Computers in Biology and Medicine 164 (2023) 107295

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Early diagnosis and personalised treatment focusing on synthetic data modelling: Novel visual learning approach in healthcare

Ahsanullah Yunas Mahmoud *, Daniel Neagu, Daniele Scrimieri, Amr Rashad Ahmed Abdullatif

Faculty of Engineering and Informatics, University of Bradford, Bradford, England, United Kingdom

ARTICLE INFO

Keywords:
Personalised and early diagnosis
Machine learning
Imbalanced UCI data
Generative Adversarial Network
Random Forest
Synthetic data
Visualisations
Healthcare

ABSTRACT

The early diagnosis and personalised treatment of diseases are facilitated by machine learning. The quality of data has an impact on diagnosis because medical data are usually sparse, imbalanced, and contain irrelevant attributes, resulting in suboptimal diagnosis. To address the impacts of data challenges, improve resource allocation, and achieve better health outcomes, a novel visual learning approach is proposed. This study contributes to the visual learning approach by determining whether less or more synthetic data are required to improve the quality of a dataset, such as the number of observations and features, according to the intended personalised treatment and early diagnosis. In addition, numerous visualisation experiments are conducted, including using statistical characteristics, cumulative sums, histograms, correlation matrix, root mean square error, and principal component analysis in order to visualise both original and synthetic data to address the data challenges. Real medical datasets for cancer, heart disease, diabetes, cryotherapy and immunotherapy are selected as case studies. As a benchmark and point of classification comparison in terms of such as accuracy, sensitivity, and specificity, several models are implemented such as k-Nearest Neighbours and Random Forest. To simulate algorithm implementation and data, Generative Adversarial Network is used to create and manipulate synthetic data, whilst, Random Forest is implemented to classify the data. An amendable and adaptable system is constructed by combining Generative Adversarial Network and Random Forest models. The system model presents working steps, overview and flowchart. Experiments reveal that the majority of data-enhancement scenarios allow for the application of visual learning in the first stage of data analysis as a novel approach. To achieve meaningful adaptable synergy between appropriate quality data and optimal classification performance while maintaining statistical characteristics, visual learning provides researchers and practitioners with practical human-in-the-loop machine learning visualisation tools. Prior to implementing algorithms, the visual learning approach can be used to actualise early, and personalised diagnosis. For the immunotherapy data, the Random Forest performed best with precision, recall, f-measure, accuracy, sensitivity, and specificity of 81%, 82%, 81%, 88%, 95%, and 60%, as opposed to 91%, 96%, 93%, 93%, 96%, and 73% for synthetic data, respectively. Future studies might examine the optimal strategies to balance the quantity and quality of medical data.

1. Introduction

Machine learning and data-driven analytics frequently use supervised and unsupervised approaches [1,2], such as classification [3] or clustering machine learning algorithms, to extract beneficial patterns for disease diagnosis. Health data are challenging because these can be small and imbalanced, containing irrelevant attributes, which affect the performance of the applied algorithms. In addition, data analysts only consider preprocessing, cleaning, and preparation of the data before moving on to feature selection or algorithm implementation [4]. Therefore, starting the data mining process by assessing other crucial visualisation methods could save both time and resources. The absence of diagnosis results judgement by human experts has created a gap in the published literature.

Combining the skills of human medical experts and the capabilities of a data-driven perspective may reveal more accurate diagnoses, improved patterns, and insightful efficient solutions. Data are at the heart of machine learning, and the better the understanding of the data, the better the performance. Therefore, it is important to concentrate on finding and addressing the fundamental causes of data challenges. Furthermore, to identify the improved balance between data quality and classification efficiency, a novel visual learning approach is proposed in this study.

* Corresponding author.
E-mail address: A.Y.Mahmoud@bradford.ac.uk (A.Y. Mahmoud).

https://doi.org/10.1016/j.compbiomed.2023.107295
Received 12 April 2023; Received in revised form 26 July 2023; Accepted 28 July 2023
Available online 2 August 2023

Experiments are conducted to uncover similarities and differences between visualisations of the original and synthetic data, the visualisations are compared and contrasted, including statistical characteristics such as means and standard deviations, cumulative sums, histograms, principal component analysis (PCA), correlation matrix, and root mean square error (RMSE). Adapting visualisation experiments will facilitate the proposal of better algorithms as well as the improvement of those already in use, as a comparison of original and synthetic data visualisations may reveal early diagnosis and personalised treatment insights for human-in-the-loop machine learning research. Prior to considering any data preparation or preprocessing, such as data cleaning, the concern is determining whether the current observations, features, and data size are appropriate. A customised version of the Generative Adversarial Network (GAN) [5,6] is used to construct synthetic data with the same statistical properties as the original data. GAN is designed to process data with continuous and binary variables. The generated and original data are classified by comparing multiple algorithms: J48, Zero Rule, Support Vector Machines (SVM), Multi Scheme, k-Nearest Neighbours, Artificial Neural Network, Naive Bayes, Random Forest [7], and Decision Trees [8,9]. Random Forest performed better, therefore, it is mainly used to classify the data. The most effective number of observations and features for a dataset are made based on how the Random Forest performed on both real and synthetic data.

This study combines supervised learning in the form of Random Forest with unsupervised learning through the use of GAN, as well as providing visualisation experiments for human experts. A new observation's fingerprints can be extracted more effectively by visualising features in detail, considering how they are distributed. In addition, a visual learning approach is introduced to evaluate the data before using time and resource-intensive techniques. To provide high-quality healthcare, it is necessary to accurately diagnose patients and identify personalised treatments [10]. Visualisation exercises are crucial [11] for selecting the best treatment method for early diagnosis. Consequently, visual learning may improve data-driven performance and should be used in conjunction with other learning approaches, including supervised learning.

Visualisation experiments combined with predictive models are crucial tools which aid healthcare professionals in identifying people who are at high risk of developing diseases or the precise health condition of the patient. However, data quality may impact the diagnosis outcome or the selection of the most suitable treatment. This results in earlier diagnosis and more effective management of the disease, potentially preventing or delaying complications by adjusting interventions and treatments according to specific risk factors. Early diagnosis and personalised treatment can lead to improved resource allocation, better health outcomes, and lower healthcare costs. Flexible visual learning diagnosis system is amendable both at the data and algorithm levels, and it can be tailored to data, aiming to obtain the optimal data quality to meet the requirements of a diagnostic health system. Personalised and early diagnosis are made possible by a combination of data visualisations, classification, and human judgement. Visual learning focuses on the idea that a classification approach should be based on treatment choices offered by actionable insights from visual experiments, to understand the data better before commencing data analysis.

The objectives of this study are to propose a novel machine learning approach in healthcare for personalised treatment and early diagnosis, to increase collaborative complementarity between data-driven perspective and human-in-the-loop. This is done in order to address the underlying causes of data challenges and find a balance between suitable data quality and optimal diagnosis performance. Hybrid visual learning approach is introduced, to combine supervised and unsupervised approaches with statistical machine learning visualisation experiments, and human-in-the-loop. Key strategies and experiments are conducted considering data challenges and data modelling simultaneously.

First, the unsupervised algorithm Generative Adversarial Network is implemented to transform the original medical data into synthetic data, to construct a learning approach based on comparing visualisations of original and synthetic data.

Second, visualisation experiments are conducted to compare the original and synthetic data, assessing which classification models might be more appropriate. A more flexible system is proposed by adjusting the data, to validate and evaluate the results meanwhile keeping the statistical properties of the data. To determine the general applicability of the visual learning approach in various medical domains, visualisations such as histograms, standard deviations (SDs) and means for numerous medical datasets of heart [12], cancer [13,14], immunotherapy [15], cryotherapy [16], exasens data [17] and diabetes [18] are compared.

Third, based on the results of visualisation experiments, several suitable classification models are put into practice such as the supervised algorithm Random Forest. Algorithms are adapted to update the data, bridging the data challenges such as small size and imbalanced classes. Fourth, to validate the performance of models on original and synthetic data, the classification results are compared with published studies. The innovations and contributions of this study are described below:

1. The introduction of a hybrid approach to visual learning that combines supervised and unsupervised methods with experiments in statistical machine learning visualisation enables human-in-the-loop with informed decisions making choices about early diagnosis and personalised treatments.
2. An adaptable and personalised system is provided, facilitating crucial data amendments and visualisations to obtain the optimal early diagnosis.
3. A distinctive system is presented, by first visualising the original and synthetic data through experiments and then accordingly applying the personalised appropriate classification solution. For application and generalisation, the system is used in a variety of health domains.
4. Instead of treating the symptoms of data challenges, the underlying causes of suboptimal classification are addressed.
5. Models are integrated to automate the evaluation of the visual learning approach. Generative Adversarial Network (GAN) is developed to create and manipulate synthetic data and Random Forest is used to classify the data. GAN+RF combination addresses data challenges optimally.

The remainder of this paper is organised as follows: Section 2 explores the related published literature. Section 3 presents the methodology. The visual learning approach is proposed, constructing a system model, flowchart and working steps. The unsupervised algorithm Generative Adversarial Network (GAN) and supervised algorithm Random Forest (RF) are described in detail. The medical datasets used as case studies are indicated. The comparison of developed machine learning algorithms is performed. The implementation of GAN and RF is presented. Machine learning and statistical experiments are provided. Sections 4 and 5 summarise and discuss the implications of the findings with recommendations for further research.

2. Literature review

Several researchers have developed various machine-learning-based approaches [19–22] to predict the response of patients to immunotherapy and cryotherapy treatment options using immunotherapy [15] and cryotherapy [16] datasets. Two Adaptive Neuro-fuzzy Inference Systems (ANFIS) were proposed [23], to assist in deciding between cryotherapy and immunotherapy for the treatment of warts, obtaining accuracies of 83.3% and 80.7%, respectively. To forecast the effectiveness of wart treatment methods, a Decision Tree (DT) based approach was proposed [24], which was transformed into fuzzily informative images for immunotherapy and cryotherapy datasets, achieving accuracies of 90% and 94.4%, respectively. In another study, a Classification and Regression Tree (CART) was deployed to create predictive models,

Table 1
Novel classification studies on immunotherapy data taking efficiency, effectiveness, and personalised machine learning experiments into consideration.

Immunotherapy study	Focus
Literature review	Application Domains, Datasets, Algorithms and Software Tools [28] Tools, Current Trends and Resources
Implementing Random Forest and Decision Trees	Efficiency of immunotherapy treatments [29]
Novel Algorithm: Pareto Principle	Multi-objective optimisation and ABC analysis [30]
Experiments:	
Personalised and adaptable machine experiments	Converting small data to big data [31]
Five novel machine learning classification experiments	To plan, conduct, and evaluate experiments, addressing the issues of imbalanced immunotherapy and medical [32] data

and the model outperformed other classifiers in terms of accuracy (100% for both immunotherapy and cryotherapy). In addition, a genetic programming-based Decision Tree was used to improve accuracy [25]. Fuzzy Rough Set (FRS), Classification and Regression Tree (CART), and Naive Bayes (NB) were combined for feature selection in another study [26] deploying algorithms such as Random Forest and Support Vector Machines to assess the effectiveness of both immunotherapy and cryotherapy treatments, resulting in an accuracy of 96%. The traditional pruning algorithm was replaced by Particle Swarm Optimisation (PSO) tuning the CART hyperparameters [27], obtaining an accuracy of 100%.

Because immunotherapy and health-related datasets are frequently imbalanced and small, machine learning algorithms typically perform suboptimally when classifying these datasets. To address the data challenges, numerous studies on classification modelling have been conducted, including literature reviews, efficiency analyses, and personalised machine learning experiments, as shown in Table 1.

An exploration of the published literature reveals that the main focus of data-driven studies is typically algorithm implementation, feature selection [33,34], and data preprocessing [35], which may involve small data modifications, to reveal useful insights. However, to address the fundamental impacts of data challenges, new approaches which consider both the data and algorithm levels are beneficial for discovering a balance between the desired diagnosis and the quality of the data while maintaining the statistical properties of the data.

After studying the published littérateur, the following observations are made: First, the human in the loop is missing, which is referred to as judgement by human medical experts. This gap is addressed by conducting visual experiments presenting a proper overview of the medical data at hand and the utilised algorithms assessing personalised and early diagnosis. Second, data challenges are not considered by assessing the data and algorithm levels together; for instance, in the case of immunotherapy datasets of warts. Third, no re-evaluation of the data was performed after conducting the machine learning tasks. The visual learning approach, that is proposed in this study, focuses on the re-evaluation of the diagnosis system, which is flexible at both data and algorithm levels. Fourth, the symptoms are treated and not the root impacts of data challenges. In addition, visual learning offers an understanding of the actual data condition or the level of data challenge to address and personalise the diagnosis accordingly. Fifth, a single learning approach, either supervised or unsupervised, is usually utilised to reveal diagnosis patterns and insights, while visual learning employs both supervised and unsupervised approaches or a combination of these to uncover the optimal early diagnosis in a given situation, and a new system is proposed to manage and handle the data better. By combining the approaches in this manner, the best can be taken from each, this can help health professionals and researchers improve the allocation of resources accordingly to adapt and individualise the diagnosis to the data at hand.

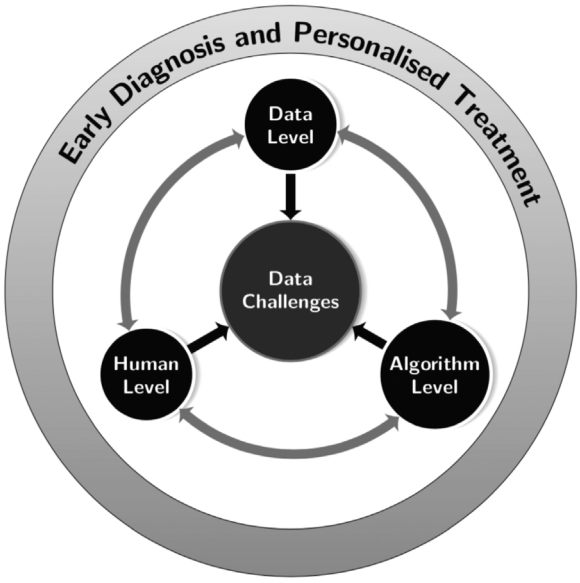

Fig. 1. Tackling data challenges for personalised and early diagnosis at the data, algorithm and human levels.

3. Methodology

Hybrid visual learning approach is based on decisions and considerations involved at data, algorithm and human levels as described in Fig. 1. Data level is the foundation of machine learning and the cause of suboptimal diagnosis performance. Medical data challenges are considered to address the causes and the symptoms, involving small sample sizes, imbalanced classes and irrelevant features. Decisions are based on multiple perspectives of data quality such as the number of features and statistical characteristics. At the algorithm level, synthetic data are generated, manipulated and classified using GAN and RF. After modelling the data the performance is considered whether it is optimal, otherwise, data adjustments are made to discover the personalised treatment and early diagnosis for the data and disease.

To increase collaborative complementarity between data-driven perspective and human-in-the-loop, a novel machine learning approach is presented with respect to healthcare for individualised treatment and early diagnosis. A hybrid visual learning approach is introduced, combining supervised and unsupervised approaches with statistical machine learning visualisation experiments for the judgement of human medical experts, which is the human level, to address the underlying causes of data challenges and find a balance between the amount of data required and the optimal classification for improved diagnosis performance. Important experiments and strategies are carried out

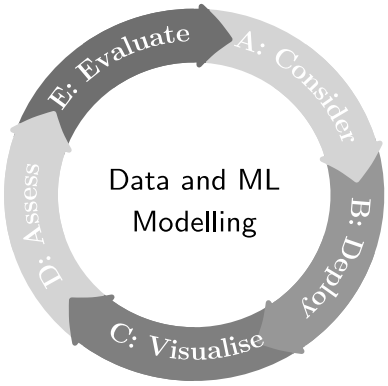

Fig. 2. The process of visual learning approach, a focus on data and machine learning modelling.

while taking data modelling and data challenge issues into consideration. First, the Generative Adversarial Network, an unsupervised algorithm, is used to create synthetic data from the original medical data. Second, based on the findings of the visualisation experiments, a number of suitable classification models are applied, such as the supervised algorithm Random Forest. In visual learning proposed work: system model, steps of working and flowchart are indicated.

3.1. Visual learning proposed work

Supervised and unsupervised approaches [2] typically use fixed data, or only minor data changes are made, such as feature selection or sampling. By contrast, the data in the case of visual learning are flexible in terms of quality and quantity, because the number of observations, features, and other parameters can be changed in accordance with an analysis of data.

3.1.1. System model and flowchart

Visual learning is implemented in two steps. The first step is illustrated in Fig. 2 [36]. The process's default order of action sequence is $A \longrightarrow B \longrightarrow C \longrightarrow D \longrightarrow E$. However, the process parameters can be changed according to a dataset to obtain personalised visual learning. The basis of the entire perspective is data to allow effective addressing of the primary causes of data challenges. The second step demonstrates the planning and execution of visual learning, as shown in the flowchart in Fig. 3.

3.1.2. Working steps

The implementation steps of the adoptable visual learning approach are described below and illustrated in Figs. 1 and 2.

(1) The pre-algorithm stage, or data level, consists of step A. (2) The algorithmic level, Generative Adversarial Network (GAN) and Random Forest (RF) are combined, integrated, and automated. This is done to evaluate the outcomes of the visual learning performance in step B. (3) Human judgement level, involving visualisation experiments of original and synthetic data in steps C, D and E.

1. Step A considering: calculate the percentage of the majority and minority classes of data. Consider whether the number of observations and features of data are suitable, less or more than necessary to find the point of balance and to discover the maximum performance potential.
2. Step B deploying: convert original data to synthetic data applying Generative Adversarial Network (GAN) according to needs, availability or actual circumstances. Deploy Random Forest for classification, meanwhile, manipulating and simulating the data simultaneously according to the required personalised treatment and early diagnosis.

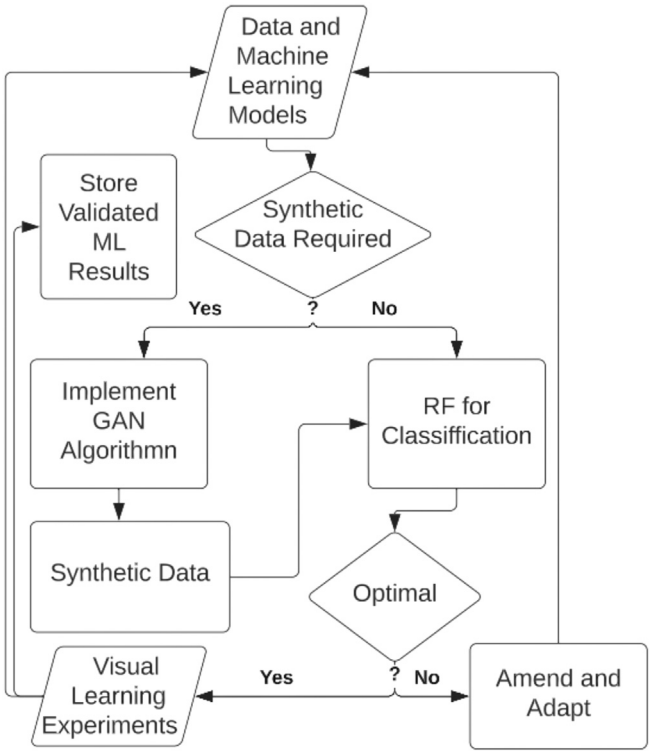

Fig. 3. The flowchart of the proposed visual learning approach.

3. Step C visualising and comparing: the original and synthetic data to consider for example if more data are beneficial.
4. Step D assessing: statistical characteristics such as means and standard deviations, cumulative sums, histograms, principal component analysis (PCA), correlation matrix, and root mean square error (RMSE).
5. Step E evaluating: Repeat and modify the steps described above of the visual learning approach as necessary for further improvement of the outcomes. To get a personalised learning experience for the available data, the steps are scalable, adaptable and amendable independently.

3.1.3. Overview

The ability to create simulations is facilitated by the fact that the data are no longer fixed; instead the data can be modified. Thus, a dynamic system is created which has the potential to be changed and improved in many ways, making it competitive with other learning approaches. The advantage of the visual learning approach is that it is adaptable to challenging data as in Fig. 4.

3.2. Generative Adversarial Network (GAN)

The minimax two-player game is the source of the fundamental concept behind Generative Adversarial Network (GAN) [5]. A basic GAN consists of a generator G that replicates the distribution of real data and a discriminator D which attempts to separate the real data from the data produced by G. In a GAN, the two models are trained with the goal of minimising the difference between synthetic and real samples and maximising the confidence in differentiating between synthetic and real samples. During training, the two models simultaneously compete with one another to improve their ability to generate and discriminate data and find a Nash equilibrium. Therefore, the minimax two-player game, dependent on G and D, is assessed using the cost function V (G, D) as in Eqs. (1) and (2).

$$min_G max_D V(G, D) = \mathbb{E}_{X \sim P_{data}}[log D(X)]$$

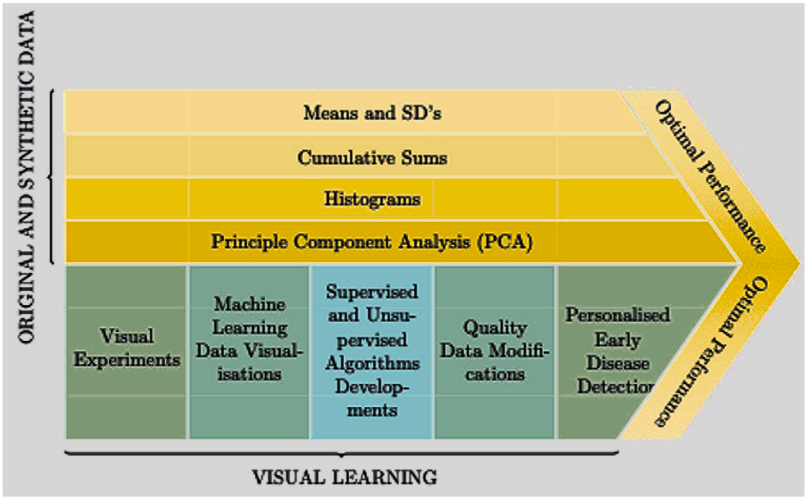

Fig. 4. Overview of the visual learning approach.

$$+\mathbb{E}_{Z\sim P_Z}[log(1-D(G(Z)))] \quad (1)$$

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z))] \quad (2)$$

Where:

x = a sample from the real data distribution p_{data}(x)
E(.) = the expectation
z = derived from a priori input noise variables $p_z(z)$
$p_g(x)$ = the generated data distribution
$G(z)$ = the data generated by G and subject to distribution p_{data}
$G(x)$ = data generated by G
$D(x)$ = the likelihood that x is sampled from p_{data}

One model is fixed during the GAN training process, whereas the other is optimised [37]. To maximise the discrimination accuracy, the generator is first fixed, and the discriminator divides the real samples into positive and generated samples as negatively as possible. Consequently, the ideal solution for the discriminator is obtained as in Eq. (3).

$$D_x^* = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)} \quad (3)$$

The generator is then trained by minimising log(1 - D (G(z))) for a fixed D. G is trained by maximising an alternative log D(G(z)) to account for the saturation situation early in learning. Ideally, $p_g = p_{data}$, which is equivalent to D(x), is attained to achieve a global optimum. For finite datasets, the discriminator may not achieve ideal optimisation. Instead, several iterations of training D and one iteration of training G are alternated during the optimisation process. Generative models have expanded significantly and have been used successfully in a wide range of practical applications [38]. When computing the density estimation, the generative models use the model distribution p_{model} to approximate the learned data distribution p_{data}. The selection of an appropriate objective (loss) function and appropriate formulation for the density function of the p_{model} are the two main issues in density estimation methods. The maximum likelihood estimation theory, in which the model parameters maximise the likelihood of the training data, is the de facto standard for the most commonly used objective.

One perspective for dealing with the marginal likelihood intractability problem is to forego computing it ever and instead learn model parameters through an indirect method. GAN accomplish this by having a strong D, which can differentiate samples from p_{data} and p_{model}. If D is unable to do so, the model learns to produce samples which are similar to the samples from the real data. The use of an explicit density function in which the maximum likelihood framework is used to estimate the parameters is a potential method for formulating the density function of p_{model}. Another option is to estimate the data distribution while excluding analytical forms of the p_{model} using an implicit density function, that is, training a G, where real and generated data are contained within the same sphere [39,40] if they are mapped to the feature space. However, the most notable class of potential solutions is GAN.

The ability to support both exact sampling and approximate estimation renders GAN an expressive class of generative models. GAN automatically picks up high-dimensional distributions over images, audio, and data, which are difficult to explicitly model. Basic GAN is an algorithmic structure which pits two neural networks against one another to capture the true distribution of data. To determine (globally) the Nash equilibrium in a zero-sum game, both neural nets attempt to optimise various opposing objective (loss) functions. The network architecture, objective (loss) function, and optimisation algorithm comprise the main building blocks for the design and optimisation of GAN. Numerous efforts have been made to enhance GAN through re-engineering the architecture [41] better objective functions [42], and different optimisation algorithms.

3.3. Random Forest

Random Forest (RF) is a classification and regression technique built on the compilation of numerous decision trees [43]. It is an ensemble of trees built from a training set and internally validated to predict the response given the predictors for upcoming observations. The construction of each tree, the method used to create the modified datasets on which each tree is based, and the method used to combine the predictions of each tree to produce a singular consensus prediction. The RF method uses the so-called Decrease of Gini Impurity (DGI) as a splitting criterion, where each tree is a standard Classification or Regression Tree (CART), which chooses the splitting predictor from a randomly chosen subset of predictors (the subset is different at each split). Every tree is built using a bootstrap sample taken with replacement from the original dataset, and the predictions from every tree are then combined through majority voting. Most software currently in the market, as described below, uses this version of the RF.

3.3.1. Classification evaluation

The following are definitions for precision, recall, f-measure, accuracy, sensitivity and specificity as in Eqs. (4), (5), (6), (7), (8), and (9), respectively.

TP = True Positive, TN = True Negative
FP = False Positive, FN = False Negative

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Table 2
Information on the immunotherapy dataset's statistics. * Time before start of treatment. **The surface area of the largest wart. SD (standard deviation), mm (millimetre), Ca. (categorical), Nu. (numerical), Pl. (plantar), Co. (Common)

Number	Attributes	Kind	Immunotherapy results	
			Quantity	**Mean /SD**
1	Sex	Ca.	Male (41) Female (49)	
2	Age (years)	Nu.	15–56	31.04/12.23
3	*Time	Nu.	0–12	7.23/3.10
4	Number of warts	Nu.	1–19	6.16/4.2
5	Type of warts	Ca.	Pl. (22) Co. (47) Both (21	
6	** Area (mm^2)	Nu.	6–900	95.7/136.61
7	Induration (mm)	Nu.	2–70	14.33
8	Success of treatment	Ca.	Yes (71) No (19)	

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - measure = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6}$$

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

3.4. Medical datasets

Multiple medical and health-related datasets from UCI machine learning repository and Kaggle are used as case studies: heart [12], cancer [13,14], immunotherapy [15], cryotherapy [16], exasens data [17] and diabetes [18]. To represent potential data challenges the datasets are chosen from various health domains to consider the general applicability of the visual learning approach. For example, the immunotherapy and cryotherapy datasets are comprehensively explained and used. The immunotherapy dataset sample consists of 90 patients older than 15 years old who were receiving immunotherapy treatment and were part of a two-year dataset collection in a hospital clinic. Immunotherapeutic methods work by boosting the host cell-mediated immunity to neutralise the virus [44]. The patients received immunotherapy, with a maximum of three concurrent sessions, and a break in between. Table 2 provides information on the results, clinical characteristics, and demographics of the data.

3.4.1. Challenges

Real medical data are limited, difficult to access, imbalanced and contain irrelevant features. The complexities of medical data make it challenging to predict outcomes [45] because personalised and early disease detection are essential for treating patients, necessitating the creation of effective and high-performance algorithms, techniques, and tools for the analysis of big data in medicine. For tasks involving health research or clinical applications, considerable time and expertise are required to use and interpret a variety of high-dimensional data types [46]. Additionally, the interpretation of multiple data types requires more computing power than the interpretation of each data type, and modelling algorithms which can learn from a staggering number of intricate features are required. The use of machine learning algorithms to automate these processes and support disease detection and diagnosis has grown in popularity [34,47]. Interestingly, deep learning models may be able to take advantage of this complexity by revealing insightful information and locating pertinent features from various data types [48,49]. A branch of artificial intelligence (AI) known as machine learning (ML) focuses on developing predictions by spotting patterns in data.

Data synthesis is a statistical disclosure limitation (SDL) technique in which the true values of sensitive variables are replaced with randomly generated values [50]. To create partially synthetic records, the conditional distribution of the sensitive variables given the non-sensitive variables is modelled and then sampled. These include imputed values for the sensitive variables as well as the underlying values for the non-sensitive variables from the dataset.

3.5. Algorithms comparison

To compare the classification performance of Random Forest to other machine learning algorithms, baseline models and other models are used as a benchmark and point of comparison. This shows how much classification is improved in comparison to the baseline models and enables the assessment of absolute performance improvements over the baseline models. The training and testing ratio for the utilised models is 70% and 30%, respectively. An experiment is carried out on the immunotherapy dataset [3] using the following supervised classification models: J48, Zero Rule, Support Vector Machines (SVM), Multi Scheme, k-Nearest Neighbours, Artificial Neural Network, Naive Bayes, Random Forest, and Decision Trees [8,9]. The experiment's objective is to apply and contrast several algorithms to ascertain which algorithms are most effective at identifying successful and ineffective therapies in the immunotherapy dataset. Table 3 displays the results of the machine learning algorithms. The Random Forest (RF) algorithm outperformed other algorithms, achieving classification accuracy, sensitivity, and specificity values of 88.88%, 95.45%, and 60.0%, respectively.

3.6. GAN and RF implementation

Generative Adversarial Network (GAN) can be applied to data from various diseases, depending on the criteria chosen for implementation, for instance, the number of observations and features, the degree of imbalance, and the sample size. In this study, small and imbalanced medical data with irrelevant features served as the data selection criteria for implementing the GAN. To consider the general applicability and generalisation of the visual learning approach, various real medical datasets from medical fields are chosen. In a real-world scenario, the implementation criteria may vary and should be according to the actual degree of complexity of the data. The human medical expert judgement based on visualisation experiments, such as the test of statistical characteristics as in experiment one Section 3.7.1, is a crucial component of using GAN on the datasets.

First, an unsupervised algorithm called Generative Adversarial Network (GAN) is created to produce synthetic data from the original data. Then, Random Forest (RF) [51,52], a supervised classification algorithm, is utilised to assess how well it performs on both real and synthetic data. To determine the best visual learning strategy, the supervised and unsupervised methods are combined. The desired quality of synthetic data can be altered by adjusting the parameters of the Generative Adversarial Network (GAN), which makes the system modifiable, while, maintaining the statistical characteristics of the data. Thus, after considering a data challenge, GAN is applied appropriately. For example, if the data are small, imbalanced, and contain irrelevant features, the best performance can be obtained by adjusting the data quality according to the challenge at hand. By developing Random Forest the average precision, recall, and f-measure for the immunotherapy data are 91%, 96%, and 93% for real data, compared to 81%, 82%, and 81% for synthetic data, respectively, as in Table 6.

A hyperparameter tuning procedure is used to systematically alter the following parameters in order to get the optimal performance out of the defined Generative Adversarial Network (GAN) in Table 4 and Random Forest in Table 5.

The Generator and Discriminator's individual losses as well as the combined loss have been selected as metrics for a tactical assessment of the defined hyperparameters. The average and maximum correlation

Table 3
The accuracy, sensitivity, and specificity of various algorithms implementation on imbalanced immunotherapy data.

Algorithm	Classification performance			Specificity ranking	Sensitivity ranking	Accuracy ranking
	Accuracy	Sensitivity	Specificity			
J48	85.18%	85.18%	0%	5	4	2
Zero Rule	85.18%	85.18%	0%	5	4	2
Support Vector Machines (SVM)	85.18%	91.30%	50%	2	2	2
Multi Scheme	85.18%	85.18%	0%	5	4	2
k-Nearest Neighbours	70.37%	95.45%	25%	4	1	4
Artificial Neural Network	77.77%	90.47%	33.33%	3	3	3
Bayes Network	77.77%	84%	0%	5	5	3
Random Forest	88.88%	95.45%	60%	1	1	1
Decision Trees	85.18%	85.18%	0%	5	4	2

Table 4
Hyperparameter tuning of Generative Adversarial Network (GAN).

Parameter	Value
The discriminator's and the generator's rate of learning	lr_d = 0.0005 and lr_g = 0.0005
The size of the hidden feature space	hidden_feature_space = 200
Size of input noise	binary_noise = 0.2
A single bag of rows' total number of rows	nr_of_rows = 25
Batch size	batch_size = 100

Table 5
Hyperparameters of Random Forest.

Parameter	Value
The random number of seed	1
The preferred number of instances	100
The number of features	int(log_2(predictors) + 1)
The number of trees	100
The number of execution slots	1
The maximum depth of the tree	unlimited

Table 6
Classification of immunotherapy data using Random Forest on original and generated data.

Random Forest	Precision	Recall	F-measure
Classification of original data			
Successful Treatment	83%	94%	88%
Average	81%	82%	81%
Classification of generated data			
Successful Treatment	96%	100%	98%
Average	91%	96%	93%

errors as well as the distribution's average error have also been considered. While Pearson correlation coefficients were used to calculate the error of the correlations, the error of the distributions was calculated by comparing means and standard deviations.

In order for a GAN's discriminator to properly learn to complete its task, it requires original input data. The statistical characteristics of a real clinical trial with 90 patients are used to simulate immunotherapy data. For the model's discriminator, 90 observations are used as input. Table 6 shows eight binary and continuous features, which are simulated. Random noise is used as the input for the generator network, which is then transformed based on the feedback from the discriminator to take into account the statistical characteristics of the original data.

Generative Adversarial Networks for data generation are incredibly efficient at producing unstructured data objects like images, however, GAN can also be used to produce structured, tabular data, which is frequently found in clinical trials. GAN is generally appropriate for the reliable generation of synthetic but realistic clinical trial data, producing more satisfying synthetic patients which closely resemble the original data, whereas some other networks exhibit lengthy training times. The customised GAN version is built on long short-term memory (LSTM) layers and is thus able to maintain underlying data properties, such as correlations and variable distributions, producing more satisfying results, even in small-sized samples, with a sufficient training speed.

3.7. Visual experiments

The visualisation experiments are conducted to allow human medical experts to select diagnosis and treatment decisions on a well-informed basis, this is referred to the human level in visual learning approach in Fig. 1. Additionally, the diagnosis system can be evaluated by human experts who may decide to make critical data or algorithmic changes. In this way, individualised treatments are based on appropriate and flexible tactics. The visualisation experiments act as the fingerprints of the entire visual learning approach, illustrating the actual screenshot of the diagnosis to obtain a real overview of data to achieve optimal, more precise and effective performance.

To better understand the data and identify the challenges at the data level, experiments are conducted using various scenarios of visualisations to compare and contrast the real data with the synthetic data. The data are also plotted side by side for improved treatment decision support. Before implementing an algorithm, comprehensive data visualisations help to better understand the data challenges and to design and implement personalised solutions for early diagnosis.

The phrase "human-in-the-loop machine learning" (HILML) refers to the interaction of human and machine learning (ML) processes to address one or more of the following issues: enhancing ML accuracy, quickening ML's ascent to the desired accuracy, improving accuracy and efficiency in humans [53,54]. HILML is used in this study to enhance data quality, which allows the applied algorithms to operate with greater accuracy and efficiency, which has a positive domino effect on obtaining a better health system. The original data are converted into synthetic data using the Generative Adversarial Network (GAN) algorithm [55] because health data are small, imbalanced, and contain irrelevant features. The data are scaled and normalised to have a range from −1 to 1. The following criteria are chosen so that some similarities and differences could be identified: to gain insights and determine whether the data are appropriate for efficient diagnosis performance or more data are needed. A variety of visualisation experiments have been developed: statistical characteristics, cumulative sums, histograms correlation matrix and RMSE and principal component analysis.

Before beginning the algorithm implementation, taking exploratory data analysis as a starting point could be useful for spotting patterns in the data and determining how the features relate to one another for

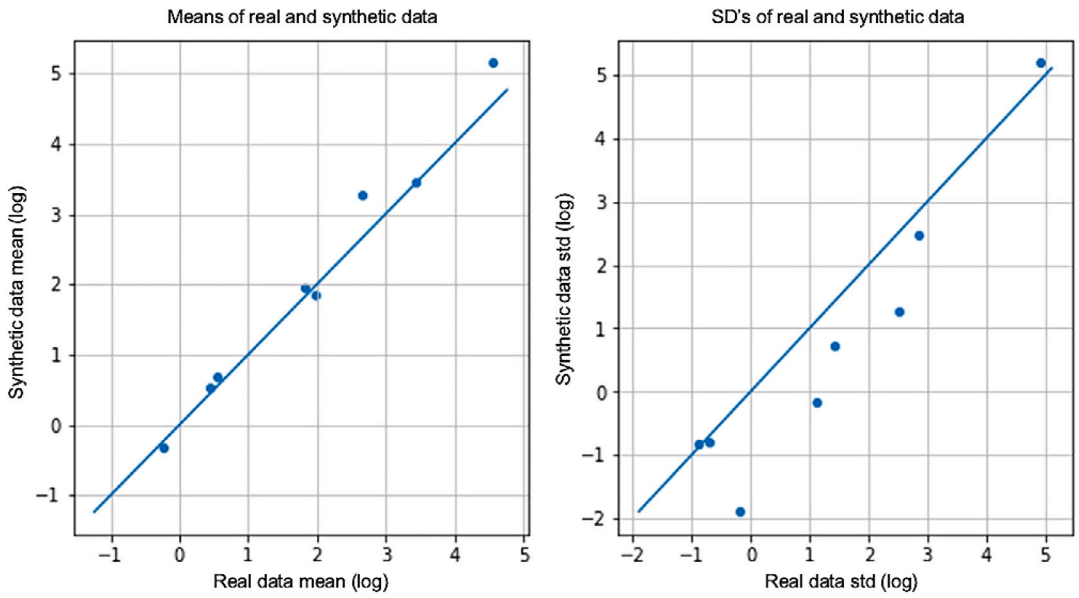

Fig. 5. Means and SDs of features of immunotherapy data.

data elimination [56]. In a variety of clinical and experimental diseases, histogram variables can be useful indicators of treatment response [57]. The features and characteristics of a dataset can be described by using descriptive statistics [58]. This includes measures of central tendency, such as the mean and standard deviation. Real-world medical data are frequently challenging in terms of implementing machine-learning algorithms. Applying analytical machine learning algorithms to raw data directly may result in less optimal performance [59]. Excluding patients with incomplete data, replacing the missing value with the mean or the most frequent value of the corresponding predictor, and imputation based on correlation models are methods for data pre-processing. Another option is visual learning, which offers systematic additional options such as cumulative sums and histograms.

3.7.1. Experiment 1: Statistical characteristics

The mean and standard deviation (SD) of the features of the immunotherapy data are shown in Fig. 5. To compare the statistical characteristics of the real and synthetic data, visualisations are constructed such that the real data are on the x-axis and the synthetic data on the y-axis are used to assess the means and SDs. In Fig. 5 the line shows the relationship between the means of real and synthetic data, indicating that both datasets have approximately the same statistical characteristics in terms of means. All points of the data are situated on a straight line, demonstrating a significant correlation among the features. SDs, on the other hand, deviate from a straight line, suggesting that the data points in synthetic data are more dispersed than the real data. The standard deviation is the average distance between each value in the dataset and the mean. When the standard deviation of a dataset is high, the values are typically spread from the mean, whereas when it is low, the values are typically grouped close to the mean. Eq. (10) demonstrates that the mean of x is equal to the summation of all x values divided by the number of values N. The general equation for determining the mean of a set of numbers:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{10}$$

Where:

$\bar{x}$= mean value of the observations
N = number of observations (measurements)
x_i = observed values of a sample item
$\sum$ = *summationnotation*

The formula for the sample standard deviation is given by Eq. (11).

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{11}$$

Where:

s = sample standard deviation
$\bar{x}$= mean value of the observations
N = number of observations (measurements)
x_i = observed values of a sample item $\sum$= *summationnotation*

The experiment shows that, after a test, the statistical characteristics of the original immunotherapy data are still present in the synthetic data, however, the data points are more scattered in the synthetic data. The two datasets have approximately equivalent statistical properties when means are compared, however, dissimilar SD's. This test is an important component of visual learning because, to support clinical diagnosis and treatment decisions, the statistical synthetic data are used as a mirror and must be identical to the original data.

3.7.2. Experiment 2: Cumulative sums

To show the total sum of data as it increases over time or in any parameter cumulative sums, also known as running totals, are used. This enables the plotting of a given measure's overall total contribution to a feature. The immunotherapy dataset contains eight features: time, induration diameter, age, result of treatment, sex, type, area, and number of warts [23]. The commutative sums for the features of the data are provided in Figs. 6 and 7 for both the original and synthetic data. Assume a vector in an n-dimensional space in Eq. (12):

$$\mathbf{v} := \langle v_1, v_2, \ldots, v_n \rangle \tag{12}$$

Where the definition of the projections is defined in Eq. (13).

$$\pi_k(\mathbf{v}) := v_k \tag{13}$$

Then, the cumulative sum vector w is determined using Eq. (14):

$$\mathbf{w} := \left\langle \pi_1(\mathbf{v}), \pi_1(\mathbf{v}) + \pi_2(\mathbf{v}), \ldots, \sum_{k=1}^{n} \pi_k(\mathbf{v}) \right\rangle \tag{14}$$

The results of the experiment reveal how a feature's structure affects the cumulative sums of the feature. For instance, because the cumulative sums of the area, time, and age features are nearly identical, these features may be suitable for classifying the immunotherapy data even before any data modelling is performed.

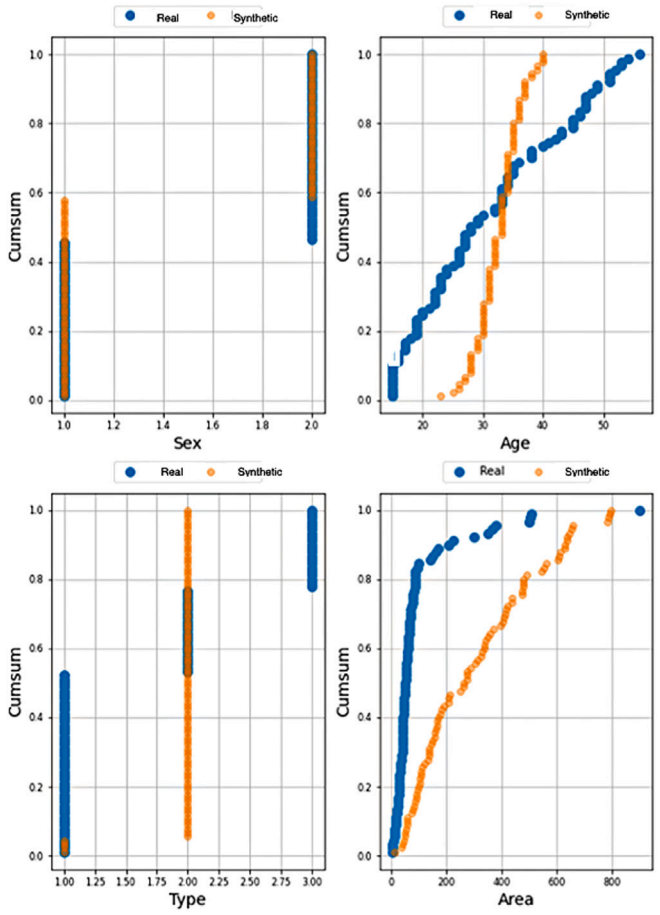

Fig. 6. Cumulative sums for sex, age, type and area features of immunotherapy data.

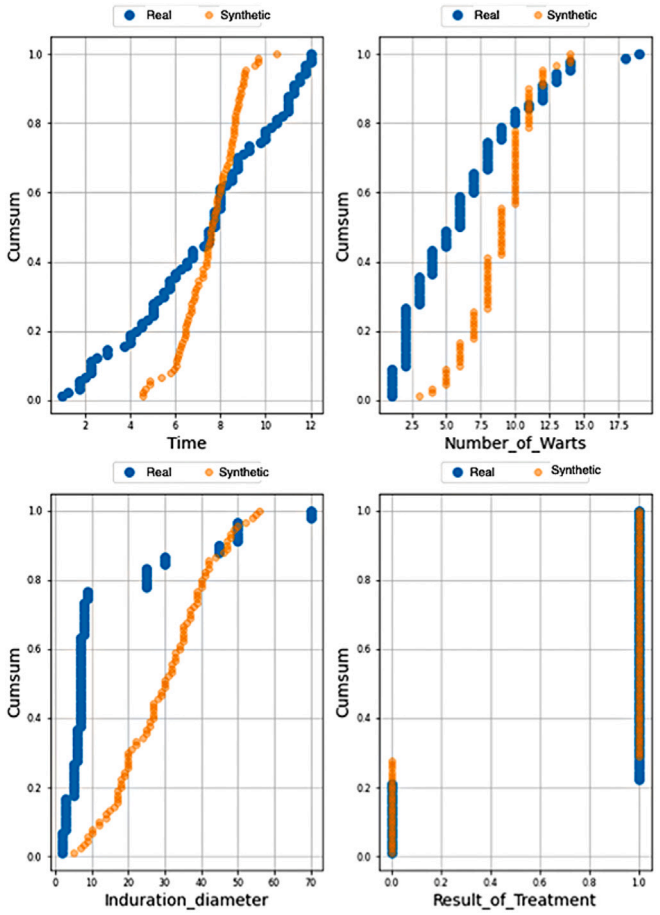

Fig. 7. Cumulative sums for number of warts, time, induration diameter and result of treatment features of immunotherapy data.

3.7.3. Experiment 3: Histograms

A histogram is a vertical bar graph in which the heights of the values represent the frequencies, which are also known as counts. The frequency distribution of the selected class range determines the bar width. Even before applying any machine learning model, histograms can be used to understand features and datasets to identify potentially useful features for classification, for instance, time and area features in the case of immunotherapy data. The histograms for the features of the immunotherapy data, for which a frequency distribution is created, would resemble Figs. 8 and 9. Histograms are constructed, showing that the distribution is altered for synthetic data, and the distributions are more normally distributed compared to the original data, as shown in Figs. 8 and 9.

3.7.4. Experiment 4: PCA

Principal component analysis (PCA) is used to create predictive models and conduct exploratory data analysis. PCA creates new variables called principal components from linear combinations of the original variables for dimensionality reduction. Dimensionality reduction involves projecting each data point into only the first few principal components to obtain lower-dimensional data while retaining as much variation in the data as possible. Data exploration and feature selection are essential tools for selecting the most relevant features. The dataset size needed for actual modelling can be minimised by retaining only the essential data. Fig. 10 shows the PCA of the immunotherapy data characteristics for both original and synthetic data.

The experiment demonstrates how correlation-based data analysis is discovered among the features of the dataset by plotting the PCA of both the original and synthetic data. First, the PCA of original data reveals that the dataset's observations are not highly correlated with one another, contrary to what the PCA of synthetic data shows, as the variance among features in the PCA of the synthetic data, is high. Second, significant variance exists among the observations in the case of real data, which might be due to an insignificant correlation between the features of the datasets. The results of the experiment show that it is possible to identify correlations between the features of a dataset by examining the variance among the observations by comparing the PCA's visualisation of the original and synthetic datasets. In addition, after this experiment is conducted and PCA is identified, correlation-based algorithms may be avoided when using algorithms to save computational resources.

3.7.5. Experiment 5: Correlation matrix

A statistical measure known as correlation expresses how closely two features are linearly related. The scatter matrix in Fig. 11 is plotted as a heat diagram to better visualise the correlation among the data features. The scatter matrix uses colours to illustrate the correlations among the features of the dataset. Navy blue and yellow indicate high and low correlations, respectively, and the darker the colour, the greater the correlation among the attributes. Therefore, the synthetic data reveal more correlations. The statistical measure of how well the changes in the value of one variable predict changes in the value of another is called the correlation coefficient. When the fluctuation of one feature accurately predicts a similar tendency in another feature, it is said that a change in one variable is the result of a change in another. The correlation coefficient between two random features X and Y is

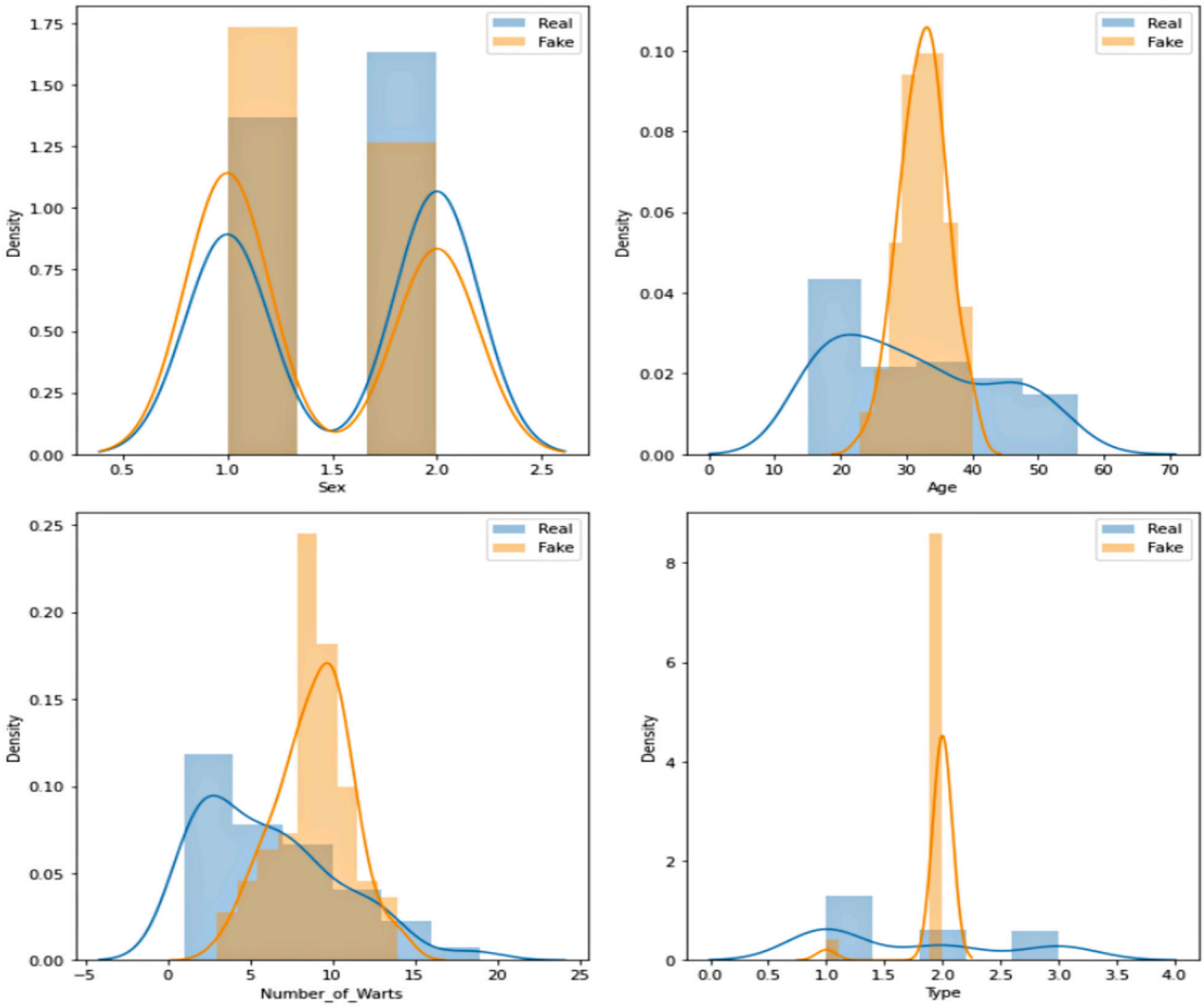

Fig. 8. Histograms for age, sex, number of warts and type features of immunotherapy data.

described by Eq. (15):

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}} \tag{15}$$

Where:

$\rho(X,Y)$ = correlation between the variables Y and X
Cov = covariance
$Cov(X,Y)$ = covariance between Y and X
$Var(X)$ = variance of X
$Var(Y)$ = variance of Y

The sample correlation coefficient r between two samples x and y is determined using Eq. (16):

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \tag{16}$$

Where:

S_x = sample standard deviation of variable x
S_y = sample standard deviation of variable y
S_{xy} = sample covariance

The intra-grid structure of the features is based on the correlation effect coefficient [60], which is a number between [−1,1]. The more the number goes toward 1, the greater the correlation among the features. The exact values of the correlation matrix illustrate the relationship between features. Table 7 lists the most correlated features in decreasing order. When the function is used, the age, type, number of warts, and area are found to be features which correlate most strongly with the result of treatment.

Table 7
Descending order correlation among features of the immunotherapy data and result of treatment.

Feature	Coefficient of Correlation
Age	0.188314
Type	0.083396
Number of Warts	0.047160
Area	0.043349

If two features move together in the same direction are said to have a positive correlation. When two different features change and move in opposite directions is called inverse correlation. The experiment indicates that a visual scatter matrix can assist researchers and health professionals in identifying useful patterns, such as correlations, in choosing the most appropriate and personalised treatment features.

3.7.6. Experiment 6: RMSE

The root mean square error (RMSE) method is based on the Euclidean distance to assess the accuracy of the predictions. The RMSE compares the quality between the predicted synthetic data values and the measured original data values, a metric for determining the similarity between two datasets. When evaluating a model's performance

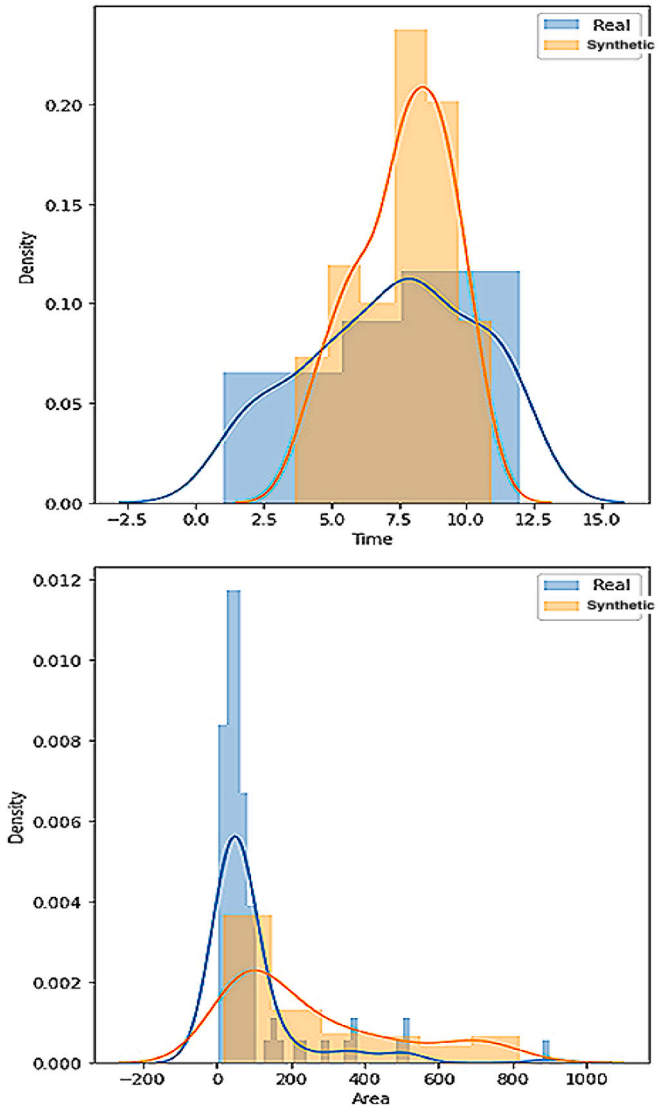

Fig. 9. Histograms for time and area features of immunotherapy data.

during training, cross-validation, or monitoring after deployment, it is useful to have a single number. The RMSE is an evaluation tool to aid comprehension and is consistent with some of the most widely used statistical assumptions. The RMSE is defined by Eq. (17) [61]:

$$RMSE = \sqrt{\left(\frac{1}{n_x n_y}\right) \sum_{i,j}^{n_x n_y} \left[\frac{r(i,j) - t(i,j)}{r(i,j)}\right]^2} \quad (17)$$

Where:

(i, j) = pixel

r(i, j) = the planning image's pixel (i, j) value in original data

t(i, j) = the value of pixel (i, j) in the synthetic data

$n_x n_y$ = total number of pixels

For numerical data, the RMSE is described by Eq. (18):

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - x_i)^2} \quad (18)$$

Where:

RMSE = root mean square error

i = variable i n = number of non-missing data points

x_i = actual observations time series

y_i = estimated time series

Table 8
RMSE for all features of the immunotherapy data.

Feature	Root mean square error (RMSE)
sex	0.699205898780101
age	12.433824833895642
Time	3.327561183544157
Number of Warts	4.389381125701739
Type	0.8563488385776752
Area	177.32857136463437
Induration diameter	20.407514955688914
Result of Treatment	0.5055250296034367

The scale of the data has an impact on how well models compare because RMSE is not scale-invariant; therefore, the use of RMSE on standardised data is optimal. The experiment shows the RMSE values for each feature of the data, enabling feature-level modifications for personalised healthcare and early disease detection. The RMSE value should be as low as possible. In the case of the immunotherapy data, sex and type are the features which obtained the lowest RMSE values of 0.69 and 0.85. The root mean square errors (RMSE) of the dataset for each feature are listed in Table 8.

4. Results and discussion

Normally, studies consider only a single approach to data analysis, for instance, supervised [62] or unsupervised. Supervised and unsupervised approaches are combined in visual learning. The Generative Adversarial Network (GAN), an unsupervised learning algorithm, is implemented to generate synthetic data, and Random Forest, a supervised learning algorithm, is applied to classify medical datasets. In this manner, many useful aspects can be supplemented to obtain the best of multiple learning approaches. Another advantage of the visual learning approach is swapping from one approach to another by moving through the different methods forward and back, manipulating and adapting the data, and finding the perfect personalised solution [63] required for the data at hand, considering the challenges which should be addressed, such as imbalanced small data.

The machine learning algorithm does not perform as well as it should due to data challenges like imbalanced, irrelevant features and small samples. By implementing necessary suitable changes to data quality, it is possible to identify the appropriate data for the current case. The performance of classification improves as data quality increases. The process of diagnosis or classification is made more robust, individualised, and effective by implementing the necessary changes at the appropriate time, which is the early stages of data analysis. When both original and synthetic data are available, the benefit is the option of selecting between imperfect original or improved synthetic data to produce better data quality for analysis. When considering disclosure risk, synthesis is typically safer. To avoid noisy data, visual learning adds more value to the data. More data-level initiatives improve the direction of data analysis and application, making diagnosis more resource-efficient and cost-effective to reduce the burden of diseases and associated costs. To conduct data analysis on an informed basis from the beginning to obtain an early diagnosis.

Using these scalable visual learning tools, patients can be divided into subgroups with greater accuracy, according to their estimated disease risk [64]. One use of these predictive analytics is the identification of a subgroup of patients who are at a higher risk of hospital admission and are therefore likely to be responsible for the majority of healthcare costs and the implementation of prompt preventive interventions based on such predictions. If successfully validated and used, these risk stratification tools could significantly lower the cost and morbidity related to avoidable readmission.

The outcomes of the visual approach are compared with those documented in previous studies. Experiments demonstrate that when visual learning is developed, performance can be enhanced, enabling

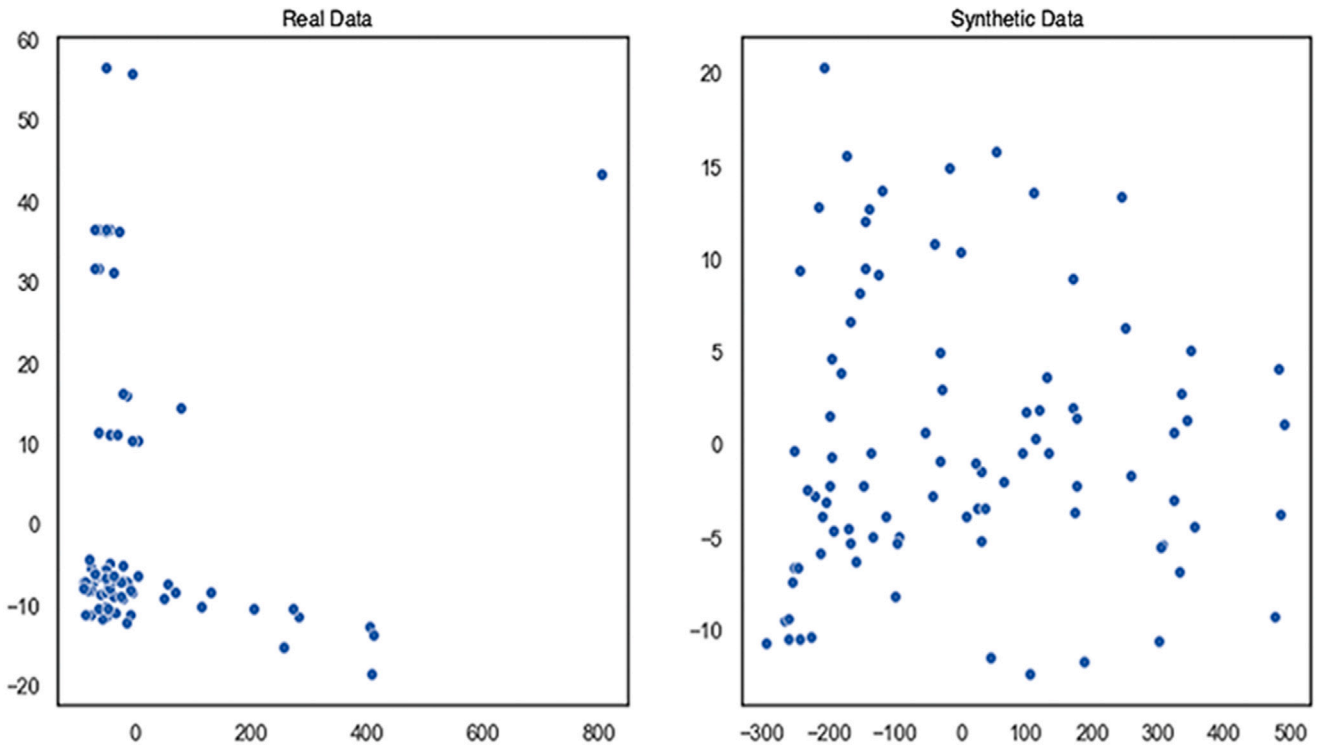

Fig. 10. PCA of the original and synthetic immunotherapy data.

prompt and accurate treatment because false positives and negatives are expensive to treat [65]. The first finding indicates that the development of additional methods is rarely necessary and the visual approach is usually sufficient. Visual learning methods are essential for the early detection of diseases and can be described as efficient and cost-effective for avoiding errors when analysing data as in Fig. 4.

To determine the applicability scope of the visual learning approach, it is generalised to several healthcare domains using different datasets from the UCI machine learning repository and Kaggle: heart [12], cancer [13,14], immunotherapy [15], cryotherapy [16], Exasens data [17] and Diabetes [18]. Both original and synthetic data are subjected to Random Forest development, and the classification results are compared with those from published studies as shown in Table 9. Experimental work and classification results show that the visual learning approach can be used in other medical fields.

The evaluation metrics used for classification are such as precision, recall, and f-measure. By using GAN to create the synthetic data, the performance is additionally enhanced. For instance, the obtained precision, recall and f-measure for the immunotherapy data are improved to 91%t, 96%, and 93%, respectively, from the original data's 81%, 82%, and 81%. However, each dataset has specific challenges, therefore the type of challenge should be first identified, and then visual learning applied for personalised diagnosis and treatment. Before applying visual learning, the degree and type of data challenges should be considered first, as in the case of synthetic cryotherapy data the performance decreased, as the cryotherapy data are balanced.

Data quality is usually improved through data preparation, also referred to as data preprocessing, in the initial stage of data analysis. This phase typically accounts for a significant workload [62,72]. Data cleaning, transformation, and reduction are the three most frequently used methods of data preparation. By implementing the visual learning approach, a systematic perspective on data pre-processing is introduced to enhance the quality of data, offering more convenient data visualisation for health professionals and researchers to improve the accuracy and efficiency of treatment decisions.

A machine learning algorithm must be properly engineered, similar to any tool, to be truly effective. For machine learning applications in the healthcare industry, clinical challenges must serve as both inspiration and benchmark. The ability of visual learning to assimilate and analyse huge, diverse datasets made up of various types of clinical data, with the clinical problem as the focal point, makes it an invaluable tool for clinicians to use when making decisions regarding the care of patients. Clinicians can consider more pieces of evidence with the help of this tool than otherwise process and analyse the data manually.

In addition, the challenging data issues are addressed at the data and algorithm levels in the visual learning approach, this has the advantage of not affecting the runtime of the classification model while improving its effectiveness and precision in handling difficult medical data. For example, focusing on the imbalanced learning issue at the data level, that is, using the same classification model in the follow-up analysis without considering the influence of the classifier and adopting a different strategy by using the RF and GAN algorithms to produce synthetic data and classify medical data to increase prediction accuracy.

5. Conclusion

The aim of this study was to consider data challenges, which are the gross root level in data analysis. Given that medical data are usually imbalanced and small, a novel visual learning approach is presented to determine the best solution at the fundamental level. To discover the improved balance between data quality and required performance, numerous methods, including synthetic data have been introduced, as visualising the data at hand results in better understanding and treatment decisions. The original health data are converted into synthetic data, which is the basic procedure when using the visual learning approach. Visual learning can be used as pre-supervised and unsupervised learning to provide an overview of the data, which is the most central element of machine learning.

Multiple visualisation methods are used to improve the quality of the treatment decision support for researchers and health professionals, offering judgement of the findings by medical experts while conserving time and other computational resources. Means and SDs, cumulative

Table 9
Comparing the implementation of Random Forest on original and synthetic data in various health domains with published literature.

Random Forest	Precision	Recall	F-measure	Accuracy	Sensitivity	Specificity
Classification of original data						
Immunotherapy:						
This study	81%	82%	81%	88.88%	95.45%	60%
Publications	77.08% [66]	73.26% [67]	79.12% [67]	100% [3]	100% [3]	100% [3]
Cancer:						
This study	82.88%	88.05%	88.07%	90%	94%	96%
Publications	95.65% [68]	98% [69]	97.77% [68]	97.14% [68]	97.19% [70]	99.71% [70]
Cryotherapy:						
This study	94%	95%	91%	97%	100%	98%
Publications	100% [71]	100% [71]	100% [71]	100% [32]	100% [32]	100% [32]
Classification of synthetic data						
Immunotherapy	91%	96%	93%	93%	96%	73%
Cancer	92.88%	94%	96%	98%	97%	94%
Cryotherapy	92.12%	94.22%	90.10%	95.34%	99.40%	88.09%

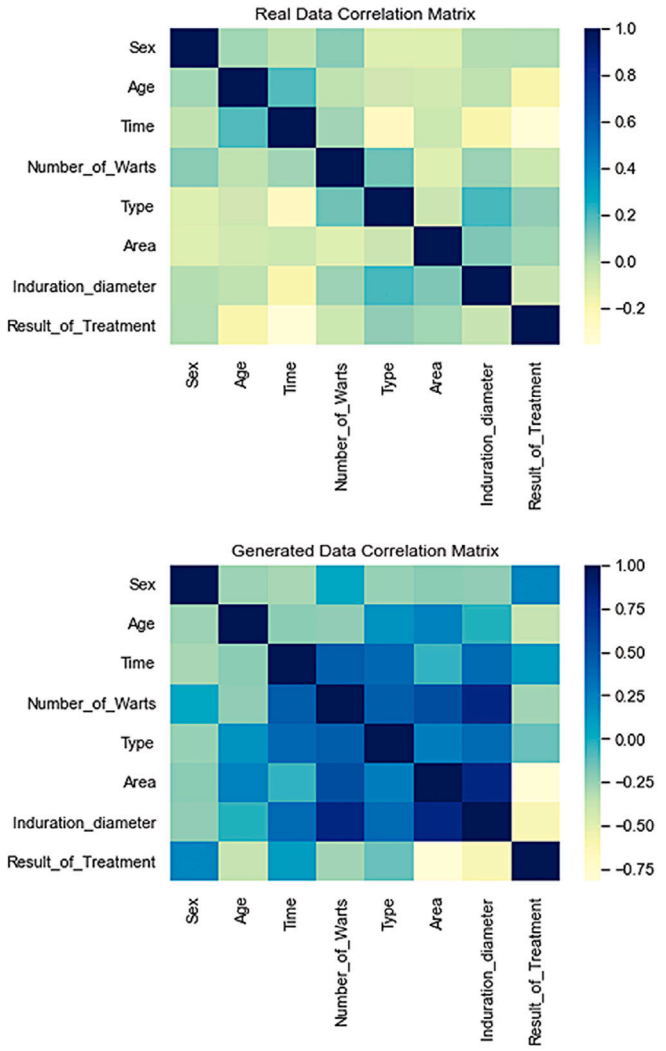

Fig. 11. Comparison of correlation matrices for immunotherapy treatments.

sums, histograms, principal component analysis (PCA), correlation matrix, and root mean square error (RMSE) of both the original and synthetic data are constructed.

The means and standard deviations (SDs) of both datasets are compared to determine whether the statistical properties of the original data are preserved in the synthetic data. The area, time, and age features are potentially reasonable for feature selection because their cumulative sums are almost identical. An unidentified observation can be identified by using a histogram as a sample for a feature. To determine whether features of a dataset are significantly correlated and to avoid using correlation-based algorithms to conserve computational resources, a scatter matrix is used to identify helpful patterns, such as correlations, in selecting the most suitable and personalised treatment features. A feature-level modification for individualised healthcare and early disease detection is made possible by the experiment's display of the RMSE values for each data feature. The features with the lowest RMSE values of 0.69 and 0.85 in the immunotherapy data were sex and type. These visualisation methods are useful for exploring data before implementing an algorithm. Instead of deploying complex methods for information extraction, focusing on foundations can facilitate early disease detection, enhance efficiency, and improve personalised treatment decisions.

Data as a central aspect of machine learning, is considered by applying Generative Adversarial Network (GAN) and Random Forest (RF) algorithms. The methods described in the visual learning approach can be considered as a pre-algorithm implementation process to provide an overview of the data and data analysis as a prototype. Visual learning offers an initial effective, cost-effective and efficient approach to address the data challenges before implementing other learning approaches, such as a supervised approach. The most relevant calculations for a better overview of robust, durable, effective treatments, early detection, and automated processes can be achieved using visual learning. The degree and nature of data challenges should be considered before applying visual learning.

An essential component of machine learning is pursued using Generative Adversarial Network (GAN) and Random Forest (RF) algorithms. With precision, recall, f-measure, accuracy, sensitivity and specificity for the immunotherapy data of 81%, 82%, 81%, 88%, 95%, and 60%, respectively, compared to 91%, 96%, 93%, 93%, 96%, and 73% for synthetic data, the Random Forest model performed best. Supervised and unsupervised techniques are combined to determine the most effective visual learning approach. The Generative Adversarial Network (GAN) parameters can be changed to obtain the desired quality of synthetic data, making this system modifiable, whilst maintaining the statistical properties, for instance, variable distributions and correlations of the data. Thus, GAN is used in the right way after considering data challenges. If the data are small, imbalanced, and contain irrelevant features, optimal performance can be obtained by adjusting the data quality in accordance with the task at hand.

Before moving on to other stages of data analysis, the methods described in the visual learning approach can be considered a pre-warming procedure. Before beginning the implementation of an algorithm, visual learning can be implemented as a starting point to find

Table 10
Suggestions for further research.

Number	Considering
1	Efficient solutions for missing values
2	Communication of understandable and multi-disciplinary data-driven results to non-technical stakeholders, and health professionals
3	Re-organise the gathered data from various formats
4	Improve the structure, mapping and formats of data
5	Coordination of data generation from multiple sources
6	Interpretability, saving cost, management and computational resources

interesting patterns in the data. Visual learning offers cost-effective and efficient methods for addressing data challenges before implementing other learning approaches. The visual learning approach provides the optimal combination of the data, algorithm and human levels to achieve the most important medical calculations and treatments: a better overview of personalised treatments, early disease detection, and automated procedures.

By enabling learning through visualisations for improved interactions of human experts with machines, the so-called human-in-the-loop and machine-in-the-loop machine learning [73,74], visual learning offers a systematic, personalised approach that enables medical experts and clinicians to collaborate with machines more effectively. This offers learning to both experts and machines to make informed decisions when human and machine abilities are limited, thereby resulting in more effective early treatment decisions. Consequently, human judgement verification [75] and machine efficiency complement one another. To address data which are the root cause of less-than-optimal data analysis and enhance diagnosis performance, complex data issues necessitate visual learning as a multifaceted solution. It is advantageous to use the diagnostic and treatment options provided by machine learning in cooperation with human medical experts to create the best healthcare system.

5.1. Limitations and future recommendations

The medical datasets are typically collected from a particular area, in a variety of formats, are challenging to access, and have undergone extensive preprocessing. Additionally, the datasets in real-world health scenarios are typically not ready for data analysis and machine learning models implementation. To make informed decisions for early diagnosis and personalised treatment when performing real diagnosis on a dataset, the actual circumstances under which the data were gathered, cleaned, and processed must be aligned with the diagnosis task at hand.

Each data should be taken into account separately when creating synthetic data from original data because data are all unique and may contain more or fewer challenges than the data from published research these are compared to. As a result, while general comparisons with published studies regarding data analysis are considered, decisions regarding specific treatments should always be made in coordination with the input of human experts. However, the decisions of experts may be influenced by competing interests, geographic ties, schools of thought, different types of treatments, and differing opinions.

The generated data might, to a certain extent, maybe devoided of statistical characteristics, which would have an impact on how diagnoses and treatments are determined. The limitation of visual learning is that at one point the statistics characteristics of the synthetic data will be lost, therefore it is essential to perform the first experiment to check that these are preserved. To determine whether a dataset has the potential to be further manipulated, visual experiments should be carried out regularly after any modification. The first experiment provides a valuable resource for evaluating a dataset's statistical properties.

In future research, the data distribution and processing considerations will be added to the prediction model at the algorithm level and applying other algorithms to generate synthetic data. This could provide more opportunities for applying high-performance machine learning techniques which require big data in downstream data processing, focusing on the intersection of data and algorithms for disease prediction. In addition, it is important to investigate and trace the interpretability and relevance of the generated data in future studies to compare human decisions based on visualisations and machine-generated solutions. Future research may focus on exploring visual learning, for example formalising and analysing the corresponding time complexity of visual learning, furthermore, considering the issues in Table 10.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.W. Berry, A. Mohamed, B.W. Yap, Supervised and Unsupervised Learning for Data Science, Springer, 2019.
[2] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.
[3] M.M. Ghiasi, S. Zendehboudi, Decision tree-based methodology to select a proper approach for wart treatment, Comput. Biol. Med. 108 (2019) 400–409.
[4] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artif. Intell. Med. 104 (2020) 101822.
[5] Y. Xiao, J. Wu, Z. Lin, Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data, Comput. Biol. Med. 135 (2021) 104540.
[6] A. Aggarwal, M. Mittal, G. Battineni, Generative adversarial network: An overview of theory and applications, Int. J. Inf. Manag. Data Insights 1 (1) (2021) 100004.
[7] M.F. Faruque, I.H. Sarker, et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in: 2019 International Conference on Electrical, Computer and Communication Engineering, ECCE, IEEE, 2019, pp. 1–4.
[8] K.M. Almustafa, Prediction of heart disease and classifiers' sensitivity analysis, BMC Bioinform. 21 (1) (2020) 1–18.
[9] S. Abbas, Z. Jalil, A.R. Javed, I. Batool, M.Z. Khan, A. Noorwali, T.R. Gadekallu, A. Akbar, BCD-WERT: A novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm, PeerJ Comput. Sci. 7 (2021) e390.
[10] M.M. Ali, B.K. Paul, K. Ahmed, F.M. Bui, J.M. Quinn, M.A. Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Comput. Biol. Med. 136 (2021) 104672.
[11] F. Engelberger, P. Galaz-Davison, G. Bravo, M. Rivera, C.A. Ramírez-Sarmiento, Developing and Implementing Cloud-Based Tutorials That Combine Bioinformatics Software, Interactive Coding, and Visualization Exercises for Distance Learning on Structural Bioinformatics, ACS Publications, 2021.
[12] UCI, UCI machine learning repository: heart disease dataset, URL https://bit.ly/44W8zAR.
[13] UCI, UCI machine learning repository: breast cancer dataset wisconsin (Diagnostic), URL https://bit.ly/3pSsRMV.
[14] UCI, Original Wisconsin breast cancer database, URL https://bit.ly/3Dto07X.
[15] UCI, UCI machine learning repository: immunotherapy dataset, URL https://bit.ly/3q3fOrV.

[16] UCI, UCI machine learning repository: cryotherapy dataset, URL https://bit.ly/44TjfQI.

[17] UCI, UCI machine learning repository: exasens dataset, URL https://bit.ly/43HFYhw.

[18] Kaggle, Pima Indians diabetes database, URL https://bit.ly/3Y2kquM.

[19] A.G. Qasem, S.S. Lam, Prediction of wart treatment response using a hybrid GA-ensemble learning approach, Expert Syst. Appl. 221 (2023) 119737.

[20] K.C. Asanya, M. Kharrat, A.U. Udom, E. Torsen, Robust Bayesian approach to logistic regression modeling in small sample size utilizing a weakly informative student'st prior distribution, Comm. Statist. Theory Methods 52 (2) (2023) 283–293.

[21] D.P. Alamsyah, Y. Ramdhani, T. Arifin, F. Febrilla, S. Setiawan, Prediction of immunotherapy success rate: Particle swarm optimization approach, in: 2022 2nd International Conference on Intelligent Technologies, CONIT, IEEE, 2022, pp. 1–5.

[22] U. Erdiansyah, A.I. Lubis, K. Erwansyah, Komparasi metode K-nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil, J. Media Inf. Budidarma 6 (1) (2022) 208–214.

[23] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, S. Nahavandi, An expert system for selecting wart treatment method, Comput. Biol. Med. 81 (2017) 167–175.

[24] S.B. Akben, Predicting the success of wart treatment methods using decision tree based fuzzy informative images, Biocybern. Biomed. Eng. 38 (4) (2018) 819–827.

[25] S. Khatri, D. Arora, A. Kumar, Enhancing decision tree classification accuracy through genetically programmed attributes for wart treatment method identification, Procedia Comput. Sci. 132 (2018) 1685–1694.

[26] A. Mishra, U.S. Reddy, Machine learning approach for wart treatment selection: Prominence on performance assessment, Netw. Model. Anal. Health Inf. Bioinform. 9 (2020) 1–14.

[27] J. Hu, X. Ou, P. Liang, B. Li, Applying particle swarm optimization-based decision tree classifier for wart treatment selection, Complex Intell. Syst. (2021) 1–15.

[28] A.Y. Mahmoud, D. Neagu, D. Scrimieri, A.R.A. Abdullatif, Review of immunotherapy classification: Application domains, datasets, algorithms and software tools from machine learning perspective, in: 2022 32nd Conference of Open Innovations Association, FRUCT, IEEE, 2022, pp. 152–161.

[29] A.Y. Mahmoud, Efficiency of immunotherapy treatments of warts utilising random forest and decision trees, Intell.-Based Med (2023) Under review.

[30] A.Y. Mahmoud, Preliminary introduction and implementation of novel machine learning algorithm utilising Pareto principle: classification of small biomedical health-related datasets, in: Advances in Computational Intelligence Systems - Contributions Presented At the 21st UK Workshop on Computational Intelligence, September 7-9, 2022, Sheffield, UK, Springer.

[31] A.Y. Mahmoud, D. Neagu, D. Scrimieri, A.R.A. Abdullatif, Machine learning experiments with artificially generated big data from small immunotherapy datasets, in: 2022 21st IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2022, pp. 986–991.

[32] A.Y. Mahmoud, Classification of imbalanced immunotherapy and health-related data utilising novel machine learning experiments, in: Advances in Computational Intelligence Systems - Contributions Presented At the 21st UK Workshop on Computational Intelligence, September 7-9, 2022, Sheffield, UK, Springer.

[33] B. Saravi, F. Hassel, S. Ülkümen, A. Zink, V. Shavlokhova, S. Couillard-Despres, M. Boeker, P. Obid, G.M. Lang, Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models, J. Personal. Med. 12 (4) (2022) 509.

[34] G. Varoquaux, V. Cheplygina, Machine learning for medical imaging: methodological failures and recommendations for the future, NPJ Digit. Med. 5 (1) (2022) 48.

[35] T. Ramesh, U.K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, M. Hamdi, Predictive analysis of heart diseases with machine learning approaches, Malaysian J. Comput. Sci. (2022) 132–148.

[36] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and AI for breast cancer screening: A retrospective analysis, Lancet Digit. Health 4 (7) (2022) e507–e519.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[38] D. Saxena, J. Cao, Generative adversarial networks (GANs) challenges, solutions, and future directions, ACM Comput. Surv. 54 (3) (2021) 1–42.

[39] L. Jin, F. Tan, S. Jiang, Generative adversarial network technologies and applications in computer vision, Comput. Intell. Neurosci. 2020 (2020).

[40] M. Wiatrak, S.V. Albrecht, A. Nystrom, Stabilizing generative adversarial networks: A survey, 2019, arXiv preprint arXiv:1910.00927.

[41] M. Lee, J. Seok, Regularization methods for generative adversarial networks: An overview of recent studies, 2020, arXiv preprint arXiv:2005.09165.

[42] X. Guo, J. Hong, T. Lin, N. Yang, Relaxed wasserstein with applications to GANs, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 3325–3329.

[43] A.-L. Boulesteix, S. Janitza, J. Kruppa, I.R. König, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 2 (6) (2012) 493–507.

[44] S. Salman, M.S. Ahmed, A.M. Ibrahim, O.M. Mattar, H. El-Shirbiny, S. Sarsik, A.M. Afifi, R.M. Anis, N.A.Y. Agha, A.I. Abushouk, Intralesional immunotherapy for the treatment of warts: A network meta-analysis, J. Acad. Dermatol. 80 (4) (2019) 922–930.

[45] P. Saranya, P. Asha, Survey on big data analytics in health care, in: 2019 International Conference on Smart Systems and Inventive Technology, ICSSIT, IEEE, 2019, pp. 46–51.

[46] K.A. Tran, O. Kondrashova, A. Bradley, E.D. Williams, J.V. Pearson, N. Waddell, Deep learning in cancer diagnosis, prognosis and treatment selection, Genome Med. 13 (1) (2021) 1–17.

[47] M. Mahmud, M.S. Kaiser, T.M. McGinnity, A. Hussain, Deep learning in mining biological data, Cogn. Comput. 13 (2021) 1–33.

[48] Z. Chen, M. Wu, R. Zhao, F. Guretno, R. Yan, X. Li, Machine remaining useful life prediction via an attention-based deep learning approach, IEEE Trans. Ind. Electron. 68 (3) (2020) 2521–2531.

[49] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, A. Telenti, A primer on deep learning in genomics, Nature Genetics 51 (1) (2019) 12–18.

[50] D. Smith, M. Elliot, J.W. Sakshaug, To link or synthesize? An approach to data quality comparison, ACM J. Data Inf. Qual. (2023).

[51] M.Z. Alam, M.S. Rahman, M.S. Rahman, A random forest based predictor for medical data classification using feature ranking, Inform. Med. Unlocked 15 (2019) 100180.

[52] T.M. Alam, M.A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T.I. Baig, A. Hussain, M.A. Malik, M.M. Raza, S. Ibrar, et al., A model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.

[53] B.F. Yuksel, P. Fazli, U. Mathur, V. Bisht, S.J. Kim, J.J. Lee, S.J. Jin, Y.-T. Siu, J.A. Miele, I. Yoon, Human-in-the-loop machine learning to increase video accessibility for visually impaired and blind users, in: Proceedings of the 2020 ACM Designing Interactive Systems Conference, 2020, pp. 47–60.

[54] R. Munro, R. Monarch, Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI, Simon and Schuster, 2021.

[55] L. Krenmayr, R. Frank, C. Drobig, M. Braungart, J. Seidel, D. Schaudt, R. von Schwerin, K. Stucke-Straub, GaNerAid: Realistic synthetic patient data for clinical trials, Inform. Med. Unlocked 35 (2022) 101118.

[56] N. Omar, N.N. Nazirun, B. Vijayam, A.A. Wahab, H.A. Bahuri, Diabetes subtypes classification for personalized health care: A review, Artif. Intell. Rev. (2022) 1–25.

[57] M.E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, G. Cook, Introduction to radiomics, J. Nucl. Med. 61 (4) (2020) 488–495.

[58] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, M.M. Hoffman, Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, Inf. Fusion 50 (2019) 71–91.

[59] M.J. Willemink, W.A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L.R. Folio, R.M. Summers, D.L. Rubin, M.P. Lungren, Preparing medical imaging data for machine learning, Radiology 295 (1) (2020) 4–15.

[60] P. Asgari, M.M. Miri, F. Asgari, The comparison of selected machine learning techniques and correlation matrix in ICU mortality risk prediction, Inform. Med. Unlocked 31 (2022) 100995.

[61] D. Kawahara, A. Saito, S. Ozawa, Y. Nagata, Image synthesis with deep convolutional generative adversarial networks for material decomposition in dual-energy CT from a kilovoltage CT, Comput. Biol. Med. 128 (2021) 104111.

[62] N. Shehab, M. Badawy, H. Arafat, Big data analytics and preprocessing, in: Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges, Springer, 2021, pp. 25–43.

[63] C.D. Cantwell, Y. Mohamied, K.N. Tzortzis, S. Garasto, C. Houston, R.A. Chowdhury, F.S. Ng, A.A. Bharath, N.S. Peters, Rethinking multiscale cardiac electrophysiology with machine learning and predictive modelling, Comput. Biol. Med. 104 (2019) 339–351.

[64] K.Y. Ngiam, W. Khor, Big data and machine learning algorithms for health-care delivery, Lancet Oncol. 20 (5) (2019) e262–e273.

[65] P.R. Magesh, R.D. Myloth, R.J. Tom, An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery, Comput. Biol. Med. 126 (2020) 104041.

[66] A. Ali, M. Abu-Elkheir, A. Atwan, M. Elmogy, Missing values imputation using fuzzy K-top matching value, J. King Saud Univ.-Comput. Inf. Sci. 35 (1) (2023) 426–437.

[67] M.T. Islam, H.A. Mustafa, Multi-Layer Hybrid (MLH) balancing technique: A combined approach to remove data imbalance, Data Knowl. Eng. 143 (2023) 102105.

[68] M.M. Islam, M.R. Haque, H. Iqbal, M.M. Hasan, M. Hasan, M.N. Kabir, Breast cancer prediction: A comparative study using machine learning techniques, SN Comput. Sci. 1 (2020) 1–14.

[69] O.J. Egwom, M. Hassan, J.J. Tanimu, M. Hamada, O.M. Ogar, An LDA–SVM machine learning model for breast cancer classification, BioMedInformatics 2 (3) (2022) 345–358.

[70] V.J. Kadam, S.M. Jadhav, K. Vijayakumar, Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression, J. Med. Syst. 43 (8) (2019) 263.

[71] Y.F. Hernández-Julio, M.J. Prieto-Guevara, W. Nieto-Bernal, I. Meriño-Fuentes, A. Guerrero-Avendaño, Framework for the development of data-driven mamdani-type fuzzy clinical decision support systems, Diagnostics 9 (2) (2019) 52.

[72] W. Sun, Z. Cai, F. Liu, S. Fang, G. Wang, A survey of data mining technology on electronic medical records, in: 2017 IEEE 19th International Conference on E-Health Networking, Applications and Services, Healthcom, IEEE, 2017, pp. 1–6.

[73] T. Haesevoets, D. De Cremer, K. Dierckx, A. Van Hiel, Human-machine collaboration in managerial decision making, Comput. Hum. Behav. 119 (2021) 106730.

[74] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Gener. Comput. Syst. (2022).

[75] M. Saleem, S. Abbas, T.M. Ghazal, M.A. Khan, N. Sahawneh, M. Ahmad, Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques, Egypt. Inf. J. 23 (3) (2022) 417–426.