl C (2018) Adverse Childhood Experiences and the Consequences on Neurobiological, Psychosocial, and Somatic Conditions Across the Lifespan. Front Psychiatry 9:. https://doi.org/10.3389/FPSYT.2018.00420
308. Pisano T, Gori S, De Luca L, Fiorentini G, Minghetti S, Nocentini A, Menesini E (2023) Peer victimization and developmental psychopathology in childhood and adolescence Italian psychiatric emergency unit. A single center retrospective observational study. Psychol Health Med 28:2147–2155. https://doi.org/10.1080/13548506.2020.1810721
309. Kaufman Joan (2019) K-SADS-PL DSM-5® : intervista diagnostica per la valutazione dei disturbi psicopatologici in bambini e adolescenti
310. American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association
311. Posner K, Brown GK, Stanley B, Brent DA, Yershova K V., Oquendo MA, Currier GW, Melvin GA, Greenhill L, Shen S, Mann JJ (2011) The Columbia-suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. American Journal of Psychiatry 168:1266–1277. https://doi.org/10.1176/APPI.AJP.2011.10111704/ASSET/IMAGES/LARGE/AJP_168_12_1266_F004.JPEG
312. McHugh ML (2013) The chi-square test of independence. Biochem Med (Zagreb) 23:143–149. https://doi.org/10.11613/BM.2013.018
313. Shih JH, Fay MP (2017) Pearson's chi-square test and rank correlation inferences for clustered data. Biometrics 73:822–834. https://doi.org/10.1111/BIOM.12653
314. Ayat S, Farahani HA, Aghamohamadi M, Alian M, Aghamohamadi S, Kazemi Z A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. https://doi.org/10.1007/s00521-012-1086-z
315. Rojas R (1996) The Backpropagation Algorithm. Neural Networks 149–182. https://doi.org/10.1007/978-3-642-61068-4_7
316. Granitto PM, Furlanello C, Biasioli F, Gasperi F (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems 83:83–90. https://doi.org/10.1016/J.CHEMOLAB.2006.01.007
317. Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26:217–222. https://doi.org/10.1080/01431160412331269698
318. Epidemiology of Impulse Control Disorders and Association With Dopamine Agonist Exposure, Active Component, U.S. Armed Forces, 2014–2018 | Health.mil. https://www.health.mil/News/Articles/2019/08/01/Epidemiology-of-Impulse-Control-Disorders. Accessed 26 Oct 2023
319. Martin G, Richardson AS, Bergen HA, Roeger L, Allison S (2005) Perceived academic performance, self-esteem and locus of control as indicators of need for assessment of adolescent suicide risk: implications for teachers. J Adolesc 28:75–87. https://doi.org/10.1016/j.adolescence.2004.04.005
320. Brådvik L (2018) Suicide Risk and Mental Disorders. Int J Environ Res Public Health 15:. https://doi.org/10.3390/IJERPH15092028
321. Cibis A, Mergl R, Bramesfeld A, Althaus D, Niklewski G, Schmidtke A, Hegerl U (2012) Preference of lethal methods is not the only cause for higher suicide rates in males. J Affect Disord 136:9–16. https://doi.org/10.1016/J.JAD.2011.08.032
322. Miché M, Hofer PD, Voss C, Meyer AH, Gloster AT, Beesdo-Baum K, Lieb R (2018) Mental disorders and the risk for the subsequent first suicide attempt: results of a community study on adolescents and young adults. Eur Child Adolesc Psychiatry 27:839–848. https://doi.org/10.1007/S00787-017-1060-5/TABLES/2
323. Haavisto A, Sourander A, Ellilä H, Välimäki M, Santalahti P, Helenius H (2003) Suicidal ideation and suicide attempts among child and adolescent psychiatric inpatients in Finland. J Affect Disord 76:211–221. https://doi.org/10.1016/S0165-0327(02)00093-9
324. Stoyanov GS, Lyutfi E, Georgieva R, Georgiev R, Dzhenkov DL, Petkova L, Ivanov BD, Kaprelyan A, Ghenev P (2022) Reclassification of Glioblastoma Multiforme According to the 2021 World Health Organization Classification of Central Nervous System Tumors: A Single Institution Report and Practical Significance. https://doi.org/10.7759/cureus.21822
325. Rong L, Li N, Zhang Z (2022) Emerging therapies for glioblastoma: current state and future directions. Journal of Experimental & Clinical Cancer Research 2022 41:1 41:1–18. https://doi.org/10.1186/S13046-022-02349-7

326. Casanova-Carvajal O, Urbano-Bojorge AL, Ramos M, Serrano-Olmedo JJ, Martínez-Murillo R (2019) Slowdown intracranial glioma progression by optical hyperthermia therapy: study on a CT-2A mouse astrocytoma model. Nanotechnology 30:355101. https://doi.org/10.1088/1361-6528/ab2164
327. Gul S, Khan MS, Bibi A, Khandakar A, Ayari MA, Chowdhury MEH (2022) Deep learning techniques for liver and liver tumor segmentation: A review. Comput Biol Med 147:105620. https://doi.org/10.1016/J.COMPBIOMED.2022.105620
328. Rao CS, Karunakara K (2021) A comprehensive review on brain tumor segmentation and classification of MRI images. Multimed Tools Appl 80:17611–17643. https://doi.org/10.1007/S11042-020-10443-1/TABLES/4
329. Magherini R, Mussi E, Volpe Y, Furferi R, Buonamici F, Servi M (2022) Machine Learning for Renal Pathologies: An Updated Survey. Sensors 2022, Vol 22, Page 4989 22:4989. https://doi.org/10.3390/S22134989
330. Ma D, Cardoso MJ, Modat M, Powell N, Wells J, Holmes H, Wiseman F, Tybulewicz V, Fisher E, Lythgoe MF, Ourselin S (2014) Automatic Structural Parcellation of Mouse Brain MRI Using Multi-Atlas Label Fusion. PLoS One 9:e86576. https://doi.org/10.1371/JOURNAL.PONE.0086576
331. Hsu LM, Wang S, Walton L, Wang TWW, Lee SH, Shih YYI (2021) 3D U-Net Improves Automatic Brain Extraction for Isotropic Rat Brain Magnetic Resonance Imaging Data. Front Neurosci 15:801008. https://doi.org/10.3389/FNINS.2021.801008/BIBTEX
332. De Feo R, Hämäläinen E, Manninen E, Immonen R, Valverde JM, Ndode-Ekane XE, Gröhn O, Pitkänen A, Tohka J (2022) Convolutional Neural Networks Enable Robust Automatic Segmentation of the Rat Hippocampus in MRI After Traumatic Brain Injury. Front Neurol 13:820267. https://doi.org/10.3389/FNEUR.2022.820267/FULL
333. Valverde JM, Shatillo A, De Feo R, Tohka J (2023) Automatic Cerebral Hemisphere Segmentation in Rat MRI with Ischemic Lesions via Attention-based Convolutional Neural Networks. Neuroinformatics 21:57–70. https://doi.org/10.1007/S12021-022-09607-1
334. Yu Z, Han X, Xu W, Zhang J, Marr C, Shen D, Peng T, Zhang XY, Feng J (2022) A generalizable brain extraction net (BEN) for multimodal MRI data from rodents, nonhuman primates, and humans. Elife 11:. https://doi.org/10.7554/ELIFE.81217
335. Yu Z, Han X, Zhang S, Feng J, Peng T, Zhang XY (2023) MouseGAN++: Unsupervised Disentanglement and Contrastive Representation for Multiple MRI Modalities Synthesis and Structural Segmentation of Mouse Brain. IEEE Trans Med Imaging 42:1197–1209. https://doi.org/10.1109/TMI.2022.3225528
336. Koch BL, Hamilton BE, Hudgins PA, Ric Harnsberger H (2017) Diagnostic Imaging: Head and Neck. Elsevier
337. 3D Slicer image computing platform | 3D Slicer. https://www.slicer.org/. Accessed 26 Oct 2023
338. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-net: Learning dense volumetric segmentation from sparse annotation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS:424–432. https://doi.org/10.1007/978-3-319-46723-8_49/TABLES/3
339. Myronenko A 3D MRI brain tumor segmentation using autoencoder regularization
340. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, Murrey B, Myronenko A, Zhao C, Yang D, Nath V, He Y, Xu Z, Hatamizadeh A, Myronenko A, Zhu W, Liu Y, Zheng M, Tang Y, Yang I, Zephyr M, Hashemian B, Alle S, Darestani MZ, Budd C, Modat M, Vercauteren T, Wang G, Li Y, Hu Y, Fu Y, Gorman B, Johnson H, Genereaux B, Erdal BS, Gupta V, Diaz-Pinto A, Dourson A, Maier-Hein L, Jaeger PF, Baumgartner M, Kalpathy-Cramer J, Flores M, Kirby J, Cooper LAD, Roth HR, Xu D, Bericat D, Floca R, Zhou SK, Shuaib H, Farahani K, Maier-Hein KH, Aylward S, Dogra P, Ourselin S, Feng A (2022) MONAI: An open-source framework for deep learning in healthcare
341. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, Wasserthal J, Köhler G, Norajitra T, Wirkert S, Maier-Hein KH (2018) nnU-Net: Self-adapting framework for u-net-based medical image segmentation. ArXiv
342. Li K, Kong L, Zhang Y (2020) 3D U-Net Brain Tumor Segmentation Using VAE Skip Connection. 2020 IEEE 5th International Conference on Image, Vision and Computing, ICIVC 2020 97–101. https://doi.org/10.1109/ICIVC50857.2020.9177441