Deep Science Publishing
https://doi.org/10.70593/978-93-49307-76-6_3

DeepScience
Open Access Books

Chapter 3: Machine learning models for predictive genomics: from variant interpretation to early risk stratification

3.1 Introduction

There is a growing focus on how to translate genomic data into new clinical applications. While the increased availability of genomic data allows the identification of genetic mutation carriers in their early years, the heterogeneity of genetic background makes it hard to understand the relationship between genetic variants and phenotypes. Machine learning models have proven to be a promising approach for a broader and highly worthwhile prediction of genetic-related genomic data. Some of the studies focused on variant interpretation for Mendelian diseases, subsetting input at first, and training a model for the mutation impact classification, because they interpret the variants of unclear significance and this could narrow down a significant percentage of SNV in most individuals.

Early risk stratification of chronic diseases has shown to be more effective for therapeutic and reversible intervention. One approach is analyzing gene sets and constructing a proxy SNP set, focusing on genes that are meaningful for a specific disease. Another is using SNP features directly. Studies creating a predictive feature of SNP data focused on early-stage diseases, where it is hard to determine affected genes beforehand. Those can be useful for various polygenic diseases, complex traits, and common disease susceptibility as well. Early risk stratification studies have been conducted primarily on retrospectively collected case-control data, using prevalent cases, with most studies just reporting the AUC. To apply these models in real world clinical practice, it is desirable to conduct a prospective cohort study, possibly utilizing cohort consortium data from multi-centers or countries, examining incident cases, and showing more comprehensive evaluation studies. With the trends towards preventative precision medicine, such prediction models could become prerequisites for participation

in chronic disease preventative programs or for initiation of preventative medication. In this direction, there are some studies that have started with drug repurposing on at-risk individuals.

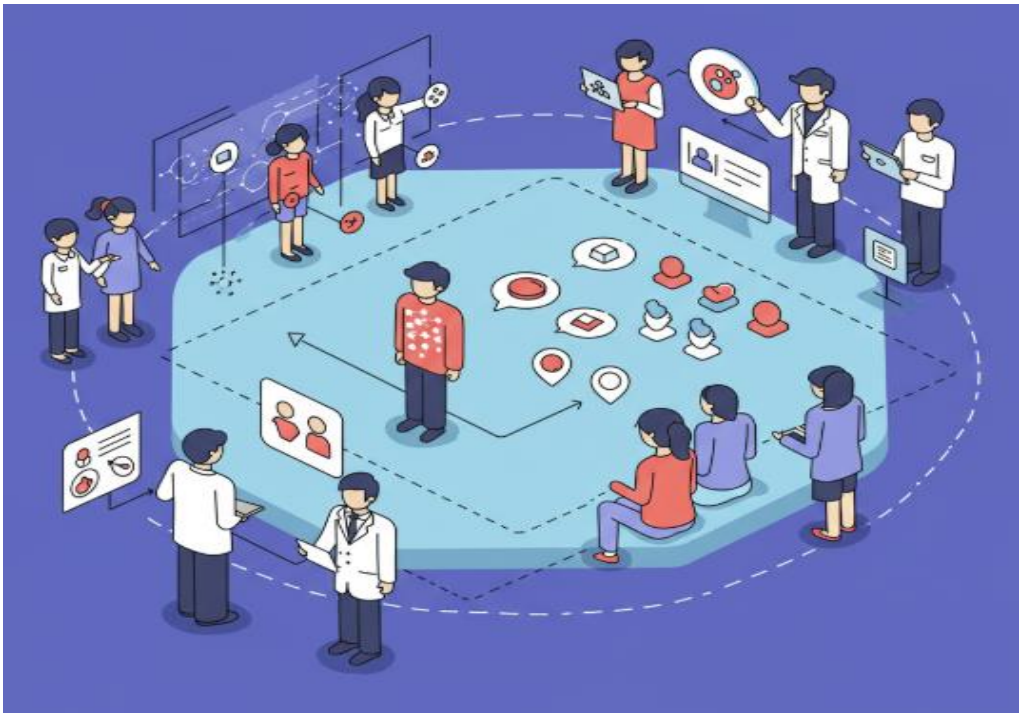

Fig 3.1: Integrating AI/ML Models for Patient Stratification

3.1.1. Background and Significance

Rule-based methods and scoring algorithms are used in increasing frequency to predict the effects of single variants. Furthermore, as a first step toward modeling variant interactions, functional networks, random forests and rule-based models propose relevant questions for the context of complex genetic disorders. High-quality ground-truth data are still lacking for the interpretation of non-coding mutations and for population-level variant consequences. Well-established modeling and validation strategies from other machine learning driven research fields may provide valuable guidance for the improvement of predictive genomics methodology and its wide-spread application in healthcare and biotechnology. Machine learning (ML) will likely be a cornerstone of biology and medicine in the 21st century. ML-based approaches are powerful mathematical tools that can be used to analyze and model complex datasets, such as those generated in genomics and medicine. Traditionally, ML algorithms have been used in bioinformatics to handle large volumes of molecular biological data, extract

patterns and relationships from those data, and make predictions— such as classification, regression, or clustering workflows—based on these learnt patterns. The on-going revolution in DNA sequencing technology has greatly increased the scale of genomics and has led to an ever-growing demand for ML applications in this field. This is especially so in the context of the new initiatives and large-scale genome projects. Machine learning-based predictive models range from variant interpretation and gene/protein function prediction through drug-target interaction prediction and clinical patient stratification up to population-wide early-risk prediction. Variant interpretation, risk gene prediction, polygenic risk scores, gene expression, splicing, DNA methylation, histone modification, transcription factor binding, chromatin accessibility, allele-specific binding and expression. Much work has also gone into understanding the comparative performance of these features for the family of methods configuring feature-based machine learning risk prediction.

3.2. Model Development

Machine learning SNP based prediction of individual risks of prominent chronic diseases such as type II diabetes, ischemic stroke and some malignant tumors will be highly beneficial not only in drafting a strategy for minimizing genetic vulnerability through lifestyle, but also from a perspective of optimizing the efficiency of the health care resources. Methodologically, as the clinical applicability of machine learning models, the focus is on the characterized effects of common genetic variants with moderate to small impact on the disease risk and the elaboration of a framework for early risk stratification in a time perspective longer than what is currently practiced. Bioinformatic feature selection methodologies enhancing the sensitivity in predictive modeling and novel experimental verification of case-control seed SNPs with potential mediating link to more widespread targets in the cellular response to genome damage are described. It encompasses the construction of suitable custom epidemiologic databases, including harmonization of the inter-study data and extensive literature review on the functional SNP effects, as well as the computational modeling of dose dependent transcription factor occupancy by genotype specific enhancers. Subsequently delineated effective connectivity pathways involving the top prioritized SNPs are presently assessed by a dedicated laboratory experiment. The paper is expected to assist the future implementation of the personalized genome information in a new generation of public health supportive prevention strategies.

3.2.1. Training and Testing Datasets

In the predictive genomics research domain, a numerous number of training and testing datasets is of interest, both for single-nucleotide polymorphism (SNP) related phenotypes (e.g., diseases, drug responses) and also for "classical" phenotypes (e.g., eye color, height) related to a given genotype. Analyses workflows and solutions on modeling injustices in predictive biomarkers for early detection of high-grade serous ovarian cancer (HGSOC). The former represents the main goals of different machine learning models, independently from the used techniques and methodologies. Surveys on related articles and scientific contributions on the used datasets are also provided to aid the scientific community in exploring novel research ideas and results. Finally, overall insights, future leads and challenges on prospective research directions are discussed.

In the fast-evolving era of data science, there is a growing integration of different disciplines such as biostatistics, computer science, mathematics, and also business intelligence, offering enhanced decision-making, biomarkers discovery, risk management, and more. In this perspective, machine learning approaches are widely spreading in the predictive genomics research domain, focusing on to highlight potential driver interactions in the breast cancer subtypes context, sensitive to the hormone estrogen and to the growth factor receptor HER2. Supplying the best knowledge, no previous works have attempted to implement such models. Nevertheless, in the study of, the authors focus on a comprehensive analysis comparing 20 selected methods between ensembles and novel deep learning approaches, in the prediction and interpretation of the impact of over 9,000,000 possible SNVs on gene expression levels.

3.2.2. Cross-Validation Techniques

Cross-validation is a statistical method useful in estimating the field efficacy of a predictive model. Many different methods have been proposed to perform cross-validation and, second, the forgetting factor in a long pattern. When the forgetting factor is equal to one (form of the leave-N-out cross-validation) the efficiency of cross-validation strongly depends on the choice of the subsequence of points used for training and for validation. The most used methods are the leave-one-out, the leave-p-out, and the K-fold. In some other papers the designed approach of cross-validation leads to better results in terms of generalization than the mentioned approaches: for this reason, the simplest method—a holdout technique with the results of small improvement—has been mainly considered in this and in what follows.

Genomic and proteomic technologies have recently led to an increase in the size and complexity of datasets, so that it can identify a function relating the Clinical and

Genomic variables to the Risk of Disease. Due to the high dimensionality of such datasets and the limited sample size, the variable selection process is a difficult task, so they have proved the importance of performing variable selection within resampling methods like cross-validation (CV). However, it has been objected that CV can be overly optimistic as a tool to select the most significant variables since the same selection process is repeated several times. A point of view based on deviations from automation. Overly optimistic interpretation of the performances obtained in several simulated experiments is also taken in this paper. Two recently proposed methods for selection of the optimal set of variables are (1) a reformulation of the ordinary least squares approach by a modification of the multiple correlation index; such a reformulation was made to be well-suited for the error estimation and (2) a new proposal admitting the use of CV both for coefficient estimation and for the choice of the set of variables .

3.3. Machine Learning Fundamentals

In the past decade, genetic data have become more complex. Researchers now have access to whole-genome sequence (WGS), whole-exome sequence (WES), RNA sequencing (RNA-seq), and other forms of gene expression and proteomics data. As the data size has increased, the questions geneticists have begun to answer have increased. For example, researchers now want to know whether disease risk is caused by gene–gene interactions between genetic variants or what effect a single missense variant might have on the function of a protein. Geneticists have begun to turn to machine learning (ML) techniques (Burugulla, 2022; Challa, 2023; Kumar et al., 2025).

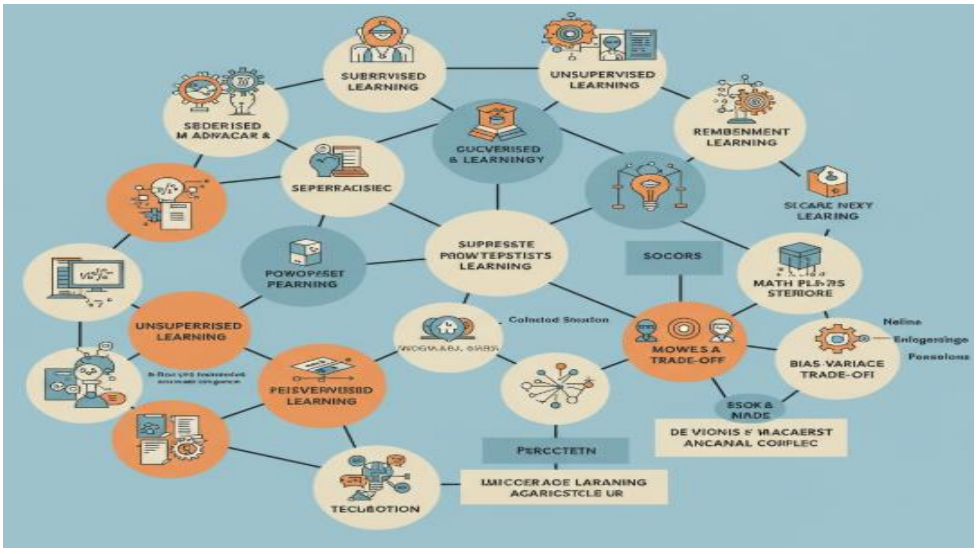

Fig 3.2: Fundamentals of machine learning

Machine learning methods can be broadly defined as the algorithms that can learn from data to make predictions or identify patterns. Given a set of features, ML can learn a function that maximizes prediction accuracy or identifies underlying patterns within the data. Similar to GWAS, ML uses a training set to construct the prediction model. There are multiple types of ML techniques that have been adapted to genetic data, including linear regression, support vector machines, neural networks, and random forests. The majority of ML algorithms in genetic analysis are used in prediction, using input data to predict an outcome. In genetic data, this could be using SNP genotypes to predict whether a subject is a case or control. For example, different signatures of gene expression in relevant tissues can be used to classify different tumor types. Importantly, ML algorithms are able to model complex interactions between input features, which enables detection of the non-additive effects that current methods may be missing.

Historically and currently, the implementation of precision medicine in the clinical setting has been defined by the analysis of an individual's genetic details. In this setting, one genome sequence with extensive clinical records or family histories is profiled as the standard clinical endpoint. With the exponential increase in sample sizes and breadth of phenotyping coverage, the information available for a given patient is likely to exponentially increase before the complexity of the disease and clinical action required for the action go down. Thus, there is an urgent need to develop efficient modeling frameworks that can be easily scaled to interpret the current volume data, deliver actionable results to both healthy providers and patients, and can be readily integrated into current clinical practice.

3.3.1. Types of Machine Learning

In precision medicine, predictive models of health outcomes from the genome is one key for integrating a standard clinical workflow that will be compatible with existing practice. This is a tough challenge, as it requires new developments in estimating risk from the genome, diverse data sources beyond genotypes, and new approaches for delivering risk stratification.

Machine learning methods offer accurate and potentially efficient disease risk prediction models that operate directly on genotypic information. Such models are an active development area in the pursuit of methods that will allow for the efficient identification of patients at high risk of disease. A key to these developments is the use of clinical outcome data and large predictive models of genetic data. Early meetings of data and method developers from many centers and hospitals will be necessary to agree on standards for the specification of cohort data and the continuing assessment of risk simulation. These collaborations will provide estimates of the portability of disease risk models across the population and the true ability of models to improve patient outcomes.

Platform development is underway for carrying these models into use in the clinic, including the incorporation of results with the electronic term record, the creation of an approach to support patient counseling, and testing of models for responding care providers and patient education.

Disease risk modeling within the precision medicine paradigm stands at a remarkable transition point, providing enormous potential to transform generalized prediction of an individual's risk of disease.

3.3.2. Common Algorithms Used in Genomics

Machine learning (ML)-based approaches can be an effective way of predicting an individual's risk of disease. Unlike other popular predictive models in genomics, there is potential for ML to account for complex interactions between features. ML algorithms utilize a set of advanced function-approximation algorithms to create a model that accurately maps the association between a set of risk single nucleotide polymorphisms (SNPs) and a particular phenotype. The genotype data of a patient can be used as an input to the predictive ML algorithm to predict their risk of developing a disease. The prediction of disease risk using SNP genotype data can be considered as a binary classification problem within the discipline of supervised learning. The approach can be broken down into several steps. In the first instance, the data are pre-processed by applying quality control and feature selection. The high-dimensional pre-processed data are fed into a specific learning algorithm, along with algorithm-specific settings. The pipeline then learns an association between the features and class labels, stores the trained model parameters, and validates the predictive performance of the model using unseen data. The output of the entire process is a binary classifier that classifies the input into different classes, along with the model's associated performance metrics. The model and its trained parameters can then be used to predict the mutation's effect size. Within the model building process, the application between the features and class labels is trained probabilistically. There are also hyperparameters that are specific to each algorithm. The set of algorithm hyperparameters impacts the potency of the pipeline functioning, so it is important to pick their values in an informed way. Given a fixed set of hyperparameters, 10-fold cross-validation is used to partition the dataset into ten equally-sized parts. The training and validation process is repeated on the partitions iteratively, with the learning occurring on nine of the partitions and the performance being evaluated on the tenth. This process is referred to as a 'fold'. The outputs of the process are ten models, along with their ten sets of parameters.

3.4. Genomic Data Types

Structural genomic variation is largely confined to deletions and insertions. Deletions and insertions are rarely observed in human genes, and different forms of insertions and deletions can produce similar phenotypes (such as a frameshift in a downstream open reading frame). This often makes phenotypic impact of insertions and deletions unpredictable, as such do not have much clinical utility. Instead, medical geneticists have spent their time on other forms of DNA variant, in particular single nucleotide variants (SNVs). At this level, advances in gene expression and epigenetic assays suggest that most genetic effects (outside of simple monogenic disorders) involve regulatory activity, usually mediated by DNA accessibility and the presence of transcription factor binding sites.

From the medical perspective, researchers would all like to be able to intervene at the first presentation of a disorder, but particularly in some specialties (e.g. neurology; cancer) earlier diagnosis is not necessarily very helpful. For neurodegenerative conditions, most of the damage is thought to be set in decades before the condition is first noticed by the patient. The first signs of many kinds of tumours are when a stage-4 cancer has already formed. Even where early stage interventions are possible, the cost–benefit analysis of aggressive action on a patient with no apparent illness will often look poor, particularly when viewed through a population lens.

Each sickness has a particular list of genomic data and in this manner it is being looked into each one in turn. This dataset incorporates treatment approaches for every one of the disorders. Similarly as with the advancement in the identification of sicknesses, a considerable lot of the sickness treatment processes are being used for sickness healing, as it were, diminish or remove the symptoms and not really kill. Meds are regularly endorsed for sickness mending that have not been put through FDA-approval for that specific sickness. From a researcher's viewpoint, the clinical goal is the discovery of an efficient treatment for any of the sicknesses. It is also expected to have an in-depth exploration of multiple aspects of the treatment. From a machine learner stand-point, the goal includes assessing the effectiveness of various treatment approaches and the modeling process of the treatment.

3.4.1. Genomic Variants

The sequence of the human genome in 2001 has paved the way for new technologies and information in biology, medicine, and beyond. The ability to collect, store and analyse nucleotide sequences is growing at an exponential rate. Encouraged by large-scale collaborative projects, aggregated genomic data releases have reached a petabyte size. Availability of this goldmine of information was one of the cornerstones in the

advancement and popularisation of personalised medicine. According to the “disease triangle” concept, individual health and well-being are influenced by the genetic make-up and lifestyle of a person. Late- and post-genomic research has made staggering progress in the identification and understanding of genetic variants, that is, the first corner of the triangle. Until now, tens of millions of single nucleotide polymorphisms (SNPs) and small insertions and deletions have been associated with human traits and disorders. However, identifying a genetic link to a certain condition is only a small step in the grand scheme of things. In order to comprehend and ultimately treat complex traits, failure of this third corner is considered to have a profound negative impact on public health, especially in high-income countries. Subjecting large amounts of data to machine learning algorithms is currently regarded as holding the greatest promise in many fields, including bioinformatics.

3.4.2. Transcriptomic Data

Great strides have recently been made in understanding the molecular fingerprint of diseases such as cancer. The increased availability of high-throughput "omics" data such as genomics, epigenomics, and transcriptomics, has given researchers the possibility to directly observe and interpret different levels of biological information. Among the molecular "omics" data, transcriptomics is used to describe the landscape of transcription within individual cells and tissue types. Changes in the transcriptome profile can reflect the state of underlying disease and inform diagnosis and prognosis. This has led to a considerable effort to collect and study large-scale transcriptomic profiles across different diseases and healthy controls. (Pamisetty, A., 2022; Pamisetty, V., 2023)

Transcriptomic data can be measured using different experimental protocols like RNA-seq and expression microarrays. With the advent of high-throughput technology, the generation of large amounts of the transcriptomic data has enabled the development of predictive models based on machine learning. A recent review of the currently ongoing projects in this field discusses how transcriptomic-based models have a variety of applications in the context of health management: starting from the development of the new predictive methods on epidemics scalability and general safety, remote monitoring of chronically ill patients (e.g., anxiousness, epilepsy, and cardiac weaknesses), development of transformative technologies enabling early disease detection ancestral to advanced (plus personalized) restorative interventions, and (ideally) the prevention of untimely death. While these predictive approaches have traditionally been linked to biometric or diagnostic data (e.g., temperature, glucose levels or heart-rate), genomics data has opened up new opportunities for predictive models based on genetic predisposition. In this context, the same work provides the results of a large scale and complex project on the development of predictive models based on genomic data that

aims to help patients to better understand different types of genetic information and its potential impact on health, predict their disease susceptibility (e.g., with respect to Alzheimer, Type 2 Diabetes or some specific forms of cancer), and provide actionable advice towards lifestyle changes to reduce risk.

3.4.3. Proteomic Data

We report a study that involved building multi-layer classifier models to automate the translation of healthy human genome data into useful medical information. The models were trained to predict quantitative traits, such as gene expression or neutrophil count, from germline clinical variants. The same models can also be interpreted in their ability to rapidly stratify patients early in life to identify those at high risk of later developing diseases or disorders. Risk assessment was performed by translating genotypes at individual positions within the genome into positional scores equal to the model-predicted change in a clinical QT value. The scores are combined across a subset of genomic positions into a single risk percentage. The top 10% of individuals under twenty years of age, ranked by this risk percentage, were subdivided into five equal groups. Over a lifetime, individuals in the top category have, on average, a 4-fold higher risk of being diagnosed with the associated disease compared to those in the bottom fifth. Stratification models are presented for obesity, anemia, neutropenia, thrombocytopenia, and gene expression in thirty healthy tissues. Many potentially clinically useful scores are reported, including for associations, loci. All scores and risk percentiles on all traits are posted so that the results can be explored interactively using a datastore. The stratification models, and the large set of scores and risk percentiles, are intended to serve as an initial resource and foundation for the development of more advanced predictive biology approaches and the study of related ethical, clinical, and legal considerations.

3.5. Variant Interpretation

The dramatic proliferation of genome-wide genotyping and sequencing data presents many challenges to variant interpretation, from the association of specific variants to their effects on biological functions, health and complex traits. Much information can support variant interpretation, including population allele frequency, conservation parameters, gene region annotation, variant curation and disease knowledge bases. The interpretation of the wealth of information can both face methodological difficulties and is expected to involve considerable subjectivity. Early ability to predict the functional impact engendered by genetic variants could obviate such constraints and conundrums

and greatly expedite knowledge conveyance between research and practical applications such as medicine, genetic counseling, pharmaco-genomics and bio-industry.

Traditionally researchers have employed association tests in clinical genetics and genome-wide association study fields, and molecular biologists have used cellular and biochemical screens to discover how genetic variation influences particular molecular events. A central goal of these diverse efforts has been to identify the subset of all genetic variants that alter gene functions and phenotype. In view of the landmark of the Human Genome Project and the recent deluge of genomic data there lies an unprecedented opportunity to transition from such correlative approaches to the identification and characterization of the sub-set of genetic events that are causally associated with specific phenotypes. Remarkably these goals are considered formulation closed, with pretty much all causal variants having already been identified, at least in humans.

Prediction of the functional effects of genetic variants is an essential and challenging task for the successful discovery of the downstream processes connecting genotype and phenotype. The rapidly increasing amount and diversity of genomic variation have highlighted the central role of elucidating the impact of mutations on gene expression regulation in the functional interpretation of genetic variants. Efforts to characterize the impact of genetic variation have, therefore, expanded from the traditional association setting to the development of novel hypothesis-driven and data-driven approaches that uncover the mechanistic basis of allelic influences on gene expression, often through the mediation of chromatin interaction and modifying protein affinity.

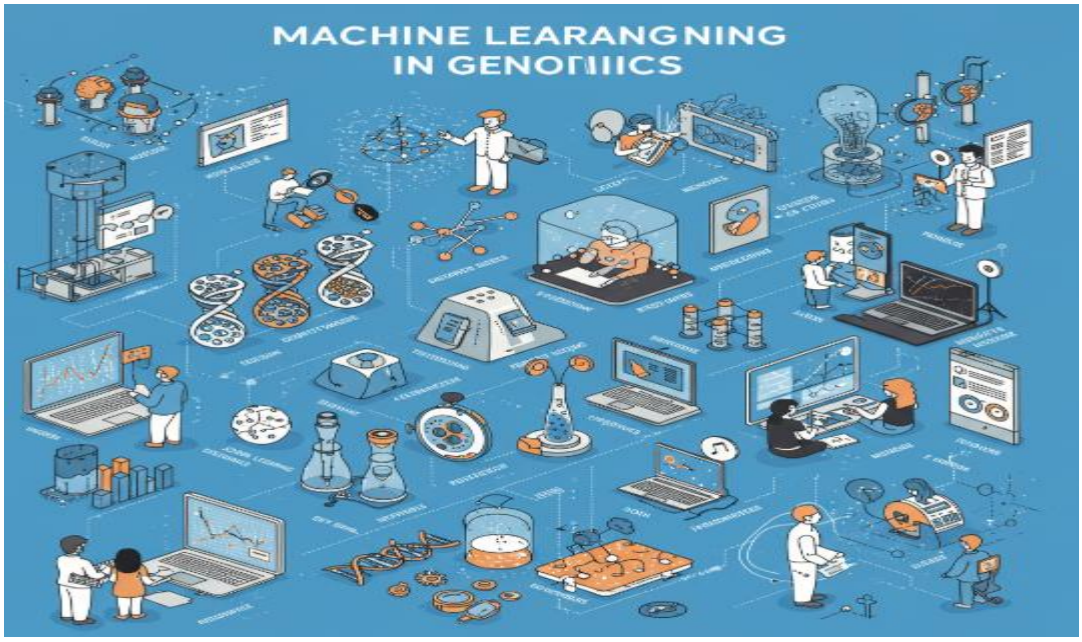

Fig : Trends of Machine Learning in Genomics

3.5.1. Clinical Significance of Variants

Since the completion of the Human Genome Project in 2001, the DNA of many individuals and other species have been sequenced, yielding many new natural genome sequences. Numerous genetic studies have been published on the genetic variants of different human populations. The clinical significance of genetic variations relies on the traits they underpin. Some genetic variations, such as non-synonymous single-nucleotide polymorphisms (nsSNPs) that alter the amino-acid sequence of an encoded protein, are more likely to be linked to the traits of an organism and are consequently more frequently reported in the scientific literature. In a study of the human DNA sequence from 650 randomly selected individuals across the globe, 20 million genetic variants were identified, including 3 million nsSNPs. Among those nsSNPs, 123,843 were reported to be associated with 3,078 traits. Clinical significance of human variants has been highlighted by the strong association of pathogenic variants previously identified in patients with cardiomyopathy.

An interpretation of a relationship between individual genetic variations or overall genetic profiles of patients and the manifestation or severity of a disorder can lead to the practice of personalized medicine using genetic testing and non-invasive risk stratification. Here, a method and system is described for the early risk stratification of health conditions, in particular, disorders manifested in facial features, forms, and malformations of an individual. The method starts with determining a loss of function genetic variant associated with facial features. Thereafter potential mutation is determined based on the loss of function genetic variant associated with facial features. Finally, the phenotype facial profile of an individual pertaining to the facial features is evaluated to determine a manifestation of the potential mutation. It is shown that the early stratification of conditions manifested in facial features can be practiced years before the development of the symptoms.

3.5.2. Tools for Variant Annotation

Recently, significant resources have been allocated to large-scale sequencing studies to identify genetic polymorphisms that may be linked to a variety of diseases and health conditions. Many of these polymorphisms affect non-coding regions and are of a regulatory nature. Computational analysis of these polymorphisms is hence crucial for prioritizing them for further experimental screening. However, extracting meaningful information from these polymorphisms is challenging due to the lack of robust functional annotations. In addition, the polymorphisms in regulatory regions, unlike the missense ones, do not have a strong in silico discernible effect and are not subject to any simple

nomenclature. Here, a new framework, called PredictSNP2, is presented that is capable of accurately evaluating the effect of single nucleotide variants and has been developed. This is a one-stop source that is able to address the different characteristics of a variant that are dictated by its location in distinct genomic regions, accounting for the effect of the surrounding sequence context.

3.6. Data Preprocessing Techniques

As a new field of healthcare analytics that merges advanced analytics and genomics, predictive genomics has an indispensable need for interpreting extensive genomic data to predict clinical phenotypes. Specifically, given a new genetic variant, the prediction task is to reveal potential associations between the variant and important clinical phenotypes of interest such as disease risk, stage or progression of a disease, treatment response, etc. Variant–trait associations are commonly estimated using "predictive features" as an intermediate representation. Generic machine learning models can then be built to predict these features, followed by interpreting them into human-interpretable knowledge, such as functional impact and causal effects underlying key genetic and genomic components using appropriate tools and databases developed in biology and bioinformatics. This workflow of variant interpretation is at the core to accelerate numerous efforts including but not limited to biomarker discovery and development, pharmacogenomics, gene editing, and identification of personalized treatment and therapy.

Looking into the workflow, extant studies have established the machine learning models and tools to predict general or generic predictive features. Nevertheless, the quality of genomics data has an essential impact on the prediction performance. Various issues like noise and batch effects are caused by subtle bias in the experimental design and data acquisition process. Although the prediction and interpretation workflow is designed as a framework of Bayesian decision graph such that it is more robust to address noise influences in raw data, subtler bias is however hard to model directly from raw data. For this purpose, a series of data preprocessing techniques and tools for addressing batch effects, a specific type of bias, are critically needed in the biological community.

3.6.1. Normalization Methods

There are three normalization methods. There are quantile normalization (QN), median normalization (Med), and variance stabilizing normalization (VSN). QN is regarded as a popular non-parametric statistical method that ensures the same distribution of observation in the sample by aligning the corresponding preprocessed raw counts with an empirical cumulative distribution function, often in the form of matched quantiles.

On the other hand, Med annotates a normalization method that strikes the whole data distribution to focus on the gene-expression levels' changes, ensuring that the median expression preserves for all the samples in the processed data by adjusting the non-steady and/or edge effect bias. Med is known to better predict two-sample Welch's t-statistic test or hallmark gene-set enrichment score than all other random non-redundant comparison methods in the absence of truncated genes' edges. Finally, VSN is a parameterized normalization embedment. VSN may achieve comparable risk prediction performance to the new normalization method after survival-based risk prediction, but its underlying biology method is more sophisticated with no tuning parameter or intuitive interpretation in its gene-expression fitting process, while it may have no superiority on the biological outcomes compared with the new normalization method when the study population is stratified by an external biological group.

3.6.2. Feature Selection and Extraction

Feature selection techniques find out the most useful variables or features defining the underlying trait to be predicted in machine learning models. Those variables can belong to different sources such as genetics, epigenetics, proteomics, metabolomics and environment, among others. In the context of genomics, thousands of genetic variants such as SNPs can be measured. Gaining biological insight from genome-wide association studies is an area of interest in predictive genomics, especially in complex diseases such as cancer. On the other hand, as medical costs are increasing, detecting individuals with high risk in early stages is becoming more relevant. Therefore, a new methodology is proposed to automatically search the most relevant genetic signatures, measured with a single-nucleotide polymorphism (SNP) panel, associated with a given clinical variable of interest. To do it, six classification methods have been selected after studying the impact of diverse configuration parameters on performance. Classification methods are a general set of approaches used to learn how to predict a binary or multinomial clinical outcome from genomic data. To search the most relevant genetic signatures, a large-scale empirical evaluation is carried out using 8 publicly available clinical SNP datasets. Two appropriate methodologies are described to produce generalizable results rather than to being trapped in a specific cohort.

Genomic data and 200 clinical datasets were obtained from the catalog of SNPs associated with cancer and public repositories. To find out the most relevant genetic signatures in clinical context with a high degree of performance, 200 real SNP datasets have been selected. To assess any algorithm (classifier and/or signature search engine) properly in the context of genomic analysis, a suitable methodology is proposed. Additionally, this methodology is also designed to provide generalizable results in independent cohorts. This methodology is therefore strongly recommended to be

followed by any researcher who endeavours to get relevant results. Moreover, some recommendations are offered to bring to light any defects that the methodologies may have, in order to get more reliable and generalizable results.

3.7. Conclusion

Early adopter examples indicate that Systems Medicine initiatives will primarily focus on Machine Learning models developed for Predictive Genomics tasks defined in narrower application scopes. Despite the rapid development of these models and methodologies, adequate consideration of the overarching System landscape they will be integrated into is currently missing. There are emerging endeavours of utilizing Machine Learning models for Predictive Genomics in large and integrated Systems Medicine workflows targeting early Risk Stratification of multifactorial, chronic pathologies, in which they have been spotlighted here. Similarly, from a broader view of application, Feature Selection methods that address Multi-Omics settings were entertained. Feature Engineering is crucial for the effectiveness of Machine and Deep Learning models. Here, pertinent background and simple illustrative examples in the form of a narrative served to catalyze a Systems view as a source of motivation for the development of Machine Learning models. Substantial attention, dedication and cognisance of relevant research gaps and challenges that must be embraced by the endeavours aimed at further advancing the field.

3.7.1. Future Trends

Firstly, the prediction of the risk of common diseases using single nucleotide polymorphism (SNP) genotyping data is introduced. This is a form of early risk stratification that abnormal tissue of future cases may be found at cancer initiation. Using SNP genotyping data, machine learning (ML) models can be built for the accurate prediction of the disease risk. Interpretation of hundreds of common variants as predicting the risk of a complex disease based on ML models built from genotype measurements of the known risk variants is then seen. Finally, I discuss possible future trends in the machine learning models for predictive genomics emerged. The concept of predicting an individual's risk of disease is increasingly discussed. When atypical histology is considered as the beginning of cancer, abnormal tissue of future cases may have been detected at cancer initiation. Feature measurements of such abnormal tissues can be made, and computational models have been built to predict the future risk of cancer by learners. The model proposed, however, considers the case that abnormal tissues of future cases have not been observed or cannot be easily obtained. In this case, one alternative approach is that one might expect to predict an individual's risk of disease

based on some easily measurable data, such as blood samples, which would not be affected by the presence of the future abnormal tissues. The most promising approaches include the prediction of the risk of a few common diseases, possibly using a patient's prior family history and other known risk factors. The prediction of the risk of common diseases can be considered as an early risk stratification, and early detection is expected to improve the prognosis. The deviated tissue under the early risk stratification concept is not yet tumorous, but the patient biopsy is performed at the cancer initiation time. A comprehensive understanding of gene mutation and gene copy number changes based on the feature measurement of biopsy is required. Based on such multi-omics data, early risk stratification may be made between the indolent and threatening nascent cancer, which may guide personalized medicine. Nevertheless, it is still of interest to predict the risk of a complex disease based on single data. In particular, the interpretation of hundreds of common variants may lead to the prediction of risk as low, average, and high value. Various machine learning (ML) models built to interpret the hundreds of common variants as accurately predicting the risk of a complex disease based on ML models built from the genotype measurements of the known risk variants are discussed.

References

Challa, S. R. (2023). The Role of Artificial Intelligence in Wealth Advisory: Enhancing Personalized Investment Strategies Through DataDriven Decision Making. International Journal of Finance (IJFIN), 36(6), 26-46.

Burugulla, J. K. R. (2022). The Role of Cloud Computing in Revolutionizing Business Banking Services: A Case Study on American Express's Digital Financial Ecosystem. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3720.

Pamisetty, A. (2022). Enhancing Cloudnative Applications WITH Ai AND Ml: A Multicloud Strategy FOR Secure AND Scalable Business Operations. Migration Letters, 19(6), 1268-1284.

Pamisetty, V. (2023). Optimizing Public Service Delivery through AI and ML Driven Predictive Analytics: A Case Study on Taxation, Unclaimed Property, and Vendor Services. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 124-149.

Kumar, S. S., Singireddy, S., Nanan, B. P., Recharla, M., Gadi, A. L., & Paleti, S. (2025). Optimizing Edge Computing for Big Data Processing in Smart Cities. Metallurgical and Materials Engineering, 31(3), 31-39.