

# AI Application in Rare Diseases

Subjects: Others

Contributor: Anna Visibelli, Bianca Roncaglia, Ottavia Spiga, Annalisa Santucci

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.

Keywords: rare disease ; machine learning ; artificial intelligence

---

## 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 <sup>[1]</sup>, based on the requirement of providers to document associated phenotypic information to support a diagnosis, researchers hypothesize that patients' phenotypic data stored in electronic medical records can be used to speed up disease diagnosis. The preliminary results obtained demonstrated that the use of collaborative filtering with phenotypic information can stratify patients with relatively similar rare diseases. In <sup>[2]</sup>, 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 <sup>[3]</sup> based on the fact that clinical symptoms in children with pulmonary diseases are frequently non-specific, researchers 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.

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 [4], 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 [5] 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 [6], 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 [7], 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 [8][9] 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 [10] 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 [11], 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, researchers 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 [12]. 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 [13], 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. Hypophosphatasia is a rare genetic disease in which patients may have stress fractures, bone and joint pain, or premature tooth loss. In [14], 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. [15] 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 [16] 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 [17] 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 [18][19], 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 [18] to identify mucormycosis disease by comparing images of patients with and without mucormycosis, a rare fungal infection caused by a group of molds. In [19], 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 [20][21]. 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 [22], researchers 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 [23], researchers studied multiple osteochondromas, an autosomal dominant disease characterized by the formation of osteochondromas or exostoses.

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 [24], 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 [25], 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 [26], 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 [27] 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.

## **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 [28]. 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 [29].

AI can play a significant role in the prognosis of rare disorders by helping to fill in the gaps in data and experience [30]. 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 [31]. 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 [32].

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 [33][34] 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 [33], 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 [34], 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 [35] has implemented a digital platform, ApreKure, 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 ApreKure 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 [36] aimed to predict QoL scores based on patient's clinical data using the XGBoost algorithm and a k-NN algorithm.

ALS is another rare and very serious disease that has been studied with AI methods. In [37], 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 [38], 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 [39] 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. [40] 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 [41], 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. [42] 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 condition) 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 [43], 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. [44] 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 nidus 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 [45], 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. [46] and An et al. [47]. In [46], 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 [47], 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.

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 <sup>[48][49]</sup>, they applied RF algorithms to identify prognostic factors in pediatric myocarditis patients. In their first study <sup>[48]</sup>, 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 <sup>[49]</sup> 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.

### **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 <sup>[50]</sup>. 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 <sup>[51]</sup>, 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 RHOTB2 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 <sup>[52]</sup>, Bakkar et al. implemented the IBM Watson® <sup>[53]</sup> 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 <sup>[54]</sup> 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 <sup>[55]</sup>, 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 <sup>[56]</sup> 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 <sup>[57]</sup>, 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 [58] 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 [59]. Thanks to these models, over 20 potential therapeutic targets were detected. In the last study [60], 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.

---

## References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.

16. 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.
17. 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.
18. 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.
19. 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.
20. 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. Physics* 2019, 46, 3133–3141.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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).
30. 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.
31. 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.
32. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* 2019, 380, 1347–1358.
33. 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.
34. 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.
35. Spiga, O.; Cicaloni, V.; Bernini, A.; Zatkova, A.; Santucci, A. ApreKure: An approach of Precision Medicine in a Rare Disease. *BMC Med. Inform. Decis. Mak.* 2017, 17, 42.
36. 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.



37. 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.
38. 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.
39. Gordon, J.; Lerner, B. Insights into Amyotrophic Lateral Sclerosis from a Machine Learning Perspective. *J. Clin. Med.* 2019, 8, 1578.
40. 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.
41. 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.
42. Lauraitis, A.; Maskeliūnas, R.; Damaševičius, R. ANN and Fuzzy Logic Based Model to Evaluate Huntington Disease Symptoms. *J. Healthc. Eng.* 2018, 2018, 4581272.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. Chou, F.-S.; Ghimire, L.V. Machine Learning for Mortality Prediction in Pediatric Myocarditis. *Front. Pediatr.* 2021, 9, 644922.
50. Tabor, H.K.; Goldenberg, A. What Precision Medicine Can Learn from Rare Genetic Disease Research and Translation. *AMA J. Ethics* 2018, 20, 834–840.
51. 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 medKanren. *Front. Artif. Intell.* 2022, 5, 910216.
52. Bakkar, N.; Kovalik, T.; Lorenzini, I.; Spangler, S.; Lacoste, A.; Sponaugle, K.; Ferrante, P.; Argentinis, E.; Sattler, R.; Bowers, 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.

59. Taniguchi, T.; D'Andrea, A.D. Molecular pathogenesis of Fanconi anemia: Recent progress. *Blood* 2006, 107, 4223–4233.
60. 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.

---

Retrieved from <https://encyclopedia.pub/entry/history/show/95718>