AI and Machine Learning in Genomic Research

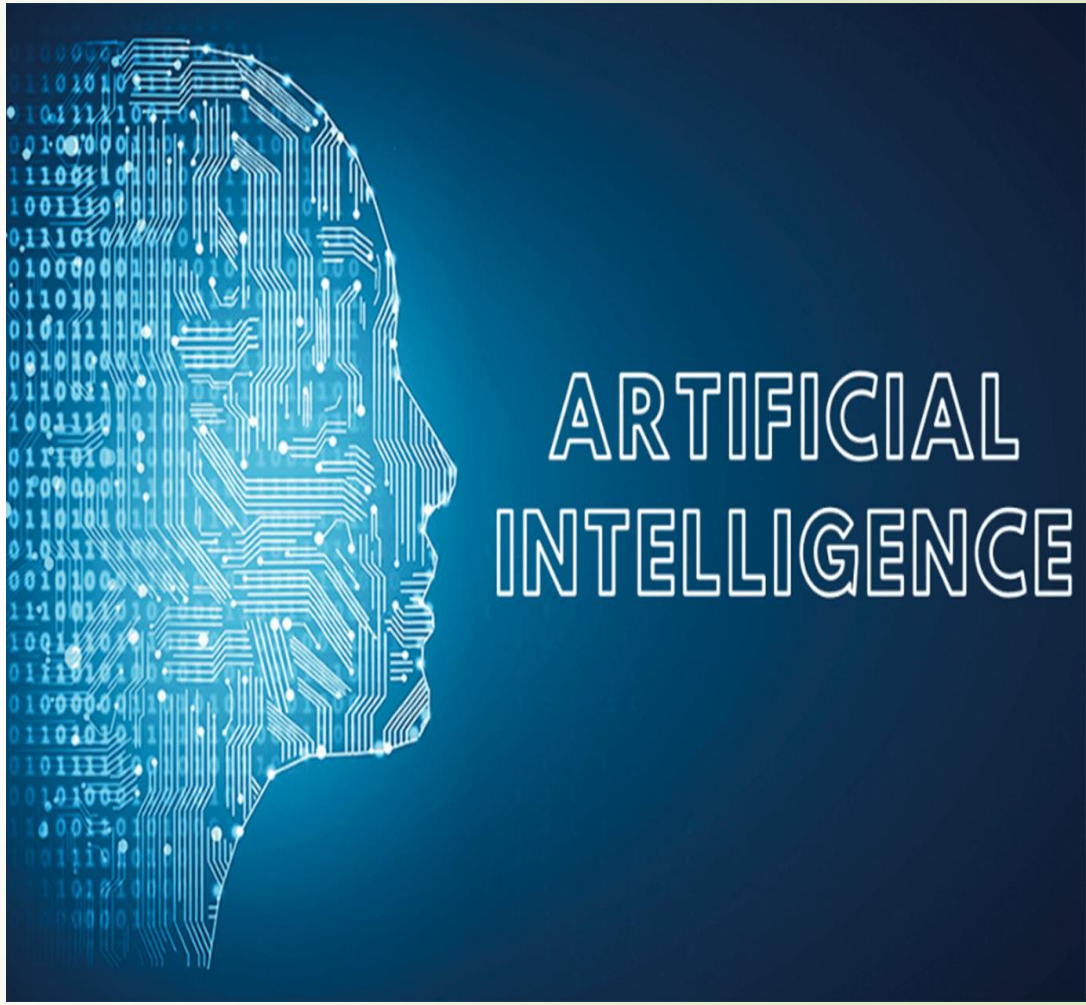
PRESENTED DURING THE 2025 CENTRE FOR GENOMICS OF NON-COMMUNICABLE DISEASES AND PERSONALIZED HEALTHCARE (CGNPH) SUMMER SCHOOL

PRESENTED BY DR. COMFORT FOLORUNSO
DEPARTMENT OF SYSTEMS ENGINEERING,
UNIVERSITY OF LAGOS

What is Artificial Intelligence?

Artificial intelligence (AI) is a branch of computer science that focuses on creating systems or machines that can perform tasks that normally require human intelligence.

It is about building computer systems that can **think**, **learn**, **and make decisions**, similar to how humans do — but using algorithms and data.

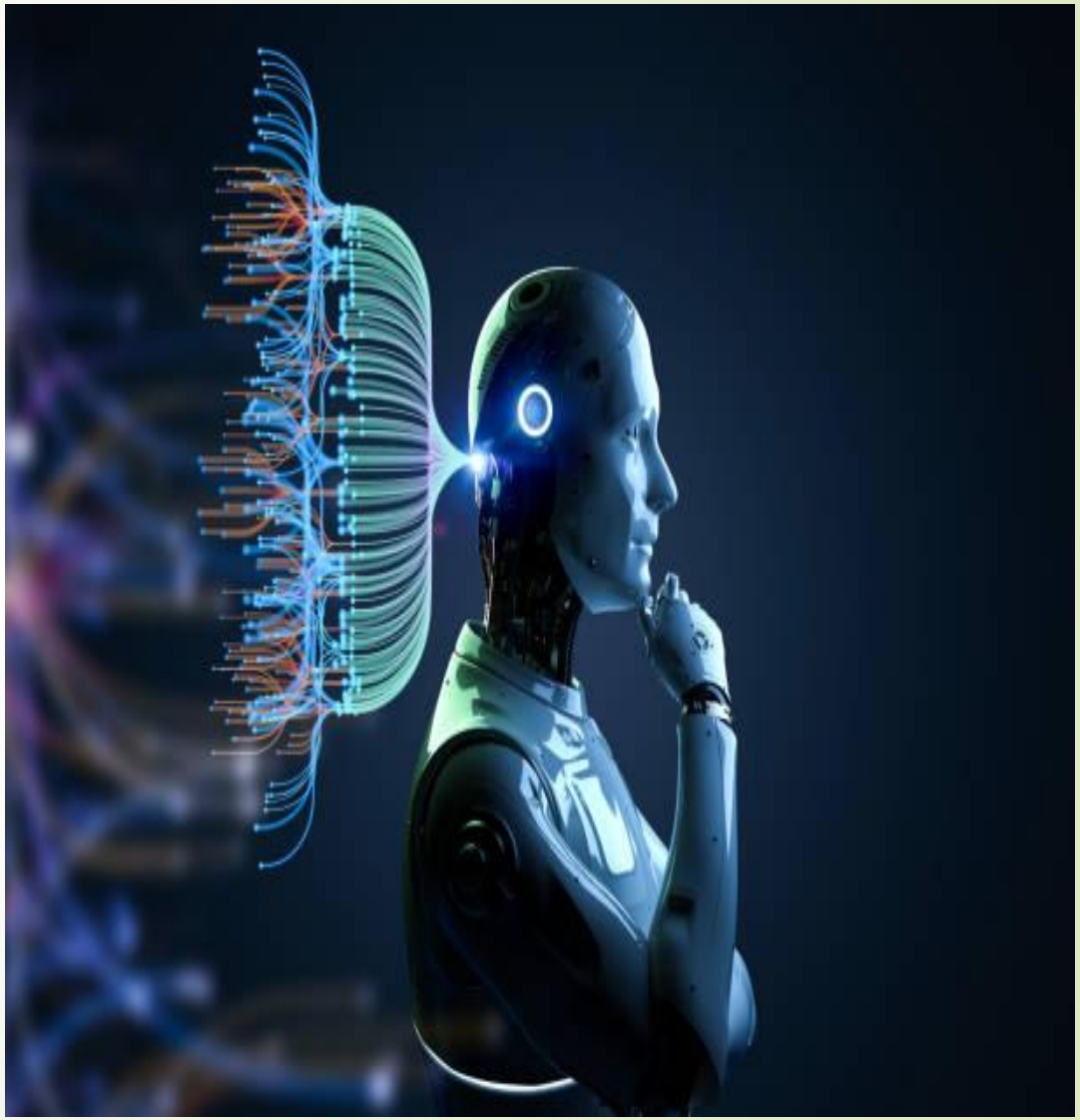


The tasks performed by AI include:

- **Learning:** acquiring knowledge or skills from data and experiences (like a child learning from examples).
- **Reasoning:** drawing conclusions or solving problems based on rules or data.
- **Perception:** understanding the world through sensory inputs, such as seeing (computer vision) or hearing (speech recognition).
- **Language understanding:** interpreting and generating human language (natural language processing).
- **Decision-making:** choosing actions to achieve goals, often by evaluating possible outcomes.

What is Machine Learning?

- **Machine learning (ML)** is a branch of artificial intelligence (AI) that focuses on developing systems that can **learn from data**, identify patterns, and make decisions or predictions **without being explicitly programmed for each specific task**.



Types of machine learning

- **Supervised learning:** learning from labeled data (e.g., examples with the correct answers).
- **Unsupervised learning:** finding patterns in data without labels (like grouping similar customers, grouping data into clusters).
- **Reinforcement learning:** learning by trial and error to maximize rewards (like training a robot or playing a game).

MACHINE LEARNING IN MEDICAL RESEARCH

What is Genomic Research

- Genomic research is about **studying all the DNA in an organism to understand how genes work**, how they influence traits (like eye colour or risk of disease), and how changes in genes (mutations) can lead to health problems.

What is genomic data

- Genomic data is the **collection of all the genetic information** found in an organism's DNA.

- The **DNA letters** are the **four chemical bases (nucleotides)** that make up DNA. They are often represented by their initials:

  A – Adenine
  C – Cytosine
  G – Guanine
  T – Thymine

- These letters pair up to form the **rungs of the DNA double helix ladder**:

  A pairs with T (A–T)
  C pairs with G (C–G)

- The **sequence** of these letters (like AGCTTACG...) is what carries all the genetic information needed to build and run an organism.

- The DNA data is like a **massive instruction manual** written in A, T, C, and G (the four DNA letters) that tells cells how to build and run the body.

How long is DNA?

- If you take the DNA **from just one human cell** and **stretch it out**, it would be about:
- **2 meters (6 feet) long**
- Even though it's so long, it all fits inside the cell's **nucleus**, which is only about **0.000006 meters (6 micrometers) wide**, because DNA is **tightly coiled and packed** into chromosomes.

- Now, imagine this:
- Your body has **~37 trillion cells**.
- If you lined up the DNA from all your cells end to end, it would stretch about **twice the diameter of the solar system!**

How big is DNA in terms of data?

- The **human genome** (all the DNA) is made of **about 3 billion base pairs (letters A, T, C, G)**.
- If you stored it as text (1 letter = 1 byte), the genome would take **~3 gigabytes (GB)** of data.

Transforming Genomic Research with AI and Machine Learning

Data Analysis

AI and Machine Learning enable the analysis of vast amounts of genetic data.

Understanding Genetic Patterns

These technologies enhance our understanding of complex genetic patterns and their implications for health and disease.

Personalized Medicine

The integration of AI and ML into genomics promises to accelerate discoveries and improve personalized medicine.

The Role of AI and ML in Genomic Research

- **Data Association**: Identify relationships between genes and diseases.
- **Outcome Prediction**: Predict potential outcomes based on genetic data.
- **Personalized Treatment**: Tailor treatment strategies to individual genetic profiles.

Applications of AI & ML in Genomic Research

1. **Genomic Variant Interpretation**: AI and ML are used to interpret genetic variants by predicting their potential pathogenicity. These algorithms analyze vast databases of genomic data and clinical outcomes to assess the significance of specific variants.

2. **Drug Discovery and Development**: AI technologies are revolutionizing drug discovery by predicting how different compounds will interact with specific genetic targets and optimizing their efficacy.

3. **Precision Medicine**: AI and ML enable the development of precision medicine strategies by tailoring treatments based on an individual's genetic makeup, particularly in oncology.

Data Challenges in Genomics

- **Volume and Complexity:** The rapid advancement of sequencing technologies generates enormous volumes of data, making traditional analysis methods inadequate.

- **Data Standardization**: The lack of standardization in genomic data formats and annotations hinders data sharing and integration.

- **Ethical Considerations:** The use of genomic data raises ethical concerns regarding privacy, consent, and data ownership.

Machine Learning Techniques in Genomics

1. Deep Learning Applications: Deep learning models can uncover complex patterns in genomic sequences, enhancing our understanding of gene function and regulation.

2. Natural Language Processing (NLP): NLP can improve the identification of gene-disease associations and enhance literature reviews in genomic research.

3. Predictive Modeling: Predictive modeling is instrumental in identifying individuals at high risk for genetic disorders, guiding early intervention strategies.

USING MACHINE LEARNING TO DETECT NATURAL SELECTION

- Genomic data is first processed to extract key features such as allele frequencies, linkage disequilibrium, haplotype structures, and measures of genetic diversity. These features capture the evolutionary signals present in the genome.

- Traditional machine learning methods, such as random forests, support vector machines, and logistic regression, can then be trained on these features to classify genomic regions as either under selection or neutral.

- Additionally, deep learning models like RNNs can be applied directly to genomic sequences to learn complex, hierarchical patterns indicative of selection.

AI IN GENOMICS

- Artificial Intelligence (AI) is revolutionizing genomics by enabling the rapid analysis of large-scale genetic data, uncovering hidden patterns, and accelerating discoveries in precision medicine, disease prediction, and drug development.

- AI-powered algorithms, particularly deep learning and machine learning techniques, help interpret the vast complexity of genomic sequences more efficiently than traditional bioinformatics approaches.

Key Applications of AI in Genomic

1. Variant calling & genome annotation

- AI models help identify genetic variants (mutations, insertions, deletions) from raw sequencing data more accurately.
- They also predict the function of genes and regulatory regions in the genome.

2. Variant calling & genome annotation

- AI models help identify genetic variants (mutations, insertions, deletions) from raw sequencing data more accurately.
- They also predict the function of genes and regulatory regions in the genome.

3. Identifying genes under selection & evolutionary studies

- AI detects signatures of natural selection in genomes, revealing genes linked to adaptation or disease resistance.

4. Drug discovery & pharmacogenomics

- AI models predict **which genes or mutations affect drug response**, helping design more effective drugs and avoid adverse reactions.
- They can also scan genomic data to find new drug targets.

5. Understanding gene expression & regulation

- Deep learning models (like CNNs, RNNs, transformers) analyze DNA sequences to predict **which genes are turned on/off**, in which tissues, and under what conditions.

6. Cancer genomics & tumor profiling

- AI identifies mutations and molecular signatures in tumor DNA to guide diagnosis and therapy choices.

7. Single-cell genomics

- AI helps process complex single-cell RNA-seq data, clustering cells, inferring lineages, and discovering new cell types.

8. Population genomics & ancestry inference

- Machine learning infers population structure, migration patterns, and ancestry components from genomic data.

9. **Genome Sequencing**:

- It is the process of determining the exact order of the DNA bases (A, T, C, and G) in the entire genome of an organism in order to identify mutations that cause genetic diseases

10. **Disease Prediction**

- It is using data (especially genetic, clinical, or lifestyle data) to estimate a person's risk of developing a disease in the future, or to predict the likely course of a disease they already have.
- It aims to identify who is at higher risk, so that prevention, early detection, or targeted treatments can be applied.

11. Precision medicine: (also called personalised medicine) is an approach to healthcare that uses information about an individual's genes, environment, and lifestyle to tailor prevention, diagnosis, and treatment to that person.

- Instead of a "one-size-fits-all" approach, it aims to deliver the right treatment to the right person at the right time.

12. Gene editing: is a set of techniques that allow scientists to change the DNA of living organisms. It means adding, removing, or altering genetic material at specific locations in the genome.

- It's like using molecular scissors to cut the DNA at a chosen spot and then making changes, such as fixing a typo in the genetic instruction manual.