

Chapter 6: Deep learning algorithms in rare disease identification: Enhancing diagnostic yield from genetic testing

6.1 Introduction

Although not specifically acknowledged as such, approximately 80% of rare diseases are genetic and may have a great impact on either patient or family. Rare disease, often referred to as an orphan disease, may have yet unclear cause and could have various signs and symptoms in its different time course. Sometimes, it takes a longer time for the right diagnosis to be made due to the rareness of the condition, given that most physicians are often unfamiliar with rare diseases. This leads to the misguidance on therapy selection. Genetic testing is usually employed to support the diagnosis, or to find out some part of the trigger, by assisting the specialist in narrowing down the diagnostic possibilities. However, the diagnostic yield of the genetic test is not always 100%. There are some challenges in diagnosing the rare disease, namely the variability in presenting symptoms, the scarcity of physicians who are familiar with rare conditions, and the lack of access to the proper medical facilities required for the necessary diagnostic tests. Regarding the genetic causation of certain conditions, for example, genetic predispositions to complex diseases, genetic influences on drug response, or different genetic profiles between patients, the rare diseases can perhaps now be better diagnosed. More widespread availability of genetic testing has made it easier to identify a better treatment. Various genetic aspects could be obtained from the patient's genetic data, ranging from the simplest one, such as the type of variant, and the type of effects brought by the variant, up to more complex profiles, like the pathway involved or the possible associations with other diseases. Traditional algorithms, however, failed to take into account the complexity of the genetic profile, and instead mostly only involved a simplistic binary decision-making process. To enhance the diagnostic yield, more advanced algorithms may be required to assist the specialist in exploiting the genetic profile. In recent years, the rapid technical advances in artificial intelligence have made

it possible to use deep learning techniques to improve diagnostic capabilities. This has resulted in saving some lives for patients whose disease remains undiagnosed for years or until it is too late. Moreover, the utilization of these algorithms may also help reduce the stress experienced by both patient and family and potentially lower the cost burden.

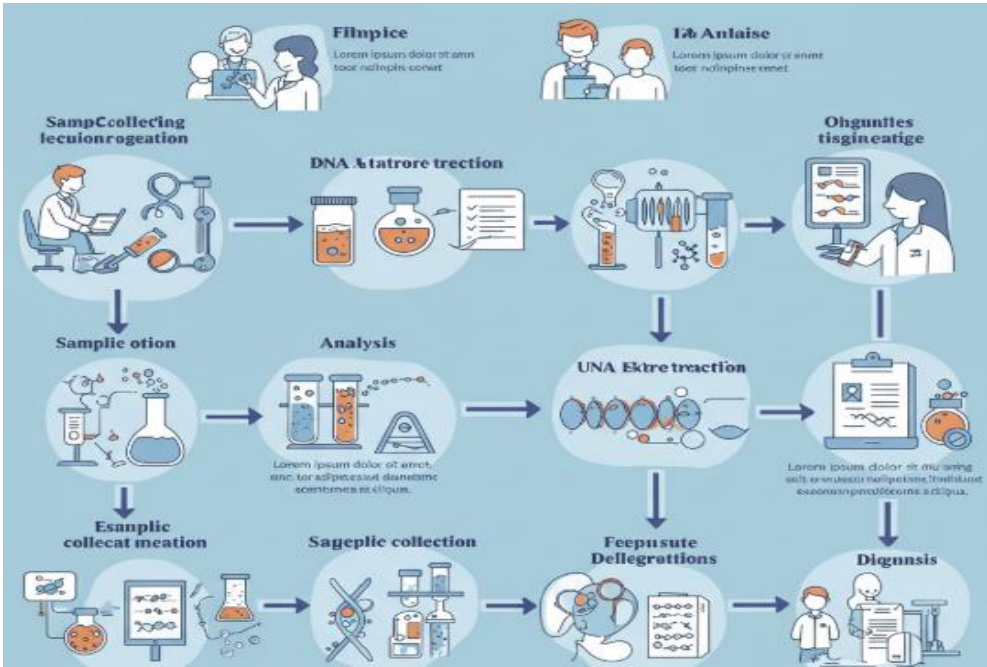


Fig 6.1: Genetic Testing in Rare Diseases

6.2. Background on Rare Diseases

A rare disease, as defined by the United States Orphan Drug Act, is one that affects fewer than 200,000 individuals in the country. In other words, approximately 1 out of 2000 people are diagnosed with a rare disease. Since the development of personalized medicine, rare disease prevalence is expected to be raised from the low 1% to around 7%. This raises the estimated lower limit of the prevalence of rare diseases from 75 to around 500 per 100,000 population in Western, industrialized countries. Currently, on the RD-Connect platform, around 8400 rare diseases are recognized. Rare diseases can be classified into a range of countries. Moreover, the spectrum of rare diseases is wide; they include genetic, infectious, and autoimmune diseases, along with a long tail of other conditions. Particularly, the population comprises undiagnosed and complex cases due to, for example, overlapping phenotypes and multiple etiologies. Nonetheless, because of the relatively large count of diseases, and there being very few cases per disease, hand research of possible diseases and treatments are limited. New diagnostic methods can

help reduce the time from symptom onset to the first treatment as well as reduce the number of misdiagnoses. Nor would people no longer face years of searching for an explanation for their symptoms, with limited access to the appropriate services, i.e., knowledge and expertise, such as doctors, tests, support groups and equipment. Despite the urgency for improvement, diagnostic testing is less efficient for rare diseases. Many patients either never obtain a diagnosis or do so after a delay of more than five years post initial consultation. Even in cases with a diagnosis, patients still have a 25% probability of misdiagnosis, and this may result in inappropriate treatments, while common symptoms remain untreated. It is imperative that the impact of rare diseases is re-engineered. One particular need is for new methods that would better utilize the information contained in the large volume of patient data generated by existing healthcare systems. It is believed that deep learning algorithms that incorporate broad medical expertise are particularly promising.

6.2.1. Definition and Classification

Rare diseases are characterized as conditions that affect only a small fraction of the population. For the majority of nations worldwide, the number cited most frequently is a prevalence threshold of 1 per 2000 individuals. Alongside the exact prevalence, this definition is usually accompanied by a range of additional criteria. Rare diseases can have inherent risks which cause altered biological, developmental, or cognitive mechanisms for survival. Careful collection and interpretation of data might be necessary to distinguish a rare disease from common diseases manifesting through common variations, or diseases of environmental cause. Moreover, a rare disease diagnosis for one patient does not necessarily mean a rare diagnosis for other patients with similar characteristics. As in the case of Bochdalek diaphragmatic hernia, Dolichocephaly and various other conditions, many disorders adversely but indirectly affect their body mechanisms. Epidemiological data about these diseases have been collected for years, particularly concerning their phenotypical protuberances. Nevertheless, such diseases are most challenging from a hidden data perspective because no obvious genotypes are available and they are inconsistently listed on death certificates.

Rare disease classification is a means of categorizing diseases into related entities, enabling a distinction between a wide range of different diseases. These schemes help to extract demographics and evidence along the unique symptoms and illness of the population. Broad classifications also influence health decision-making, integrating data about several patients to establish a comprehensive likelihood of exposure or best procedures for treating similar patients. Rare disease classification can be done based on etiology, heredity and manifestation. Upon known information, rare diseases are often

classified based on the causality of the conditions. This can be genetic or acquired, and may comprise diseases for which no origin is identified and mapped (though complex diseases usually feature a known etiology). Meanwhile, several rare diseases are not wholly genetic or sporadic in their etiology, with a mixture of biological, environmental, and societal factors adding to the danger. Scientists classify these diseases separately, often with further sub-categorizations specifying the relative contribution of the elements in a multifactorial simulation. Finally, some rare diseases are wholly ignored due to the multiple insufficiencies of medical science, understanding of environmental factors, and general understanding of the disease. Consulting the sum of all the underlying biological components of a disease will be extensive and take years of research (Koppolu et al., 2022; Kaulwar, 2023)

6.3. Overview of Genetic Testing

Genetic testing is defined as the testing of human tissues and/or bodily fluids in order to identify changes or mutations in a patient's genome. These mutations are then assessed to determine the patient's risk of developing certain inherited diseases. The purpose of genetic testing is to provide clinical diagnosis and prognosis to patients with diseases that are caused by alterations in specific genes. The medical significance of genetic testing has been increasing since the completion of the human genome project, and it plays an essential role in the diagnosis and understanding of a wide range of phenotypes. Clinical application of genetic testing may include confirmation of a suspected diagnosis, identification of at-risk family members for disease predisposition, guidance for specific treatment, etc. Genetic testing has been recognized as an indispensable tool in the diagnosis and understanding of rare monogenic diseases. This has long been an area of effort by medical researchers in the hope of identifying new drugs, vaccines, and therapeutic responses.

Over the years, the effort has obtained limited success with merely 5% of more than 7,000 rare diseases with associated genotype information. The main methods of genetic testing include Sanger sequencing, custom arrays, whole genome sequencing (WGS), and whole exome sequencing (WES). Sanger sequencing and custom arrays are used in DNA sequencing and genotyping of certain regions of the genome, while WGS and WES refer to sequencing the entire genome and the exons of genes. The vast majority of genes encode proteins that form the building blocks of cells. Exons are the parts of a gene that contain the coding information for proteins. Alterations can and frequently do occur in the genome of healthy individuals, collectively referred to as single-nucleotide polymorphism (SNP), insertion, or deletion. A change can interfere with the DNA replication process or lead to disruption of the normal coding sequence of genes resulting in different diseases. Mutations have been categorized broadly into three classes, where,

in general, the genetic sequence: gains, losses, or modifies gene function. Consistent with this hypothesis, there have appeared matching studies with gene level and protein diseases. The translation of the genetic code, perhaps not surprisingly, can cause either gain-of-function or loss-of-function of the translated species. Also, frequently, multiple mutations are needed to produce a disease state.

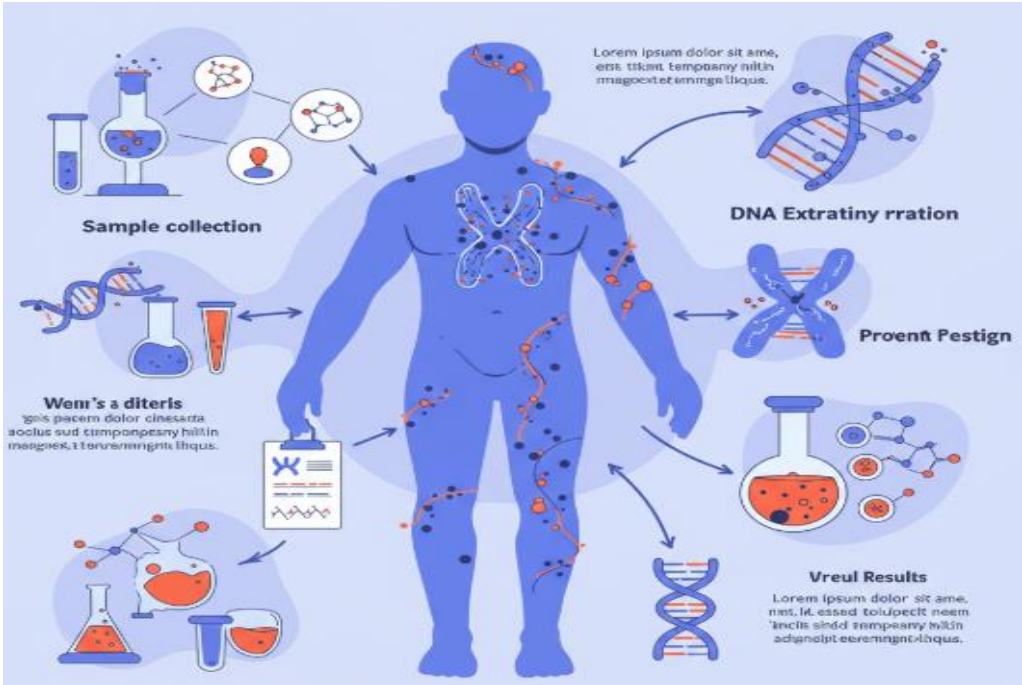


Fig 6.2: Overview of Genetic Testing

6.3.1. Types of Genetic Tests

Genetic testing has become an essential component in the diagnosis and management of a wide range of clinical conditions. The advent of new technologies, such as next-generation sequencing, has enabled the generation of high-throughput genomic data, thus offering hope for clinically relevant findings. Advances in genetic tests include diagnostic tests to detect genetic mutations that give rise to a clinical phenotype, carrier tests to identify genetic lesions in individuals who are asymptomatic but still have an increased risk of transmitting the disease, prenatal tests of pregnancies in affected families, and newborn screening. In the clinical-genetic field for rare diseases with ultra low prevalence, the development of these tests that complement deep learning models have expedited numerous relevant, eligible, and network assistance-based findings for accurate rare disease diagnosis. Despite the promise of high-throughput screening technologies, the massive expansion of genetic testing has led to an unmet interpretation

bottleneck in a myriad of fields such as cancer genomics, pharmacogenomics, and rare genetic diseases (Singireddy, J., 2024; Singireddy, S. et al., 2024).

Advancements in genetic testing technologies have made it possible to use a variety of tests to identify genetic variations linked to rare diseases. Genetic tests can be developed for a variety of purposes to identify genetic mutations responsible for phenotypes. It is possible to note a subset of DNA alterations that have a strong causal relationship with the condition and accurately screen the variant on a randomized search of the fantastic roster at the right amount of times. If screening is conducted exhaustively hard, the number of candidate genes (or variants of standing) will still be returned for any phenotype and unaccomplished. With declining costs, the widespread adoption of the approach, particularly in population-based strategies, has fuelled new development. Deep-learning models have had success in a variety of applications across various biomedically-related tasks. Efforts to identify rare phenotype-associated variants have largely been focused on those that are PTVs, as they can be linked to a change in the stability, accumulation, or expression of messenger RNA; they are also more efficiently screened, because they typify a majority of the disease-causing mutations. This is particularly advantageous for single-gene disorders amenable to treatment that is efficacious only when initiated early in the disease course. Such a model does not currently exist in the clinical perception field. However, as genetic data in the human population accumulates, concerns about the phenotypic meaning of genetic variants will increase and machine learning methods will need to be developed to address these concerns to a greater degree.

New gene variants likely underlie Mendelian diseases with an unknown etiology each year. However, the diagnostic yield continues to be less than 25% on average in studies involving sequencing of the entire genome or exome. Data collected across 5 research initiatives worldwide showed that a modest 7.7% of unsolved probands could benefit from appropriate treatment when mutations in the candidate gene were finally identified. This compares to only 1% in the entire unsolved sys-4 cohort. Ongoing analyses on over 42,000 novel gene-disease relationships discovered through 6,454 studies have shown that across 5 major disease groups, a typical Mendelian disease now traces to nearly 20 genes. Deep-learning algorithms have achieved outstanding progress in a range of biomedical applications. Inaccuracies in the genetic testing procedure led to grievous tragedy in the discovery and decline of protein-coding genes. Efforts to seal the candidate cpDNA (or novel nuclear gene) were made considerably harder by the involvement of two female offspring as a sibling pair. Profiling the entire shared ups of parental and offspring mitochondrial variants using either the approach or meta-analysis would have immediately narrowed down the genomic scope of interpretation. In data from a suitably powered deep learning algorithm, it would be within the reach of a candidate discovery screen by amplification akin to those used in previous studies on the

genetic transmission of mitochondrial disease. However, similar knowledge of the actual mutation's prior probability could easily have caused a review of the tragic loss of time.

6.4. Deep Learning Fundamentals

This part aims to introduce the basic and new concepts such as deep learning methods on medical data and the motivation. In addition, it explains the progression of machine learning methods and the foundation of deep learning. Ahead of the hope of support, a reader can easily understand the following parts along with those descriptions.

The ability of computers to autonomously learn patterns from data has revolutionized many aspects of life as well as clinical practice, notably in diagnostic testing. Deep learning is a subset of machine learning methods that has demonstrated particular promise in medical diagnostics of rare diseases. Most deep learning methods build upon neural networks, which are computer algorithms inspired by the arrangement of connections in biological systems. Traditional neural networks are relatively shallow, consisting of 1-2 “hidden layers” of neurons that modulate the relationship between input data and target outputs. Deep learning methods, in contrast, consist of more complex neural network architectures with many layers that capture intricate nuances in the data more effectively.

In a clinical context, deep learning is often used for either supervised or unsupervised learning. In supervised learning, the input data and target values are used to train a model to predict those targets from a new data set. In the case of diagnosing rarity disease patients, a very large collection of features on many examples from different patients is used to train the algorithm. The trained algorithm can later be applied to a new patient's data to predict a diagnosis. Because deep learning models are able to handle very large datasets very efficiently, this approach is particularly well suited to the unique challenges of rare disease identification tasks in the healthcare sector.

6.4.1. Introduction to Deep Learning

Deep learning is a significant area of machine learning and is based on artificial neural networks, i.e., a set of algorithms designed to detect patterns – in this case, patterns from genetic data. The breakthrough of deep learning is primarily due to big data, increased computational power, and the innovation of some techniques. Neurons are the building blocks of artificial neural networks. Each neuron is connected with other neurons through synapses. Every connection is associated with a weight, signifying the importance of the input. Neurons sum up all the inputs and produce one output. Generally, neural networks have an input layer, one or more hidden layers, and output

layers. Hidden layers boost the ability of neural networks to learn the representations of input feature spaces. When there are two or more hidden layers, the network is known as a deep neural network. Complex series of transformations are made by deep neural networks on input data to learn and represent the underlying distribution effectively, which greatly distinguishes them from traditional predictive models. Capacity is boosted by adding more hidden layers, and the network can automatically extract hierarchical features through these hidden layers. Deep learning architectures can gain insights at different levels of abstraction or timescales from data. The feature space of data will be transformed with the mechanisms integrated in the network before it is delivered into each layer. In this way, traditional methods heavily rely on a priori knowledge of data, whereas deep learning architectures can automatically learn feature spaces and data representation. According to the network structure and the mode of information propagations, there are many specialized types of neural networks. Feedforward networks are the simplest type of neural networks where information is only propagated forward. Recurrent networks are capable of managing time series data generated in sequencing times, which can be interpreted as cyclic graphs.

6.5. Deep Learning Algorithms in Medical Diagnostics

With the increasing availability of medical data, traditional analytical methods often reach their limits when analyzing large datasets with a high level of complexity. Deep learning algorithms provide new opportunities to detect patterns in such datasets, which are hardly recognizable by humans or traditional analytical methods. Essentially, deep learning algorithms learn to model high-dimensional relationships from data, often without the need for prior knowledge or a precisely defined algorithm. The applications of deep learning algorithms are manifold, especially in the field of medical diagnostics. They are classified into two categories: First, there are classification algorithms. They are trained to assign data to certain groups. On the other hand, regression algorithms are used, which can determine certain continuous values on the basis of the data. When used correctly, both types of algorithms have the potential to increase diagnostic accuracy by leveraging large amounts of medical data. In recent years, many scientific studies have shown that the performance of deep learning algorithms in several tasks even exceeds the capabilities of experts. The results obtained in healthcare are particularly promising. Time-consuming diagnostic processes can be automated by deep learning algorithms, thereby improving both diagnostic yield and patient care. Despite this, experts continue to face challenges in transferring deep learning research results from academia to the healthcare sector. Special challenges in the medical field relate to the quality of the data or the interpretability of the algorithms used. Challenges pertaining to the application of CNNs to rare disease diagnostics are also present. They include the requirement of vast datasets, as well as a need for flawless data labelling. Moreover, a considerable number

Convolutional Neural Networks (CNNs) are an application of neural networks utilized primarily for processing multidimensional data, such as images. They usually consist of multiple layers modulating primarily convolutions and pooling. Owing to the specific structure of convolutional layers, CNNs are able to identify features within data. These features are not predefined as opposed to subsequent image processing technique applications. They consist of patterns characteristic for a specific set of training data. CNNs have shown significant performance in medical imaging diagnostics and have been applied, among others, in diabetic retinopathy detection, pathological pulmonary nodule detection in chest x-ray files, and breast cancer diagnostics based on mammography results.

6.6. Application of Deep Learning in Rare Disease Identification

In this rapidly evolving scientific and technological landscape, this paper and the interactive diagnostic platform presented within it aim to assist academics and clinicians in understanding the practical application and translation of deep learning algorithms in the time-efficient identification of rare diseases (RDs) and other conditions of underlying genetic origin. By using genetic and clinical phenotypes, six different types of deep learning approaches can be used to enhance the diagnostic yield of genetic testing panels and other genetic health tests. Genomic and phenotypic sentence encoders are introduced, the former derived from novel sequence-to-sequence architectures for genetic variants and the latter from the modification/rescaling of a publicly available clinical sentence encoder.

There is mounting evidence of the significant contribution of deep learning algorithms to the field of precision medicine, through improving the interpretation of integrated genomic and medically-derived patient information. Affecting a considerable number of people, tens of millions in the US and hundreds of millions globally, RDs can be categorized as a form of precision medicine. Importantly, this type of condition is caused by genetic alterations, so the underlying genetic cause can be identified following the performance of next-generation sequencing technologies to sequentially read off a person's DNA at single-base resolution. Moreover, many RDs have specific or highly specific clinical manifestations, so clinical characteristics may also suggest genetic analyses for suspected patients. In this way, beyond appreciating the advantages of precise treatment decisions taking advantage of genetic and phenotypic information, there is a strong academic interest in understanding the practical impact of different machine learning methodologies on real-world diagnostic yield increment after utilising pan-genomic and, particularly, clinical seq-data. In what follows, at first, as background information necessary to understanding the practical application of learned methodologies and the design and descriptions of the presented interactive tool, a

discussion will be done on the fundamentals and current approaches in the mathematical formulation of the RD identification problem and a review of the relevant literature on the topic.

6.6.1. Data Sources and Preprocessing

Rare diseases, by definition, indicate medical conditions with very low prevalence in the population. However, due to the numerosity of uncommon diseases, it is estimated that up to 300 million people worldwide are at the same time affected by a rare disease. The early identification of rare diseases is still a primary need, but it is often more complex and often involves a longer time span compared to pathologies with higher prevalence. Deep learning-based tools hold great potential to help clinicians. One of the factors that mainly limits the performance of such tools is the reduced size of labelled datasets. Several factors are impeding the generation of these sets, including privacy concerns, complex data access issues, and non-rigorous data governance. Several case studies are discussed that showcase the effectiveness of the designed data sources and methodologies as well as current practice pointers to significantly enhance strategies for sourcing or generating data.

When using deep learning algorithms to train predictive or generative models for rare diseases, it is crucial to provide data sets composed of genetic data, if available, together with clinical and demographic information. Optimal data quality and the balanced representation of different groups will affect the success of training algorithms. Barcode data generated by or commercially available sequencing becomes ubiquitous in research and healthcare. This data is portable and can be codified as images, which is ideal for the application of recent computer vision techniques. However, such data is also challenging for reliable deep learning-based bioinformatics analyses because of its high dimensionality, and also often leads to biased models, thus further limiting its diagnostic value.

Based on a common set of DNA samples, a non-targeted deep learning study is employed here to underline the biases found in the genetic data, and to subsequently discuss their consequences on the analytical results. In addition, on-target sequencing studies are discussed and recommendations are made that might allow to minimize the biases described here while fostering a better understanding of the underlying biology, and eventually realizing the full potential of DNA data to study and diagnose rare diseases. It is highlighted that improved data sourcing and preprocessing can contribute significantly to the success of deep learning-based initiatives. When using effectively preprocessed, shared and commonly cited benchmarks, generated by large datasets, algorithms can achieve F1-scores that are comparable or even exceed those of other healthcare applications. On the other hand, initial explorations with heterogeneous,

unprocessed datasets often lead to models with very poor performance. Therefore, before embarking on deep learning technologies, strict protocols for the collection, curation and sharing of the data tailored to the intended applications should be considered. Addressing all of the above issues typically requires interdisciplinary collaboration between computational, life, social and medical scientists, as well as legal and privacy experts.

6.7. Conclusion

Artificial intelligence (AI) has become increasingly important in improving the diagnostics of knowing rare diseases. For hospitals and private health care providers, the improved diagnostic validity of deep learning algorithms are transforming the genetic testing market. In recent years, the global adoption and integration of AI networks in assorted sectors, most notably in healthcare and life sciences, has expanded swiftly. In the area of genetic disease diagnostics, this has had a substantial impact. With the support of deep learning networks, genetics technicians and medical professionals are able to diagnose illnesses much more precisely and successfully. This is of particular interest when it comes to rare disease diagnosis.

Provided that deep learning methodologies are integrated into the customary diagnostic practice, they repay the variegated difficulties in the genetic diagnostics of rare diseases. A total of 8% of genetic diseases affect people. Since mutations in the DNA occur, the majority of genetic disorders lead to a disease pattern that is handed down from parent to kid. In addition to frequent genetic diseases, there are also uncommon genetic disorders, generally referred to as rare diseases. The number of individuals with rare diseases totals 400 million worldwide, and more than half of rare maladies affect children. The demographically significant quantity demonstrates the demand for deep learning diagnostics in the field of rare disease diagnosis. It is anticipated that because of these circumstances approximately 29 million Americans are affected by a rare genetic disorder. In order to accurately diagnose a rare genetic disease, a next-generation sequencer must be done before a geneticist can examine the results. However, distinguishing between disease and non-disease mutation might be difficult even then.

The final method, the third implementation, develops a deep learn-based model for genetic diagnostics. The NGS cohort is divided by deep neural nets that train groups composed of disease-full and self-trained, insurance-free datathon inspections. By using the model learned over the self-trained data, the clinical geneticist can determine whether a novel NGS positive genetic variant is related to the medical condition of the patient by comparing it with the insurance-free data of the positive sets learned. The method behind the recommendations section to inspect genetic variance and deploy recommendations gradually. There are so many research studies which have been attempted to improve the genomic and genetic methods for the identification of rare diseases.

6.7.1. Future Trends

This paper has examined a few use cases for deep learning applications in the genetic testing process for rare diseases. By analyzing the most successful applications, and the few that are not, future trends in technology, methodology, and challenges to tackle have been identified. Data processing will play a key role in continuing to improve the use of AI in this domain, as the arrival of more organized and refined datasets opens the door to more extensive applications of deep learning. On the clinical side, opportunities to facilitate the delivery of rare disease interpretation services by healthcare providers have been identified. Partnering with patients and increasing their preparedness for DNA testing has the potential to significantly increase the diagnostic yield from the sequenced genetic data. Finally, future efforts are discussed that will be necessary to further investigate how compliance with professional practice guidelines can be better supported and fostered by the use of AI tools. The development of artificial intelligence (AI) takes on a multitude of, often novel abilities. A field of AI, machine learning (ML), further drills down to create deep learning, whose DCNN format is herein succinctly referred to as just deep learning. Beyond improved diagnostics, with the potential of expanded telehealth, this methodology has application to the realization of cutting-edge therapies that had no prior feasible pathway. AI is not constrained by paradigm, permitting reconsideration of how complex diseases might be approached in favor of novel, potentially more viable solutions. For instance, thalidomide-induced disease had been viewed as irreversible due to extensive limb damage; however, AI found promise in epothilone. Historic instances of malign bias underscore the need to prevent a repeat outcome. Emerging AI-based diagnostics of rare disease may soon surpass abilities to maintain a pace with subsequent drug availability. Without advanced intervention, AI has potential to widen the gulf serving as an eventual detriment to afflicted individuals.

References

- Challa, S. R., Challa, K., Lakkarasu, P., Sriram, H. K., & Adusupalli, B. (2024). Strategic Financial Growth: Strengthening Investment Management, Secure Transactions, and Risk Protection in the Digital Era. *Journal of Artificial Intelligence and Big Data Disciplines*, 1(1).
- Suura, S. R. (2024). Generative AI Frameworks for Precision Carrier Screening: Transforming Genetic Testing in Reproductive Health. *Frontiers in Health Informa*, 4050-4069.
- Annapareddy, V. N., & Sudha Rani, P. (2024). AI and ML Applications in RealTime Energy Monitoring and Optimization for Residential Solar Power Systems. Available at SSRN 5116062.
- Kannan, S. (2025). Transforming Community Engagement with Generative AI: Harnessing Machine Learning and Neural Networks for Hunger Alleviation and Global Food Security. *Cuestiones de Fisioterapia*, 54(4), 953-963.

Sriram, H. K. (2023). Harnessing AI Neural Networks and Generative AI for Advanced Customer Engagement: Insights into Loyalty Programs, Marketing Automation, and Real-Time Analytics. *Educational Administration: Theory and Practice*, 29(4), 4361-4374.