Journal of Biomedical Informatics 143 (2023) 104429

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases

P. Roman-Naranjo [b,c,d,*], A.M. Parra-Perez [b,c,d], J.A. Lopez-Escamez [a,b,c,d,*]

[a] Meniere's Disease Neuroscience Research Program, Faculty of Medicine & Health, School of Medical Sciences, The Kolling Institute, University of Sydney, Sydney, New South Wales, Australia
[b] Division of Otolaryngology, Department of Surgery, Instituto de Investigación Biosanitaria, ibs.GRANADA, Universidad de Granada, Granada, Spain
[c] Otology and Neurotology Group CTS495, Department of Genomic Medicine, GENYO - Centre for Genomics and Oncological Research - Pfizer, University of Granada, Junta de Andalucía, PTS, Granada, Spain
[d] Sensorineural Pathology Programme, Centro de Investigación Biomédica en Red en Enfermedades Raras, CIBERER, Madrid, Spain

ARTICLE INFO

Keywords:
Artificial intelligence
Rare diseases
Precision medicine
Rare variants
DNA-sequencing
Genomics

ABSTRACT

Background: The diagnosis of rare genetic diseases is often challenging due to the complexity of the genetic underpinnings of these conditions and the limited availability of diagnostic tools. Machine learning (ML) algorithms have the potential to improve the accuracy and speed of diagnosis by analyzing large amounts of genomic data and identifying complex multiallelic patterns that may be associated with specific diseases. In this systematic review, we aimed to identify the methodological trends and the ML application areas in rare genetic diseases.

Methods: We performed a systematic review of the literature following the PRISMA guidelines to search studies that used ML approaches to enhance the diagnosis of rare genetic diseases. Studies that used DNA-based sequencing data and a variety of ML algorithms were included, summarized, and analyzed using bibliometric methods, visualization tools, and a feature co-occurrence analysis.

Findings: Our search identified 22 studies that met the inclusion criteria. We found that exome sequencing was the most frequently used sequencing technology (59%), and rare neoplastic diseases were the most prevalent disease scenario (59%). In rare neoplasms, the most frequent applications of ML models were the differential diagnosis or stratification of patients (38.5%) and the identification of somatic mutations (30.8%). In other rare diseases, the most frequent goals were the prioritization of rare variants or genes (55.5%) and the identification of biallelic or digenic inheritance (33.3%). The most employed method was the random forest algorithm (54.5%). In addition, the features of the datasets needed for training these algorithms were distinctive depending on the goal pursued, including the mutational load in each gene for the differential diagnosis of patients, or the combination of genotype features and sequence-derived features (such as GC-content) for the identification of somatic mutations.

Conclusions: ML algorithms based on sequencing data are mainly used for the diagnosis of rare neoplastic diseases, with random forest being the most common approach. We identified key features in the datasets used for training these ML models according to the objective pursued. These features can support the development of future ML models in the diagnosis of rare genetic diseases.

1. Introduction

Rare diseases (RDs) continue to be a challenge to the healthcare system due to the difficulty of reaching an accurate diagnosis. Although there is no uniform international criteria, RDs are usually defined as those affecting fewer than 4–5 cases out of 10,000 individuals [1]. Considering them as a whole, RDs can be regarded as a common event, with 7,265 different RDs (http://www.orphadata.org/data/xml/en_pro

* Corresponding author at: Division of Otolaryngology, Department of Surgery, Instituto de Investigación Biosanitaria, ibs.GRANADA, Universidad de Granada, Granada, Spain (P. Roman-Naranjo). Meniere's Disease Neuroscience Research Program, Faculty of Medicine & Health, School of Medical Sciences, The Kolling Institute, University of Sydney, Sydney, New South Wales, Australia (J.A. Lopez-Escamez).
E-mail addresses: romanjo@ugr.es (P. Roman-Naranjo), alberto.parra@genyo.es (A.M. Parra-Perez), jose.lopezescamez@sydney.edu.au (J.A. Lopez-Escamez).

https://doi.org/10.1016/j.jbi.2023.104429
Received 28 January 2023; Received in revised form 5 June 2023; Accepted 17 June 2023
Available online 22 June 2023

duct7.xml, updated on November 28, 2022) with an estimated accumulated prevalence of 3.5–5.9% and affecting more than 400 million people worldwide [2,3].

Most RDs appear to be caused or modified by genetic factors; up to 80% of them are thought to have a genetic etiology [4]. Our current knowledge on this aspect is limited, existing 3,868 RDs (53.2%) linked to, at least, a gene that cause or modify the disease phenotype (http://www.orphadata.org/data/xml/en_product6.xml, updated on November 28, 2022) [2]. The improved performance and the price reduction of Next-generation sequencing (NGS) technologies in recent years have made them more attractive for clinical applications in RDs, increasing rapidly the number of phenotype-genotype associations [5]. This has resulted in an accurate molecular diagnosis in many patients suffering from monogenic RDs, which has occasionally led to personalized treatments and improved disease management. Nevertheless, other patients with more complex disorders receive an inconclusive genetic diagnosis, placing the diagnostic yield of DNA-based NGS technologies in most studies at 40–50% [6,7]. This is mainly caused by the absence of pathogenic or likely pathogenic variants in known disease-causing genes, finding instead variants of unknown significance (VUS) or variants in novel genes not previously associated with the disease.

In this scenario of rare and complex genetic disorders where a diagnosis is not reached or a prognosis is not accurate enough, more sophisticated methods should be applied to analyze large-scale genomic data. The use of artificial intelligence (AI) and, particularly, machine learning (ML) algorithms has raised great interest in recent years due to its potential to uncover complex patterns in genomic data [8]. These ML algorithms have shown the capacity to learn from and act on large, heterogeneous datasets to extract new biological insights, improving the accuracy of the diagnosis of RDs [9–12].

Compared to previous reviews in the field of ML and RDs, such as Schaefer *et al.* [9] or Brasil *et al.* [13], in this systematic review we used a different approach, investigating the role of AI/ML algorithms in the diagnosis and prognosis of RDs using genomic data. The range of options when it comes to choosing a learning algorithm or a DNA-based NGS technique to address RDs is highly variable. On the one hand, ML methods are usually divided into two main categories: supervised and unsupervised learning. Supervised ML algorithms require labeled data to solve mainly regression and classification tasks, whereas unsupervised ML algorithms address classification tasks based on unlabeled data by seeking common patterns. The review from Libbrecht *et al.* describes these algorithms in more detail and provides examples applied to genomic data [14]. On the other hand, regarding NGS techniques, there are mainly two strategies: a) to sequence the entirety of the DNA sequence (whole genome sequencing, WGS), or b) to just sequence some regions of the DNA, such as coding regions (exome sequencing, ES), or certain disease-causing genes (gene panel). Nevertheless, the raw data generated in these experiments can be processed in many ways, with different workflows depending on the aim of the study.

This systematic review presents a thorough overview of the existing evidence on the application of AI/ML algorithms to the diagnosis of RDs using DNA-based sequencing data. We conducted a comprehensive search of the literature and included studies that used a variety of ML approaches and sequencing data sources in different research settings. Our analysis focused on the evaluation of trends in the field, the ability of these approaches to identify genetic variations associated with RDs, and the potential of AI/ML to improve their diagnosis.

2. Methods

2.1. Systematic literature search and data sources

We performed a literature search using PubMed, Web of Science, and Scopus to identify relevant publications on the use of AI/ML for the diagnosis and prognosis of RDs using genomic data. We also used citation and hand searching to ensure that potentially relevant studies were retrieved. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed to design and perform this systematic review [15], and its protocol was registered in PROSPERO (registration number CRD42022360247).

A search in the selected databases using the search terms ‘rare AND (“artificial intelligence” OR “machine learning” OR “deep learning”) AND ((exome OR genome OR panel) AND sequencing)’ and considering publications from 2012 onward resulted in 296 abstracts. The citation and hand searching resulted in 10 additional records. The date of the last search was September 29, 2022.

The list of abstracts was screened for inclusion using the following inclusion criteria: (i) an application of AI/ML methods; (ii) a diagnostic or prognosis application using a DNA-based NGS technique (panel, exome, or genome sequencing); and (iii) an application to a RD within the orpha.net database. Non-English articles, review articles, conference papers, duplicate records, and studies not relevant to any RD or AI/ML were excluded. To narrow our focus to clinical applications, we excluded animal studies as well as publications that only reported methodological aspects of AI/ML without presenting clinical data from the study population. For all articles considered relevant, the full text was reviewed using the same screening procedure as in the first stage.

2.2. Data extraction

All the selected articles were evaluated to gather data on five main aspects: i) study characteristics and study population (subjects included, RDs studied, study design, use of secondary data), ii) characteristics of the applied AI/ML techniques (selected ML model, programming languages used, input data, associated features, feature selection methods, model evaluation), iii) information about the DNA-based NGS technology used (type, sample collected, DNA sequencing kit, sequencing platform, read length, mean coverage), iv) the variant discovery approach (alignment method, used SNV/Indel/CNVs callers, variant annotation software, variant filtering criteria), and v) authors (number of authors and institutions involved, authors’ countries) and journal details (name, category, journal impact factor, journal citation indicator).

2.3. Data analysis

The data collected from selected articles were summarized and analyzed using a variety of approaches. Journal Impact Factor (JIF) scores were obtained from the Journal Citation Report (JCR) database. Bibliometric networks, including data from authors and abstracts, were constructed and visualized using VOSviewer [16]. Similarly, full-text articles were analyzed using WordStat 9.0 (Provalis Research, Montreal, Quebec, Canada) to extract main topics and keywords.

Selected articles were divided into “rare neoplastic diseases” and “other rare diseases” to enable comparisons. AI/ML models were categorized into three categories: supervised, unsupervised, and deep learning models. Input variables that the model uses to make predictions (features) were classified in 1) “clinical features”, which include information about patients’ clinical characteristics; 2) “phenotype-related features”, including data about the association between genes and phenotypes (e.g., Human Phenotype Ontology); 3) “read alignment features”, which include the properties related to read mapping and sequencing quality; 4) “genotype-related features”, including details of variants found in patients (e.g., variant allele frequency, count of variants in a certain gene, length of indel); 5) “sequence region and structural features”, including information about the region where the variant is located (e.g., gene size, GC content); 6) “network features”, which include details about the known pathways in which a particular gene is involved (e.g., number of pathway, network neighbors); 7) “evolutionary/pathogenicity features”, which include pathogenicity and evolutionary conservation scores of variants (e.g., CADD, PolyPhen-2);

8) "gene expression features", including data on gene expression; 9) "tissue-specific features", including features which are specific for certain types of tissues; and 10) "disease-specific features", including features which are specific for certain types of diseases. The co-occurrence of these features in the datasets used for training AI/ML models was examined and plotted using UpSetR [17].

3. Results

3.1. Included studies

The literature search in databases identified 494 studies, with 296 remaining after removing duplicates (Supplementary Table 1). Among them, 93 studies were selected for full-text review, and 14 were included in the final analysis. In addition, 11 studies were identified through hand and citation searching. After screening, 8 further studies met the selection criteria of this systematic review. Thus, 22 studies were included in the final analysis (Supplementary Table 2). Fig. 1 shows the PRISMA flow diagram for article selection, including the reasons for excluding records.

3.2. Temporal trends and bibliometrics

To assess the temporal trends in the use of AI/ML methods for the diagnosis and prognosis of RDs using sequencing data, *meta*-data from included articles was retrieved (Supplementary Table 3). In recent years, we noticed a relative rise in the number of studies that address this challenge using AI/ML (Fig. 2A). Most of these articles were published in journals belonging to the first quartile (90.9%) and within the "Genetics & Hereditary" JCR category (31.8%) (**Supplementary** Fig. 1). It should be noted that the count for 2022 is based on studies published up to September 29, 2022.

A total of 318 authors contributed to the selected articles. The bibliometric analysis showed a low level of collaboration between authors of different articles, creating 19 clusters where only 3 authors participated in 2 or more articles (**Supplementary** Fig. 2A). The term co-occurrence analysis of abstracts found 100 relevant terms divided into 3 clusters that summarize the main topics of this research field. These clusters group together terms mainly associated with genetics (cluster 1), cancer (cluster 2), and methodology terms (cluster 3) (Fig. 2B). The most frequently occurring terms in these abstracts were "genetics" (18 occurrences), "machine learning" (15 occurrences) and "whole-exome sequencing" (10 occurrences). These key terms were also among the most frequently used terms in the analysis of full-text articles, where terms such as "random forest" (59.1% of studies), "somatic mutations" (54.5% of studies), or "rare variants" (54.5% of studies) were also in a significant proportion of studies (**Supplementary** Fig. 2B).

3.3. Application areas for AI/ML techniques

The most common disease scenario was rare neoplastic diseases (59%). The remaining studies investigated different kinds of RDs, such as developmental, neurological, or circulatory diseases **(**Fig. 3A**).** Exome sequencing was the most used NGS method in both rare neoplastic diseases (61.5%) and other RDs (55.5%) **(**Fig. 3B**).** Of note, 63.6% (14/22) of the studies employed sequencing data stored in external databases, primarily The Cancer Genome Atlas (TCGA), but also the Myocardial Genetics Consortium (MIGEN), or the Undiagnosed Diseases Network (UDN). These studies showed larger sample sizes than those using their own cohorts (**Supplementary** Fig. 3), but they also showed higher intra-method variability, as seen by the mixed sample processing methods they employed (**Supplementary** Fig. 4). Supplementary Table 4 summarizes the NGS-related and sequencing data processing methods in detail.

Supervised machine learning methods were chosen in 86.3% of the studies, with Random Forest (RF) being the most employed algorithm within this group (54.5%) **(**Fig. 3C**).** One study discarded the genetic features after the feature selection process, and three studies did not describe the selected features in detail. In terms of model performance evaluation, 20 studies clearly stated the evaluation metrics of the selected model, while 2 studies directly referred to the results obtained by the selected model without reporting evaluation metrics. Finally, 12

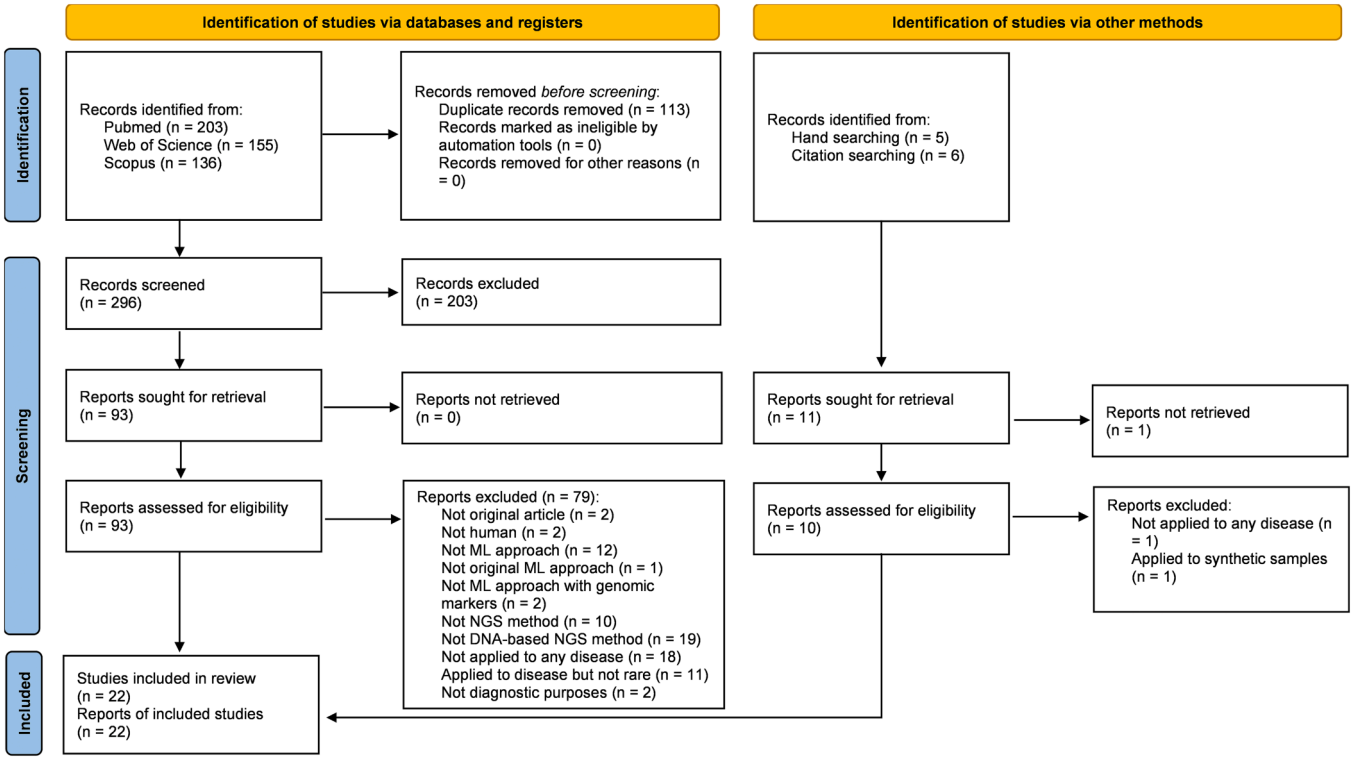

Fig. 1. PRISMA flow diagram for the identification, screening and selection of genetic studies using AI/ML for the diagnosis of rare diseases.

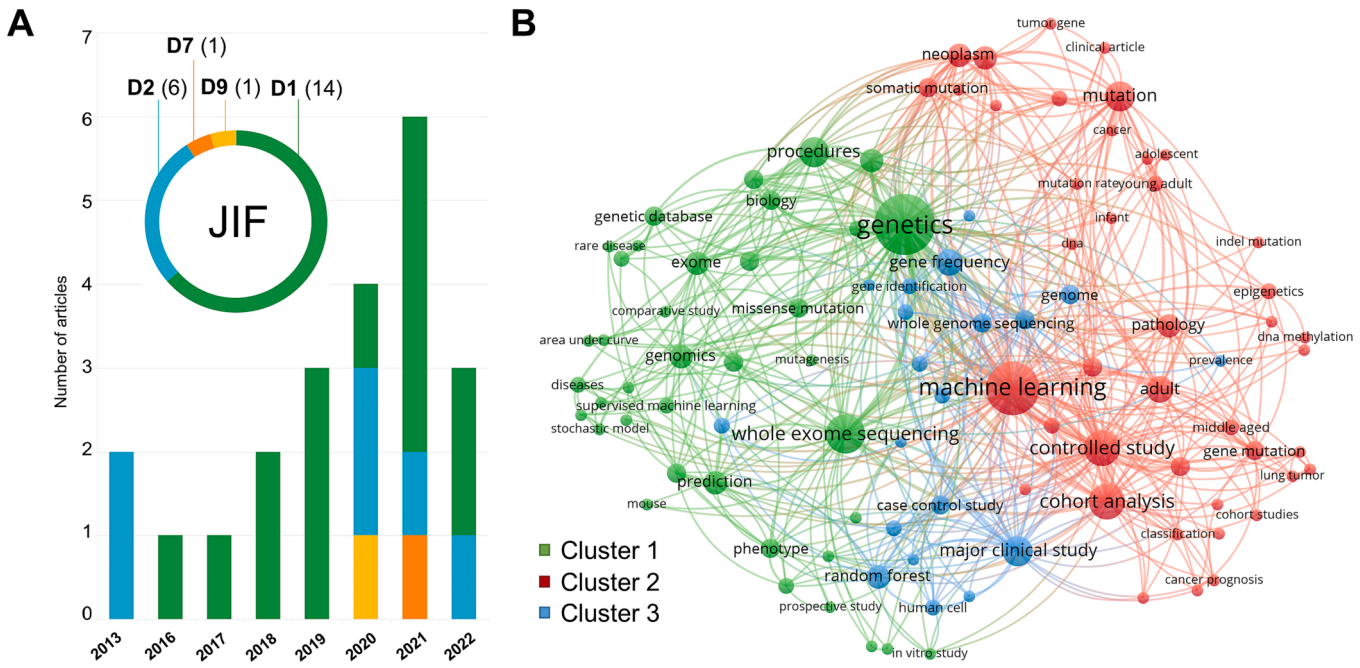

Fig. 2. Visualization of temporal trends and bibliometrics. Panel **A)** shows the selected studies distributed per year and divided into deciles (D) according to journal impact factors (JIF). Panel **B)** displays a keyword co-occurrence network using abstracts of selected studies.

of the 22 included studies compared the efficacy of their models with at least one existing approach (comparators). These studies reported varying degrees of improvement over the comparators. Supplementary Table 5 summarizes the AI/ML methods chosen by the included studies and their performances.

3.4. AI/ML in the study of rare genetic diseases

The objectives of AI/ML approaches in the different studies were investigated. It was found that the primary goal of using AI/ML in rare neoplastic diseases was the differential diagnosis of patients (5/13), followed by the identification of somatic mutations when a matched normal tissue was not available (4/13). In contrast, the major goals in other RDs were to prioritize variants and candidate genes (5/9) and to identify biallelic or digenic inheritance (3/9). To date, the use of AI/ML for the differential diagnosis of patients with non-neoplastic diseases is uncommon (1/9) (Fig. 4A).

Looking at the types of instances (labels) and features (attributes) of datasets used for training these AI/ML models, we found that they were distinctive and different depending on the goal pursued (Table 1 and Fig. 4B). For the differential diagnosis of patients, most datasets included only features related to the genotype of patients. These features primarily contained mutational load data for each gene or genomic window using collapsing methods. Models trained to predict the prognosis of RDs included clinical features (e.g., sex, age, exposure to certain substances) in addition to genotype features. The four AI/ML models aimed at finding possible pathogenic combinations of genes (digenic) or variants (biallelic) shared the usage of features related to biological networks or pathways (e.g., the associated pathway of each gene in KEGG or Reactome, network neighbors). Datasets focused on training models for variant or gene prioritization were distinguished by using features linked to predictors of variant pathogenicity at protein level and conservation across the genome of different species. Finally, for the identification of somatic mutations without a matched normal sample, the AI/ML models combined genotype features (e.g., variant allele frequency) with characteristics of the genome region where the variant is located (e.g., GC-content) or sequencing and mapping quality scores (e.g., coverage). Supplementary Table 6 contains further information regarding these types of features and how they were selected.

3.5. Data and code access for reproducibility

When it comes to studies that define ML models, reproducibility is a key factor. Of the selected articles, 16 studies (72.7%) provided access to the data used during the analysis; 3 studies did so only upon data request; 2 did not explicitly declare in the text that data were available; and one stated that data were not available. In terms of the code of AI/ML models, 16 studies (72.7%) had made it publicly available. With respect to the variant discovery approaches, all studies specified the software used for sequence alignment; 21 studies (95.5%) included information about the variant calling step; 17 studies (77.3%) did not mention the use of copy number variations (CNVs) during the analysis, and 3 studies did not state how the variants were annotated. Supplementary Table 7 summarizes data availability and reproducibility information.

4. Discussion

AI/ML involve the use of algorithms to process and gain insights from data with the aim of making predictions or decisions that can be applied to a wide range of fields, including healthcare and genetics. In this systematic review, we have evaluated the latest developments in AI/ML when it comes to rare genetic conditions and examined the ways in which the use of DNA sequencing data can improve their diagnosis. In addition, we have identified some challenges and opportunities for future research in this area.

4.1. Exome sequencing and rare neoplastic diseases as main topics

Although to a lesser extent than in other types of diagnostic methods, such as medical imaging, AI/ML are increasingly being used in the field of RDs [9,18,19]. This trend was also found when focusing only on those studies that use DNA sequencing data to improve the diagnostic process. Through the bibliometric study carried out in this review, and the subsequent manual analyses, we found that exome sequencing was the most prevalent sequencing approach in the field, and that rare neoplastic

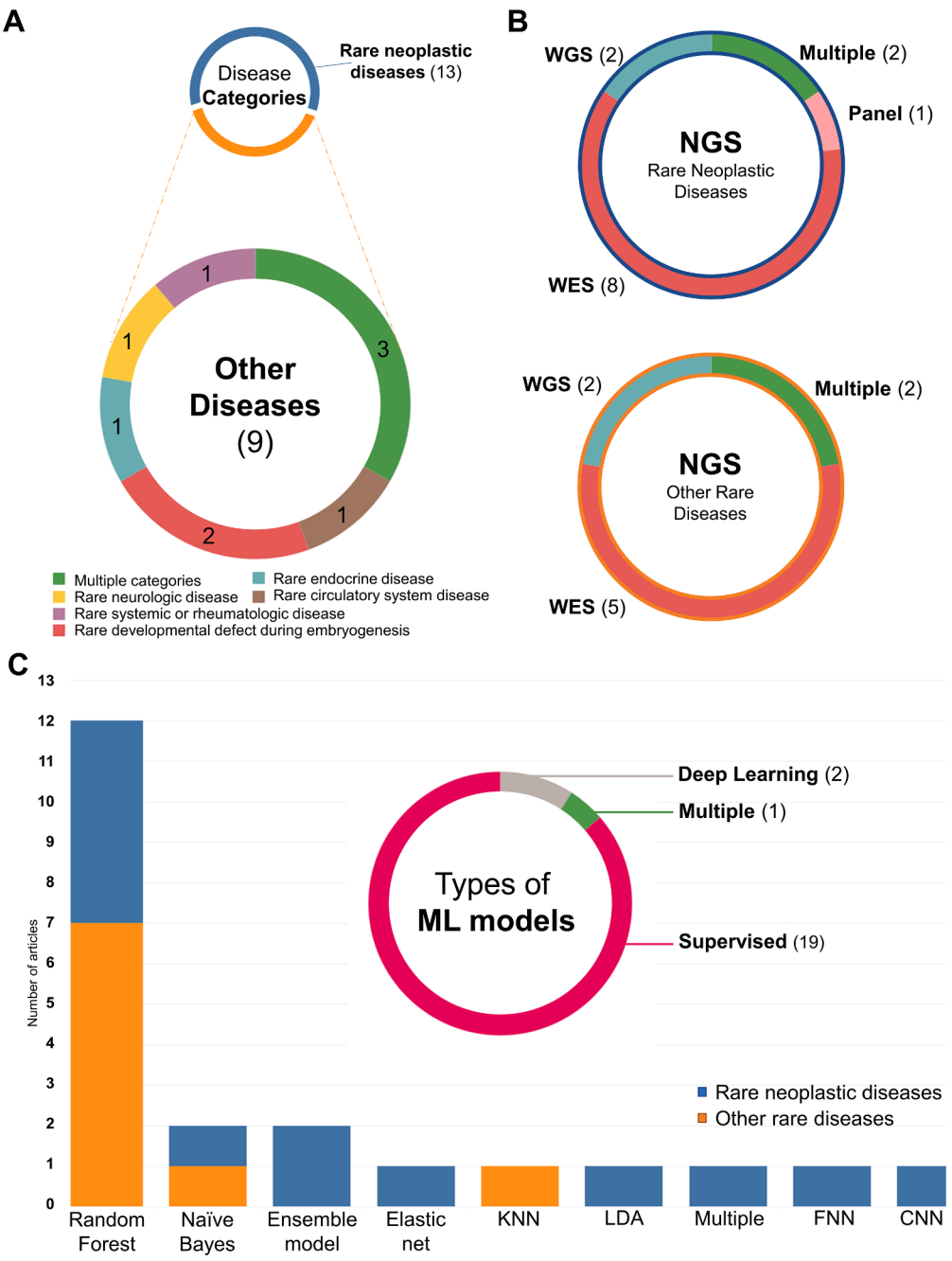

Fig. 3. Methods and areas of application. Panel **A)** displays the distribution of rare diseases identified in selected studies. Panel **B)** shows the next-generation sequencing (NGS) methods used in studies targeting rare neoplastic diseases and other rare diseases. Panel **C)** summarizes the types of machine learning algorithms applied in selected studies. Footer: KNN: K-Nearest Neighbors; LDA: Linear Discriminant Analysis; FNN: Feedforward neural network; CNN: Convolutional Neural Networks.

diseases were the most prevalent clinical scenario. Exome sequencing continues to be a good starting point for the genetic diagnosis of RDs, as it provides a cost-effective and efficient way to identify disease-causing variants [20]. However, depending on the specific rare disease context, genome sequencing may be necessary to provide a complete diagnosis, including the analysis of non-coding variations, CNVs, or chromosomal rearrangements [21,22].

Rare neoplastic diseases generally have a worse diagnosis and higher funding opportunities than other RDs, making them the type of rare disease in which AI/ML are used the most [19,23]. This is also due to the existence of public databases such as TCGA, which allow researchers to access a large amount of genomic data and use AI/ML techniques to identify patterns and make predictions [24]. When we analyzed the data on which these AI/ML models were trained, we saw that many of them (63.6%) were based on sequencing data from external sources, such as TCGA. These studies showed larger sample sizes, but also a greater diversity in sequencing technology characteristics, such as read depth, different length of reads or different sequencing kits and platforms. Mixing sequencing data from different technologies, qualities, and batches may lead to several biases that can influence the variant calling results, affecting in turn the results of downstream analyses, and making difficult to draw accurate conclusions from the data [25]. The precision when taking clinical decisions must be maximized, so these studies must have control over these factors [26]. Different studies have shown how to approach this process [25,27].

4.2. AI/ML algorithms and feature selection in genetic studies

Most of the methods utilized in the selected studies fall into the category of supervised learning (86.7%), with RF being the most common algorithm among them (73.7%). RF algorithm offers a combination of properties that makes it one of the most widely used and suitable

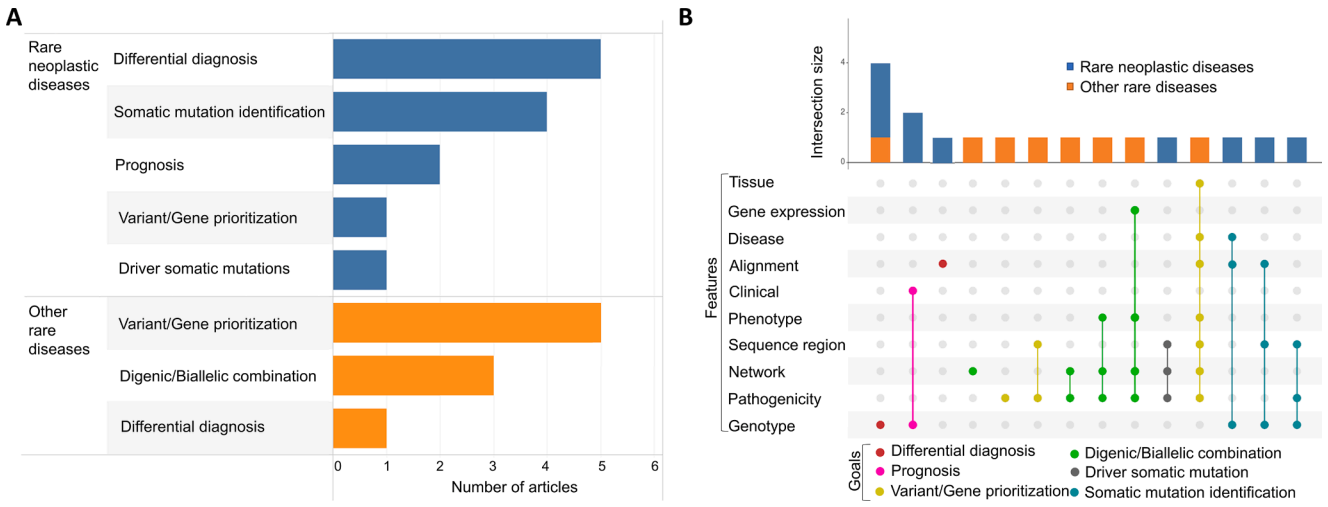

Fig. 4. Objectives and settings of AI/ML models. Panel **A)** displays the goals of AI/ML models in rare neoplastic diseases (blue) and other rare diseases (orange). Panel **B)** contains an upset plot showing the different combinations of features in the training datasets of AI/ML models depending on the objective pursued. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Distinctive features identified in the datasets used for training the ML models of included studies, based on the specific goal being pursued.

Objective AI/ML algorithm	Type of instances	Distinctive dataset feature/s	Example of feature	Use cases (ref)
Stratification/ Differential diagnosis	Patients	Genotype features	Burden value	[48–52]
Prognosis of patients	Patients	Genotype + Clinical features	Burden value + age	[53,54]
Variant/Gene prioritization	Genes/ variants	Pathogenicity features	CADD score	[55–58]
Identification of digenic/biallelic combinations	Pair of genes/ variants	Network features	Number of pathways shared	[59–61]
Identification of somatic mutations	Variants	Genotype + Sequence features	VAF + GC-content	[62–64]

CADD: Combined Annotation Dependent Depletion; VAF: Variant allelic frequency.

algorithms for the study of genetic variants [28,29]. RF combines multiple decision trees (forest) that can handle high-dimensional data, capturing interactions and complex relationships between features by creating random subsets of both, data and features, at each tree. In addition, RF also allows to compute feature importances, which can be used to identify the most relevant features for the prediction task, providing interpretable models [30]. All this makes RF well suited for complex genetic problems and explains its popularity among genetic studies.

Conversely, the application of deep learning and unsupervised algorithms in the analysis of DNA-based NGS data for RDs was limited. While these types of models have shown remarkable success in other domains, such as image or gene expression analyses, their adaptation to genomic data could be a complex task. In general, deep learning algorithms need a large amount of data to be trained, which can be a challenge when dealing with RDs. Although deep learning might be a powerful approach, the use of these methods with small datasets can result in suboptimal models. In addition, due to the complex transformations made on data, the interpretability of decisions made by deep learning models can be difficult, behaving as black boxes. For these and other reasons, researchers may choose other available ML models over deep learning. On the other side, none of the included studies utilized unsupervised algorithms. This may be due to the complexity and genetic heterogeneity of RDs. Unsupervised methods are based on the identification of repeated patterns in the data. However, patients with RDs often carry very rare or novel variants, which makes it difficult to find clusters of patients with the same features.

The structure of the dataset is a fundamental and key aspect of any AI/ML model, as it is the data that the model uses to learn and make predictions. The processes of feature selection and feature engineering can have a substantial effect on the performance of the model; hence, it is essential that the final features possess relevance to the problem at hand [31]. In this systematic review, we have identified the features used by each of the selected studies and found that these features were specific to each of the objectives pursued. This insight can be valuable in understanding the current state of research in the field, and it can serve as a starting point for creating new datasets in future studies.

The results suggested that collapsing or burden methods seem to be crucial for setting up the features of datasets used to train models for the stratification or differential diagnosis of patients. These methods divide the genome into portions (bins or genes) and summarize the information contained in these segments into a burden value, which can be calculated in different ways [32,33]. This approach has shown its usefulness in finding candidate genes in different complex RDs with both genome and exome sequencing data [34–36]. Thus, applied to AI/ML tasks, this process helps to decrease the dimensionality of datasets based on genetic variants by grouping them into one value per gene or bin, which helps to reduce the curse of dimensionality and improve interpretability [37].

On the other hand, models focused on predicting patient prognosis integrate clinical and genomic data to obtain a more complete picture of the patient and assess the risk of disease progression. Previous studies, particularly in cancer, have shown how this integration of data provides a more comprehensive and accurate assessment of patient outcome [38,39]. Alternatively, models aimed at predicting possible pathogenic combinations of genes use features that summarize the association of these genes with the biological pathways in which they participate. The use of these features is supported by the fact that digenic diseases are usually caused by variants in genes that are functionally related and have a common pathway [40,41].

4.3. Future challenges

From the results of this review, we identified some challenges that need to be addressed in future studies. When we analyzed the type of genomic data used to train the AI/ML models reviewed, we realized that

most of them (77.3%) were based exclusively on single nucleotide variants or short indels, not including the analysis of CNVs. CNVs are a significant source of genetic diversity in humans that has remained understudied due to the difficulty of detection. However, today there are different algorithms for CNV detection that simplify the task considerably, as well as guidelines that help us to interpret them [42,43]. This allows the possibility of evaluating its effect on the pathogenesis and outcome of RD. On the other hand, when we examine the goals pursued in the analysis of neoplastic RDs, we can see that the differential diagnosis or stratification of patients stands out above the other objectives. This is totally different in other RDs, where, in fact, this objective is the least pursued of the 3 objectives identified, and, therefore, a field where the contribution of genetic variation to the phenotype is not well understood. The use of AI/ML algorithms on rare disease sequencing data can support the identification of novel genetic interactions, uncovering patterns and relationships that may not be immediately apparent, and providing a better understanding of the regulatory mechanisms mediated by these variants in the phenotype. The use of unsupervised methods would be a possible first approach to achieve the objective of identifying clusters of patients according to their genetic background [44].

4.4. Limitations

Our review is limited by the design of the systematic search and the exclusion of purely methodological articles, focusing only on those studies that applied machine learning methods to data retrieved from patients with a rare disease. Because of the limited number of studies available on the topic, and although it has been studied, articles have not been discarded because of the quality of the journal in which they were published (i.e., JIF), and this may have influenced, in some way, the results of this review. In addition, to reduce variability in study methodology and facilitate the analysis, we have only focused on those studies using DNA-based sequencing, not including other NGS methodologies such as RNA-seq, which are widely used in conjunction with AI/ML methodologies [45–47]. Finally, by restricting the literature search to those articles whose title, abstract or main text includes the terms “artificial intelligence”, “machine learning” or “deep learning”, we may have missed some studies that used statistical models to make predictions and did not use any of these terms.

5. Conclusions

We have conducted a systematic review of ML algorithms to the diagnosis of RDs using DNA-based sequencing data, providing an overview of the current state of the field and the potential of these methods to improve diagnostic accuracy. Exome sequencing is the most widely used sequencing technology and rare neoplastic diseases are the most common disease scenario. On the other hand, the goals of AI/ML algorithms in RDs using sequencing data are broad, ranging from patient stratification to the identification of possible pathogenic combinations of variants. However, we found common patterns in these goals when configuring the datasets with which these models are trained, identifying key features for each of the objectives. Finally, we identified possible future challenges, such as the use of CNV to train the AI/ML models, or the application of AI/ML for the stratification of patients with non-neoplastic RDs. Thus, this systematic review can be used as a reference for further studies, supporting the development of future ML models in the diagnosis of rare genetic diseases.

Fundings

JALE has received funds from Instituto de Salud Carlos III (Grant# PI20-1126), CIBERER (Grant# PIT21_GCV21), Andalusian University, Research and Innovation Department (PY20-00303, EPIMEN), Andalusian Health Department (Grant# PI027-2020), Asociación Síndrome de Meniere España (ASMES) and Meniere’s Society, UK. PRNV is supported by PY20-00303 Grant (EPIMEN). AMPP is a PhD student in the Biomedicine Program at Universidad de Granada and his salary was supported by Andalusian University, Research and Innovation Department (Grant# PREDOC2021/00343).

Declaration of Competing Interest

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

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104429.

References

[1] T. Richter, S. Nestler-Parr, R. Babela, et al., Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group, Value Health. 18 (6) (2015) 906–914, https://doi.org/10.1016/j.jval.2015.05.008.
[2] Orphadata: Free access data from Orphanet. © INSERM 1999. Available on http://www.orphadata.org. Data version (XML data version).
[3] S. Nguengang Wakap, D.M. Lambert, A. Olry, et al., Estimating cumulative point prevalence of : analysis of the Orphanet database, Eur. J. Hum. Genet. 28 (2) (2020) 165–173, https://doi.org/10.1038/s41431-019-0508-0.
[4] 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med*. 2021;385(20):1868-1880. 10.1056/NEJMoa2035790.
[5] A.L. Wise, T.A. Manolio, G.A. Mensah, et al., Genomic Medicine for Undiagnosed Diseases, Lancet Lond Engl. 394 (10197) (2019) 533–540, https://doi.org/10.1016/S0140-6736(19)31274-7.
[6] M. Vinkšel, K. Writzl, A. Maver, B. Peterlin, Improving diagnostics of rare genetic diseases with NGS approaches, J. Community Genet. 12 (2) (2021) 247–256, https://doi.org/10.1007/s12687-020-00500-5.
[7] Dai P, Honda A, Ewans L, et al. Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis. Genet Med. Published online May 14, 2022. 10.1016/j.gim.2022.04.021.
[8] E. Routhier, J. Mozziconacci, Genomics enters the deep learning era, PeerJ. 10 (2022) e13613.
[9] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, S. Thun, The use of machine learning in rare diseases: a scoping review, Orphanet J Rare Dis. 15 (1) (2020) 145, https://doi.org/10.1186/s13023-020-01424-6.
[10] S.T. Setty, M.P. Scott-Boyer, T. Cuppens, A. Droit, New Developments and Possibilities in Reanalysis and Reinterpretation of Whole Exome Sequencing Datasets for Unsolved Rare Diseases Using Machine Learning Approaches, Int. J. Mol. Sci. 23 (12) (2022) 6792, https://doi.org/10.3390/ijms23126792.
[11] A.S.A. Cohen, E.G. Farrow, A.T. Abdelmoity, et al., Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes, Genet. Med. Off J. Am. Coll. Med. Genet. 24 (6) (2022) 1336–1348, https://doi.org/10.1016/j.gim.2022.02.007.
[12] A. Okazaki, J. Ott, Machine learning approaches to explore digenic inheritance. *Trends Genet TIG*, Published online May 14 S0168–9525 (22) (2022) 00105–00106, https://doi.org/10.1016/j.tig.2022.04.009.
[13] Brasil S, Pascoal C, Francisco R, dos Reis Ferreira V, A. Videira P, Valadão G. Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter? *Genes*. 2019;10 (12):978. 10.3390/genes10120978.
[14] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332, https://doi.org/10.1038/nrg3920.
[15] M.J. Page, J.E. McKenzie, P.M. Bossuyt, et al., The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, PLOS Med. 18 (3) (2021) e1003583.
[16] N.J. van Eck, L. Waltman, Software survey: VOSviewer, a computer program for bibliometric mapping, Scientometrics. 84 (2) (2010) 523–538, https://doi.org/10.1007/s11192-009-0146-3.
[17] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, UpSet: Visualization of Intersecting Sets, IEEE Trans. Vis. Comput. Graph. 20 (12) (2014) 1983–1992, https://doi.org/10.1109/TVCG.2014.2346248.
[18] Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health*. 2020;2(9):e486-e488. 10.1016/S2589-7500(20)30160-6.
[19] J. Lee, C. Liu, J. Kim, et al., Deep learning for rare disease: A scoping review, J. Biomed. Inform. 135 (2022), 104227, https://doi.org/10.1016/j.jbi.2022.104227.
[20] J. Klau, R. Abou Jamra, M. Radtke, et al., Exome first approach to reduce diagnostic costs and time – retrospective analysis of 111 individuals with rare

neurodevelopmental disorders, Eur. J. Hum. Genet. 30 (1) (2022) 117–125, https://doi.org/10.1038/s41431-021-00981-z.
[21] S. Marwaha, J.W. Knowles, E.A. Ashley, A guide for the diagnosis of rare and undiagnosed disease: beyond the exome, Genome. Med. 14 (1) (2022) 23, https://doi.org/10.1186/s13073-022-01026-w.
[22] E. Souche, S. Beltran, E. Brosens, et al., Recommendations for whole genome sequencing in diagnostics for rare diseases, Eur. J. Hum. Genet. 30 (9) (2022) 1017–1021, https://doi.org/10.1038/s41431-022-01113-x.
[23] Z. Dlamini, F.Z. Francies, R. Hull, R. Marima, Artificial intelligence (AI) and big data in cancer and precision oncology, Comput. Struct. Biotechnol. J. 18 (2020) 2300–2311, https://doi.org/10.1016/j.csbj.2020.08.019.
[24] J.N. Weinstein, E.A. Collisson, G.B. Mills, et al., The Cancer Genome Atlas Pan-Cancer Analysis Project, Nat. Genet. 45 (10) (2013) 1113–1120, https://doi.org/10.1038/ng.2764.
[25] R. De-Kayne, D. Frei, R. Greenway, S.L. Mendes, C. Retel, P.G.D. Feulner, Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets, Mol. Ecol. Resour. 21 (3) (2021) 653–660, https://doi.org/10.1111/1755-0998.13309.
[26] R.L. Goldfeder, J.R. Priest, J.M. Zook, et al., Medical implications of technical accuracy in genome sequencing, Genome Med. 8 (1) (2016) 24, https://doi.org/10.1186/s13073-016-0269-0.
[27] K. Ellrott, M.H. Bailey, G. Saksena, et al., Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines, Cell Syst. 6 (3) (2018) 271–281.e7, https://doi.org/10.1016/j.cels.2018.03.002.
[28] B.A. Goldstein, E.C. Polley, F.B.S. Briggs, Random Forests for Genetic Association Studies, Stat. Appl. Genet. Mol. Biol. 10 (1) (2011) 32, https://doi.org/10.2202/1544-6115.1691.
[29] X. Chen, H. Ishwaran, Random Forests for Genomic Data Analysis, Genomics. 99 (6) (2012) 323–329, https://doi.org/10.1016/j.ygeno.2012.04.003.
[30] L. Breiman, Random Forests, Mach Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[31] Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinforma.* 2022;2. Accessed January 21, 2023. https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312.
[32] Dering C, König IR, Ramsey LB, Relling MV, Yang W, Ziegler A. A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. *Front Genet.* 2014;5. Accessed January 21, 2023. https://www.frontiersin.org/articles/10.3389/fgene.2014.00323.
[33] D.L. Nicolae, Association Tests for Rare Variants, Annu. Rev. Genom. Hum. Genet. 17 (1) (2016) 117–130, https://doi.org/10.1146/annurev-genom-083115-022609.
[34] P. Roman-Naranjo, A. Gallego-Martinez, A. Soto-Varela, et al., Burden of Rare Variants in the OTOG Gene in Familial Meniere's Disease, Ear Hear. 41 (6) (2020) 1598–1605, https://doi.org/10.1097/AUD.0000000000000878.
[35] A.A. Dilliott, A. Abdelhady, K.M. Sunderland, et al., Contribution of rare variant associations to neurodegenerative disease presentation, NPJ Genomic Med. 6 (2021) 80, https://doi.org/10.1038/s41525-021-00243-3.
[36] Lin J, Li C, Cui Y, et al. Rare variants in IMPDH2 cause autosomal dominant dystonia in Chinese population. *J Neurol.* Published online January 17, 2023. 10.1007/s00415-023-11564-x.
[37] N. Altman, M. Krzywinski, The curse(s) of dimensionality, Nat. Methods. 15 (6) (2018) 399–400, https://doi.org/10.1038/s41592-018-0019-x.
[38] B. Lobato-Delgado, B. Priego-Torres, D. Sanchez-Morillo, Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis, Cancers. 14 (13) (2022) 3215, https://doi.org/10.3390/cancers14133215.
[39] J. Gonzalez-Bosquet, S. Gabrilovich, M.E. McDonald, et al., Integration of Genomic and Clinical Retrospective Data to Predict Endometrioid Endometrial Cancer Recurrence, Int. J. Mol. Sci. 23 (24) (2022) 16014, https://doi.org/10.3390/ijms232416014.
[40] A. Gazzo, D. Raimondi, D. Daneels, et al., Understanding mutational effects in digenic diseases, Nucleic Acids Res. 45 (15) (2017) e140.
[41] A.A. Schäffer, Digenic inheritance in medical genetics, J. Med. Genet. 50 (10) (2013) 641–652, https://doi.org/10.1136/jmedgenet-2013-101713.
[42] E.R. Riggs, E.F. Andersen, A.M. Cherry, et al., Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen), Genet. Med. 22 (2) (2020) 245–257, https://doi.org/10.1038/s41436-019-0686-8.
[43] V. Gordeeva, E. Sharova, G. Arapidi, Progress in Methods for Copy Number Variation Profiling, Int. J. Mol. Sci. 23 (4) (2022) 2143, https://doi.org/10.3390/ijms23042143.
[44] A.O. Basile, M.D. Ritchie, Informatics and Machine Learning to Define the Phenotype, Expert Rev. Mol. Diagn. 18 (3) (2018) 219–226, https://doi.org/10.1080/14737159.2018.1439380.
[45] L. Wang, Y. Xi, S. Sung, H. Qiao, RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes, BMC Genom. 19 (1) (2018) 546, https://doi.org/10.1186/s12864-018-4932-2.
[46] C. Gunavathi, K. Sivasubramanian, P. Keerthika, C. Paramasivam, A review on convolutional neural network based deep learning methods in gene expression data for disease diagnosis, Mater Today Proc. 45 (2021) 2282–2285, https://doi.org/10.1016/j.matpr.2020.10.263.
[47] W.A. Figgett, K. Monaghan, M. Ng, et al., Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus, Clin. Transl. Immunol. 8 (12) (2019) e01093.
[48] L. Parida, C. Haferlach, K. Rhrissorrakrai, et al., Dark-matter matters: Discriminating subtle blood cancers using the darkest DNA, PLoS Comput. Biol. 15 (8) (2019) e1007332.
[49] S. Parvandeh, L.A. Donehower, K. Panagiotis, et al., EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants, Nucleic Acids Res. 50 (12) (2022) e70.
[50] P. Peneder, A.M. Stütz, D. Surdez, et al., Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden, Nat. Commun. 12 (1) (2021) 3230, https://doi.org/10.1038/s41467-021-23445-w.
[51] Y. Li, Y. Luo, Performance-weighted-voting model: An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation, Quant. Biol. Beijing China. 8 (4) (2020) 347–358, https://doi.org/10.1007/s40484-020-0226-1.
[52] Aguiar-Pulido V, Wolujewicz P, Martinez-Fundichely A, et al. Systems biology analysis of human genomes points to key pathways conferring spina bifida risk. *Proc Natl Acad Sci U S A.* 2021;118(51):e2106844118. 10.1073/pnas.2106844118.
[53] M.A. Chaix, N. Parmar, C. Kinnear, et al., Machine Learning Identifies Clinical and Genetic Factors Associated With Anthracycline Cardiotoxicity in Pediatric Cancer Survivors, JACC CardioOncol. 2 (5) (2020) 690–706, https://doi.org/10.1016/j.jaccao.2020.11.004.
[54] M.G. Zauderer, A. Martin, J. Egger, et al., The use of a next-generation sequencing-derived machine-learning risk-prediction model (OncoCast-MPM) for malignant pleural mesothelioma: a retrospective study, Lancet Digit Health. 3 (9) (2021) e565–e576, https://doi.org/10.1016/S2589-7500(21)00104-7.
[55] H. Carter, C. Douville, P.D. Stenson, D.N. Cooper, R. Karchin, Identifying Mendelian disease genes with the variant effect scoring tool, BMC Genom. 14 (Suppl 3) (2013) S3, https://doi.org/10.1186/1471-2164-14-S3-S3.
[56] D. Vitsios, S. Petrovski, Mantis-ml: Disease-Agnostic Gene Prioritization from High-Throughput Genomic Screens by Stochastic Semi-supervised Learning, Am. J. Hum. Genet. 106 (5) (2020) 659–678, https://doi.org/10.1016/j.ajhg.2020.03.012.
[57] A.R. Majithia, B. Tsuda, M. Agostini, et al., Prospective functional classification of all possible missense variants in PPARG, Nat. Genet. 48 (12) (2016) 1570–1575, https://doi.org/10.1038/ng.3700.
[58] K.J. Carss, A.A. Baranowska, J. Armisen, et al., Spontaneous Coronary Artery Dissection: Insights on Rare Genetic Variation From Genome Sequencing, Circ. Genomic Precis. Med. 13 (6) (2020) e003030.
[59] N.A. Davis, C.A. Lareau, B.C. White, et al., Encore: Genetic Association Interaction Network centrality pipeline and application to SLE exome data, Genet. Epidemiol. 37 (6) (2013) 614–621, https://doi.org/10.1002/gepi.21739.
[60] S. Mukherjee, J.D. Cogan, J.H. Newman, et al., Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network, Am. J. Hum. Genet. 108 (10) (2021) 1946–1963, https://doi.org/10.1016/j.ajhg.2021.08.010.
[61] M. Laan, L. Kasak, K. Timinskas, et al., NR5A1 c.991-1G > C splice-site variant causes familial 46, XY partial gonadal dysgenesis with incomplete penetrance, Clin. Endocrinol. (Oxf). 94 (4) (2021) 656–666, https://doi.org/10.1111/cen.14381.
[62] B.J. Ainscough, E.K. Barnell, P. Ronning, et al., A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data, Nat. Genet. 50 (12) (2018) 1735–1743, https://doi.org/10.1038/s41588-018-0257-y.
[63] D.E. Wood, J.R. White, A. Georgiadis et al. A machine learning approach for somatic mutation discovery Sci. Transl. Med. 10 457 2018 eaar7939 10.1126/scitranslmed.aar7939.
[64] I. Kalatskaya, Q.M. Trinh, M. Spears, J.D. McPherson, J.M.S. Bartlett, L. Stein, ISOWN: accurate somatic mutation identification in the absence of normal tissue controls, Genome Med. 9 (1) (2017) 59, https://doi.org/10.1186/s13073-017-0446-9.