الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila

كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics

قسم الإعلام الآلي
Department of Computer Science

Domain: Mathematics and Computer Science

Thesis Presented to Fulfill the Partial Requirement
for **Master's Degree** in Computer Science

Specialty: Information Systems and Software Engineering

Prepared By: Amina Boufissiou

Supervised By:

Lamri Sayad

Entitled

Gene-disease association using Machine Learning

Jury Members

Bilal Lounnas	President
Lamri Sayad	Supervisor
Makhlouf Benazi	Examiner

Academic year: 2024/2025

Dedication

"In the name of Allah, the Most Gracious, the Most Merciful, I express my profound gratitude for His guidance and strength throughout this academic journey. Every effort undertaken has been with sincere devotion to Him.

This thesis stands as a testament of love to my parents, whose unwavering support and belief in me have been the foundation of my success. Their sacrifices have illuminated my path forward. Not forgetting my siblings, whose steadfast presence and encouragement have sustained me through challenges and celebrated my triumphs. With a special thanks to my youngest brother, who contributed the most and made a significant difference in my projects.

This gratitude extends to the innovative AI tools that provided valuable assistance throughout my research process, helping to refine my ideas and enhance the quality of this work.

I am especially grateful to all those who have played a role in my life's journey —friends and mentors. Each moment of support has played a part in making this work possible."

Acknowledgement

Above all, I thank Allah, the Almighty, who has bestowed upon me the courage, strength, and perseverance to complete this scholarly endeavor.

I would like to extend my profound gratitude to our esteemed professor Dr. Lamri Sayad for his exceptional guidance, enduring support, and invaluable expertise throughout the development of this thesis. His insightful feedback and constructive critiques have significantly enhanced the quality of this research.

With sincere gratitude, I thank my SIGL batch of 2025, a distinguished batch to which I proudly belong.

I wish to express my thanks to the honorable members of the thesis defense committee, who graciously agreed to evaluate this work and provide their expert assessment.

Furthermore, I acknowledge with gratitude all individuals who have contributed, whether directly or indirectly, to the successful completion of this thesis.

It is my earnest hope that this humble contribution meets the academic standards and adequately reflects the knowledge and skills acquired during this research period.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ORMDL3	ORMDL sphingolipid biosynthesis regulator 3
TTN	Titin gene
GWAS	Genome-Wide Association Study
SNP	Single Nucleotide Polymorphism
NGS	Next-Generation Sequencing
ATP	Adenosine Triphosphate
WGS	Whole Genome Sequencing
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DEG	Differentially Expressed Gene
WGCNA	Weighted Gene Co-expression Network Analysis
DCM	Dilated Cardiomyopathy
IgE	Immunoglobulin E antibody
GEO	Gene Expression Omnibus
NCBI	National Center for Biotechnology Information
MIAME	Minimum Information About a Microarray Experiment
MAGNet	Molecular Anatomy of the Genogram Network
NumPy	Numerical Python
SHAP	SHapley Additive exPlanations
GNN	Graph Neural Network
SVM	Support Vector Machine
S3VM	Semi-Supervised Support Vector Machine
PCA	Principal Component Analysis
DQN	Deep Q-Network
LASSO	Least Absolute Shrinkage and Selection Operator
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic

General Introduction

1. Study Context

The convergence of genomics and artificial intelligence has really changed the game for biomedical research, particularly when it comes to unraveling the genetic basis of complex diseases. Basically, gene-disease association studies are the first step to a revolution in the precision medicine field as they pinpoint the relationships between genetic variants and disease susceptibility, progression, and therapeutic response. But the steep increase in the amount of genomic data, coupled with the complex interactions of genes, and the environment, create quite a big problem for analysis.

Machine Learning (ML) has been the main factor which allowed researchers to decipher the biological patterns behind the issues in the genome at an unprecedented scale. These include decision trees, ensemble methods, and neural networks that can go beyond the limits of traditional statistical ways by finding new connections, disease risk predictions and therapeutic target justified path. This project is about using ML to analyze asthma, and cardiomyopathy. The main focus is on both coding and non-coding genetic variants (e.g., ORMDL3 and TTN). The main reasons are to deepen the mechanistic understanding and to open the way for clinical translation.

2. Research Problem

Despite enormous progress in genomic technologies and still existing the main gaps in gene-disease association studies remain:

Computational Limitations: Traditional approaches (e.g., GWAS) tend to be inadequate for polygenic interaction modeling and to give functional interpretation of non-coding parts.

Clinical Translation Barriers: Diverse data sets and ML models that are often quite opaque ("black-box" phenomenon) create the trust issue for the clinicians.

Representation Bias: The underrepresented diverse populations in the genomic databases will not only be underrepresented in the future but are a source of systematic bias in the research findings and in their applicability in the real world.

This study tries to find the answers to the three basic question:

Can ML techniques solve the problem of noisy and high dimensional data in asthma, and cardiomyopathy genomic data?

Which hybrid analytical methods (e.g., SHAP paired with GNN) best fit maintaining high performance while increasing the model explicability?

How can theoretical computer predictions from the algorithm be practically applied and thus become parts of clinical practice?

3. Project Objectives

This thesis aims to:

Construct a reliable robust machine learning framework that integrates GWAS data for asthma and cardiomyopathy.

Evaluate algorithms systematically (XGBoost, MLP, and possibly Graph Neural Networks) in terms of predictive performance and interpretability.

4. Manuscript Organization

Our thesis is organized into three chapters

Chapter 01: points out the key notions of Artificial Intelligence (AI) and Machine Learning (ML) fields, as well as the main types of learning - supervised, unsupervised, and reinforcement. Moreover, it provides definitions of most frequently used ML algorithms such as Decision Trees and Neural Networks, and examples of their applicability in genomics. This chapter is all about the emerging machine learning role in the process of scientific research promoting.

Chapter 02: consists of the literature review and basic information about genomics and a brief description of the main genetic diseases is given. The information is presented in the form of classification and characteristics. In addition to this, Chapter 2 is devoted to two most common genetically complicated diseases (asthma, cardiomyopathy) chosen as an example for our research and it gives the summary of the scientific publications, which are related to the genetic and computational parts of these diseases.

Chapter 03: details the method of creating and the actual running of the multi-genetic prediction system. It explains the parts of the process: data acquisition, feature selection, and model evaluation. Besides this, the chapter releases the technical information including the libraries and the hardware/software equipment, the conditions that the models adhere to, as well as verification activities and the performance of the implemented models.

This research attempts to incorporate significant innovation in the field of deep ML with such research of genomic lessons has rapidly led deep learning to accelerate the path of personalized remedies of respiratory and cardiovascular disease.

CHAPTER 1

Introduction to Artificial Intelligence and Machine Learning

CHAPTER 1

Introduction to Artificial Intelligence and Machine Learning

1.1 Introduction

Artificial Intelligence and its intersection with biology has revolutionized genomics, especially our ability to comprehend complex disease mechanisms. With the exponential growth of genomic data generation via high-throughput sequencing technologies, classical analytical methods are not readily applicable for gaining coherent information from large, multidimensional data sets. The importance of machine learning algorithms in genomic data analysis cannot be overstated as it provides an indispensable means to deal with genomic challenges, thereby facilitating the discovery of gene-disease associations, disease susceptibility, and personalized treatment strategies. This chapter also lays solid grounds to understand AI and ML concepts, what type of algorithms is available and used in practice in this field (including some typical applications in genomic research) and the theoretical fundamentals on which gene-disease association strength is based in order to contribute to these studies.

1.2 A Brief Definition of Artificial Intelligence (AI)

Artificial Intelligence (AI) is essentially a multi-variety academic discipline centered on intelligent computation systems that function in ways that mimic the human intellect. It includes learning, reasoning, decision-making, problem solving and perception of adapting to a new environment. This conceptual framework of AI is characterized on two dimensions: (1) systems that function in a human-like manner by mimicking cognitive processes, as opposed to systems that act intelligently; and (2) systems, whose performance is measured against human performance, as compared to those compared against theoretical notions of rationality – where rationality corresponds to a system's capability to select optimally actions given its knowledge and its goals. AI has come a long way since it emerged as an official field of study in the 1950s from a theoretical concept to a crucial aspect of today's technology in practical use from virtual

assistants and autonomous vehicles to sophisticated data analysis frameworks, while continuously expanding the boundaries of what constitutes machine intelligence, as shown in Figure 1.1. [1]

1.3 What is Machine Learning

Machine Learning (ML) is an advanced sub-domain of artificial intelligence (AI) that focuses on the creation of computational systems that can learn from data rather than through explicit programming. Arthur Samuel described ML as the "field of study that gives computers the ability to learn without being explicitly programmed" in 1959, and Tom Mitchell (1997) formalized it as any program that improves performance (P) on some tasks (T) with experience (E). ML algorithms don't go by rote commands; they learn patterns and then refine their own performance through iterative exposure to information. This approach allows systems to learn and enhance themselves without human interventions and perhaps explains why ML is being used in areas such as image recognition, language processing, and predictive analysis. The process of learning includes the use of training sets, which are sets of examples (each example is a training instance or sample). For example, spam filters are enhanced by examining already labeled e-mail. This basic ability to adapt to the data-based patterns delineates ML as a science, but also almost an art, what makes it the basis of AI today.
. [2]

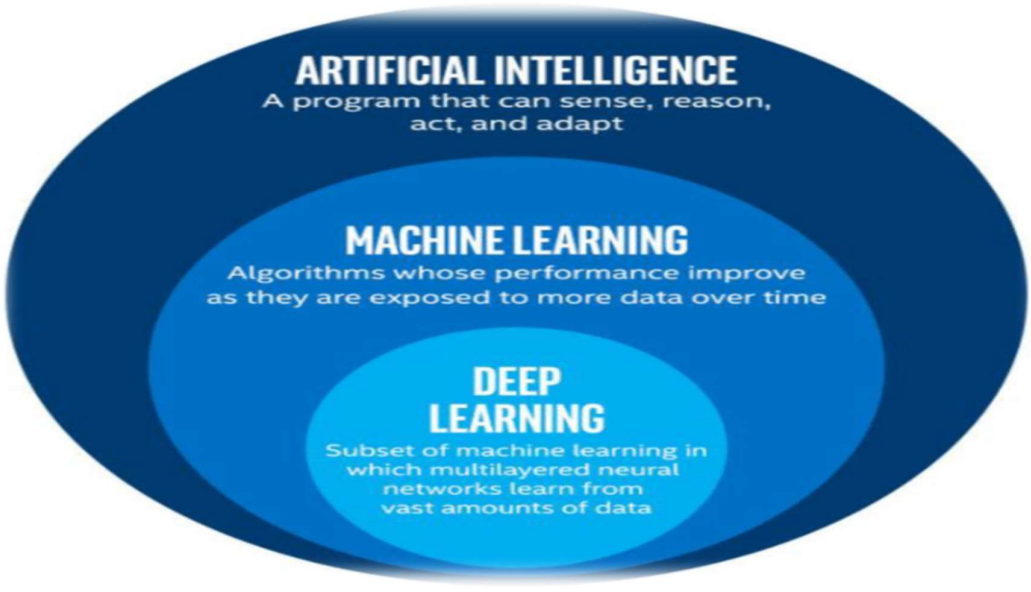

Figure 1.1: AI, ML, and Deep Learning Hierarchy -Venn Diagram [3]

1.4 Types of Machine Learning

Machine learning algorithms are organized into four main algorithm types, which include:

1.4.1 Supervised Machine Learning

Supervised learning is a machine learning paradigm in which an algorithm learns from labeled training data that is, mapping input features to known output labels and generalizing this mapping to make accurate predictions on unseen data. The latter forms the basis for most applications in classification spam detection, and regression continuous outputs like house price prediction. Some of the major algorithms available within this framework include Linear Regression Predicting numerical values, Logistic Regression Binary or multi-class classification, Decision Trees Rule-based hierarchical models, Random Forests An ensemble of decision trees to reduce overfitting Support Vector Machines (SVM) effective for high-dimensional data). The strength of supervised learning lies in situations where ample labeled data is available leading also to generally higher accuracy and interpretability however when labeling is a scarce or expensive resource it does not perform well. Common applications span healthcare (disease diagnosis), finance (credit scoring), and natural language processing (sentiment analysis). Challenges include overfitting (memorizing noise instead of learning patterns) and dependency on high-quality labeled datasets. [4]

1.4.2 Unsupervised Machine Learning

Unsupervised learning refers to the derivation of latent patterns from data that has not been labeled and for which outputs have not been specified. This type of learning is mainly applied in clustering (grouping similar data points, e.g., customer segmentation), dimensionality reduction (compressing data while preserving its structure, e.g., PCA), and association rule mining (discovering co-occurrence patterns, e.g., market basket analysis). Some important algorithms are K-Means Clustering (partitioning data into *k* clusters), Hierarchical Clustering (building nested clusters), DBSCAN (density-based clustering for irregular shapes), and Principal Component Analysis (PCA) the feature space reducer. Unsupervised learning proves useful in exploratory data analysis and fields where labeling is impractical—such as anomaly detection(fraud identification) or genomics (gene expression analysis)—but suffers from problems with subjective evaluation since there is no ground truth against which to validate it, and also interpretability especially with complex methods like deep clustering. [4]

1.4.3 Semi-Supervised Learning

Semi-supervised learning sits between supervised and unsupervised; it uses both labeled and unlabeled data. It works well when labeled data is not enough, but unlabeled data is plentiful; hence, it reduces annotation cost while improving generalization. Self-Training, Co-Training, and Label Propagation are the techniques of this method. In Self-Training, high-confidence predictions of one model are used for labeling. In Co-Training Long et al., multiple models train on different subsets of features. Label Propagation spreads the label information through graphs of similarity. Algorithms like Semi-Supervised SVM (S3VM) adapt traditional fields to incorporate unlabeled data. Some applications include medical imaging (where expert annotation is limited) and speech recognition (where large audio datasets need transcription). Semi-supervised learning can powerfully enhance a model's robustness but with pseudo-labels generated from unlabeled data that are noisy or incorrect.

1.4.4 Reinforcement Learning (RL)

Reinforcement learning, aka trial-and-error learning, is a method by which an agent learns to make sequential decisions through interaction with the environment so as to maximize cumulative rewards. It is characterized by delayed rewards (long-term strategic optimization e.g., game AI), exploration vs. exploitation (choice of whether to take novel actions or stick to known rewards), and policy optimization (e.g., Q-learning, Deep Q-Networks). RL algorithms like Q-Learning (model-free value iteration), DQN (Deep Q-Networks) (Q-learning joined with deep neural networks) and Policy Gradient Methods perform well in dynamic domains such as robotics (autonomous navigation) and gaming (AlphaGo). However, they are computationally very expensive and require properly engineered rewards. In contrast to supervised learning, reinforcement learning (RL) depends on environmental feedback rather than labeled input, which makes it perfect for complicated, adaptive systems. [5]

These categories are designed for different tasks; Supervised Learning works well when labeled data is available and needs to produce exact predictions while it is limited when data is unlabeled. Unsupervised Learning helps to identify hidden patterns in unlabeled data but there is no ground truth to assess performance on. Semi-Supervised Learning aims to achieve a better cost-performance tradeoff by using few labeled and many unlabeled examples. Reinforcement Learning learns optimal sequential decisions based on feedback from the environment, useful for complex and dynamic problems but computationally expensive. Several hybrid methods (e.g. semi-supervised RL) are also being developed to achieve adaptive AI.

1.5 Popular Machine Learning Algorithms

1.5.1 Decision Trees

Decision tree is a standard non-parametric supervised learning tool which aims to design a hierarchical tree-based form for decision making. This algorithm works by recursively dividing the input feature space into disjoint regions via binary splits, where each internal node is a decision rule on feature values and each leaf node stores a prediction. The architecture construction decides on which split criteria which maximize information gain or minimize impurity indices like Gini coefficient, or entropy.

Decision trees are especially useful in fields that need clear decision-making processes because of their exceptional interpretability, low data preparation requirements, and ability to handle both numerical and categorical variables with ease. [6]

1.5.2 Support Vector Machines (SVM)

Support Vector Machines are powerful learning algorithms formulated in terms of structural risk minimization and developed in statistical learning theory. The basic goal is to find the best hyperplane that separates the classes of data as much as possible in the feature space, which separates the decision boundaries as widely as possible from the nearest data points, called support vectors. SVMs are extremely flexible through kernel functions: the transformation of non-linearly separable problems to higher dimensions in which separation is possible. This approach demonstrates excellent generalization. SVMs are a fundamental approach for both classification and regression problems since they are especially efficient in high-dimensional spaces and are computationally efficient even when training with small quantities of data. [7]

1.5.3 Neural Networks

Neural network are computational structures loosely modelled after the central nervous system, involving a set of relatively simple interconnected processing units, also called neurons, arranged in multiple layers which work together to approximate complex non-linear functions. Such systems are based on weighted connections between neurons, in which information flows using activation functions to compute the output of multiple input signals. Learning is iterative updating of the weights using backward propagation algorithms that minimize prediction errors across training sets. Neural networks are very powerful in pattern recognition, feature extraction, and function approximation, deep architectures have capacity

to model the increasingly complex relationship within high dimensional data space, especially good at computer vision, natural language processing and complex prediction tasks. [8]

1.5.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors stands for a non-parametric, instance-based learning algorithm that goes over the local neighborhood structure in the feature space to make predictions. The method is based on the assumption that similar instances will have similar target values. It measures the distance (typically Euclidean) to find the k closest training examples to a query point. Classification decisions are made by the majority vote of those neighbors, while regression predictions use the averaging method. The simplicity of KNN is the source of its power - it needs no explicit training phase, and it changes dynamically according to local data patterns. On the other hand, computational complexity grows monotonically with the dataset size, and performance is very sensitive to the choice of a suitable distance metric and the best value of k obtained during cross-validation procedures. [9] [10]

1.5.5 Machine Learning Algorithms in Genomics Research

Random Forest (RF): Random Forest (***based on decision trees***) is an ensemble algorithm that briefly combines multiple decision trees for regression tasks, or uses a majority vote for classification tasks. It also significantly improves the performance of a single decision tree by way of reducing the overfitting problem, which happens to be a result of the randomness introduced during the creation of individual trees. In genomics, RF is especially good for gene expression analysis and biomarker identification. [11]

Advantages*:* Can easily work with high-dimensional genomic data and resistant to overfitting with the help of built-in feature selection, computationally efficient, and gives feature importance rankings.

Limitations*:* Decreased interpretability than with the case of single decision trees, it can be biased toward categorical variables, and it does not perform well when dealing with complex gene-gene interactions.

XGBoost (Extreme Gradient Boosting): XGBoost (***based on decision trees***) applies regularization methods that are at an advanced level to decrease the weights, avoid overfitting, and improve thereby its performance in real-world problems. The algorithm implementation makes it possible to cache data and use several CPU cores for quick processing. It is a gradient boosting framework that is specifically designed for genomic prediction problems and is optimized for speed and performance. [12]

***Advantages**:* Outperformance in accuracy of genomic applications, inbuilt regularization to prevent overfitting, efficient parallel processing, and automatic handling of missing values.

***Limitations**:* Needs a lot of efforts in tuning of hyperparameters, more computationally expensive compared to Random Forest, and could be overfitting-prone when working with small genomic datasets.

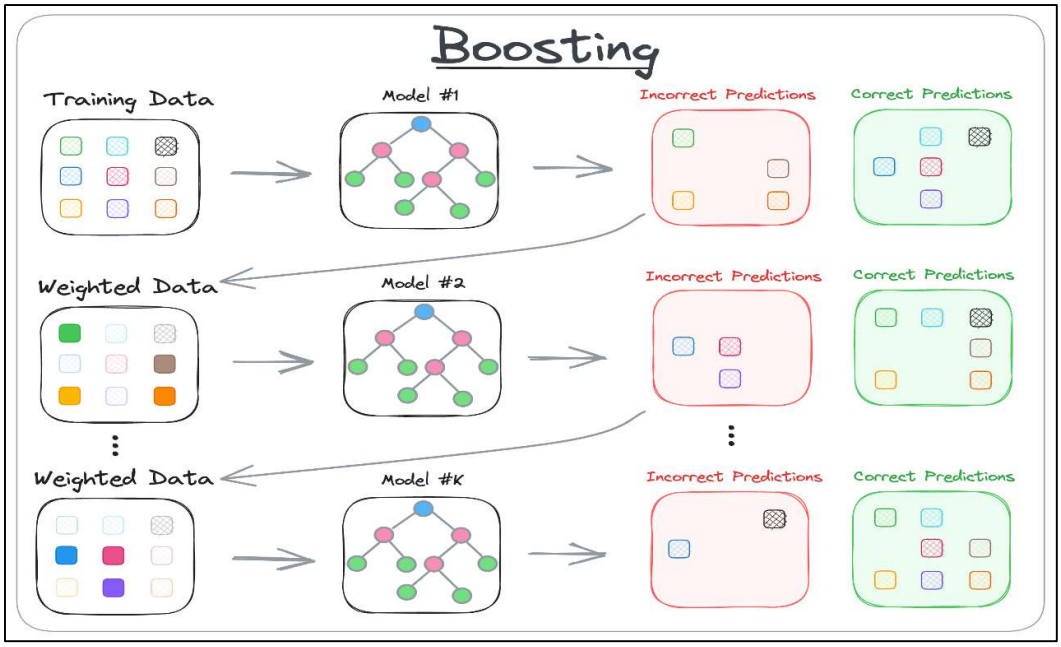

Figure 1.2: XGBoost Algorithm Flowchart [13]

Multilayer Perceptron (MLP): MLP (***is a type of neural network***) is a feedforward artificial neural network with at least three node levels; an input, one or more hidden layers, and an output layer. MLP is a type of feedforward neural network consisting of fully connected neurons with a nonlinear activation function, most commonly used to separate the data that is not linearly separable. The MLPs are used in genomics for the purpose of gene expression and sequence analysis, where they perform complex pattern recognition. [14]

***Advantages**:* Can model complex nonlinear relationships in genomic data, universal function approximator, and the architecture is flexible and can be changed to fit different genomic problems.

***Limitations**:* It needs a lot of data in order to train stably, it is overfitting-prone with only a few genomic samples, and the computation is too expensive since hyperparameter optimization is very necessary.

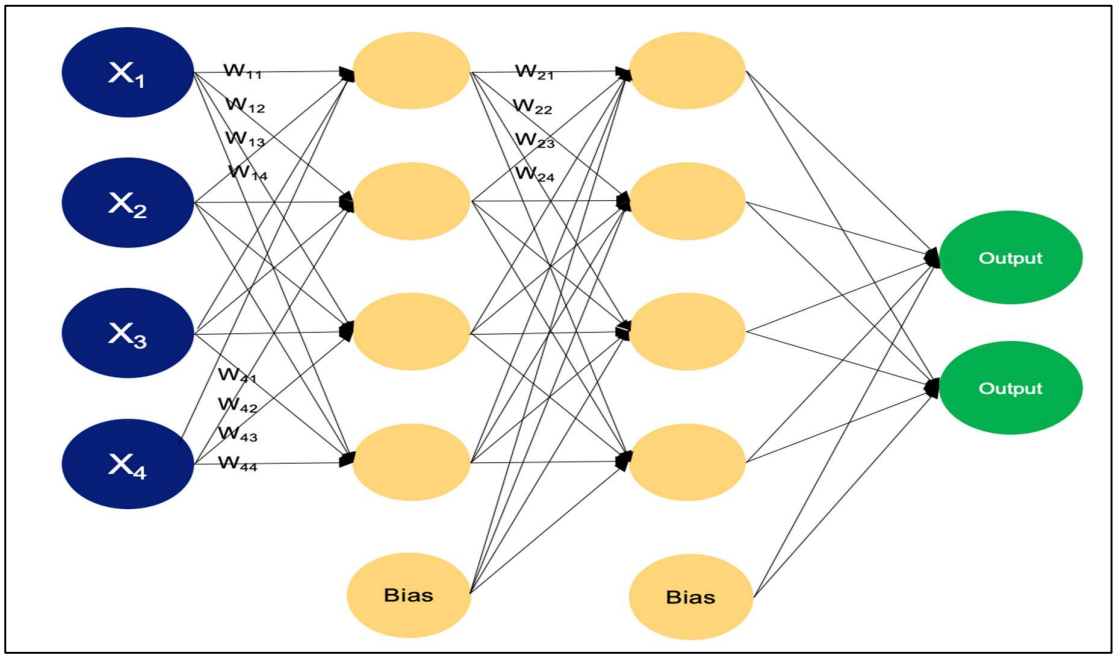

Figure 1.3: Multi-Layer Perceptron Architecture (MLP) [15]

1.6 Preprocessing and Evaluation Methods

1.6.1 Feature Selection Methodologies

Feature Transformation: The numerical alteration of original gene data that have the greatest potential to improve model performance such as log-transformation, normalization, and standardization, all of which are techniques applied to gene expression values.

Feature Engineering: The procedure of designing new features or changing those that are already in order to increase a machine learning model's accuracy. This also involves the use of StandardScaler for standardization which changes features to have zero mean and unit variance ($z = (x - \mu)/\sigma$).

SelectKBest Algorithm: A statistical test involving only one variable at a time that selects the highest-ranking k- features, according to the given score function.

Statistical Feature Selection: A method of choosing the relevant features based on performance of statistical tests for the purpose of identifying the genes that have a significant differential expression between the disease and control groups. For instance, the F-test (f_classif) can be used through SelectKBest.

Dimensionality Reduction: It is the process of lowering the number of input variables while at the same time keeping the major information. This study utilizes Principal Component Analysis (PCA) with 150 components as one of the alternatives in case SelectKBest fails, changing the

high dimensional gene expression data into a more understandable lower dimensional representation.

Synthetic Minority Oversampling Technique (SMOTE): The method of data augmentation is employed to tackle class imbalance by creating synthetic examples for the minority classes via interpolation among the existing minority class instances, thus implementing it to the asthma and cardiomyopathy datasets.

1.6.2 Evaluation Tools

ANOVA F-test: Analysis of Variance Features test is a method that is based on statistics and works by calculating the proportion between variance that comes from different groups and that which is from the same group, hence it is used to evaluate the significance of the gene expression differences between disease phenotypes. The F-statistic formulates feature importance as F = MSB/MSW, where MSB is mean square between groups and MSW is mean square within groups.

Genetic Biomarker Extraction: A genetic trait that can be measured reliably is going to be the indicator of normal biological processes, disease, or a reaction to the drug after an intervention. Here, certain gene expression patterns act as symptom-related markers of asthma and cardiomyopathy.

Feature Importance Ranking: A technique applied to determine the extent of each genetic variable's impact on the predictive model's accuracy. The XGBoost feature importance was the basis for cardiomyopathy prediction, while F-test scores were employed for the ranking of asthma-related genes.

Cross-Disease Feature Analysis: The examination of genetic factors over several disease types in order to uncover common pathogenic routes and disease-specific indicators thus allowing the construction of multi-disease prediction models.

Pathway-Based Feature Grouping: The pooling of the genetic features on the basis of the biological pathways they are part of which makes it easier to detect functionally related genes ones that are driving disease pathogenesis.

Information-Theoretic Methods: These are approaches for feature selection which are based on principles of information theory such as mutual information and information gain and which are used for quantification of the statistical dependence between the features and the target variables. On this basis of these numbers the most informative genetic markers are obtained.

1.7 The Importance of Machine Learning in Scientific Research

Machine Learning is now the only option for scientists to do their research in a better way and in the fields that require them to deal with large datasets, genomics being the one here. ML helps researchers to identify complex issues with the data, to be able to make correct predictions, and also to automate analyses which will free up their time for more important tasks that would not be possible to do manually. Among others, an example here is that ML is being applied in the studies of association between genes and diseases in order to find the genetic markers that are related to a particular disease. This will make it possible for the researchers to handle the endless genomic datasets and find the hidden links between the gene mutations and the manifestation of the diseases.

1.8 Challenges in Machine Learning

Machine learning is encountering major issues that are severely limiting its capabilities and the widespread use of the technology in various areas. Although there are great algorithmic achievements, these impediments to progress are still there and they slow down the creation of solid and trustworthy systems. The main problems mentioned were data-related issues such as the lack of volume, bad quality, and class imbalance, which in turn influence the extent of the model's generalization ability. Besides that, the intricacy of present-day algorithms makes it difficult to understand them, especially those areas, that are of highest impact and hence need transparent decision-making processes.

Key challenges include:

- **Data Quality Issues**: Number of data points less than necessary, missing issues noise caused by data, and imbalance in the classes hurt model performance.
- **Interpretability Problems**: Complex algorithms have a "black box" characteristic that hinders their use in important areas.
- **Overfitting and Generalization**: Cases where models succeed on observed data but are not able to predict new data.
- **Computational Complexity**: Problems related to scaling with high-dimensional datasets and shortages of resources in the environment.
- **Algorithmic Bias:** Propagating social inequity through biased datasets or those lacking suitable assumptions might be a risk that.
- **Concept Drift**: There are changes in distribution that led to reduced performance of a model if it relies only on the past. [16]

1.9 Conclusion

This chapter focused on Artificial Intelligence and Machine Learning basics and we emphasized their value to genomics field. We spoke about different machine learning paradigms like supervised, unsupervised, semi-supervised, and reinforcement learning and listed the most popular algorithms to be decision trees, support vector machines, neural networks, and k-nearest neighbors. Then, we illustrated that machine learning has changed the face of genomics as a result of its applications in the process of variant calling, genome-wide association studies, and disease prediction, however, we also pointed out that challenges are still there such as data quality issues, interpretability problems, and computational complexity. This overall outline sets the theoretical base for gene-disease association studies. In the next chapter, the previous research on gene-disease association with the help of Machine Learning will be reviewed, their findings will be analyzed, and the limitations of existing approaches will be discussed.

CHAPTER 2

Literature Review and Background in Genomics

CHAPTER 2

Literature Review and Background in Genomics

2.1 Introduction

The genomics area has been changed drastically after the introduction of the high-throughput sequencing technologies and computational biology methods. One of the tasks of the molecular understanding of genetic diseases is expounding the entire genome principle, the disease's mechanisms, and the advanced analytical methods. This chapter is a logical and detailed presentation of the genomics' basics, the diseases' classifications by the genes, and the machine learning's main part in the genomics' informatics. We explore the history of the computer methods in the studies of gene-disease relationships, discuss the consistent research ways, and find out the obstacles, which affect the new work direction. We first provide the theoretical foundation here by carrying out a comprehensive literature review, which we have found that still has a number of shortcomings, especially when it comes to our chosen subject, the use of machine learning methods for searching for genetic associations of diseases such as asthma and cardiomyopathy.

2.2 An overview of Genomics

Genomics is an interdisciplinary field of molecular biology that is involved with the structure, function, evolution, mapping, and editing of genomes. A genome is the complete set of an organism's DNA, without ignoring the genes and the three-dimensional hierarchical structural configuration of the DNA. The field of genomics is no longer restricted to genetics alone. It now deals with gene interactions, regulatory elements, and non-coding sequences that collectively contribute to biological functions. NGS is capable of sequencing millions of DNA fragments simultaneously, revealing the genomes' structures, genetic differences, gene expression and fluctuations in gene behavior at a very detailed level. The breakthrough in genomics owes its existence to the next-generation sequencing platforms that, while keeping the costs low, have enormously increased accuracy and throughput. Whole genome sequencing (WGS) is emerging as the main technique for molecular genetic diagnosis of rare and

unidentified diseases, as well as for the discovery of cancer driver genes. At present, genomics is the central pillar of contemporary biological and medical research as it is from the basic to the clinical sectors that personalized medicine is established through genomics. [17]

Here are some definitions of some of the most important concepts in genomics:

2.2.1 DNA

DNA is a long molecule that contains the genetic instructions for the growth, functioning, and development of a living organism. It works just like a blueprint of the body and is present in almost every cell. [18]

2.2.2 RNA

RNA carries out the instructions of DNA. It plays a vital role in protein synthesis by acting as a message bearer between the DNA and other parts of the cell. [19]

2.2.3 Protein

Proteins are large molecules made up of amino acids. They are very crucial in our body as they help build body structures, carry out different chemical reactions, help in immune function and communication within the body. [20]

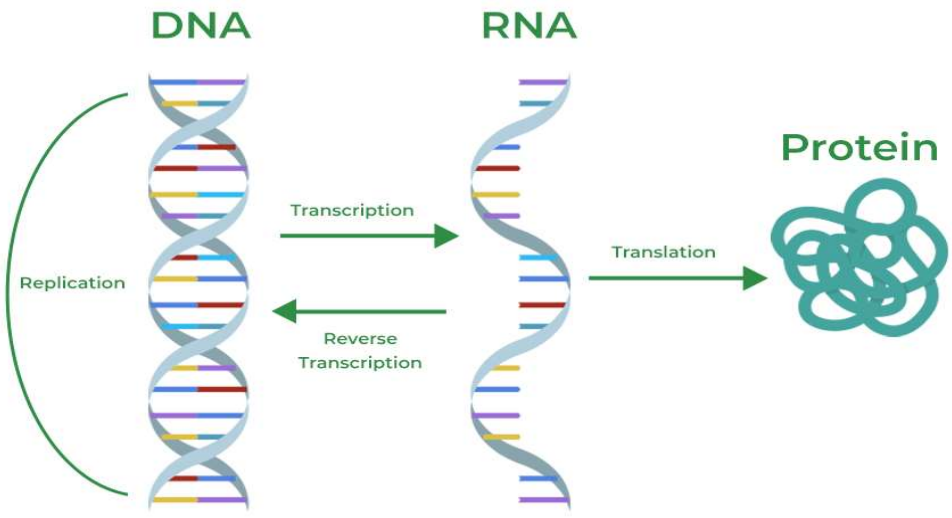

Figure 2.1: DNA Structure and Gene Expression - Central dogma illustration [21]

2.2.4 Chromosome

A chromosome is a thread-like structure of DNA and protein within the cell nucleus. It compacts and orders genetic material; humans normally have 46 chromosomes per cell. [22]

2.2.5 Mitochondria

Mitochondria are small structures within cells. They are known as the "powerhouses" of the cell because they generate energy in a form that is usable by the cell. That energy is ATP, which allows the cell to carry out its functions. [23]

2.3 Fundamentals of Genetic Diseases

A genetic disorder is a health challenge that comes from the presence of one or more faults in the genome. Thus, a mutation in one gene (monogenic) or different genes (polygenic) or a chromosome abnormality can be the cause of the disorder. Genetic diseases are still the most diverse group of conditions, which are caused by changes in the DNA sequence, chromosomal structure, or gene regulation that make cellular processes go wrong. The cause of a genetic disorder can be a mutation of one gene (a monogenic disorder), several genes (multifactorial disorder), a combination of mutations of genes and environmental factors, or damage to chromosomes, which forms a complex classification system based on the molecular mechanisms of the disorders. Genetic disorders are caused by changes in a person's DNA sequence. Our DNA sequence gives instructions for making the proteins that are needed for the function of the cells, and if there are interruptions, the proteins are not working properly, and diseases will appear. The pathophysiology consists of the loss-of-function mutations, gain-of-function mutations, and dominant-negative effects... mutant proteins interfere with normal cellular processes. [24]

2.4 Types of Genetic Diseases and Their Characteristics

Although a genetic disorder can be a disease, it is not always the same. A DNA sequence change in an individual's DNA, a mutation in one gene (monogenic) or in a few genes (polygenic), or an abnormal chromosome can be the cause of a genetic disorder. [25]

2.4.1 Monogenic Diseases

A monogenic disease means that only one gene can determine a genetic disease, due to the alteration of a single gene as a result of mutations in the gene's protein which lead to the gene being dysfunctional. These diseases choose to be rare, more severe in manifestation, and the number of affected individuals being significantly lower compared to polygenic diseases. Such

diseases follow exact Mendelian inheritance patterns (autosomal dominant, autosomal recessive, or X-linked)

2.4.2 Polygenic Diseases

Polygenic diseases are the illnesses that are produced by the action of a lot of genes going on simultaneously, and usually giving the environmental factors, they are combo. The polygenic disorders are the most abundant, but it is hard to figure out their hereditary models, and usually they are the result of changing many genes plus a small effect to the total risks of a disease. [26] [27]

2.4.3 Chromosomal Disorders

A chromosomal abnormality, or chromosomal aberration, is a disorder characterized by a morphological or numerical alteration in single or multiple chromosomes, affecting autosomes, sex chromosomes, or both. A numerical abnormality means an individual is either missing one of the chromosomes from a pair or has more than two chromosomes instead of a pair. A structural abnormality means the chromosome's structure has been altered. Deviation from the normal diploid complement of 46 chromosomes is referred to as aneuploidy; an extra chromosome results in trisomy, whereas a missing chromosome results in monosomy. [28]

2.4.4 Mitochondrial Diseases

Mitochondrial disorders are a set of monogenic diseases which are genetically and phenotypically quite diverse. The principal feature of mitochondrial diseases is an abnormal oxidative phosphorylation. Mitochondrial diseases are monogenic conditions that have a defect in the process of oxidative phosphorylation and that are caused by pathogenic variants in one of over 340 different genes. These disorders affect mitochondrial DNA which is inherited only from the mother and as a result, depending on the number of mitochondria, the disruption of one system can lead to the involvement of several organ systems since energy is necessary for all cells to function. [29]

2.5 Relation of Machine Learning and Genomics

Machine learning has completely changed genomics research as it has brought specialized computational methods that can handle enormous genetic datasets and produce biological knowledge. Growth in data technologies that generate high-throughput data in human genomics has made AI, in particular, deep learning methods, an indispensable tool in managing the complexity of genomic datasets. Machine learning has thus found its place in the field of genomics through deep learning technologies delivering remarkable accomplishments. The main usages involved are variant calling, where Deep Variant, an AI model, is the best among

the whole community of tools in accuracy. And genome-wide association studies (GWAS), where machine learning methods such as gradient boosting, random forest, SVM, and neural networks are employed to prioritize the most important GWAS SNPs and genes; identify epistasis among the selected variants. Moreover, ML algorithms make it easier to detect and diagnose various human diseases through genomics and proteomics that have been technically evolved. At the same time, these big and intricate data set open new windows for the understanding of the stage and pathogenesis of genetic diseases, the rules for health and the prediction of well-being. These instances of the use of the machine learning capacity indicate that genomics play a crucial role in the transformation of the clinical application of the genomic findings and the personalized medicine methods. [30] [31]

2.6 The Applications of Machine Learning in Genomic Analysis

Machine learning today is a major computational paradigm for solving the exponential increase and complexity of genomic information. The high-throughput data-generating technologies in human genomics have left us with a vast amount of genomic data. To unveil the knowledge and pattern from this genomic data, artificial intelligence in particular deep learning methods has been the lead. With the genome data becoming more and more complicated, researchers are relying on artificial intelligence and machine learning as tools to uncover hidden patterns for the benefit of health and science. [32]

2.6.1 Pattern Recognition in Large-Scale Genetic Datasets

Machine learning algorithms are perfect at discovering complex patterns in high-dimensional genomic datasets that are beyond the reach of traditional statistical methods. Gene signatures that are given by a set of genes contain the predicting power that may be used for our treatment strategies that are more personalized by the patients that are the most likely to be identified as benefiting from the treatment. The core of the problem is the large data dealing, where the features (genes) often outnumber the samples, hence, the requirement for specialized algorithms that are capable of managing the high-dimensional sparse data. [33]

2.6.2 Identification of Biomarkers for Disease Diagnosis and Prognosis

The high-throughput profiling of gene expression technologies, DNA microarrays, and RNA sequencing provide vast gene expression data sets that allow the discovery of biomarkers through data analysis. Different genes that are significantly expressed between two or more conditions are often identified as biomarkers for particular diseases using traditional tests. The feature selection technique has been at the forefront in this aspect during these past few years. To demonstrate how accurate the biomarkers are, researchers usually do the receiver operating

characteristic (ROC) curve analysis that is a plot of the true positive rate or sensitivity versus the false positive rate or 1-specificity. Machine learning methods that include the LASSO regression algorithm and the SVM-RFE algorithm have been employed to superpose the candidate biomarkers with the top accuracy through the (ROC) curve analysis. [34]

2.7 Gene-disease association prediction

The development of the medical care precision medicine concept has displaced the traditional symptom-driven treatment process by enabling early risk prediction of disease through better diagnostics and the customization of more effective treatments. ML models can make gene-disease association predictions by extracting patterns from the present genomic databases, pinpointing the disease susceptibility genes that are not known so far, and forecasting the pathogenic variants. [35]

2.7.1 Related Work and Previous Studies

Machine learning (ML) algorithms can diagnose more accurately and faster by going through a huge amount of genomic data and detecting complex multiallelic patterns that are possibly related to certain diseases. Research papers exhaustively in this domain look at the methodological approaches, performance evaluation metrics, and comparative analysis of different computational frameworks. Publications try to address the disproportionately few numbers of samples compared to the enormous number of variants, discussing which ML techniques are utilized and presenting newly obtained results with deep neural networks. [36]

In respiratory diseases: In respiratory diseases, particularly asthma, differentially expressed genes (DEGs) in asthmatic subjects compared with controls were identified using significance threshold adjusted P-value < 0.05, followed by weighted gene co-expression network analysis (WGCNA) to construct gene co-expression modules. WGCNA has thus associated the key genes with the neutrophilic asthma. These genes are “IRFG, IRF1, STAT1, IFIH1, IFIT3, GBP1, GBP5, IFIT2, CXCL9, and CXCL11”. Asthma diagnosis and risk stratification have been facilitated by constructing predictive genetic panels, utilizing machine learning algorithms like support vector machines, random forest, and elastic net regression. [37]

In cardiovascular diseases: In regard to cardiovascular diseases, machine learning applications have been concentrating on the discovery of genetic variants which are linked to different cardiomyopathy subtypes. A study has shown that the variants of the gene located

upstream of the MYH7 enhancer gene (rs875908) are with dilated cardiomyopathy (DCM) possibly associated. Recently, research has implemented (WGCNA) together with machine learning algorithms for the investigation of the major genes and immune cells in combination with filtration in heart failure resulting from ischemic cardiomyopathy, thus they are able to identify the most important molecular pathways that participate in disease progression. In addition to that, the research on genome-wide association of hypertrophic cardiomyopathy discovered common genetic variants and changed risk factors that support disease susceptibility and expressivity. [38]

However, those studies are limited often by small sample size, the need of population stratification bias, and focus mainly on the single nucleotide polymorphisms while forgetting the complex structural variants and the interactions between the genes.

2.7.2 Challenges of Previous Studies and Research Gaps

This part of the paper assesses the main defects and methodological limitations, which have hampered the gene-disease association models development.

Methodological Limitations: Most research to date is focused on single-disease prediction models that limit clinical applicability in situations where patients exhibit multiple comorbidities. Besides that, insufficient feature selection techniques as well as an inadequate number of machine learning algorithm comparison remain the main issues. [39]

Validation Issues: Generally, there is still a lack of support that the results found will be confirmed elsewhere through external datasets, long-term follow-ups, and clinical trials for predicted biomarkers. Without standardized evaluation metrics, it is challenging to match the results of various methodologies with each other.

Research Gaps: At this point of time, research is far behind the situation when multi-disease prediction models become comprehensive and there are standardized evaluation frameworks for genomic machine learning applications. [40]

2.7.3 Motivation for Current Research

This research is motivated by the need to Isolate the limitations of the current gene-disease association studies, that is to say, the lack of extensive multi-disease prediction models.

This study focuses on **Asthma** and **Cardiomyopathy** due to three key reasons. First, there is a lack of research addressing both diseases together, despite strong evidence of their link (research shows that the patients who are suffering from asthma are the ones who have the

highest risk of cardiovascular diseases since the same inflammatory pathways). Studies show that asthma patients have a (32% to 42%) higher risk of heart failure and cardiac diseases. Another large-scale study reported a 2.14 times higher hazard of facing heart failure in adults with asthma. Second, these diseases are widespread and chronic. Asthma affects over 300 million people globally, this disease is the cause of around 450,000 deaths per year, while cardiomyopathy affects around 2.5 million, often leading to severe outcomes. Third, early prediction using genetic data could help save lives by allowing for earlier medical intervention.

More specifically, the reason for coming up with a multi-disease prediction system for asthma and cardiomyopathy using MLP and XGBoost algorithms lies in the fact that these two diseases have the same inflammatory pathways and are exposed to similar genetic risk factors.

This research is attempting to make an innovative input by conducting comprehensive feature selection techniques and implementing rigorous validation procedures, thereby removing the methodological limitations that have been pointed out in the current study.

2.7.4 Disease Genetic Background and Clinical Relevance of asthma and cardiomyopathy

Understanding the genetic complexity and clinical significance of asthma and cardiomyopathy is essential for developing effective machine learning-based prediction models. Both diseases exhibit distinct genetic architectures that necessitate different computational approaches for accurate risk assessment and clinical decision support.

2.7.5 Asthma: Pathophysiology and Genetic Complexity

Asthma is a complex chronic inflammatory disease characterized by heterogeneous phenotypes and multifactorial etiology involving genetic predisposition, specific IgE to respiratory allergens, and overactive immune responses. The disease exhibits significant genetic complexity with multiple susceptibility loci contributing to its polygenic nature, where asthma susceptibility is strongly influenced by genetics alongside environmental factors.

The polygenic architecture of asthma, involving hundreds of genetic variants with small individual effects, creates challenges for traditional statistical methods but provides opportunities for machine learning algorithms to identify meaningful patterns. This genetic heterogeneity justifies the use of complex modeling approaches such as multi-layer perceptron, which can capture intricate relationships between multiple genetic markers in disease prediction. [41]

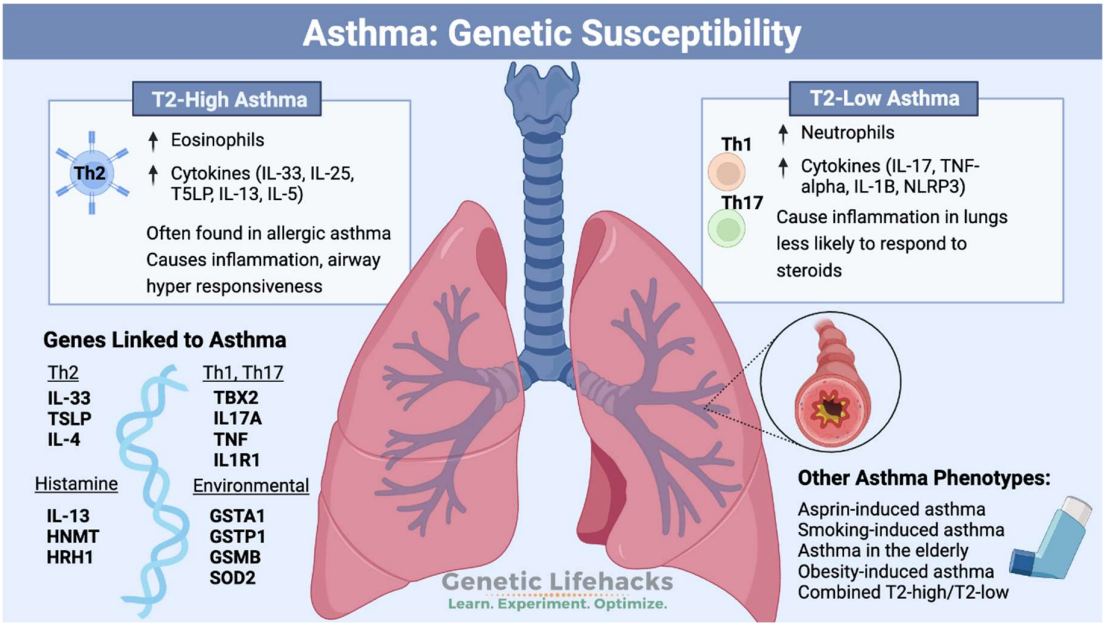

Figure 2.2: Asthma Pathophysiology Diagram [42]

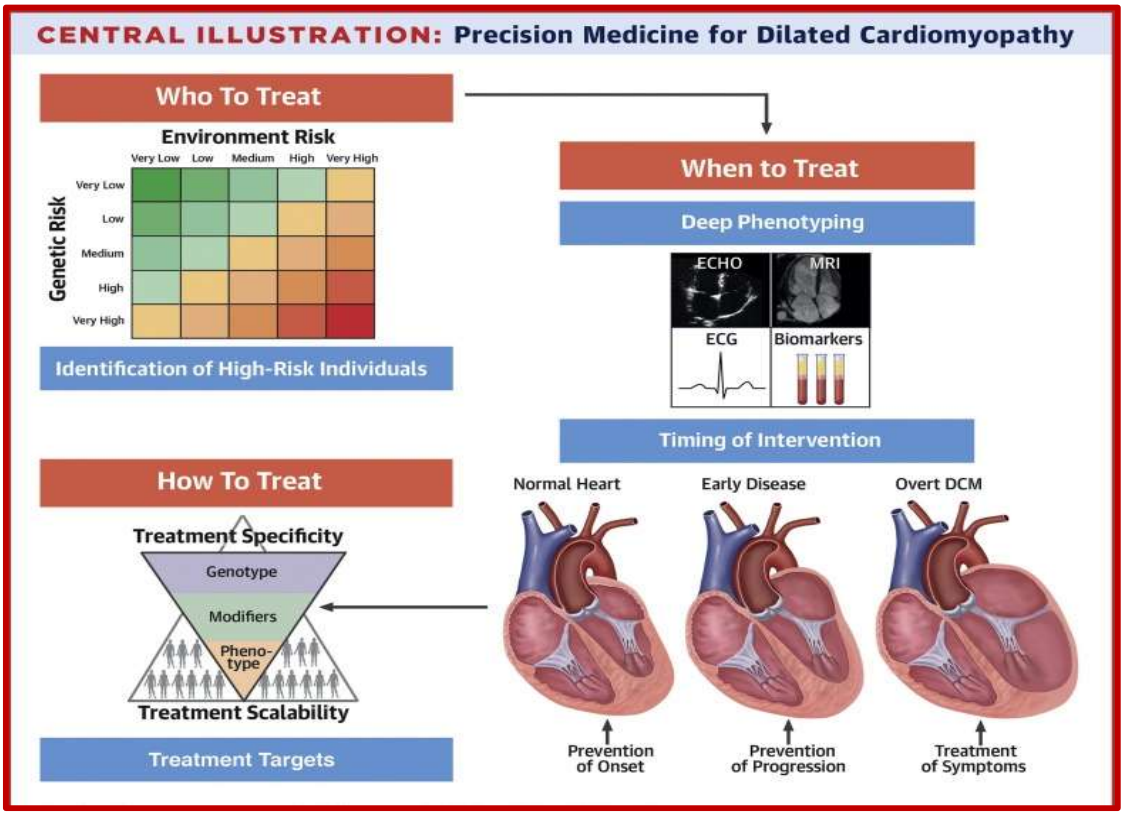

Figure 2.3: Cardiomyopathy Types Illustration [43]

2.7.6 Cardiomyopathy: Genetic Mechanisms and Clinical Subtypes

Cardiomyopathy encompasses a diverse group of heart muscle diseases marked by structural and functional abnormalities not primarily caused by coronary artery disease. The disease exhibits both monogenic and polygenic inheritance patterns, with primary subtypes including dilated, hypertrophic, and restrictive forms, each having distinct genetic profiles.

MYH7 and MYBPC3 are the most common causal genes, responsible for approximately 40% of hypertrophic cardiomyopathy cases. The relatively high penetrance of pathogenic variants, combined with clinical urgency for risk stratification and genetic counseling implications, necessitates interpretable prediction models. This genetic architecture supports tree-based algorithms like XGBoost, which provide both high predictive accuracy and interpretable feature importance rankings essential for clinical decision-making. [44]

2.8 Conclusion

This chapter laid down an extensive foundation for the understanding of genomics and genetic diseases. Furthermore, we explored the basic principles of gene-disease associations and also looked at the various classes of disorders, that is, from a single-gene to the complex polygenic disorders. We went on to explain how machine learning has changed genomic analysis with its powerful algorithms that are able to handle high-dimensional genetic data. The systematic review we conducted has exposed considerable advances in gene-disease association studies, particularly in respiratory and cardiovascular diseases while also pinpointing the persistent limitations in these studies including methodological issues, dataset biases, and the disproportionate representation of single-disease prediction models.

These research gaps serve as the scientific premise for the creation of an integrated prediction system for asthma and cardiomyopathy employing complementary machine learning algorithms. Our next chapter will feature the presentation of our suggested methodology for multi-disease genetic prediction; we will be giving a detailed account of the algorithmic approaches and the validation strategies that are consistent with the literature that the limitations identified in this review.

CHAPTER 3

Methodology and Implementation of Multi-Disease Genetic Prediction System

CHAPTER 3

Methodology and Implementation of Multi-Disease Genetic Prediction System

3.1 Introduction

This chapter delves into the methodology used in the development of machine learning models that can effectively identify the gene-disease relations for asthma and cardiomyopathy. We employ the notion outlined in the earlier chapters in illustrating our systematic approach which involves activities such as data harvesting, purification, feature extraction, and model construction.

The method is a mix of clinical knowledge about disease pathophysiology and computational techniques for building predictive models. We discuss in detail the use of the XGBoost and Multi-Layer Perceptron architectures to the same framework, setting up evaluation criteria that provide model trustworthiness and clinical use.

3.2 Overview of the Proposed Methodology

This methodology introduces a dual-algorithm framework for multi-disease prediction aimed at asthma and cardiomyopathy. Machine learning is a new tool that is expected to catch disease very quickly, however, it still faces some problems like lack of labeled data and dataset imbalance which prevent it from being able to build correct prediction models.

3.2.1 Multi-Disease Approach

The outlined model has three stages: (a) Data normalization(b) Weighted normalized feature extraction and (c) prediction. It not only eliminates the restrictions of single-disease prediction models but also provides comprehensive health evaluation.

3.2.2 Disease Selection Rationale

Asthma and cardiomyopathy were selected in an effort to compare the two diseases with a focus on the similarities and differences as well as selecting suitable samples for the study. Along the line. The epidemiological linkage is what makes the pair of these diseases the most suitable ones for integrated prediction modeling instead of a random one.

3.2.3 Dual Algorithm Framework

XGBoost was used for cardiomyopathy prediction whereas MLP was for asthma prediction. A disease prediction model is developed for model training and performance evaluation with other algorithms. In a single evaluation framework, the model achieved a high accuracy of in medical diagnosis applications.

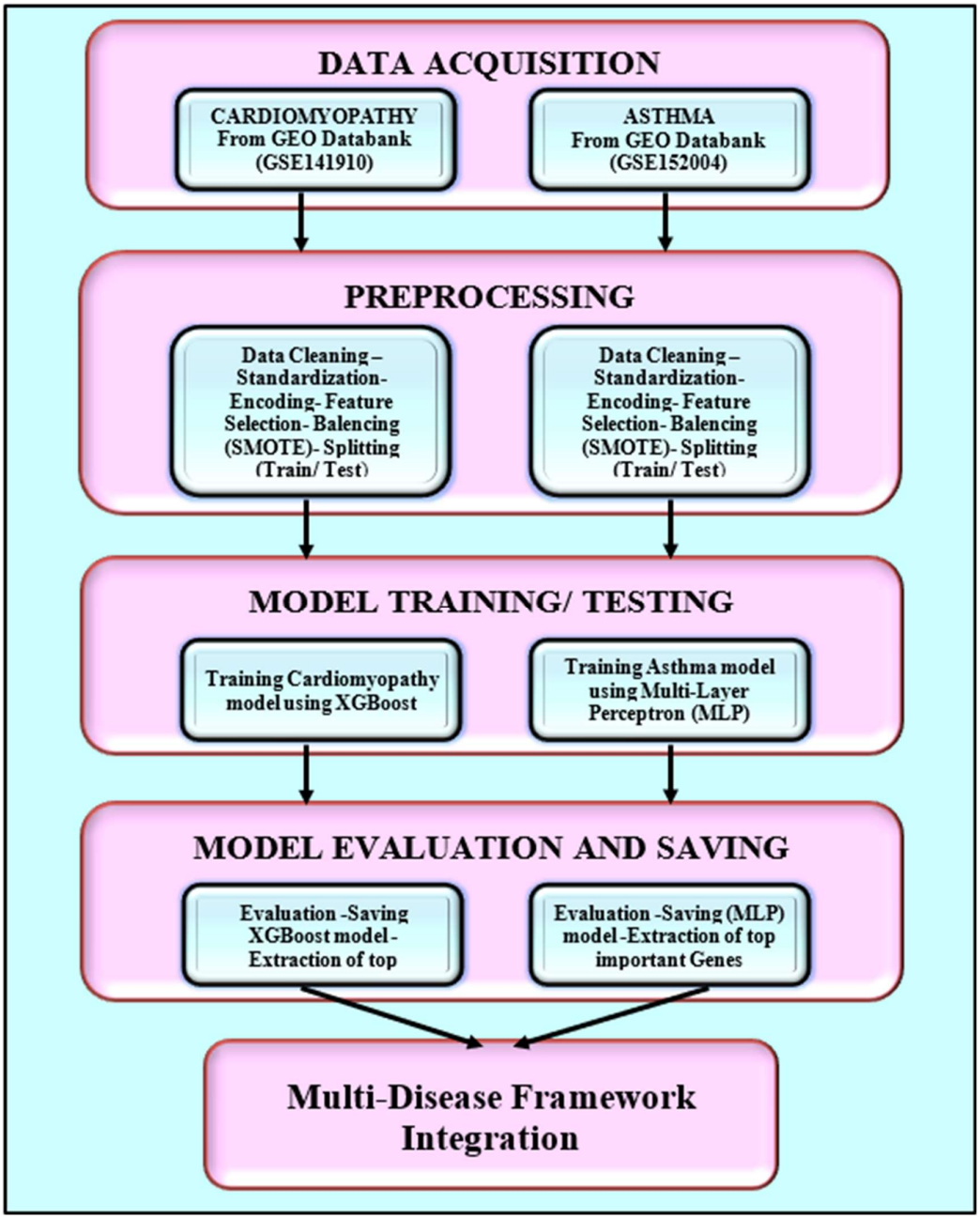

Figure 3.1: Multi-Disease Genetic Prediction Diagram

3.3 Implementation

3.3.1 Implementation Tools and Technical Infrastructure

Software and Libraries:

Programming Environment: Python 3.10+ interpreter executing on Windows 11 IoT Enterprise platform, providing cross-platform compatibility and extensive machine learning ecosystem support.

Core ML Libraries:

- scikit-learn 1.3+ for MLP neural networks, data preprocessing, and evaluation metrics
- XGBoost 1.7+ for gradient boosting classification with optimized performance
- imbalanced-learn for SMOTE implementation addressing class imbalance

Data Processing Stack:

- pandas 2.0+ for DataFrame operations and CSV handling
- NumPy 1.24+ for numerical computations and array manipulations
- GEOparse for NCBI GEO database integration.

Visualization Tools: matplotlib 3.7+ and seaborn 0.12+ for statistical plotting and performance visualization, which allows the graphical representation of model evaluation.

Deployment Framework: Gradio 3.0+ for building user-friendly web interfaces that enable the execution of gene expression analysis and disease prediction in real-time.

```
# ========== 📦 استيراد وتثبيت المكتبات المطلوبة ==========
!pip install GEOparse xgboost imbalanced-learn --quiet

import GEOparse
import pandas as pd
import numpy as np
import urllib.request

import gzip
import shutil
import os

import tarfile
import glob
import joblib

import xgboost as xgb
import matplotlib.pyplot as plt
import seaborn as sns

# ✅ استيراد المكتبات

from tqdm import tqdm  # لشريط التقدم
from sklearn.neural_network import MLPClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score, roc_auc_score, confusion_matrix, precision_score, recall_score
from sklearn.feature_selection import SelectKBest, f_classif
from imblearn.over_sampling import SMOTE
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split, RandomizedSearchCV
```

Figure 3.2: Code snippet for libraries import

Computational Resources and Workflow:

Hardware Configuration:

Device: HP Laptop 15-bs1xx (DESKTOP-1E80L07).

Processor: Intel Core i3-5005U @ 2.00GHz (64-bit, x64-based).

Memory: 4.00 GB RAM.

Operating System: Windows 11 IoT Enterprise (64-bit).

Performance Considerations: Limited RAM (4GB) induced implementation of memory management policy which included feature selection (500 features for asthma, 200 for cardiomyopathy) and batch-processing for large gene expression datasets.

Workflow Pipeline: Serial processing design carrying out activities like data download → preprocessing → feature selection → model training → evaluation → deployment, targeted for single-thread execution on dual-core processor.

Model Persistence: joblib serialization provides a way to save and load the trained model without retraining which is very convenient during the deployment process on hardware with less resource available.

Reproducibility Protocol: Conserved random seeds (random_state=42) in any procedure of randomness make sure experimental results are always matching thus promoting research replication.

3.3.2 Datasets Sources

Gene expression data for this study were taken from the Gene Expression Omnibus (GEO), which is a public functional genomics data repository supporting MIAME-compliant data submissions. The two main datasets that were used for this purpose are: GSE152004 for asthma analysis containing normalized gene expression profiles and GSE141910 for cardiomyopathy analysis providing comprehensive cardiac tissue expression data.

Dataset	Disease	Samples	Genes	Control	Disease	Data Source
GSE152004	Asthma	695	~20,000	347	348	GEO Database
GSE141910	Cardiomyopathy	166	~15,000	91	75	GEO Database

Table 3.1: Dataset Characteristics Summary

The asthma dataset (GSE152004) consists of control and asthmatic patient samples with binary classification labels that were taken from the GEO metadata. For cardiomyopathy, phenotypic

data were obtained from the MAGNet phenoData repository, where the samples were categorized according to etiology classifications, e.g., "NF" stood for control subjects, while other etiologies represented the cardiomyopathy cases. Samples that fulfilled a certain set of criteria were those with complete gene expression profiles, and the disease status was verified, while those that were missing phenotypic information or had incomplete expression data were excluded.

```
[ ]  # ========== 🛠️ تحميل ومعالجة بيانات الربو ==========
     os.makedirs("data", exist_ok=True)
     url_asthma = "https://ftp.ncbi.nlm.nih.gov/geo/series/GSE152nnn/GSE152004/suppl/GSE152004_695_expr_norm.txt.gz"
     urllib.request.urlretrieve(url_asthma, "data/GSE152004.txt.gz")

     with gzip.open("data/GSE152004.txt.gz", 'rb') as f_in:
         with open("data/GSE152004.txt", 'wb') as f_out:
             shutil.copyfileobj(f_in, f_out)

     df_asthma_expr = pd.read_csv("data/GSE152004.txt", sep='\t', index_col=0)

     # تحميل بيانات GEO
     gse_asthma = GEOparse.get_GEO(geo="GSE152004", destdir="data")

     # استخراج التسميات المصححة
     labels_asthma_corrected = []
     for gsm in gse_asthma.gsms.values():
         status = str(gsm.metadata["characteristics_ch1"]).lower()
         if 'control' in status or 'healthy' in status:
             labels_asthma_corrected.append(0)
         else:
             labels_asthma_corrected.append(1)

     print("✅ توزيع الفئات:", pd.Series(labels_asthma_corrected).value_counts().to_dict())
```

Figure 3.3: Asthma data Downloading

```
[ ]  # ✅ تحميل تسميات مرض القلب (phenoData)
     # تحميل جدول التسميات
     df_labels = pd.read_csv("https://raw.githubusercontent.com/mpmorley/MAGNet/master/phenoData.csv")
     print("تم تحميل التسميات. عدد العينات الكلي:", len(df_labels))

     # ✅ الاحتفاظ فقط بالعينات من نوع 'C'
     df_labels = df_labels[df_labels['sample_name'].str.startswith("C")]
     df_labels = df_labels.set_index("sample_name")

     # ✅ تصنيف الحالة المرضية
     def label_disease(etiology):
         etiology = str(etiology).strip().upper()
         if etiology == "NF":
             return "control"
         else:
             return "cardiomyopathy"

     df_labels["Label"] = df_labels["etiology"].apply(label_disease)
     df_labels = df_labels[df_labels["Label"].isin(["control", "cardiomyopathy"])]

     print("توزيع الفئات:")
     print(df_labels["Label"].value_counts())
```

Figure 3.4: Cardiomyopathy data Downloading

3.3.3 Preprocessing Workflow

For the sake of processing efficiency, a comprehensive preprocessing pipeline in agreement with the data requirements and compatible with machine learning algorithms was implemented.

Data cleaning:

The missing values were completed by the numpty function **nan_to_num** that acted as a numerical imputation. Furthermore, the Standard Scaler was applied to normalize the gene expression data, that is to say, the gene expression values will have a zero mean and unit variance. We also used **StandardScaler** from the Scikit-learn library to make data in same range and normally distributed.

```
[ ]  # ✅ ترميز التصنيفات
     y = df_labels["Label"].map({"control": 0, "cardiomyopathy": 1}).values
     X = df_expr.T.values

     # ✅ معالجة القيم المفقودة
     X = np.nan_to_num(X)

     # ✅ تسوية البيانات
     scaler = StandardScaler()
     X_scaled = scaler.fit_transform(X)
```

Figure 3.5: data cleaning

feature selection:

The feature selection step used **SelectKBest** with **f_classif** scoring to pick the most differentiating genes. According to dataset characteristics, 500 features for asthma, and 200 for cardiomyopathy, correspondingly, were selected.

```
     # ✅ اختيار أفضل 200 جين (تم زيادة العدد لتحسين الأداء)
     try:
         selector_cardio = SelectKBest(f_classif, k=min(200, X_scaled.shape[1]))
         X_selected = selector_cardio.fit_transform(X_scaled, y)
         print(f"\n✅ تم اختيار {X_selected.shape[1]} ميزة")
     except Exception as e:
         print(f"\n⚠️ خطأ في اختيار الميزات: {str(e)}")
         print("الاستمرار بجميع الميزات...")
         X_selected = X_scaled

     # ✅ تقسيم البيانات
     X_train_cardio, X_test_cardio, y_train_cardio, y_test_cardio = train_test_split(
         X_selected, y,
         test_size=0.2,
         stratify=y,
         random_state=42
     )
```

Figure 3.6: Feature Selection and data splitting using SelectKBest with f_classif

Train-Test Split: 85-15% split for asthma prediction and 80-20% split for cardiomyopathy prediction maintaining the original class distribution using train_test_split (stratify=y, random_state=42). We also used PCA (Principal Component Analysis) reducing dimensionality by creating new components that contain as much variance in the data as possible, as an alternative solution in the Ashma disease if SelectKBest was not possible or if it did not work.

```
# ========== 1. إعادة اختيار الميزات (زيادة k) ==========
try:
    selector_asthma = SelectKBest(f_classif, k=min(500, X.shape[1]))
    X_selected = selector_asthma.fit_transform(X_scaled, y)
except:
    #////////////////////////////////////////////////////////////////////////////
    pca = PCA(n_components=150)
    X_selected = pca.fit_transform(X_scaled)

# ========== 2. تقسيم جديد أكثر توازنًا (اختبار أقل بنسبة 15%) ==========
X_train_asthma, X_test_asthma, y_train_asthma, y_test_asthma = train_test_split(
    X_selected, y, test_size=0.15, stratify=y, random_state=42
)
```

Figure 3.7: Feature Selection Alternative (PCA)

Class Imbalance Handling with SMOTE: To solve the class imbalance problem we have employed SMOTE, only training sets were oversampled using SMOTE (Synthetic Minority Oversampling Technique) thus, the original test distribution was preserved while the minority classes were balanced.

The quality control procedures included outlier identification by means of statistical thresholds and data validation checks that were performed along the whole preprocessing pipeline to guarantee dataset integrity.

```
# ========== 3. SMOTE لموازنة التدريب ==========
if len(np.unique(y_train_asthma)) > 1:
    smote = SMOTE(random_state=42)
    X_res_asthma, y_res_asthma = smote.fit_resample(X_train_asthma, y_train_asthma)
else:
    X_res_asthma, y_res_asthma = X_train_asthma, y_train_asthma
```

Figure 3.8: Data Balencing with SMOTE

3.3.4 Models Training

Asthma Prediction Model (MLP) implementation:

- Algorithm Rationale: Neural networks are the best choice for a complicated polygenic asthma problem.
- Network Architecture: Description of the number of hidden layers, neurons in each layer, and activation functions used.
- Training Protocol: Backpropagation, optimization algorithms (Adam).
- Regularization Techniques: early stopping.
- Complexity Handling: Non-linear interactions between genes and pathways networks are represented by the model.
- **Architecture:** Multi-layer perceptron with hidden layers (256, 128, 70)
- **Optimizer:** Adam with learning_rate=0.001, max_iter=800
- **Dataset:** GSE152004 (695 samples)

```
# ========== 4. تدريب MLP ==========
model_asthma = MLPClassifier(
    hidden_layer_sizes=(256, 128, 70),  # طبقات بدلاً من واحدة
    activation='relu',
    solver='adam',
    learning_rate_init=0.001,
    max_iter=800,
    early_stopping=True,
    random_state=42
)

model_asthma.fit(X_res_asthma, y_res_asthma)
print("✅ تم تدريب MLP !")
```

Figure 3.9: Asthma Prediction Model (MLP) training

Cardiomyopathy Prediction Model (XGBoost) implementation:

- Algorithm Rationale: A discussion about the decision tree ensemble methods suitable for a cardiomyopathy genetic study.
- Model Architecture: Boosting settings, max depth, learning rate fine-tuning.
- Biological Relevance: The treatment of gene-gene interactions and the non-linear part of genetics.
- **Parameters:** max_depth=8, learning_rate=0.2, n_estimators=500
- **Class balancing:** scale_pos_weight = negative/positive sample ratio
- **Dataset:** GSE141910 (166 samples) with phenoData labels

```
[ ]  # ✅ تدريب نموذج XGBoost
     params = {
         'objective': 'binary:logistic',
         'eval_metric': 'logloss',
         'max_depth': 8,
         'learning_rate': 0.2,
         'n_estimators': 500,
         'subsample': 0.8,
         'colsample_bytree': 0.9,
         'scale_pos_weight': np.sum(y == 0) / np.sum(y == 1),  # موازنة الفئات
         'random_state': 42,
         'n_jobs': -1
     }

     model_cardio = xgb.XGBClassifier(**params)

     try:
         model_cardio.fit(X_res_cardio, y_res_cardio)
         print("\n🎯 تم تدريب النموذج بنجاح!")
     except Exception as e:
         print(f"\n❌ فشل تدريب النموذج: {str(e)}")
         raise
```

Figure 3.10: Cardiomyopathy Prediction Model (XGBoost) training

3.3.4 Models Testing and evaluation

For testing and evaluation, we used the next performance Metrics for both models: Classification accuracy, precision, recall, AUC score, confusion matrix, F1-score and classification report that used sklearn.metrics library for the control and disease classes respectively.

```
[ ]  # ========== 5. التقييم ==========
     y_pred = model_asthma.predict(X_test_asthma)
     print("\n📊 تقرير الأداء - MLP :")
     print(classification_report(y_test_asthma, y_pred, target_names=["control", "asthma"]))

     #////////////////////////////////////////////////////////////////////////////
     accuracy = accuracy_score(y_test_asthma, y_pred)
     print(f"🎯 دقة النموذج المحسّن: {accuracy * 100:.2f}%")
```

Figure 3.11: Testing and Evaluating Asthma (MLP) model

```
[ ]  # ✅ التقييم
     y_pred = model_cardio.predict(X_test_cardio)
     y_proba = model_cardio.predict_proba(X_test_cardio)[:, 1]

     print("\n📊 تقرير الأداء - مرض القلب:")
     print(classification_report(y_test_cardio, y_pred, target_names=["control", "cardiomyopathy"]))
     print(f"🎯 الدقة: {accuracy_score(y_test_cardio, y_pred):.4f}")
```

Figure 3.12: Testing and Evaluating Cardiomyopathy (XGBoost) model

Achieved Accuracy: **79% for Asthma (MLP)** model and **96% for cardiomyopathy (XGBoost)** model.

class	Precision	Recall	f1-score	Support	Accuracy	Macro avg	weighted avg	AUC score
Control	0.79	0.58	0.67	38	0.79	0.76	0.78	0. 7905
asthma	0.79	0.91	0.85	67				

Table 3.2: MLP model performance summary for Asthma

class	Precision	Recall	f1-score	Support	Accuracy	macro avg	weighted avg	AUC score
Control	0.80	1.00	0.89	4	0.96	0.93	0.96	0.9615
Cardiom -yopathy	1.00	0.95	0.98	22				

Table 3.3: XGBoost model performance summary for Cardiomyopathy

3.3.5 Multi-Disease Framework Integration

Model Saving and Key Gene Identification:

```
# ✅ استرجاع الجينات المختارة
selected_genes_cardio = df_expr.index[selector_cardio.get_support()]
# ✅ الأهمية
importances_cardio = model_cardio.feature_importances_

# ✅ إنشاء DataFrame مرتب
gene_importance_df = pd.DataFrame({
    "Gene": selected_genes_cardio,
    "Importance": importances_cardio})

# ✅ الترتيب التنازلي
gene_importance_df = gene_importance_df.sort_values(by="Importance", ascending=False)
# ✅ حساب النسبة المئوية
gene_importance_df["Importance (%)"] = 100 * gene_importance_df["Importance"] / gene_importance_df["Importance"].sum()
# ✅ أخذ أهم 20 جين فقط
top_20_genes_df = gene_importance_df.head(20).reset_index(drop=True)
# ✅ عرض النتائج
print("🔬 أهم 20 جينًا مرتبة حسب الأهمية:")
print(top_20_genes_df)
```

Figure 3.13: most important genes extraction for Cardiomyopathy (XGBoost) model

The trained models are saved with joblib.dump() that can be loaded now to make prediction in real time without retraining. Then we extracted the most important genes top the ranked using SelectKBest f_classif scores for both Asthma and cardiomyopathy. The biological interpretation of the results is given in the framework

Framework Integration:

To finalize the implementation, the two disease models; an MLP for asthma and an XGBoost for cardiomyopathy were combined into a unified framework. The framework thus allows parallel processing of gene expression data for each condition with its respective disease-specific classifier. The integration supports end-to-end predictions after preprocessing and feature selection, it also enables this multi-disease configuration to be an efficient, scalable diagnosis support system based on genetic markers.

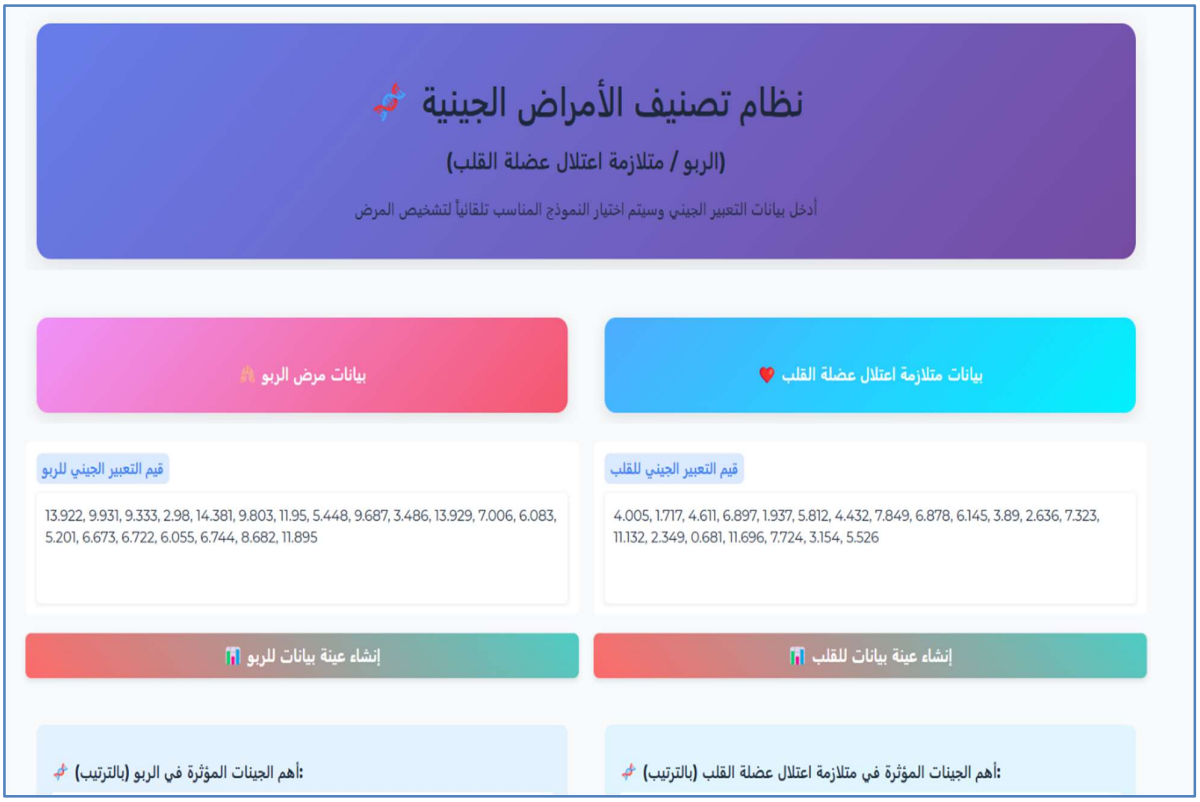

Figure 3.14: Multi-Disease Genetic Prediction System

Figure 3.15: Multi-Disease Genetic Prediction System 2

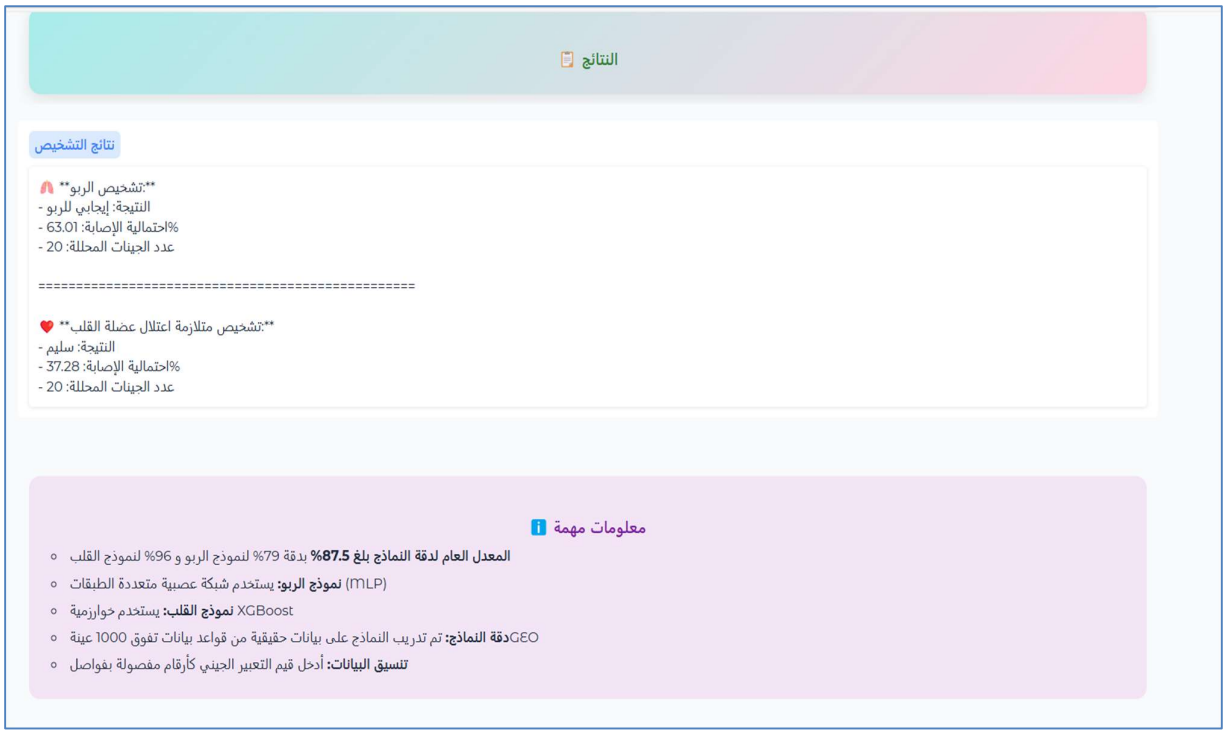

Figure 3.16: Multi-Disease Genetic Prediction System 3

3.4 Conclusion

In this chapter, we had the performance of the model for the target diseases and also outlined a detailed approach to conducting gene-disease machine learning-based research for asthma and cardiomyopathy. They indirectly laid out the clinical and genetic bases for the two diseases, designed good data preprocessing pipelines, and then selected features strategically implementing the feature selection strategies with disease-specific genetic markers.

Our method utilizes XGBoost for cardiomyopathy and Multi-Layer Perceptron for asthma prediction in a single framework. We have delineated rigorous training routines and evaluation indicators while being aware of computational and interpretability limitations for the future research community.

General conclusion

This thesis represents a thorough examination gene-disease association prediction via machine learning, which has laid down both the theoretical and practical bases for healthcare applications in genomics. In three closely connected chapters, we were able to show how artificial intelligence has the potential to revolutionize genetic medicine and personalized healthcare.

In the initial chapter, we reaffirmed the basic concepts of AI and ML, delving into different algorithmic methods and their implementation in genomics research. Machine learning paradigms like supervised, unsupervised, and reinforcement learning were utilized to tackle complicated genomic puzzles, thus providing the theoretical base required for advanced gene-disease association research.

Chapter 2 endeavored to survey in depth the existing literature on genomics as well as machine learning applications, specifying the gaps in the research that can be filled with novel ideas. We sketch the research agenda that led our present approach to the building of prediction models that are more reliable and clinically useful.

Chapter 3 proposed some of our fresh ideas on the genetic prediction of multiple diseases with an emphasis on asthma and cardiomyopathy, presenting the constructions of a programmatic system that combines the XGBoost and Multi-Layer Perceptron architectures to demonstrate that machine learning not only provides a means of extracting genomic data but also facilitates the translation of it into practical clinical applications while granting it the ability to face computational as well as interpretational challenges.

This work aims to be a significant contribution to the arena of precision medicine as it delivers the proven means for the prediction of gene-disease association. Our multi-disease schematic thus equips us with solutions that are not only scalable but also extendible to other genetic illnesses, which in turn may be the fastest track to the conception of personalized treatment strategies and up to-date patient outcomes through early diagnosis and risk evaluation. The methods and results shared in this thesis delineate a springboard for computational genomics undertaking in the future.

While genomic datasets further develop and the availability of computational resources increases, our framework remains to be vitally significant; it holds the potential to be improved through the incorporation of multi-omics, the conducting of clinical validation studies, and the designing of more complex AI models that may close up genetic research and the clinical practice farther more.

This research marks a notable advancement in the application of machine learning for genomics, thereby supporting the aim of producing highly personalized, genetics-based diagnostic and treatment solutions.

Our future direction for this project emphasizes clinical translation through work with hospitals and clinics to validate our models in real-world healthcare settings. Our next steps is to make health professionals decision support tools, since they can be used directly in the patient care workflows. Further, we want to take a step forward in more advanced methodological approaches by studying multi-omics integration—genomic, transcriptomic, and proteomic information integrated together—to understand disease from one vantage point. We intend to investigate further Graph Neural Networks (GNNs) as a method of modeling complex gene-protein interaction networks for the discovery of new therapeutic targets by way of unused network regions. Increased model transparency is also on our agenda through the application of tools from the field of Explainable Artificial Intelligence.

Bibliography

[1] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[2] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[3] P. Khalkar et al., "Handwritten Text Recognition using Deep Learning (CNN & RNN)," International Advanced Research Journal in Science, Engineering and Technology, vol. 8, no. 6, Jun. 2021. [Online]. Available: https://www.researchgate.net/publication,/353939315_Handwritten_Text_Recognition_using_Deep_Learning_CNN_RNN. [Accessed: may. 10, 2025].

[4] H. Bhasin, Machine Learning for Beginners. France: BPB Publications, 2020.

[5] Types of Machine Learning," GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/types-of-machine-learning/. [Accessed: 05-03-2025].

[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009.

[7] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.

[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533-536, Oct. 1986.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013.

[11] A. Sharma and S. Kumar, "Random Forest ensemble learning for genomic data analysis and biomarker discovery," Computational Biology and Chemistry, vol. 89, pp. 107-115, 2020.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794.

[13] *D. Dose, "Formulating and Implementing XGBoost from Scratch," Daily Dose of Data Science, Apr. 8, 2023. [Online]. Available: https://www.dailydoseofds.com/formulating-and-implementing-xgboost-from-scratch/.*

[14] "Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning applications in genomics: Multilayer perceptron networks for gene expression analysis," Nature Reviews Genetics, vol. 16, no. 7, pp. 321-332, 2015.".

[15] *N. Lunge, "A Deep Architecture: Multi-Layer Perceptron," Medium, Jun. 30, 2020. [Online]. Available: https://medium.com/@nlunge786/a-deep-architecture-multi-layer-perceptron-164bc5ff3842.*

[16] A review of deep learning applications in human genomics using next-generation sequencing data," Human Genomics. [Online]. Available: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-022-00396-x. [Accessed: 15-03-2025].

[17] "Next-Generation Sequencing Technology: Current Trends and Advancements," PMC, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10376292/.

[18] "National Human Genome Research Institute, "Deoxyribonucleic acid (DNA)," Genome.gov, 2023. [Online]. Available:

https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid".

[19] National Human Genome Research Institute, “Ribonucleic acid (RNA),” Genome.gov, 2023. [Online]. Available: https://www.genome.gov/genetics-glossary/RNA.

[20] National Human Genome Research Institute, “Protein,” Genome.gov, 2023. [Online]. Available: https://www.genome.gov/genetics-glossary/Protein".

[21] Central Dogma (Steps & Guide),” GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/central-dogma-steps-guide/. [Accessed: Jun. 5, 2025].

[22] G. 2. [. A. h.-g. National Human Genome Research Institute.

[23] w. i. u. b. t. c. t. p. i. v. f. Mitochondria are small structures inside cells known as the "powerhouses" of the cell. They generate energy in the form of ATP.

[24] "Genetic Disorders," Genome.gov, Mar. 2019. [Online]. Available: https://www.genome.gov/For-Patients-and-Families/Genetic-Disorders.

[25] National Human Genome Research Institute, "Genetic disorders," U.S. Department of Health and Human Services, Mar. 2019. [Online]. Available: https://www.genome.gov/For-Patients-and-Families/Genetic-Disorders.

[26] P. M. Visscher et al., "10 years of GWAS discovery: Biology, function, and translation," American Journal of Human Genetics, vol. 101, no. 1, pp. 5-22, Jul. 2017.

[27] T. A. Manolio et al., "Finding the missing heritability of complex diseases," Nature, vol. 461, no. 7265, pp. 747-753, Oct. 2009.

[28] T. Hassold and P. Hunt, "Genetics, nondisjunction," in StatPearls. Treasure Island, FL: StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482240/.

[29] C. Garone and S. Rahman, "Genetics of mitochondrial diseases: Current approaches for the molecular diagnosis," in Mitochondrial Disorders Caused by Nuclear Genes, 1st ed., S. Rahman and C. Garone, Eds. Amsterdam: Elsevier, 2018, pp. 247-278.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013.

[31] "Artificial intelligence and machine learning in precision and genomic medicine," PMC. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9198206/. [Accessed: 07-03-2025].

[32] J. Singh, K. Hanson, and S. T. Sarasohn-Kahn, "A review of deep learning applications in human genomics using next-generation sequencing data," Human Genomics, vol. 16, no. 26, 2022.

[33] A. Author, "Machine learning methods for pattern recognition analysis of genomic and molecular data," Ph.D. dissertation, New Jersey Institute of Technology, 2024.

[34] L. L. Kalliauer et al., "Artificial intelligence in clinical and genomic diagnostics," Genome Medicine, vol. 11, no. 70, 2019.

[35] X. Li et al., "Analysis of potential genetic biomarkers using machine learning methods and immune infiltration regulatory mechanisms underlying atrial fibrillation," BMC Medical Genomics, vol. 15, no. 112, 2022.

[36] P. Roman-Naranjo, A. M. Parra-Perez, and J. A. Lopez-Escamez, "A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases," Journal of Biomedical Informatics, vol. 143, pp. 1-8, 2023.

[37] Y. Zhang et al., "A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation," BMC Bioinformatics, vol. 25, no. 1, pp. 1-18, 2024.

[38] K. Kim et al., "Machine Learning to Advance Human Genome-Wide Association Studies," Genes, vol. 15, no. 1, pp. 1-20, 2024.

[39] P. Roman-Naranjo, A. M. Parra-Perez, and J. A. Lopez-Escamez, "A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases," Journal of Biomedical Informatics, vol. 143, pp. 1-8, 2023.

[40] M. Johnson et al., "Genetic diversity in disease association studies: Current limitations and future directions," Human Genetics, vol. 142, no. 4, pp. 445-462, 2024.

[41] R. Anand et al., "Decoding the genetic and epigenetic basis of asthma," PubMed, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36727912/.

[42] D. Rhonda, “Asthma Genes: The Genetics of Asthma,” Genetic Lifehacks, [Online]. Available: https://www.geneticlifehacks.com/asthma-genes/. [Accessed: Jun. 1, 2025].

[43] R. P. Patel et al., “Cardiomyopathy in genetic and molecular studies,” *Journal of the American College of Cardiology*, vol. 75, no. 24, pp. 2990–3002, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735109719379343.

[44] S. R. Ommen et al., "Molecular Genetic Basis of Hypertrophic Cardiomyopathy," PubMed, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33983830/.

الملخص:

تتناول هذه الأطروحة كيفية استخدام تقنيات التعلم الآلي للتنبؤ بالارتباطات بين الجينات والأمراض، مع التركيز على مرضي الربو و متلازمة اعتلال عضلة القلب. غالبًا ما تواجه الأساليب الجينومية التقليدية صعوبات في التعامل مع البيانات المعقدة وعالية الأبعاد، لا سيما في المناطق غير المشفّرة من الجينوم. لمعالجة ذلك، قمنا بتطوير واجهة تنبؤية باستخدام نماذج مثل (XGBoost & MLP) بهدف تحقيق الدقة وقابلية التفسير معا. تظهر نتائجنا كيف يمكن ان يسهم التعلم الآلي في تحسين تحسين التشخيص المبكر ودعم اتخاذ القرار السريري من خلال نموذج تنبؤ متعدد الأمراض، حيث تم تحقيق دقة عالية لكلا الحالتين (79% للربو و96% لاعتلال عضلة القلب). تشمل التوجهات المستقبلية دمج بيانات متعددة الأوميكس، والتحقق السريري، وتطبيق الذكاء الاصطناعي القابل للتفسير لتعزيز الثقة وسهولة الاستخدام.

الكلمات المفتاحية:

الارتباط بين الجينات والأمراض، التعلم الآلي، الجينوميات، الربو، اعتلال عضلة القلب ، XGBoost ، MLP
الذكاء الاصطناعي القابل للتفسير ،اتخاذ القرارات السريرية.

Abstract:

This thesis examines how machine learning can be used to forecast gene-disease associations, focusing on asthma and cardiomyopathy. Traditional genomic approaches often struggle with complex, high-dimensional data, particularly in non-coding regions. To address this, we developed a predictive framework using models such as XGBoost and MLP, aiming for both accuracy and interpretability. Our results show how machine learning (ML) might enhance early diagnosis and aid in clinical decision-making through a multi-disease prediction model, achieving high accuracy for both conditions (79% for asthma and 96% for cardiomyopathy). Future directions include multi-omics integration, clinical validation, and the application of explainable AI to improve trust and usability.

Keywords:

Gene-Disease Association, Machine Learning, Genomics, Asthma, Cardiomyopathy, XGBoost, MLP, Clinical Decision, Explainable AI.

Résumé :

Cette thèse examine comment l'apprentissage automatique peut être utilisé pour prédire les associations gènes-maladies, en particulier la cardiomyopathie et l'asthme. Les approches génomiques traditionnelles peinent souvent à gérer des données complexes et de grande dimension, notamment dans les régions non codantes. Pour y remédier, nous avons développé un cadre prédictif utilisant des modèles tels que XGBoost et MLP, visant à la fois la précision et l'interprétabilité. Nos résultats montrent comment l'apprentissage automatique (AM) pourrait améliorer le diagnostic précoce et faciliter la prise de décision clinique grâce à un modèle de prédiction multi-maladies, atteignant une précision élevée pour les deux pathologies (79 % pour l'asthme et 96 % pour la cardiomyopathie). Les orientations futures incluent l'intégration multi-omique, la validation clinique et l'application de l'IA explicative pour améliorer la confiance et la convivialité.

Mots-clés :

Association gène-maladie, apprentissage automatique, génomique, asthme, cardiomyopathie, XGBoost, MLP, médecine de précision, modélisation prédictive, IA explicative, intégration multi-omique, aide à la décision clinique.