Review

The Impact of Artificial Intelligence in the Odyssey of Rare Diseases

Anna Visibelli [1,*,†], Bianca Roncaglia [1,†], Ottavia Spiga [1,2,3,*] and Annalisa Santucci [1,2,3]

1   Department of Biotechnology, Chemistry and Pharmacy, University of Siena, 53100 Siena, Italy
2   Competence Center ARTES 4.0, 53100 Siena, Italy
3   SienabioACTIVE—SbA, 53100 Siena, Italy
*   Correspondence: anna.visibelli2@unisi.it (A.V.); ottavia.spiga@unisi.it (O.S.)
†   These authors have contributed equally to this work.

Abstract: Emerging machine learning (ML) technologies have the potential to significantly improve the research and treatment of rare diseases, which constitute a vast set of diseases that affect a small proportion of the total population. Artificial Intelligence (AI) algorithms can help to quickly identify patterns and associations that would be difficult or impossible for human analysts to detect. Predictive modeling techniques, such as deep learning, have been used to forecast the progression of rare diseases, enabling the development of more targeted treatments. Moreover, AI has also shown promise in the field of drug development for rare diseases with the identification of subpopulations of patients who may be most likely to respond to a particular drug. This review aims to highlight the achievements of AI algorithms in the study of rare diseases in the past decade and advise researchers on which methods have proven to be most effective. The review will focus on specific rare diseases, as defined by a prevalence rate that does not exceed 1–9/100,000 on Orphanet, and will examine which AI methods have been most successful in their study. We believe this review can guide clinicians and researchers in the successful application of ML in rare diseases.

Keywords: rare disease; machine learning; artificial intelligence; precision medicine; data analysis

1. Introduction

The term rare diseases refers to a vast set of diseases that affect a small proportion of the total population; there are more than 7000 known disorders, and an estimated 250 rare new diseases are discovered annually [1]. In addition, diseases with a prevalence of <1 case per 50,000 population are defined as ultra-rare [2]. This definition is related to a prevalence threshold [3], and it differs depending on the jurisdiction. In Europe, the European Medicines Agency considers a prevalence of less than 5 in 10,000 people (less than 1 in 2000) [4], while in the United States, diseases affecting less than 200,000 people in the country were defined as rare by the Orphan Drug Act in 1983 [5]. In Japan, the Ministry of Health, Labor, and Welfare defines a threshold of fewer than 50,000 individuals in the country (equivalent to less than 1 in 2500 people) [6]. Therefore, an international definition of rare disease is lacking. Although individually they can be considered rare, they collectively afflict more than 500 million people worldwide [7]. Most of these disorders have characteristics that pose serious challenges for both researchers and public health professionals, especially for patients who face not only a loss in terms of health and quality of psychological and social well-being, but also financial burdens [8].

First, the process of diagnosing a rare disease is often long and exhausting. In 25% of patients, it takes between 5 and 30 years after disease onset to receive a correct diagnosis, which requires the participation of a competent and comprehensive clinical team [9]. A survey conducted from October 2019 through March 2020 by the National Organization for Rare Disorders [10] on 1108 individuals found that 50% of patients and caregivers attribute diagnostic delays to a lack of knowledge about the disease, while 42% believe that

delays are caused by limited medical specialization. Many patients identified the problem of doctors not being able to link symptoms, particularly between different organ systems, in addition to the fact that waiting times to consult specialists are long and there would be a need for more tests. A diagnostic delay can have tremendous effects on the patient's clinical picture, so prompt and accurate diagnosis are the starting point for being able to find therapeutic interventions and resources that can ensure a good clinical outcome [11].

What complicates the rare disease odyssey is that the diagnosis is never the end of the journey, since even from a prognostic and therapeutic point of view, there are huge gaps to be filled [12]. The difficulties encountered at the prognostic level are related to the lack of valid parameters and/or biomarkers, since the molecular pathophysiological mechanisms are still largely unknown. Moreover, the small number of patients does not allow statistically significant parameters to be derived [13]. Thus, the prognosis of patients may change depending on various genetic and environmental factors, but it is complicated to arrive at standards of care for treatment and rehabilitation because health research is necessarily conducted on a small scale and cannot be based on evidence or experience [14]. Conventionally, it takes 10 to 15 years to bring a drug to market, with an average R&D cost of $2.6 billion [15]. These two factors represent a bottleneck in the drug discovery pipeline for rare disorders, as research costs are high while revenues are low due to the small number of patients. This implies that the development of new drugs and treatments can be time-consuming and hindered by the lack of data and funding [16]. The heterogeneous patient populations, often unknown etiology and pathogenesis, the timing of disease progression, and the lack of exhaustive clinical studies make the search for specific drugs very difficult [17]. The key problems related to the development of drugs and therapies for rare diseases are [13]:

- only small cohorts of patients are interested in purchasing these drugs, making them so-called orphan drugs because they are not competitive for pharmaceutical companies.
- difficulties in treatment because most rare diseases are caused by genetic errors and/or have a degenerative nature.
- significant percentages of patients do not respond to available therapies due to partial or complete loss of response.

Therefore, rare diseases are often referred to as orphans as they fail to attract political, financial, and research interests, even though laws have been passed over the years to address this problem; the US Orphan Drug Act in 1983 and the European Union Regulation on Orphan Medicines in 2000 have rewarded innovation and focused on the value of healthcare for rare disease patients. Nevertheless, for most of them, there are no adequate therapeutic options. Over the years, specialized interdisciplinary centers for rare disorders have been established, where doctors and researchers can exchange opinions and ideas, creating networks of knowledge and experience that can help patients [9].

Possible innovative answers to biomedical and clinical challenges come from the world of information technology, and a striking example has been the fight against COVID-19. Since the beginning of the pandemic, artificial intelligence (AI) has played a crucial role in the battle against the virus, and several methods have been applied for various purposes [18]. Machine learning (ML) and deep learning (DL) models have been used for the early detection and diagnosis of COVID-19 by monitoring the demographic, clinical, and epidemiological characteristics of patients, and for developing diagnostic tools that can quickly analyze CT scans and X-rays to identify patterns indicative of the disease [19]. AI has also been used to predict patient vulnerability, in order to administer appropriate drugs and treatments [20], as well as being decisive in accelerating the discovery of potential vaccines. Similarly, AI has been an essential ally for public health policies in contact tracing, monitoring the spread of the virus, and creating predictive models that have helped to identify potential outbreaks. Thus, it is clear how AI is increasingly coming to the aid of physicians at every stage of disease management, to evaluate the efficacy of medical treatments or deeply investigate the correlation between patients and treatments according to their own molecular characteristics. The precision medicine approach is widely applied

to the healthcare area, in particular to rare diseases with the creation of patient registries leveraging large amounts of data to discover potential links. It is a comprehensive and prospective approach to prevention, diagnosis, treatment, and monitoring, built on the genetic characteristics of the individual. Harmonizing databases and including registries are the major facilitators to understand the complexity of diseases, to conduct clinical trials, to improve the drug development process, and to assign the right treatment to the right individual after reliable patient stratification. AI is an ally that can integrate and analyze heterogeneous data (e.g., multi-omics data as well as images). However, first-generation AI systems, which rely on the development of algorithms for diagnosis and treatment that are trained on big data, are not always adequate to meet the needs of rare diseases. Data scarcity and sparsity characterize these disorders, due to fragmented knowledge and the limited number of data and specimens available [21]. Phenotype and disease severity as well as pharmacogenomic and pharmacokinetic factors are the elements on which successful diagnosis and treatment depend [13]. Diagnostic decision support systems (DDSS) already exist, i.e., expert systems that support doctors in facilitating the diagnostic process by incorporating medical knowledge. These systems have been proven effective and have improved clinical diagnosis by compiling lists of appropriate differential diagnoses for a given sample of tests [22,23]. For rare diseases, these systems need to be implemented. Networks and registries have been built to bring together data and expertise on rare diseases, making them free, accessible, and shareable worldwide. One example is Orphanet, which over the past 20 years has become the go-to source for information on rare diseases, facilitating access to information and means to identify potential patients, and contributing to the development and sharing of knowledge. Other available datasets are the Online Mendelian Inheritance in Man and Human Phenotype Ontology. The knowledge deposited in these databases is used by DDSSs built for rare disorders, some examples of which are FindZebra [24], PhenoTips [25], Rare Disease Discovery [26], and Ada DX [27].

Second-generation AI systems were designed to fill the diagnostic, prognostic, and therapeutic gaps that must be overcome to achieve patient-centricity for patients with rare diseases [28] (Figure 1). These systems use a personalized closed-loop system designed to enhance end-organ function, overcome problems of tolerance or loss of efficacy, and improve patients' responses to chronic drugs [13] in a precision medicine perspective.
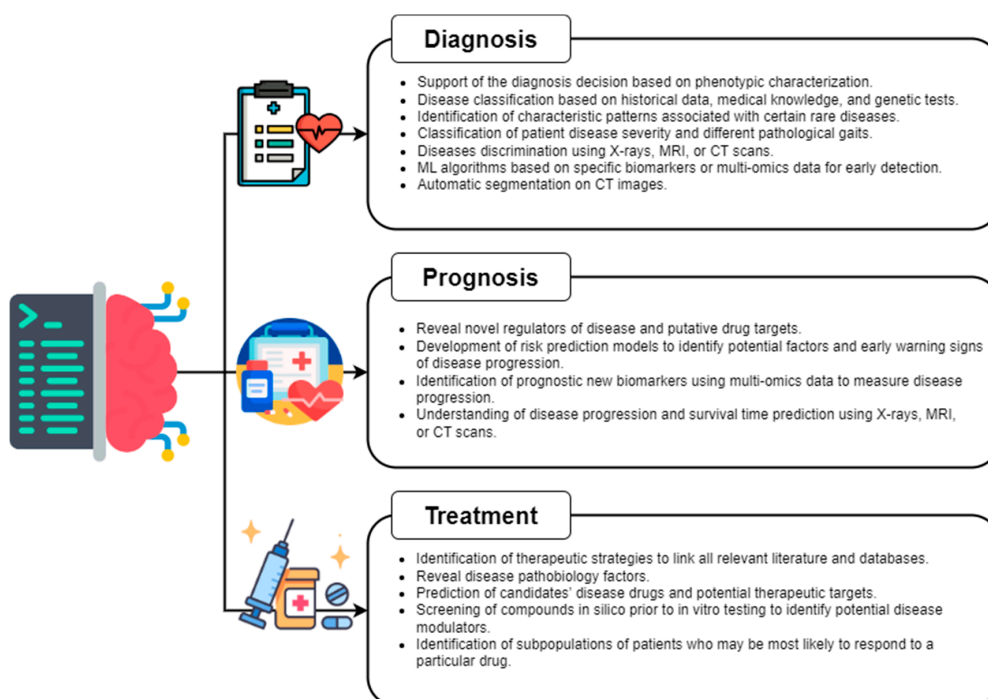

Diagnosis
- Support of the diagnosis decision based on phenotypic characterization.
- Disease classification based on historical data, medical knowledge, and genetic tests.
- Identification of characteristic patterns associated with certain rare diseases.
- Classification of patient disease severity and different pathological gaits.
- Diseases discrimination using X-rays, MRI, or CT scans.
- ML algorithms based on specific biomarkers or multi-omics data for early detection.
- Automatic segmentation on CT images.

Prognosis
- Reveal novel regulators of disease and putative drug targets.
- Development of risk prediction models to identify potential factors and early warning signs of disease progression.
- Identification of prognostic new biomarkers using multi-omics data to measure disease progression.
- Understanding of disease progression and survival time prediction using X-rays, MRI, or CT scans.

Treatment
- Identification of therapeutic strategies to link all relevant literature and databases.
- Reveal disease pathobiology factors.
- Prediction of candidates' disease drugs and potential therapeutic targets.
- Screening of compounds in silico prior to in vitro testing to identify potential disease modulators.
- Identification of subpopulations of patients who may be most likely to respond to a particular drug.

Figure 1. Examples of ML applications within diagnosis, prognosis, and treatment.

So far, two scoping reviews [29,30] have been written on the use of ML and DL in rare diseases. The purpose of this review is to highlight the successes of AI algorithms in the study of rare diseases in the past decade and provide researchers with guidance on which methods have proven to be most effective. The review will focus on specific rare diseases, as defined by a prevalence rate that does not exceed 1–9/100,000 on Orphanet, and examine which AI methods have been most successful in their study.

2. Artificial Intelligence

The scientific field called AI [31] tries to develop robots that can mimic human perception and acquire the ability to solve problems for themselves. The fundamental subsets of AI are ML, which is based on the premise that computers can learn to execute certain jobs by gaining experience and improving their skills, and DL, which includes models of increasing complexity and abstraction. At the base of ML there are a series of different algorithms which, starting from primitive notions, learn to make a specific decision or to perform actions learned over time. Only a collection of data (training set) is given to the machine, which is iteratively evaluated to extract information, similarly to how humans learn.

Based on how the computer learns data and information, four distinct learning approaches can be identified: supervised learning [32], unsupervised learning [33], semi-supervised learning [34], and reinforcement learning [35]. Furthermore, the ML process consists of six components regardless of the algorithm adopted [36]. Data collection and pre-processing refer to the preparation of data, which is generally unstructured and sparse, in a format suitable for the algorithm's input. The second step includes the standardization of the dataset, which helps to learn algorithms to avoid bias in the results. Later, feature selection can be applied to limit the number of input variables and reduce computational costs, thus improving efficiency. Even so, not all ML techniques are appropriate for all issues; rather, specific algorithms are better suited to a particular class of challenges. Any ML model's ultimate goal is to learn from examples in a way that allows it to apply what it has learned to novel situations never encountered before. The model should then be trained on a subset of the total dataset, and its performance should then be measured against unknown data. Moreover, an ML tool may also do a wide range of tasks:

- Classification: two or more classes are created with the input data, and the learning system aims to produce a model capable of assigning a class to each input.
- Regression: conceptually similar to classification, with the difference that the output is continuous.
- Clustering: data are divided into groups about which there is no prior knowledge.

Decision trees, genetic and boosting algorithms, and metric techniques, such as the K-nearest neighbor algorithm (k-NN), Support Vector Machines (SVM), statistical approaches, Bayesian models (BM), Artificial Neural Networks (ANN), and Ensemble methods are all examples of ML techniques, which have a wide range of applications across various disciplines.

Numerous instances of multidisciplinary research in the particular context of experimental biology and biomedicine have shown the potential effectiveness of these methods. The most common AI techniques used include the DL subclass, which has opened the door to exploring tasks that would be difficult to address using shallow approaches [37]. Although it has a wide range of applicability, AI is still not a common technology employed by scientists. To be precise and effective, ML algorithms require large volumes of data, which is a challenging task in the biological field because digital information is scattered and not always accessible, making it difficult to make accurate predictions.

3. AI Application in Rare Diseases

To identify scientific articles that apply ML in the field of rare diseases, papers containing ML data on rare diseases were searched on Pubmed. The search strings we used include general terms related to AI ("artificial intelligence", "machine learning", "deep learning") and diseases ("rare disease", "ultra-rare disease", "orphan disease"). Thus, we chose to

include papers on diseases defined by a prevalence rate that does not exceed 1–9/100,000 on Orphanet. We included papers published between 1 January 2013, and 31 December 2022, and the studies identified in the search had to fulfill the following eligibility criteria: rare disease topic and use of at least one ML method with sufficient detail to extract the basic information analyzed in this review. Twenty-nine unique rare diseases were identified from the reviewed articles. After having selected relevant studies according to the eligibility criteria, we divided them based on the medical application, i.e., diagnosis, treatment, and prognosis. Then, the medical study, input data, and algorithm type and performances were assessed in detail for each study.

3.1. Diagnosis

Accurate diagnosis of rare diseases is an important task in patient triage, risk stratification, and targeted therapies. Rare disease symptoms often appear unfamiliar and atypical to a clinician due to their infrequency, and the likelihood that patients will not get an appropriate diagnosis and subsequent successful therapy is highest. The variability of rare diseases also makes it difficult to identify corresponding diseases in a timely manner due to the lack of clinical diagnostic procedures accessible.

A typical approach for the diagnosis of a rare disease includes a thorough medical history, physical examination, and genetic testing, which may identify specific mutations that are associated with the disease. Additionally, imaging studies such as X-rays, MRI, or CT scans may also be used. In this context, AI has the potential to play a significant but challenging role, through the development of ML algorithms that can analyze large amounts of data to identify patterns and markers that are characteristic of specific rare diseases. Moreover, AI-based diagnostic tools can also help to reduce the time and costs associated with diagnosing rare diseases by identifying potential diagnoses more quickly and accurately. Many ML techniques have been created to help in standardizing and sharing clinical and medical words through diverse medical resources, in order to improve inter-operability in the field of rare diseases. However, ML algorithms often require a significant number of training examples to achieve a good generalization performance, while the number of relevant clinical records in this field is bounded by the size of the population.

New strategies have been used to compensate for the lack of training data for rare disease diagnosis. For example, in [38], based on the requirement of providers to document associated phenotypic information to support a diagnosis, authors hypothesize that patients' phenotypic data stored in electronic medical records can be used to speed up disease diagnosis. In this study, they suggested a collaborative filtering method enhanced with natural language processing and semantic techniques to help with phenotypic characterization-based rare disease identification. The preliminary results obtained demonstrated that the use of collaborative filtering with phenotypic information can stratify patients with relatively similar rare diseases. In [39], the phenotype-based Rare Disease Auxiliary Diagnosis system was developed, adopting both the traditional phenotypic similarity method and a new ML method to build four diagnostic models to support the diagnosis of rare diseases. Each model provides, with high diagnostic precision, a list of the top 10 candidate diseases as the prediction outcome. In another study [40] based on the fact that clinical symptoms in children with pulmonary diseases are frequently non-specific, authors developed and tested a questionnaire-based and data mining-supported tool, providing diagnostic support for selected pulmonary diseases. Eight different classifiers and an ensemble classifier were developed and trained to categorize any given new questionnaire and suggest a diagnosis. All questionnaires of patients suffering from cystic fibrosis, asthma, primary ciliary dyskinesia, acute bronchitis, and the healthy control group were correctly diagnosed by the fusion algorithm and exhibited good results in arriving at diagnostic suggestions. Moreover, due to the very nature of rare diseases, the lack of historical data poses a great challenge to ML-based approaches in accurately identifying rare diseases based on symptom descriptions. This work [41] used medical knowledge

in automatically constructed knowledge graphs to develop a rare disease classification algorithm, delivering robust performance and outperforming a wide range of baselines.

More than one method has been applied to Huntington's Disease (HD). This is a rare, inherited, neurodegenerative disorder that causes the progressive breakdown of nerve cells in the brain and leads to the loss of cognitive, behavioral, and physical abilities. It typically develops between the ages of 30 and 50, and the most visible symptom is chorea, which consists of involuntary movements of the upper and lower extremities, face, or body, and occurs in about 90% of patients. There is currently no cure for HD, but treatments are available to manage symptoms and improve quality of life. Reliable markers measuring disease progression in HD, before and after disease manifestation, may guide a therapy aimed at slowing or halting disease progression. ML methods have been widely used for gait assessment through the estimation of spatio-temporal parameters, demonstrating that the application of supervised classification methods is a valuable and promising approach to the automatic detection of disease stages in HD. In [42], Zhang et al. investigate the potential of classifying patient disease severity based on individual footstep pressure data using DL techniques. Using the Motor Subscale of the Unified HD Rating Scale as the gold standard, the experiments performed showed that use of VGG16 and similar modules can achieve high classification accuracy. The objective of the work described in [43] was instead to propose a validated SVM classifier that takes advantage of Hidden Markov Model-derived information for the classification of different pathological gaits. Specifically, the presented methodology allowed for proper discrimination against gait data from HD patients and healthy elderly controls using data from inertial measurement units placed at the shank and waist. Furthermore, alterations in oculomotor performance are among the first observable physical alterations during the pre-symptomatic stages of HD. In the pre-symptomatic and early symptomatic stages of HD, quantifiable assessments of oculomotor function have been investigated as potential markers of disease state and development. In [44], Miranda et al. reported the application of the SVM algorithm to oculomotor features pooled from a four-task psychophysical experiment. They were able to automatically distinguish control participants from pre-symptomatic HD participants and HD patients with high accuracy. Finally, quantitative electroencephalography (qEEG) may also provide a quantification method for possible sub-cortical dysfunction occurring before, or concomitant with, motor or cognitive disturbances observed in HD. In this pilot study [45], the authors constructed an automatic classifier, distinguishing healthy controls from HD gene carriers using qEEG. Derived qEEG features that correlated with clinically known markers represented new potential biomarkers of HD disease progression.

Starting from the assumption that bio-imaging technologies are increasingly impacting life sciences, and that sharing of image data is required to enable innovative future research, there are several rare disease studies that use images as input data. Parkinson's disease (PD) and multiple system atrophy (MSA) are two neurodegenerative diseases that can have overlapping clinical manifestations. MSA is a progressive rare neurodegenerative disorder characterized by a combination of symptoms that affect both the autonomic nervous system and movement. This is caused by the progressive degeneration of neurons in several parts of the brain and spinal cord. The objective of the studies described in [46,47] were to assess the potential of SVM techniques to distinguish between PD and MSA patients at the single-patient level. Measures of cerebellar-brain network and cerebellar-striatal connectivity and subcortical edge-wise tractography data were used as predicting features in the articles respectively. Convolutional neural networks (CNN) were used in [48] to distinguish each representative parkinsonian disorder using a single midsagittal MRI. CNN enabled accurate discrimination among PD, progressive supranuclear palsy, MSA with predominant parkinsonian features, and normal status, although the dataset was limited.

Amyotrophic lateral sclerosis (ALS) is also a neurodegenerative rare disorder that affects nerve cells in the brain and spinal cord. The disease is progressive and leads to increasing disability, with patients eventually losing the ability to speak, swallow, and breathe. There is no known cure for ALS, and treatment options are focused on managing

symptoms and prolonging survival. In [49], a deep CNN was developed for the classification of ALS patients and healthy individuals. Based on the recent insight that regulatory regions harbor the majority of disease-associated variants, authors employed a two-step approach: promoter regions that are likely associated with ALS have been identified, and individuals were classified based on their genotype in the selected genomic regions to identify potentially ALS-associated promoter regions. The application of a new advanced neuroimaging method, which delineates the profile of tissue properties along the corticospinal tract of patients with ALS using diffusion tensor imaging (DTI), was described in [50]. RF was used to assess the clinical utility of DTI in discriminating ALS from controls, with the potential to be of diagnostic utility in ALS. Finally, in [51], the authors utilized independent component analysis to derive brain networks based on resting-state functional magnetic resonance imaging and used those derived networks to build an ALS disease state classifier using SVM.

More generally, SVM methods have been widely and differently applied in the field of rare diseases. In this study, Palstrøm et al. [52] aimed to improve the diagnosis of amyloidosis by developing unbiased models based on proteomics data for the recognition of amyloid-containing biopsies, followed by accurate subtyping of amyloidosis. The authors demonstrated that using SVM on proteomics data can identify and classify patients with high accuracy. Hypophosphatasia is a rare genetic disease in which patients may have stress fractures, bone and joint pain, or premature tooth loss. In [53], the authors developed several ML algorithms based on specific biomarkers of this disease, determining the best way to diagnose this condition. SVM was the ML algorithm that provided the best predictive models in terms of classification. Nguyen, et al. [54] proposed a measuring instrument based on ML to quantitatively assess impairment levels while engaged in daily activity, for monitoring the progression of neurodegenerative conditions of Friedreich ataxia. Movement patterns during a simulated eating task were captured and kinematic biomarkers were extracted that were consistent with the frequently used clinical rating scales. SVM and other methods have been shown to accurately classify individuals with Friedreich ataxia and control subjects. The work in [55] aimed to assess the feasibility of a supervised ML algorithm for the assisted diagnosis of patients with clinically diagnosed progressive supranuclear palsy (PSP), a rare neurodegenerative disorder that shares similar clinical symptoms with PD. Morphological MRI of PD patients, PSP patients, and healthy control subjects was used as the input of a supervised ML algorithm based on the combination of PCA as a feature extraction technique and SVM as a classification algorithm. The authors in [56] characterized the 3D structure of the cortical bone in high-resolution micro-CT images to analyze the micro-structural properties of bone in cases of osteogenesis imperfecta (OI), a genetic disorder of connective tissues caused by an abnormality in the synthesis or processing of collagen. Numerous features computed from the image were used in an SVM model to classify between healthy and OI bone.

ANN and DL models have been shown to be highly effective in identifying and classifying diseases, and are becoming increasingly popular in the medical field as a tool for accurate and efficient diagnosis. In both [57,58], NN models were applied to eye photographs with the aim of identifying rare diseases. A hybrid learning-based neural network classifier (HLNNC) was implemented in [57] to identify mucormycosis disease by comparing images of patients with and without mucormycosis, a rare fungal infection caused by a group of molds. In [58], the discrimination ability of a deep CNN for ultrawide-field pseudocolor imaging and ultrawide-field autofluorescence was demonstrated for the detection of retinitis pigmentosa, a complex hereditary eye condition that causes cells in the light-sensitive retina to degenerate. Using the proposed model, retinitis pigmentosa was distinguished from healthy eyes with high sensitivity and specificity on ultrawide-field pseudocolor and ultrawide-field imaging. Automatic segmentation was instead implemented in [59,60]. In the first study, a deeply supervised 3D V-Net was used to automatically segment the arteriovenous malformations volume on CT images, demonstrating its clinical feasibility by validating the shape, positional accuracy, and dose coverage of

the automatic volume. In the second study, a DL approach based on a holistically-nested network reliably segmented the lung across the breathing cycle to accurately analyze the lung and respiratory muscle movement in Duchenne muscular dystrophy. This is a severe form of childhood muscular dystrophy that affects 1 in 5000 boys, characterized by progressive muscle degeneration caused by alterations in a protein that helps to keep muscle cells intact. In [61], authors constructed an ANN diagnostic model capable of differentiating primary immune thrombocytopenic purpura (pITP) patients and established a potential pITP diagnosis platform. pITP is defined as isolated autoimmune thrombocytopenia with idiopathic low platelet count, normal bone marrow, and unexplained causes of thrombocytopenia. In a recent study described in [62], authors studied multiple osteochondromas, an autosomal dominant disease characterized by the formation of osteochondromas or exostoses. The aim of this study was to create an efficient system based on a switching neural networks approach to characterize multiple osteochondromas patients in three classes, according to the number of bone segments affected, the presence of skeletal deformities, and functional limitations. Finally, due to the urgent need for biomarkers for the early detection of neurodevelopmental spectrum disorders, in [63], authors applied a trained neural network, ConvNetACh, with heart rate variation data of Rett syndrome patients, capable of distinguishing them from subjects showing typical development.

Ensemble learning (EL) can help to improve the accuracy of rare disease diagnosis by combining the predictions of multiple models and leveraging the strengths of each individual model. This can be particularly useful in the context of rare diseases, where the number of cases is limited and the diagnostic criteria can be complex. Pulmonary arterial hypertension (PAH) is a rare but progressive cardiopulmonary disease that leads to heart failure and premature death. MicroRNAs are small, non-coding molecules of RNA, previously shown to be dysregulated in PAH, and contribute to the disease process in animal models. In [64], EL techniques were used to select miRNAs able to distinguish PAH and healthy controls. These circulating miRNAs and their target genes may provide insight into PAH pathogenesis and reveal novel regulators of disease and putative drug targets. Primary sclerosing cholangitis (PSC) is a rare, chronic, cholestatic liver disorder characterized by inflammation and fibrosis in the bile ducts, and it is known for its frequent concurrence with inflammatory bowel disease. Dysbiosis of the gut microbiota in PSC was reported in several studies, but the microbiological features of the salivary microbiota in PSC have not been established. In [65], Iwasawa et al. implemented a random forest (RF) algorithm able to distinguish the salivary microbial communities of PSC patients, ulcerative colitis patients, and healthy controls, indicating the potential of salivary microbiota as biomarkers for the non-invasive diagnosis of PSC. In [66], an ML method based on RF was developed to automatically detect the early deterioration of photoreceptor integrity caused by inherited retinal degenerative diseases. An application example is choroideremia, which is an X-linked chorioretinal dystrophy characterized by progressive degeneration of the choroid. This tool can be used for choroidal flow assessment in order to provide a more comprehensive description of disease progression. Finally, authors in [67] used RF methodology in patients with three groups of rare myopathic conditions, which includes any disease that affects the muscles that control voluntary movement, showing that the methodology was able to classify myotonic dystrophy type 1 and inflammatory myopathy. Table 1 summaries the described above ML methods applied for the diagnosis of rare diseases.

Table 1. Summary of studies investigating ML methods applied for the diagnosis of rare diseases.

Disease	Methods	Data Type	Sample Size	Model Performance	References
Huntington disease	ANN	footstep pressure data	180	ACC: 89%	[42]
	SVM	gait data	15 (post stroke patients), 17 (HD patients), 10 (Controls)	ACC: 90.5%	[43]
	SVM	eye tracking data	22 (Controls), 14 (pre-HD patients), 14 (HD patients)	ACC: 73.47%, TPR: 74.31%, TNR: 72.64% pre-HD vs. Controls ACC: 81.84%, TPR: 76.19%, TNR: 87.48% HD vs. Controls ACC: 83.54%, TPR: 92.62%, TNR: 74.45 % pre-HD vs. HD	[44]
	BM	quantitative electroencephalography data	26 (preHD patients), 25 (Controls)	TNR: 83%, TPR: 83%, ACC: 83%.	[45]
Multiple system atrophy	SVM	MRI data	30 (MSA patients), 62 (PD patients), 59 (Controls)	ACC: 77.17%	[46]
	SVM	MRI data	31 MSA patients, 65 PD patients, 54 (Controls)	ACC: 78%	[47]
	ANN	MRI data	125 (PD patients), 98 (PSP patients), 54 (MSA-P patients), 142 (Controls)	PD ACC: 96.8%, AUC: 0.995 PSP ACC: 93.7%, AUC: 0.982 MSA-P ACC: 95.2%, AUC: 0.990 Controls ACC: 98.4%, AUC: 1.000	[48]
Amyotrophic lateral sclerosis	ANN	whole-genome data	4511 (ALS patients), 7397 (Controls)	ACC: 77%	[49]
	RF	brain MRI data	24 (ALS patients), 24 (Controls)	ACC: 80%	[50]
	SVM	fMRI data	32 (ALS patients), 31 (Controls)	ACC: 71%	[51]
Amyloidosis	SVM	proteomics data	75 (amyloid positive), 78 (Controls)	ACC: 96–99%	[52]
Hypophosphatasia	SVM	clinical data	23 (HPP patients), 22 (Controls)	ACC: 90%, TPR: 87%, TNR: 93%, AUC: 0.936	[53]
Friedreich ataxia	SVM	kinematic data	30 (Friedreich patients), 14 (Controls)	ACC: 91%, TPR: 90%, TNR: 93%, AUC: 0.91	[54]
Progressive supranuclear palsy	SVM	MRI data	28 (PSP patients), 28 (PD patients), 28 (Controls)	ACC: 85.8%, TPR: 86%, TNR: 86% PD vs. Controls ACC: 89.1%, TPR: 89.5%, TNR: 89.1% PSP vs. Controls PSP vs. PD 84.7 87.5 83.8	[55]
Osteogenesis imperfecta	SVM	μCT images	21 (specimens of 13 OI patients), 19 (specimens from 15 Controls)	AUC: 96%	[56]
Mucormycosis disease	ANN	eye photographs	Not clear	ACC: 99.5%	[57]
Retinitis pigmentosa	ANN	ultrawide-field images	150 (RP patients), 223 (Controls)	AUC: 0.998, TPR: 99.3%, TNR: 99.1% of the ultrawide-field pseudocolor group AUC: 1.000, TPR: 100%, TNR: 99.5% of the ultrawide-field autofluorescence	[58]
Cerebral arteriovenous malformation	ANN	Brain CT images	80	DSC: 85.2% TPR: 88% TNR: 99%	[59]
Duchenne muscular dystrophy	ANN	cine MRI data	15 (Duchenne patients), 16 (Controls)	DSC: 97.2 for the sagittal view DSC: 96.1 for the axial view DSC: 96.6 for the coronal view	[60]

Table 1. *Cont.*

Disease	Methods	Data Type	Sample Size	Model Performance	References
Immune thrombocytopenic purpura	ANN	proteomics data	64 (pITP patients), 70 (sITP patients), 82 (Controls)	TPR: 87.5%, TNR 69.7%, ACC: 75.0%	[61]
Multiple osteochondromas	ANN	clinical data	96 (class I), 137 (class II), 56 (class III)	Class I ACC: 94% Class II ACC: 80% Class III ACC: 79%	[62]
Rett syndrome	ANN	heart rate variation data	35 (Rett patients), 40 (Controls)	ACC: 88%	[63]
Pulmonary arterial hypertension	EL	microRNA expression data	64 (PAH patients), 43 (Controls)	TPR: 91%, TNR: 64%, ACC: 81%, AUC: 0.85	[64]
Primary sclerosing cholangitis	RF	bacterial 16S rRNA gene sequence data	24 (PSC patients), 16 (UC patients), 24 (Controls)	PSC AUC: 0.7423 UC AUC: 0.8756	[65]
Choroideremia	RF	OCT/OCTA data	16 (eyes with choroideremia), 5 (Controls eyes)	J: 0.876 ± 0.066	[66]
Dermatomyositis	RF	sonographic muscle images	11 (IBM patients), 19 (DM1 patients), 21 (PM-DM patients)	ACC: 78.4%	[67]

Performance Metrics Abbreviations: true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), mean reciprocal rank (MRR), Relative Absolute Error (RAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Concordance index (C-index), area under the curve (AUC), Coefficient of determination (R^2), Confidence interval (CI), Sørensen–Dice coefficient (DSC), Jaccard similarity index (J).

3.2. Prognosis

The prognosis includes information about the likely or expected evolution, duration, and outcome of the condition. In most cases, the possibility of a cure is also mentioned; however, most rare conditions are chronic and lifelong, so the goal is to manage the condition rather than cure it [6]. The difficulty in making a predictive prognosis not only affects the physical health of the patient, but also their mental health, leading to stress, anxiety, and depression [68].

AI can play a significant role in the prognosis of rare disorders by helping to fill in the gaps in data and experience [69]. By analyzing large amounts of data, such as electronic health records, genomic data, and imaging studies, ML algorithms can identify patterns and predict outcomes for individuals with rare diseases, providing valuable insights that can inform prognoses and guide decisions [70]. Additionally, AI can be used to develop new prognostic tools, such as risk prediction models, which can identify potential factors and early warning signs of disease progression, allowing for early intervention and potentially improving patient outcomes [71].

The commonly used AI approaches in the prognosis phase are supervised learning with EL, ANN, and SVM as the most widely used methods. Unsupervised methods, such as clustering, are used less frequently.

Two recent studies [72,73] used ML to identify new biomarkers that could be employed for prognostic purposes for adrenocortical carcinoma (ACC), a rare and aggressive cancer that arises from the cells of the outer layer of the adrenal gland. The prognosis for ACC is generally poor, with a 5-year survival rate of only about 10–20%, so early detection is crucial for improving the chances of survival, as well as identifying new markers. In [72], the authors applied a simple and unsupervised ML method called uniform manifold approximation and projection (UMAP) to mRNA expression data from the TCGA-ACC study, the largest multi-platform study of ACC. UMAP is a dimension reduction technique, and it found two distinct clusters that strongly correlated with patient prognosis. They then used an RF algorithm to identify the transcriptional differences between the two clusters, finding 100 genes that could serve as new biomarkers or novel targets for treatment. In [73], the authors performed a proteomic analysis of ACC at different stages and identified 7000 individual proteins. They selected 117 differentially expressed proteins (DEPs) using three feature selection algorithms (ReliefF, infoGain, and ANOVA) and conducted a survival analysis to assess the effect of the identified DEPs on patient survival. They were able to identify five new candidate protein biomarkers as prognostic factors, which can help in defining new therapeutic targets. Both studies highlight the importance of using ML with multi-omic data to better understand the biology of ACCs and to identify biomarkers for the disease.

The study of alkaptonuria is an example of how multiple ML techniques have been applied to an ultra-rare disease. Alkaptonuria (AKU) is an autosomal, recessive, and metabolic disorder caused by a defect in the enzyme homogentisic acid oxidase. As a result, homogentisic acid accumulates in the body and causes the formation of ochronotic pigments, and this can lead to various symptoms such as arthritis, amyloidosis, and kidney stones. Due to the rarity of the disease and the lack of a standardized method of assessment, studying AKU can be challenging. A recent study [74] has implemented a digital platform, ApreciseKUre, which is designed to collect, integrate, and analyze data for patients with AKU. The platform includes a wide range of data, including genetic, biochemical, histopathological, clinical, therapeutic resources, and quality of life (QoL) scores, which can be shared among researchers and clinicians to create a precision medicine ecosystem. The authors describe how ML applications were used to analyze and interpret the data in ApreciseKUre to achieve patient stratification, and tailor care and treatment to specific subgroups of patients. Two specific studies show the potential of ML in the context of AKU data. The first study [75] aimed to predict QoL scores based on patient's clinical data using the XGBoost algorithm and a k-NN algorithm. The second study [76] aimed to compare different algorithms (K-means and hierarchical clustering) to explore phenotype-

genotype relationships that were previously unknown in AKU. Both studies showed that ML successfully predicted clinical outcomes and QoL scores, and also identified new biomarkers and subgroups of patients with AKU. These studies highlight the need for the development of databases for rare diseases, helping to optimize the benefit-risk ratio, and improving overall patient outcomes.

ALS is another rare and very serious disease that has been studied with AI methods. In [77], the authors used pharmacometabolomics approaches and ML algorithms to identify metabolic changes in patients with ALS and the effects of two different treatments: riluzole and olesoxime. They applied multivariate statistical techniques such as partial least squares regression, orthogonal partial least squares discriminant analysis, and a novel algorithm called Biosigner. This algorithm, which is based on bootstrapping and different methods like RF and SVM, was found to have better predictive power than other approaches. The study found that certain lipids and amino acids were differentially expressed in the two treatment groups, and that these changes might be linked to changes in energy metabolism and glutamate metabolism, which are known to be important in ALS pathophysiology. In [78], Huang et al. present a novel non-parametric survival analysis method called GuanRank that aims to improve the reliability and robustness of survival predictions in clinical trials. This method is based on the Kaplan-Meier estimator and transforms the problem into a general regression problem that can be solved by ML regression algorithms such as Lasso regression, Gaussian process regression, and RF. The method was validated on the PRO-ACT database, a large de-identified dataset of patients in ALS clinical trials, and it demonstrated superior performance over the traditional survival models such as the Cox proportional hazard model. Gordon & Lerner [79] also used data from the PRO-ACT database to predict the state of ALS patients. They used RF, XGBoost, cumulative link models, ordinal decision trees, and cumulative probability trees as the prediction models and BM for knowledge representation. They found that ordinal classification models improved predictive performance and identified variables that were not previously known to be related to ALS, such as creatinine, CK, and phosphorus. In addition, data related to language and MRI images of ALS patients can be used to better understand the progression of the disease. Wang et al. [80] aimed to develop an automated assessment tool for speech impairment in ALS to improve the early detection and monitoring of bulbar dysfunction in ALS patients. They proposed the use of ML to detect abnormal speech patterns in ALS from both acoustic and articulatory samples and to help in the assessment of disease progression. The speech data is in the form of features extracted from speech recordings, which can be done using open-source algorithms such as openSMILE. Gradient boosting was used as the feature selection technique and SVM was used to predict intelligible speaking rate from speech acoustic and articulatory samples. In [81], the authors aimed to use DL to predict the survival time of ALS patients based on clinical characteristics and advanced MRI metrics. They collected high-resolution diffusion-weighted and T1-weighted images from 135 ALS patients at their first visit, and then monitored each patient's survival time until death. Then, they used DL to create four different networks: one based on clinical data, one based on structural connectivity MRI data, one based on morphology MRI data, and one based on a combination of the three sources of information. The results showed that MRI data alone can provide valuable predictions of survival time and that combining clinical characteristics and MRI data into a DL approach can further improve predictions about a patient's survival time. These studies on ALS highlight the importance of combining multiple sources of data such as clinical characteristics and MRI metrics to improve the accuracy of predictions.

As already seen for diagnosis, AI can be of great help in the prognostic phase of HD as well. Lauraitis et al. [82] proposed a hybrid model that uses artificial ANN and a Fuzzy Logic expert system (FLS) to predict, through finger-tapping tests, the deterioration of reaction state in individuals with neurological movement disorders such as hand tremors and non-voluntary movements. This model is composed of four sub-models (dataset formation, ANN prediction, FLS, and a decision module for determining the person's con-

dition) and was tested on a dataset of 3032 records from 20 test subjects. Results show that the feed-forward backpropagation neural network model achieved the best performance results. The authors plan to validate the proposed system using a larger dataset including data from PD and Alzheimer's patients, as well as using more sophisticated finger-tapping features and comparing ANN results with those of SVM regression. In [83], they present a new approach that uses a combination of brain function and structure imaging data to identify whether a person with HD will receive a clinical diagnosis within 5 years, known as premanifest HD (preHD). The researchers used an SVM to classify individuals with preHD from controls. The input data were resting-state functional connectivity, subcortical gray matter volume, and cortical thickness. The SVM was trained using a linear kernel and a weighted cost function to account for class imbalances, and then the models were evaluated using leave-one-out validation and permutation testing. They also applied independent validation to test the generalizability of the findings. Asadi et al. [84] also wanted to predict the progression of a disease, i.e., cerebral arteriovenous malformations (cAVMs). They noticed that the lack of large observational studies on the long-term outcome of unruptured cAVMs has made it difficult to determine the best course of action. cAVMs are rare, abnormal connections between the arteries and veins in the brain that typically form before birth. They can vary in size and location, and may cause a rupture, leading to hemorrhage and reduced blood flow to the brain. Since cAVMs can present symptoms at any age, the goal is to identify factors that can be used to predict hemorrhagic risk and to develop a risk stratification model that can be used to guide treatment decisions. They used ANN and SVM to predict the outcome of cAVMs post-endovascular treatment with relatively high accuracy and precision. The ANN was found to be the strongest predictor of fatal outcome, with the presence or absence of nidal fistulae having the greatest predictive power. The study also found out that the classical regression model had mediocre accuracy in predicting the outcome of mortality, with the type of treatment-related complication being the most important predictor. In [85], the authors developed an ML algorithm based on DTI to predict the clinical severity of PSP. The algorithm was trained on data from a cohort of PSP patients and was found to be accurate in predicting the severity of the disease as measured by various clinical scales. Moreover, the algorithm identified regions of the brain related to motor function, such as the thalamus, and regions related to psychomotor interactions, such as the parahippocampus gyrus, that are associated with the severity of the disease.

Other examples of where SVMs have been successfully applied include the works of Zhutovsky et al. [86] and An et al. [87]. In [86], they wanted to determine the prognostic accuracy of clinical and structural MRI data of patients with a behavioral variant of frontotemporal dementia (bvFTD) presenting late-onset behavioral changes. This disorder presents with behavioral and cognitive symptoms that overlap with other neurological and psychiatric disorders, so the authors suggest that predictive biomarkers could facilitate early detection. They used data from 73 patients, divided into three groups based on 2-year follow-up diagnosis: probable/definite bvFTD, neurological, and psychiatric. They then used SVM classifiers to perform classification tasks and evaluated performance using cross-validation. They found that the combination of clinical and voxel-wise whole brain data showed the best performance overall, and concluded that the results show the potential for automated early confirmation of bvFTD using ML analysis of clinical and neuroimaging data in a diverse and clinically relevant sample of patients. In [87], the authors used the SVM model to study mutations that cause Diamond-Blackfan anemia (DBA), a rare hereditary disorder characterized by failure of erythropoiesis. They first conducted a comprehensive study on the structural basis of human RPS19 mutations that occur in DBA, based on its 3D structures, and then used this knowledge to train an SVM model to predict the pathogenicity of all possible missense mutations of RPS19. They used 29 DBA mutations (positive samples) and 30 neutral ones (negative samples) as training data, and extracted 8 features to be used for each mutation, such as interaction with rRNA, structural stability, and conservation. After five-fold cross-validation, the best hyperparameters were

identified and the SVM model was able to predict 26 of the 29 DBA mutations correctly, with a significantly reduced false-positive rate compared to other prediction tools.

As mentioned, the most widely used type of AI algorithms for the prognosis of rare diseases are EL algorithms. Here, we give some examples of where these have been successfully applied. In [88], they used a bootstrap aggregation (bagging) ensemble technique with a reduced-error pruning regression tree as the underlying classifier to predict the energy expenditure (EE) of children with DMD. Existing ML algorithms developed for healthy populations do not accurately estimate EE for children with DMD, so they develop a custom ML model specifically for this population. Bagging is an ensemble meta-algorithm that improves the stability and accuracy of statistical regression obtained by a regression tree. They demonstrate that this technique has proven to be superior to other models such as multilayer perceptron, SVM, linear regression, naive Bayes, and reduced-error pruning regression tree. Another example of boosting algorithm application is [89], where the authors present a novel ML model named the PSC risk estimate tool (PREsTo) that predicts the 5-year risk of hepatic decompensation in patients with PSC. Due to the rarity of the disease, improved biomarkers are necessary to risk-stratify patients in clinical trials and serve as surrogate endpoints. The PREsTo model used gradient boosting machines, a step-wise method to create an ensemble of weak prediction models, typically decision trees. Each decision tree may have different variables, and the ones with the strongest predictive power are used more often and earlier in the model-building process. The study reported that the model performed well compared to the existing prognostic markers, had an excellent performance when applied to a later point in the disease course, and had good performance among various PSC subgroups. The advantages of this model are that it uses readily available clinical data, and it is non-invasive, inexpensive, and accurate. In [90], Robinson et al. used immune cell frequency profiles, clinical, and serological data from patients with juvenile-onset systemic lupus erythematosus (jSLE) to identify predictive disease outcome signatures using RF and sparse partial least squares-discriminant analysis (sPLS-DA).

BRF was used to overcome difficulties in obtaining validation datasets because it does not overfit to training data, and it was used to further define and validate the pathological immune cell profile of the disease. sPLS-DA was used as a secondary validation method to rank and validate the immunological variables by their distribution in patients with jSLE and healthy controls. The analyses identified 8 immune cell subtypes that were consistently correlated with jSLE patients, compared with healthy controls. Lastly, in the works of Chou & Ghimire [91,92], they applied RF algorithms to identify prognostic factors in pediatric myocarditis patients. In their first study [91], they used an RF algorithm on 500 factors from a publicly available pediatric hospitalization database (Kids' Inpatient Database) to identify mortality risk factors, and validated these factors using linear and binomial regression models. They also used negative binomial regression models to study the association between the length of hospitalization and risk factors. The goal of the second study [92] was to develop a model to predict in-hospital mortality among patients hospitalized for pediatric myocarditis, since traditional logistic regression models have low sensitivity. A total of 14 variables were included in model development and an RF algorithm was applied because of the nature of the predictors, which are all two-level categorical variables. Based on the importance scores of the risk factors, the top 5 variables were selected as MV, ECMO use, cardiac arrest, ventricular fibrillation, and AKI. Table 2 summaries the described above ML methods applied for the prognosis of rare diseases.

Table 2. Summary of studies investigating ML methods applied for the prognosis of rare diseases.

Disease	Methods	Data Type	Sample Size	Model Performance	References
Adrenocortical carcinoma	clustering, RF DT	mRNA expression data differentially expressed proteins	79 117	ACC: 96%	[72] [73]
Alkaptonuria	XGBoost, k-NN custering	genetic, biochemical, histopathological, clinical, therapeutic resources, and QoL scores	129 112	RAE: 0.25, R2: 0.94	[75] [76]
Amyotrophic lateral sclerosis	RF, SVM	metabolomics data	38 (treated), 36 (placebo)	TPR: 71.4%, TNR: 71.4%, PPV: 71.4%, NPV: 70.0%	[77]
	LASSO, RF	a subset of the PRO-ACT dataset (survival and clinical data)	6565	C-index: 0.7355	[78]
	RF, XGBoost, BM, DT GBoost, SVM ANN	a subset of the PRO-ACT dataset speech acoustic and articulatory data Clinical and MRI data	3772 1832 135	ACC: 71–84.7% R2: 0.712, RMSE: 37.562 ACC: 84.4%.	[79] [80] [81]
Huntington disease	ANN, FLS SVM, EL	finger-tapping tests data Neuroimaging data	3032 19 (preHD), 21(Controls)	R2: 0.98, MSE: 0.08 F1: 74%	[82] [83]
Cerebral arteriovenous malformation	ANN, SVM, Log Reg	clinical and imaging data	199	ACC: 97.5%	[84]
Progressive supranuclear palsy	LASSO, Lin Reg	Imaging data	53	R2: 0.892	[85]
Behavioral variant of frontotemporal dementia	SVM	clinical and structural MRI data	73	ACC: 72–82%, TPR: 67–79%, TNR; 77–88%, AUC: 0.80–0.9	[86]
Diamond-Blackfan anemia	SVM	structural data of missense mutation	29 (positive samples), 30 (negative samples)	ACC: 95%, TPR: 90%, TNR; 98% F1: 94%	[87]
Duchenne muscular dystrophy	EL	inertial sensor (accelerometer) data	7	RMSE: 0.017	[88]
Primary sclerosing cholangitis	GBoost	clinical and laboratory data	509	C-index: 0.90	[89]
Juvenile-onset systemic lupus erythematosus	RF	Immunophenotyping data	67 (jSLE), 39 (Controls)	ACC: 86·8%	[90]
Pediatric myocarditis	RF RF, Log Reg	Diagnoses/procedures data from the Kids' Inpatient Database	7241 4144	CI: 95% TPR: 89.9% TNR: 85.8%, ACC: 87.9%	[91] [92]

3.3. Treatment

There is an urgent need to identify novel treatment options for rare diseases, which is a difficult challenge due to the lack of essential data including drug molecules, genes, and protein structure information. The speed at which new biomedical knowledge is being discovered makes it particularly challenging to connect disease mechanisms to drug action. Almost 95% of rare diseases do not have FDA-approved drug treatment and the increasing number of rare diagnoses puts pressure on scientists and clinicians to characterize these conditions and match patients with appropriate treatments [93]. As biomedical discoveries continue to generate big amounts of data, an opportunity emerges for AI to help in translating biomedical knowledge into a format that can be used to identify therapeutic strategies for patients. Recently, The Hugh Kaul Precision Medicine Institute created mediKanren [94], an AI platform based on knowledge graphs that uses the mechanistic insight of genetic disorders to identify therapeutic strategies, enabling an efficient way to link all relevant literature and databases. The method was tested by analyzing genetic data and publications of two rare disorders related to missense variants in the TMLHE and RHOBTB2 genes, revealing molecular mechanisms and pathways which have provided new therapeutic targets.

Currently, AI methods for treatment belong mostly to supervised learning, which uses labeled datasets to train algorithms able to classify or predict outcomes accurately. In [95], Bakkar et al. implemented the IBM Watson® [96] method to screen RNA-binding proteins (RBPs) in the genome and identify additional RBPs involved in ALS. Numerous RBPs have been shown to be altered in ALS, making them a contributing factor in disease pathobiology. IBM Watson extracts domain-specific text features from published literature to identify new connections between entities of interest. From these annotated documents, Watson created a semantic model of the set of RBPs with known mutations that cause ALS, and then applied that model to a candidate set of all other RBPs to cluster all the candidates by similarity to the known set using a graph diffusion algorithm. Gated Recurrent Unit Cooperation-Attention-Network (GCAN) was used in [97] to predict drugs for rare diseases, with particular attention to Gaucher disease, a rare metabolic disorder in which deficiency of the enzyme glucocerebrosidase results in the accumulation of toxic quantities of certain lipids. Two heterogeneous networks were built for information enhancement; one network contains the father nodes of the rare disease, while the other network contains information on the son nodes. A biased random walk approach was used to collect data from the father and son nodes, where nodes were linked in a hierarchical relationship with two hop distances. The effectiveness of two Gaucher disease drugs predicted by GCAN has been established. In [98], authors showed interest in sialidosis, an ultra-rare lysosomal storage disease characterized by an excessive accumulation of glycoprotein-derived oligosaccharides. J. Klein et al. applied the so-called Assay Central software [99] to build Bayesian ML models to screen compounds in silico before in vitro testing. This approach has been applied to identify new compounds that can act as a potential disease modulator in the treatment of sialidosis. In [100], the authors used an RF classifier for the prediction of cell-penetrating peptides, which can facilitate the intracellular delivery of large therapeutically-relevant molecules. The goal was to deliver phosphorodiamidate morpholino oligonucleotides, a type of antisense therapy recently approved by the FDA for the treatment of DMD. Multi-output regression ML methodologies were implemented in [101] to predict the potential effect of external proteins on the signaling circuits that trigger Fanconi anemia-related cell functionalities. This rare condition causes genomic instability and a range of clinical features, including developmental abnormalities in major organ systems and a high predisposition to cancer [102]. Thanks to these models, over 20 potential therapeutic targets were detected. In the last study [103], Spiga et al. developed an RF model that performs a prediction of the QoL scores based on data deposited in ApreciseKUre. Predicted QoL scores were then correlated with the drugs taken by AKU patients, revealing that drugs typically used to treat AKU patients were effective in reducing pain, but some common drugs not related to specific AKU symptoms also showed a correlation

with some QoL scores. Table 3 summaries the described above ML methods applied for the treatment of rare diseases.

Table 3. Summary of studies investigating ML methods applied for the treatment of rare diseases.

Disease	Methods	Data Type	Sample Size	Model Performance	References
Amyotrophic lateral sclerosis	IBM Watson®	RNA-binding proteins	1478	AUC: 0.935	[95]
Multiple diseases, exp. Gaucher disease	ANN	disease, gene, and drug data	7000	Hits@10: 0.454, MRR 0.231	[97]
Sialidosis	BM	disease targets and accessible bioactivity data	57	AUC: 0.737, PPV: 34.5%, TPR: 91%, TNR: 58.7%, F1: 50%	[98]
Duchenne muscular dystrophy	RF	PMO activity data	64	ACC:72%, PPV:75%, TPR:69%	[100]
Fanconi anemia	EL	gene expression data	>11,000	R2: 0.62–0.97	[101]
Alkaptonuria	RF	genetic, biochemical, histopathological, clinical, therapeutic resources, and QoL scores	129	ACC: 70%	[103]

4. Discussion

In this review, we studied the scientific literature of the last decade to understand which AI methods are most widely used in the field of rare and ultra-rare diseases, revealing that the most applied algorithms are SVM, RF, and ANN. One of the reasons these methods are so popular is their ability to handle complex, high-dimensional data, and indeed images were found to be the most widely used type of input. Medical imaging techniques, such as PET, CET, MRI, and ultrasound, produce large amounts of this type of data, which can be standardized and therefore easily processed by AI methods. Moreover, these methods can also learn from small datasets, as is often the case in rare diseases due to their low prevalence, and are able to identify important features, which can help researchers better understand the mechanisms underlying the disease and potentially identify new targets for treatment. Another point to highlight is that most studies have focused on using AI methods for the diagnosis and prognosis of rare diseases, which is a typical application of classification and prediction, and this is also understandable given the challenges associated with identifying and predicting rare disorders. However, it is worth noting that ML could also play a significant role in improving the treatment, accelerating drug development by identifying potential drug candidates, predicting their efficacy and optimizing their dosages, as well as identifying drugs already on the market and repurposing them for other diseases.

However, the application of ML models can also present various challenges, starting with the difficulty of applying them to unstructured and poorly standardized textual data, such as medical records. This is because the presence of label-noise and sparsity can lead to model overfitting, meaning that the method has high prediction accuracy on training data and low prediction accuracy on new evaluation data. In addition, the aforementioned most widely used methods focus on classification tasks while few algorithms are applied to the study of the biological mechanisms of these diseases. This is probably due to the complexity of the systems biology that are these disorders, which sometimes have molecular bases that are still unknown, or a pathological picture that is unclear and overlaps with that of other diseases. This implies that there is a scarcity of already developed and validated methods that can tolerate these constraints, but there is also a high demand for AI approaches from the biomedical world. It is therefore essential to intentionally develop methods and analytic workflows that can address these challenges, and it is up to the developers of such methods to ensure that the resulting models are reliable. Thus, to increase confidence in the end user (i.e., clinicians and researchers), explanations of the behavior of the developed algorithms must be provided, and robust error analysis must be conducted.

Another issue that needs to be addressed is that only few studies have validated their models on external datasets, thus assessing their potential in clinical practice. External validation studies can help assess the generalizability and reliability of the algorithm and determine its potential use in clinical settings, but unfortunately only a few of the reported

studies have conducted this type of validation. This is partly due to the difficulty in this field of obtaining large and diverse datasets and the lack of standardized methods for data collection and analysis, but without adequate validation, there is a risk that the algorithms will produce unreliable or inaccurate results when applied to new datasets or patient cohorts. Solving this problem requires collaboration among researchers, clinicians, and data scientists to develop standardized methods, as well as sharing of data and algorithms to facilitate external validation studies.

5. Conclusions

ML methods have shown promise in the identification and diagnosis of rare diseases. With the vast amounts of data now available through electronic health records and heterogeneous databases, AI algorithms can help quickly identify patterns and associations that would be difficult or impossible for human analysts to detect. For example, researchers use ML algorithms to analyze patient data and identify characteristic patterns associated with certain rare diseases. By doing so, these tools can help to narrow down the list of possible diagnoses, making it more likely that patients will receive a correct diagnosis in a timely manner.

Predictive modeling techniques, such as DL, have been used to forecast the progression of rare diseases, allowing for earlier interventions and better treatment planning. This could potentially lead to a more accurate classification of rare diseases and enable the development of more targeted treatments. From a precision medicine perspective, by identifying biomarkers associated with a particular rare disease, AI algorithms can help to develop personalized treatment plans, helping to improve patient outcomes and reduce the risk of side effects.

AI has also shown promise in the field of drug development for rare diseases. AI algorithms can be used to analyze patient data and identify subpopulations of patients who may be most likely to respond to a particular drug. This can help to make clinical trials more efficient and increase the chances of a drug being approved for use in patients with rare disorders. Despite the potential benefits of ML, there are still challenges that must be overcome to fully realize its potential in the field of rare diseases. These include a lack of large, well-annotated datasets, and the need for interpretable models that can be easily understood and trusted by clinicians. Rare diseases usually present an unusually big data regime, which is characterized by huge omics data but a limited number of patients. The rarity of orphan patients, despite the presence of registries, still has a large impact on ML analyses, and thus open data can contribute significantly to support the modeling attempt. In the future, patient registries and open data need to be integrated, translating the largest amounts of data available into potential connections. Finally, a current limitation is the lack of interpretability, which makes it difficult for clinicians and researchers to understand algorithm outputs. In fact, explainability is one of the most debated topics for the application of AI in healthcare. While AI-based systems have been shown to outperform humans in certain analytical tasks, the lack of explainability continues to attract criticism. In this context, explainable AI approaches are the new frontier of ML applications in healthcare, in order to ensure the understanding, by both clinicians and patients, of the "mental process" followed by the artificial brain to reach a certain decision.

In conclusion, the application of ML techniques can greatly assist rare disease research and treatment, but to use it effectively, it needs to be implemented under the right ethical principles, avoid biases, and also be transparent for the patient. To fully tap into its potential, AI needs to be validated through clinical trials and real-world evidence. Furthermore, it needs to be accompanied by regulatory frameworks that ensure the safety and reliability of AI-based medical devices and diagnostic tools. More work is needed to overcome data-related challenges, ensure fair and trustworthy models, and help translate the research findings into practical applications that can benefit patients and their families.

References

1. Vickers, P.J. Challenges and Opportunities in The Treatment of Rare Diseases. *Drug Discov. World* **2013**, *14*, 9–16.
2. Hughes, D.; Tunnage, B.; Yeo, S. Drugs for exceptionally rare diseases: Do they deserve special status for funding? *QJM Int. J. Med.* **2005**, *98*, 829–836. [CrossRef] [PubMed]
3. Nguengang Wakap, S.; Lambert, D.M.; Olry, A.; Rodwell, C.; Gueydan, C.; Lanneau, V.; Murphy, D.; Le Cam, Y.; Rath, A. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *Eur. J. Hum. Genet.* **2020**, *28*, 165–173. [CrossRef] [PubMed]
4. European Union. Regulation (EC) N°141/2000 of the European Parliament and of the Council of 16 December 1999 on Orphan Medicinal Products. 2000. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000R0141 (accessed on 18 November 2022).
5. National Institute of Health. Public Law 97–414 97th Congress. 1983. Available online: https://www.govinfo.gov/content/pkg/STATUTE-96/pdf/STATUTE-96-Pg2049.pdf (accessed on 14 December 2022).
6. Richter, T.; Nestler-Parr, S.; Babela, R.; Khan, Z.M.; Tesoro, T.; Molsen, E.; Hughes, D.A. Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value Health* **2015**, *18*, 906–914. [CrossRef] [PubMed]
7. Stoller, J.K. The Challenge of Rare Diseases. *Chest* **2018**, *153*, 1309–1314. [CrossRef]
8. Navarrete-Opazo, A.A.; Singh, M.; Tisdale, A.; Cutillo, C.M.; Garrison, S.R. Can you hear us now? The impact of health-care utilization by rare disease patients in the United States. *Anesth. Analg.* **2021**, *23*, 2194–2201. [CrossRef]
9. Willmen, T.; Völkel, L.; Ronicke, S.; Hirsch, M.C.; Kaufeld, J.; Rychlik, R.P.; Wagner, A.D. Health economic benefits through the use of diagnostic support systems and expert knowledge. *BMC Health Serv. Res.* **2021**, *21*, 947. [CrossRef]
10. NORD. Barriers to Rare Disease Diagnosis, Care and Treatment in the US: A 30-Year Comparative Analysis. 2020. Available online: https://rarediseases.org/wp-content/uploads/2020/11/NRD-2088-Barriers-30-Yr-Survey-Report_FNL-2.pdf (accessed on 18 November 2022).
11. Mueller, T.; Jerrentrup, A.; Bauer, M.J.; Fritsch, H.W.; Schaefer, J.R. Characteristics of patients contacting a center for undiagnosed and rare diseases. *Orphanet J. Rare Dis.* **2016**, *11*, 81. [CrossRef]
12. Tambuyzer, E.; Vandendriessche, B.; Austin, C.P.; Brooks, P.J.; Larsson, K.; Needleman, K.I.M.; Valentine, J.; Davies, K.; Groft, S.C.; Preti, R.; et al. Therapies for rare diseases: Therapeutic modalities, progress and challenges ahead. *Nat. Rev. Drug Discov.* **2019**, *19*, 93–111. [CrossRef]
13. Hurvitz, N.; Azmanov, H.; Kesler, A.; Ilan, Y. Establishing a second-generation artificial intelligence-based system for improving diagnosis, treatment, and monitoring of patients with rare diseases. *Eur. J. Hum. Genet.* **2021**, *29*, 1485–1490. [CrossRef]
14. Valdez, R.; Ouyang, L.; Bolen, J. Public Health and Rare Diseases: Oxymoron No More. *Prev. Chronic Dis.* **2016**, *13*, E05. [CrossRef] [PubMed]
15. Alves, V.M.; Korn, D.; Pervitsky, V.; Thieme, A.; Capuzzi, S.J.; Baker, N.; Chirkova, R.; Ekins, S.; Muratov, E.N.; Hickey, A.; et al. Knowledge-based approaches to drug discovery for rare diseases. *Drug Discov. Today* **2021**, *27*, 490–502. [CrossRef] [PubMed]
16. Baxter, K.; Horn, E.; Gal-Edd, N.; Zonno, K.; O'Leary, J.; Terry, P.F.; Terry, S.F. An End to the Myth: There Is No Drug Development Pipeline. *Sci. Transl. Med.* **2013**, *5*, 171cm1. [CrossRef] [PubMed]
17. Kakkis, E.D.; O'Donovan, M.; Cox, G.; Hayes, M.; Goodsaid, F.; Tandon, P.K.; Furlong, P.; Boynton, S.; Bozic, M.; Orfali, M.; et al. Recommendations for the development of rare disease drugs using the accelerated approval pathway and for qualifying biomarkers as primary endpoints. *Orphanet J. Rare Dis.* **2015**, *10*, 16. [CrossRef]
18. Pradhan, A.; Prabhu, S.; Chadaga, K.; Sengupta, S.; Nath, G. Supervised Learning Models for the Preliminary Detection of COVID-19 in Patients Using Demographic and Epidemiological Parameters. *Information* **2022**, *13*, 330. [CrossRef]
19. Khanna, V.V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Chadaga, R.; Umakanth, S. Diagnosing COVID-19 using artificial intelligence: A comprehensive review. *Netw. Model. Anal. Health Inform. Bioinform.* **2022**, *11*, 25. [CrossRef]
20. COVID-19 Mortality Prediction using Machine Learning: A Deep Forest Approach. In Proceedings of the 2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Shivamogga, India, 14–15 October 2022; pp. 245–250. [CrossRef]

21. Ilan, Y. Second-Generation Digital Health Platforms: Placing the Patient at the Center and Focusing on Clinical Outcomes. *Front. Digit. Health* **2020**, *2*, 569178. [CrossRef]
22. Sim, I.; Gorman, P.; Greenes, R.A.; Haynes, R.B.; Kaplan, B.; Lehmann, H.; Tang, P.C. Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *J. Am. Med. Inform. Assoc.* **2001**, *8*, 527–534. [CrossRef]
23. Porat, T.; Delaney, B.; Kostopoulou, O. The impact of a diagnostic decision support system on the consultation: Perceptions of GPs and patients. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 79. [CrossRef]
24. Winther, O.; Svenstrup, D.; Henningsen, P.P.; Kristiásson, R.; Jørgensen, H.L. FindZebra-the search engine for difficult medical cases. *Orphanet J. Rare Dis.* **2014**, *9*, O5. [CrossRef]
25. Girdea, M.; Dumitriu, S.; Fiume, M.; Bowdin, S.; Boycott, K.M.; Chénier, S.; Chitayat, D.; Faghfoury, H.; Meyn, M.S.; Ray, P.N.; et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Hum. Mutat.* **2013**, *34*, 1057–1065. [CrossRef] [PubMed]
26. Alves, R.; Piñol, M.; Vilaplana, J.; Teixidó, I.; Cruz, J.; Comas, J.; Vilaprinyo, E.; Sorribas, A.; Solsona, F. Computer-assisted initial diagnosis of rare diseases. *PeerJ* **2016**, *4*, e2211. [CrossRef]
27. Ronicke, S.; Hirsch, M.C.; Türk, E.; Larionov, K.; Tientcheu, D.; Wagner, A.D. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet J. Rare Dis.* **2019**, *14*, 69. [CrossRef]
28. Lanar, S.; Acquadro, C.; Seaton, J.; Savre, I.; Arnould, B. To what degree are orphan drugs patient-centered? A review of the current state of clinical research in rare diseases. *Orphanet J. Rare Dis.* **2020**, *15*, 134. [CrossRef]
29. Lee, J.; Liu, C.; Kim, J.; Chen, Z.; Sun, Y.; Rogers, J.R.; Chung, W.K.; Weng, C. Deep learning for rare disease: A scoping review. *J. Biomed. Inform.* **2022**, *135*, 104227. [CrossRef] [PubMed]
30. Schaefer, J.; Lehne, M.; Schepers, J.; Prasser, F.; Thun, S. The use of machine learning in rare diseases: A scoping review. *Orphanet J. Rare Dis.* **2020**, *15*, 145. [CrossRef] [PubMed]
31. Mcculloch, W.S.; Pitts, W.H. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
32. Cunningham, P.; Cord, M.; Delany, S.J. Supervised Learning. In *Machine Learning Techniques for Multimedia: Cognitive Technologies*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 21–49.
33. Ghahramani, Z. Unsupervised learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 72–112.
34. Zhu, X.J. Semi-Supervised Learning Literature Survey. 2005. Available online: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf (accessed on 18 November 2022).
35. Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education Limited, Prentice Hall Press: Upper Saddle River, NJ, USA, 2016.
36. Alzubi, J.; Nayyar, A.; Kumar, A. Machine Learning from Theory to Algorithms: An Overview. *J. Phys. Conf. Ser.* **2018**, *1142*, 012012. [CrossRef]
37. HamediRad, M.; Chao, R.; Weisberg, S.; Lian, J.; Sinha, S.; Zhao, H. Towards a fully automated algorithm driven platform for biosystems design. *Nat. Commun.* **2019**, *10*, 5150. [CrossRef]
38. Shen, F.; Liu, S.; Wang, Y.; Wang, L.; Afzal, N.; Liu, H. Leveraging Collaborative Filtering to Accelerate Rare Disease Diagnosis. *AMIA Annu. Symp. Proc.* **2018**, *2017*, 1554–1563.
39. Jia, J.; Wang, R.; An, Z.; Guo, Y.; Ni, X.; Shi, T. RDAD: A Machine Learning System to Support Phenotype-Based Rare Disease Diagnosis. *Front. Genet.* **2018**, *9*, 587. [CrossRef]
40. Rother, A.K.; Schwerk, N.; Brinkmann, F.; Klawonn, F.; Lechner, W.; Grigull, L. Diagnostic Support for Selected Paediatric Pulmonary Diseases Using Answer-Pattern Recognition in Questionnaires Based on Combined Data Mining Applications—A Monocentric Observational Pilot Study. *PLoS ONE* **2015**, *10*, e0135180. [CrossRef] [PubMed]
41. Li, X.; Wang, Y.; Wang, D.; Yuan, W.; Peng, D.; Mei, Q. Improving rare disease classification using imperfect knowledge graph. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 238. [CrossRef] [PubMed]
42. Zhang, S.; Poon, S.K.; Vuong, K.; Sneddon, A.; Loy, C.T. A Deep Learning-Based Approach for Gait Analysis in Huntington Disease. *Stud. Health Technol. Inform.* **2019**, *264*, 477–481. [CrossRef]
43. Mannini, A.; Trojaniello, D.; Cereatti, A.; Sabatini, A.M. A Machine Learning Framework for Gait Classification Using Inertial Sensors: Application to Elderly, Post-Stroke and Huntington's Disease Patients. *Sensors* **2016**, *16*, 134. [CrossRef] [PubMed]
44. Miranda, Â.; Lavrador, R.; Júlio, F.; Januário, C.; Castelo-Branco, M.; Caetano, G. Classification of Huntington's disease stage with support vector machines: A study on oculomotor performance. *Behav. Res. Methods* **2015**, *48*, 1667–1677. [CrossRef] [PubMed]
45. Odish, O.F.F.; Johnsen, K.; van Someren, P.; Roos, R.A.C.; van Dijk, J.G. EEG may serve as a biomarker in Huntington's disease using machine learning automatic classification. *Sci. Rep.* **2018**, *8*, 16090. [CrossRef] [PubMed]
46. Baggio, H.C.; Abos, A.; Segura, B.; Campabadal, A.; Uribe, C.; Giraldo, D.M.; Perez-Soriano, A.; Muñoz, E.; Compta, Y.; Junque, C.; et al. Cerebellar resting-state functional connectivity in Parkinson's disease and multiple system atrophy: Characterization of abnormalities and potential for differential diagnosis at the single-patient level. *NeuroImage Clin.* **2019**, *22*, 101720. [CrossRef]
47. Abos, A.; Baggio, H.C.; Segura, B.; Campabadal, A.; Uribe, C.; Giraldo, D.M.; Perez-Soriano, A.; Muñoz, E.; Compta, Y.; Junque, C.; et al. Differentiation of multiple system atrophy from Parkinson's disease by structural connectivity derived from probabilistic tractography. *Sci. Rep.* **2019**, *9*, 16488. [CrossRef]

48. Kiryu, S.; Yasaka, K.; Akai, H.; Nakata, Y.; Sugomori, Y.; Hara, S.; Seo, M.; Abe, O.; Ohtomo, K. Deep learning to differentiate parkinsonian disorders separately using single midsagittal MR imaging: A proof of concept study. *Eur. Radiol.* **2019**, *29*, 6891–6899. [CrossRef]

49. Yin, B.; Balvert, M.; van der Spek, R.A.A.; Dutilh, B.E.; Bohté, S.; Veldink, J.; Schönhuth, A. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. *Bioinformatics* **2019**, *35*, i538–i547. [CrossRef] [PubMed]

50. Sarica, A.; Cerasa, A.; Valentino, P.; Yeatman, J.; Trotta, M.; Barone, S.; Granata, A.; Nisticò, R.; Perrotta, P.; Pucci, F.; et al. The corticospinal tract profile in amyotrophic lateral sclerosis. *Hum. Brain Mapp.* **2016**, *38*, 727–739. [CrossRef] [PubMed]

51. Welsh, R.C.; Jelsone-Swain, L.M.; Foerster, B.R. The Utility of Independent Component Analysis and Machine Learning in the Identification of the Amyotrophic Lateral Sclerosis Diseased Brain. *Front. Hum. Neurosci.* **2013**, *7*, 251. [CrossRef] [PubMed]

52. Palstrøm, N.B.; Rojek, A.M.; Møller, H.E.H.; Hansen, C.T.; Matthiesen, R.; Rasmussen, L.M.; Abildgaard, N.; Beck, H.C. Classification of Amyloidosis by Model-Assisted Mass Spectrometry-Based Proteomics. *Int. J. Mol. Sci.* **2021**, *23*, 319. [CrossRef] [PubMed]

53. Garcia-Carretero, R.; Olid-Velilla, M.; Perez-Torrella, D.; Torres-Pacho, N.; Darnaude-Ortiz, M.T.; Bustamate-Zuloeta, A.D.; Tenorio, J.A. Predictive modeling of hypophosphatasia based on a case series of adult patients with persistent hypophosphatasemia. *Osteoporos. Int.* **2021**, *32*, 1815–1824. [CrossRef]

54. Nguyen, K.D.; Corben, L.A.; Pathirana, P.N.; Horne, M.K.; Delatycki, M.B.; Szmulewicz, D.J. An Instrumented Measurement Scheme for the Assessment of Upper Limb Function in Individuals with Friedreich Ataxia. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019. [CrossRef]

55. Salvatore, C.; Cerasa, A.; Castiglioni, I.; Gallivanone, F.; Augimeri, A.; Lopez, M.; Arabia, G.; Morelli, M.; Gilardi, M.C.; Quattrone, A. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *J. Neurosci. Methods* **2014**, *222*, 230–237. [CrossRef]

56. Abidin, A.Z.; Jameson, J.; Molthen, R.; Wismüller, A. Classification of micro-CT images using 3D characterization of bone canal patterns in human osteogenesis imperfecta. *Proc. SPIE Int. Soc. Opt. Eng.* **2017**, *10134*, 1013413. [CrossRef]

57. Karthikeyan, S.; Ramkumar, G.; Aravindkumar, S.; Tamilselvi, M.; Ramesh, S.; Ranjith, A. A Novel Deep Learning-Based Black Fungus Disease Identification Using Modified Hybrid Learning Methodology. *Contrast Media Mol. Imaging* **2022**, *2022*, 4352730. [CrossRef]

58. Masumoto, H.; Tabuchi, H.; Nakakura, S.; Ohsugi, H.; Enno, H.; Ishitobi, N.; Ohsugi, E.; Mitamura, Y. Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. *PeerJ* **2019**, *7*, e6900. [CrossRef]

59. Wang, T.; Lei, Y.; Tian, S.; Jiang, X.; Zhou, J.; Liu, T.; Dresser, S.; Curran, W.J.; Shu, H.; Yang, X. Learning-based automatic segmentation of arteriovenous malformations on contrast CT images in brain stereotactic radiosurgery. *Med. Phys.* **2019**, *46*, 3133–3141. [CrossRef]

60. Kovacs, W.; Hsieh, N.; Roth, H.; Nnamdi-Emeratom, C.; Bandettini, W.P.; Arai, A.; Mankodi, A.; Summers, R.M.; Yao, J. Holistic segmentation of the lung in cine MRI. *J. Med. Imaging* **2017**, *4*, 041310. [CrossRef]

61. Zhang, H.-W.; Zhou, P.; Wang, K.-Z.; Liu, J.-B.; Huang, Y.-S.; Tu, Y.-T.; Deng, Z.-H.; Zhu, X.-D.; Hang, Y.-L. Platelet proteomics in diagnostic differentiation of primary immune thrombocytopenia using SELDI-TOF-MS. *Clin. Chim. Acta* **2016**, *455*, 75–79. [CrossRef] [PubMed]

62. Mordenti, M.; Ferrari, E.; Pedrini, E.; Fabbri, N.; Campanacci, L.; Muselli, M.; Sangiorgi, L. Validation of a new multiple osteochondromas classification through Switching Neural Networks. *Am. J. Med. Genet. Part A* **2013**, *161*, 556–560. [CrossRef] [PubMed]

63. Artoni, P.; Piffer, A.; Vinci, V.; LeBlanc, J.; Nelson, C.A.; Hensch, T.K.; Fagiolini, M. Deep learning of spontaneous arousal fluctuations detects early cholinergic defects across neurodevelopmental mouse models and patients. *Proc. Natl. Acad. Sci. USA* **2019**, *117*, 23298–23303. [CrossRef] [PubMed]

64. Errington, N.; Iremonger, J.; Pickworth, J.A.; Kariotis, S.; Rhodes, C.J.; Rothman, A.M.; Condliffe, R.; Elliot, C.A.; Kiely, D.G.; Howard, L.S.; et al. A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach. *Ebiomedicine* **2021**, *69*, 103444. [CrossRef] [PubMed]

65. Iwasawa, K.; Suda, W.; Tsunoda, T.; Oikawa-Kawamoto, M.; Umetsu, S.; Takayasu, L.; Inui, A.; Fujisawa, T.; Morita, H.; Sogo, T.; et al. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. *Sci. Rep.* **2018**, *8*, 5480. [CrossRef] [PubMed]

66. Wang, Z.; Camino, A.; Hagag, A.M.; Wang, J.; Weleber, R.G.; Yang, P.; Pennesi, M.E.; Huang, D.; Li, D.; Jia, Y. Automated detection of preserved photoreceptor on optical coherence tomography in choroideremia based on machine learning. *J. Biophotonics* **2018**, *11*, e201700313. [CrossRef]

67. Nodera, H.; Sogawa, K.; Takamatsu, N.; Hashiguchi, S.; Saito, M.; Mori, A.; Osaki, Y.; Izumi, Y.; Kaji, R. Texture analysis of sonographic muscle images can distinguish myopathic conditions. *J. Med. Investig.* **2019**, *66*, 237–247. [CrossRef]

68. Shire. Rare Disease Impact Report: Insights from Patients and the Medical Community. 2013. Available online: https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf (accessed on 18 November 2022).

69. Bohr, A.; Memarzadeh, K. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 25–60. [CrossRef]

70. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big Data in Health Care: Using Analytics to Identify And Manage High-Risk and High-Cost Patients. *Health Aff.* **2014**, *33*, 1123–1131. [CrossRef]

71. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]

72. Marquardt, A.; Landwehr, L.-S.; Ronchi, C.L.; di Dalmazi, G.; Riester, A.; Kollmannsberger, P.; Altieri, B.; Fassnacht, M.; Sbiera, S. Identifying New Potential Biomarkers in Adrenocortical Tumors Based on mRNA Expression Data Using Machine Learning. *Cancers* **2021**, *13*, 4671. [CrossRef] [PubMed]

73. Jang, H.; Moon, S.; Jung, K.; Kim, S.; Kim, H.; Han, D.; Kim, J. Mass Spectrometry-Based Proteomic Discovery of Prognostic Biomarkers in Adrenal Cortical Carcinoma. *Cancers* **2021**, *13*, 3890. [CrossRef] [PubMed]

74. Spiga, O.; Cicaloni, V.; Bernini, A.; Zatkova, A.; Santucci, A. ApreciseKUre: An approach of Precision Medicine in a Rare Disease. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 42. [CrossRef]

75. Spiga, O.; Cicaloni, V.; Fiorini, C.; Trezza, A.; Visibelli, A.; Millucci, L.; Bernardini, G.; Bernini, A.; Marzocchi, B.; Braconi, D.; et al. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet J. Rare Dis.* **2020**, *15*, 46. [CrossRef]

76. Spiga, O.; Cicaloni, V.; Dimitri, G.M.; Pettini, F.; Braconi, D.; Bernini, A.; Santucci, A. Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease. *Brief. Bioinform.* **2021**, *22*, bbaa434. [CrossRef]

77. Blasco, H.; Patin, F.; Descat, A.; Garçon, G.; Corcia, P.; Gelé, P.; Lenglet, T.; Bede, P.; Meininger, V.; Devos, D.; et al. A pharmaco-metabolomics approach in a clinical trial of ALS: Identification of predictive markers of progression. *PLoS ONE* **2018**, *13*, e0198116. [CrossRef]

78. Huang, Z.; Zhang, H.; Boss, J.; Goutman, S.A.; Mukherjee, B.; Dinov, I.D.; Guan, Y. For the Pooled Resource Open-Access ALS Clinical Trials Consortium Complete hazard ranking to analyze right-censored data: An ALS survival study. *PLOS Comput. Biol.* **2017**, *13*, e1005887. [CrossRef] [PubMed]

79. Gordon, J.; Lerner, B. Insights into Amyotrophic Lateral Sclerosis from a Machine Learning Perspective. *J. Clin. Med.* **2019**, *8*, 1578. [CrossRef]

80. Wang, J.; Kothalkar, P.V.; Kim, M.; Bandini, A.; Cao, B.; Yunusova, Y.; Campbell, T.F.; Heitzman, D.; Green, J.R. Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples. *Int. J. Speech-Lang. Pathol.* **2018**, *20*, 669–679. [CrossRef]

81. van der Burgh, H.K.; Schmidt, R.; Westeneng, H.-J.; de Reus, M.A.; Berg, L.H.V.D.; Heuvel, M.P.V.D. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage Clin.* **2016**, *13*, 361–369. [CrossRef]

82. Lauraitis, A.; Maskeliūnas, R.; Damaševičius, R. ANN and Fuzzy Logic Based Model to Evaluate Huntington Disease Symptoms. *J. Healthc. Eng.* **2018**, *2018*, 4581272. [CrossRef]

83. Mason, S.L.; Daws, R.; Soreq, E.; Johnson, E.B.E.B.; Scahill, R.; Tabrizi, S.J.S.J.; Barker, R.A.; Hampshire, A. Predicting clinical diagnosis in Huntington's disease: An imaging polymarker. *Ann. Neurol.* **2018**, *83*, 532–543. [CrossRef] [PubMed]

84. Asadi, H.; Kok, H.K.; Looby, S.; Brennan, P.; O'Hare, A.; Thornton, J. Outcomes and Complications After Endovascular Treatment of Brain Arteriovenous Malformations: A Prognostication Attempt Using Artificial Intelligence. *World Neurosurg.* **2016**, *96*, 562–569. [CrossRef] [PubMed]

85. Chen, Y.-L.; Zhao, X.-A.; Ng, S.-H.; Lu, C.-S.; Lin, Y.-C.; Cheng, J.-S.; Tsai, C.-C.; Wang, J.-J. Prediction of the Clinical Severity of Progressive Supranuclear Palsy by Diffusion Tensor Imaging. *J. Clin. Med.* **2019**, *9*, 40. [CrossRef]

86. Zhutovsky, P.; Vijverberg, E.G.; Bruin, W.B.; Thomas, R.M.; Wattjes, M.P.; Pijnenburg, Y.A.; van Wingen, G.A.; Dols, A. Individual Prediction of Behavioral Variant Frontotemporal Dementia Development Using Multivariate Pattern Analysis of Magnetic Resonance Imaging Data. *J. Alzheimer's Dis.* **2019**, *68*, 1229–1241. [CrossRef]

87. An, K.; Zhou, J.-B.; Xiong, Y.; Han, W.; Wang, T.; Ye, Z.-Q.; Wu, Y.-D. Computational Studies of the Structural Basis of Human RPS19 Mutations Associated with Diamond-Blackfan Anemia. *Front. Genet.* **2021**, *12*, 650897. [CrossRef]

88. Pande, A.; Mohapatra, P.; Nicorici, A.; Han, J.J.; Altini, M.; Voos, M. Machine Learning to Improve Energy Expenditure Estimation in Children with Disabilities: A Pilot Study in Duchenne Muscular Dystrophy. *JMIR Rehabil. Assist. Technol.* **2016**, *3*, e7. [CrossRef] [PubMed]

89. Eaton, J.E.; Vesterhus, M.; McCauley, B.M.; Atkinson, E.J.; Schlicht, E.M.; Juran, B.D.; Gossard, A.A.; LaRusso, N.F.; Gores, G.J.; Karlsen, T.H.; et al. Primary Sclerosing Cholangitis Risk Estimate Tool (PREsTo) Predicts Outcomes of the Disease: A Derivation and Validation Study Using Machine Learning. *Hepatology* **2018**, *71*, 214–224. [CrossRef]

90. Robinson, G.A.; Peng, J.; Dönnes, P.; Coelewij, L.; Naja, M.; Radziszewska, A.; Wincup, C.; Peckham, H.; Isenberg, D.A.; Ioannou, Y.; et al. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: Patient stratification using a machine-learning approach. *Lancet Rheumatol.* **2020**, *2*, e485–e496. [CrossRef]

91. Chou, F.-S.; Ghimire, L.V. Identification of prognostic factors for pediatric myocarditis with a random forests algorithm-assisted approach. *Pediatr. Res.* **2020**, *90*, 427–430. [CrossRef]

92. Chou, F.-S.; Ghimire, L.V. Machine Learning for Mortality Prediction in Pediatric Myocarditis. *Front. Pediatr.* **2021**, *9*, 644922. [CrossRef] [PubMed]

93. Tabor, H.K.; Goldenberg, A. What Precision Medicine Can Learn from Rare Genetic Disease Research and Translation. *AMA J. Ethics* **2018**, *20*, 834–840. [CrossRef]

94. Foksinska, A.; Crowder, C.M.; Crouse, A.B.; Henrikson, J.; Byrd, W.E.; Rosenblatt, G.; Patton, M.J.; He, K.; Tran-Nguyen, T.K.; Zheng, M.; et al. The precision medicine process for treating rare disease using the artificial intelligence tool mediKanren. *Front. Artif. Intell.* **2022**, *5*, 910216. [CrossRef]

95. Bakkar, N.; Kovalik, T.; Lorenzini, I.; Spangler, S.; Lacoste, A.; Sponaugle, K.; Ferrante, P.; Argentinis, E.; Sattler, R.; Bowser, R. Artificial intelligence in neurodegenerative disease research: Use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. *Acta Neuropathol.* **2017**, *135*, 227–247. [CrossRef]

96. Chen, Y.; Argentinis, J.E.; Weber, G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin. Ther.* **2016**, *38*, 688–701. [CrossRef]

97. Cao, H.; Zhang, L.; Jin, B.; Cheng, S.; Wei, X.; Che, C. Enriching limited information on rare diseases from heterogeneous networks for drug repositioning. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 304. [CrossRef] [PubMed]

98. Klein, J.J.; Baker, N.C.; Foil, D.H.; Zorn, K.M.; Urbina, F.; Puhl, A.C.; Ekins, S. Using Bibliometric Analysis and Machine Learning to Identify Compounds Binding to Sialidase-1. *ACS Omega* **2021**, *6*, 3186–3193. [CrossRef]

99. Zorn, K.; Lingerfelt, M.; Clark, A.; Ekins, S. Assay Central: A new approach to compiling Big Data and preparing machine learning models for drug repurposing. In Proceedings of the American Chemical Society National Meeting, Washington, DC, USA, 23 August 2017. [CrossRef]

100. Wolfe, J.M.; Fadzen, C.M.; Choo, Z.-N.; Holden, R.L.; Yao, M.; Hanson, G.J.; Pentelute, B.L. Machine Learning to Predict Cell-Penetrating Peptides for Antisense Delivery. *ACS Cent. Sci.* **2018**, *4*, 512–520. [CrossRef]

101. Esteban-Medina, M.; Peña-Chilet, M.; Loucera, C.; Dopazo, J. Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. *BMC Bioinform.* **2019**, *20*, 370. [CrossRef]

102. Taniguchi, T.; D'Andrea, A.D. Molecular pathogenesis of Fanconi anemia: Recent progress. *Blood* **2006**, *107*, 4223–4233. [CrossRef]

103. Spiga, O.; Cicaloni, V.; Visibelli, A.; Davoli, A.; Paparo, M.; Orlandini, M.; Vecchi, B.; Santucci, A. Towards a Precision Medicine Approach Based on Machine Learning for Tailoring Medical Treatment in Alkaptonuria. *Int. J. Mol. Sci.* **2021**, *22*, 1187. [CrossRef] [PubMed]