A Synthetic Dataset Generation for the Uveitis Pathology Based on MedWGAN Model

Heithem Sliman, Imen Megdiche, Sami Yangui, Aida Drira, Ines Drira, Elyes Lamine

A Synthetic Dataset Generation for the Uveitis Pathology Based on MedWGAN Model

Heithem Sliman
Institut National Universitaire JF Champollion, ISIS Castres, France
heithem.sliman@univ-jfc.fr

Imen Megdiche
Institut National Universitaire JF Champollion, ISIS Castres, IRIT, University of Toulouse, France
imen.megdiche@irit.fr,imen.megdiche@univ-jfc.fr

Sami Yangui
LAAS-CNRS, Toulouse University, INSA Toulouse,France
yangui@laas.fr

Aida Drira
CHU de Nice, France
aidadrira@gmail.com

Ines Drira
CHU de Toulouse, France
drira.ines@gmail.com

Elyes Lamine
Institut National Universitaire JF Champollion, ISIS Castres, CGI Mines Albi, University of Toulouse, France
elyes.lamine@univ-jfc.fr

ABSTRACT

Artificial Intelligence (AI) has undergone considerable development in recent years in the field of medicine and in particular in decision support diagnostic. However, the development of such algorithms depends on the presence of a sufficiently large amount of data to provide reliable results. Unfortunately in medicine, it is not always possible to provide so much data on all pathologies. This problem is particularly true for rare diseases. In this paper we focus on uveitis, a rare disease in ophthalmology which is the third cause of blindness worldwide. This pathology is difficult to diagnose because of the disparity in prevalence of its etiologies. In order to provide physicians with a diagnostic aid system, it would be necessary to have a representative dataset of epidemiological profiles that have been studied for a long time in this domain. This work proposes a breakthrough in this field by suggesting a methodological framework for the generation of an open source dataset based on the crossing of several epidemiological profiles and using data augmentation techniques. The results of these generated synthetic data have been qualitatively validated by specialist physicians in ophthalmology. Our results are very promising and consist in a first brick to promote research in AI on Uveitis disease.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Machine Learning*;

KEYWORDS

Synthetic Datasets, Data Augmentation, Decision Support Systems

ACM Reference Format:
Heithem Sliman, Imen Megdiche, Sami Yangui, Aida Drira, Ines Drira, and Elyes Lamine. 2023. A Synthetic Dataset Generation for the Uveitis Pathology Based on MedWGAN Model. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27-March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3555776.3577648

1 INTRODUCTION

Artificial Intelligence (AI) has undergone considerable development in recent years in the field of medicine and in particular in decision support systems. However, the development of such algorithms depends on the presence of a sufficiently large amount of data to provide reliable results. Unfortunately in medicine, it is not always possible to provide so much data on all pathologies. This problem is particularly true for rare diseases. In this paper, we are interested in a rare disease known in ophthalmology which is Uveitis. The Uveitis corresponds to the inflammation of the intermediate tunic of the eye called uvea, as shown in Figure 1, which is composed of the choroid extended anteriorly by the ciliary body and by the iris. Inflammatory damage to the retina, secondary to primary inflammatory damage to the uvea, is considered to be a full fledged uveitis [22].

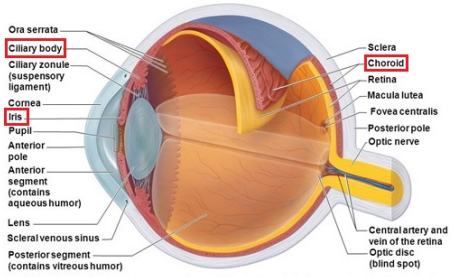

Figure 1: Sagittal section of the human eye

Uveitis is eye affection located at the crossroads of several medical specialties and represents a real diagnostic and therapeutic challenge. It may belong to the manifestations of a general disease

or may affect only the eye. Causes of Uveitis are multiple and heterogeneous, including purely ophthalmological diseases, infectious diseases, systemic diseases, and even drug causes. Sixty possible etiologies are described and classified into 5 groups of unequal importance by the authors of [40].

The Uveitis mainly affects young adults, with 70 to 90% of patients between the ages of 20 and 60, and are responsible for 5% of legal blindness, thus ranking third in the causes of blindness worldwide [4]. Affecting mainly professionally active people, uveitis represents a major public health problem with medico-economic consequences [35]. It is a relatively rare pathology with an incidence of 7 to 52/100,000 people per year, and a prevalence of 38 to 284/100,000 people per year [3, 29, 40]. The incidence of uveitis is estimated in the countries of the northern hemisphere at just over 50 cases per 100,000 inhabitants per year and their prevalence at just over 100 cases per 100,000 inhabitants [5]. In France, an old study carried out in the department of "Savoie" estimated the annual incidence of uveitis at 17 per 100,000 people per year [13].

The low number of cases of uveitis as well as the multidisciplinary management of uveitis has prompted doctors to look for tools to help diagnose these pathologies, in order to shorten the delays in establishing etiological diagnosis. Clinical decision-making can be supported by a set of IT tools called CDSS (Clinical Decision Support System), and there are two main types of these support systems. On the one hand, knowledge-based systems (expert systems) where the intervention of the expert is essential in order to set up the rules, and which require a communication interface between the machine and the user. On the other hand, systems using machine learning.

The first expert systems in medicine appeared in the 70s, defined as " computer systems imitating the approach of the expert in a given field, whatever the method of reasoning used [28] ". MYCIN [47] is one of the first expert systems developed by a group of researchers from Stanford University in 1970, which makes it possible to diagnose and treat certain infectious pathologies based on rules pre-established by specialist physicians , or even INTERNIST-1 [36], an expert system which helps internists to diagnose a range of pathologies.

To help diagnose uveitis, several systems have been developed since the 1990s. We can distinguish the 3D shell expert system [44] which was developed by Wiehler et al., and whose objective was to guide the differential diagnosis of secondary forms of uveitis. We found also a Bayesian network for the differential diagnosis of anterior uveitis [15] developed in 2016 by Gonzalez-Lopez et al. Another system is Uvemaster [12], a medical decision support application, developed and regularly updated since 1992, whose objective is to facilitate the differential diagnosis of uveitis. The development of expert systems requires close collaboration between the business expert and the IT specialist. It can lead to undesirable results if the need of the expert is implicit. On the other hand, the greater the number of criteria involved in the rules, the more complicated it will be to implement these rules in a computer system.

These systems are showing modest results in terms of accuracy [44], [15], [12], and Bayesian networks drop in performance as soon as the number of criteria used is increased [26], hence the interest in using Machine Learning. However, in order to train algorithmic models, it is necessary to have large databases representing the electronic health records for patients who have been diagnosed and followed for uveitis, which is not obvious due to medical data protection policies.

In the recent years, research has focused on distributed privacy preserving data mining (aggregating distributed analytics results) and machine learning model (federated learning) training as a means to avoid data sharing, in addition to traditional data anonymization techniques for privacy preserving data publishing (PPDP) that could allow for data pooling.

Anonymization techniques aim to strike a balance in the final published data between disclosure risk and data utility, resulting in a modified version of the original dataset that no longer identifies individuals. The utility of data anonymized using these methods, on the other hand, is frequently harmed, and the data remains vulnerable to disclosure [23]. A potential solution to these limitations is the generation of fully synthetic data (SD) as an alternative to real data. One of the most promising but underutilized technologies for enabling PPDP and distributed privacy preserving analytics is synthetic data generation (SDG). SD is generated from a model that fits to a real data set. Despite the fact that this model contains no data from the original set, it allows us to generate data which is similar to original data. Research on this direction has been ongoing, with promising results in various application domains such as healthcare, biometrics, and energy consumption, and the need for a robust solution to capitalize on advances in Big Data and AI technology has never been greater [23]. Furthermore, a recent publication describes cases of re-identification in anonymized individual-level data shared in the COVID-19 context, resulting in a decrease in critical information sharing [23]. The use of synthetic tabular data generation (STDG) is proposed as a solution to provide researchers with realistic databases, which can substitute real datasets.

The contribution of this paper consists in a synthetic data generation approach for uveitis disease based on an aggregated epidemiological profile from the literature. An initial data generation algorithm has been developed and the created sample has been validated by specialist physicians . Using this sample of 200 synthetic patient records, we aimed to develop a synthetic data generation approach based on medWGAN model, while respecting the distributions and taking into account the problem of imbalance related to the rarity of some uveitis etiologies. We choose n equal to 2000 for a validation with physicians but the approach is generic, the number of synthetic samples can vary according to the need of the AI algorithms.

The paper is organized as follows : first we present related work on synthetic data generation, then we expose the materials corresponding to the epidemiological profile of Uveitis and our method to generate our synthetic dataset for this rare disease. In section 4, we present the results of the validation of our dataset, and we discuss these results. Finally, we conclude with the perspectives of this work.

2 RELATED WORK

This Section describes and reviews the related work in the literature. We evaluate and classify through categories the relevant approaches for synthetic data generation, as well as, the associated models that could be associated to process them.

Synthetic Data (SD) are data created by a model that has been trained or built to replicate real data (RD) based on its distributions (i.e., shape and variance) and structure (i.e., correlations among attributes) [27].

SD has two main applications: (i) data augmentation, which is used to balance datasets or supplement existing data before training a machine learning model, and (ii) privacy preservation, which is used to allow secure and private sharing of sensitive data.

Synthetic Data Generation (SDG) has been studied in healthcare for a variety of modalities, including biological signals [24], medical pictures [21], free-text content in electronic health records (EHR) [18], time series smart-home activity data [9], and EHR tabular data [45] on which we focus in this paper.

Synthetic tabular data generation approaches can be divided into three categories [23]:

- Classical approaches: Among them, we can distinguish baseline methods, statistical and probabilistic models and ML models.
  - Baseline methods are used for anonymisation; they include simple techniques based on replacing some values, these techniques are based on replacing values, deleting sensitive attributes and adding noise to the data [34].
  - Statistical and probabilistic models synthesize the data using statistical and probabilistic models that attempt to simulate the real data [41].
  - ML models in particular supervised ML models [37].

  These approaches have shown some weakness in generating high quality tabular data that guarantees the privacy of the original data, as they frequently attempt to memorize real data and the correlations between attributes. However, they have frequently served as a benchmark to evaluate more advanced technologies [23].
- Deep Learning approaches: within this group, we find Autoencoders, GANs and Ensembles
  - Autoencoders are unsupervised neural network that learn how to reconstruct data given an encoded representation of the real data [39].
  - GANs consists of two antagonistic neural networks: generator and discriminator, which learn to generate high quality SD by an adversarial training process [19].
  - Ensemble methods in which two different types of DL models are used to generate synthetic data [10].

  These approaches have shown better performance in learning real data patterns and in generating more diverse data, and their use has led to the generation of a higher quality and better privacy preserving tabular data. This is why they have seen a substantial rise in popularity in recent years for synthetic tabular data generation while preserving the privacy of real data [23] particularly GAN-based approaches.
  Other approaches which include personalized methods, techniques, or frameworks that are developed to generate synthetic data for a specific application: (1) Content Modeling for Synthetic E-Health Records (CoMSER) [30] : As a two-step technique, this method was developed to produce synthetic EHRs using publicly available Health Information Statistics and information gathered from experienced doctors without using the real data. (2) Other methods like Aten Framework [31], SynSys [9], Synthea [42], Prophet [25], that have been used to generate realistic synthetic data.
  These approaches have generally behaved as accurate, and had shown good results for each application they were suited for. Once developed, these systems allow us to generate an unlimited number of synthetic patient records. However, there are some limitations to consider. In fact, each method was designed for a specific application, and we need to produce major modifications to turn it suitable for other applications. In addition, these methods require the close collaboration of specialist physicians for the development of data schemes, which takes a considerable time.

For the reasons described above, GANs have gotten a lot of attention after their inception in 2014 [16]. They are considered one of the most interesting developments in AI in recent years, and have shown to be excellent at creating synthetic image data [20]. Because of this promising performance, the development of GANs for alternative data types, particularly tabular data, is a hot topic in AI research right now. Furthermore, in most papers, the suggested GAN-based strategy for synthetic tabular data generation outperformed the other approaches compared using different assessment methodologies [23]. The authors of [23] dressed a comparative table on the different works conducted based on GAN Models.

The majority of these models were tested on de-identified health-related datasets [2, 6, 7], from which we can distinguish:

- A dataset from Sutter Palo Alto Medical Foundation (PAMF), which consists of 10-years of longitudinal medical records of 258K patients
- The MIMIC-III dataset [14], which is a publicly available dataset consisting of the medical records of 46K intensive care unit (ICU) patients over 11 years old.
- The heart failure study dataset from Sutter, which consists of 18-months observation period of 30K patients.
- Longitudinal Health Insurance Database of 498K from Taiwan National Health Insurance Research Database (NHIRD).

There are no standardized metrics or methods to evaluate and benchmark the different approaches for resemblance, utility and privacy dimensions. To overcome this issue, an alternative categorization methodology was proposed and some criterion were established to evaluate the "Poor", “Good” or “Excellent” performance in each of the analyzed dimensions for each publication [23]. They calculated the mean and standard deviation Mean absolute error (MAE) between real data (RD) and synthetic data (SD). The lower values have been classified as "Excellent”, the medium values as "Good" and the higher values as "Poor”. As we will use binary data for the data augmentation, we have focused only in the comparison of GAN based models which were tested on binary data.

The comparison method in [23] has shown that healthGAN [11] presented an excellent level of resemblance , medWGAN [2] and SMOOTH-GAN [38] showed a good level of resemblance, while the resemblance for medGAN [8] and DP-GAN [46] was poor, between RD and SD. Once the table was completed, we were interested in best three models and decided to study them.

The first model we were interested in was the HealthGAN proposed by [11], since its use led to an excellent resemblance between RD and SD. This model was in fact a WGAN model originally developed by [1], which generally facilitates stable training but generates low quality samples or fails to converge in some settings due to the the use of the weight-clipping technique [2].

The second studied model was SMOOTH-GAN, a novel model proposed by [38]. It is a conditional GAN based on WGAN-GP, adapting it for healthcare data. To overcome the issues encountered with WGAN model, the authors of the WGAN-GP model offered an alternative method of weight clipping called gradient penalty, which includes penalizing the norm of the gradient of the discriminator (critic) with respect to its input. Using this technique, the WGAN-GP showed a better performance than many GAN architectures, including the standard WGAN [2].

Finally, we analysed the third model that is the medWGAN, proposed by [2]. This model is based on the medGAN model developed by [8], which resolved some limitations of the original GAN by adding an autoencoder to the model architecture, and by using the minibatch averaging technique that significantly improves the model performance. In the proposed medWGAN, Baowaly et al. employed the same architecture of medGAN, but instead of using the general GAN, they replaced it by the WGAN-GP. In this way, they succeeded to combine the advantages of two of the most relevant models: medGAN and WGAN-GP.

For the reasons described above, we opted for the medWGAN model, and we used the official source code for this model, publicly available at Github[1].

3 MATERIALS AND METHODS

3.1 Materials

It is essential to consider the epidemiology of uveitis because the diagnostic approach will be oriented towards the search for the most frequent etiologies in the population studied, which will have important consequences on the quality of the therapeutic management. Indeed, the Causal epidemiology varies according to genetic factors (HLA-B27 antigen, in the first place), environmental (outbreaks of tuberculosis), the definition of the disease (i.e. sarcoidosis), the inclusion of certain ophthalmological entities in the group of idiopathic uveitis or ophthalmological entities (i.e. pars planitis), paraclinical investigations carried out (i.e. nuclear imaging) and method of patient recruitment (i.e. tertiary centers). This accounts for the great heterogeneity of the series reported in the literature [5].

It is interesting to note that epidemiology of uveitis changes over time in the same geographical region. Thus, in Japan, Behcet's disease, which was the first cause of uveitis 30 years ago, now occupies sixth place, behind sarcoidosis and Vogt-Koyanagi-Harada (VKH) disease. The place of epigenetic factors is also better and better identified. Indeed, the risk of uveitis associated with Behcet's disease is very high in Turkey. Surprisingly, the incidence of the disease in patients of Turkish origin migrating to Germany quickly reaches that of the German population [17]. We aim to generate a realistic database that could represent French patients treated and followed for uveitis, and to respect the epidemiological characteristics already mentioned, we have sought to draw up an epidemiological profile of uveitis extracted from the most recent French descriptive studies. To do that , we identified three interesting retrospective studies on french patients followed for uveitis:

(1) The first study was conducted on 121 patients treated for uveitis in the ophthalmology department of the Croix-Rousse hospital in Lyon [33], from January 2002 to December 2006. Uveitis associated with the virus human immunodeficiency and post-traumatic or post-surgical endophthalmitis were excluded from the cohort. Those lost to follow-up during the 4-year of the assessment were also excluded.
(2) The second study is a retrospective epidemiological study of 690 patients with a diagnosis of uveitis, examined for the first time at the ophthalmology consultation at the Nancy regional university hospital center [32] between 1st January 2005 and 31 December 2016, and sent at the Regional Competence Center dedicated to systemic and autoimmune diseases for diagnostic and/or therapeutic management. The non-inclusion criteria were as follows: patients aged fewer than 18, patients with a first episode of acute anterior uveitis responding well to topical treatment, patients for whom the etiological diagnosis could be made by the ophthalmologist after clinical examination without the need for initiation of systemic treatment.
(3) The third study included 960 patients aged at least 18 years treated at the specialized "uveitis" consultation of the Montpellier University Hospital [35] between January 2003 and August 2018.

After discussing with our specialist physicians in ophthalmology, we considered the following points to generate the profile that we will use for our synthetic data generation :

- Our reference profile is constructed based on the three studies described above, We calculated for each etiology an average prevalence weighted by the size of each of the three populations.
- all AS-type uveitis are included in HLA-B27 uveitis, then all AS (Ankylosing spondylitis) and HLA-B27 patients can be compiled under the same etiology, which we called HLA-B27/AS.
- To represent the clinical examination results with each of the etiologies identified, we used the documents of "THE STANDARDIZATION OF UVEITIS NOMENCLATURE (SUN) WORKING GROUP" published in 2021 [17]. We used the clinical description of 15 etiologies treated by the SUN working group: HLA-B27/AS, Sarcoidosis, Multifocal choroiditis, serpiginous choroiditis, Toxoplasmosis, Herpes simplex virus(HSV), Fuchs, Birdshot, Behcet, Syphilis, Varicella zoster virus(VZV), VKH, Tuberculosis, Tubulointerstitial Nephritis and Uveitis Syndrome(TINU), Multiple sclerosis (MS). We then included all remaining etiologies in the group of idiopathic uveitis.

This work results on a profile containing all the etiologies mentioned by these three studies and taking into account the recommendations of our specialist physicians . We show in table 1 the first six etiologies from the total of 43 etiology.

[1] https://github.com/baowaly/SynthEHR

Etiology	Percentage
idiopathic	42.521%
HLA-B27 / AS	18.181%
sarcoidosis	6.657%
Multifocal choroiditis	5.962%
Toxoplasmosis	4.888%
HSV	4.28%

Table 1: Extract from the resulting epidemiological profile of uveitis

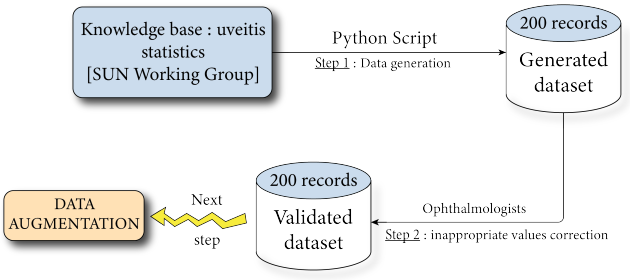

Figure 2: Synthetic data generation methodology

3.2 Method

In this section, we describe our methodology to generate a Synthetic Dataset for uveitis pathology. Our open dataset available in this link [2] is intended for the scientific community to accelerate research and innovation on the diagnosis of Uveitis using advanced AI techniques. This dataset contains simulated data for patients treated for uveitis, based on the epidemiological profile already established in Table 2. To achieve our goal, we had requested the collaboration of specialist physicians in ophthalmology whose participation is essential in order to validate the dataset that we are going to be created.

3.2.1 Data Generation Protocol. To generate the new synthetic dataset, we divided our work into four essential steps:

(1) **Initial dataset creation:** Since we do not have a real dataset of uveitis patients, we opted for the generation of an initial realistic dataset which will then serve as a training base for the data augmentation model(Figure 2). To generate this dataset we use two elements: (1) the epidemiological profile of uveitis to fix the number of patients by etiology and (2) the articles of " THE STANDARDIZATION OF UVEITIS NOMENCLATURE (SUN) WORKING GROUP " [17] recently published to describe the results of the clinical examination within each uveitis etiology. Our generated dataset includes 200 lines, each line presenting the result of the clinical examination of a patient followed for uveitis, while presenting the associated etiology.
(2) **First expert validation:** The three ophthalmologists participating in this work have validated the initial database of 200 patients, which was generated using Python script while taking into account the clinical characteristics of each etiology, and saved in an Excel document. Each expert has examined line by line the document in particular the values generated for each clinical observation. At the end of this step we obtained a database of 200 synthetic patients whose etiological diagnosis is labeled by experienced doctors.
(3) **Data augmentation:** During this step we used the medWGAN model to generate a new database of 2000 patients. The model learn rules from the initial database already validated by ophthalmologists, then allows us to generate the desired number of records.
(4) **Second expert validation:** The specialist physicians selected randomly 200 synthetic patients from the 2000 generated dataset. This represents 10% of validation. The objective is to check whether a random sample of synthetic data is realistic.

3.2.2 Original dataset creation algorithm. To generate our original data we initially created a new dataframe with one column for etiologies and 27 columns for examination results.

Subsequently we fixed the size of the dataset which was 200 rows (line 2); this size allowed us to distribute the different etiologies on the etiology column, each according to its corresponding frequency on the epidemiological profile already defined (lines 3 to 5) of Algorithm 1. Then we proceeded by etiology; for all rows of each etiology we have gone through the columns, and we have filled each column with the corresponding values according to their frequencies described in the knowledge base (lines 6 to 9), while taking into account the conditions that exist between some columns, and which were explained to us by the specialist physicians .

Once this work was done, we saved the created dataset to an Excel file that we then sent to the ophthalmologists for correction and validation.

Algorithm 1: Original dataset generation algorithm

```
VARIABLES : df : DATAFRAME
            len_df : INTEGER
            Etiology : COLUMN
            etiology : STRING
            etiology_records : LIST
            etiology_frequency : FLOAT
            unique_value : STRING
            value_list : LIST
            value_frequency : FLOAT
INPUT     : empty dataframe with named columns
OUTPUT    : generated dataframe
1  begin
2    len_df ← 200
3    Foreach etiology do
4      etiology_records ← etiology
       ×etiology_frequency × len_df
5      Etiology ← etiology_records
6    Foreach etiology do
7      Foreach column ≠ Etiology do
8        Foreach unique_value do
9          value_list
         ← unique_value×value_frequency×etiology_records
10   end
```

[2] https://github.com/heithemsliman/uveitis_dataset_generation.git

3.2.3 Data augmentation algorithm. We used here a GAN-based model to generate our synthetic data, which is medWGAN. The original GAN is made up of two parts: a generator (G) that tries to generate realistic, but fake data, and a discriminator (D) that tries to discern the difference between the generated fake data and the real data. The generator can learn the distribution of real samples by playing an adversarial game against the discriminator if both the generator and the discriminator are sufficiently expressive [7]. The data augmentation model medWGAN is an improved version of medGAN already proposed by Edward Choi et al. in 2017. MedGAN uses a combination of an autoencoder (Enc + Dec) and an adversarial framework to learn the distribution of discrete features such as diagnosis. The autoencoder aids the original GAN in learning the distribution of multi-label discrete variables in this case 3.

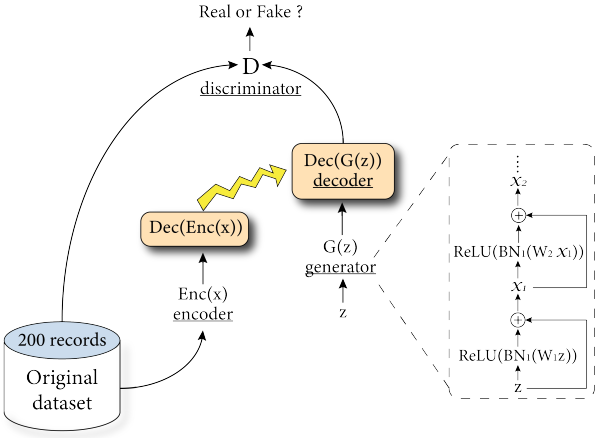

Figure 3: The architecture of the medGAN Algorithm[8]

Architecture of medGAN: The discrete x comes from the source data (original dataset), z is the random prior for the generator G; G is a feed-forward network with shortcut connections (right-hand side figure); An auto-encoder (Enc and Dec) is learned from x; The same decoder Dec is used after the generator G to construct the discrete output. The discriminator D tries to differentiate real input x and discrete synthetic output Dec(G(z)).

As a contribution, medWGAN have used an improved generative network called WGAN-GP (Wasserstein GAN with gradient penalty) instead of the general GAN. The rest of the structure is the same as that of medGAN shown in 3 [2]. The loss function of the original GAN measures the JS (Jensen–Shannon: measure of similarity between two probability) divergence between the distributions of real and generated data. Wasserstein Distance is a measure of the distance between two probability distributions; it is proposed to replace JS divergence because it has a much smoother value space [43].

4 RESULTS

In this section we compared etiologies and features distribution in the original dataset to those in the generated dataset.

4.1 Description of the generated database

After achieving the first generation of 2000 synthetic records by the medWGAN model, we noticed that the error rate is higher within the class of granulomatous uveitis. This is due to the unbalanced nature of our dataset, given that granulomatous uveitis accounted for a quarter of the dataset compared to non-granulomatous uveitis which represent the remaining three quarters. To correct this constraint, we proceeded to balance the data by adding new records of granulomatous uveitis to our original dataset before training the model. These records were generated by the medWGAN network and validated by our specialist physicians , and then we concatenated them to the original dataset initially validated by the specialist physicians . This allowed us to create a balanced dataset of 290 patients that we used for the initial training of our model within 1000 epochs. Once our model has learned the rules, we performed a second training using the original dataset within 500 epochs, having the aim to generate a dataset that reproduces the same distribution of etiologies as that on the original dataset, to respect the epidemiological profile of uveitis used in this work. This method effectively allowed us to keep a correct distribution of etiologies, which is quite evident in Figure 4.

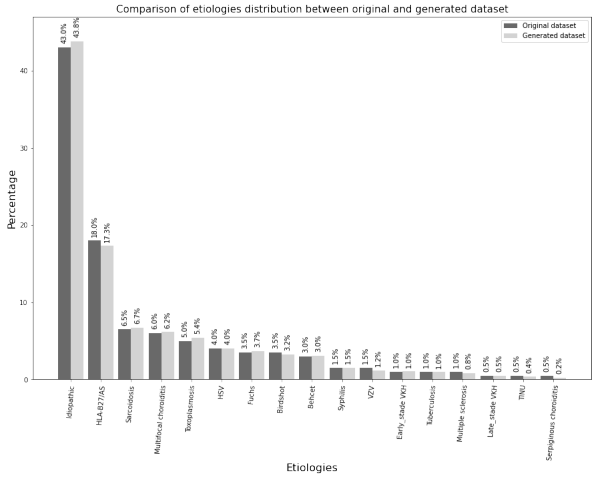

Figure 4: Comparison of etiologies distribution between original and generated dataset

We compared also the values distribution in some columns, and we found that the model we used have kept almost the same distribution on the new generated dataset, which is clearly shown in Table 2.

4.2 Qualitative evaluation

For the purposes of assessing the quality of our generated data, we asked the three ophthalmologists, who participated in this study, to evaluate a sample of 10% from 2000 total records. This step aim to provide a qualitative evaluation of the GAN network used in our work.

To achieve our sampling, we randomly picked 200 samples from the generated dataset, randomly shuffled the order, and then presented them to the ophthalmologists. Our specialist physicians are

		Original	Generated
Age	Age < 30	25%	25.50%
	30 ≤ Age ≤ 60	62%	60.9%
	Age > 60	12%	12.75%
Gender	Female	52.5%	54.2%
	Male	47.5%	45.8%
Uveitis	granul.	28%	29.35%
	non-granul.	72%	70.65%
Duration	chronic	40.5%	47.55%
	acute	33.5%	33.95%
	recurrent	21%	16.1%
	undetermined	5%	2.4%
Laterality	unilateral	54%	54.35%
	bilateral	36.5%	37.9%
	Alternating	9.5%	7.75%

Table 2: Values distribution in original and generated datasets

then asked to determine how realistic those records are, using three classes of description : "Poor", "good", or "excellent" (excellent being most realistic). Due to time constraint, we finally got the assessment for 170 records, and the results have shown that 78 records got "poor" label, 68 records were labeled as "good", and the remaining 24 records as "excellent" Figure 5.

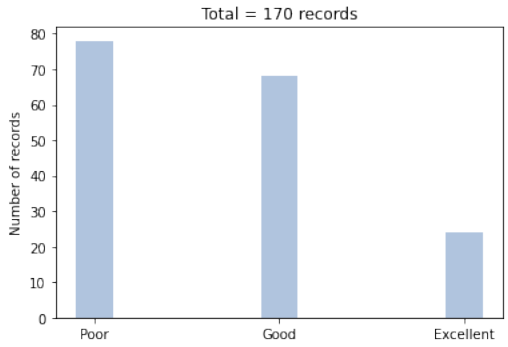

Figure 5: Assessment results for the generated dataset's sample

Discussion. . We have performed a quantitative and qualitative evaluation of our generated dataset.

- The quantitative evaluation has shown that our generated data was of a good quality. The used model kept a similar distribution of etiologies for the generated data as in the original dataset, and this was also noticed when examining the values distribution within the features. The similarity of distributions between real data and generated data is important since we are dealing with a rare disease which requires a specific distribution. Taking these information into account, we can affirm that our generated data meets the requirements of a realistic dataset.
- After the qualitative evaluation, we noticed that about 45% of our randomly selected sample was labeled by our specialist physicians as "poor". These results are not surprising, in fact we used 27 relevant attributes that were available in the SUN articles [17] to describe examination results for each etiology. However, our specialist physicians have noticed the absence of some important attributes that can help the algorithm to properly differentiate the etiologies, but it was difficult to provide additional description, which can be consistent with our data, from the literature. Thus the "poor" qualification is rather interpreted as a lack of information that would be necessary to integrate by additional columns. On the other hand, We can add that 40% of records which got a "poor" label belonged to rare etiologies having prevalence less than 4%, and among all the picked rare etiologies 78% of them were labeled as "poor" and 22% as "good". The remaining 55% of evaluations were classed between good and excellent, these results are very satisfactory for a first version of our dataset.

5 CONCLUSION

In this paper we presented a methodological framework to accelerate Artificial Intelligence in the context of rare diseases. Our methodological framework was developed in collaboration with ophthalmologists on the uveitis disease by drawing on the scientific expertise on the profiles of this disease. We have generated a synthetic dataset based on an epidemiological profile representative of France population through two steps: the first step is a sample generation validated by doctors and a second step is an automatic generation by MedWGAN mechanisms. Our dataset has been evaluated by specialist physicians . The results we obtained are very promising, it is a first dataset available to accelerate the development of AI algorithms to help diagnosis this rare disease. In perspectives of this work, we are going to increase the sample of data validated by doctors and also try to have more points of view of other doctors in the working group. Our objective is to open our work to form a community of volunteer ophthalmologists in order to generate an open source and reliable dataset. A second objective is to compare our sample with real dataset samples extracted from hospitals, this process is long since the prior agreement of the patient in the RGPD framework must be obtained. The third objective is to encourage the creation of a computer science community that would participate in the experimentation of other data augmentation approaches such as SMOTE and the development of AI classification uveitis approaches.

6 ACKNOWLEDGEMENT

We would like to thank the doctors of the CHU of Nice and the CHU of Toulouse for their advice and collaboration during this project.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN.
[2] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2018. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association* 26, 3 (12 2018), 228–241.
[3] Pierre-Jean Bertrand, Yvan Jamilloux, René Ecochard, Gaelle Richard-Colmant, Mathieu Gerfaud-Valentin, Martin Guillaud, Philippe Denis, Laurent Kodjikian,

and Pascal Sève. 2019. Uveitis: Autoimmunity… and beyond. *Autoimmunity Reviews* 18, 9 (2019), 102351.

[4] C. Bonnet and A. Brézin. 2020. Uvéites, éléments d'orientation diagnostique. *Journal Français d'Ophtalmologie* 43, 2 (2020), 145–151.

[5] Antoine P. Brézin. 2012. Uvéites. *La Presse Médicale* 41, 1 (2012), 10–20.

[6] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. 2017. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records.

[7] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research)*, Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.), Vol. 68. PMLR, 286–305.

[8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks.

[9] Jessamyn Dahmen and Diane Cook. 2019. SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors* 19, 5 (2019). https://www.mdpi.com/1424-8220/19/5/1181

[10] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin P. Bennett. 2020. Medical Time-Series Data Generation Using Generative Adversarial Networks. In *Artificial Intelligence in Medicine*, Martin Michalowski and Robert Moskovitch (Eds.). Springer International Publishing, Cham, 382–391.

[11] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin P. Bennett. 2020. Medical Time-Series Data Generation Using Generative Adversarial Networks. In *Artificial Intelligence in Medicine*, Martin Michalowski and Robert Moskovitch (Eds.). Springer International Publishing, Cham, 382–391.

[12] Jose A. Gegundez-Fernandez; Jose I. Fernandez-Vigo; David Diaz-Valle; Rosalia Mendez-Fernandez; Ricardo Cuiña-Sardiña; Enrique Santos-Bueso; Jose M. Benitez del Castillo. 2017. Uvemaster: A Mobile App-Based Decision Support System for the Differential Diagnosis of Uveitis. *Investigative Ophthalmology and Visual Science* 58, 10 (2017), 3931–3939.

[13] Vadot E. 1992. Epidemiology of intermediate uveitis: a prospective study in Savoy. *Developments in Ophthalmology* 23, - (1992), 33–34.

[14] Johnson Alistair E.W, Pollard Tom J, Shen Lu, Lehman Li wei H, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Anthony Celi Leo, and Mark Roger G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 2052–4463.

[15] Garcia-Aparicio AM Gonzalez-Lopez JJ, Munoz-Sanz N Sanchez-Ponce D, Fernandez-Ledo, Beneyto P, and et al. 2016. Development and validation of a Bayesian network for the differential diagnosis of anterior uveitis. *Eye* 30, 64 (2016), 865-872.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc., 2672–2680.

[17] THE STANDARDIZATION OF UVEITIS NOMENCLATURE (SUN) WORKING GROUP. 2021. *American Journal of Ophthalmology* 228 (2021), 1–280.

[18] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of Synthetic Electronic Medical Record Text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 374–380.

[19] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2020. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications.

[20] Alqahtani H., Kavakli Thorne M., and Kumar G. 2021. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Archives of Computational Methods in Engineering* 28 (2021), 525–552.

[21] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. 2018. GAN-based synthetic brain MR image generation. In *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 734–738.

[22] H.A.S Haute Autorité de Santé. 2020. Uvéites Chroniques Non Infectieuses de l'enfant et de l'adulte. *Protocole National de Diagnostic et de Soins, MAI 2020* – (2020).

[23] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic Data Generation for Tabular Health Records: A Systematic Review. *Neurocomputing* 493, C (2022), 28–45.

[24] Andres Hernandez-Matamoros, Hamido Fujita, and Hector Perez-Meana. 2020. A novel approach to create synthetic biomedical signals using BiRNN. *Information Sciences* 541 (2020), 218–241.

[25] Jayun Hyun, Seo Hu Lee, Ha Min Son, Ji-Ung Park, and Tai-Myoung Chung. 2020. A Synthetic Data Generation Model for Diabetic Foot Treatment. In *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications*, Tran Khanh Dang, Josef Küng, Makoto Takizawa, and Tai M. Chung (Eds.). Springer Singapore, Singapore, 249–264.

[26] Yvan Jamilloux, Nicolas Romain-Scelle, Muriel Rabilloud, Coralie Morel, Laurent Kodjikian, Delphine Maucort-Boulch, Philip Bielefeld, and Pascal Sève. 2021. Development and Validation of a Bayesian Network for Supporting the Etiological Diagnosis of Uveitis. *Journal of Clinical Medicine* 10, 15 (2021).

[27] R. Hoptroff K. El Emam. 2019. The Synthetic Data Paradigm for Using and Sharing Data. *Digital Technol.* 19, 6 (2019).

[28] Boullier D Le Beux P. 2001. L'information médicale numérique. *Les cahiers du numérique* 2, 2 (2001), –.

[29] Prete M., Dammacco R., Fatone MC., and Racanelli V. 2016. Autoimmune uveitis: clinical, pathogenetic, and therapeutic features. *Clinical and Experimental Medicine* 16, 2 (2016), 125—136.

[30] Scott McLachlan, Kudakwashe Dube, and Thomas Gallagher. 2016. Using the CareMap with Health Incidents Statistics for Generating the Realistic Synthetic Electronic Healthcare Record. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, Vol. -. 439–448.

[31] Scott McLachlan, Kudakwashe Dube, Thomas Gallagher, Jennifer A. Simmonds, and Norman Fenton. 2019. Realistic Synthetic Data Generation: The ATEN Framework. In *Biomedical Engineering Systems and Technologies*, Alberto Cliquet Jr., Sheldon Wiebe, Paul Anderson, Giovanni Saggio, Reyer Zwiggelaar, Hugo Gamboa, Ana Fred, and Sergi Bermúdez i Badia (Eds.). Springer International Publishing, Cham, 497–523.

[32] E. Neiter, J.-B. Conart, C. Baumann, H. Rousseau, S. Zuily, and K. Angioi-Duprez. 2019. Caractéristiques épidémiologiques et étiologiques des uvéites dans un centre hospitalier universitaire. *Journal Français d'Ophtalmologie* 42, 8 (2019), 844–851.

[33] A.-M. Nguyen, P. Sève, J. Le Scanff, J. Gambrelle, J. Fleury, C. Broussolle, J.-D. Grange, and L. Kodjikian. 2011. Aspects cliniques et étiologiques des uvéites : étude rétrospective de 121 patients adressés à un centre tertiaire d'ophtalmologie. *La Revue de Médecine Interne* 32, 1 (2011), 9–16.

[34] Benjamin Nguyen. 2014. Techniques d'anonymisation. *Statistique et Société* 2, 4 (2014), 53–60.

[35] Sarah PEREZ-ROUSTIT. Octobre 2018. *Epidémiologie, caractéristiques cliniques et étiologiques des uvéites prises en charge au CHU de Montpellier.* Ph.D. Dissertation. Faculté de médecine-université de Montpellier.

[36] Randolph, McNeil, Challinor, Masarie, and Jack Myers. 1986. The INTERNIST-1/QUICK MEDICAL REFERENCE Project—Status Report. *Western Journal of Medicine* 145, 6 (1986), 816–822.

[37] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, and Gorka Epelde. 2020. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Med Inform* 8, 7 (20 Jul 2020), e18910.

[38] Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. 2020. SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation. In *Artificial Intelligence in Medicine*, Martin Michalowski and Robert Moskovitch (Eds.). Springer International Publishing, Cham, 37–48.

[39] Mohit Sewak, Sanjay Sahay, and Hemant Rathore. 2020. An Overview of Deep Learning Architecture of Deep Neural Networks and Autoencoders. *Journal of Computational and Theoretical Nanoscience* 17 (01 2020), 182–188.

[40] P. Sève, B. Bodaghi, S. Trad, J. Sellam, D. Bellocq, P. Bielefeld, D. Sène, G. Kaplanski, D. Monnet, A. Brézin, M. Weber, D. Saadoun, P. Cacoub, C. Chiquet, and L. Kodjikian. 2018. Prise en charge diagnostique des uvéites: recommandations d'un groupe d'experts. *La Revue de Médecine Interne* 39, 9 (2018), 676–686.

[41] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine* 3 (12 2020).

[42] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25, 3 (08 2017), 230–238.

[43] Lilian Weng. 2019. From GAN to WGAN.

[44] Wiehler, U., Schmidt, R., Skonetzki, and S. et al. 2006. Optimierung der differenzialdiagnostischen Strategie bei Patienten mit sekundären Uveitisformen mit einem computergestützten System. *Ophthalmologe* 103, 5 (2006), 406-409.

[45] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* 416 (2020), 244–255.

[46] Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. 2020. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics* 24, 8 (2020), 2378–2388.

[47] Victor L. Yu, Bruce G. Buchanan, Edward H. Shortliffe, Sharon M. Wraith, Randall Davis, A.Carlisle Scott, and Stanley N. Cohen. 1979. Evaluating the performance of a computer-based consultant. *Computer Programs in Biomedicine* 9, 1 (1979), 95–102.