Rare Variant Pathogenicity Classification and Phenotypic Pattern Recognition: AI-Driven Computational Approaches to Early Detection of Genetic Disorders

Dr. Giovanna Di Guglielmo, Associate Professor of Information Engineering, University of Pisa, Italy

1. Introduction

Genetic disorders are becoming increasingly common, with 4–6% and 36% of children being born with a major or minor genetic disorder, respectively. These disorders affect an individual's mental and physical health and lead to significant levels of stress in families. Early detection is essential as it can lead to early intervention that can prevent or ameliorate many of these disorders. The last few decades have seen a significant increase in the development and availability of genetic tests, which are helping to detect genetic disorders at an early stage. Development in artificial intelligence is changing our lifestyles, which can be a potential savior for human society in the healthcare sector. Technological developments in machine learning and artificial intelligence can have a wide range of applications in the healthcare sector, including the discovery of potential drugs, advancements in diagnosing diseases, and helping to predict possible targets for drug intervention. The widespread applicability of next-generation sequencing has reduced diagnostic price tags but, at the same time, has increased the diagnostic dilemmas. Accurate clinical diagnosis of patients is critical, mainly when almost less than 50% of cases have a detectable mutation. This has led to improvements and the creation of revised methodologies for disease diagnostics. This text will cover the current challenges in the early detection of genetic disorders, which can be addressed by the use of machine learning frameworks, suggesting a roadmap for research.

1.1. Background and Significance

With the first genetic mutation described in 1902, there was a dire need to elucidate and classify genetic disorders. A major breakthrough in their delineation was made in the 1960s with the discovery of the structure of DNA. The Human Genome Project marked a

further milestone in the understanding of our genetic makeup and potential for sequence-based disease predisposition in its wake. Since 2001, the knowledge of such mutations could potentially alter the expected course of a phenotype or stratify therapeutic options. The development of high-throughput sequencing technology has theoretically provided the possibility for early detection. Early molecular confirmation can also prevent harm for those presenting with symptoms of the disease as the first manifestation. In contrast to what is conceived by disorder, patients may benefit from genetic counseling, including recurrence risk calculations for their children, in addition to the current advances in novel genetic therapies.

In the spirit of the Hippocratic Oath, pregnant women have the right to choose between prenatal testing or molecular confirmation, possibly through artificial intelligence. Few have had a complete exome or genome; hence, having a panel targeted against their ethnic background was common. It is our belief that current applications must not be overlooked, and traditional studies must continue to be combined wherever appropriate as knowledge is based on cumulative evidence. However, next-generation sequencing enables a semiconductor-based chip coupled with bioinformatics developed to genotype hundreds of thousands of the human genome and assess the inheritance of millions of common genetic markers. The need for a competent AI-driven data-effective interpretation machine is becoming increasingly relevant. The potential of computational imaging and artificial intelligence in terms of data efficiencies has enabled many ethical conundrums to be tackled.

2. Genetic Disorders: An Overview

Genetic disorders affect millions of families around the world. They can occur at any stage of life and may present themselves in a number of ways. An examination of the genetic material of the person suffering from a disorder is needed to confirm that the disorder is genetic in origin. Genetic disorders may be classified by specific information about them as well. One type of disorder is single-gene disorders, in which a single gene is responsible for telling the body to produce a protein in a way that is different from what is needed. Some examples include sickle cell disease, cystic fibrosis, hereditary spherocytosis, and hemophilia. Chromosomal abnormalities make up another category of genetic disorders that occur when DNA is missing or in the wrong place. A few examples of these types include Turner syndrome, Klinefelter syndrome, and Down

syndrome. Multifactorial inheritance also leads to genetic disorders when a wide variety of genetic differences involve multiple genes in combination with environmental factors. Some examples include diabetes, high blood pressure, and cancer.

Genetic mutations are responsible for causing genetic disorders. Every gene produces something called a protein. Proteins are responsible for the color of your eyes, the strength of your muscles, and the digestion of foods. In order for our genes to work properly, they must build proteins that are capable of performing a function effectively. When a gene mutation occurs, which is when the DNA sequence for a gene is changed, a person will develop a disorder resulting from an inability to make a protein correctly. Approximately 14% of infants and 34% of children are born with genetic diseases, with the incidence of fatal and new diseases being 10% and 30% respectively. Many genetic conditions are serious and cause lifelong disabilities. In the not-so-positive domain, genetic disorders can exponentially spread into a population where a small number of people contracting unusual diseases can lead populations to contract a genetic disorder. Early diagnoses for those affected are essential, especially given the prevalence of such conditions.

2.1. Types and Causes

Genetic disorders can be classified into various types. Taking into account the mode of inheritance, genetic disorders can be grouped as autosomal dominant disorders, autosomal recessive disorders, X-linked disorders, Y-linked disorders, sporadic disorders, multifactorial disorders, and polygenic disorders. Similarly, genetic disorders can also be classified based on the involvement of a single gene or multiple genes, chromosomal aberrations, inherited complex multifactorial influences, environmental influences, inborn errors, malformation sequences, and polymalformation syndromes. The availability of this classification system makes it easier to understand and interpret the possible causes and patterns of disorders that parents carry and pass on to their offspring.

Genetic disorders are caused by alterations in genes, and mutations can occur at many levels. The gene may be completely missing, not functioning properly, or functioning inappropriately. Consequently, it may be initiating a harmful and unnecessary action in the body, creating an associated harmful effect. In general, a multiple gene disorder is concerned with gene abundance at many loci. A trait may show multiple gene

inheritance because of environmental influences. Some traits that are controlled by gene abutment only may be influenced by environmental stimuli. In the presence of a given genotype, an environmental agent may be a necessary cause for a certain phenotype to appear. The results of adoption and twin studies provide evidence of the presence of gene-environment interactions. Insufficient or adverse environmental agents may bring about deleterious effects in genetically distinct organisms.

3. Current Challenges in Early Detection

Although various diagnostic methods and approaches have been developed and are in use in the healthcare system today, there still exist difficulties in early detection. DNA sequencing combined with computational methods allows for the detection of more than 6,700 single-gene genetic diseases but contributes only substantially limited to the detection of cases with common syndromic conditions. Only in the case of early detection of the disease, screening, preventive and treatment measures, lifestyle changes, behavior, and decisions around pregnancy can be made in order to ensure positive health outcomes. Additionally, more than 38% of the general population has no awareness of their genetic history, and less than 25% have at least minor knowledge of the history. Advances in genetic testing for various diseases lead not only to foretelling a person's susceptibility to a disease but also open the door for preventive care for newborns and help them avert the effects of genetic disorders in a reasonable time frame, thus reducing the suffering of the individual and society as well as the financial burden of care.

It is difficult for physicians and the public to understand the usefulness of the available genetic testing. The majority of healthcare professionals have poor self-confidence in their skills to integrate genetic information into the decision-making process. Misconceptions and lack of information among the public and healthcare providers are important obstacles in early detection. It is also possible that there are carrier individuals among the disease-causing genetic mutations, so it is necessary to carefully examine the benign and pathogenic markers in the genetic test report. In addition, accessibility and quality of genetic testing differ from state to state. The quality of commercial DNA testing services across seven countries showed that the quality and accessibility of the results of these tests, added to the varying prices paid in different countries, had a greater range. Despite the increasing success in the clinical genetic testing market,

increased use of genetic testing requires increased accuracy and validation tools. Given the changing role of genetic testing for prenatal diagnosis in newborns, the inherent problems in early detection appear to have entered a time when further precise methodologies will be able to unduly and completely complement traditional methods.

3.1. Limitations of Traditional Methods

Traditional diagnostic methods for the detection of genetic disorders have several limitations. The process of identifying genetic disorders is primarily performed using visual and manual interpretation of genetic testing results. However, the drawback of this methodology is that the interpretation of the results is reliant upon the subjective experience of the healthcare professional and their knowledge of the presented disease. Due to the lack of objectivity, an additional side effect is the possibility of inconsistencies in the results depending upon the experience of the analyst. For most healthcare institutions, the next step in the process is to conduct tests such as karyotype testing, chromosomal microarray testing, aCGH, and/or FISH to identify if the root cause of the symptoms could be due to chromosomal anomalies. However, the tests are generally chosen by the practitioner based on their experience and knowledge. Within smaller institutions, the available tests are often limited, which can sometimes cause conditions with rare chromosomal diseases to go undiagnosed over a long period of time. Besides the availability of the tests, the main factor in choosing which tests to perform is usually the cost. A comprehensive chromosomal microarray test can often result in an out-of-pocket expense of $1,000 to $3,000. Given that many patients are either uninsured or underinsured, this vital test is often skipped because of financial constraints. Even if the financial aspect is accounted for, the ensuing waitlist for carrier screening or comprehensive genetic testing often has a long turnaround time of several weeks to several months. This waiting period can result in an increased risk of delayed critical care. With the continuous advancements of technology, there is an increasing demand for low-cost, high-sophistication diagnostic tools. This is also true for genetic testing and therapeutics as the demand for genetic tests and gene therapies continues to expand.

4. Machine Learning in Healthcare

Machine learning (ML), a subset of artificial intelligence, is increasingly transforming the healthcare sector. By examining health data patterns, ML algorithms can offer new insights into diagnostics, treatment design, and patient risk. The key strength of such

algorithms lies in their capacity to integrate and process vast amounts of health data, such as clinical notes, imaging studies, and omics profiles, from multiple data sources. ML algorithms are becoming critical tools for disease detection and diagnosis, offering enhanced decision support, exact risk quantification, predictive modeling, and valuable treatment personalization. Success stories in the application of ML can be seen in a variety of domains, such as the identification of cancer biomarkers in oncology, patient phenotyping in critical care, predicting patient treatment response in cardiology and orthopedics, and many other examples. ML algorithms are making a difference across the entire patient journey, personalized for every step of their care. For AI applications to realize potential in day-to-day clinical care, high-quality data captured within robust clinical infrastructure, such as electronic health record platforms, are recognized as essential to jump-start machine learning (ML) applications. As these technologies become more advanced, AI-driven approaches are likely to be effective in the efficient onset of genetic disorders and are considered an important aspect moving forward.

4.1. Applications in Genetic Disorder Detection

4.1. Introduction

In the previous subsection, we detailed the different strategies and methodologies for the identification of human genetic disorders. In this subsection, we focus on the application of machine learning techniques for identifying and diagnosing patients with human genetic disorders at an early stage. Here, we share a few applications of machine learning techniques widely used for handling genetic data. Most of these techniques have been used for pattern recognition in genetic data for applications like identifying and matching individuals based on their genetic information, predictive modeling for gender and age, and handling large datasets. Examples include support vector regression, random forest, back nearest search, Fisher linear discriminant, extreme learning machine, and AdaBoost classifier. Also, various machine learning classifiers have been used for predicting individuals' subpopulation affiliation based on genetic data, which have been shown to be superior to the best similarity measures used for this purpose.

Further, there are reports in recent literature that demonstrate that machine learning classifiers, especially deep learning models, which complement feature extraction with classification in the same framework, are able to outperform existing predictive models

in the field of genomics. A deep learning algorithm has been developed for annotating pathogenic variants and interpreting their outcomes with high accuracy, which is better than existing computational predictors. Also, deep learning classifiers have been developed for the differential diagnosis of rare genetic defects from heterogeneous clinical presentations with an accuracy ranging from 74% to 99.5%. It has been shown that many of the known Mendelian pathogenic variants are situated either at binding sites of transcription factors or at other conserved regulatory elements, which has motivated the development of deep learning techniques for predicting variants at the non-coding regions. Finally, a deep learning algorithm has been developed for correctly classifying pathogenic variants with state-of-the-art performance of high sensitivity and specificity, outperforming existing methods that have a reported performance of lower sensitivity and specificity. These AI classifiers, which have demonstrated better precision, recall, and F1-score in identifying affected individuals' genetic disorders at an early stage, can help bridge the gap between existing detection rates and the proposed values targeted at reducing disorder prevalence in the global population. This will require the creation of a genetic health blueprint involving AI-driven data, analytics, and machine learning for collective computing involving patients, genomes, disorders, and the environment. However, there are some challenges in implementing AI-based detection in clinical practice. The classifiers are found to perform suboptimally with datasets in which all the pathogenic variants are already known and reported, due to overfitting of the model on such data, which interferes with its ability to generalize on new data. It is also important for healthcare professionals to know how the AI model is implemented and the way specific biological features in the datasets are used by the classifier. This is more important from the standpoint of result interpretation without expecting healthcare professionals to be data scientists, as more and new bioinformatics models and pipelines continue to evolve for the AI diagnostic purpose of genetic disorders. Thus, effective interpretation of AI-based model predictions of genetic disorders forms an essential part of reporting the AI-based classifier results for genetic disorders.

5. AI-Driven Approaches for Early Detection

Early Detection

The early detection of genetic disorders is an important approach for managing the onset of phenotypes during the initial stages of disease progression. Importantly, artificial intelligence-driven methodologies have demonstrated the potential to enhance early detection. AI tools mainly focus on the strategic anti-interventions with the aim of reducing future disease-related damages. Approaches such as predictive analytics offer the potential for clinical health monitoring, and data visualization technologies provide a method for efficient data representation for early understanding and data dependency of phenotypic and genotypic manifestations.

Artificial intelligence is a rapidly growing field that assumes processing numerous amounts of data should be real-time. For this purpose, the quality data must be timely acquired and pre-processed for further usage. In healthcare, timely data may save patients' lives. For an optimal training set, the quality of the input data is a strong determining factor for the quality of the final outcomes. Data integration systems are gaining in prevalence. Using AI for the analysis of numerous patient images or genetic data without transmitting the data outside of the hospital IT system is authorized. The following AI systems are tools and frameworks applied in some clinical diagnostic laboratories: Setting the right priorities; Introducing current results to make AI a part of laboratory workflow; and Building trust with first AI applications. Enhancing early detection of genetic disorders using AI systems that can also be used for patient follow-up is an opportunity to avoid welfare exitus and to limit incorrect early interventions. False negatives are at the forefront of managing data to avoid welfare exitus.

Artificial intelligence approaches may be used to improve the early diagnosis of genetic health issues. AI-driven techniques are tools or systems consisting of complex modules and data that foster the exploration of larger datasets, fusing molecular and clinical data faster, in a more coherent and homogeneous way. It will improve interpretation and understanding of complex genotype–phenotype data.

5.1. Data Collection and Preprocessing

With the popularization of big omics data, data collection and preprocessing have become hot research topics in the application of AI in genetic disorder detection. A series of studies stated that the success of a model largely relies on the quality and representativeness of its datasets. Applying diverse types of genetic data ensures that the synthetic datasets adapt to more application scenarios.

The value of AI algorithms mainly depends on the nature and organization of their underlying data. During the last decades, data preprocessing has emerged as a critical stage to improve intimacy with data by reducing or removing irrelevant and redundant information, which causes the AI model to slow down, be inaccurate, and become unmanageable. For raw genetic data like DNA sequences and SNP data, first select the ranges of the data and collect the dataset that contains specific genetic disorders of interest. In general, this will expand training data research and improve the performance of detection and classification models. After collecting the data, it is necessary to take steps to ensure that the dataset is clean and organized in an appropriate format.

In this process, it is common for data to be incomplete due to noise, outliers, or missing values, and misclassification errors, which adversely affect the performance of AI learning algorithms. For AI-driven approaches, the first important step is to preprocess the data, which have different formats and quality levels. Particularly, the data of different genetic disorder databases are encoded in diverse formats. The raw data of genetic disorders need to be preprocessed with techniques that include the gene selection process. Further, various missing data completion mechanisms can be used on experimental data during the data preprocessing phase. However, extreme and inconsistent values are ignored. Additionally, it is important to standardize the data into a predefined format to achieve high performance through generalization, but not through data memorization. It should also be noted that, with the development of AI-driven approaches, the data preprocessing parameters can be learned from the associated datasets, but still present some limitations and challenges in terms of the mathematical operations required. Moreover, in the next subsection, the fundamental body of work in AI-driven approaches for early detection of genetic disorders is presented.

6. Future Direction

There are several emerging trends and technologies that could completely revolutionize early detection of genetic disorders. 1. Tribrid systems for classification of genetic disorders: With the increase in computational power and the availability of data, we see more and more the inclusion of end-to-end approaches for genetic disorder detection. As more advances are made in multimodal and sequential data processing, we expect these methods to be extended for DNA and RNA data. 2. Fusion of different types of

data: A focus on multimodal biomarkers has been a clear trend in the development of AI techniques for early morphological detection of genetic disorders. We suggest that more studies should be made for combining different data types such as multimodal imaging with clinical data, behavioral data, and other models. 3. The US NIH is funding a wide range of research projects in omics using multi-omic data types in the genetic research of common complex disorders. We expect that in the next decade, the development of the best approaches and strategies in genetics will have improved and will start being applied for the earlier diagnosis of genetic disorders. We encourage data scientists to interact more with biomedical scientists and lab researchers to develop more general and robust strategies and approaches for the development of all human genetic research. Challenges and Opportunities: The main opportunities and challenges are related to computational cost. Challenges Involving Ethical Considerations and the Regulatory Landscape: Considering the massive potential of AI as part of many future healthcare tools, it is critical to be engaged with the ethical, moral, and regulatory discussions that will shape future research. In the healthcare setting, the introduction of deep learning models should proceed cautiously, especially in rare genetic disorders because, in general, if a healthcare professional does not have domain knowledge in genetics, they would be unable to assess the evidence base and would likely depend heavily on the model. Future Perspective: As these areas continue to grow, the research priorities for AI-driven early detection of genetic disorders should continue to be: (1) refining algorithms and techniques and (2) the enhancement of interpretability. We need to push forward the limit of what is possible in the realm of early detection and the inclusion of many disorder subtypes. These will enhance the possibilities of genome-based, personalized medicine.

7. Conclusion

The early diagnosis of a genetic disorder is essential to prevent fatalities and other complications. Although some therapies exist, most genetic disorders remain untreatable. Furthermore, traditional clinical diagnosis can be expensive, time-consuming, and invasive. In contrast, non-invasive, low-cost methods exist, but they may require prompt disease surveillance to maintain patient survival, and may be considered as an intrusion of privacy. Recent computer-aided diagnostic approaches rely on AI technologies, which can accurately, quickly, and inexpensively predict genetic disorders from various diagnostic materials. To be impactful, AI-driven diagnostic tools

need to address the needs of various stakeholders and must comply with established regulations and guidelines, necessitating complex multi-disciplinary integration of computer engineering, medical practice, psychoeducational counseling research, and ethics. Importantly, AI-driven methods should aggregate extensive high-quality data on patients and their family members, and this then calls for real-time data ownership, sharing, and the following security and privacy policies. We believe that precision therapy for most genetic disorders remains promising, and large-scale multicenter research should be conducted on conditions currently lacking in early detection approaches. Conventional, karyotype-based early detection methods are challenged by a high false-negative rate and inability to evaluate disease risks. We have drafted a review on non-invasive, high-throughput early detection methods for genetic disorders, like other body fluids-based, cell-free DNA, and AI-driven approaches. AI technologies as a process of intelligently making use of big data hold great promise for the transformation of health care. We have learned that big data increases the likelihood of naturally occurring detection bias. Healthcare research in digital, data-driven environments may benefit from increased peer collaboration, prompt data sharing, and collaborative efforts involving patients, in order to predict, prevent, and quickly detect health outcomes, including genetic disorders. In this review, we have explored the strengths and weaknesses of AI technologies, and are encouraged to yield robust, and rapidly developing kinds of recorded pre-symptomatic genetic diagnostic diseases in need of portable diagnostics in a limited resource setting.